**РОССИЙСКАЯ АКАДЕМИЯ НАУК**
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ИНСТИТУТ НАУЧНОЙ ИНФОРМАЦИИ
ПО ОБЩЕСТВЕННЫМ НАУКАМ
(ИНИОН РАН)
ФУНДАМЕНТАЛЬНАЯ БИБЛИОТЕКА

**А.Б. АНТОПОЛЬСКИЙ**

# ЛИНГВИСТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

***Монография***

**МОСКВА**
**2022**

УДК 81’33; 81:022
ББК 81.1; 81 ф
А 72

Печатается по решению Ученого совета ИНИОН РАН

Научно-исследовательский отдел библиотековедения

Серия «Наука, образование и технологии»

Рецензенты:
**В.А. Плунгян**; действительный член РАН, д-р филол. наук, профессор, зам. дир. ИРЯ им. В.В. Виноградова РАН
**Л.Р. Комалова**; д-р филол. наук
**В.П. Захаров**, канд. филол. наук, доц. каф. математической лингвистики СПбГУ

**А 72** **Антопольский, А.Б.**
**Лингвистические информационные ресурсы :** монография / РАН. ИНИОН, Фундам. б-ка ; науч. ред. Д.В. Ефременко. – Москва : ИНИОН РАН, 2022. – 464 с.

**ISBN 978-5-248-01030-1**

Рассматриваются проблемы и тенденции цифровых лингвистических информационных ресурсов (ЛИР), создаваемых как в целях научных исследований, так и для прикладных задач: обработки естественного языка, обучения языкам, машинного и традиционного перевода, редактирования и др. Особое внимание уделяется вопросам типологии ЛИР, принципам каталогизации ЛИР, метаданным, форматам представления, методам разметки и аннотирования ЛИР. В качестве наиболее перспективной модели коллаборации при разработке, поддержке и развитии ЛИР предлагается платформа Семантического веба и лингвистических открытых связанных данных. Отдельные главы посвящены описанию основных категорий ЛИР – корпус, словари, лингвистические базы данных, языковые карты, образовательные ЛИР и др. Книга представляет собой справочно-аналитическое издание: кратко описываются основные объекты данной области с минимальными авторскими комментариями.

Для специалистов по проблемам научной информации и цифровизации науки, научных библиотек, преподавателей вузов, студентов и аспирантов.

УДК 81’33; 81:022
ББК 81.1; 81 ф



ISBN 978-5-248-01030-1

# ОГЛАВЛЕНИЕ

# ПРЕДИСЛОВИЕ

Предлагаемая читателю монография посвящена лингвистическим информационным ресурсам (далее – ЛИР), т.е. организованным языковым данным в цифровой форме. ЛИР в последние десятилетия стали важнейшим инструментом в самых различных компьютерных, человеко-машинных технологиях и процессах современной индустрии обработки данных и интеллектуальных систем.

Вопросы создания и использования ЛИР как важной части языковой политики становятся не только техническими, но в значительной степени социальными и политическими. Особенно важен европейский опыт управления ЛИР, потому что для Евросоюза преодоление языкового барьера при сохранении равенства языков является магистральной политической задачей.

Не претендуя на полноту, перечислим некоторые сферы применения ЛИР.

*Общие задачи:*

- терминологическая деятельность;
- переводческая деятельность;
- редакторская деятельность;
- контролируемая коммуникация с использованием ограниченного естественного языка;
- поддержка и помощь в изучении и преподавании родного и иностранных и неродных языков;
- сохранение исчезающих и находящихся в опасности языков;
- поддержка и помощь в проведении языковой политики при взаимодействии языков;
- поддержка и помощь в коммуникации для людей с ограниченными возможностями.

*Задачи в сфере ИТ и искусственного интеллекта*:

- машинный перевод;
- речевые технологии (в частности, автоматический анализ и синтез устной речи);
- голосовое общение с системами искусственого интеллекта (ИИ);
- лингвистическое обеспечение информационного поиска;
- автоматическое извлечение данных (Data Mining);
- автоматическое реферирование текстов;

- создание электронных лексикографических ресурсов;
- корпусная лингвистика (создание и использование электронных корпусов текстов);
- разработка диалоговых систем.

Широкий фронт применения ЛИР вызвал к жизни и массовое производство различных ЛИР. Действительно, в крупнейших языковых архивах счет идет на десятки, и даже сотни тысяч ЛИР. Современные технологии позволяют формировать и обрабатывать языковые корпуса в десятки и сотни миллионов слов. Однако масштабы языковой индустрии влекут и многочисленные проблемы, которые в настоящее время волнуют лингвистическое сообщество. Речь, прежде всего, идет о необходимости перехода к промышленным технологиям создания, обработки и использования ЛИР, в том числе повторного использования. Это требует внедрения методов управления, стандартизации, применения типовых программных продуктов, архивации ЛИР и т.д. В то же время значительная часть ЛИР создается в академической среде научными коллективами, которые не готовы к масштабированию этих ресурсов в промышленных объемах и не очень склонны использовать в своих исследованиях чужие разработки, несмотря на иногда очевидный экономический эффект.

В этих условиях актуальным в языковой индустрии становится создание информационной инфраструктуры, которая обеспечивала бы сохранность и повторное использование ЛИР, взаимодействие между разработчиками ЛИР, в основном учеными-лингвистами и IT-компаниями, распространение передовой практики, внедрение промышленных методов и прочее. Такая инфраструктура создается в Евросоюзе, и этот опыт следует внимательно изучить в России, тем более что вопрос о российской информационной инфраструктуре индустрии ЛИР практически не поставлен.

Несколько слов о назначении книги. Настоящая монография ориентирована не столько на профессиональных лингвистов, занимающихся созданием ЛИР, сколько на создателей информационной инфраструктуры, для которых важна общая картина информационного пространства. Книга представляет собой справочно-аналитическое издание: основная ее часть – это описание основных объектов данной сферы с небольшими авторскими комментариями.

В своей работе нам бы хотелось ответить на следующие вопросы:

- как устроено информационное пространство ЛИР;
- какие организации этим занимаются, какие проекты они реализуют;
- какие технологические решения предлагаются в настоящее время;
- какие основные ресурсы создаются в каждой категории ЛИР;
- какое место в корпоративном сообществе и общем информационном пространстве занимают российские лингвисты и российские ЛИР;
- каковы тенденции развития ЛИР.

Рассматривая тенденции развития языковой индустрии, мы пришли к выводу, что наиболее перспективным направлением является платформа лингвистических связанных открытых данных (LLOD), создаваемая в идеологии

Семантической сети. Именно платформа LLOD обеспечивает наилучшие возможности для международной коллаборации в области ЛИР. Следует также учесть, что платформа LLOD развивается в рамках открытой науки, которую автор считает перспективой развития научной коммуникации в целом. Это направление рассмотрено более подробно.

Важно также учитывать, что ЛИР, как и языковые технологии в целом, тесно связаны с развитием цифровой гуманитаристики. Поэтому нам представляется, что информационная инфраструктура должна создаваться для цифровой гуманитаристики, в рамках которой создание и развитие ЛИР будет происходить наиболее естественным образом. Здесь можно обратиться к опыту Франции и ее национальной программы TGIR Hum-Num.

Связь ЛИР с цифровой гуманитаристикой касается, например, такой важной перспективы, как создание онтологии научного знания, которая, по мнению автора, будет определять развитие как многих систем искусственного интеллекта, так и различных направлений цифровой гуманитаристики.

В заключение о структуре монографии.

Монография состоит из 21 главы, разделенных на четыре части, в которых рассмотрены соответственно: организация деятельности в области ЛИР, основные технологические аспекты создания ЛИР, отдельные категории ЛИР и перспективы их развития. В главах, посвященных категориям ЛИР, по возможности отдельно освещался российский опыт.

Использованная литература организована по главам. Интернет-ресурсы, содержащие электронные публикации, а также приравненные к ним неопубликованные документы (стандарты, методики, отчеты) отнесены к литературе. Остальные ссылки на интернет-ресурсы (сайты организаций, проектов, базы данных, программные продукты и др.) оформлены в виде подстрочных ссылок.

Большое количество используемых в тексте сокращенных наименований проектов, стандартов, ресурсов и других информационных объектов продиктовало необходимость сформировать *Указатель акронимов*. Он представляет собой алфавитный указатель акронимов и приравненных к ним наименований, причем после каждого стоит № Приложения, где имеется расшифровка акронима на русском и английском языках, адрес и, в некоторых случаях, аннотация. Русские (кириллические) сокращения приведены отдельно с расшифровкой и без ссылок.

Приложения представляют собой алфавитные перечни некоторых категорий информационных объектов: каталогов и других собраний ЛИР, организаций, работающих в этой сфере, нормативных документов, программных инструментов. При этом в приложении 4 собраны акронимы проектов, систем, технологий, моделей, непосредственно ЛИР, а также некоторые акронимы общего назначения. В отдельные приложения вынесены международные терминологические БД, а также зарубежные центры русистики.

# ЧАСТЬ 1
# ОРГАНИЗАЦИЯ ДЕЯТЕЛЬНОСТИ В СФЕРЕ ЛИР

## ГЛАВА 1. ЛИР – ОПРЕДЕЛЕНИЕ И ТИПОЛОГИЯ

### Вводные замечания

Центральным и важным при исследовании ЛИР является вопрос определения и типологии ЛИР. От этого зависит отнесение тех или иных информационных объектов к категории ЛИР. Составители многих порталов, каталогов, справочных систем, репозиториев и иных собраний ЛИР либо сведений о них придерживаются существенно различных взглядов на этот вопрос.

Можно утверждать, что существует два основных подхода к определению и типологии ЛИР.

При первом из них под ЛИР понимаются ресурсы, которые содержат языковые данные и / или непосредственно используются в языковых технологиях. Это прежде всего корпуса, лексиконы, банки синтаксических деревьев, лингвистические процессоры, описания языков и др. Назовем такой подход узким, а класс ресурсов, который относят к ЛИР сторонники этого подхода, – *специальными* ЛИР. За рубежом для этого класса ЛИР применяется также термин *языковые ресурсы* (*Language Resources*).

Второй подход определяет ЛИР более широко и включает в него не только специальные, но и любые ресурсы, создаваемые или используемыми лингвистами в профессиональной деятельности. Назовем такой подход широким, а ЛИР, которые включают в свое рассмотрение сторонники этого подхода, – *тематическими*, поскольку эти ЛИР, как правило, выделяются по тематическому принципу из универсальных или широкотематических информационных систем и ресурсов. К ним относят, например, электронные библиотеки, библиографии, труды конференций, периодику, энциклопедии, сведения о лингвистических учреждениях и персонах и тому подобные ресурсы.

В настоящей главе рассмотрим различные подходы к определению и типологии ЛИР.

### Типологии специальных ЛИР

Вначале рассмотрим узкий подход.

Англоязычная Википедия предлагает следующую типологию ЛИР:

«Важные классы языковых ресурсов включают:

1. Данные
   - лексические ресурсы, например, машиночитаемые словари;
   - лингвистические корпуса, т.е. цифровые коллекции данных на естественном языке;
   - лингвистические базы данных, такие как коллекция кросс-лингвистических связанных данных.
2. Инструменты
   - лингвистические аннотации и инструменты для создания таких аннотаций в ручном или полуавтоматическом режиме (например, инструменты для аннотирования подстрочного сглаженного текста, такие как Toolbox и FLEx, или другие инструменты языкового документирования);
   - приложения для поиска и извлечения таких данных (системы управления корпусом), для автоматического аннотирования (разметка частей речи, синтаксический анализ, семантический анализ и т.д.).
3. Метаданные и словари
   - словари, репозитории лингвистической терминологии и языковых метаданных, например, META-SHARE (для метаданных языковых ресурсов), реестр категорий данных ISO 12620 (для лингвистических функций, структур данных и аннотаций в языковом ресурсе) или база данных Glottolog (идентификаторы для языковых разновидностей) и библиографическая база данных»[1].

Среди имеющихся предложений по типологии ЛИР Википедия упоминает LREMap, META-SHARE и, для данных, классификацию LLOD. Приведем описания этих типологий ЛИР.

***Типология LLOD***[2] (подробно о LLOD см. главу 19):

- корпуса
- лексиконы и словари
- терминологические ЛИР, тезаурусы, базы знаний
- метаданные ЛИР
- категории лингвистических данных
- типологические базы данных
- другие

Одна из наиболее авторитетных организаций в области ЛИР – это *ELRA (Европейская ассоциация лингвистических ресурсов)*[3], которая разработала ряд сервисов для индустрии ЛИР.

[1] Языковой ресурс – Language resource. – URL: https://ru.qaz.wiki/wiki/Language_resource (дата обращения: 01.12.2021).

[2] Linguistic Linked Open Data. – URL: http://linguistic-lod.org/ (дата обращения: 01.12.2021).

[3] European Language Resources Association. – URL: http://www.elra.info/en/ (дата обращения: 01.12.2021). Подробное описание ELRA см. в гл. 4.

Один из основных сервисов ELRA – *META-SHARE*[1] – открытая инфраструктура, включающая сеть репозиториев для обмена языковыми данными, инструментами и связанными с ними веб-сервисами. В META-SHARE используется типология, включающая два фасета: тип ЛИР и тип Медиа. Этот подход позволяет использовать в обоих фасетах достаточно общую классификацию.

Тип ЛИР:

- корпуса (включая письменные / текстовые, устные / речевые, мультимодальные / мультимедийные корпуса);
- лексические / концептуальные ресурсы (включая терминологические ресурсы, списки слов, семантическую лексику, онтологии и т.д.);
- языковая документация (включая грамматики);
- инструмент / сервис (включая базовые средства обработки, приложения, веб-сервисы и т.д., необходимые для обработки информационных ресурсов).

Тип Медиа:

- текст
- аудио
- изображение
- видео
- textNumerical
- textNgram

Подробное описание системы метаданных META-SHARE представлено в документе [1], а также будет дано ниже в главе 6, посвященной метаданным ЛИР.

Заметим, что сведения о ЛИР, которые мы называем тематическими, например, публикации, в системе META-SHARE рассматриваются как дополнительные сведения, характеризующие ЛИР, а не как самостоятельные ЛИР.

***Карта оценочного описания ЛИР (LRE map)***[2], предназначенная для осуществления мониторинга ЛИР, также разработана в ELRA. В этой карте выделено три основных типа ЛИР (данные, документация и инструменты). Списки видов для этих типов ЛИР приводятся в гл. 2.

В рамках ELRA создана также служба идентификации ЛИР – *Международный стандартный номер ЛИР (ISLRN)*[3], где используется типология ЛИР, принятая в OLAC.

В ELRA имеются также собственные каталоги ЛИР. В основном каталоге[1] выделяется всего четыре типа ЛИР:

---

[1] META-SHARE (Search & exchange language resources). – URL: http://www.meta-share.org/ (дата обращения: 01.12.2021).

[2] LRE (Linguistic Resource Evaluation) Map. – URL: http://lremap.elra.info/ (дата обращения: 01.12.2021). См. также гл. 6.

[3] International Standard Language Resource Number (ISLRN). – URL: http://www.elra.info/en/islrn/ (дата обращения: 01.12.2021).

- корпуса
- лексика и концептуальные ЛИР
- инструменты и сервисы
- языковая документация

Кроме основного каталога в ELRA имеется каталог ЛИР научно-исследовательского назначения[2], где предлагается иная типология:

1. Устные ЛИР
   - телефонные записи
   - микрофонные записи
   - вещательные ресурсы
   - фонетические ресурсы
2. Письменные ЛИР
   - корпуса
   - одноязычные лексиконы
   - многоязычные лексиконы
3. Терминологические ЛИР.
4. Мультимодальные / мультимедийные ЛИР.

Приведем еще несколько типологий ЛИР, в аспекте узкого подхода к проблеме.

Типология ЛИР авторитетной европейской инфраструктуры *CLARIN*[3] включает следующие семейства ЛИР:

**Корпуса**

- корпуса компьютерных сетей
- корпуса научных текстов
- исторические корпуса
- корпуса учебных текстов
- литературные корпуса
- аннотированные вручную корпуса
- мультимедийные корпуса
- газетные корпуса
- параллельные корпуса
- парламентские корпуса
- справочные корпуса
- корпуса устной речи

**Лексические ресурсы**

- лексика
- словари
- концептуальные ресурсы
- глоссарии
- списки слов

---

[1] 1412 Language Resources (Page 1 of 71) // ELRA. – URL: http://catalog.elra.info/en-us/repository/search/?q= (дата обращения: 01.12.2021).

[2] R&D Catalogue of Language Resources // ELRA Home Catalogue. – URL: http://catalogue-old.elra.info/retd/ (дата обращения: 01.12.2021).

[3] CLARIN. – URL: https://www.clarin.eu/ (дата обращения: 01.12.2021).

**Инструменты**

- нормализация
- распознавание именованных сущностей
- маркировка и лемматизация частей речи
- инструменты для анализа эмоционального восприятия

Известная лингвистическая сеть *OLAC (Консорциум открытых лингвистических архивов)* использует систему метаданных Дублинского ядра (DC)[1]. При этом для типов ресурсов DC предлагается расширение, включающее всего три квалификатора. В результате типология ЛИР в этом каталоге представлена следующим образом[2]:

- лексиконы
- первичные тексты
- языковая документация

Существуют локальные типологии ЛИР, ориентированные на определенные программные продукты. Например, известная система *Обшая архитектура обработки текста* (*GATE)*[3] содержит три типа ЛИР (данных): документы, корпуса и графы аннотаций.

*Document / Blank Document* – документ Gate, загруженный из файла или пустой. Новый документ создается через Language Resources > New > Gate Document. Документ можно сохранить в формате XML.

*Gate Corpus* – корпус для хранения документов. Корпус создается через Language Resources > New > Gate Corpus. Наполнить корпус можно, указав список документов при создании, или добавив документы в интерфейсе уже созданного корпуса, или с помощью команды Populate. Корпус можно сохранить в XML.

*Аннотации* организованы в виде графов, которые моделируются как Java-наборы. Аннотации представлены в виде дуг с начальным и конечным узлами, ID, присвоенным типом и FeatureMap (набором объектов). Узлы содержат указатели на источники в документе.

## Типологии в рамках широкого подхода к ЛИР

Наиболее полная типология ЛИР как специальных, так и тематических представлена в популярном ресурсе ***LINGUIST List***[4]. Основные разделы этого ресурса выглядят следующим образом:

- люди и организации
- вакансии

[1] OLAC Metadata. – URL: http://www.language-archives.org/OLAC/metadata.html (дата обращения: 01.12.2021).

[2] OLAC Linguistic Data Type Vocabulary. – URL: http://www.language-archives.org/REC/type.html (дата обращения: 01.12.2021).

[3] General Architecture for Text Engineering (GATE). – URL: https://gate.ac.uk/ (дата обращения: 01.12.2021). Подробнее см. в гл. 9.

[4] LINGUIST List. – URL: https://linguistlist.org/ (дата обращения: 01.12.2021).

- конференции и другие мероприятия
- публикации
- языковые ресурсы
- словари
- языки
- области лингвистики
- лингвистические компьютерные средства

Еще один пример широкой типологии ЛИР – это ***Метаиндекс лингвистики, естественного языка и компьютерной лингвистики*****,** созданный в Стэнфордском университете[1]. Он включает следующие типы ЛИР:

- лингвистические теории и области
- списки конференций по лингвистике
- лингвистические журналы и другие материалы в Интернете
- онлайн-журналы открытого доступа
- онлайн-библиографии
- лингвистические общества
- грамматики и словари
- избранные языки
- кафедры и программы компьютерной лингвистики
- компании

***Навигатор информационных ресурсов по языкознанию* (*НИРЯЗ*)**[2], разработанный при участии автора, в отличие от большинства каталогов ЛИР включает не только цифровые, но и бумажные ЛИР, в частности библиотечные фонды, архивные и музейные документы. НИРЯЗ включает около 1,2 тыс. ЛИР, созданных в учреждениях РАН.

Сокращенная типология ЛИР этого каталога выглядит следующим образом:

- библиотеки
- архивы
- музеи
- каталоги
- электронные коллекции и библиотеки
- информационные системы
- справочники, энциклопедии
- персональные ресурсы
- лингвистические ресурсы
  - корпуса текстов
  - словарные БД и электронные картотеки
  - лингвистические процессоры
  - грамматические ресурсы

[1] Linguistics, Natural Language, and Computational Linguistics Meta-index. – URL: https://nlp.stanford.edu/links/linguistics.html (дата обращения: 01.12.2021).

[2] Навигатор информационных ресурсов по языкознанию. – URL: http://niryaz2.alexo.beget.tech/ (дата обращения: 01.12.2021).

    - описания языков, реестры языков
    - лингвистические атласы
    - этно- и социолингвистические БД
    - комплексные лингвистические АИС (сайты)
    - информационные языки
- периодика
- библиографии
- мероприятия
- неопубликованные материалы
- медиаресурсы
- прочие интернет-ресурсы

Легко видеть, что здесь специальные ЛИР выделены в отдельный тип, остальные типы ЛИР выделены по тематическому принципу.

Приведем еще несколько примеров широкого подхода к типологии ЛИР в некоторых российских каталогах. Иногда классификация ЛИР приводится с сокращениями.

***NLPub – каталог ресурсов для обработки естественного языка***[1]

Методы и инструменты

- Обработка текста
    - графематический анализ
    - морфологический анализ
    - синтаксический анализ
    - проверка правописания
    - расстановка переносов
    - построение конкордансов
    - извлечение ключевых слов
    - автоматическое реферирование
    - тематическая классификация
    - тематическое моделирование
    - извлечение именованных сущностей
    - извлечение отношений
    - анализ тональности
    - информационный поиск
    - машинный перевод
    - обнаружение дубликатов
    - сегментация текста
    - интегрированные пакеты
- Обработка речи
- Утилиты
    - конечный преобразователь
    - обработка языковых моделей
    - редактор тезауруса
    - анализ текстовых корпусов

[1] NLPub. – URL: https://nlpub.ru/

- Методы
  - варианты категориальной грамматики
  - варианты (типизированного) лямбда-исчисления и линейная логика
  - варианты с использованием комбинаторной логики
  - связи с алгеброй, теорией категорий, теорией игр
- Алгоритмы
  - языковые модели
  - морфологический анализ
  - синтаксический анализ
  - извлечение именованных сущностей
  - извлечение ключевых слов
  - автоматическое реферирование
  - кластеризация графов
  - генерация текста
  - алгоритмы общего назначения
- Ресурсы
  - словари
  - тональный словарь
  - тезаурусы
  - корпусы
  - коллекции n-грамм
  - банки данных
  - размеченные коллекции изображений
  - журналы
- Эксперты и мероприятия
- Образование
- Проекты

***Компьютерная лингвистика. Портал знаний***[1]

В этом проекте классификация ЛИР разработана наиболее подробно и фундаментально; фактически построена – онтология понятий, относящихся к компьютерной лингвистике. Описание проекта можно найти в работе [2].

Приведем верхние уровни этой классификации и полностью раздел, непосредственно касающийся специальных ЛИР.

I. Деятельность – проекты
II. Интернет-ресурсы
  - Информационные ресурсы
  - Сайты организаций, персон, проектов

III. Методы и средства исследования
IV. Научные результаты и продукты
  - Лингвистические ресурсы

[1] Компьютерная лингвистика. Портал знаний. – URL: https://uniserv.iis.nsk.su/cl/ (дата обращения: 01.12.2021).

- корпуса
  - корпуса текстов
  - речевые корпуса
- лингвистические БД
  - грамматические ресурсы
  - лексико-семантические ресурсы
  - морфологические БД
  - речевые БД
  - семантико-синтаксические ресурсы
  - синтаксические ресурсы
- онтологии
- словари и тезаурусы

- Прикладные системы
- Технологии и программные продукты

V. Объекты исследования

VI. Структурные языковые единицы

***Металингвистическая БД С. Крылова***

Оригинальным проектом по типологии лингвистических знаний является работа известного российского лингвиста С.А. Крылова, которую он назвал металингвистической БД и которая размещена на информационном портале Starling[1]. Процитируем С.А. Крылова:

«Металингвистические базы данных (МБД), служат инструментом систематизации знаний о лингвистике (а не напрямую о языке), однако косвенно способствуют также систематизации сведений о языке. Можно выделять две разновидности МБД:

(1) метанаучные (МН-) МБД (входы в которые являются металингвистическими проекциями научных текстов по лингвистике) и

(2) метаобъектные (МО-) МБД (входы в которые являются металингвистическими проекциями языковых сущностей).

Входами в МО-МБД служат, например, характеристики языковых общностей (лингвонимические, этнонимические, топонимические, хронологические); нарицательные лингвистические термины; имена языковых единиц (в том числе имена таксономических классов внеязыковых сущностей).

Следует прежде всего проводить различие между онтологическим (материальным) уровнем, на котором можно выделить объектное множество (оригинал, универсум) с существующими в нем отношениями, и гносеологический (эпистемологический, идеальный) уровень, на котором выделяется модельное множество (модель, теория) с заданными на нем отношениями. Эту модель и строит металингвист, воплощающий ее в виде грамматики, словаря, предметного или именного указателя, таблицы, графа, дерева, карты, атласа, базы данных и т.п.» [3].

[1] Вавилонская башня. Проект «Эволюция языка». – URL: https://starling.rinet.ru/program.php?lan=ru (дата обращения: 01.12.2021).

В данной работе предлагается развернутая система понятий, представляющих предметную область; мы приводим верхние уровни этой классификации.

I. *Универсум языковых явлений*
IA. Общелингвистический универсум
(IA. 1.) Мир языковой системы
(IA. 1.0.) Языковая система и ее подсистемы
(IA. 1.1.) Языковые единицы (ЯЕ)
(IA. 1.2.) Отношения между ЯЕ
(IA. 1.3.) Члены отношений между ЯЕ
(IA. 1.4.) Функции ЯЕ
(IA. 1.5.) Способы выражения значений
(IA. 1.6.) Классы ЯЕ
(IA. 1.7.) Члены классов ЯЕ
(IA. 1.8.) Языковые структуры
(IA. 1.9.) Части языковых структур
(IA. 1.10.) Языковые процессы
(IA. 1.11.) Логические связи языковых явлений
(IA. 2.) Речевая динамика
(IA. 3.) Речевая способность (типы, аспекты и компоненты)
(IA. 4.) Речевое варьирование (типы и проявления)
(IA. 5.) Языковое функционирование (типы)
(IA. 6.) Языковые изменения (типы, аспекты и компоненты)
(IA. 7.) Языковые сходства и различия (типы)
(IA. 8.) Исторические отношения между языковыми общностями
IБ. Частнолингвистический универсум
(IБ. 1.) Универсум исторических языковых общностей
(IБ. 2.) Универсум ареалов распространения языков: континенты, регионы, страны, населенные пункты
(IБ. 3.) Универсум частнолингвистических единиц
IB. Универсум речевых событий
(IB1.) Универсум словесности (множество текстов)
(IB1.2.) Универсум памятников письменности
(IB1.3.) Универсум высказываний
IB2. Универсум вхождений речевых знаков-экземпляров (tokens)
II. *Универсум собственно лингвистики*
(II. 1) лингвисты (в том числе лингвисты-непрофессионалы)
(II. 2) Лингвистические школы и направления
(II. 3.) Лингвистические кружки, общества, ассоциации и т.п.
(II. 4) Места, где протекает деятельность лингвистов (континенты, страны, провинции, населенные пункты)
(II. 5) Учреждения, где протекает деятельность лингвистов
(II. 6.) Универсум лингвистических работ
III. *Мир лингвистических моделей*
III. 1. Описания языков (словари, грамматики и т.п.)

III. 2. Описания речевых отрезков: транскрипции, хрестоматии текстов, издания памятников, продукты транскрипции и транслитерации, переводы текстов, фонетические сонограммы, комментарии, глоссы, формальные представления текстов в виде морфологических и синтаксических «разборов», синтаксических графов (в частности, деревьев зависимостей и составляющих), цепочки трансформационного вывода, толкования отдельных примеров и т.п.

III. 3. Описания ЯЕ: словарные статьи, правила, законы, исключения к правилам и т.п.

К сожалению, этот проект не получил продолжения, и его результаты не используются при разработке российских ЛИР.

Приведем еще несколько описаний российских порталов, в которых систематизированы ссылки на ЛИР на основе более или менее подробных классификаций.

***Информационные ресурсы по лингвистике***[1]

На портале, созданном И.П. Сусовым из Тверского государственного университета, собрано большое количество (около 400) ссылок на ЛИР, который автор понимает весьма широко. Ниже приводятся разделы этого портала.

СИСТЕМАТИЗИРОВАННАЯ ЛИНГВИСТИЧЕСКАЯ ИНФОРМАЦИЯ ON-LINE

- Журналы лингвистического профиля
- Библиотеки, каталоги и архивы лингвистического профиля
- Лингвистические справочники и энциклопедии
- Прочие лингвистические ресурсы в Интернете
- Подписка на рассылку лингвистических сообщений (linguistics mailing lists)

МАТЕРИАЛЫ К УЧЕБНЫМ КУРСАМ ПО ЛИНГВИСТИКЕ

- Общие курсы по теоретическому языкознанию и смежным наукам
- Материалы к истории языкознания

ИССЛЕДОВАТЕЛЬСКИЕ КОЛЛЕКТИВЫ

- Зарубежные лингвистические учреждения и коллективы
- Лингвистические учреждения и коллективы в России и СНГ
- Объединения (ассоциации) и конференции по проблемам языка и речи

ИНДИВИДУАЛЬНЫЕ ИССЛЕДОВАТЕЛИ

- Зарубежные исследователи языка и речи

(адреса, домашние страницы, персоналии, личные архивы, публикации)

- Отечественные лингвисты в Интернете

(домашние страницы, персоналии, личные архивы, публикации)

ДОПОЛНИТЕЛЬНЫЕ РАЗДЕЛЫ

- Лингвистическая гостиная
- Каталог языков мира (более 6700 языков)

[1] Информационные ресурсы по лингвистике. – URL: http://homepages.tversu.ru/~ips/InfoSeek.htm (дата обращения: 01.12.2021).

- Каталог языковых семей
- Индоевропейские языки
- Неиндоевропейские языки
- Словари, тезаурусы и переводчики on-line
- Общие и специальные энциклопедии
- Газеты на германских языках
- Газеты на романских языках
- Газеты на славянских языках
- Российские и зарубежные библиотеки
- В мире книг и журналов
- Новости и обзоры

***Лингвистика***[1]

Портал ссылок по филологии и лингвистике включает следующие разделы сайта по лингвистике:

- предмет лингвистики – лингвистика в энциклопедиях и словарях
- порталы и каталоги ссылок о лингвистике
- известные лингвисты в Сети
- лингвистические журналы
- статьи по лингвистике
- учебные кафедры лингвистики
- лингвистические научные центры
- лингвистическая экспертиза

На портале имеется указатель материалов сайта по разделам лингвистики, расположенных по алфавиту понятий:

**Г** Гендерная лингвистика // Генеративная лингвистика // Грамматика

**З** Зарубежная лингвистика

**И** Известные лингвисты // Интерлингвистика // История лингвистических учений // История языков

**К** Когнитивная лингвистика // Компьютерная лингвистика // Контрастивная лингвистика // Корпусная лингвистика

**Л** Лексикография // Лексикология // Лингвистика текста // Лингвистические методы // Лингвистические учения // Лингводидактика // Лингвисты // Литературоведение

**Н** Нейролингвистика

**О** Общее языкознание // Ономастика // Онтолингвистика

**П** Паралингвистика // Переводоведение // Прикладная лингвистика // Психолингвистика

**С** Семантика // Семиотика // Социолингвистика // Сравнительно-историческое языкознание // Стилистика // Структурная лингвистика

**Т** Теоретическая лингвистика // Терминоведение

**Ф** Филология // Фонетика // Фразеология

[1] Лингвистика © Юрий Новиков (2009–2021). – URL: http://filologia.su/lingvistika (дата обращения: 01.12.2021).

**Э** Этимология
**Я** Языкознание сравнительно-историческое

***Каталог лингвистических программ и ресурсов в Сети***[1]

Данный каталог включает в себя описание программ, связанных с анализом текстов и вычислительной лингвистикой, а также соответствующих ресурсов, доступных сегодня в глобальной сети Интернет. Упор при составлении каталога делался на бесплатные программы, доступные для загрузки или использования в режиме on-line. Также описаны коммерческие версии некоторых наиболее интересных программ. Тематически каталог разбит на следующие разделы:

- программы анализа и лингвистической обработки текстов
- программы преобразования текстов
- психолингвистические программы
- генераторы текстов
- системы обработки естественного языка и машинного перевода
- каталоги и коллекции ресурсов
- словари и тезаурусы
- поисковые машины и системы полнотекстового поиска
- системы синтеза и распознавания речи

***Лингвистические ресурсы в Интернете***[2]

КОРПУСА
- Славянские языки
- Другие языки

СЛОВАРИ
- Одноязычные словари
- Двуязычные и многоязычные словари

ДРУГИЕ РЕСУРСЫ
- Информационные сайты и рассылки
- Архивы
- Блоги
- Отдельные проекты

Из приведенных примеров очевидно, что общего подхода к типологии ЛИР ни в России, ни в мире нет, хотя пересечения типов ЛИР весьма велики. В дальнейшем наш анализ будет в основном сосредоточен на специальных ЛИР. Соответственно в следующем разделе будут рассмотрены собрания специальных ЛИР, а если в собрании применяется широкий подход – то соответствующие разделы таких собраний.

[1] Каталог лингвистических программ и ресурсов в Сети. – URL: https://rvb.ru/soft/catalogue/index.html (дата обращения: 01.12.2021).

[2] Лингвистические ресурсы в Интернете. – URL: http://rusling.narod.ru/q_res.htm

## Литература к главе 1

1. Documentation and User Manual of the META-SHARE Metadata Model / E. Desipri [et al.] ; ed. : P. Labropoulou, E. Desipri // META-NET. – URL: http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf (дата обращения: 01.12.2021).

2. Разработка портала знаний по компьютерной лингвистике / О.И. Боровикова, Ю.А. Загорулько, Г.Б. Загорулько, И.С. Кононенко, Е.Г. Соколова // Труды 11-й национальной конференции по искусственному интеллекту с международным участием КИИ – 2008 (г. Дубна, Россия). – Москва : ЛЕНАНД, 2008. – Т. 3. – С. 380–388.

3. Крылов С.А. Из каких элементов состоит метаязык лингвистики? // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.) / гл. ред. А.Е. Кибрик. – Москва : Изд-во РГГУ, 2010. – Вып. 9(16). – С. 248–253.

# ГЛАВА 2. СОБРАНИЯ ЛИР

## Введение

В мировом Интернете имеется достаточно много различных собраний ЛИР, либо сведений о них. Собрания могут быть следующих видов:

- порталы, содержащие ссылки на ЛИР
- каталоги и навигаторы, содержащие, кроме ссылки, минимальные сведения о ЛИР
- поисковые и справочные системы, где возможен поиск описаний и / или ЛИР по различным критериям и с использованием различных фильтров
- архивы и репозитарии ЛИР, где собраны не только описания ЛИР, но и сами ресурсы

Всего нами обнаружено свыше 160 таких собраний, перечень которых приводится в Приложении 1. Значительная часть этих собраний включена в регистр лингвистических архивов OLAC[1], перечень лингвистических мета-сайтов Linguist list[2] или каталог репозиториев научных данных RE3[3]. Однако эти перечни существенно пересекаются, поэтому мы сочли полезным сделать общий список.

Российские каталоги и порталы ЛИР представлены в приложении 2. Заметим, что в 2012 году Д. Усталов опубликовал первый анализ российских каталогов ЛИР [1], причем в сферу его рассмотрения вошло всего пять каталогов. Сейчас их значительно больше, в нашем списке их свыше 40, причем в этот список не вошли каталоги образовательных ресурсов по русскому языку. Этот класс ЛИР, включая их каталоги, рассмотрен отдельно в главе 17.

В России создан пока единственный архив ЛИР, а именно архив корпусов уральских и алтайских языков на платформе ЛингвоДок. Его описание приводится в конце настоящей главы. Еще некоторые российские интеграционные проекты рассмотрены в главах, посвященных отдельным категориям ЛИР.

---

[1] Participating Archives // OLAC: Open Language Archives Community. – URL: http://www.language-archives.org/archives (дата обращения: 01.12.2021).

[2] Language Meta Sites // The Linguist Llist: International Linguistics Community Online. – URL: https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=25 (дата обращения: 01.12.2021).

[3] Registry of research data repositories. – URL: https://www.re3data.org/ (дата обращения: 01.12.2021).

Особый тип ЛИР представляют терминологические базы и банки данных (ТБД). Их каталог можно найти, например, по адресу[1]. Специфика ТБД заключается в том, что их создают не столько лингвистические, сколько международные организации – отраслевые или универсальные (ООН, ЕС, ISO. ФАО и др.). ТБД используются в основном для переводческой и редакторской деятельности; они рассмотрены в главе 12.

Далее приводятся более подробные описания нескольких наиболее известных каталогов и архивов ЛИР. В отдельных разделах представлены мировые и европейские собрания. Российский архив ЛингвоДок отнесен к европейским собраниям.

## Мировые собрания ЛИР

### *Связанные лингвистические открытые данные LLOD*

Центральным способом сбора, архивации и интеграции ЛИР и организации эффективных коллабораций в этой области являются, по мнению автора, платформа Семантического веба и основанный на ней проект связанных лингвистических открытых данных LLOD[2]. В связи с важностью и перспективностью этой платформы ее описание приводится отдельно, в главе 19.

### *Сообщество открытых языковых архивов OLAC*

Наиболее полным собранием ЛИР является собрание языковых архивов OLAC[3].

OLAC представляет собой международное партнерство учреждений и частных лиц, которые создают всемирную виртуальную библиотеку ЛИР. OLAC решает две задачи:

- выработки консенсуса в отношении лучшей современной практики цифрового архивирования ЛИР
- развития сети взаимодействующих хранилищ, служб для обеспечения сохранности ресурсов и доступа к ним

Архивы, входящие в OLAC (их в апреле 2021 г. было 63), в совокупности содержат свыше 400 тыс. ЛИР.

Каталог в OLAC обеспечивает поиск и сортировку найденных ЛИР по следующим параметрам:

- по языкам и семействам языков
- по наличию ЛИР в онлайне
- по странам и регионам
- по наименованию архива
- по типу ЛИР (см. гл. 1)

---

[1] Terminology websites & blogs // Terminology coordination. – URL: https://termcoord.eu/terminology-websites (дата обращения: 01.12.2021).

[2] Linguistic Linked Open Data (LLOD) Cloud. – URL: https://linguistic-lod.org/llod-cloud (дата обращения: 01.12.2021).

[3] OLAC: Open Language Archives Community. – URL: http://www.language-archives.org/ (дата обращения: 01.12.2021).

- по типу дискурса
- по области лингвистики
- по типу ЛИР по Дублинскому ядру метаданных
- по формату
- по предметным рубрикам Библиотеки Конгресса
- и еще по ряду признаков

Представляет интерес перечень областей лингвистики, к которым отнесены ЛИР, представленные в OLAC. Приведем список этих областей. В скобках указано количество ЛИР, относящихся к данной области.

- Антропологическая лингвистика (1970)
- Прикладная лингвистика (185)
- Когнитивная лингвистика (22)
- Компьютерная лингвистика (2305)
- Дискурс-анализ (348)
- Судебная лингвистика (45)
- Общее языкознание (6078)
- Историческая лингвистика (152)
- Изучение языка (390)
- Документирование языков (25 067)
- Лексикография (5051)
- Лингвистические теории (28)
- Языкознание и литературоведение (1)
- Математическая лингвистика (29)
- Морфология (1926)
- Нейролингвистика (76)
- Фонетика (5167)
- Фонология (3309)
- Прагматика (7)
- Психолингвистика (36)
- Семантика (2690)
- Социолингвистика (543)
- Синтаксис (5309)
- Корпусная лингвистика и лингвистика текста (22 277)
- Письменный и устный перевод (378)
- Типология (6424)
- Системы письменности (3049)

Система OLAC теперь интегрирована с облаком лингвистических связанных открытых данных (LLOD). Это открывает путь для того, чтобы содержимое 63 участвующих архивов было взаимосвязано и доступно для поиска и использования.

OLAC содержит детальный регистр участвующих архивов ЛИР[1]. Каждый архив снабжен подробной анкетой. Приведем содержание анкеты:

[1] Participating Archives // OLAC: Open Language Archives Community. – URL: http://www.language-archives.org/archives (дата обращения: 01.12.2021).

- объем (количество ЛИР)
- название репозитория
- учреждение
- URL архива
- местонахождение
- краткое местоположение
- синопсис
- условия доступа
- администратор
- участники
- базовый URL
- идентификатор репозитория
- OAI-версия
- OLAC-версия
- записи в архиве
- фасетный поиск
- последнее пополнение
- дата сведений
- дата последней коррекции
- отчеты

Кроме того, в регистре содержатся сравнительные количественные данные об архивах, включая сведения об использовании каждого элемента метаданных, изложения правил или политики депонирования ЛИР в отдельных архивах. Данные регистра имеются также в виде XML-файла.

Общая статистика по архивам OLAC по состоянию на октябрь 2021 г. приводится в таблице 1.

*Таблица 1*

**Статистика архивов OLAC**

| Показатель | Значение |
|---|---|
| Количество архивов | 62 |
| Архивы с обновляемыми каталогами | 28 |
| Архивы с пятизвездочными метаданными | 21 |
| Количество ресурсов | 446 070 |
| Количество ресурсов, доступных онлайн | 391 872 |
| Различные языки | 8157 |
| Различные лингвистические подполя | 28 |
| Различные лингвистические типы | 3 |
| Различные типы DCMI | 12 |
| Среднее число элементов в записи | 18,5 |
| Среднее число схем кодирования на запись | 7,5 |
| Средний показатель качества метаданных | 6,8 |
| Число просмотренных записей в месяц | 8608 |
| Клики в месяц | 2172 |
| Последнее обновление | ежедневно |

***Справочная система для поиска информации об языковых ресурсах Linghub***[1]

В этой системе содержится информация о более чем 100 тыс. ЛИР по свыше 1000 языков. Система создана на основе объединения данных VLO CLARIN, META-SHARE, LRE Map, DATAHUB.

Стандартное описание ЛИР в *Linghub* включает следующие реквизиты:
*создатель* – Contributor
*описание* – Description
*права доступа* – Rights
*источник* – Source
*предмет* – Subject
*наименование* – Title

Пользователю предлагаются списки значений следующих полей и поиск по ним (при всех значениях указывается количество ЛИР, списки упорядочены по частоте встречаемости):
*язык* – список
*права доступа* – перечислены различные ограничения
*тип ЛИР* – перечислены типы ЛИР, включая данные и инструменты
*создатель* – перечисляются лица и организации
*источник* – перечисляются источники
*поставщик* – перечисляются лица и организации
*предмет* – перечисляются жанры документов, тематические области и др.

***Каталог Консорциума лингвистических данных LDC***[2]

Каталог LDC включает по данным на февраль 2021 г. около 900 ЛИР, по большей части корпусов и лексиконов, созданных в научных или исследовательских целях в университетах, входящих в Консорциум.

Поиск в каталоге LDC предлагается в двух вариантах:
Свободный поиск по параметрам
- названию ЛИР или публикации
- автору
- номеру в каталоге
- ключевым словам

С просмотром допустимых значений по параметрам
- языку
- году создания ЛИР
- типу ЛИР по DCMI
- источнику данных
- проекту, в рамках которого создан ЛИР
- назначению ЛИР

В каталоге имеются подробные правила использования ЛИР, доступных через каталог LDC с учетом лицензионных соглашений, членства в

[1] Linghub. – URL: http://linghub.org/ (дата обращения: 01.12.2021).
[2] LDC Catalog. – URL: https://catalog.ldc.upenn.edu/ (дата обращения: 01.12.2021).

LDC или на коммерческих условиях. LDC также предоставляет набор программных инструментов.

***Архив языков и культуры SIL***[1]

SIL – это глобальная некоммерческая организация, которая работает с местными сообществами по всему миру над разработкой языковых решений для улучшения жизни. Подробнее о ней – см. главу 3.

Среди проектов SIL важное место занимает архив языков и культур. Коллекция была организована в 1947 году как средство отслеживания и публикации «корпоративной библиографии». Девять изданий «SIL Bibliography» и пять приложений были напечатаны в период с 1948 по 1992 год. «Библиография» была опубликована в Интернете начиная с 1997 года. Данный архив включает как специальные ЛИР, так и тематические, прежде всего в виде публикаций.

Основная часть коллекции находится в Далласе, но некоторые физические ресурсы разбросаны по всему миру в библиотеках и офисах SIL. Значительная часть коллекции все еще требует каталогизации и / или оцифровки, прежде чем ее можно будет опубликовать в Интернете. Онлайн доступно в настоящее время 48 тыс. объектов. Возможен поиск по следующим признакам:

- условия доступа к ЛИР
- язык документа
- создатель ЛИР
- страна
- тематика изучения
- тип ЛИР
- источник
- предметная область ЛИР
- код языка
- язык как предмет
- дата создания ЛИР

***Архив исчезающих языков ELAR***[2]

Это цифровой архив, хранящий и публикующий мультимедийные коллекции исчезающих языков. В архиве собраны коллекции со всего мира с региональными опорными пунктами в Африке, на Ближнем Востоке, в Азии, Австралии и Латинской Америке. На сегодняшний день в ELAR можно найти записи, охватывающие более 450 языков. Коллекции в ELAR содержат аудио- и видеозаписи обыденного использования языка, словесного искусства, песен, рассказов, ритуалов и многое другое. Коллекции также содержат словари, педагогические материалы, такие как буквари для преподавания языка, транскрипции и переводы записей на основные контактные языки, такие как

[1] Language & Culture Archives // SIL. – URL: https://www.sil.org/resources/language-culture-archives (дата обращения: 01.12.2021).

[2] Endangered Languages Archive. – URL: https://www.elararchive.org/ (дата обращения: 01.04.2022).

испанский, мандаринский, английский или русский. Эти коллекции можно просмотреть и получить доступ к ним через онлайн-каталог ELAR. Все материалы являются цифровыми и доступны бесплатно (после бесплатной регистрации).

Миссия ELAR состоит в том, чтобы:

- обеспечить безопасное долгосрочное хранилище коллекций языковой документации;
- обучать и поддерживать участников и партнеров в создании и сохранении коллекций;
- сделать коллекции бесплатными для исследователей, сообществ и общественности;
- помочь пользователю в поиске записей и доступе к ним.

***Проект архивирования лингвистических данных LACITO[1]***

Целью проекта архивирования лингвистических данных LACITO является сохранение и распространение речевых данных. С этой целью были разработаны нормы подготовки и использования документов, включающих звук и текст, с использованием международно признанных стандартов, в частности SGML (Standard Generalized Markup Language).

Основным источником данных для проекта является множество документов, записанных и расшифрованных в полевых условиях членами LACITO за последние тридцать лет. Эти уникальные записи, в основном спонтанной речи на бесписьменных языках, служат основой для исследований соответствующих языков и культур. Некоторые из транскрипций и переводов были опубликованы, но оригинальные звуковые записи никогда не публиковались и не архивировались должным образом. Документы, подготовленные в рамках проекта, включают в себя как звук, так и текст – как минимум фонологическую транскрипцию и свободный перевод, а также, где это возможно, пословные глоссы, примечания и т.д. Текст индексируется по звуку на уровне «предложения» или интонационной конструкции. Доступ к документам можно получить либо локально на компакт-диске, либо по Сети.

Для текстовых материалов была принята разметка XML. Для кодирования символов использовался Unicode.

Формат звукового файла, используемый в проекте, – RIFF (WAV). Это формат Windows, но он может быть использован на других платформах или преобразован в другие форматы. В проекте используется оцифровка на частоте 44,1 кГц с разрешением 16 бит, стерео или моно, в зависимости от исходной записи (обычно моно). Эти параметры, возможно, чрезмерны, учитывая качество оригинальных записей, но они были выбраны, чтобы избежать дальнейшего ухудшения часто незаменимых документов.

[1] Langues et Civilisations à Tradition Orale (LACITO). Linguistic Data Archiving Project. – URL: http://xml.coverpages.org/lacitoAR-desc-english.html#Resume (дата обращения: 01.12.2021).

## Европейские собрания ЛИР

***Европейская координация языковых ресурсов ELRC***[1]

Данный проект ЕС осуществлялся в 2014–2017 гг. и предусматривал сбор языковых ресурсов для машинного перевода в рамках программы Connecting Europe Facility (CEF)[2].

ELRC охватывал все страны, связанные с CEF, т.е. 28 государств – членов ЕС плюс Норвегию и Исландию. Общая цель ELRC заключается в сборе языковых ресурсов от администраций государственных служб и для них во всех странах, входящих в CEF, с тем чтобы улучшить качество, охват и производительность системы машинного перевода CEF (eTranslation)[3] в контексте текущих и будущих цифровых онлайн-сервисов CEF (CEF DSIS).

Таким образом, все данные, собранные ELRC, должны использоваться Европейской комиссией для поддержки разработки eTranslation CEF и его адаптации к соответствующим цифровым сервисам CEF.

Основные результаты, достигнутые с помощью ELRC, включали 225 языковых ресурсов, собранных, проверенных и доставленных. В целом ELRC собрала 138 двуязычных / многоязычных корпусов, 50 терминологий и 37 моноязычных корпусов.

В соответствии с требованиями контракта, ELRC удалось охватить все языки необходимыми типами языковых ресурсов для каждого языка. Кроме того, ELRC провела необходимую оценку и валидацию языковых ресурсов, чтобы обеспечить их качество и пригодность для целей машинного перевода. Все языковые ресурсы, собранные ELRC, были загружены в репозиторий ELRC-SHARE.[4]

ELRC организовала 29 страновых семинаров с участием национальных или региональных организаций, центров языковой компетенции, европейских учреждений и других потенциальных держателей языковых ресурсов. Привлечение ELRC в каждую страну и вовлечение на национальном уровне было ключевым фактором для укрепления ответственности на местном уровне, на которых строится ELRC. Семинары ELRC обеспечили контакты с потенциальными держателями данных, которые были ключевыми для последующего процесса сбора данных**.**

***Европейская исследовательская инфраструктура языковых ресурсов и технологий CLARIN***[5]

CLARIN – это сетевая федерация репозиториев языковых данных, сервисных и экспертных центров. Подробнее деятельность CLARIN будет

---

[1] European Language Resource Coordination – supporting Multilingual Europe. – URL: https://lr-coordination.eu/ (дата обращения: 01.12.2021).

[2] Connecting Europe Facility programme. – URL: https://ec.europa.eu/inea/en/connecting-europe-facility (дата обращения: 01.04.2022).

[3] eTranslation. – URL: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation (дата обращения: 01.12.2021).

[4] ELRC-SHARE Repository. – URL: https://elrc-share.eu/ (дата обращения: 01.12.2021).

[5] CLARIN. – URL: https://www.clarin.eu/ (дата обращения: 01.12.2021).

рассмотрена ниже, в главе 4. Здесь мы приведем некоторые сведения о структурах CLARIN, обеспечивающих доступ к ЛИР и их хранение.

*Виртуальная языковая обсерватория (VLO)*[1]

Одним из важнейших сервисов CLARIN является виртуальная языковая обсерватория (VLO), которая предоставляет средства для изучения ЛИР. Ее цель – обеспечить простой в использовании интерфейс, позволяющий осуществлять единый процесс поиска и обнаружения большого количества ЛИР из самых разных областей. Фасетная организация VLO позволяет легко исследовать доступные ресурсы и получать к ним доступ. Мощный синтаксис запросов позволяет также выполнять более целенаправленный поиск. Он также позволяет легко просматривать параметры обработки обнаруженных ресурсов с помощью коммутатора ЛИР и создавать виртуальные коллекции на основе результатов поиска с помощью реестра виртуальных коллекций.

Всего во VLO 1,2 млн записей, из них 800 тыс. уникальных. Далее описываются основные функциональные возможности VLO.

*Фасетный поиск VLO*. Перечисляются фасеты, и для каждого фасета приводится топ 10 значений этого фасета (в скобках – количество ЛИР для данного значения).

*Языки*

Английский (144 784)
Голландский (121 225)
Немецкий (59 801)
Несколько языков (35 142)
Словенский (30 255)
Французский (27 039)
Испанский; кастильский (16 726)
Болгарский (14 423)
Польский (8398)
Африкаанс (7871)

*Коллекции (приводится собственное название)*

Meertens collection: Liederenbank (128 988)
The Language Archive (115 889)
TextGrid Repository (86 949)
ELAR (85 740)
TalkBank (70 743)
Early English Books Online (Phase 2) (28 347)
Early English Books Online (Phase 1) (25 196)
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (19 149)
Bavarian Archive for Speech Signals (BAS) (17 421)
Варненски периодичен печат в края на 19 и началото на 20 век (14 077).

[1] CLARIN Virtual Language Observatory. – URL: https://vlo.clarin.eu/https://vlo.clarin.eu/help (дата обращения: 01.12.2021).

*Тип ресурса* Топ 10 из примерно 300 вариантов типа ресурсов
- Текст (412 447)
- Аудио (300 835)
- Изображение (132 386)
- Сессия (84 935)
- Аннотация (68 552)
- Стих (53 518)
- Видео (45 777)
- Другое (29 781)
- Периодические издания (29 634)
- Структурированные данные (8937)

*Модальность*
- Разговорные ЛИР (99 557)
- Письменные ЛИР (4523)
- Речь (2421)
- Жесты (1307)
- Указательные жесты (454)
- Мимика (452)
- Эмоциональное состояние (451)
- Письменность (350)
- Письменные языки (220)
- Знаки (191)

*Форматы*
- Text / html (257 337)
- Unknown type (157 940)
- Audio / x-wav (127 214)
- Text / tg.edition+tg.aggregation+xml (86 949)
- Text / xml (85 648)
- Image / jpeg (75 396)
- Application / pdf (74 108)
- Text / x-chat (68 943)
- Text / plain (68 844)
- Application / octet-stream (61 344)

*Доступность*
- Государственные (326 105)
- Академические (1962)
- Ограничения для индивидов (45 737)
- Не определено (399 987)

В пределах одного фасета можно выбрать несколько значений, что позволит расширить выбор. Кроме того, можно сузить область значений внутри фасета, например для поиска многоязычных корпусов, которые охватывают ряд конкретных языков.

*Синтаксис поиска*

В самом простом виде поисковый запрос состоит из одного или нескольких терминов, разделенных пробелами. Такие запросы приводят к получению

всех документов, которые имеют один или несколько вхождений всех включенныхтерминов. Другими словами, оператор AND подразумевается по умолчанию. Можно построить более конкретный запрос, используя расширенные синтаксические функции, поддерживаемые VLO. В системе используется синтаксис анализатора запросов *Lucene query parser*. Полный обзор синтаксических функций, включая параметры нечеткого поиска, диапазоны и повышение термина, можно найти на странице описания синтаксиса *Lucene*[1].

*Поля таргетинга*

В дополнение к логическим операторам синтаксис также позволяет осуществлять поиск вхождений термина в определенном поле, таком как язык или модальность. Доступны следующие имена полей: язык, страна, континент, модальность, жанр, тема, формат, организация, тип ресурса, ключевое слово, ресурсы.

*Интерпретация результатов поиска*

По умолчанию для каждого элемента в результатах поиска отображается заголовок записи и фрагмент ее описания (если таковой имеется). Кроме того, ряд иконок показывают данные, указывающие количество и тип доступных ресурсов (ресурса), а также уровень доступности (публичный, академический или ограниченный), лицензия и / или условия использования.

*Выделение поисковых запросов*

Пользователь может развернуть результаты поиска, чтобы увидеть более подробную информацию. В дополнение к полному описанию отображается ряд дополнительных свойств записи, таких как коллекция, язык и организация (если они доступны), а также список до десяти ресурсов, связанных с записью. Результаты поиска отображаются в определенном порядке, отражают релевантность по отношению к запросу.

*Доступ к ресурсам и другим ссылкам*

Щелчок по заголовку результата поиска (или записи) приводит вас на новую страницу, содержащую информацию о доступности ЛИР. Страница состоит из нескольких вкладок, отображающих различные типы информации, относящиеся к записи.

*Таблица ссылок*

Список всех связанных ресурсов можно найти, выбрав вкладку *Ссылки*. Этот список представлен в виде таблицы, в которой показаны имена файлов всех ресурсов вместе с их типом. Дополнительные сведения для отдельного ресурса можно найти, развернув строку.

*Иерархия записей (вкладка «Иерархия»)*

В некоторых случаях вы не найдете ни ресурсов, ни ссылок на связанные страницы. Это в основном касается записей, которые не указывают на сами ресурсы, а являются частью иерархии. В этом случае доступна вкладка

[1] Apache Lucene – Query Parser Syntax. – URL: https://lucene.apache.org/core/2_9_4/queryparsersyntax.html (дата обращения: 01.12.2021).

*Иерархия*, содержащая дерево, позволяющее просматривать эту иерархию и находить базовые записи, которые могут содержать ссылки на конкретные ресурсы или страницы.

*Обработка ресурсов с помощью инструментов CLARIN*

Многие ресурсы, которые могут быть обнаружены с помощью VLO, пригодны для обработки с использованием одного или нескольких специализированных инструментов. CLARIN упростил этот процесс для ряда инструментов и определенного набора типов ресурсов, позволяя легко обнаружить инструменты, которые могут быть применены к определенному ресурсу, и в случае совпадения немедленно приступить к применению одного или нескольких инструментов к выбранному ресурсу.

*Коммутатор языковых ресурсов*

В разделе *Ссылки* страницы записи (страницы, на которую ссылается заголовок результатов поиска) отображается таблица отдельных ресурсов, совместно описываемых метаданными записи. Из меню «*Параметры*» выбирается опция *Process with Language Resource Switchboard*. Это приведет вас к коммутатору языковых ресурсов (LRS). Здесь можно либо настроить тип файла и языковые значения содержимого, либо перейти к значениям, обнаруженным службой LRS. Подробную информацию о LRS можно найти в сводном документе CLARIN-PLUS[1].

*Предоставление данных в VLO*

Пользователь может получить доступные в цифровом виде ЛИР или инструменты / сервисы, которые либо обрабатывают, либо производят такие ресурсы. Для этих пользователей предлагается стандартизированная процедура «сбора» (извлечения и агрегирования) метаданных, описывающих ресурсы по протоколу OAI-PMH. Метаданные по протоколу OAI-PMH собираются автоматически.

Приведем краткие описания еще нескольких архивов, входящих в инфраструктуру CLARIN.

***Языковой архив TLA***[2]

TLA является частью Психолингвистического института Общества Макса Планка в Неймегене. Он содержит различные типы материалов, в том числе: аудио и видео ЛИР языков со всего мира; фотографии, заметки, экспериментальные данные и другую информацию, необходимую для документирования и описания языков и того, как люди их используют; записи речи при повседневном общении в семьях и сообществах; разговоры взрослых на исчезающих и малоизученных языках, а также языковые явления. Всего TLA включает около 150 тыс. ЛИР.

[1] CLARIN-PLUS Deliverables // CLARIN. – URL: https://www.clarin.eu/content/clarin-plus-deliverables (дата обращения: 01.12.2021).

[2] The Language Archive (TLA). – URL: https://archive.mpi.nl/tla/?fbclid=IwAR1z9MEqta8IeLV3bzNRJHN3QCN43k85jOQRnih1b5UPca7ovnQnZpiKbo0 (дата обращения: 01.12.2021).

TLA включает как самостоятельную часть архив по проекту DOBES[1], который содержит коллекции языковой документации по 68 проектам, что финансировались в рамках программы DOBES. Они включают аудиовизуальные, текстовые и другие связанные ресурсы более чем на 100 исчезающих языках со всего мира.

Просмотр архива TLA возможен по следующим признакам:

- уровень доступа
- коллекции
- автор (создатель)
- страна
- формат
- жанр
- язык

TLA на своем сайте размещает также изложение политики депонирования ЛИР в каталоге, подробную инструкцию депонирования[2] и таблицу допустимых типов и форматов файлов ЛИР[3].

TLA предлагает пользователям ELAN – инструмент аннотирования для аудио- и видеозаписей[4].

***Оксфордский текстовый архив OTA***[5]

OTA предоставляет услуги репозитория для литературных и лингвистических наборов данных. В этой роли OTA собирает, каталогизирует, сохраняет и распространяет высококачественные цифровые ресурсы для научных исследований и преподавания. В настоящее время OTA располагает 64 тыс. текстов на более чем 25 языках и активно работает над расширением фонда. OTA является частью CLARIN; он зарегистрирован как CLARIN-центр, и услуги OTA являются частью вклада Оксфордского университета в консорциум CLARIN-UK.

OTA осуществляет каталогизацию, проверку и долгосрочное хранение электронных текстов, языковых корпусов и других цифровых ЛИР.

OTA предоставляет депонированные электронные тексты, языковые корпуса и другие ЛИР пользователям с минимальными административными препятствиями, соблюдая при этом правовые и этические ограничения на распространение или использование.

---

[1] DOkumentation BEdrohter Sprachen (DOBES). – URL: http://dobes.mpi.nl (дата обращения: 01.12.2021).

[2] Deposit Manual TLA // The Language Archive. – URL: https://archive.mpi.nl/tla/deposit-manual-tla (дата обращения: 01.12.2021).

[3] Accepted file types and formats // The Language Archive. – URL: https://archive.mpi.nl/tla/accepted-file-formats (дата обращения: 01.12.2021).

[4] An annotation tool for audio and video recordings (ELAN) // The Language Archive. – URL: https://archive.mpi.nl/tla/elan (дата обращения: 01.12.2021).

[5] Oxford Text Archive (OTA). – URL: https://ota.bodleian.ox.ac.uk/repository/xmlui/ (дата обращения: 01.12.2021).

ОТА разрабатывает и поддерживает подключение архивных коллекций к соответствующим инфраструктурным службам, например для обнаружения ресурсов и доступа веб-служб к содержимому ресурсов.

ОТА использует методы представления информации в соответствии с правилами TEI[1]. ЛИР, принятые на хранение и соответствующие TEI, будут доступны в режиме онлайн следующими способами:

- метаданные TEI, Dublin Core и OLAC, доступные через OAI-PMH;
- тексты, доступные по соответствующей лицензии Creative Commons в следующих форматах: XML; HTML; ePub; mobi (Kindle); обычный.

***Центр лингвистических исследований LINDAT/CLARIN***[2]

LINDAT / CLARIN предоставляет техническую поддержку и помощь учреждениям или исследователям, которые хотят поделиться, создать и модернизировать свои инструменты и данные, используемые для исследований в области лингвистики или смежных областях. Проект также предоставляет открытый цифровой репозиторий и архив, открытый для всех ученых, которые хотят, чтобы их работа была сохранена, продвинута и широко доступна. Репозиторий содержит свыше 1,1 тыс. ЛИР.

Поиск в репозитории возможен по следующим признакам:

- автор
- предмет
- права
- язык (ISO)
- тип
- содержит ф айлы
- сообщество

***Многоязычные текстовые инструменты и корпуса для языков Центральной и Восточной Европы MULTEXT-East***[3]

Ресурсы MULTEXT-East представляют собой многоязычный набор данных для лингвистических инженерных исследований и разработок. Он включает:

- морфосинтаксические спецификации, определяющие категории (части речи), их морфосинтаксические особенности (атрибуты и значения) и компактные представления наборов тегов;
- морфосинтаксическую лексику;
- аннотированный параллельный корпус «1984»;
- некоторые сопоставимые текстовые и речевые корпуса.

Спецификации доступны для следующих языков и разновидностей языков: албанский, болгарский, чеченский, чешский, дамаскини, английский,

[1] Text Encoding Initiative (TEI). – URL: https://tei-c.org/ (дата обращения: 01.12.2021). Подробнее см. в гл. 3.

[2] LINDAT/CLARIAH-CZ. – URL: https://lindat.cz/repository (дата обращения: 01.12.2021).

[3] MULTEXT-East. – URL: http://nl.ijs.si/ME (дата обращения: 01.12.2021).

эстонский, венгерский, македонский, персидский, польский, румынский, русский, сербскохорватский, словацкий, словенский, торлак и украинский.

***Европейская ассоциация языковых ресурсов ELRA***[1]

Деятельность ELRA подробнее описана в главе 4. Здесь описываются собрания ЛИР и / или сведений о ЛИР, имеющиеся в ELRA. Кроме этих собраний, ELRA ведет базу данных сообщений о разработке ЛИР[2], упорядоченную по хронологии, а также перечень свободно распространяемых ЛИР[3].

*META-SHARE*

Центральным проектом ELRA является инфраструктура по обмену ЛИР, известная под названием META-SHARE[4].

META-SHARE – это механизм открытого обмена ЛИР, предназначенный для устойчивого совместного использования и распространения ЛИР и направленный на расширение доступа к таким ресурсам в глобальном масштабе. META-SHARE – это открытое, интегрированное, безопасное и совместимое средство совместного использования и обмена данными для ЛИР (наборов данных и инструментов). META-SHARE реализуется в рамках сети передового опыта META-NET, состоящей из 60 исследовательских центров 34 стран. META-NET занимается созданием технологических основ многоязычного Европейского информационного общества. Создатель META-NET – META, Многоязычный европейский технологический альянс.

META-SHARE спроектирована как сеть распределенных хранилищ ЛИР, включая языковые данные и основные средства обработки языка (например, морфологические анализаторы, POS-метчики, распознаватели речи и т.д.).

*Компоненты архитектуры*

Каждый репозиторий META-SHARE, участвующий в сети, содержит:

- локальный репозиторий, состоящий из его собственных ЛИР и соответствующих метаданных, которые будут следовать схеме META-SHARE;
- регистр мета-ресурсов, состоящий из метаданных ЛИР, хранящихся во всех репозиториях, участвующих в сети.

Участники META-SHARE берут на себя обязательство следовать схеме метаданных, экспортировать свои метаданные и разрешать их сбор в соответствии с протоколом OAI-PMH. Они могут использовать все предлагаемые услуги, такие как поиск ЛИР, доступ, загрузка, отчетность и т.д.

META-SHARE предлагает следующий спектр услуг:

- регистрация, авторизация и аутентификация пользователей
- описание инструмента ЛИР и загрузка

---

[1] ELRA. – URL: http://www.elra.info (дата обращения: 01.12.2021). Подробнее см. в гл. 4.

[2] Language Resources Announcements // ELRA. – URL: http://www.elra.info/en/catalogues/language-resources-announcements/ (дата обращения: 01.12.2021).

[3] ELRA releases free Language Resources // ELRA. – URL: http://portal.elda.org/en/catalogues/free-resources/ (дата обращения: 01.12.2021).

[4] META-SHARE. – URL: http://portal.elda.org/en/catalogues/meta-share/http://www.meta-net.eu/ (дата обращения: 01.12.2021).

- просмотр, поиск и загрузка ЛИР
- валидация ЛИР, техническое обслуживание, оценка
- архивирование ЛИР, управление версиями и сохранение
- статистика, отчетность, рекомендации
- доступ, распространение и юридические услуги, включая охрану интеллектуальной собственности
- оформление счетов и оплата

*Метаданные*

В рамках META-SHARE подготовлен фундаментальный нормативный документ, описывающий систему метаданных для ЛИР [2]. Его содержание изложено в главе 6.

*Каталог ELRA*

Основным собранием ЛИР в рамках ELRA является каталог ЛИР[1]. В настоящее время каталог ELRA содержит около 1400 ЛИР.

Поиск в каталоге возможен по перечисленным ниже признакам, причем для некоторых признаков указаны допустимые значения.

- Язык
- Тип ресурса
  - корпуса
  - лексика и концептуальные ЛИР
  - инструменты и сервисы
  - языковая документация
- Тип носителя
  - текст
  - аудио
  - изображение
  - видео
  - текст цифровой
  - текст N-граммы
- Доступность
  - доступно
  - доступно через другого дистрибьютора
- Лицензия
  - ограничения использования
  - проверено
  - предусмотрено использование
  - использование специфично для NLP
- Распознавание речи
  - синтез речи
  - языковое моделирование
  - идентификация говорящего
  - верификация говорящего

[1] Catalogue of Language Resources // ELRA. – URL: http://portal.elda.org/en/catalogues/catalogue-language-resources/ (дата обращения: 01.12.2021).

- Тип лингвальности
  - одноязычие
  - двуязычие
  - многоязычие
- Тип многоязычия
- Тип MIME
- Соответствие стандартам
- Область знаний
- Географический охват
- Временной охват
- Содержание текста
- Диалект

В составе каталога ELRA в качестве его самостоятельного подмножества выделен Каталог R&D[1], который содержит ЛИР, полученные в результате академических исследований и доступные бесплатно или по низким ценам и предназначенные для научного использования. В этом каталоге используется собственная типология ЛИР, которая приводится ниже:

- Устные ЛИР

*A – телефонные записи*

Базы данных, каталогизированные в этом разделе, были созданы с использованием записей динамиков, сделанных по телефонной (фиксированной или мобильной) сети или через микрофон. В разделе имеются речевые ресурсы, записанные в различных средах и охватывающие большое количество европейских и неевропейских языков, например базы данных, созданные в рамках проекта SpeechDat.

*B – микрофонные записи*

Базы данных, каталогизированные в этом разделе, были созданы с использованием записей ораторов, сделанных через микрофон, например, базы данных, созданные в рамках баз данных проекта BABEL.

*C – вещательные ресурсы*

Базы данных, каталогизированные в этом разделе, были созданы с помощью записей динамиков, сделанных по радио, телевидению или Интернету, таких как итальянский новостной корпус вещания.

*D – ресурсы, связанные с речью*

В этом разделе представлены ЛИР, отражающие произношение и фонетические лексиконы, такие как базы данных PHONOLEX[2] и MHATLex[3].

- Письменные ЛИР

[1] R&D Catalogue of Language Resources. – URL: http://catalogue-old.elra.info/retd/ (дата обращения: 01.12.2021).

[2] PHONOLEX. – URL: https://www.phonetik.uni-muenchen.de/forschung/abgeschlossene_projekte/phonolex.html (дата обращения: 01.12.2021).

[3] Pérennou G., Calmès M. de. MHATLex: Lexical Resources for Modelling the French Pronunciation. – URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/37.pdf (дата обращения: 01.12.2021).

- корпуса – этот раздел содержит одноязычные и многоязычные корпуса, параллельные или нет, которые также могут быть аннотированы
- одноязычные лексиконы – раздел, посвященный одноязычной лексике, содержит различные типы словарей
- многоязычные лексиконы – двуязычные или многоязычные словари и лексиконы, такие как базы данных EuroWordNet[1]

- Терминологические ЛИР
  - одноязычные, двуязычные и многоязычные терминологические базы данных. Они охватывают большое количество специализированных областей, таких как автомобилестроение, страхование, лингвистика, финансы и т.д.
- Мультимодальные / Мультимедийные ЛИР

Ресурсы этого раздела были созданы в различных модальностях, включая устную речь.

### Карта LRE

Еще одно собрание ЛИР в рамках сообщества ELRA – это массив описаний ЛИР, сделанных на основе Оценочной карты лингвистических ресурсов (карта LRE)[2]. Инициированная ELRA и FLaReNet[3] на выставке LREC 2010, где было заполнено почти 2000 бланков ЛИР, карта LRE стала реальным механизмом мониторинга использования и создания ЛИР путем сбора информации как о существующих, так и о вновь созданных ресурсах. Эта функция была настолько успешной, что была реализована также на других крупных конференциях.

В настоящее время массив карт LRE составляет 6143 ЛИР. В этом массиве возможен поиск с использованием следующих фильтров (для большинства поисковых признаков приводятся несколько самых частых значений с указанием числа ЛИР):

- *Тип ресурса*
  - Способы оценки ЛИР
    - оценочные данные (230)
    - инструмент оценки (71)
    - оценочный пакет (25)
    - методология оценки / стандарты / руководства (15)
    - другое (712)
  - Ресурс-данные
    - корпуса (2920)
    - лексиконы (666)
    - онтологии (162)

[1] EuroWordNet. – URL: https://archive.illc.uva.nl/EuroWordNet/ (дата обращения: 01.12.2021).

[2] LRE Map. – URL: http://lremap.elra.info/ (дата обращения: 01.12.2021).

[3] FLaReNet. – URL: http://www.flarenet.eu/, http://www.elra.info/en/projects/archived-projects/flarenet/ (дата обращения: 01.12.2021).

        - модели грамматики языка (82)
        - терминология (66)
        - банки деревьев (42)
    - Ресурс-руководство
        - представление – аннотационный формализм / руководства (62)
        - языковые ресурсы / технологическая инфраструктура (20)
        - метаданные (10)
    - Ресурс-инструмент
        - таггер / парсер (400)
        - аннотирование (245)
        - корпусные инструменты (83)
        - распознавание именованных сущностей (60)
        - инструмент машинного перевода (51)
        - программный инструмент (41)
        - токенизатор (35)
        - инструмент машинного обучения (32)
        - инструмент языкового моделирования (29)
        - распознаватель многозначности (17)
        - распознаватель речи / транскрибер (14)
        - обработка сигналов / извлечение функций (14)
        - веб-сервис (9)
        - синтезатор текста в речь (9)
        - идентификатор языка (6)
        - динамик Распознавания (4)
        - инструмент анализа настроений (4)
        - просодический анализатор (3)
        - анализатор изображений (3)
        - инструмент устного диалога (1)
- *Состояние производства ЛИР*
    - существующий / используемый (2587)
    - вновь созданный / в процессе разработки (1408)
    - вновь созданный – законченный (1290)
    - существующий – обновленный (487)
- *Доступность*
    - в свободном доступе (2772)
    - другое (1232)
    - от собственника (1229)
    - из дата-центра (ов) (580)
- *Модальность*
    - письменность (4355)
    - другое (500)
    - речь (430)
    - мультимодальный / мультимедийный (286)
- *Использование ресурсов*
    - другое (1566)
    - извлечение информации / поиск информации (608)

- машинный перевод / синхронный перевод (532)
- синтаксический анализ и маркировка (289)

- *Тип ЛИР по языкам*
  - одноязычный (2507)
  - независимо от языка (2206)
  - многоязычный (961)
  - двуязычный (380)
- *Язык (top 4)*
  - английский (961)
  - немецкий (216)
  - французский (180)
  - испанский (130)

***Универсальный каталог***[1]

Универсальный каталог содержит информацию о ЛИР, идентифицированных по всему миру. Эти ЛИР, как правило, находятся командой ELRA, но также включены сведения от членов ELRA, сотрудников и посетителей вебсайта.

Универсальный каталог является общедоступным для широкого круга языковых сообществ. Он предназначен облегчить как поиск ЛИР, так и хранение для всех пользователей ЛИР. Оба вида деятельности выполняются в упрощенном порядке.

Нынешний универсальный каталог представляет собой первый этап работы по международному сотрудничеству. Второй этап будет осуществляться в сотрудничестве с другими партнерами, которые проводят аналогичные инициативы в США и Японии.

Основная типология ЛИР в этом каталоге:

- Инструменты
- Устные ЛИР
  - телефон
  - десктоп / микрофон
  - ЛИР, связанные речью
  - вещательные ЛИР
- Письменные ЛИР
  - корпуса
  - одноязычные лексиконы
  - многоязычные лексиконы
- Терминологические ЛИР
- Мультимодальные / мультимедийные ЛИР

Предлагается форма для пополнения каталога.

[1] Universal Catalogue // ELRA. – URL: http://www.elra.info/en/catalogues/universal-catalogue/ (дата обращения: 01.12.2021).

***Российский архив уральских и алтайских языков на платформе ЛингвоДок***[1]

Начиная с 2013 года под руководством доктора филологических наук Ю.В. Норманской ведется создание платформы *ЛингвоДок*, на которой в настоящее время собраны аудиословари и корпуса более чем по 900 исчезающим диалектам уральских и алтайских языков России.

Помимо места для хранения данных и поиска данных, на этой платформе есть возможность одновременной распределенной обработки материала и программы для его анализа, в частности, выявления в онлайн-режиме фонетического сходства языков, употребления тех или иных морфологических параметров в определенном значении, возможности построения карт фонетических, морфологических или лексических изоглосс в синхронии и их изменений в диахронии.

Платформа ЛингвоДок позволяет размещать данные пользователей из различных организаций с сохранением всех прав создателей словарей и корпусов, дает возможности работы с данными в режиме, когда материалы открыты только ограниченному числу пользователей, выбранных создателем словаря или корпуса. Но при этом для каждого пользователя ЛингвоДока появляется возможность сравнения данных его словарей по любым параметрам с данными других диалектов с помощью авторских программ сотрудников Лаборатории. Благодаря тому, что на платформе уже сейчас представлены материалы в едином цифровом формате по 900 диалектам уральских и алтайских языков России, суммарный объем которых превышает 2 млн словоформ, анализ сравнительный-исторический, фонетический, морфологический проводится методом обработки больших данных, что значительно повышает точность полученных результатов.

Очевидно, что в представленном обзоре описаны лишь несколько собраний ЛИР, отобранных по субъективной оценке автора. Список хранилищ ЛИР, включенных в известные каталоги, а также обнаруженных автором, представлен в приложении 1, которое, впрочем, также не претендует на исчерпывающую полноту.

## Литература к главе 2

1. Усталов Д.А. Каталоги лингвистических ресурсов: состояние и перспективы // Молодой ученый. – 2012. – № 12. – С. 148–152.

2. Documentation and User Manual of the META-SHARE Metadata Model / E. Desipri [et al.] ; ed.: P. Labropoulou, E. Desipri // META-NET. – URL: http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf (дата обращения: 01.12.2021).

[1] Lingvodoc 3.0. – URL: http://lingvodoc.tsu.ru/ (дата обращения: 01.12.2021).

# ГЛАВА 3. МЕЖДУНАРОДНАЯ ДЕЯТЕЛЬНОСТЬ В ОБЛАСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ЦИФРОВОЙ ГУМАНИТАРИСТИКИ

## Общие замечания

В настоящей главе описывается международная деятельность, связанная с созданием и развитием ЛИР, языковых технологий, а также цифровой гуманитаристики в целом. Описываются международные организации и реализуемые ими проекты в данных областях, кроме европейских, которые вынесены в главу 4.

Описание включает следующие разделы:

- Профессиональные ассоциации по компьютерной и прикладной лингвистике
- Консорциумы
- Защита и сохранение исчезающих языков
- Терминологическая и переводческая деятельность
- Организации по цифровой гуманитаристике

Общие списки международных организаций в области компьютерной лингвистики и цифровой гуманитаристики, а также проектов, реализуемых этими организациями, включая европейские проекты, вынесены в приложения 3 и 4.

## Профессиональные ассоциации

Общее взаимодействие между специалистами разных стран осуществляется через профессиональные ассоциации. Они проводят конференции, издают журналы, создают порталы и сайты, организуют учебу и семинары, находят другие формы взаимодействия. В настоящем обзоре описывается деятельность следующих ассоциаций:

Ассоциация компьютерной лингвистики (ACL):

- Международная ассоциация прикладной лингвистики (AILA)
- Международный комитет по координации и стандартизации речевых баз данных и методов оценки (COCOSDA)
- Международный комитет компьютерной лингвистики (ICCL – COLING)
- Международная ассоциация квантитативной лингвистики (IQLA)
- Международная ассоциация речевых коммуникаций (ISCA)

***Ассоциация компьютерной лингвистики ACL***[1]

ACL является самой известной профессиональной организацией в данной области, в нее входит около 2 тыс. членов из всех регионов мира. Антология ACL в настоящее время содержит 64 тыс. статей по изучению компьютерной лингвистики и обработки естественного языка. ACL проводит десять ежегодных конференций и издает два журнала.

ACL включает 24 рабочие группы (SIG), их состав отражает направления деятельности ассоциации:

- SIGANN: аннотирование
- SIGBioMed: биомедицинские аспекты NLP
- SIGDAT: лингвистические данные и корпусные подходы к NLP
- SIGDIAL: дискурс и диалог
- SIGEDU: образовательные приложения
- SIGEL: языки, находящиеся под угрозой исчезновения
- SIGFSM: методы и модели с конечным состоянием в NLP
- SIGGEN: генерация естественного языка
- SIGHAN: китайский язык
- SIGHUM: языковые технологии для социально-экономических и гуманитарных наук
- SIGLEX: лексиконы
- SIGMOL: математика языка
- SIGMORPHON: вычислительная морфология, фонология и фонетика
- SIGMT: машинный перевод
- SIGNLL: изучение естественного языка
- SIGPARSE: парсинг
- SIGREP: репрезентативное обучение
- SIGSEM: вычислительная семантика
- SIGSLAV: обработка естественного языка славян
- SIGSLPAT: обработка речи для вспомогательных технологий
- SIGSLT: устный перевод
- SIGTYP: лингвистическая типология
- СИГУР: уральские языки
- SIGWAC: сеть как корпус

***Международная Ассоциация прикладной лингвистики AILA***[2] AILA основана в 1964 году во Франции. В состав AILA входят более 8 000 человек по всему миру, которые в качестве исследователей, политиков или практиков активно работают в области прикладной лингвистики.

Прикладная лингвистика – это междисциплинарная область, занимающаяся практическими проблемами языка и коммуникации, которые могут быть выявлены, проанализированы или решены путем применения существующих лингвистических теорий, методов, средств и результатов

[1] Association for Computational Linguistics. – URL: https://www.aclweb.org/portal (дата обращения: 01.12.2021).

[2] Association Internationale de Linguistique Appliquée. – URL: https://aila.info/ (дата обращения: 01.12.2021).

исследований или путем разработки новых теоретических и методологических основ для прикладных лингивстических задач.

Проблемное поле прикладной лингвистики довольно широко: от аспектов языковой и коммуникативной компетентности личности, таких как овладение первым или вторым языком, грамотность, языковые расстройства и т.д., до проблем, связанных с языком и коммуникацией в обществах и между языковыми личностями, таких как, например, языковые вариации и языковая дискриминация, многоязычие, языковые конфликты, языковая политика и языковое планирование и т.д.

***Международный Комитет по координации и стандартизации речевых баз данных и методов оценки COCOSDA***[1]

COCOSDA поддерживает развитие ресурсов устной речи и оценку речевых технологий. В первом случае COCOSDA способствует разработке разных типов корпусов данных устной речи с целью построения и / или оценки текущей или будущей технологии устной речи. Для оценки технологий COCOSDA предлагает координацию научно-исследовательских проектов с целью повышения их эффективности.

COCOSDA поддерживает разработку новых тематических областей, но только если новая область оправдана новым приложением речевой технологии, т.е. если это приложение предъявляет новые требования к форме корпусов данных.

Тематические области COCOSDA:

- распознавание речи
- синтез речи
- устные корпуса
- инструменты аннотации корпусов
- локальные языки

***Международный комитет компьютерной лингвистики ICCL (COLING)***[2]. ICCL был создан в качестве постоянного органа для проведения международных конференций по компьютерной лингвистике COLING, но без постоянного секретариата, подписки или фондов. Отличительной чертой COLING всегда было разнообразие подходов к вычислительной технике и лингвистике.

***IQLA***[3] ***– Международная ассоциация количественной лингвистики***

IQLA была основана в 1994 году для содействия использованию математических и статистических методов в лингвистическом моделировании, текстовом анализе и смежных областях. Количественная лингвистика охватывает

[1] International Committee for the Co-ordination & Standardisation of Speech Databases and Assesment Techniques. – URL: http://www.cocosda.org/ (дата обращения: 01.12.2021).

[2] International Committee on Computational Linguistics (ICCL). – URL: https://wiki2.org/en/International_Committee_on_Computational_Linguistics (дата обращения: 01.12.2021).

[3] The International Quantitative Linguistics Association (IQLA). – URL: http://www.iqla.org (дата обращения: 01.12.2021).

весь спектр теоретических и эмпирических исследований в этих областях, включая наблюдение и описание лингвистических данных, формулирование и эмпирическую проверку количественно сформулированных лингвистических и текстовых моделей относительно таких данных, а также обсуждение связанных с ними методологических и эпистемологических проблем.

За последние три десятилетия количественная лингвистика претерпела бурное развитие, как в теоретическом, так и в прикладном отношении. Вводя в лингвистику количественные методы и модели научных дисциплин из естественных и социальных наук, она способствовала развитию новых и захватывающих теоретических перспектив, а также решению практических проблем в различных отраслях лингвистических и текстологических исследований.

***ISCA*[1] *– Международная ассоциация речевых коммуникаций***

Цель ассоциации состоит в том, чтобы продвигать в международном мировом контексте деятельность и обмен опытом во всех областях, связанных с наукой и техникой речевой коммуникации. Ассоциация предназначена всем лицам и учреждениям, заинтересованным в фундаментальных исследованиях и технологических разработках, направленных на описание, объяснение и воспроизведение различных аспектов человеческого общения с помощью речи. В рамках ISCA действуют рабочие группы (SIG), организованные по тематикам или по языкам:

*Тематические группы*

- Синтез речи – SynSig
- Аудиовизуальная речь – AVISA
- Недостаточно обеспеченные ресурсами языки – SIGUL
- Характеристика говорящего и языка – SpLC
- Речевая просодия – SPROSIG
- Дискурс и диалог – SIGDIAL
- Речевые и языковые технологии в образовании – SLaTE
- Машинное обучение – SIGML
- Речь и язык в мультимедиа – SLIM
- Обработка речи и языка для вспомогательных технологий – SLPAT
- История наук о речевой коммуникации – SIG-HIST
- Взаимодействие с компьютером ребенка – SIG-CHILD
- Надежная обработка речи – SIG-RoSP
- Безопасность и конфиденциальность в речевой коммуникации – SIG-SPSC
- Устный перевод – SIG-SLT

*Языковые группы*

- Обработка китайского разговорного языка – SIG-CSLP
- Французская ассоциация устной речи – AFCP
- Итальянская ассоциация устной речи – AISV
- Иберийские языки – SIG-IL

[1] International Speech Communication Association (ISCA). – URL: https://www.isca-speech.org/iscaweb/index.php/about-isca (дата обращения: 01.12.2021).

- Обработка речи на индийском языке – SIG-ILSP
- Анализ русской речи – SIGRU
- Недостаточно обеспеченные ресурсами языки – SIGUL

## Консорциумы

В отличие от ассоциаций, обычно включающих специалистов как физических лиц, консорциумы объединяют организации, в том числе университетские лаборатории. Приведем описания некоторых наиболее известных консорциумов и основных реализуемых ими проектов.

***Консорциум лингвистических данных LDC***[1]

*LDC* – это открытый консорциум университетов, библиотек, корпораций и государственных исследовательских лабораторий. Первоначально основная роль LDC заключалась в создании репозитория и распространении языковых ресурсов. Сейчас и с помощью своих членов LDC превратилась в организацию, которая создает и распространяет широкий спектр языковых ресурсов. LDC также поддерживает спонсируемые исследовательские программы и оценки языковых технологий, предоставляя ресурсы и внося свой организационный вклад. LDC находится в Университете Пенсильвании.

LDC поддерживает образование, исследования и развитие языковых технологий путем создания и обмена ЛИР. Основные типы ЛИР:

- *Данные.* Предоставление ЛИР этого типа сообществу – основа деятельности LDC. Ежегодно адаптируется и рассылается подписчикам и членам LDC 30–36 ЛИР.
- *Инструменты*. Программное обеспечение, разработанное и распространяемое LDC. Все инструменты доступны бесплатно по лицензии с открытым исходным кодом. Условия лицензий могут отличаться.
- *Документы*. Публикации (с 1998 г.) сотрудников LDC, а также доклады, презентации, статьи и книги.

Основные функции LDC:

*Получение данных*. Членство в LDC обеспечивает организациям права, привилегии (и скидки) на доступ. Организациям, не являющимся членами, большинство ЛИР доступны с ограничениями.

*Управление данными*. Эффективность NLP в немалой степени зависит от доступности ЛИР, удобства использования и возможности архивирования. LDC предоставляет информацию о «распаковке» ЛИР, извлечении файлов и других способах использования данных. Предлагается простая, проверенная временем модель лицензии LDC. В сервисе управления данными объясняется, как LDC помогает исследователям в разработке и реализации планов управления данными для конкретных проектов.

[1] Linguistic Data Consortium (LDC). – URL: https://www.ldc.upenn.edu (дата обращения: 01.12.2021).

Перечислим текущие проекты LDC.

- *AIDA*. Цель – разработать многогипотезный семантический движок, который генерирует явные альтернативные интерпретации событий, ситуаций и тенденций из множества неструктурированных источников. LDC поддерживает AIDA, собирая, создавая и аннотируя мультимодальные лингвистические ресурсы на нескольких языках.
- *KAIROS*. Разработка системы искусственного интеллекта на основе схем, способствующая идентификации сложных событий и доведению их до сведения пользователей.
- *NIEUW*. Новые стимулы и рабочие процессы в сборе и аннотации лингвистических данных. Цель состоит в том, чтобы создать основу для разработки многоязычных языковых ресурсов с использованием методов краудсорсинга, доказавших свою эффективность во многих научных дисциплинах.
- *SRE* (*Speaker Recognition Evaluation*). Оценка систем распознавания речи по заказу Национального института стандартов.
- *LRE* (*Language Recognition Evaluation*). Оценка систем распознавания текста по заказу Национального института стандартов.

***OLAC – Сообщество открытых лингвистических архивов***[1]

OLAC является международным партнерством учреждений и частных лиц, которые создают всемирную виртуальную библиотеку языковых ресурсов путем:

- выработки консенсуса в отношении лучшей современной практики цифрового архивирования языковых ресурсов
- создания сети взаимодействующих хранилищ и служб для обеспечения доступа к таким ресурсам

OLAC объединяет 63 лингвистических архива из всех регионов мира, в которых содержатся до 450 тыс. описаний ЛИР, на более 8 тыс. языков. OLAC поддерживает фундаментальный каталог ЛИР самых разных типов, который был описан выше. Предлагается сводный анализ метаданных для ЛИР, а также инструменты поддержки и использования ЛИР. Реализована развитая поисковая машина для каталога, позволяющая получать различную статистику.

OLAC уделяет большое внимание технологиям работы с ЛИР, протоколам и процедурам обмена ЛИР, а также разработке и поддержке системы метаданных. В основе метаданных лежит известный международный стандарт «Дублинское ядро» (*Dublin Core*).

OLAC поддерживает модель лингвистических связанных открытых данных с использованием технологий Семантического веба.

Приведем перечень основных документов OLAC (в квадратных скобках приводится обозначение документа).

[1] OLAC: Open Language Archives Community. – URL: http://olac.ldc.upenn.edu/ (дата обращения: 01.12.2021).

*Стандарты*

Принципы и процессы OLAC [2006-04-05]. Этот документ обобщает руководящие идеи OLAC (т.е. цель, видение и основные ценности), а затем описывает, как OLAC организован и как он работает.

Метаданные OLAC [2008-05-31]. Этот документ определяет формат, используемый сообществом открытых языковых архивов OLAC для обмена метаданными в рамках инициативы открытых архивов OAI. Набор метаданных основан на полном наборе Дублинских основных терминов метаданных DCMT, но формат позволяет использовать расширения для выражения специфичных для сообщества квалификаторов.

Репозитории OLAC [2008-07-28]. Этот документ определяет стандарты, которым должны следовать архивы OLAC при реализации хранилища метаданных.

*Рекомендации*

Рекомендации по лучшей практике описания языковых ресурсов [2008-07-11]

Словарь «Тип дискурса» [2012-02-04]

Словарь «Лингвистические типы данных» [2006-04-06]

Лингвистический предметный словарь [2006-04-06]

Словарь ролей (Contributor в формате Dublin Core) [2006-04-06]

Рекомендуемые расширения метаданных [2008-02-22]

*Методики и проекты*

Запросы для выборочного сбора метаданных OLAC [2003-07-29]

Создание записей OLAC и репозиториев в Toolbox [2006-09-28]

Метрики метаданных OLAC [2009-06-29]

Руководство по использованию метаданных OLAC [2008-07-11]

Спецификации формата отображения метаданных OLAC и перехода OLAC в OAI_DC [2009-07-23]

Дополнительные расширения метаданных [2002-12-04]

Viser – средство отображения метаданных OLAC [2003-07-29]

***Международное сообщество лингвистов онлайн LINGUIST List***[1]

Проект вырос из форума по обмену информацией академических лингвистов, организованного в Университете Индианы. Сейчас это самый посещаемый лингвистический портал, который включает сведения по ЛИР, публикациям, конференциям, учреждениям, проектам, специалистам, вакансиям, учебным программам, геоданные по языкам и многое другое. Для каждой категории лингвистической информации предлагается поиск с использованием разнообразных фильтров. На базе возникшего вокруг портала сообщества LINGUIST List была организована реализация ряда проектов, в том числе перечисляемых ниже.

[1] The Linguist List International Linguistics Community Online. – URL: https://linguistlist.org (дата обращения: 01.12.2021).

*Mailing Lists*[1]. *Архив лингвистических сайтов* – проект по созданию единого постоянного и доступного для поиска архивного сайта для сотен языковых форумов и дискуссий прошлого и настоящего, чтобы информация, которую они содержат, могла быть свободно доступна любому специалисту в этой дисциплине.

*LL-MAP*[2]. *Лингвистические геоданные.* Проект, предназначенный для интеграции языковой информации с данными естественных и социальных наук с помощью географической информационной системы (ГИС). Подробное описание см. в главе 15.

*MultiTree*[3]. *Генеалогические деревья.* В рамках проекта создается цифровая библиотека научных гипотез о языковых отношениях и подгруппах. Эта информация систематизируется в базе данных. Имеется возможность поиска с помощью веб-интерфейса. Каждая гипотеза представляется графически в виде интерактивного гиперболического отображения генеалогического древа, сопровождаемого информацией обо всех задействованных языках, а также об авторах и библиографических источниках гипотезы.

***SIL International***[4]

SIL – международная некоммерческая организация (бывший Летний институт лингвистики – Summer Institute of Linguistics). Миссия SIL – служение языковым сообществам в качестве международного защитника. SIL обслуживает языковые сообщества по всему миру для устойчивого развития языка посредством исследований, переводов, обучения и разработки материалов.

По состоянию на 2020 год SIL участвует примерно в 1350 активных языковых проектах в 104 странах. Эти проекты затрагивают более 1,1 млрд человек в 1600 местных общинах. Работа SIL объединяет более 4300 сотрудников из 89 стран, которые работают вместе с тысячами других местных партнеров и общественных волонтеров по всему миру.

SIL работает с сообществами этнолингвистических меньшинств, чтобы они наращивали свой потенциал для устойчивого развития собственных языков. Развитие языка включает в себя действия, которые языковое сообщество предпринимает для того, чтобы его язык продолжал служить его меняющимся социальным, культурным, политическим, экономическим и духовным потребностям и целям. Опыт SIL в области языкового развития включает обучение и консультации по таким направлениям, как лингвистический анализ, орфография и разработка систем письма, разработка литературы, многоязычное образование и грамотность. Поскольку носители языка стремятся использовать свой язык в письменных материалах, может потребоваться разработка

[1] Mailing Lists // The Linguist List. – URL: https://old.linguistlist.org/lists/ (дата обращения: 01.12.2021).

[2] LL-Map: Language and Location – Map Accessibility Project. – URL: http://llmap.org/ (дата обращения: 01.12.2021).

[3] MultiTree // The Linguist List. – URL: https://old.linguistlist.org/projects/multi-tree.cfm (дата обращения: 01.12.2021).

[4] SIL International. – URL: https://www.sil.org/ (дата обращения: 01.12.2021).

алфавита. Сотрудники SIL обладают техническими знаниями для определения звуков языка, которые должны быть представлены, и выработки рекомендаций по системе письма. Лингвисты SIL основывают свои рекомендации на понимании психолингвистики и лучших практиках в теории грамотности. Основные ресурсы SIL:

- образовательные ресурсы по английскому языку
- глоссарии лингвистических терминов
- языковые и культурные архивы
- публикации SIL
- программы и шрифты

SIL является лидером в области идентификации и документирования мировых языков. Он поддерживает авторитетную базу данных *Ethnologue* (Этнолог: языки мира)[1]. Ethnologue объединяет больше данных, чем любой другой ресурс подобного рода – от населения до карт, диалектов, сведений об использовании языков, тенденции развития языков и многое другое. Имеется статистика, публикации, справочные данные. В ресурсе возможен поиск по странам, названиям и кодам языков, языковым семьям. Имеются коммерческие сервисы.

*Общая онтология для лингвистических описаний GOLD*[2]. Проект, начавшийся в 2004 году в Институте информации и технологий Восточного Мичиганского университета, вырос в сообщество, цель которого – объединить ученых, заинтересованных в наилучшей практике кодирования лингвистических данных. Проект продвигает передовую практику, развивает интероперабельность данных за счет использования онтологии GOLD, облегчает поиск по разрозненным наборам данных и предоставляет платформу для обмена существующими данными и инструментами. Таксономия лингвистических терминов GOLD основана на принципах онтологической инженерии, что обеспечивает богатую аксиоматизацию классов и отношений.

Вначале GOLD была построена сверху вниз с использованием онлайнового глоссария лингвистических терминов SIL International и стандартных лингвистических источников. Позже GOLD была сопоставлена с Объединенной онтологией верхнего уровня (SUMO). В настоящее время GOLD внедряется в качестве строительного блока в инфраструктуру компьютерной лингвистики. На базе сообщества GOLD реализуется ряд проектов, создаются программные инструменты и методики.

*Электронная лингвистика*[3]. Функция консорциума – содействие использованию электронных данных и связанных с ними инструментов в области лингвистики. Разработанный пакет e-Linguistics содержит код Python для:

---

[1] Ethnologue: Languages of the World. – URL: https://www.ethnologue.com/about (дата обращения: 01.12.2021).

[2] General Ontology for Linguistic Description (GOLD) Community. – URL: http://linguistics-ontology.org/ (дата обращения: 01.12.2021).

[3] The e-Linguistics Toolkit (ELTK). – URL: http://purl.org/linguistics/eltk (дата обращения: 01.12.2021).

- преобразования устаревших данных в различных рабочих форматах в совместимый формат;
- хранения и объединения преобразованных данных;
- вывода лингвистических данных в различных форматах.

*Метасхема языка SIL*[1]. Это – язык семантической интерпретации, который используется для определения значения элементов и атрибутов в схеме разметки XML в терминах понятий, определенных в формальной семантической схеме (такой, как схема RDF или онтология OWL).

*FIELD. Среда ввода лингвистических данных*[2]. FIELD – это веб-инструмент для полевых лингвистов, позволяющий размещать свои языковые данные в полностью доступной для поиска базе данных.

*LEGO. Расширение лексикона с помощью онтологии GOLD*[3]. Это проект по созданию инструментов и стандартов для облегчения обмена и взаимодействия лексических данных. Создается сеть взаимодействия данных путем сопоставления с понятиями GOLD лексических элементов из нескольких лексиконов и списков слов. Пользователи могут осуществлять поиск по морфосинтаксической информации, а также по определению, языку, орфографии и т.д. LEGO также предлагает набор требований к данным, которые создатели лексиконов могут реализовать для того, чтобы присоединиться к сети интероперабельности.

*ODIN. Онлайновая база данных межлинейных примечаний*[4] формируется на основе сбора лингвистических данных из Интернета. ODIN содержит данные по различным языкам мира, подвергнувшиеся лингвистическому анализу по методу IGT[5]. IGT содержит фонетическую транскрипцию, морфосинтаксический анализ (морфемный глоссарий, грамматическую информацию различного рода), а также свободный перевод. В настоящее время ODIN дает ссылки на научные статьи в Интернете, которые содержат примеры IGT. Имеются разные виды поиска. ODIN в настоящее время интегрируется с лингвистической онтологией, чтобы пользователи могли искать IGT, используя терминологию GOLD.

Следует указать также на совместный проект Института психолингвистики Общества Макса Планка (MPI), Университета Франкфурта и Института языковой информации и технологий Восточного Мичиганского университета *RELISH. Обеспечение совместимости лексиконов языков, находящихся под угрозой исчезновения, с помощью гармонизации стандартов.*

[1] SIL: A Metaschema Language. – URL: http://www.sil.org/~simonsg/metaschema/ (дата обращения: 01.12.2021).

[2] Field Input Environment for Linguistic Data. – URL: http://emeld.org/tools/fieldinput.cfm (дата обращения: 01.12.2021).

[3] Lexicon Enhancement via the GOLD Ontology. – URL: http://lego.linguistlist.org/ (дата обращения: 01.12.2021).

[4] The Online Database of Interlinear Text. – URL: http://odin.linguistlist.org/ (дата обращения: 01.12.2021).

[5] IGT (interlinear glossed text) – межлинейные текстовые примечания.

Этот проект существенно способствовал развитию системы стандартов на ЛИР. Содержание этого проекта изложено в главе 5.

***Азиатская федерация по обработке естественного языка (AFNLP)***[1]

Кроме организаций мирового охвата, в сфере компьютерной лингвистики действуют международные региональные организации. В качестве примера приведем описание *Азиатской федерации по обработке естественного языка*. AFNLP призвана содействовать распространению информации и научно-исследовательскому сотрудничеству между исследователями региона, а также координировать инициативы региона с инициативами других регионов в области обработки естественного языка и смежных областях.

Официальными членами AFNLP являются либо профессиональные ассоциации, либо научно-исследовательские институты / университеты в странах или территориях региона, которые представляют исследователей в этих странах / территориях или берут на себя ответственность представлять их. Приглашаются также другие организации, такие как организаторы конференций в регионе, международные профессиональные организации, профессиональные ассоциации исследовательских областей, связанных с NLP.

AFNLP учреждает подкомитеты, которые занимаются конкретными вопросами, такими как совместное использование лингвистических ресурсов и т.д.

Исследовательские области, которые охватывают организации – члены AFNLP, являются областями связанными с NLP, включая, но не ограничиваясь ими:

- использование компьютеров в различных областях лингвистики, таких как прагматика, дискурс, семантика, синтаксис, морфология, лексика и др.;
- теории NLP, такие как когнитивные модели понимания языка, теория перевода, машинное обучение, статистические языковые модели, коммуникационные модели, нейронные модели обработки языка и т.д.;
- языковые технологии, такие как анализ и генерация устной и письменной речи, машинный перевод, перевод речи, поиск и извлечение информации, интерфейс естественного языка и диалоговые системы, анализ текста и т.д.;
- приложения, связанные с NLP, такие как мультимодальное взаимодействие, интеграция и обмен знаниями, инструменты для интернета, XML и его расширения, построение онтологий и т.д.;
- ресурсы и методы оценки систем NLP.

## Защита и сохранение исчезающих языков

Одним из важнейших направлений деятельности, связанной с созданием ЛИР, является деятельность, направленная на сохранение исчезающих языков и языков, находящихся в опасности. Этим занимается ЮНЕСКО, а также специализированные организации.

[1] Asian Federation of Natural Language Processing. – URL: http://www.afnlp.org/wp/ (дата обращения: 01.12.2021).

***Фонд исчезающих языков FEL***[1]

Это некоммерческая организация, базирующаяся в Нью-Хейвене, штат Коннектикут. FEL поддерживает проекты поддержки языков и документации, находящихся под угрозой исчезновения, которые направлены на сохранение языков мира, одновременно предоставляя редкие лингвистические данные научному сообществу. С 1997 года Фонд спонсировал более 100 языковых проектов в 30 странах, а недавно начал разработку большого цифрового архива языковых данных, находящихся под угрозой исчезновения. Основным механизмом поддержки ***FEL*** является финансирование отдельных лиц, племен и музеев. Поддерживаемые программы – это проекты по развитию радиопрограмм коренных народов, записи речи пожилых людей и последних живых носителей языков, находящихся под угрозой исчезновения, а также производство материалов для использования в программах обучения языкам во всем мире.

***Программа документирования исчезающих языков ELDP***[2]

Целью Программы ELDP, основанной в 2002 году, является сохранение находящихся под угрозой исчезновения языков во всем мире. С этой целью ELDP предоставляет гранты по всему миру отдельным лицам или группам для документирования этих языков. ELDP предоставляет финансирование для проектов документирования, возглавляемых лингвистами, лингвистическими антропологами и членами сообществ, обладающих навыками лингвистического документирования. Каждый год выделяется от 30 до 40 грантов на проекты документирования по всему миру. Эти средства позволяют грантополучателям проводить полевые работы по фиксации речи носителей языков, находящихся под угрозой исчезновения, на аудио- и видеозаписи, составлению документального описания по языку или жанру, находящемуся под угрозой исчезновения. Эти документальные коллекции затем архивируются и сохраняются, и становятся свободно доступными через цифровой онлайн-архив исчезающих языков (ELAR), который является частью библиотеки Школы восточных и африканских исследований (SOAS) Лондонского университета[3].

***Документирование исчезающих языков DOBES***[4]

Языковое документирование – это реакция языкового сообщества на имманентное исчезновение большинства мировых языков. Оно преследует три основные цели:

- техническое обслуживание и ревитализация;

[1] The Foundation for Endangered Languages. – URL: https://ogmios.org/ (дата обращения: 01.12.2021).

[2] The Endangered Languages Documentation Programme. – URL: https://www.eldp.net/en/about (дата обращения: 01.12.2021).

[3] SOAS Library. – URL: https://www.soas.ac.uk/library/ (дата обращения: 01.12.2021).

[4] DOBES: Documentation of endangered languages. – URL: https://dobes.mpi.nl/research/ (дата обращения: 01.12.2021).

- сохранение информации о языковом разнообразии и культурных ценностях человечества для будущих поколений носителей языка и исследователей;
- введение отчетности в лингвистические исследования.

В настоящее время во всем мире говорят примерно на 7000 языках (многие из них имеют ряд диалектов). Однако предполагается, что к концу XXI века только одна треть, а может быть только одна десятая этих языков будет продолжать существовать. Поскольку язык является уникальным выражением интеллектуального наследия и культурных знаний каждого говорящего сообщества, способы концептуализации окружающей среды и социальной структуры будут безвозвратно утрачены со смертью языка.

В 2000 году Фольксвагенфонд начал программу DOBES (*Dokumentation bedrohter Sprachen*) с целью документирования языков, которые потенциально находятся под угрозой исчезновения. В 2000 году был начат экспериментальный этап с участием семи групп документации и одной группы архивирования с намерением выработать рекомендации относительно того, как может работать языковая документация и как лучше всего осуществлять цифровое архивирование. С тех пор ежегодно отбирались новые группы документирования, чтобы в течение трех-пяти лет провести значительную работу по документированию. Было профинансировано 67 проектов документирования. Ежегодно проводятся рабочие совещания, в ходе которых участники прошлых и нынешних проектов встречаются для обмена опытом и результатами.

***Электронная метаструктура данных об исчезающих языках E-MELD***[1]

Цель проекта, который реализует Linguist list, – создание архитектуры эффективного сотрудничества между лингвистами, работающими над исчезающими языками. Одной из основных целей проекта является выработка консенсуса в отношении стандартов для метаданных, лингвистических аннотаций и языковой идентификации, что позволит обеспечить широкий доступ к данным в максимально полезной форме.

Члены научного сообщества сталкиваются с двумя неотложными ситуациями: число языков в мире стремительно сокращается, в то время как число инициатив по оцифровке языковых данных стремительно увеличивается. Последнее может показаться чистым благом перед лицом первой ситуации, но есть две проблемы, из-за которых все может пойти не так без адекватного сотрудничества между архивистами, полевыми лингвистами и лингвистическими инженерами.

Во-первых, общий стандарт оцифровки лингвистических данных не согласован, а возникающие в результате этого различия в методах архивирования и языковой репрезентации серьезно затрудняют доступ к данным, поиск и межъязыковое сравнение.

[1] Electronic Metastructure for Endangered Languages Data (E-MELD). – URL: http://emeld.org/ (дата обращения: 01.12.2021).

Во-вторых, стандарты могут разрабатываться без руководства со стороны дескриптивных лингвистов, людей, которые лучше всего знают диапазон структурных возможностей человеческого языка. Если лингвистические архивы хотят обеспечить максимально широкий доступ к данным и предоставить их в максимально полезной форме, необходимо достичь консенсуса по некоторым аспектам архивной инфраструктуры. Целью проекта является выработка рекомендаций лучшей практики по следующим темам:

- метаданные
- разметка
- идентификация языка
- лингвистическая онтология
- сервер метаданных языковых архивов по всему миру
- демонстрационный проект по десяти исчезающим языкам

***Проект по архивированию лингвистических данных LACITO***[1]

Целями проекта по архивированию лингвистических данных LACITO являются сохранение и распределение речевых данных. Для этого разработаны нормы подготовки и использования документов, включающих звук и текст, с использованием международных стандартов, в частности SGML (стандартный обобщенный язык разметки).

Основным источником данных для проекта является множество документов, зафиксированных в полевых условиях членами LACITO за последние 30 лет. Это уникальные записи, в основном спонтанной речи на бесписьменных языках, служат основой для исследования соответствующих языков и культур. Транскрипции и переводы некоторых из них были опубликованы, но оригиналы никогда не публиковались и должным образом не архивировались.

Документы, подготовленные в рамках проекта, включают в себя как звук, так и текст, как минимум фонологическую транскрипцию и свободный перевод, а также, где это возможно, пословные глоссы, заметки и т.д. Текст индексируется по звуку на уровне «предложения» или интонационной группы. Доступ к документам возможен либо локально на компакт-диске, либо по Сети.

***Фонд библиотеки долговременного хранения естественных языков Rosetta***[2]

Проект Rosetta – это глобальное сотрудничество языковых специалистов и носителей языка, работающих над созданием общедоступной цифровой библиотеки естественных языков. Проект Rosetta – это первое исследование Фонда в области долгосрочного архивирования. Он служит средством сосредоточения внимания на проблеме цифрового устаревания и способах решения этой проблемы с помощью креативных методов архивного хранения. Прототип долгосрочного архива – *Rosetta Disk* – никелевый диск диаметром три дюйма с почти 14 000 страниц информации, микроскопически

[1] Linguistic Data Archiving Project (LACITO). – URL: http://xml.coverpages.org/lacitoARdesc-english.html (дата обращения: 01.12.2021).

[2] The Rosetta Project. – URL: https://rosettaproject.org/ (дата обращения: 01.12.2021).

выгравированной на его поверхности. Поскольку каждая страница представляет собой изображение, а не цифровое кодирование 1 и 0, она может быть прочитана человеческим глазом с помощью оптического увеличения мощностью 500. Диск покрыт сферой из нержавеющей стали и стекла, что защищает его от случайных ударов и истирания. При минимальном уходе он мог бы легко сохраняться и читаться в течение тысяч лет.

Фонд решил начать с создания ключа, своего рода декодера, для любой информации, которую можно было бы оставить в письменной форме на любом языке. В основе коллекции дисков Rosetta лежит набор «параллельной» информации – одни и те же тексты, один и тот же набор слов, одни и те же типы описаний – для более чем 1000 человеческих языков. Идея собрать параллельные тексты была вдохновлена оригинальным «Розеттским» камнем, который имел один и тот же основной текст (декрет), написанный тремя различными письменами. С тех пор коллекция проекта Rosetta выросла до более чем 100 000 страниц документов, а также языковых записей на более чем 2500 языках. В настоящее время коллекция размещена в интернет-архиве и продолжает расширяться за счет новых материалов и вкладов.

Будучи лингвистической коллекцией, проект Rosetta призван привлечь внимание к утрате мировых языков. Точно так же, как глобализация угрожает культурному разнообразию человечества, языки небольших, уникальных, локализованных человеческих обществ подвергаются серьезному риску. Фактически лингвисты предсказывают, что в течение следующего столетия мы можем потерять до 90% мирового языкового разнообразия. Чтобы остановить эту волну и помочь обратить вспять эту тенденцию, идет работа над поощрением культурного и языкового разнообразия человека, а также над тем чтобы ни один язык не исчез бесследно.

***Программа ЮНЕСКО по поддержке языкового разнообразия и многоязычия в Интернете***[1]

ЮНЕСКО убеждена, что культурное разнообразие и многоязычие в Интернете призваны сыграть ключевую роль в укреплении плюралистического, справедливого, открытого и инклюзивного общества знания. ЮНЕСКО призывает государства-члены разрабатывать всеобъемлющую языковую политику, выделять ресурсы и использовать соответствующие инструменты для поощрения и облегчения языкового разнообразия и многоязычия, включая Интернет и средства массовой информации. В этой связи организация поддерживает включение новых языков в цифровой мир, создание и распространение контента на местных языках в Интернете и каналах массовой коммуникации, а также поощряет многоязычный доступ к цифровым ресурсам в киберпространстве. Программа поддерживает следующие ресурсы по проблеме языкового разнообразия:

- нормативные документы
- языковую политику

[1] Linguistic diversity and multilingualism on Internet. – URL: http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/linguistic-diversity-and-multilingualism-on-internet/ (дата обращения: 01.12.2021).

- продвижение местного контента в Интернете
- измерение языкового разнообразия в Интернете
- атлас языков, находящихся в опасности
- интернационализированные доменные имена
- специальные инициативы и мероприятия
- глоссарий по управлению Интернетом

***Атлас ЮНЕСКО по языкам мира, находящимся под угрозой исчезновения***[1]

Этот ресурс призван повысить осведомленность о языковой угрозе и необходимости сохранения языкового разнообразия мира среди политиков, профессиональных сообществ и широкой общественности, а также стать инструментом мониторинга состояния языков, находящихся под угрозой исчезновения, и тенденций в области языкового разнообразия на глобальном уровне.

Последнее издание Атласа насчитывает около 2500 языков (из которых 230 – языки, вымершие с 1950 года), т.е. приближается к общепринятым оценкам, что около 3000 языков находится под угрозой. Для каждого языка атлас содержит его название, степень опасности исчезновения и страны, в которых на нем говорят. Онлайн-издание предоставляет дополнительную информацию о количестве говорящих, соответствующих политиках и проектах, источниках, кодах ISO и географических координатах. Эта бесплатная интернет-версия атласа впервые обеспечивает широкий доступ, интерактивность и своевременное обновление информации на основе отзывов.

***Обеспечение функциональной совместимости лексиконов языков, находящихся под угрозой исчезновения, посредством гармонизации стандартов RELISH***[2]

Проект RELISH направлен на гармонизацию ключевых европейских и американских стандартов, установление единого способа представления структуры лексики в ЛИР, а также разработку процедуры миграции гетерогенных лексиконов в совместимый со стандартами формат XML.

Проект RELISH способствует языковым исследованиям, решая двуединую проблему: во-первых, гармонизацию между цифровыми стандартами лексической информации в Европе и Америке, а во-вторых – взаимодействие существующих лексиконов исчезающих языков, созданных с использованием разрозненных программных средств. Расхождение цифровых стандартов в лексических данных до сих пор препятствовало международному сотрудничеству в области языковых технологий для создания ресурсов, анализа и вебсервисов для доступа к архивам. Проект RELISH работает над гармонизацией ключевых европейских и американских цифровых стандартов, установлением единого способа ссылки на структуру лексики и лингвистические концепции.

---

[1] UNESCO Interactive Atlas of the World's Languages in Danger. – URL: http://www.unesco.org/languages-atlas/ (дата обращения: 01.12.2021)

[2] RELISH (Rendering Endangered Language Lexicons Interoperable through Standards Harmonization) // The Linguist List. – URL: https://old.linguistlist.org/projects/relish.cfm (дата обращения: 01.12.2021).

Работа над RELISH включала интеграцию двух основных стандартов лексической разметки, используемых в настоящее время в лингвистической работе: общей онтологии лингвистического описания (GOLD) и реестра категорий данных ISOcat. Связанная с этим попытка заключается в создании формата цифрового обмена для лексиконов, который будет доступен как выходной XML-формат RELISH для лексиконов, оцифрованных проектом LEGO (улучшение лексики с помощью онтологии GOLD) в LINGUIST List. Этот формат облегчит взаимодействие между лексиконами, созданными с помощью различных программных средств, и будет способствовать более тесному сотрудничеству между лингвистами из разных частей мира.

В конечном итоге RELISH принесет значительные выгоды отдельным исследователям и организациям, поддерживающим их исследования. ЛИР занимают центральное место в большом научном сообществе, включая антропологов, археологов, историков, генетиков, социологов и лингвистов. Следовательно, обеспечение функциональной совместимости любого отдельного словаря экспоненциально увеличит его потенциальный научный вклад. Кроме того, гармонизация стандартов на этом раннем этапе упростит будущую разработку программных инструментов и веб-сервисов, используемых в лексических исследованиях, и придаст импульс другим усилиям по гармонизации стандартов, а также предложит научному сообществу гибкий и интегрированный доступ к важным новым цифровым материалам.

## Терминологическая и переводческая деятельность

Терминологическая и переводческая деятельность является предметом многих международных организаций, прежде всего ООН и ЮНЕСКО, а также многих специализированных организаций. Ряд проектов по терминологии описан в главах 4 и 12. Здесь мы приведем сведения лишь о некоторых структурах и проектах.

***Консорциум терминологии и переводов LTAC***[1]

LTAC создан для решения языковых проблем в глобальном масштабе. Это достигается путем координации усилий ее членов и партнеров. Как некоммерческий консорциум, LTAC использует обширный опыт, методы и технологии для терминологии, овладения языком и локализации / перевода. LTAC работает над продвижением управления терминологией, овладением языком и управлением переводческими проектами. LTAC предоставляет бесплатный доступ к языковым ресурсам для поддержки этой деятельности. LTAC реализует следующие проекты:

- *Linport* – формат представления переводческих материалов
- *TBX* – формат обмена терминологическими данными
- *Tranquality* – модель оценки «качества перевода» с помощью системы многомерных показателей качества (MQM)

[1] Language Terminology/Translation and Acquisition Consortium. – URL: http://ltacglobal.org/about.html (дата обращения: 01.12.2021).

- *GEvTerm* – база данных для решения лингвистических проблем
- *TerminOrg* – терминология для крупных организаций

***Международная ассоциация машинного перевода IAMT***[1]

IAMT включает три региональных ассоциации:

- Ассоциация машинного перевода в Северной и Южной Америке (AMTA)
- Азиатско-Тихоокеанская Ассоциация машинного перевода (AAMT)
- Европейская ассоциация машинного перевода (EAMT)

EAMT – это организация, которая обслуживает растущее сообщество людей, заинтересованных в МТ и инструментах перевода, включая пользователей, разработчиков и исследователей этой все более жизнеспособной технологии. EAMT вместе с AMTA и AAMT организует семинары и конференции, такие как двухлетний саммит МТ, а также ежегодные конференции EAMT и рабочие места. Под эгидой IAMT также составляются списки компаний и продуктов, которые распространяются бесплатно или по номинальной стоимости среди его членов. EAMT также поддерживает список рассылки mt-list@eamt.org как общественный форум для обсуждения технологии перевода.

Архив МТ[2] представляет собой электронное хранилище и библиографию статей, книг и документов по темам машинного перевода и компьютерных средств помощи переводу. В этом хранилище можно найти полные материалы всех конференций EAMT и IAMT.

Среди проектов EAMT – справочник коммерческих систем машинного перевода и средств поддержки перевода. Это полный список текущего программного обеспечения, коммерческих продуктов и поставщиков составлен Джоном Хатчинсом от имени EAMT. Имеется соответствующая база данных. EAMT ведет список рассылки, посвященный МТ. Этот список открыт для широкой публики и служит платформрой для обсуждения и обмена информацией по всем аспектам переводческой технологии.

***Азиатско-Тихоокеанская ассоциация машинного перевода AAMT***[3]

AAMT состоит из исследователей, производителей и пользователей систем машинного перевода. Ассоциация стремится развивать технологии машинного перевода для расширения сферы эффективных глобальных коммуникаций. AAMT занимается разработкой, совершенствованием, обучением и рекламой систем машинного перевода.

[1] International Association for Machine Translation (IAMT). – URL: http://eamt.org/international-association-for-machine-translation/ (дата обращения: 01.12.2021).

[2] Ingestion of the MT-archive website. – URL: https://github.com/mardub1635/mt-archive/blob/master/README.md (дата обращения: 01.12.2021).

[3] Asia-Pacific Association for Machine Translation (AAMT). – URL: https://aamt.info/ (дата обращения: 01.12.2021).

***Терминология для крупных организаций TerminOrg***[1]

TerminOrg – это консорциум терминологов, которые продвигают терминологический менеджмент как неотъемлемую часть корпоративного стиля, разработки контента, управления контентом и глобальных коммуникаций в крупных компаниях. Миссию TerminOrg можно сформулировать следующим образом:

- предоставление руководящих принципов и рекомендаций по управлению терминологией;
- повышение осведомленности о роли терминологии для эффективной внутренней и внешней коммуникации, передачи знаний, образования, снижения рисков, управления контентом, перевода и присутствия на мировом рынке, особенно в крупных организациях;
- представление заинтересованным сторонам терминологических стандартов и инструментов;
- определение и пропаганда экономической ценности управленческой терминологии.

***Глобальные терминологические вызовы GEvTerm***[2]

Цель GEvTerm – решить лингвистические проблемы во всем мире. Подход GEvTerm к преодолению лингвистических вызовов заключается в создании большой компьютеризированной языковой базы данных, которую люди могут пополнять, извлекать из нее выгоду и получать к ней доступ по всему миру через любое подключенное к Интернету устройство.

GEvTerm включает в себя множество проектов, разработанных для решения отдельных лингвистических проблем, например создание многоязычной базы данных названий продуктов питания.

***UNGEGN***[3]***, группа экспертов ООН по географическим названиям***

К терминологической деятельности правильно относить и международное сотрудничество в области топонимики. После Второй конференции в 1972 году специальная группа экспертов была официально преобразована в группу экспертов Организации Объединенных Наций по географическим названиям для продолжения программы сотрудничества между конференциями. UNGEGN организует конференцию ООН по стандартизации географических названий, созываемую каждые пять лет с целью:

- поощрять стандартизацию национальных и международных географических названий;
- содействовать международному распространению информации о национальных стандартизированных географических названиях;

[1] Terminology for Large Organizations (TerminOrgs). – URL: http://www.terminorgs.net (дата обращения: 01.12.2021).

[2] Global Event Terminology. – URL: http://gevterm.net/gevterm/index.php (дата обращения: 01.12.2021).

[3] United Nations Group of Experts on Geographical Names. – URL: https://unstats.un.org/unsd/ungegn (дата обращения: 01.12.2021).

- принять единые системы латинизации для преобразования каждой нелатинской системы письма в латинский алфавит.

*Рабочие группы* UNGEGN

1. Рабочие группы по названиям стран.
2. Рабочая группа по географическим названиям управления данными.
3. Рабочие группы по топонимической терминологии.
4. Рабочие группы по рекламе и финансированию.
5. Рабочие группы по системам латинизации.
6. Рабочие группы по учебным курсам по топонимии.
7. Рабочая группа по оценке и реализации.
8. Рабочая группа по экзонимам.
9. Рабочая группа по географическим названиям культурного наследия:

A. целевая группа для Африки

B. топонимические руководящие принципы для картографических и других редакторов для международного использования

*Ресурс:* Всемирные географические названия UNGEGN. Многоязычный набор данных названий стран, столиц и крупных городов[1].

***Портфолио проекта языковой совместимости Linport***[2]

Это один из проектов, инициированных LTAC. Цель – разработка открытого независимого от поставщика формата, который может быть использован многими различными переводческими сервисами для представления переводческих материалов. Область применения проекта Linport не включает разработку форматов для конкретных видов данных; Linport, однако, включает сотрудничество с группами, которые разрабатывают такие форматы (например, XLIFF). Формат Linport должен частично основываться на существующих форматах пакетов для конкретных инструментов или организаций.

В проекте сформулирован следующий набор требований.

*Автономная информация о проекте перевода*. Пакет должен содержать или указывать на всю информацию, необходимую для выполнения задач в рамках переводческих проектов.

*Осуществимость реализации*. Должна быть обеспечена возможность разработки процедур импорта / экспорта для существующих средств перевода без необходимости внесения существенных изменений в их базовую конструкцию.

*Независимость платформы / инструмента перевода*. Пакет Linport не должен быть привязан к какой-либо конкретной платформе, инструменту перевода или языку программирования.

*Спецификации структурированного перевода*. Пакет Linport должен включать способ передачи спецификаций, совместимых со стандартом ISO TS 11669.

---

[1] Geographical Names Database // UNGEGN World Geographical Names. – URL: https://unstats.un.org/unsd/geoinfo/geonames/ (дата обращения: 01.12.2021).

[2] Linport: The Language Interoperability Portfolio Project. – URL: www.linport.org (дата обращения: 01.12.2021).

*Поддержка удаленного и локального доступа к ресурсам*. Ресурсы, на которые ссылается пакет Linport, могут быть локальными (например, на жестком диске) или удаленными (например, на сервере, доступ к которому осуществляется через веб-службы), и оба режима доступа должны поддерживаться.

*Открытый интерфейс*. Абстрактное описание интерфейса прикладного программного обеспечения, а также несколько специфичных для языка программирования API (все производные от абстрактного описания интерфейса) и библиотека для каждого API, так что приложения могут использовать библиотеку, а не обращаться непосредственно к пакету Linport.

*Эталонная реализация*. Недостаточно определить только формат или интерфейс. Необходимо предоставить справочную реализацию Linport с открытым исходным кодом.

*Модульная организация пакетов.* Пакеты должны поддерживать операции объединения и разделения, необходимые для поддержки требований рабочего процесса. Например, сложный многоязычный пакет может быть разделен на несколько двуязычных пакетов, которые затем могут быть изменены (например, путем добавления целевого текста) и рекомбинированы позже.

*Профили*. Linport должен допускать профили, подмножества общей архитектуры Linport, которые поддерживают потребности конкретных аудиторий или целей.

## Цифровая гуманитаристика

В данном разделе мы опишем лишь несколько организаций по цифровой гуманитаристике (DH), главным образом те, которые ставят своей целью международную координацию. Дополнительно деятельность в этой сфере будет рассмотрена в главах 4 и 20.

***Альянс организаций цифровой гуманитаристики ADHO***[1]

ADHO – это зонтичная организация цифровых гуманитарных наук, созданная в 2005 году для координации деятельности нескольких региональных организаций DH, называемых учредительными организациями, а именно:

- Европейская ассоциация цифровых гуманитарных наук (EADH)
- Ассоциация компьютеров и гуманитарных наук (ACH)
- Канадское общество цифровых гуманитарных наук (CSDH/SCHN)
- Австралийская ассоциация цифровых гуманитарных наук (AADH)
- Японская ассоциация цифровых гуманитарных наук (JADH)
- Humanistica – франкоязычная ассоциация цифровых гуманитарных наук
- Тайваньская ассоциация цифровых гуманитарных наук (TADH)

Цели ADHO заключаются в продвижении и поддержке цифровых исследований и преподавания различных гуманитарных дисциплин, объединяя гуманитариев, занимающихся цифровыми и компьютерными исследованиями,

[1] Alliance of Digital Humanities Organizations. – URL: https://wiki2.org/en/Alliance_of_Digital_Humanities_Organizations (дата обращения: 01.12.2021).

преподаванием, созданием, распространением цифровых ресурсов. ADHO поддерживает инициативы в области публикации, презентации, сотрудничества и профессиональной подготовки; признает и поддерживает передовой опыт в этих усилиях; консультирует различные коллективы. Членами ADHO-обществ являются те, кто находится на переднем крае таких областей, как текстовый анализ, электронные публикации, кодирование документов, текстовые исследования и теория, новые медиаисследования и мультимедиа, цифровые библиотеки, прикладная дополненная реальность, интерактивные игры и многое другое. В ее состав входят исследователи и преподаватели гуманитарных компьютерных наук и академических кафедр, а также специалисты по ресурсам, работающие в библиотеках, архивных центрах и с гуманитарными компьютерными группами.

*Конференция ADHO*[1]

Альянс курирует совместную ежегодную конференцию, которая начиналась как конференция ACH/ALLC (или ALLC/ACH), а теперь известна как конференция Digital Humanities.

*Специальные группы интересов* (SIGs)

- AVinDH. Аудиовизуальные материалы и их использование в DH
- GO:: DH (Global Outlook: Digital Humanities). Цель – развитие глобальной коммуникации и сотрудничество
- GeoHumanitie: пространственные и временные аспекты DH
- Библиотеки и DH
- Связанные открытые данные; DH и сообщество Semantic Web

*Рецензируемые журналы ADHO*

- *DSH: Digital Scholarship in the Humanities*, печатный журнал, издательства Oxford University Press
- *Digital Studies/Le champ numérique*, открытый доступ, рецензируемый электронный журнал от CSDH/SCHN, основанный в 2008 году
- *Digital Humanities Quarterly*, открытый доступ, рецензируемый электронный журнал от ADHO
- *DH Commons*, открытый доступ, рецензируемый электронный журнал от centerNet
- *Humanités numériques*, открытый доступ, рецензируемый электронный журнал Humanistica
- *Journal of the Text Encoding Initiative* – официальный журнал Консорциума TEI
- *Journal of Digital Archives and Digital Humanities*, открытый доступ, рецензируемый электронный журнал от TADH

В качестве примера организации, входящей в ADHO, рассмотрим деятельность франкоязычной ассоциации цифровой гуманитаристики ***Humanistica***[2]**.**

[1] Digital Humanities conference. – URL: https://wiki2.org/en/Digital_Humanities_conference (дата обращения: 01.12.2021).

[2] L'association francophone des humanités numériques/digitales (Humanistica). – URL: http://www.humanisti.ca/ (дата обращения: 01.12.2021).

Humanistica объединяет сообщество участников научных исследований и высшего образования (исследователей, инженеров, студентов, преподавателей, специалистов в области научно-технической информации и т.д.), которые используют цифровые инструменты в своей работе в области гуманитарных и социальных наук или имеют целью изучение цифровых технологий в своих соответствующих областях. Это сообщество объединяет людей и учреждения, независимо от границ, которые используют французский язык для проведения исследований в своей практике. То есть, эта организация является международной.

Humanistica ориентирована, с одной стороны, на франкоязычные академические институты, для которых проблемы и практика DH не всегда очевидны; с другой стороны – связана с другими международными организациям, занимающимся DH, в которых франкоязычная специфика может и должна быть учтена.

Не ограничиваясь тем, что она является местом для дискуссий цифровых гуманитариев, ассоциация является инициатором конкретных проектов и организует несколько рабочих групп:

- *инструменты и ноу-хау* – готовит франкоязычную базу данных инструментов DH;
- *цифровое искусство, дизайн и гуманитарные науки*. Целью этой группы является создание путем организации семинаров, учебных дней и опросов сети исследователей в разных областях: архитектура, прикладное искусство, пластическое искусство, исполнительское искусство, эстетика, музыковедение, музыка, искусствоведение и т.д.;
- *образование*. Рабочая группа разрабатывает рекомендации по обучению студентов и молодых ученых в области DH. Группа систематически выявляет имеющиеся в настоящее время курсы по DH за пределами университетского образования.

***Консорциум Инициативы по кодированию текстов TEI***[1]

TEI – наиболее известная и авторитетная международная организация, разрабатывающая методологию в области DH. Миссия TEI заключается в разработке и поддержании руководящих принципов цифрового кодирования литературных, лингвистических и исторических текстов. Консорциум публикует *Руководящие принципы для электронного кодирования текста и обмена*. Это международный и междисциплинарный стандарт, который широко используется библиотеками, музеями, издателями и отдельными учеными для представления всех видов текстовых материалов для онлайн-исследований и преподавания.

В дополнение к руководящим принципам консорциум предоставляет различные ресурсы и услуги: учебные мероприятия по изучению TEI, информацию о проектах, использующих TEI, библиографию публикаций, связанных с TEI, а также программное обеспечение, разработанное для TEI или адаптированное к методам TEI.

[1] Text Encoding Initiative. – URL: https://tei-c.org/ (дата обращения: 01.12.2021).

Сообщество TEI включает членов в Северной Америке, Европе, Австралии и Азии, в сотнях университетов, библиотек, научных организаций по всему миру. Материалы, на кодирование которых рассчитано руководство TEI, столь же разнообразны, как и применяющие его специалисты из широкого спектра гуманитарных и социальных наук. Помимо широкого распространения в электронных библиотеках, TEI используется для представления рукописей, исследовательских работ, исторических архивов, ранних печатных изданий, книг, лингвистических сборников, антологий, критических изданий, древних надписей и множества других литературных, исторических и культурных материалов. Сфера охвата TEI постоянно расширяется, и руководящие принципы находятся в постоянном развитии, чтобы идти в ногу с возникающими потребностями сообщества TEI.

***Национальная инициатива по сетевому культурному наследию NINCH***[1]

Это коалиция организаций, созданная для обеспечения лидерства культурного сообщества в эволюции цифровой среды. Особенно полезно руководство NINCH по передовой практике в области цифрового представления и управления материалами культурного наследия [1]. Однако в последние годы новых материалов на сайте NINCH не появлялось.

***Международная сеть центров цифровых гуманитарных наук CenterNet***[2]

CenterNet – международная сеть, созданная для совместных действий в интересах цифровых гуманитарных наук и смежных областей в целом, а также гуманитарной цифровой инфраструктуры в частности. Опираясь на свою новую публикацию DHCommons, CenterNet позволяет отдельным Центрам DH объединяться в международные сети, обмениваясь проектами, инструментами, персоналом и опытом.

Центр включает около 200 организаций, включая российского участника – Пермский государственный национальный исследовательский университет.

CenterNet реализует различные инициативы, такие как День цифровой гуманитаристики[3], Ресурсы для создания и поддержания центров DH[4], а также предоставляет виртуальный центр DH для изолированных проектов DH и платформу для обучения более широкого научного сообщества цифровым гуманитарным наукам.

---

[1] The National Initiative for a Networked Cultural Heritage (NINCH). – URL: http://www.ninch.org/ (дата обращения: 01.12.2021).

[2] An international network of digital humanities centers (CenterNet). – URL: https://dhcenternet.org/ (дата обращения: 01.12.2021).

[3] Day of DH 2020. – URL: https://dhcenternet.org/initiatives/day-of-dh/2020 (дата обращения: 01.12.2021).

[4] Resources for Starting and Sustaining Digital Humanities Centers. – URL: https://dhcenternet.org/resources-for-starting-and-sustaining-dh-centers/ (дата обращения: 01.12.2021).

***Консорциум гуманитарных центров и институтов CHCI***[1]

Консорциум гуманитарных центров и институтов – это глобальный форум, который укрепляет работу гуманитарных центров и институтов посредством пропаганды, предоставления грантов и инклюзивного сотрудничества. CHCI продвигает межинституциональные партнерства, признает региональные гуманитарные культуры и мобилизует коллективный потенциал гуманитарных наук для решения наиболее актуальных проблем современного общества.

***Коалиция гуманитарных и художественных инфраструктур и сетей CHAIN***[2]

Коалиция была создана в 2009 году на основе Мэрилендского института технологий для гуманитарных наук. Объявленные цели:

- пропаганда совершенствования цифровой исследовательской инфраструктуры для гуманитарных наук;
- разработка устойчивых бизнес-моделей;
- содействие технической интероперабельности ресурсов, инструментов и услуг;
- пропаганда передовой практики и соответствующих технических стандартов;
- развитие инфраструктуры совместного обслуживания;
- координация подходов к правовым и этическим вопросам;
- взаимодействие с другими соответствующими инициативами в области вычислительной инфраструктуры;
- расширение географического охвата коалиции.

На сайте коалиции[3] представлен обширный (не менее 100) перечень исследований, проведенных членами коалиции по различным сферам цифровой гуманитаристики.

***Проект Bamboo***[4]

Этот проект (2008–2012) был инициативой по развитию инфраструктуры для искусств и гуманитарных наук. Возглавляемый Калифорнийским университетом в Беркли, проект включает Австралийский национальный университет, Индианский университет, Северо-Западный университет, Университет Тафтса, Чикагский университет, Иллинойский университет в Урбана-Шампейне, Мэрилендский университет, Оксфордский университет и Висконсинский университет в Мэдисоне.

Мэрилендский институт технологий для гуманитарных наук (МИТ) разрабатывает в рамках проекта программную среду, в которой ученые могут

[1] The Consortium of Humanities Centers and Institutes. – URL: http://chcinetwork.org/ (дата обращения: 01.12.2021).

[2] Coalition of Humanities and Arts Infrastructures and Networks. – URL: https://mith.umd.edu/news/chain/ (дата обращения: 01.12.2021).

[3] The Maryland Institute for Technology in the Humanities (MITH). – URL: https://mith.umd.edu/research/ (дата обращения: 01.12.2021).

[4] Project Bamboo. – URL: https://www.projectbamboo.org/ (дата обращения: 01.12.2021).

открывать, анализировать и сопровождать цифровые тексты на протяжении 450 лет печатной культуры на английском языке – с 1473 по 1923 год, а также тексты из классического мира. Архив проекта доступен по адресу[1].

***Гуманитарные и социальные науки онлайн H-Net***[2]

Это независимая некоммерческая научная ассоциация, которая предлагает открытое академическое пространство для ученых, преподавателей, продвинутых студентов и других специалистов. Цифровая платформа H-Net, *The Commons*, предоставляет динамический набор функций, которые позволяют ученым взаимодействовать друг с другом, совместно производить знания и распространять информацию среди своих подписчиков и широкой общественности. Созданная на основе онлайн-интегратора сетей, модерируемых сертифицированными редакторами, H-Net имеет уникальные возможности для поощрения технологических инноваций в гуманитарных и социальных науках, сохраняя при этом передовой академический опыт. В настоящее время на платформе H-Net размещается 180 бесплатных онлайн-сообществ (сетей), с примерно 200 тыс. подписчиков, каждая сеть контролируется консультативным советом экспертов. Публичные архивы H-Net датируются с 1994 годом, и, как и весь контент H-Net, они находятся в свободном для общественности доступе.

Вот некоторые разделы ресурсов H-Net:

- статьи и журналы, представляющие интерес
- библиография
- новые дискуссии
- медиаархивы и галереи
- информационные бюллетени и обзоры
- ресурсы для исследований
- учебные ресурсы

## Литература к главе 3

1. The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials / the Humanities Advanced Technology and Information Institute, University of Glasgow, the National Initiative for a Networked Cultural Heritage // NINCH. – URL: https://www.ninch.org/guide.pdf (дата обращения: 01.12.2021).

[1] Bamboo Download Archives. – URL: https://www.atlassian.com/software/bamboo/download-archives (дата обращения: 01.12.2021).

[2] Humanities and Social Sciences Online (H-Net). – URL: https://networks.h-net.org/node/513/pages/59033/mission-statem (дата обращения: 01.12.2021).

# ГЛАВА 4. ИНФРАСТРУКТУРА ЯЗЫКОВЫХ РЕСУРСОВ И ТЕХНОЛОГИЙ: ЕВРОПЕЙСКИЙ ОПЫТ

## Введение

Общая структура управления наукой и инновациями в ЕС представлена на сайте Европейской комиссии[1]. Она образует чрезвычайно сложную и развитую систему, включающую множество различных органов и проектов. В настоящей главе мы рассмотрим те структуры, которые связаны с организацией научной коммуникации и созданием информационных ресурсов, прежде всего для задач NLP и DH в целом.

Научные коммуникации и ресурсы вместе с другими инструментами и сервисами входят в научную инфраструктуру. Для создания и поддержания научной инфраструктуры в ЕС создали специфическую правовую форму, которую назвали *Европейские консорциумы научной инфраструктуры (ERIC)*[2]. ERIC предоставляют исследовательским сообществам ресурсы и услуги для проведения исследований и стимулирования инноваций, в том числе:

- основное научное оборудование, или наборы инструментов;
- коллекции, архивы или научные данные;
- вычислительные системы и коммуникационные сети;
- другие компоненты инфраструктуры, открытые для внешних пользователей.

Руководящий орган ЕС – *Европейская комиссия* – определяет, оценивает и реализует стратегии и инструменты для создания в Европе устойчивой научной инфраструктуры мирового уровня. Комиссия гарантирует, что все ресурсы и услуги ERIC открыты и доступны для всех исследователей в Европе и за ее пределами. Комиссия также разработала хартию доступа к научной инфраструктуре[3].

[1] Research and innovation // European Commission. – URL: https://ec.europa.eu/info/research-and-innovation_en (дата обращения: 01.12.2021).

[2] European Research Infrastructure Consortium (ERIC). – URL: https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en (дата обращения: 01.12.2021).

[3] European charter of access for research infrastructures. Principles and guidelines for access and related services. – URL: https://op.europa.eu/en/publication-detail/-/publication/78e87306-48bc-11e6-9c64-01aa75ed71a1/ (дата обращения: 01.12.2021).

Ключевые цели ERIC:

- сокращение дублирования в разработках;
- координация разработки и использования научных ресурсов и услуг;
- разработка стратегий для новых инфраструктур;
- объединение усилия на международном уровне для создания и управления крупными, сложными или дорогостоящими инфраструктурами;
- содействие объединению навыков, данных и усилий лучших ученых мира;
- стимулирование инновационного потенциала путем повышения осведомленности промышленности о новых возможностях.

Всего в настоящее время создано 18 ERIC, их перечень имеется на странице ERIC Landscape[1]. Из 18 ERIC два непосредственно относятся к языковым технологиям и цифровой гуманитаристике, они будут рассмотрены ниже.

## Инициативы и стратегии по научной инфраструктуре

Кратко перечислим различные европейские структуры, определяющие создание и поддержку научной инфраструктуры. Нужно заметить, что таких структур в ЕС достаточно много. Вообще, на взгляд стороннего наблюдателя, число постоянных или временных органов, имеющих отношение к управлению наукой в Евросоюзе, явно избыточно.

*Европейский стратегический форум по научным инфраструктурам (ESFRI)*[2]. ESFRI разрабатывает стратегическую дорожную карту, определяющую инвестиционные приоритеты ERIC на ближайшие 10–20 лет.

*Группа старших должностных лиц*[3]. Глобальная группа экспертов, которая подводит итоги существующей ситуации с ERIC и исследует новые возможности сотрудничества.

*EIROforum*[4] – соглашение о сотрудничестве по объединению ресурсов, возможностей и опыта организаций-членов для поддержки европейской науки.

*Ассоциация исследовательских инфраструктур европейского уровня (ERF-AISBL)*[5]. Некоммерческая ассоциация, способствующая развитию и популяризации европейской инфраструктуры, которая предоставляет доступ

[1] ERIC Landscape. Members of the European Research Infrastructure Consortium (ERIC). – URL: https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric/eric-landscape_en (дата обращения: 01.12.2021).

[2] European Strategy Forum on Research Infrastructures (ESFRI). – URL: https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/esfri_en (дата обращения: 01.12.2021).

[3] Group of Senior Officials (GSO) on global Research Infrastructures. – URL: https://www.gsogri.org/ (дата обращения: 01.12.2021).

[4] EIROforum. – URL: https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eiroforum_en (дата обращения: 01.12.2021).

[5] The Association of European-Level Research Infrastructures Facilities. – URL: https://erf-aisbl.eu/ (дата обращения: 01.12.2021).

внешним пользователям. Члены ERF открыты на международном уровне и включают в себя национальные инфраструктуры, а также европейские сети и консорциумы исследовательских инфраструктур. Ежегодно организации – члены ERF обслуживают более 20 тыс. академических и промышленных пользователей из Европы и других стран.

*Глобальный научный форум ОЭСР (GSF)*[1]. Европейская комиссия входит в состав двух рабочих групп:

- Рабочая группа GSF по устойчивости, с задачей обеспечить согласованность и взаимодополняемость с долгосрочными усилиями по обеспечению устойчивости, проводимыми на европейском уровне.
- Рабочая группа GSF по социально-экономическому воздействию научных инфраструктур: разработка принципов и процессов, которым необходимо следовать при рассмотрении социально-экономического воздействия научной инфраструктуры.

*Долгосрочная устойчивость научных инфраструктур*[2]. В Еврокомиссии разработаны стратегические документы и созданы экспертные группы для поддержки долгосрочных устойчивых инфраструктур, и определяются вопросы и меры для достижения этой цели:

- синхронизация национальных научных инфраструктур, дорожных карт по их развитию и бюджетов;
- обеспечение широкого доступа к научной инфраструктуре путем создания схем трансграничного доступа;
- запуск крупномасштабных пилотных проектов с участием научных инфраструктур и промышленности;
- использование данных, полученных европейскими научными инфраструктурами;
- создание эффективных механизмов управления на европейском уровне;
- поиск путей все более эффективного использования европейских структурных инвестиционных фондов;
- содействие международному распространению научных инфраструктур и их роли флагманов европейской научной политики.

### Информационные ресурсы и проекты открытой науки ЕС

Для организации научных коммуникаций, информационных ресурсов и научной инфраструктуры важнейшую роль играет переход к открытой науке. В ЕС функционирует ряд информационных структур, ориентированных на открытую науку.

[1] OECD Global Science Forum. – URL: https://www.oecd.org/sti/inno/global-science-forum.htm (дата обращения: 01.12.2021).

[2] Long-term sustainability. – URL: https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/long-term-sustainability_en (дата обращения: 01.12.2021).

***Европейское открытое научное облако EOSC***[1]

Это среда для размещения и обработки исследовательских данных для поддержки науки ЕС. EOSC обеспечивает создание доверенной, виртуальной, федеративной межстрановой и междисциплинарной среды для хранения, обмена, обработки и повторного использования научных цифровых объектов (таких как публикации, данные и программное обеспечение). В соответствии с *Европейской стратегией обработки данных*[2], EOSC является ядром информационного пространства науки, исследований и инноваций. График развития EOSC, установленный в европейской стратегии данных, предусматривает следующие этапы:

- после 2020 года создать обновленную, ориентированную на заинтересованные стороны структуру управления EOSC, возможно, в связи с запуском соответствующего Европейского партнерства EOSC в первом квартале 2020 года;
- к 2025 году развернуть операции EOSC для обслуживания исследователей ЕС;
- начиная с 2024 года объединить EOSC за пределами исследовательских сообществ, с более широким государственным и частным секторами.

EOSC является центральным элементом, поддерживающим циркуляцию, распространение и использование знаний в обновленном Европейском исследовательском пространстве, адекватном цифровой эпохе. Страны ЕС и страны, связанные с Horizon 2020, представленные в правлении EOSC, единогласно согласились запустить EOSC в качестве совместного программного Европейского партнерства в рамках Horizon Europe[3] с 2021 года.

В Европе ведущую роль по этому направлению играет программа ***OpenAIRE***[4], которая объединяет ряд проектов ЕС по открытой науке. Миссия OpenAIRE: обеспечить неограниченный, безбарьерный, открытый доступ к результатам исследований, оплачиваемых за счет государственного финансирования в Европе.

Рассмотрим далее структуры Евросоюза, непосредственно занятые языковыми технологиями и ЛИР.

***Объединенный исследовательский центр JRC***

В рамках JRC создан *Центр компетенций JRC по интеллектуальному анализу текста*[5]. JRC разработал инструменты языковой технологии для

[1] European Open Science Cloud (EOSC). – URL: https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en (дата обращения: 01.12.2021).

[2] European data strategy. – URL: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en (дата обращения: 01.12.2021).

[3] Horizon Europe – это программа финансирования исследований и инноваций комиссии, следующая за Horizon 2020 с 2021 года.

[4] Open Access Infrastructure for Research in Europe. – URL: https://www.openaire.eu/ (дата обращения: 01.12.2021).

[5] Competence Centre on Text Mining and Analysis. – URL: https://knowledge4policy.ec.europa.eu/text-mining_en (дата обращения: 01.12.2021).

более чем 20 языков, и с 2004 года анализирует до 300 000 онлайн-новостных статей в день, создавая таким образом ценные метаданные.

JRC также помогает распространять лингвистические ресурсы, выпускаемые другими организациями Евросоюза. Вот перечень ресурсов, поддерживаемых JRC:

- *DGT-Acquis*: Многоязычный параллельный корпус официального журнала ЕС
- *DCEP*: Цифровой корпус Европарламента
- *DGT:* Память перевода[1] Генеральной дирекции переводов
- *ECDC:* Память перевода Европейского центра профилактики и контроля заболеваний
- *EAC:* Память перевода Генеральной дирекции образования и культуры
- *JRC-Names*: Имена и варианты их написания
- *JRC Eurovoc:* Тезаурус Eurovoc

JRC призван обслуживать непосредственно органы управления ЕС и их терминологические службы. Что же касается координации деятельности по науке и технологиям в странах – членах ЕС, то для этого созданы упомянутые выше ERIC.

### Инфраструктурные консорциумы ERIC

***Общеевропейская исследовательская инфраструктура для языковых ресурсов и технологий CLARIN***[2]

CLARIN является специализированным ERIC в области языковых технологий. CLARIN обеспечивает доступность цифровых ЛИР для ученых, исследователей, студентов разных дисциплин, преимущественно гуманитарных и социальных наук, с помощью единого входа. CLARIN предлагает долгосрочные решения и технологические услуги для развертывания, подключения, анализа и поддержки цифровых ЛИР. CLARIN поддерживает ученых, которые хотят участвовать в передовых исследованиях на основе данных, внося свой вклад в действительно многоязычное европейское исследовательское пространство.

CLARIN руководствуется принципом, что все цифровые языковые ресурсы и инструменты во всей Европе и за ее пределами должны быть доступны через единый онлайн-вход для поддержки исследователей.

CLARIN также предлагает расширенные инструменты для поиска, исследования, использования, аннотирования, анализа или объединения ЛИР. Эти функции реализуются через сетевую федерацию центров: хранилищ языковых данных, сервисных центров и центров знаний с единым входом для всех членов академического сообщества во всех участвующих странах. Инструменты и данные из разных центров являются функционально совместимы-

[1] Подробнее о Памяти переводов см. в гл. 12.

[2] Common Language Resources and Technology Infrastructure. – URL: https://www.clarin.eu/ (дата обращения: 01.12.2021).

ми, поэтому сбор данных можно объединять, а инструменты из разных источников можно интегрировать для выполнения сложных операций. Инфраструктура CLARIN полностью функционирует во многих странах и продолжает создаваться в странах, которые присоединились совсем недавно. Перечислим основные сервисы CLARIN.

*Консорциумы-участники*. Большая часть операций, услуг и центров инфраструктуры CLARIN обеспечивается и финансируется членами CLARIN. Членами и наблюдателями могут быть страны или межправительственные организации. Они создают национальные консорциумы, обычно включающие университеты, исследовательские институты, библиотеки и публичные архивы, из которых по крайней мере один имеет статус центра CLARIN.

Все консорциумы придерживаются одинаковых критериев совместимости данных и услуг, условий доступа, качества данных и услуг. Функциональная совместимость обеспечивается с помощью стандартов, принятых в рамках CLARIN. Доступ к ЛИР соответствует принципам открытой науки. Члены и наблюдатели свободны решать, что именно они вносят в инфраструктуру CLARIN.

*Депозитные услуги*. Одним из основных сервисов инфраструктуры CLARIN является обеспечение архивирования ЛИР и предоставление их сообществу верифицированным способом, чтобы помочь исследователям устойчиво хранить свои ЛИР (например, корпуса, лексиконы, аудио- и видеозаписи, грамматики и т.д.). Многие центры CLARIN предлагают услугу депозита.

*Виртуальная языковая обсерватория (VLO)*. Цель – предоставить простой в использовании интерфейс, обеспечивающий единый процесс поиска большого количества ЛИР. Описание VLO см. выше, глава 2.

*Легкий доступ к защищенным ресурсам*. Благодаря федеративному входу в систему, защищенные приложения и наборы данных доступны всем, у кого есть учетная запись компьютера из многих европейских стран. Однако если нужно получить доступ к этим услугам из другой страны или из учреждения, которое не участвует в этих федерациях идентификации, можно запросить учетную запись CLARIN.

*Коммутатор языковых ресурсов*. Инструмент, который поможет найти соответствующий язык обработки веб-приложения для ваших данных. После загрузки файла или ввода URL-адреса вы можете выбрать, какую задачу выполнять. Затем коммутатор предоставит вам список доступных инструментов CLARIN для анализа ввода.

*Виртуальные коллекции*. Представляют собой последовательные наборы ссылок на цифровые объекты (например, размеченный текст, видео). Ссылки могут происходить из разных архивов, отсюда и термин «виртуальный».

*Реестр*. CLARIN предоставляет реестр, где ученые могут создавать и публиковать свои виртуальные коллекции. Он тесно интегрирован с инфраструктурой и обеспечивает постоянные идентификаторы и федеративный вход в систему. Метаданные коллекции доступны через Виртуальную языковую обсерваторию.

*Инвентаризация ЛИР*. Предоставляется инструмент, удобный для инвентаризации (каталогизации) ЛИР. Данный каталог отличается от других депозитных услуг тем, что нет необходимости загружать данные или метаданные (достаточно ссылки на веб-сайт и описание), и его можно использовать немедленно, без предварительного обращения в хостинг-центр.

*Федеративный поиск***.** Чтобы обеспечить исследователям возможность расширенного поиска с использованием конкретных моделей коллекции данных, CLARIN предлагает поисковую систему (пока прототип) данных, которые доступны в центрах хранения. Сами данные остаются у владельца, поэтому поиск называется федеративным.

*CLARIN для исследователей*. CLARIN для исследователей – это онлайновая коллекция учебных материалов, тематических исследований и контактов с экспертами из всей сети CLARIN, которые предназначены для исследователей и студентов всех ступеней, что работают в области цифровой гуманитаристики.

*Ресурсные семьи*. Целью данной инициативы CLARIN является предоставление обзоров, доступных ЛИР для исследователей в области цифровой гуманитаристики и NLP.

*Реестр курсов по цифровой гуманитаристике*. Содержит выбор курсов, предлагаемых европейскими академическими организациями. Студенты, преподаватели и исследователи могут осуществлять поиск в базе данных на основе дисциплин, местоположения, кредитов ECTS или присуждаемых академических степеней.

*Мероприятия.* CLARIN ежегодно проводит 12 конференций, семинаров и др.

*Обмен знаниями.* В CLARIN создается инфраструктура обмена знаниями, включающая технические средства, которые предоставляют пользователям доступ к данным и инструментам людям, управляющим этими средствами, стандартами, условиями доступа, лицензиями, обеспечением качества и т.д. Основу инфраструктуры обмена знаниями составляют Центры знаний CLARIN[1].

*Центры знаний CLARIN (К-центры*) – краеугольный камень инфраструктуры знаний CLARIN; это учреждения, которые согласились поделиться своими знаниями и опытом по одному или нескольким аспектам области, охватываемой инфраструктурой CLARIN.

В центре внимания CLARIN находятся языковые ресурсы (во всех модальностях, из всех регионов и с любой тематической направленностью). Центры знаний обслуживают исследователей и преподавателей разных дисциплин, где язык играет одну из своих многочисленных ролей, начиная от объекта изучения и заканчивая средством общения или выражения, средством для хранения и извлечения информации, объектом обучения или преподавания, учебным источником для аналитики на основе данных и многим

[1] Knowledge Centres // CLARIN. – URL: https://www.clarin.eu/content/knowledge-centres (дата обращения: 01.12.2021).

другим. Центры знаний можно найти в основном в странах CLARIN, но не только: их виртуальное присутствие наблюдается и в других местах.

Все Центры знаний имеют свои специфические области знаний, которые могут относиться ко многим различным категориям, например:

- отдельные языки (например, датский, чешский, португальский), языковые семьи (например, южнославянские) или группы языков (например, морфологически богатые языки, языки Швеции);
- письменный текст и формы речи, отличные от письменного текста (например, устный язык, язык жестов);
- лингвистические темы (например, языковое разнообразие, изучение языков, диахронические исследования);
- темы языковой обработки (например, анализ речи, построение древовидных структур, машинный перевод);
- типы данных, отличные от корпусов (например, лексические данные, словарные сети, банки терминологии);
- использование или обработка семейств языковых данных, общих для большинства языков (например, газет, парламентских отчетов, устной истории);
- общие методы и проблемы (например, управление данными, этика, права интеллектуальной собственности, OCR).

Услуги, предлагаемые Центрами знаний, могут принимать разные формы. У всех есть служба поддержки, которая ответит на запросы в течение двух рабочих дней. Некоторые предлагают онлайн-курсы, некоторые – документы с лучшими практиками, некоторые – рекомендации по получению доступа к данным и инструментам и их использованию; некоторые готовы принять получателей грантов CLARIN на мобильность и еще много других моделей. Чтобы получить статус Центра знаний CLARIN, необходимо пройти определенную сертификацию.

Аннотированный перечень центров знаний приводится в приложении 5. Они также перечисляются на странице сайта CLARIN[1], с указанием сферы компетенции каждого из них. Более полную информацию о каждом центре можно получить, пройдя по соответствующей ссылке. Для каждого центра на портале CLARIN указываются следующие данные:

- полное имя
- короткое имя
- URL
- контактное лицо, почта для связи, адрес размещения (организация, город и страна)
- дата сертификации
- сфера компетенции, аудитории
- виды услуг
- основной язык и другие языки

[1] Overview of CLARIN K-centres, ordered by acronym. – URL: http://vonweber.elsnet.org/cgi/kcentres_page.cgi (дата обращения: 01.12.2021).

- охватываемые формы
- лингвистические темы
- технологии NLP
- типы лингвистических данных
- семейства ЛИР
- ключевые слова
- ссылка на интервью с руководителем
- дата последнего обновления

Некоторые центры знаний реализованы одновременно в нескольких учреждениях и странах.

*Стратегия развития CLARIN на 2021–2023 гг*. CLARIN предлагает план развития[1], включающий четыре приоритетные области, причем по каждой области предлагаются конкретные задачи.

- *Инфраструктура знаний.* Углубление мониторинга ресурсов, технологий и опыта структур, входящих в CLARIN; увеличение способов финансирования обмена и распространения лучших практик.
- *Техническая инфраструктура.* Поддержка и развитие системы идентификаторов; контроль качества метаданных, как автоматический, так и интеллектуальный.
- *Организационное развитие*. Поддержка сотрудничества на уровне координации, технологии и финансов, развитие инструментов обмена информацией.
- *Устойчивость.* Расширение состава, в том числе за счет неевропейских стран, финансовая диверсификация, новые формы сотрудничества с индустрией.

Кроме CLARIN, имеется еще один ERIC, который развивает языковые технологии.

***Цифровая исследовательская инфраструктура для искусств и гуманитарных наук DARIAH***[2]

DARIAH – это сеть людей, экспертных знаний, информации, знаний, контента, методов, инструментов и технологий из стран – членов организации. DARIAH разрабатывает, поддерживает и эксплуатирует инфраструктуру в научно-исследовательской деятельности, основанной на IT, и предоставляет ее исследователям для создания, анализа и интерпретации цифровых ресурсов. Более подробное описание DARIAH представлено в главе 20.

[1] CLARIN Strategy 2021–2023 at a Glance. – URL: https://www.clarin.eu/content/vision-and-strategy (дата обращения: 01.12.2021).

[2] The Digital Research Infrastructure for the Arts and Humanities (DARIAH)-EU. – URL: https://www.dariah.eu (дата обращения: 01.12.2021).

## Европейские объединения и проекты

Кроме Консорциумов ERIC, в Европе создано много других профессиональных международных объединений, имеющих форму ассоциаций, советов, консорциумов других видов и просто временных проектных сообществ, которые работают в области языковых технологий.

Причем некоторые проекты институализируются в виде постоянно действующих структур или неформальных сообществ. В настоящем разделе описаны 15 основных объединений и коллективных проектов в этой сфере.

***Европейская ассоциация лингвистических ресурсов ELRA***

Начнем с описания *ELRA*[1], тем более что недавно подписано соглашение о сотрудничестве ELRA и CLARIN, которое включает ELRA в состав органов Еврокомиссии.

Основанная в 1995 году, ELRA является некоммерческой организацией, миссия которой заключается в том, чтобы сделать общедоступными ЛИР, применяемые в технологиях NLP.

Для достижения этой цели ELRA проводит такие действия, связанные с ЛИР, как идентификация ЛИР, распространение, производство, проверка, оценка технологий, информирование о ЛИР. При ELRA создано *Агентство по оценке и распространению ЛИР*[2] (ELDA), которое является оперативным органом ассоциации и реализует выполнение задач ELRA, определенных Советом ассоциации, в том числе коммерческих проектов.

ELRA и ELDA реализуют следующие сервисы на основе ЛИР:

- идентификация и каталогизация ЛИР
- распространение ЛИР
- производство ЛИР
- проверка и оценка ЛИР
- поддержка правовых вопросов, связанных с ЛИР

*Идентификация*. За последние несколько лет усилия ELRA по идентификации ЛИР значительно возросли. Это напрямую связано с акцентом деятельности ELRA на выявление уже существующих ресурсов, благодаря чему достигается оптимизация усилий по созданию уже доступных ресурсов.

Для задач идентификации разработан *Международный стандартный номер языковых ресурсов* (*ISLRN*)[3] – уникальная и универсальная модель идентификации для ЛИР, использующая стандартизированные метаданные.

*Каталогизация.* В каталоге ELRA ЛИР распределены по четырем категориям: «Речь и смежные ресурсы», «Письменные ресурсы», «Терминологические ресурсы» и «Мультимодальные / мультимедийные ресурсы». Описание каталога имеется в главе 1.

---

[1] European Language Resources Association (ELRA). – URL: http://www.elra.info/en/ (дата обращения: 01.12.2021).

[2] The Evaluations and Language resources Distribution Agency (ELDA). – URL: http://www.elra.info/en/about/elda/ (дата обращения: 01.12.2021).

[3] International Standard Language Resource Number (ISLRN). – URL: http://www.elra.info/en/islrn/ (дата обращения: 01.12.2021).

*META-SHARE*. Под эгидой ELRA разработан META-SHARE, сервис обмена и распространения открытых ЛИР, который направлен на расширение доступа к таким ресурсам в глобальном масштабе. META-SHARE – это открытый, интегрированный, безопасный и совместимый сервис совместного использования и обмена для ЛИР (наборов данных и инструментов) для автоматической обработки естественного языка (NLP). META-SHARE спроектирована как сеть распределенных репозиториев ЛИР, включая языковые данные и базовые инструменты NLP. См. также главу 6.

*Карта LRE*. Для каталогизации ЛИР разработана модель их описания (LRE map). В соответствии с этой схемой описано свыше 6 тыс. ЛИР. Подробно карта LRE описана в главе 6.

*Конференция по языковым ресурсам и их оценке (LREC)*. Инициатива LREC «Поделитесь своими ЛИР», инициированная в 2014 году в Рейкьявике и продолжавшаяся в 2016 году в Порторожe, имела большой успех. Наборы общих ЛИР были проверены вручную, и теперь доступны очищенные версии списков ЛИР, которые включают доступные ЛИР – наборы данных.

*Многоязычный инструментарий анонимизации для государственной администрации (MAPA)*[1]

Конечная цель MAPA – разработать полностью развертываемый многоязычный набор анонимизации имен на основе распознавания именованных сущностей, применимый ко всем языкам ЕС, и с подключением к eTranslation, независимо от того, является ли текст одноязычным, двуязычным или смешанным.

Это далеко не все проекты ELRA по языковым технологиям; на портале ELRA есть список законченных проектов, в которых принимала участие ELRA[2].

***Трансевропейская инфраструктура лингвистических ресурсов TELRI***[3]

TELRI – это общеевропейский альянс, состоящий из 28 основных национальных языковых (технологических) учреждений с акцентом на страны Центральной и Восточной Европы и СНГ. От России в состав TELRI входит Институт русского языка РАН.

Текущий проект – TELRI II. Основные цели TELRI II:

- укрепить общеевропейскую инфраструктуру для сообщества многоязычных исследователей и разработчиков;
- собирать, продвигать и предоставлять одноязычные и многоязычные ЛИР, включая инструменты для NLP;
- предложить комплексный сервис для академических и промышленных пользователей;

[1] The Multilingual Anonymization Toolkit for Public Administrations (MAPA). – URL: http://www.elra.info/en/projects/current-projects/mapa/ (дата обращения: 01.12.2021).

[2] Completed and Archived Projects. – URL: http://www.elra.info/en/projects/archived-projects/ (дата обращения: 01.12.2021).

[3] Trans-European Language Resources Infrastructure (TELRI). – URL: http://telri.nytud.hu/ (дата обращения: 01.12.2021).

- организовывать НИР и ОКР в области NLP;
- обеспечить форум, где эксперты из академических кругов и индустрии делятся и оценивают инструменты, ресурсы и новые тенденции, исследуют альтернативные подходы и участвуют в совместной деятельности;
- предоставлять опыт своих партнеров исследовательскому сообществу, общественности и языковой индустрии.

Задачи и функции TELRI могут быть описаны через функции рабочих групп (РГ) этой организации:

- *РГ 1 Координация*. Финансовое, административное и договорное управление. Научный менеджмент.
- *РГ 2 TELRI Сеть*. Подготовка, продвижение и организация семинаров TELRI. Информационный бюллетень, веб-страница. Сетевая организация TELRI. Рекламная деятельность. Пленарные заседания TELRI. Сотрудничество с партнерами из стран, не являющихся членами TELRI.
- *РГ 3:* TRACTOR[1] *Сервис.* Продвижение TRACTOR для академических и промышленных языковых ресурсов / сообщества языковых технологий. Реализация сети TRACTOR с использованием сети ELAN. Администрирование сети. Сервисный телефон доверия. Сервисный справочник. Информация о ресурсах и программном обеспечении, проекты, опыт и участие членов TELRI II. Каталог метаданных существующих ресурсов, предлагаемых партнерами проекта.
- *РГ 4 TRACTOR Инструменты и ресурсы*. Координация приобретения инструментов и ресурсов для TRACTOR. В частности, установление стандартов, которым должны соответствовать инструменты и их документация. Приобретение программных средств. Приобретение ресурсов.
- *РГ 5. Организация совместных исследований*. Целью РГ 5 является разработка, структурирование и подготовка многосторонних исследовательских и опытно-конструкторских проектов, отвечающих требованиям европейского многоязычного сообщества NLP. Основные темы: извлечение лингвистических знаний из текстовых и / или лексических ресурсов для задач в области многоязычной лексикографии, средства перевода, идентификация терминологии и средства поиска информации.

***Сеть содействия языковым ресурсам FLaReNet***[2]

Миссия FLaReNet заключается в определении приоритетов, а также краткосрочных, среднесрочных и долгосрочных стратегических целей, и предоставлении согласованных рекомендаций в форме плана действий для ЕС, национальных организаций и промышленности в области языковых ресурсов и технологий. В составе FLaReNet действуют следующие рабочие группы:

- РГ 1 – Управление и распространение информации
- РГ 2 – Анализ и мониторинг устойчивости для ЛИР
- РГ 3 – Методы и модели для ЛИР

[1] TRACTOR – TELRI Archive of Computational Tools and Resources.

[2] Fostering Language Resources Network (FLaReNet). – URL: http://www.elra.info/en/projects/archived-projects/flarenet/ (дата обращения: 01.12.2021).

- РГ 4 – Гармонизация форматов и стандартов
- РГ 5 – Протоколы и процедуры оценки
- РГ 6 – Автоматическая конструкция ЛИР
- РГ 7 – Эволюционирующая дорожная карта
- РГ 8 – План действий и инфраструктура

В деятельности FLaReNet участвуют 40 организаций-партнеров, около 100 институциональных членов, свыше 400 индивидуальных членов, а также группы поддержки в 25 организациях – в основном из европейских стран, но не только. FLaReNet продолжает работу ISLE и EAGLES.

***Европейская координация языковых ресурсов ELRC***[1]

Проект ELRC реализуется с 2014 г. в рамках программы *Connecting Europe Facility* (CEF)[2]. Он нацелен на все страны, связанные с CEF, т.е. на 28 государств – членов ЕС плюс Норвегия и Исландия. Общая цель ELRC заключается в сборе ЛИР от администраций государственных служб во всех странах, входящих в CEF, с тем чтобы улучшить качество, охват и производительность создаваемой по программе CEF системы машинного перевода[3].

На первом этапе основные результаты, достигнутые с помощью ELRC, включали 225 ЛИР: 138 двуязычных / многоязычных корпусов, 50 терминологий и 37 моноязычных корпусов. В соответствии с требованиями контракта ELRC удалось охватить все языки необходимыми типами ЛИР для каждого языка. Кроме того, ELRC провела необходимую оценку и валидацию ЛИР, чтобы обеспечить их качество и пригодность для целей машинного перевода. Все ЛИР, собранные ELRC, были загружены в репозиторий ELRC-SHARE.

Органом управления ELRC является Совет языковых ресурсов. Он состоит в общей сложности из 60 членов, включая представителей от стран. В целом, национальные службы играют ключевую роль на национальном уровне в эффективной мобилизации государственного сектора и поощрения и облегчения вклада языковых ресурсов между государственными органами и министерствами в каждой стране. Они являются необходимыми мостами между консорциумом и соответствующими игроками в каждой стране, обеспечивая тем самым эффективность задач проекта через местную ответственность и связь с национальным сообществом.

2 февраля 2020 года ELRC вступил в свой третий этап – ELRC3[4]. В ходе ELRC через репозиторий ELRC-SHARE были доступны впечатляющие 1337 уникальных языковых ресурсов. Это происходит благодаря

[1] European Language Resource Coordination (ELRC) – supporting Multilingual Europe. – URL: https://lr-coordination.eu/ (дата обращения: 01.04.2022).

[2] Connecting Europe Facility (CEF)). – URL: https://ec.europa.eu/inea/en/connecting-europe-facility (дата обращения: 01.04.2022).

[3] eTranslation. – URL: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation (дата обращения: 01.12.2021).

[4] European Language Resources Coordination (ELRC) 3. – URL: http://www.elra.info/en/projects/current-projects/elrc3/ (дата обращения: 01.12.2021).

выдающейся поддержке и вкладу различных государственных администраций, исследователей и компаний по всей Европе, а также сети национальных опорных пунктов ELRC.

«Новый» ELRC выйдет далеко за рамки сбора языковых ресурсов: он преодолеет разрыв с европейской индустрией языковых технологий, поддерживая адаптацию последних исследовательских инноваций и искусственного интеллекта к местным потребностям и продвигая существующие решения в области языковых технологий.

***Европейская Языковая Сеть ELG***[1]

Основная цель проекта заключается в решении проблемы фрагментации в европейском бизнесе и исследованиях в области лингвистических технологий (LT) путем создания европейской языковой сети в качестве основной платформы для LT в Европе и укрепления европейского бизнеса в области LT в отношении конкуренции с другими континентами. Проект будет развертывать и наполнять ELG в качестве платформы для коммерческих и некоммерческих LT, как функциональных (запущенные службы и инструменты), так и нефункциональных (наборы данных, ресурсы, модели). Использование ELG для идентификации, поиска, обнаружения, получения и интеграции LT-услуг, наборов данных и ресурсов создаст множество преимуществ для компаний, использующих и / или предоставляющих LT, а также для некоммерческих, академических организаций.

Это позволит коммерческому и некоммерческому европейскому сообществу LT депонировать и загружать свои технологии и наборы данных в ELG, развертывать их через Сеть и подключаться к другим ресурсам. ELG будет стимулировать многоязычный цифровой единый рынок европейского LT-сообщества, создавая новые рабочие места и возможности. Будет оказана финансовая поддержка до 20 пилотных проектов, чтобы продемонстрировать полезность ELG.

На основе ELG будут созданы более 30 национальных центров компетенции и Европейский совет по LT для европейской координации. ELG будет способствовать развитию LT, адаптированных к языкам и культурам, а также к социальным и экономическим требованиям стран ЕС.

***Сеть передового опыта META-NET***[2]

META-NET состоит из 60 исследовательских центров из 34 стран. Сеть направлена на создание технологических основ многоязычного европейского информационного общества. META-NET создает META, многоязычный европейский технологический альянс. META объединяет исследователей, поставщиков коммерческих технологий, частных и корпоративных пользователей LT, специалистов в области языков и других участников информационного общества. META инициирует амбициозные совместные

[1] European Language Grid (ELG). – URL: http://www.elra.info/en/projects/current-projects/european-language-grid/ (дата обращения: 01.12.2021).

[2] Multilingual Europe Technology Alliance (META). – URL: http://www.meta-net.eu/ (дата обращения: 01.12.2021).

международные усилия по развитию LT в Европе, объединенной единым цифровым рынком и информационным пространством.

Для реализации приложений, обеспечивающих автоматический перевод, многоязычную информацию и управление знаниями, а также создание контента на всех европейских языках, необходимы согласованные, существенные, общеконтинентальные усилия в области исследования и разработки LT. Эти усилия также будут способствовать развитию интуитивно понятных языковых интерфейсов для различных технологий, от бытовой электроники, машин и транспортных средств до компьютеров и роботов.

***Европейская сеть лингвистических данных, ориентированная на Интернет, NexusLinguarum***[1]

Основная цель заключается в европейском сотрудничестве между лингвистами, компьютерными специалистами, терминологами и другими заинтересованными сторонами с целью исследования и расширения области лингвистической науки о данных. Основные направления:

- продвижение моделей на основе открытых связанных данных для лингвистических данных, таких как Ontolex-Lemon, в соответствии с фактическими, общественными и официальными стандартами;
- разработка расширений существующих моделей для поддержки областей, связанных с диахроническими и социальными языковыми вариациями;
- обсуждение, проверка и консолидация существующих передовых практик в области связанных открытых лингвистических данных (LLOD);
- определение технологий для преобразования ЛИР в LLOD;
- внедрение поддерживаемых сообществом процедур сравнительного анализа и оценки качества для непрерывного мониторинга качества LLOD;
- разработка новых совместных методологий для создания, связывания и улучшения LR на протяжении их жизненного цикла экономичным способом;
- разработка расширений методов на основе машинного обучения / глубокого обучения для обнаружения особенностей лингвистических данных в больших объемах различных типов многоязычных текстовых данных.

***Европейское отделение Ассоциации компьютерной лингвистики EACL***[2]

EACL – профессиональная ассоциация компьютерной лингвистики в Европе. Основная деятельность – организация конференций раз в два года, а также поддержка образовательных инициатив в этой области, например, вводные курсы по компьютерной лингвистике в летних школах, европейские программы магистров речи и языка и стипендии в специализированных мастерских.

Конференции EACL происходят регулярно, с 400–500 участниками и 75–80 докладами.

---

[1] NexusLinguarum. – URL: https://nexuslinguarum.eu/ (дата обращения: 01.12.2021).

[2] Association for Computational Linguistics, European Chapter (EACL). – URL: http://eacl.org/ (дата обращения: 01.12.2021).

***Европейская ассоциация цифровой гуманитаристики EADH***[1]

Европейская ассоциация цифровых гуманитарных наук (EADH) была основана в 1973 году под названием Association for Literary and Linguistic Computing (ALLC), с первоначальной целью поддержки применения компьютеров в изучении языка и литературы. По мере расширения спектра доступных и актуальных информационных технологий в гуманитарных науках, интересы членов ассоциации существенно расширились и охватывают не только анализ текстов и языковые корпуса, но и историю, искусствоведение, музыку, обработку рукописей и изображений, электронные издания. Новое название ассоциации, принятое в 2012 году, отражает это значительное расширение сферы охвата. Сегодня миссия EADH заключается в том, чтобы представлять цифровые методы во всех гуманитарных дисциплинах.

***Консультативная группа экспертов по стандартам языковых технологий EAGLES***[2]

EAGLES была организована по инициативе Европейской комиссии в рамках программы лингвистических исследований. EAGLES была направлена на ускорение создания и внедрения стандартов в следующих областях:

- крупномасштабные ЛИР (такие как корпуса текстов и речи, цифровые лексиконы);
- средства манипулирования ЛИР с помощью формальных моделей, языков разметки и различных программных инструментов;
- средства оценки ресурсов, инструментов и продуктов.

Многочисленные известные компании, исследовательские центры, университеты и профессиональные организации по всему Европейскому Союзу сотрудничают для разработки рекомендаций для стандартов де-факто и передовой практики в вышеуказанных областях языковой инженерии. Работу над общими техническими стандартами вели пять рабочих групп:

- Текстовые корпуса
- Вычислительные лексиконы
- Грамматические формализмы
- Оценка
- Устная речь

Продолжателем деятельности EAGLES является консорциум FLaReNet.

***Европейское лингвистическое общество SLE***[3]

SLE была основана как платформа для свободного обмена мнениями и дискуссий идей, чтобы обеспечить рост и обогащение лингвистики, без привязки к какой-либо конкретной научной школе. SLE организует ежегодные совещания и выпускает публикации. На сайте общества приводятся сведения о членах, президентах и собраниях SLE.

---

[1] European Association for Digital Humanities (EADH). – URL: https://eadh.org/ (дата обращения: 01.12.2021).

[2] Expert Advisory Group on Language Engineering Standards (EAGLES). – URL: http://www.ilc.cnr.it/EAGLES/home.html (дата обращения: 01.12.2021).

[3] The Linguistic Society of Europe (SLE). – URL: https://societaslinguistica.eu/ (дата обращения: 01.12.2021).

***Многоязычные текстовые инструменты и корпуса для языков Центральной и Восточной Европы MULTEXT-East***[1]

Ресурсы MULTEXT-East представляют собой многоязычный набор данных для лингвистических инженерных исследований и разработок. Он состоит из:

- MULTEXT – восточных морфосинтаксических спецификаций, определяющих категории (части речи), их морфосинтаксические особенности (атрибуты и значения) и компактные представления набора тегов MSD
- морфосинтаксической лексики
- аннотированного параллельного корпуса «1984»
- некоторых сопоставимых текстовых и речевых корпусов

Спецификации доступны для таких макроязыков, языков и разновидностей языков, как: албанский, болгарский, чеченский, чешский, дамаскини, английский, эстонский, венгерский, македонский, персидский, польский, риянский, румынский, русский, сербохорватский, словацкий, словенский, торлак и украинский.

***Многоязычный доступ к информации по Ковид-19 MLIA***[2]

Проект направлен на ускоренное создание ресурсов и инструментов для улучшения многоязычного доступа широкой публики к информации о социальных, экономических или политических аспектах, связанных с пандемией, в том числе по извлечению информации, многоязычному семантическому поиску и машинному переводу. С момента своего запуска в июне 2020 года проект достиг огромного прогресса в сборе языковых ресурсов и инструментов для поддержки разработки приложений и услуг в связи с пандемией COVID-19. В настоящее время несколько сотен ресурсов доступны в репозитории ELRC COVID-19 под международной лицензией Creative Commons Attribution-ShareAlike 4.0. Кроме того, предусмотрены учебные и тестовые корпуса для выполнения перечисленных трех задач.

Как мы уже отмечали, в предложенном обзоре в ряде случаев лишь перечислены организации и проекты, связанные с языковыми технологиями и ЛИР в Европе, без их подробного анализа. В некоторых случаях не вполне ясно соотношение этих организаций и проектов между собой, и возникает впечатление некоего дублирования в деятельности организаций. Особенно это касается бюрократических структур, которых слишком много для эффективного управления отраслью. Впрочем, излишняя бюрократизация вообще свойственна структурам Евросоюза, что отмечают многие наблюдатели. Тем не менее масштабы и результаты деятельности европейских структур в области языковых технологий, несомненно, впечатляют.

[1] Multilingual Text Tools and Corpora for Central and Eastern European Languages (MULTEXT-East). – URL: http://nl.ijs.si/ME/ (дата обращения: 01.12.2021).

[2] COVID-19 Multilingual Information Access initiative (MLIA). – URL: http://www.elra.info/en/projects/current-projects/covid-19-mlia-init/ (дата обращения: 01.04.2022).

# ЧАСТЬ 2
# ТЕХНОЛОГИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

## ГЛАВА 5. МЕЖДУНАРОДНАЯ СТАНДАРТИЗАЦИЯ ЯЗЫКОВЫХ РЕСУРСОВ И ТЕХНОЛОГИЙ

### Вводные замечания

Международная организация по стандартизации (ISO) определяет стандарт как «документ, содержащий требования, спецификации, руководящие принципы или характеристики, которые могут использоваться последовательно для обеспечения соответствия материалов, продуктов, процессов и услуг своему назначению».

Обычно стандарт, разрабатываемый органом по стандартизации, должен отвечать некоторым строгим требованиям и правилам, установленным группой экспертов. После прохождения процесса публичной проверки члены организации по стандартизации (например, представители правительственных, промышленных или академических организаций) должны согласиться с тем, что стандарт может быть опубликован. Хотя процесс стандартизации занимает много времени, он обеспечивает прозрачную и справедливую разработку стандартов с учетом множества точек зрения и потребностей всех членов стандартной организации. Официальные стандарты, разработанные в рамках одного аккредитованного органа, такие как ISO, DIN, IEEE, CEN/ISSS или NISO, обычно определяются как стандарты де-юре.

С другой стороны, спецификация – это «явный набор требований, которым должен удовлетворять материал, продукт, система или услуга». Любое частное лицо, компания или организация могут разработать спецификацию, которая обычно ограничивается конкретным приложением и определяет задачи и цели этого приложения. Когда спецификация часто используется и признается пользователями больше, чем любая другая существующая спецификация, ее называют стандартом де-факто. Многие спецификации, такие как PDF, CMDI, HTML, были разработаны вне органа стандартизации, но в силу широкого распространения они были признаны равноправными стандартам ISO. В настоящей главе будут кратко описаны международные стандарты де-юре и спецификации в области языковых технологий и ресурсов, которые широко используются международным сообществом прикладных лингвистов.

## Международные органы по стандартизации

Разработкой стандартов и спецификаций для языковых ресурсов и технологий занимаются многие международные организации. Ниже мы перечислим те из них, которые завоевали безусловный авторитет в сообществе. Главным образом мы будем опираться на информацию, представленную в Информационной системе по стандартизации CLARIN[1], но будем в некоторых случаях сокращать или дополнять эту информацию.

### *Международная организация по стандартизации ISO*[2]

Основная роль в стандартизации инструментов и ресурсов, используемых в языковой индустрии, принадлежит техническому Комитету 37 Международной организации по стандартизации «Язык и терминология» (ISO/ТК 37). Сфера деятельности этого Комитета – стандартизация описаний, ресурсов, технологий и услуг, связанных с терминологией, письменным, устным переводом и другими языковыми видами деятельности в многоязычном информационном обществе.

В состав ТК 37 входят пять подкомитетов (приводим наименования подкомитета (ПК) на русском языке, который является одним из трех официальных языков ISO):

ПК1 Принципы и методы

ПК2 Рабочий процесс терминологии и языковое кодирование

ПК3 Управление терминологическими ресурсами

ПК4 Управление языковыми ресурсами

ПК5 Письменный, устный перевод и сопутствующие технологии

В России аналогичные функции выполняет Технический комитет по стандартизации ТК 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле».

Всего в ТК 37 по состоянию на январь 2021 года разработано и действует 70 стандартов, и еще 31 находится в стадии разработки. Отмененные стандарты не учитываются.

Основными подкомитетами ТК 37, связанные со стандартизацией ЛИР, являются *ПК3 Управление терминологическими ресурсами* и *ПК4 Управление языковыми ресурсами.* Приведем состав рабочих групп этих подкомитетов.

ISO/ТК 37/ПК 3 «Управление терминологическими ресурсами» включает следующие рабочие группы (РГ):

РГ 1 Категории данных

РГ 3 Обмен данными

РГ 4 Управление базами данных

РГ 5 Извлечение терминологии

ISO/ТК 37/ПК 4 «Управление языковыми ресурсами» включает следующие рабочие группы:

---

[1] CLARIN Standards Information System. – URL: https://clarin.ids-mannheim.de/standards/index.xq (дата обращения: 01.12.2021).

[2] International Organization for Standardization (ISO). – URL: http://www.iso.org/ (дата обращения: 01.12.2021).

РГ 1 Основные дескрипторы и механизмы для языковых ресурсов
РГ 2 Семантическая аннотация
РГ 3 Многоязычное представление информации
РГ 4 Лексические ресурсы
РГ 5 Рабочий процесс управления языковыми ресурсами
РГ 6 Лингвистическая аннотация

Общая структура семейства стандартов, разработанных этими подкомитетами ТК 37, представлена ниже.

*ПК 3 Управление терминологическими ресурсами*

Системы управления терминологией, знаниями и содержанием
Аспекты, связанные с концепцией разработки и интернационализации систем классификации
Компьютерные приложения в терминологии
Структура терминологической разметки
Управление терминологическими ресурсами
Спецификации категорий данных
Категории данных (в 2 ч.)
Извлечение терминологии
Терминологические базы данных (в 3 ч.)
TBX (termbase exchange)
Обмен терминологическими базами
Представление концептуальных отношений и предметных областей в TBX
Создание схемы для ядра TBX

*ПК 4 Управление языковыми ресурсами*

Управление языковыми ресурсами
Структуры функций
Структуры компонентов
Инфраструктура метаданных компонентов (CMDI) (в 2 ч.)
Структура лексической разметки (LMF) (в 5 ч.)
Пословная сегментация письменных текстов (в 2 ч.)
Лингвистические аннотации
  Структура лингвистических аннотаций (LAF)
  Морфосинтаксическая структура аннотаций (MAF)
  Структура синтаксических аннотаций (SynAF) (в 2 ч.)
  Структура семантических аннотаций (SemAF) (в 14 ч.)
  Комплексная структура аннотаций (ComAF) (в 2 ч.)
Постоянная идентификация и устойчивый доступ (PISA)
Контролируемое человеческое общение (CHC) (в 4 ч.)
Контролируемый естественный язык (CNL)
Транскрипция разговорного языка
Корпусный язык запросов Lingua Franca (CQLF) (в 2 ч.)
Многоязычная информационная структура

Полный аннотированный перечень стандартов, разработанных ISO/ТК 37 и его подкомитетами, а также проектов стандартов доступен на сайте

ТК 37[1]. Там же имеется список членов и наблюдателей ТК 37. Всего в этом ТК – 35 членов и 27 наблюдателей. Российскую Федерацию представляет Росстандарт (ГОСТ Р).

***Консорциум Всемирной сети W3 C***[2]

Это международная организация по стандартизации для Всемирной сети, основанная в октябре 1994 года Тимом Бернерсом-Ли. Одним из важных направлений работы W3 C является разработка технических спецификаций и руководств, таких как HTML, XHTML и XML, в которых описаны протоколы связи и другие архитектурные блоки Интернета. В настоящее время существует более 90 стандартов W3 C, которые W3 C называет рекомендациями.

Среди нормативных документов W3 C есть ряд документов, регламентирующих представление ЛИР, в частности в формате связанных данных, например *Рекомендации по созданию лингвистических связанных данных: двуязычные словари* [1].

***Инициатива по кодированию текстов TEI***[3]

В настоящее время среди групп специальных интересов TEI (SIG) существует SIG «TEI для лингвистов». Ключевыми задачами являются разработка набора рекомендаций по кодированию языковых ресурсов с помощью разметки TEI. SIG «TEI для лингвистов» сотрудничает с исследователями, работающими в рамках ISO/TC 37 / SC4, что делает их работу эффективной и действенной.

***Международные стандарты для языковых технологий ISLE***[4]

Перед исполнителями этого проекта, который послужил основой для разработки ряда стандартов ISO, была поставлена цель улучшения доступности ЛИР при помощи стандарта описания метаданных мультимедийных / мультимодальных ЛИР. С помощью такого стандарта станет возможным создание видимого и доступного для поиска пространства таких ресурсов в Интернете. Это позволит заинтересованным сторонам оперативно находить подходящие ресурсы и повысит уровень их (ресурсов) повторного использования.

Многие ЛИР создаются в таких дисциплинах, как корпусная лингвистика, антропология, речевая инженерия, но далеко не все доступны через такие известные каталоги, как LDC и ELRA. Кроме того, большинства этих ресурсов вообще нет в открытом доступе, и только очень немногие люди

[1] Technical Committees ISO/TC 37: Language and terminology. – URL: https://www.iso.org/committee/48104.html (дата обращения: 01.12.2021).

[2] The World Wide Web Consortium (W3 C). – URL: http://www.w3.org/ (дата обращения: 01.12.2021).

[3] The Text Encoding Initiative (TEI). – URL: http://www.tei-c.org (дата обращения: 01.12.2021). Описание TEI было представлено выше.

[4] International Standard for Language Engineering (ISLE). – URL: https://www.mpi.nl/ISLE/ (дата обращения: 01.12.2021).

знают о них. Хорошо известно, что даже в учреждениях, где эти ресурсы генерируются, систематическое информирование об их существовании, обновлении, удалении по той или иной причине является весьма проблематичным.

Ситуация, описанная выше, стала стимулом для инициативы ISLE *Meta Data Initiative*. Сообщество языковых ресурсов нуждалось в стандарте для описания основных характеристик ресурсов, таких, как в случае корпусов: название языка, на котором говорят, возраст говорящих, пол, образование и т.д. Сообществу также требовались инструменты, помогающие легко генерировать такие метаописания. Предпочтительно, чтобы эти описания были доступными в Интернете и интегрировались в формирующееся пространство метаописаний, что позволяет пользователям просматривать информационно-профессиональное пространство, иметь возможность полифункционального поиска в нем и, наконец, получить доступ к самим ресурсам.

Эта инициатива установила требования к используемым инструментам и определила возможные практические сценарии применения, например связь метаописаний с местом хранения, гарантии их качества, методы верификации и т.д.

Проект функционировал в 2000–2003 гг.

***Инициатива по метаданным Дублинского ядра DCMI***[1]

DCMI – это открытая международная организация, целью которой является разработка интероперабельного стандарта метаданных. Предпосылкой для создания стандарта было всеобщее признание и возможность описания разнообразных электронных ресурсов в Интернете. DCMI была основана в 1995 году. Основным направлением текущей работы является дальнейшая разработка и расширение стандартов метаданных Dublin Core.

Деятельность DCMI заключается не только в разработке и поддержании стандартов, но и в разработке инструментов и услуг для использования, проектирования и управления метаданными. DCMI предоставляет полную документацию и поддержку стандартов DCMI, связанных ресурсов, учебных материалов и мероприятий. Технические характеристики, стандартизированные этим органом:

- *абстрактная модель DCMI*
- *набор элементов метаданных Dublin Core*

***Сообщество общей онтологии для лингвистических описаний GOLD***[2]

Подробнее деятельность GOLD описана в главе 2. Технические характеристики, стандартизированные этим органом:

- *общая онтология лингвистического описания*

[1] Dublin Core Metadata Initiative. – URL: http://dublincore.org/ (дата обращения: 01.12.2021).

[2] GOLD Community. – URL: http://linguistics-ontology.org/ (дата обращения: 01.12.2021).

***Международная федерация библиотечных ассоциаций и учреждений* IFLA, ИФЛА**[1]

ИФЛА – это независимая международная неправительственная некоммерческая организация, которая представляет интересы библиотечных и информационных служб и их пользователей. ИФЛА насчитывает более 1500 членов примерно из 150 стран мира в области библиотечных и информационных услуг. Целями являются разработка и обеспечение стандартов для библиотечных и информационных услуг. Технические характеристики, стандартизированные этим органом:

- *рекомендации для многоязычных тезаурусов*
- *международное стандартное библиографическое описание*

***Ассоциация отраслевых стандартов локализации LISA***[2]

С 1990 по 2011 год Ассоциация стандартов индустрии локализации (LISA) была ведущей некоммерческой ассоциацией для частных лиц, предприятий, ассоциаций и организаций по стандартизации, занимающихся языками и языковыми технологиями. Одна из целей LISA заключалась в разработке и распространении лучших методов перевода и локализации для информационной индустрии.

С 2011 года LISA больше не действует. После признания несостоятельности LISA разработанные стандарты были переданы под лицензию Creative Commons, а спецификации были перенесены в новые рабочие структуры в других организациях по стандартизации, таких как ISO, GALA и т.д. Однако разработанные стандарты все еще широко используются сегодня, в том числе для оценки качества перевода. Технические характеристики, стандартизированные этим органом:

- *обмен терминологическими БД*
- *обмен памятью переводов*
- *обмен правилами сегментации*

***Сообщество открытых языковых архивов OLAC***[3]

Это международное партнерство учреждений и частных лиц, целью которого является создание всемирной виртуальной библиотеки ЛИР. Информация о его деятельности есть также в главах 2, 3, 6. С момента создания OLAC разработал большое количество данных и инструментов для лингвистических исследований, предоставленных для общего пользования. OLAC развивает сеть взаимодействующих репозиториев и служб для размещения и доступа к языковым ресурсам, предоставляет стандарты и рекомендации по созданию, архивированию и использованию языковых ресурсов. Технические характеристики, стандартизированные OLAC:

- *метаданные открытых языковых архивов*

---

[1] The International Federation of Library Associations and Institutions (IFLA). – URL: http://www.ifla.org/ (дата обращения: 01.12.2021).

[2] Localization Industry Standards Association (LISA). – URL: https://en.wikipedia.org/wiki/Localization_Industry_Standards_Association (дата обращения: 01.12.2021).

[3] OLAC: Open Language Archives Community. – URL: http://www.language-archives.org/ (дата обращения: 01.12.2021).

***Организация по развитию стандартов структурированной информации OASIS***[1]

OASIS была основана в 1993 году и насчитывает более 5000 участников, представляющих более 600 организаций и отдельных членов в 100 странах. Это некоммерческий консорциум, работающий над разработкой и принятием открытых стандартов для информационных технологий. Стандарты, разработанные OASIS, среди прочего включают:

- *области безопасности*
- *облачные вычисления*
- *се́рвис-ориенти́рованную архитекту́ру*
- *веб-сервисы*
- *интеллектуальные сети*
- *электронные публикации*

***Ассоциация глобализации и локализации GALA***[2]

Это некоммерческая ассоциация, объединяющая и поддерживающая профессионалов и организации в глобальной языковой индустрии. В сфере деятельности этой индустрии находятся:

- *перевод*
- *синхронный перевод*
- *технология удаленного синхронного перевода*
- *локализация*
- *машинный перевод (МП)*
- *компьютерная поддержка перевода (CAT)*
- *память перевода*
- *системы управления переводами*

Участники GALA предоставляют такие языковые услуги, как создание субтитров, дублирование, многоязычные настольные издательские системы, тестирование качества, консультирование, озвучивание, разработка многоязычного контента и др. GALA поддерживает ряд спецификаций на ЛИР, в том числе унаследованных от LISA.

***Инициатива по языку разметки правил RuleML***[3]

Это открытая сеть людей и групп, представляющих промышленность и академические круги, работающих над разметкой веб-правил. Его основная цель – обеспечить основу для интегрированного подхода к разметке правил. Инициатива RuleML сотрудничает с ISO и W3 C, например в разработке таких стандартов, как «ISO Общая логика (CL): структура для семейства языков, основанных на логике» и W3 C Формат обмена правил (RIF). Технические характеристики, стандартизированные этим органом:

- *язык разметки правил*

---

[1] Organization for the Advancement of Structured Information Standards (OASIS); OASIS Open. – URL: http://www.oasis-open.org/ (дата обращения: 01.12.2021).

[2] The Globalization and Localization Association (GALA). – URL: https://www.gala-global.org/about/about-gala (дата обращения: 01.12.2021).

[3] The Rule Markup Initiative (RuleML). – URL: https://ruleml.org/index.html (дата обращения: 01.12.2021).

### Тематика международных стандартов и спецификаций

В данном разделе будет представлена тематическая классификация стандартов и спецификаций на ЛИР и языковые технологии. Мы предлагаем следующее обобщение классов объектов стандартизации (или тем), выделенных в информационной системе стандартизации CLARIN-D[1]:

- Знаковый уровень
  - кодировка символов
  - транскрипция
- Лингвистическое аннотирование
  - общая аннотация корпуса
  - морфосинтаксическая аннотация
  - синтаксическая аннотация
  - семантическая аннотация
  - аннотация многоязычных данных
- Лексиконы
  - контролируемый словарь
  - лексические знания
  - терминология
  - тезаурус
  - онтология
- Разметка
  - язык разметки
  - язык ограничений
  - сегментация
- Метаинформация
  - метаязык
  - метаданные
  - категоризация данных
  - схема
- Представление данных
  - форматы файлов
  - форматирование
  - представление знаний
  - сериализация
  - структура функций
  - язык запросов
  - трансформация

Далее описываются выделенные в этой классификации объекты стандартизации и перечисляются стандарты и спецификации, отнесенные к ним.

[1] Topics. – URL: https://clarin.ids-mannheim.de/standards/views/list-topics.xq (дата обращения: 01.12.2021).

Аннотированный перечень стандартов, упомянутых в данном разделе, доступен в информационной системе CLARIN[1]. Чтобы не перегружать раздел ссылками, мы ограничились включением в наименование стандарта официальных акронимов. Их расшифровка с переводом на русский язык и ссылкой на интернет-адрес представлены в приложении 4.

**Знаковый уровень**

***Кодировка символов***. Стандарты, относящиеся к этой теме:

- UCS Универсальный набор кодированных символов
- Unicode Стандарт Юникода
- IPA Международный фонетический алфавит

***Транскрипция***. Стандарты, относящиеся к этой теме:

- TSL Транскрипция разговорной речи
- CHAT Коды для человеческого анализа стенограмм

**Лингвистическое аннотирование**

***Общая аннотация корпуса***

Аннотированный корпус является важным ресурсом в лингвистических исследованиях. Он используется для многих целей, например, для устранения неоднозначности слов, создания словарей, лексикографических исследований, извлечения информации и т.д.

Аннотации корпуса понимаются как «практика добавления интерпретирующей (особенно лингвистической) информации к существующему корпусу устной и / или письменной речи». Эта информация может включать данные разных лингвистических уровней: прагматики, синтаксиса, семантики, грамматики и т.д. Подробное описание технологии аннотации содержится в главе 7.

Существуют два метода аннотации данных: встроенные и отделенные аннотации. Во встроенной аннотации основные данные и данные аннотации представляют собой единицу и сохраняются в одном файле данных. Недостатки этого метода:

- исходные данные усложняются при дальнейшей обработке;
- структуры аннотаций могут перекрываться;
- множественные аннотации с трудом упорядочены.

В последнее время широкое распространение получила аннотация stand-off. Концепция аннотации заключается в том, что одна или несколько аннотаций могут быть сохранены отдельно от первичных данных и связаны с ними. Данные аннотации могут храниться в одном файле с первичными данными или отдельно в другом файле. Преимущества:

- разделение первичных данных и их аннотаций позволяет сохранить первичные данные неизменными и использовать их для дальнейшей обработки. При индексировании на уровне знаков и слов аннотации напрямую

[1] Standards and Specifications. – URL: https://clarin.ids-mannheim.de/standards/views/list-specs.xq?sortBy=name&page=1 (дата обращения: 01.12.2021).

ссылаются на первичные данные. По этой причине необходимо, чтобы первичные данные оставались неизменными, иначе индексация может быть нарушена. Поэтому разумно установить разрешение только на чтение для первичного файла данных;

- множественная аннотация создается естественно;
- относительно легко дополнить первичные данные дополнительными аннотациями. Для этого создаются новые файлы;
- каждую аннотацию можно изменить отдельно;
- для автономных аннотаций обычно используется расширяемый язык разметки (XML). Выражения XPointer и XLink связывают данные аннотации с исходным текстом или ссылаются на него.

Другие типы аннотации корпуса:

- лемматизация
- частеречная разметка
- синтаксическая аннотация
- семантическая аннотация
- аннотация дискурса
- прагматическая аннотация
- фонетическая аннотация
- лексическая аннотация

Стандарты, относящиеся к этой теме:

- **NLM JATS** Набор тегов для архивирования и обмена журналов NLM
- **DITA** Дарвиновская архитектура ввода информации[1]
- **JATS** Набор тегов для статей журнала
- **TEI Guidelines** Рекомендации по кодированию и обмену электронным текстом
- **WordSeg** Словесная сегментация письменных текстов
- **MLIF** Многоязычная информационная структура
- **CES/XCES** Стандарт кодирования корпуса
- **LAF** Структура лингвистических аннотаций

При морфосинтаксической аннотации каждый лексический токен в текстовом корпусе становится тегом морфосинтаксических меток, таких как часть речи и другие морфологические характеристики.

Идентификация части речи является основным шагом не только для морфосинтаксического анализа, но и для многих других лингвистических аннотаций, таких как синтаксические, семантические и т.д. Таким образом, токены маркируются (например, как существительные, глаголы, прилагательные,

[1] Darwin Information Typing Architecture (DITA) – это основанная на XML архитектура для проектирования, написания, управления и публикации тематически ориентированного информационного контента в печатном виде и в Интернете. DITA была разработана OASIS и считается стандартом. – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecDita (дата обращения: 01.12.2021).

наречия), и выявляется синтаксическая функция слова. После этого проводится анализ морфологических характеристик. В ходе этого анализа каждый токен приводится к его основной морфологической форме, и выделяется его морфологическая структура (внутренняя структура слов) и грамматические формы (такие как число, род, падеж, лицо, время, вид, настроение и т.д.). Стандарты, относящиеся к этой теме:

- **MAF** Морфосинтаксическая структура аннотаций

***Синтаксическая аннотация***

Цель синтаксической аннотации – описать структуру предложений. Она показывает, как слова образуют фразы и предложения, а также их взаимосвязь друг с другом в предложении. Иногда синтаксически аннотированный корпус называют древовидным банком, потому что аннотации представлены в виде древовидных структур.

Различают две основные модели: аннотацию структур на основе составляющих и зависимостей. Обе структуры формируют иерархическую структуру предложения. Они отличаются друг от друга тем, как элементы иерархически соотносятся друг с другом в одном предложении. В структуре составляющих слова образуют синтаксическую единицу, такую как именная фраза и глагольная фраза. Фразы расположены иерархически и могут быть глубоко вложенными.

Однако структура зависимостей основана на зависимости между словами в одном предложении и описывает то, как слова соотносятся с другими словами. В структуре зависимостей одно слово – это главное, а другие – подчиненные. Обычно глагол является главным членом предложения, а остальные слова ему подчиняются.

Далее проводится различие между частичным и полным синтаксическим анализом. Полный синтаксический анализ описывает полное дерево синтаксического анализа предложения. Этот вид синтаксического анализа сложен и требует большого количества времени, поскольку имеет достаточно широкий грамматический охват. К тому же это достигается с очень большими техническими и временными затратами. Причина в том, что все возможные деревья синтаксического анализа должны быть рассчитаны для данного предложения.

Другой альтернативный анализ – частичный синтаксический анализ, который иногда называют разбиением на части. Он описывает частичную структуру предложения без построения полного дерева синтаксического анализа. В отличие от полного синтаксического анализа, частичный синтаксический анализ может идентифицировать фразы (например, существительное – фраза или глагол – фраза) или сегменты в предложении. Он не может анализировать сами фразы или определять, каково соотношение фраз в предложении. Частичный синтаксический анализ можно выполнить быстро и эффективно. Стандарты, относящиеся к этой теме:

- **SynAF** Структура синтаксических аннотаций
- **Penn Treebank** Банк деревьев структур фраз

- **KAF** Формат аннотаций KYOTO[1]

***Семантическая аннотация***

Существует множество статистических данных или методов машинного обучения для автоматического извлечения и идентификации информации, основанных на семантических аннотированных корпусах.

Семантическая аннотация обеспечивает описание различного рода знаний, содержащихся в документе, и их семантику в предметной области. Целью семантической аннотации является присвоение сущностей в тексте и ссылка на их семантические описания. Семантическая аннотация может предоставлять информацию о типе именованного объекта, информацию о времени и событии, информацию о дискурсе и семантическом отношении между объектами и т.д. Стандарты, относящиеся к этой теме:

- **SemAF** Структура семантических аннотаций
- **TimeML** Язык разметки для событий и временных выражений на естественном языке
- **DiAML** Язык разметки диалоговых актов
- **SemRoleML** Язык разметки семантических ролей

***Аннотация многоязычных данных***

Многоязычные данные – это идентичные или похожие данные на двух или более языках. Многоязычные данные с разными уровнями аннотации использовались во всех видах языковых или кросс-лингвистических исследований, а также в различных задачах обработки естественного языка. Их можно применять для огромного количества приложений, таких как машинный перевод, распознавание речи, поиск информации и т.д. Примерами многоязычных данных являются многоязычные корпуса (например Europarl), международные словари (например Agrovoc) или многоязычные наборы данных (например DBpedia, YAGO).

Факторами сложности построения многоязычных аннотированных данных являются размер данных, количество языков и типы аннотаций. Некоторые проблемы в многоязычной аннотации включают кодирование символов, распознавание букв, цифр и символов в данных, связывание между различными ресурсами, относящимися к одному и тому же объекту на разных языках, и заполнение любых лексических пробелов. Лексические несоответствия или совпадения в целевом языке также являются распространенной проблемой. Стандарты, относящиеся к этой теме:

- **MLIF** Многоязычная информационная структура

[1] KYOTO Annotation Format (KAF) – это многоуровневый формат аннотаций, основанный на XML. Аннотация является stand-off, что означает, что исходный документ остается неизменным и хранится только для чтения. KAF предоставляет слои аннотаций для базовой обработки естественного языка и открыт для расширения другими слоями аннотаций, необходимыми для конкретных приложений, которые могут быть стандартизированы позже. – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecKAF (дата обращения: 01.12.2021).

## Лексиконы

***Контролируемый словарь***

Контролируемый словарь – это «список терминов, которые были явно перечислены». Этот список контролируется органом регистрации контролируемого словаря и доступен в нем. Все термины в контролируемом словаре должны иметь однозначное, неизбыточное определение» (ANSI/NISO Z39.19–2005). Основные цели контролируемых словарей – организовать информацию, предоставить терминологию для каталога и повысить эффективность систем хранения и поиска информации. Стандарты, относящиеся к этой теме[1]:

- рекомендации для многоязычных тезаурусов
- структурированные словари для поиска информации
- представление словарных статей
- руководство по одноязычным контролируемым словарям
- коды стран
- коды названий языков
- тезаурусы и взаимодействие с другими словарями

***Лексические знания***

Цель лексических знаний – описать всю информацию о словах и их отношениях. Лексическое знание включает несколько информационных аспектов лингвистической теории, таких как морфосинтаксический, семантический, прагматический и лексический.

Морфосинтаксический аспект описывает слова, их составные части и определяет структурное сочетание слов. Семантический аспект описывает значение слова в терминах концепций некоторой базы знаний. Он определяет, как слова могут быть объединены, чтобы выразить несколько значений. Прагматический аспект описывает такую информацию, как коннотация, импликатура, стиль и т.д. Лексический аспект описывает внутренние лексические ограничения и отношения между различными языковыми единицами и соединениями, такими как омонимия, синонимия, антонимия, гипернимия, гипонимия и т.д. Стандарты, относящиеся к этой теме:

- LAF Структура лексической разметки

***Терминология***

Подробное описание этой категории лексиконов представлено в главе 12. Стандарты, относящиеся к этой теме:

- TBX Обмен терминологическими базами данных
- OLiA Онтологии лингвистической аннотации
- PISA Постоянная идентификация и устойчивый доступ
- DictionaryEntry-RePresentation Представление статей в словарях
- TMS Системы управления терминологией
- SimpL-1 Упрощенный естественный язык
- GOLD Общая онтология лингвистического описания

[1] Controlled Vocabulary. – URL: https://clarin.ids-mannheim.de/standards/views/view-topic.xq?id=TopicControlledVocabulary (дата обращения: 01.12.2021).

***Тезаурус***

Тезаурус – это словарь с алфавитной или тематической структурой. Термины в тезаурусе сгруппированы по сходству значения или области знаний и связаны между собой семантическими отношениями, такими как синоним, антоним, гипероним (обобщение), гипоним (спецификация) и т.д. Существуют тезаурусы с определениями слов и / или переводами, одноязычные и многоязычные. Тезаурусы используются во многих приложениях, включая поиск информации и обработку естественного языка, например для определения синонимов и других семантически подобных связанных терминов для индексации онлайн-материалов или для приложений анализа текста.

Стандарты, относящиеся к этой теме:

- **Multilingual Thesaurus** Рекомендации для многоязычных тезаурусов
- **SV** Структурированные словари для поиска информации
- **Controlled Vocabulary** Руководство по созданию, формату и управлению одноязычными контролируемыми словарями
- **SKOS** Простая система организации знаний
- **ISO-Thesauri** тезаурусы и взаимодействие с другими словарями

***Онтология***

Онтология – это формальное структурированное представление любого количества концептов и их соответствующих отношений друг с другом. Онтология описывает знания с помощью стандартной терминологии, значения отдельных понятий и семантических отношений между терминами. Вся информация в онтологии образует своего рода сеть для представления знаний. Эти знания можно использовать по-разному, например в информационном поиске, искусственном интеллекте, семантической сети, разработке программного обеспечения, биомедицинской информатике, библиотечном деле и т.д.

Стандарты, относящиеся к этой теме:

- **UMBEL** Верхний слой обмена отображением и связыванием
- **DOL** Распределенный язык онтологий
- **OLIA** Онтологии лингвистической аннотации
- **OWL** Язык веб-онтологий
- **SIMPLE** Простая базовая онтология
- **DAML+OIL** Язык разметки агента DARPA + язык интеграции онтологий
- **FLORA-2** Объектно-ориентированный язык базы знаний
- **BFO** Базовая формальная онтология
- **OpenCyc** Онтологическая база знаний
- **SUMO** Объединенная онтология верхнего уровня
- **DOLCE** Описательная онтология для лингвистической и когнитивной инженерии
- **SWRL** Язык правил семантической паутины, сочетающий OWL и RuleML
- **OntoIOp** Интеграция онтологий и взаимодействие
- **GOLD** Общая онтология лингвистического описания

## Разметка

### *Язык разметки*

Языки разметки – это языки, которые предоставляют дополнительную информацию о тексте, чтобы облегчить его автоматическую обработку, включая редактирование и форматирование для отображения или печати. Язык разметки определяет группы символов для форматирования макета, стиля, структурного представления текста, такого как заголовки, абзацы, списки и т.д. Эти специальные группы символов называются тегами. Примером наиболее распространенного и используемого языка разметки является HTML, который используется для отображения документов в веб-браузерах.

Другой наиболее популярный язык – это XML, который является метаязыком, он позволяет определять язык разметки документа и его структуру. Преимущество использования XML заключается в возможности разработки или изменения языка разметки с помощью набора правил и тегов, которые соответствуют индивидуальным требованиям и потребностям, например: имя, заголовок, адрес и т.д. XML используется скорее для хранения структурированных данных, чем для форматирования информации на странице. Стандарты, относящиеся к этой теме:

- **TBX** Обмен терминологическими базами
- **HyTime** Гипермедиа / Язык структурирования времени
- **TimeML** Язык разметки для событий и временных выражений
- **TMX** Обмен памяти переводов
- **XPath** Язык адресации частей XML-документа
- **R2 ML** Язык разметки правил REWERSE II
- **RuleML** Язык разметки правил
- **XQuery** Язык запросов XML
- **DAML** Язык разметки диалога
- **TEI** Рекомендации по кодированию и обмену электронным текстом
- **SemRoleML** Язык разметки семантических ролей

### *Язык ограничений*

Язык ограничений предоставляет формальную модель и синтаксис для определения конкретных ограничений для проверки экземпляров данного языка разметки. Языки ограничений (CL) можно разделить на типы: CL на основе грамматики и CL на основе правил. Стандарты, относящиеся к этой теме:

- **XML Schema** Схема XML
- **RELAX NG** Регулярный язык для XML следующего поколения
- **RDFS** Язык описания словарей: Схема RDF

### *Сегментация*

Сегментация – это процесс членения последовательности символов на значимые языковые единицы. В качестве единиц можно определить предложения, а также слово или тему. Различается, например, сегментация предложений и слов. Задача сегментации предложения – найти границы предложения в тексте. Точно так же задача сегментации слов – разбить текст на

границы слов. В зависимости от языка задача сегментации может быть более или менее сложной. Стандарты, относящиеся к этой теме:

- **SRX** Обмен правилами сегментации
- **WordSeg** Пословная сегментация письменных текстов

## Метаинформация

### *Метаязык*

Метаязык – это основа языка разметки. Он предоставляет, по крайней мере, синтаксис и формальную модель. Иногда метаязык также предоставляет один или несколько формальных признаков грамматики документа, определяющих конкретный язык разметки. Стандарты, относящиеся к этой теме:

- **XHTML** Расширяемый язык разметки гипертекста
- **XML** Расширяемый язык разметки
- **HTML** Язык гипертекстовой разметки
- **SGML** Стандартный обобщенный язык разметки
- **XMLNS** Расширяемое пространство имен языка разметки
- **RDF/XML** Спецификация синтаксиса RDF/XML

### *Метаданные*

Метаданные содержат информацию о других данных. Эта информация должна позволять и / или облегчать обнаружение, извлечение, использование и управление соответствующими ресурсами. Метаданные описывают основную информацию, такую как название, автор, местоположение или время и место создания ресурса, а также множество других данных. Метаданные также могут быть созданы для физических и цифровых версий ресурсов.

Метаданные для цифровых объектов могут быть встроены в цифровой объект или храниться отдельно. В этом случае метаданные должны быть связаны с объектами, которые хранятся и управляются в базе данных. Это позволяет гораздо быстрее добавлять, искать, изменять и читать информацию метаданных.

Наиболее распространенным форматом хранения метаданных являются XML, SGML и HTML. Некоторые из самых популярных схем метаданных: Dublin Core, OLAC, IMDI, TEI (-Header) и т.д. Они различаются разнообразием использования, дизайном, расширяемостью, сложностью, взаимодействием с другими схемами. Выбор той или иной схемы метаданных зависит от ресурсов, которые следует описать, и целей использования. Стандарты, относящиеся к этой теме:

- **Z39.87** Словарь данных – технические метаданные для цифровых неподвижных изображений
- **TextMD** Технические метаданные для текста
- **OLAC Metadata** Метаданные открытого языкового архива
- **DCMI** Абстрактная модель «Инициативы метаданных Дублинского ядра»
- **METS** Стандарт кодирования и передачи метаданных
- **ISBO** Международное стандартное библиографическое описание
- **NISO MIX** Метаданные NISO для изображений в XML-схеме

- **RDF** Структура описания ресурсов
- **IMDI** Инициатива метаданных ISLE
- **CMDI** Инфраструктура метаданных компонентов
- **DC** Набор элементов метаданных Дублинского ядра

***Категоризация данных***

Категоризация данных – это механизм для создания центрального глобального (или локального) реестра для обеих категорий данных (т.е. имен элементов и атрибутов или более общих концепций) и значений, используемых в процессе аннотации. Стандарты, относящиеся к этой теме:

- **TMF** Структура терминологической разметки
- **ITS** Набор тегов интернационализации
- **TMS** Разработка, внедрение и обслуживание систем управления терминологией
- **DCR** Реестр категорий данных

**Представление данных**

***Форматы файлов***

Стандарты, относящиеся к этой теме:

- **TBX** Обмен ТБД
- **PDF/A** Формат файла электронного документа для длительного хранения
- **PDF** Формат переносимого документа
- **CHAT** Коды для человеческого анализа стенограмм
- **RTF** Расширенный текстовый формат
- **TMS** Обмен памяти переводов
- **EAF** Формат аннотаций ELAN[1]

***Форматирование***

Форматирование описывает представление исходных данных для вывода в интерактивные программы чтения, печати, речевые программы и т.д. В отличие от преобразования, в результате которого одна структура данных и контент трансформируется в другую конкретную структуру или другой формат, форматирование фокусируется на средствах визуализации исходных данных. Стандарты, относящиеся к этой теме:

- **DSSSL** Семантика стиля документа и язык спецификации
- **XSL-FO** Объекты форматирования расширяемого языка таблиц стилей
- **IPA** Международный фонетический алфавит

[1] Формат аннотаций ELAN (ELAN Annotation Format, EAF), также известный как Формат аннотаций EUDICO (EUDICO Linguistic Annotator), разработан Лингвистическим архивом Института психолингвистики имени Макса Планка. EAF хранит сложные многослойные аннотации аудио- и видеозаписей. Аннотации в EAF выровнены по времени и привязаны к одной точке на временной шкале аудио- или видеозаписей. – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecEAF (дата обращения: 01.12.2021).

***Представление знаний***

Стандарты, относящиеся к этой теме*:*

- **SV** Структурированные словари для поиска информации
- **DOL** Язык распределенной онтологии
- **OWL** Язык веб-онтологий
- **RDF** Структура описания ресурсов
- **Controlled Vocabulary** Руководство по созданию, формату и управлению одноязычными контролируемыми словарями
- **SKOS** Простая система организации знаний
- **OntoIOp** Интеграция онтологий и взаимодействие
- **Topic Maps** Тематические карты

***Сериализация***

Стандарты, относящиеся к этой теме:

- **Turtle** Краткий язык RDF для TRIPLE[1]
- **RDF/XML** Спецификация синтаксиса RDF/XML

***Структура функций***

Стандарты, относящиеся к этой теме:

- **Feature structures** Структуры функций
  - Часть 1: Представление структуры функций
  - Часть 2: Объявление системы функций

***Язык запросов***

Язык запросов – это разновидность компьютерного языка, используемого для запроса информации из базы данных или других информационных систем. В случае корректного выражения запроса и грамотной инструкции пользователь достаточно оперативно достигает своей цели (требуемого набора данных), не обладая при этом определенными навыками программирования. Запросы строятся в виде строк с конкретными ключевыми словами и синтаксисом, различающихся в зависимости от языка запросов. То есть, дифференцируются языки запросов к базам данных и языки запросов поиска информации. Наиболее известными языками запросов к базам данных являются язык структурированных запросов (SQL), многомерное выражение (MDX) и так далее. Языки запросов поиска информации, например XQuery, используются для запроса коллекций XML-данных. Стандарты, относящиеся к этой теме:

- **TRIPLE** Язык запросов, выводов и преобразований RDF для семантического веба
- **CQLF** Корпусные запросы Lingua Franca
- **SeRQL** Язык запросов Sesame RDF
- **SPARQL** Протокол и язык запросов RDF
- **RDQL** Язык запросов данных RDF
- **RQL** Язык запросов RDF
- **XQuery** Язык запросов XML

[1] Terse RDF Triple Language (Turtle). – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTurtle (дата обращения: 01.12.2021).

***Трансформация***

Целью преобразования является преобразование исходных данных определенной структуры и содержимого в другую конкретную структуру или другой формат, в то время как исходные данные остаются неизменными. Существуют различные задачи, такие как изменение, переформатирование структуры данных, объединение данных или изменение представления данных. Стандарты, относящиеся к этой теме:

- **DSSSL** Семантика стиля документа и язык спецификации
- **XPath** XML Path Language[1]
- **XSLT** Преобразования расширяемого языка таблиц стилей[2]

## Информационные системы по стандартизации

Сведения о международных и национальных стандартах в области ЛИР и языковых технологий и адреса доступа к ним можно найти в различных источниках.

Прежде всего это информационная система ISO[3]. Поиск стандартов возможен тремя способами:

- с помощью МКС (Международный классификатор стандартов). МКС является средством классификации стандартов по отраслевому признаку, например, электротехника или целлюлозно-бумажная промышленность;
- с помощью ТК (технические комитеты). Щелчком мыши по ТК открывается обзор всех стандартов, опубликованных этой группой экспертов;
- с помощью каталога стандартов по ключевому слову или номеру стандарта (все стандарты ISO пронумерованы); например, чтобы найти стандарт ISO 9001 можно ввести в поисковую строку «менеджмент качества» или «9001».

Однако не все стандарты ISO открыты для бесплатного и свободного доступа: полные тексты утвержденных документов доступны только для экспертов ISO или за плату.

Напротив, все рекомендации и спецификации Консорциума W3 свободно доступны на портале Консорциума[4]. Здесь к услугам пользователя развернутая тематическая навигация и подробные комментарии к рекомендациям, спецификациям и проектам, находящимся в разработке. Документы можно получить в различных форматах.

---

[1] XPath Language – это язык, предлагающий возможность перемещаться по элементам и атрибутам и находить информацию в XML-документе. XPath Language разработан консорциумом World Wide Web Consortium (W3 C) и занимает важное место в стандарте W3 C XSLT. – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecXPath (дата обращения: 01.12.2021).

[2] Extensible Stylesheet Language Transformations (XSLT). – URL: https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecXSLT (дата обращения: 01.12.2021).

[3] Каталог стандартов ИСО. – URL: https://www.iso.org/ru/standards-catalogue/browse-by-ics.html (дата обращения: 01.12.2021).

[4] All Standards and Drafts // W3 C. – URL: https://www.w3.org/TR/ (дата обращения: 01.12.2021).

Вероятно, наиболее удобной для сообщества является Информационная система стандартов CLARIN[1], разработанная в немецком центре знаний CLARIN-D, которая была использована при подготовке настоящего обзора.

Информационная система стандартов CLARIN перечисляет связанные с языковыми технологиями стандарты, которые центры CLARIN готовы принять и признать, и визуализирует некоторые из их взаимозависимостей.

Система позволяет проводить поиск по ответственным организациям, тематике, а также по наименованиям и обозначениям нормативных документов. Данная система включает в себя нормативы не только основных организаций по стандартизации – ISO и W3 C, но и многих других, частично описанных в первом разделе настоящей главы.

Однако система в настоящее время не обладает достаточной полнотой, об этом создатели сообщают на сайте портала. В частности, отсутствие российских нормативных документов очень затрудняет работу российсских пользователей. Настоящая монография может отчасти восполнить этот пробел. Правда, российские нормативные документы можно найти в Электронном фонде правовых и нормативно-технических документов Консорциума Кодекс[2], а также в официальном каталоге Росстандарта[3], однако поиск в этих каталогах не очень удобный.

### Российская стандартизация в области ЛИР и языковых технологий

В России стандартизация сосредоточена в Федеральном агентстве по техническому регулированию и метрологии (Росстандарт). Основной Технический комитет (ТК) по интересующей нас тематике – это ТК 55 «*Терминология, элементы данных и документация в бизнес-процессах и электронной торговле*» [2]. В этом ТК переведены на русский язык и утверждены в качестве национальных стандартов несколько стандартов по управлению ЛИР из числа разработанных в ISO.

Кроме того, ряд полезных для языковых технологий стандартов разработан в других технических комитетах, прежде всего в ТК 191 (ISO/ТК 46) «*Научно-техническая информация, библиотечное и издательское дело»* и ТК 22 (ISO JTC 1) «*Информационные технологии»*.

Перечень российских национальных стандартов в области ЛИР приводится в приложении 6. В него включены национальные нормативные документы в области ЛИР, разработанные упомянутыми техническими комитетами Росстандарта.

---

[1] CLARIN Standards Information System. – URL: https://clarin.ids-mannheim.de/standards/index.xq (дата обращения: 01.12.2021).

[2] Консорциум Кодекс. Электронный фонд. Более 25 000 000 актуальных правовых и нормативно-технических документов. – URL: https://docs.cntd.ru/ (дата обращения: 01.12.2021).

[3] Каталог стандартов // Росстандарт. – URL: http://old.gost.ru/wps/portal/pages.CatalogOfStandarts (дата обращения: 01.12.2021).

Из представленного обзора следует, что международные организации по стандартизации проводят большую и весьма содержательную работу по анализу и обобщению различных ЛИР, обеспечивая тем самым возможность их интеграции и повторного использования. Результаты этой работы весьма впечатляют. Однако с реальным внедрением разработанных стандартов дело обстоит менее убедительно. Из многочисленных международных сообществ, связанных с разработкой ЛИР, пожалуй только CLARIN активно использует разработанные стандарты ISO. Большинство известных в мире ЛИР пока, очевидно, не стандартизовано.

Особенно наглядно этот недостаток деятельности по стандартизации ЛИР проявляется в России. Выбор стандартов для перевода производит впечатление случайного, а качество переводов при этом чрезвычайно низкое: кажется, что результаты автоматического перевода вообще не редактируются.

Главный недостаток деятельности российского Технического комитета 55 заключается в том, что разработанные им стандарты вообще не применяются при разработке российских ЛИР. Это неудивительно, ведь в составе этого ТК практически нет разработчиков ЛИР. Исключение только одно – возглавляет ТК 55 ФГУП Стандартинформ, который поддерживает известный российский банк терминологических данных Ростерм. Однако этот банк данных не замечен в активном сотрудничестве с другими разработчиками отечественных ЛИР, а также в разработке открытого доступа и интеграции терминологических данных.

Следует отметить также слабую координацию между техническими комитетами. Иначе трудно объяснить появление дублирующих стандартов. Например, при наличии сериального стандарта на метаданные для образовательных ресурсов на основе ISO/IEC 19788 (ГОСТ 33247–2015, ГОСТ ISO/IEC 19788–2-2015, ГОСТ ISO/IEC 19788–3-2015, ГОСТ ISO/IEC 19788–5-2015) имеется еще ГОСТ Р 55750–2013. *Метаданные электронных образовательных ресурсов*.

Очевидно, что междисциплинарное сотрудничество в этой системе не налажено. Кроме того, руководству Росстандарта можно порекомендовать обращать внимание не только на разработку, но и на внедрение национальных стандартов в данной сфере.

## Литература к главе 5

1.Gracia J., Vila-Suero D. Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries : Final Community Group Report 29 September 2015 / ed.: J. Gracia // The World Wide Web Consortium (W3 C). – URL: https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/ (дата обращения: 01.12.2021).

2.Приказ Росстандарта от 3 декабря 2010 г. № 4850 «О Техническом комитете по стандартизации ТК 55 “Терминология, элементы данных и документация в бизнес-процессах и электронной торговле”». – URL: https://docs.cntd.ru/document/902260132 (дата обращения: 01.12.2021).

# ГЛАВА 6. МЕТАДАННЫЕ ЛИР

## Краткая история

Постановка задачи разработки специальной системы метаданных для ЛИР принадлежит, по-видимому, рабочей группе EAGLES/ISLE, которая в 2001 году предложила план разработки соответствующего стандарта [1]. Этот проект получил название Инициативы IMDI.

Параллельно с этими предложениями была разработана система метаданных OLAC, которая опиралась на известную систему метаданных Дублинского ядра.

Позже появилась новая система компонентов метаданных CLARIN (CMDI), а затем, пожалуй, наиболее детально разработанная на сегодня система метаданных META-SHARE, которая была разработана в рамках ELRA. Однако в ELRA была предложена также система метаданных для ЛИР, известная как Международный стандартный номер ЛИР, которая опирается на правила OLAC, и, независимо, – карта оценки ЛИР (LRE map), использующая другую модель описания ЛИР.

Наконец, в 2015 и в 2019 гг. появились две части стандарта ISO на метаданные для ЛИР, основанные на CMDI.

Все перечисленные модели метаданных будут описаны в настоящей главе. Подробно изложены предложения Инициативы IMDI, поскольку в них сформулированы основные задачи и проблемы создания системы метаданных ЛИР, которые решались во всех последующих проектах.

Создание и использование лингвистических метаданных является центральной частью общей проблемы управления лингвистическими данными. В связи с этим в данной главе будут рассмотрены так называемые Остинские принципы цитирования лингвистических данных, разработанные Калифорнийским университетом, но широко обсуждавшиеся лингвистическим сообществом.

Модели стандартов метаданных в той или иной степени опирались на проекты создания словарей (регистров, онтологий) лингвистических категорий, которые также активно разрабатывались в последние годы международным сообществом лингвистов. Поскольку метаданные и словари лингвистических категорий теснейшим образом связаны, описания этих проектов также представлены в настоящей главе. Эти проекты собраны в отдельном разделе данной главы.

### Проект метаданных IMDI [1]

Если ЛИР – это наборы данных, представляющие примеры использования языка, либо непосредственно, как в корпусах, либо в виде производных данных, как в лексиконах и онтологиях, то фундаментальные и прикладные лингвистические исследования имеют долгую историю создания и использования текстовых ЛИР. В последнее время очень актуальны мультимедийные ЛИР. Они используются в лингвистике и смежных областях, таких как язык жестов, антропология, компьютерная лингвистика, искусственный интеллект, фонетика, психология, распознавание речи, мультимодальные исследования и человеко-машинный интерфейси-дизайн. Цели и области применения мультимедийных ЛИР самые разные: лингвисты используют их для создания и проверки новых лингвистических гипотез; инженеры по распознаванию речи используют их для тестирования устройств распознавания речи и установки параметров распознавания. Каждый проект определял свою собственную структуру заголовка и содержание, соответствующие целям проекта.

Развитие Всемирной паутины с ее связанными веб-страницами открыло новые возможности. Проект IMDI поставил задачу создать пространство связанных метаописаний с информацией о существующих ЛИР. Это пространство, снабженное соответствующими инструментами для просмотра и поиска, должно быть доступно в Интернете.

В связи с тем, что ЛИР значительно различаются и предназначены для разных пользователей со своими специфическими запросами, возникает вопрос: могут ли требования к обработке такого разнообразия приложений быть прописаны в одном стандарте?

Рабочая группа IMDI разработала подробные предложения, в которых были учтены требования сообщества разработчиков и пользователей ЛИР, существующий опыт разработки систем метаданных, в том числе Dublin Core (DC), RDF и других. Была определена сфера применения ЛИР, среди которых разработчики выделили типы ЛИР: текстовые корпуса, аннотированные корпуса, мультимедийные корпуса, лексиконы, типологические базы данных, грамматические данные, онтологии и другие.

На этой основе были определены структура метаописания, объем метаданных, элементы словаря метаданных, отображение элементов метаданных, в том числе повторное использование определений элементов метаданных из других сообществ.

Были сформулированы требования к инструментам. Необходимы редакторы метаописаний, браузеры, которые понимают структуру связанных файлов метаописания и предоставляют графические изображения, поддержка пользователя во время навигации, инструменты поиска, которые могут справиться со структурой файла метаописания и любыми элементами метаданных, полученных из других стандартов. Инструменты поиска должны эффективно использовать связи между метаописаниями и знаниями о доступных метаописаниях. В некоторых случаях могут потребоваться такие методы оптимизации, как кэширование.

Практически осуществимый сценарий внедрения стандарта должен включать такие темы:

- места хранения метаописаний
- способы регистрации и привязки метаописаний
- способы построения просматриваемых иерархий
- способы контроля за связыванием новых описаний с существующим пространством
- требования к центрам, которые могли бы создать и поддерживать пространство метаданных ЛИР

В результате деятельности рабочей группы EAGLES/ISLE появились проекты систем метаданных для лексиконов [2] и предложения по классификации и структуре словарей [3]. Наибольшее распространение метаданные IMDI получили применительно к мультимодальным ЛИР[4]. Также была разработана схема перехода от модели метаданных IMDI к стандарту метаданных OLAC [5]. Руководство пользователя для модели метаданных IMDI представлено по адресу [6]. Полный перечень документов, разработанных в рамках инициативы IMDI, доступен по адресу[1].

## Метаданные OLAC

Метаданные Сообщества открытых лингвистических ресурсов (OLAC) определены в нормативном документе *Метаданные OLAC* [7]. Этот документ определяет формат метаданных, используемый OLAC для описания ЛИР и предоставления связанных с ними услуг. OLAC использует формат XML для обмена метаданными ЛИР в рамках Инициативы открытых архивов (OAI).

*Набор метаданных*

Набор метаданных OLAC основан на наборе метаданных Dublin Core (DC) и использует все пятнадцать элементов, определенных в этом стандарте. Для обеспечения большей точности в описании ресурсов OLAC существуют рекомендации DC для квалификации элементов с помощью уточнений элементов или схем кодирования.

Цитируемый документ определяет только формальные (синтаксические) требования к описанию метаданных OLAC. Он не дает полного набора рекомендаций о том, что означают элементы метаданных, уточнения и схемы или о том, как их следует использовать. Такие советы содержатся в *информационной записке OLAC* [8].

Квалификаторы, рекомендованные DC, применимы к широкому спектру ресурсов. Однако сообщество разработчиков ЛИР имеет ряд требований к описанию ресурсов, которые не удовлетворяются этими общими стандартами. Чтобы удовлетворить эти потребности, члены OLAC разработали специальные квалификаторы для сообщества, и сообщество в целом приняло некоторые из них (следуя [OLAC-Process]) в качестве рекомендуемой передовой практики для описания ЛИР. Эти рекомендуемые квалификаторы перечисле-

[1] IMDI Documents. – URL: https://www.mpi.nl/ISLE/documents/docs_frame.html (дата обращения: 01.12.2021).

ны в разделе *Расширения OLAC*, а способ их использования в описании ресурсов описан ниже в разделе *Использование расширений OLAC*.

*Формат метаданных*

XML-реализация метаданных OLAC соответствует *Руководящим принципам реализации Дублинского ядра в XML* [9]. Схема метаданных OLAC является приложением профиля [HP2000], которое включает в себя элементы из двух схем метаданных (простой и квалифицированной). Схема метаданных OLAC и схемы для всех расширений OLAC используют две основные схемы Дублинского ядра: простая DC и квалифицированная DC.

Контейнером для записи метаданных OLAC является элемент <olac>. Если метаданные OLAC хранятся в статическом репозитории[1], имена из пространства имен могут быть удалены из отдельных записей OLAC и помещены в корневой элемент.

В дополнение к этим основным элементам метаданных DC запись может использовать квалификаторы DC в соответствии с рекомендациями, приведенными в DCXML. Квалифицированный элемент может указывать уточнение (используя элемент, определенный в пространстве имен dcterms), или схему кодирования (используя схему, определенную в dcterms как значение атрибута xsi:type), или и то, и другое.

Нормативный документ OLAC содержит определение пространства имен, словарь рекомендуемых языковых идентификаторов, порядок использования расширений, включая внешние расширения, устанавливает порядок документирования расширений.

Кроме изложенного документа в OLAC разработаны *Рекомендации по использованию метаданных OLAC* [10][2]. Приведем (с сокращениями) фрагмент этого документа, определяющий один из наиболее сложных вопросов метаописаний ЛИР, а именно рекомендации по детализации ЛИР.

*Детализация ЛИР*

При определении правильного уровня для единиц, описываемых как ЛИР в контексте OLAC, учитывается множество факторов. Уровень единицы измерения, подходящий для включения в агрегированный каталог, такой как OLAC, может отличаться (как правило, быть выше) от уровня, желательного для каталога холдингов конкретного учреждения, который в свою очередь обычно выше уровня, желательного для описания подробного содержания ресурса.

Хранилище метаданных должно рассматривать ресурсы с одним происхождением как составляющие единую единицу, и поэтому должны быть описаны в рамках одной записи.

Для ресурса, опубликованного в той или иной форме, подходящей единицей описания для записи OLAC является единица самой публикации. Коллективная работа может требовать отдельных записей для отдельных доку-

[1] OLAC Repositories. – URL: http://www.language-archives.org/OLAC/repositories.html (дата обращения: 01.12.2021).

[2] OLAC Metadata Usage Guidelines. – URL: http://www.language-archives.org/NOTE/usage.html (дата обращения: 01.12.2021).

ментов, содержащихся в ней, которые должны быть связаны с записью для работы в целом через отношения isPartOf и hasPart.

В общем случае запись OLAC соответствует цитируемому источнику. Таким образом, для опубликованных работ детализация не представляет особой проблемы.

Детализация представляет собой самую большую проблему для первичных исходных материалов (например, записей, транскрипций, аннотаций, заметок, наборов данных). Типичной практикой архивистов является сбор таких материалов в коллекции, которые в свою очередь становятся первичными единицами архивного описания (т.е., результатом являются ресурсы, которые по правилам DC могут быть отнесены к типу коллекции).

Однако эти компоненты не являются единицами описания на уровне OLAC. Это единица коллекции, которая образует базовую единицу для описания OLAC. Главным фактором при определении принадлежности материалов к единой коллекции является их происхождение.

Обычно собранные ресурсы имеют общее происхождение. Это может быть один исследователь или исследовательская группа. Это также может быть проект, который объединяет материалы из разрозненных источников для единой новой исследовательской цели, создавая таким образом новую коллекцию, основанную на вторичном использовании материалов. Общая история также имеет значение; тот факт, что набор ресурсов был перемещен или передан в другие руки или обработан в целом с момента его первоначального сбора, помогает установить его идентичность как единой единицы для архивного описания.

ЛИР общего происхождения, которые составляют коллекцию, отличаются высокой степенью общности элементов метаданных, например один и тот же исследователь, автор, предметный язык, приблизительные даты, охват, лингвистический тип.

Другие элементы метаданных, которые также могут быть важны для обнаружения ресурсов, но которые способны отличаться для элементов коллекции (например, формат или тип дискурса), могут быть повторены на уровне описания OLAC.

В качестве альтернативы коллекция может быть разделена на субколлекции по жанру дискурса, говорящему или другому значимому признаку. Затем субколлекции могут быть обработаны в различных записях OLAC (и связаны с целым через отношения isPartOf и hasPart отношения).

Коллекция, как правило, описывается более подробно в опции «помощь в поиске», которая дает подробную информацию об организации и выделяет особенности, представляющие интерес. Дальнейшие характеристики отдельных элементов в коллекции (например, тема, дополнительные участники, специфика события, формат) документируются на более тонком уровне детализации при описании коллекции.

### Метамодель META-SHARE

Платформа META-SHARE является сервисом ассоциации ELRA, предназначенным для обмена ЛИР. На этой платформе реализованы разнообраз-

ные возможности для описания ЛИР. В связи с этим приведем основные положения документа [11], регламентирующего систему метаданных META-SHARE и содержащего руководство по ее применению.

*Введение*

В настоящем документе представлена обновленная версия схемы метаданных 2.0. Документ рассматривается как живой документ, встроенный в платформу META-SHARE и отражающий прогресс и изменения в связи с событиями в этой области.

Схема метаданных META-SHARE предназначена для описания ЛИР, в частности для удовлетворения потребностей сообщества разработчиков лингвистических технологий. Схема охватывает следующие типы ресурсов / носителей:

- корпуса (текстовые, аудио-, видео-, мультимодальные / мультимедийные корпуса, включая сенсомоторные ресурсы, n-граммные ресурсы);
- лексические / концептуальные ресурсы (например, компьютерные словари, лексиконы, онтологии, терминологические ресурсы, тезаурусы, мультимодальная / мультимедийная лексика и т.д.)
- языковые описания (например, компьютерные грамматики);
- технологии (инструменты / сервисы), которые могут быть использованы для обработки информационных ресурсов.

Цитируемый документ призван выступать в качестве руководства пользователя, содержащего пояснения для поставщиков ЛИР, которые хотят описать свои ресурсы в соответствии с ним.

Документ описывает модель метаданных для описания ЛИР, открытый распределенный механизм для совместного использования и обмена ресурсами META-NET и, более конкретно, ее обновленную версию 2.0; она была реализована в виде XML-схемы.

Более подробное изложение теоретических принципов и общее введение в модель можно найти в работах [12; 13].

*Основы модели*

В контексте META-SHARE термин «метаданные» относится к описаниям ЛИР, охватывающим как данные (текстовые, мультимодальные / мультимедийные и лексические данные, грамматики, языковые модели и т.д.), так и технологии (инструменты / услуги), используемые для их обработки.

Механизм, который был принят, – это компонентный механизм (Component MetaData Infrastructure, CMDI), в соответствии с которым семантически когерентные элементы группируются вместе. Элементы используются для кодирования конкретных описательных признаков ЛИР. Чтобы обеспечить семантическую согласованность с другими связанными схемами и моделями, метамодель включает ссылки на концептуально одинаковые или аналогичные существующие элементы Дублинского ядра и Реестра категорий данных (ISO DCR)[1]; при необходимости новые элементы будут вводиться в ISO DCR.

[1] Реестр категорий данных ISO DCR будет описан ниже в данной главе.

Было введено понятие отношений для кодирования связывающих признаков между ресурсами. Отношения существуют и между различными формами ЛИР (например, первичные данные и аннотированные ЛИР), различными ЛИР (например, язык, ресурс и инструмент, который был использован для его создания, и т.д.) независимо от того, включены они в репозиторий META-SHARE или нет, а также между ЛИР и дополнительными документами. Отношения представлены как элементы в текущей версии схемы.

Совокупность всех компонентов и элементов, описывающих конкретные типы и подтипы ЛР, представляют профиль этого типа. Очевидно, что некоторые компоненты включают информацию, общую для всех типов ресурсов (например, идентификация, контакты, лицензионная информация и т.д.), и таким образом используются для всех ЛИР, в то время как другие (например, компоненты, включающие информацию о содержании, аннотации и т.д.) различаются по типам.

Элементы относятся к двум основным уровням описания:

- начальный уровень, обеспечивающий базовые элементы для описания ресурса (минимальная схема), и
- второй уровень – с более высокой степенью детализации (максимальная схема), обеспечивающий подробную информацию о ресурсе и охватывающий все этапы производства и использования ЛР.

Минимальная схема содержит те элементы, которые считаются необходимыми для описания ЛИР (с точки зрения поставщика) и идентификации ЛИР (с точки зрения потребителя). Эти два уровня содержат четыре класса элементов:

- первый уровень содержит Обязательные (M) и Зависимые от условий обязательные (MC) элементы (т.е. заполняемые при выполнении определенных условий), а
- второй уровень включает Рекомендуемые (R) и Необязательные (O) элементы.

*Онтологии META-SHARE*

META-SHARE стремится предоставить пользователям не только каталог ЛИР (данных и инструментов), но и информацию, которая может быть предназначена для оптимального их использования. Например, исследовательские работы, документирующие производство ресурса, а также используемые стандарты и методики.

В онтологии META-SHARE проводится различие между ЛИР как таковыми и другими соответствующими материалами, такими как справочная документация, связанная с ресурсом (отчеты, инструкции и т.д.), лица / организации, участвующие в их создании и использовании (создатели, дистрибьюторы и т.д.), проекты, мероприятия и лицензии (для доступа к ЛИР)[1].

[1] Согласно классификации, введенной в главе 1, это соответствует различению специальных и тематических ЛИР.

Основной интерес для META-SHARE представляют специальные ЛИР, остальные материалы – акторы, проекты, документы и т.д. – описываются тогда, когда они связаны с конкретным ресурсом. Например, META-SHARE включает в библиографический список только те документы, которые связаны с конкретными ресурсами.

*Таксономия ЛИР*

Основным элементом, используемым для классификации ЛИР по типам, является *ResourceType* со следующими значениями:

- корпус (включая письменные / текстовые, устные / речевые, мультимодальные / мультимедийные корпуса);
- лексический / концептуальный ресурс (включая терминологические ресурсы, списки слов, семантические словари, онтологии и т.д.);
- языковое описание (включая грамматики);
- инструмент / сервис (включая базовые средства обработки, приложения, веб-сервисы и т.д.).

Важное место в описании ЛИР в контексте META-SHARE занимает элемент *MediaType,* который определяет форму / физический носитель ресурса. Понятие *медиа* предпочтительнее, чем письменное / устное / мультимодальное различие, поскольку оно имеет более четкую семантику и позволяет рассматривать ЛИР как набор модулей, каждый из которых может быть описан через отличительный набор признаков. Таким образом, предусмотрены следующие значения медиа:

- текст
- аудио
- изображение
- видео
- textNumerical
- textNgram

Ресурс может состоять из частей, принадлежащих к различным типам медиа: например, мультимодальный корпус включает в себя видеочасть (движущееся изображение), аудиочасть (диалоги) и текстовую часть (субтитры и / или транскрипцию диалогов); мультимедийный лексикон включает в себя текстовую часть, но может также включать в себя видео и / или аудиочасть; ресурс языка жестов также является ресурсом с различными типами носителей (видео, изображение, текст).

Точно так же одни и те же программные инструменты могут быть применены к ресурсам различных типов медиа: например, инструмент может использоваться как для видео, так и для аудиофайлов. Таким образом, для каждой части ресурса создается соответствующий набор функций (компонентов и элементов), например для устного корпуса и его транскрипций набор звуковых функций будет использоваться для звуковой части, а набор текстовых функций – для транскрибируемой части.

*Основное содержание и структура модели*

Ядром модели является компонент *ResourceInfo*, который содержит всю информацию, имеющую отношение к описанию ЛИР. Он включает в себя «административные» компоненты, общие для всех ЛИР и «содержательные» компоненты, характерные для конкретного типа ЛИР. Такая агрегация компонентов обеспечивает наиболее полное описание ЛИР.

Все компоненты типов ЛИР расположены под *resourceComponentTypecomponent*. Аналогично для каждого типа ЛИР создаются определенные зависимые от среды компоненты, чтобы сгруппировать вместе наборы функций, относящихся к каждому типу ЛИР. Элементы *ResourceType* и *MediaType* кодируют две оси классификации схемы, в то время как каждое из значений этих двух элементов связано с соответствующим компонентом. Набор компонентов *ResourceType* и *MediaType* включает в себя:

- *corpusInfo, lexicalConceptualResourceInfo, languageDescriptionInfo, toolServiceInfo* – включают информацию, специфическую для каждого типа ЛИР, и принимают значения *corpus, lexical / conceptualResource, languageDescription and toolServicefor* соответственно;
- *corpusTextInfo, corpusAudioInfo, corpusVideoInfo, lexicalConceptualResourceTextInfo, lexicalConceptualResourceVideoInfo* – предоставляют информацию в зависимости от типа носителя каждого типа ЛИР и включают элемент *MediaType* со значениями *text, audio, video* и т.д. соответственно.

Набор из шести компонентов обладает «особым» статусом в том смысле, что они могут быть присоединены к различным компонентам, выполняющим различные роли, а именно *PersonInfo, organizationInfo, communicationInfo, projectInfo, sizeInfo и DocumentInfo*. Например, sizeInfo может использоваться либо для определения размера всего ресурса, либо в сочетании с другим компонентом для описания размера частей ресурса (например, для домена, языка и т.д.); *PersonInfo* используется для контактных лиц, создателей ресурсов, лицензиатов, аннотаторов корпуса и т.д.

Существенным является статус элемента; в метамодели принято четыре статуса:

- обязательный
- условно-зависимый (обязательный в определенных условиях)
- рекомендованный
- необязательный

Наконец, для других сущностей модели, помимо ЛИР, были разработаны специальные элементы, позволяющие осуществлять их массовое кодирование независимо от ресурса, с которым они связаны. Таким образом, например, используя такой элемент, поставщик ресурсов может загрузить соответствующие метаданные для всех персон за один раз, а затем, редактируя метаданные для ресурсов, создать соответствующие ссылки на сохраненных персон.

*Структура представления и условные обозначения*

В следующих разделах даются «специальные» компоненты модели, за которыми следуют компоненты, общие для всех типов ЛИР, а затем компоненты типа ресурсов в следующем порядке: корпусы, инструменты / сервисы, языковые описания и, наконец, лексические / концептуальные ресурсы.

Для каждого компонента предоставляется следующая информация:

- *определение*: краткое утверждение, объясняющее семантику компонента в META-SHARE;
- *тип*: обычно он принимает значение «component»; значение «special status component» используется для специальных элементов;
- *элементы*: набор элементов / компонентов, включенных в компонент, с гиперссылкой;
- *компонент*: используется вместо «элементов» для компонентов специального статуса.

Для элементов сопроводительная информация включает в себя:

- определение: краткое утверждение, объясняющее его семантику в контексте META-SHARE;
- тип: со значениями:
  - o xs:строка: свободный текст
  - o xs:integer/int/double: числовые поля
  - o xs:boolean: да / нет
  - o s:date: дата в соответствии с ISO-8601
  - o ms:myString: свободный текст на любом языке (атрибут «lang» должен использоваться для указания языка текста)
  - o ms:emailAddress: email
  - o ms:httpURI: pattern of url's
  - o ms:myStringURI: URL
  - o closed controlled vocabulary: значение должно быть выбрано из списка значений, содержащихся в контролируемом словаре
  - o open controlled vocabulary: значение может быть выбрано из списка значений, содержащихся в контролируемом словаре, но может дополняться;
- пространство значений: там, где это возможно, используется ссылка на контролируемую лексику или на стандартизированные, управляемые словари;
- значения: если используется контролируемый словарь, специфичный для META-SHARE, то набор значений перечисляется вместе с определениями, где это необходимо;
- примеры: небольшой список возможных значений для целей иллюстрации, особенно в случае текстовых элементов;
- DCLINK: имя соответствующего элемента схемы Дублинского ядра, предоставленное для целей сопоставления;
- ISOcatLINK: имя соответствующего элемента DCR ISOcat;
- комментарии: используется для заметок там, где это необходимо.

По мнению многих специалистов, к которым присоединяется автор, META-SHARE является на сегодня наиболее продуманной и подробно разработанной системой метаданных для ЛИР. Но использование этой системы весьма затратно: достаточно сказать, что Руководство META-SHARE включает свыше 200 страниц.

### Международный стандартный номер ЛИР (ISLRN)[1]

В рамках ELRA разработано еще несколько моделей метаописаний ЛИР. Одна используется в сервисе Международного стандартного номера языкового ресурса (ISLRN). Это универсальная схема идентификации ЛИР, которая обеспечивает уникальный идентификатор с использованием стандартизированной номенклатуры, и, следовательно, использование ЛИР с надлежащими ссылками в разных приложениях – в научно-исследовательских проектах, оценке продуктов, а также в документах и научных работах.

Каждый объект в мире требует своего рода идентификации, чтобы быть правильно распознанным. Традиционные печатные материалы, такие как книги, например, обычно используют Международный стандартный номер книги (ISBN), контрольный номер Библиотеки Конгресса (LCCN). Многие повседневные продукты применяют Международный / Европейский артикул (EAN), который является универсальной системой штрих-кодирования.

Для цифровых ресурсов применяется Цифровой идентификатор объекта (DOI) и другие идентификаторы в качестве уникальной схемы идентификации.

Идентификация – это важный параметр в сетевом пространстве, поскольку в нем стали активно применяться технологии и методы естественного языка. Таким образом, уникальные ресурсы должны быть идентифицированы, а метакаталогам необходим общий формат идентификации, обеспечивающий корректное управление данными. Поэтому ЛИР должны иметь тождественные схемы идентификации независимо от их способов представления, отнесений к типам и физического местоположения (на жестких дисках, в Интернете или Интранете).

ISLRN не должен заменять местные и конкретные идентификаторы, он является не обязанностью, а скорее существенной и передовой практикой. Например, ресурс, распределенный между несколькими центрами обработки данных, по-прежнему будет иметь «локальный» идентификатор центра обработки данных, но будет еще иметь уникальный ISLRN.

Основная цель схемы метаданных, используемой в ISLRN, – это идентификация ЛИР. На основе широко известной схемы OLAC был выбран минимальный набор метаданных, гарантирующий правильное распознавание и идентификацию любого ЛИР. Очевидна простота полей, которые легко и быстро заполняются, что не вызывает затруднений.

[1] International Standard Language Resource Number. – URL: http://www.islrn.org/ (дата обращения: 01.12.2021).

***Основные метаданные ISLRN***

- заглавие
- полное официальное имя
- имя, по которому ресурс упоминается в библиографии
- тип ресурса
- характер или жанр содержания ЛИР с лингвистической точки зрения
- источник / URL
- формат / тип MIME
- формат файла (тип MIME) ЛИР
- размер / продолжительность
- среда доступа
- материальный или физический носитель ЛИР
- описание
- версия
- тип СМИ
- список типов, используемых для категоризации характера или жанра содержимого ЛИР
- язык(и)
- создатель ЛИР
- распределитель
- лицо или организация, ответственные за предоставление ЛИР
- правообладатель ЛИР
- связь

## Карта LRE[1]

Еще одна модель метаданных, которая разработана в рамках ELRA, – это оценочная карта ЛИР (LRE), разработанная для мониторинга создания ЛИР в разнообразных проектах. Эта карта была впервые распространена на конференции LREC в 2010 году, имела большой успех и позже распространялась на многих других конференциях. В настоящее время при помощи LRE описано свыше 6 тыс. ЛИР. Приведем содержание этой карты. В скобках указано количество ЛИР в массиве LRE, о которых имеются соответствующие данные.

- Оценка ЛИР
  - Оценочные данные (230)
  - Инструменты оценки (71)
  - Оценочный пакет (25)
  - Методология оценки / Стандарты / Руководящие принципы (15)
- Ресурс-данные
  - Корпус (2920)
  - Лексикон (666)
  - Онтология (162)
  - Грамматика / Языковая модель (82)

[1] LRE map. – URL: http://www.elra.info/en/catalogues/lre-map/ (дата обращения: 01.12.2021).

- Терминология (66)
- Банки деревьев зависимостей (42)

- Ресурс-руководство
  - Представление. Аннотации. Формализм / Руководство (62)
  - Языковые ресурсы / Технологии. Инфраструктура (20)
  - Метаданные (10)
- Ресурс-инструмент
  - Таггер / Парсер (400)
  - Инструмент аннотации (245)
  - Корпус Инструмент (83)
  - Распознаватель именованных объектов (60)
  - Инструмент машинного перевода (51)
  - Программный инструментарий (41)
  - Токенизатор (35)
  - Инструмент машинного обучения (32)
  - Инструмент моделирования языков (29)
  - Определитель смысла слов (17)
  - Распознаватель речи / Транскриптор (14)
  - Обработка сигналов / Извлечение признаков (14)
  - Веб-сервис (9)
  - Преобразование текста в речь (9)
  - Идентификатор языка (6)
  - Определитель говорящего (4)
  - Инструмент сентимент-анализа (4)
  - Просодический анализатор (3)
  - Анализатор изображений (3)
  - Инструмент устного диалога (1)
- Состояние производства ЛИР
  - Существующие и используемые (2587)
  - Недавно созданные, в процессе (1408)
  - Недавно созданные, законченные (1290)
  - Существующие обновленные (487)
  - Другое (354)
  - Не применимо (17)
- Доступность
  - Свободно доступно (2772)
  - Другое (1232)
  - От владельца (1229)
  - Из дата-центра (580)
  - Нет в наличии (267)
  - Не применимо (63)
- Модальность (семиотический тип)
  - Письменные (4355)
  - Речь (430)
  - Мультимодальные / Мультимедиа (286)
  - Неприменимо (261)

- Устная и письменная речь (186)
- Язык жестов (64)
- Независимо от модальности и другое (561)

- Использование ресурсов
  - Извлечение информации, поиск информации (608)
  - Машинный перевод, синтез речи (532)
  - Разбор и тегирование (289)
  - Языковое моделирование (282)
  - Классификация документов, текстовая категоризация (201)
  - Распознавание / Понимание речи (185)
  - Приобретение (181)
  - Дискурс (178)
  - Открытие / Представление знаний (172)
  - Смысл словосочетания (160)
  - Признание / Поколение эмоций (154)
  - Оценка / Валидация (152)
  - Создание / Аннотация Корпуса (148)
  - Текстовое копирование (147)
  - Распознавание именнованных сущностей (144)
  - Диалог (122)
  - Подведение итогов (96)
  - Ответы на вопросы (83)
  - Морфологический анализ (80)
  - Семантическая паутина (68)
  - Веб-сервисы (67)
  - Создание / Аннотация аннотации (63)
  - Синтез речи (55)
  - Поколение естественных языков (54)
  - Машинное обучение (54)
  - Текстовое вложение и перефразирование (53)
  - Анализ мнений / Анализ настроений (40)
  - Распознавание / Поколение жестового языка (35)
  - Идентификация языка (35)
  - Идентификация личности (30)
  - Анафора, кореференция (30)
  - Семантическая ролевая маркировка (29)
  - Обнаружение и отслеживание тем (25)
  - Обработка мультимедийных документов (22)
  - Голосовое управление (3)
  - Другое (1566)
- Тип по языку
  - Одноязычный (2507)
  - Независимо от языка (2206)
  - Многоязычный (961)
  - Двуязычный (380)
  - Трехъязычный (84)

    - Не применимо (5)
- Язык (Top-4)
    - Английский (961)
    - Немецкий (216)
    - Французский (180)
    - Испанский (130)

**Инфраструктура компонентов метаданных (CMDI) CLARIN**

***Общие сведения***

Метаданные для языковых ресурсов и инструментов существуют во множестве форматов. Часто эти описания содержат специализированную информацию для конкретного исследовательского сообщества (например, заголовки TEI для текста, IMDI для мультимедийных коллекций).

Чтобы преодолеть эту дисперсию, CLARIN инициировала разработку Инфраструктуры компонентов метаданных (CMDI). Эта инфраструктра представлена на специальном разделе портала CLARIN[1]. Она обеспечивает основу для описания и повторного использования схем метаданных. Строительные блоки описания («компоненты», включающие определения полей) могут быть сгруппированы в готовый формат описания («профиль»). Оба они хранятся и совместно используются другими пользователями в реестре компонентов для повторного использования. Каждая запись метаданных затем оформляется в виде XML-файла, включая ссылку на профиль, на котором она основана.

Подход CMDI сочетает архитектурную свободу при моделировании метаданных с мощными возможностями исследования и поиска в широком диапазоне ЛИР.

На сегодняшний день существуют две поддерживаемые версии платформы компонентов метаданных CLARIN: CMDI 1.1 и CMDI 1.2. Они не взаимозаменяемы, но CMDI 1.1 метаданных может быть легко преобразован в CMDI 1.2.

Вместо единого формата метаданных CMDI предоставляет основу для создания и использования самостоятельных форматов метаданных. Она опирается на модульную модель так называемых компонентов метаданных, которые могут быть собраны вместе для оптимизации повторного использования, взаимодействия и сотрудничества между разработчиками моделей метаданных. Относительно небольшой набор компонентов (общие метаданные, метаданные для текстовых ресурсов, метаданные для мультимедиа и метаданные о людях) может быть объединен в индивидуальные профили.

CMDI – это не просто еще один формат. Это гораздо больше: как метамодель она обеспечивает четко определенную структуру для определения и использования вашего собственного формата. Он также позволяет пользова-

[1] Component Metadata. – URL: https://www.clarin.eu/content/component-metadata (дата обращения: 01.12.2021).

телю интегрировать существующие схемы (IMDI, OLAC) в качестве компонентов и таким образом обеспечивает совместимость с существующей базой.

Ни одна единая схема метаданных никогда не сможет удовлетворить все потребности разнородного сообщества исследователей гуманитарных и социальных наук: они варьируются от описания греческих текстов на вазах до анализа жестов в видеороликах YouTube и записи фонетических особенностей телефонных записей. Отсюда и необходимость гибкого решения.

Далее приводятся разделы CMDI, материалы, направленные на методическую поддержку и развитие CMDI, функциональные модули, инструменты и сервисы CMDI для создания компонентов и профилей CMDI, а также сведения и инструменты по применению CMDI в некоторых репозиториях ЛИР.

***Часто задаваемые вопросы***[1] включают следующие рубрики.

- Метаданные в CLARIN: создание и редактирование CMDI
- Метаданные в CLARIN: основы
- Метаданные в CLARIN: преобразование в CMDI
- Метаданные в CLARIN: сбор метаданных и VLO
- Метаданные в CLARIN: понятия

***Спецификация CMDI 1.2*** [14]

CMDI был разработан в европейской инфраструктуре CLARIN с участием других инициатив и экспертов. Уже на подготовительном этапе, который начался в 2007 году, CLARIN нуждалась в гибкости в области метаданных, поскольку она сталкивалась со многими типами ресурсов, которые должны были быть точно описаны. Для версии 1.0 был создан инструментарий CMDI, состоящий из XML-схем и таблиц стилей XSLT для проверки и преобразования компонентов, профилей и записей. Эта версия использовалась на протяжении всего этапа становления CLARIN. CMDI учитывает растущее число инструментов и инфраструктурных систем, которые имеют дело с записями и компонентами CMDI с общим синтаксисом и семантикой.

Новая версия CMDI 1.2 более функциональна и менее проблематична. Эти изменения освещены в документе CE-2014–0318. Переход от 1.1 к 1.2 поддерживается версией 1.2 инструментария CMDI. Описывается жизненный цикл метаданных, устанавливается содержание работ на каждом этапе. Спецификация CMDI содержит описание структуры файла CMD, правила установлений отношений между ресурсами, язык описания.

Более подробная информация об изменениях в CMDI 1.2 может быть найдена на странице[2]. Общая информация на этой странице относится как к CMDI 1.1, так и к CMDI 1.2. Далее приводятся основные разделы CMDI 1.2.

***Примеры и наборы данных[1]***

[1] Frequently Asked Questions – Metadata in CLARIN. – URL: https://www.clarin.eu/faq-page/267 (дата обращения: 01.12.2021).

[2] CMDI 1.2. – URL: https://www.clarin.eu/cmdi1.2 (дата обращения: 01.12.2021).

Раздел содержит примеры – тестовые, реальные, большие и малые наборы данных, в том числе полученные путем сбора метаданных.

***Руководство CMDI по передовой практике*** **(проект) [15]**

Руководство по передовой практике содержит сборник общих рекомендаций по моделированию и созданию метаданных CMDI CLARIN. Руководство также содержит набор описаний общих подходов и проблем.

***Руководство CMDI по детализации при описании ЛИР (проект*****) [16]**

Поскольку этот документ содержит рекомендации тем, кто хочет создать хранилище в центре CLARIN, то не все рекомендации применимы в других контекстах:

- CLARIN ориентирован на распределенное хранение ЛИР в десятках центров, где большинство из них имеют собственный репозиторий;
- ресурсы доступны в Интернете;
- метаданные для ресурсов хранятся в формате CMDI, где каждый центр может использовать свой собственный профиль ресурсов;
- все метаданные собираются через OAI PMH и впоследствии включаются в:
  - поисковые системы
  - порталы (например VLO);
- создание метаданных полностью находится под контролем центра, после сбора данных файлы метаданных должны быть полностью расширены;
- CMDI – основа для исследовательской инфраструктуры, которая должна быть реализована в относительно краткосрочной перспективе, быть надежной, хорошо масштабироваться и быть пригодной для неподготовленных пользователей.

Специальный раздел документа посвящен соотношению метаданных и лингвистических аннотаций. Делается вывод, что хотя эти подходы к описанию ЛИР в значительной степени пересекаются, метаданные отличаются большей стабильностью и универсальностью.

***Реестр компонентов***

Данная подсистема CMDI[2] имеет следующие функции:

- регистрация и хранение компонентов / профилей
- просмотр зарегистрированных компонентов / профилей
- редактирование и создание компонентов / профилей

На начальном экране приложения реестра компонентов отображается браузер компонентов. Отсюда можно получить доступ к функциям редактирования компонентов. В браузере компонентов есть выпадающее меню, которое

[1] CMDI Examples. – URL: https://www.clarin.eu/content/cmdi-examples (дата обращения: 01.12.2021).

[2] Component Registry Documentation. – URL: https://www.clarin.eu/content/component-registry-documentation (дата обращения: 01.12.2021).

позволяет выбирать между двумя типами элементов: «профили» и «компоненты». Выбор одного из них позволяет просмотреть все зарегистрированные профили или компоненты в таблице. Доступные параметры в этом меню, которые можно переключать, – это *публичное пространство*, *личное пространство пользователя* и одна или несколько *общих команд пространства*.

В публичном пространстве отображаются все опубликованные профили / компоненты. В пользовательском пространстве отображаются все профили / компоненты, расположенные в вашем собственном личном рабочем пространстве. Если вы являетесь членом одной или нескольких команд, в командных пространствах отображаются профили / компоненты, видимые и редактируемые всеми участниками этих команд. Элементы пространства пользователя и коллектива не публикуются.

Наконец, есть выпадающее меню *фильтр состояния*, позволяющий переключаться между типами состояния, для которых отображаются значения: *разработка, производство* или *устаревшие*.

***Реестр понятий CLARIN***[1]

Концептуальный (понятийный) реестр CLARIN (CCR) образует основу семантического слоя совместимости CLARIN, особенно в контексте метаданных, т.е. компонентной инфраструктуры метаданных (CMDI). Предлагается набор понятий с их постоянными идентификаторами, имеющим отношение к предметной области ЛИР. Подробное описание реестра см. ниже.

***Редактор метаданных COMEDI***[2]

COMEDI – это веб-редактор для CMDI-соответствующих метаданных, как принято в CLARIN. С помощью COMEDI можно интерактивно создавать новые записи метаданных CMDI или загружать и изменять существующие метаданные. Запись метаданных в COMEDI можно экспортировать в виде XML-файла CMDI. Его можно также передавать через OAI-PMH. COMEDI обрабатывает любой CMDI-совместимый профиль. (В настоящее время поддерживается версия 1.1; планируется поддержка CMDI 1.2.) COMEDI лучше всего работает в Safari или Chrome. Он работает в Firefox, но существуют некоторые незначительные проблемы. Другие браузеры не были протестированы.

***Coala*** [3]

Coala – инструмент для преобразования электронных таблиц метаданных для различных наборов речевых данных в стандартизированные файлы CMDI.

***CMDI Maker***[4]

[1] CLARIN Concept Registry. – URL: https://www.clarin.eu/ccr (дата обращения: 01.12.2021).

[2] COMEDI :: The COmponent Metadata EDItor. – URL: https://clarino.uib.no/comedi/page (дата обращения: 01.12.2021).

[3] BAS Web Services. – URL: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Coala (дата обращения: 01.12.2021).

[4] CMDI Maker. – URL: http://cmdi-maker.uni-koeln.de/ (дата обращения: 01.12.2021).

CMDI Maker – простое в использовании веб-приложение HTML5 для быстрого создания научных метаданных.

***Arbil редактор CMDI***[1]

Arbil – общий редактор метаданных, браузер и органайзер для IMDI, CMDI 6 и аналогичных форматов метаданных**.** Arbil сконструирован так, что его можно использовать и офлайн, и дистанционно. Данные могут быть введены на любом этапе частично или в целом. Руководство по использованию этого редактора представлено в документе [17][2].

***Просмотр, поиск и использование метаданных CMDI***

Основной инструмент для поиска ЛИР с помощью CMDI – это Виртуальная языковая обсерватория, подробно описанная в главе 2. Другой инструмент для работы с ресурсами CLARIN – поисковая машина Института Меертенса[3], при помощи которой можно искать по наименованию коллекций или по схеме профиля ЛИР.

***CMDI-to-RDF***[4]

Проект *CLARIAH CMD2 RDF* направлен на то, чтобы преобразовать записи CLARIN CMD в виде связанных открытых данных, что позволит интегрировать эти записи в облако LOD. В этом проекте предлагается три интерфейса:

- браузер для представления в RDF коллекции записей CMD;
- *RESTful* – интерфейс для доступа к представлению RDF отдельных записей или всех записей из хранилища CLARIN;
- конечная точка SPARQL для запроса на представление в RDF коллекции записей CMD.

Эти сервисы в основном интересны технической аудитории, которая может создавать новые сервисы, включая пользовательский интерфейс, на этих базовых сервисах.

Далее приведем несколько примеров практического внедрения CMDI.

***Использование Islandora (Fedora + Drupal) с CMDI***[5]

Гамбургский центр языковых корпусов (HZSK) имеет своей целью сохранение и обеспечение доступности речевых корпусов, собранных Специ-

[1] Arbil information, manuals & download. – URL: https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045 (дата обращения: 01.12.2021).

[2] Virtual Language Observatory (VLO). – URL: https://www.clarin.eu/content/virtual-language-observatory-vlo (дата обращения: 01.12.2021).

[3] Search resourcesat CLARIN in Europe. – URL: https://www.meertens.knaw.nl/cmdi/search/#q=*%3A* (дата обращения: 01.12.2021).

[4] The CLARIAH CMD2 RDF. – URL: http://cmdi2rdf.meertens.knaw.nl/cmd2rdf/ (дата обращения: 01.12.2021).

[5] Islandora. Open source digital asset management. – URL: https://www.islandora.ca/ (дата обращения: 01.12.2021).

альным исследовательским центром по многоязычию за 11 лет его существования. Очевидно, что центральную роль при этом имеют метаданные. Будучи одним из центров CLARIN, HZSK начал переход на использование CMDI.

В существующей среде корпусами управляет файловая система. Этот подход оказался негибким, небезопасным, трудным в обслуживании и расточительным из-за высокой избыточности данных. Чтобы решить эти проблемы, было проведено исследование по размещению цифрового репозитория на сервере. Описывается процесс применения для репозитория информационной системы Islandora (Fedora + Drupal). При этом было необходимо использовать CMDI. В цитируемой статье обсуждаются проблемы, возникающие при установке Islandora и CMDI, и преимущества использования этой среды для речевых данных. В результате проблемы были решены программным обеспечением репозитория, но, как и ожидалось, возникли другие сложности.

***Использование Fedora с CMDI***

В этом документе [18] в намеренно сжатой форме описывается установка и настройка Fedora, а также установка и настройка службы поставщика OAI, размещенной на ней.

Цель состоит в том, чтобы доставить метаданные в формате CMDI объектам, хранящимся в Fedora. Этот пример предполагает, что Fedora будет работать на MySQL и Tomcat.

Цель состояла в том, чтобы сформировать метаданные в формате CMDI для объектов, хранящихся в Fedora. Это было реализовано под управлением Linux, но результат можно отнести и к другим операционным системам.

***Организация репозитория с помощью Fedora и CMDI в Институте немецкого языка (IDS)***

Репозиторий IDS использует Fedora Commons в качестве базовой платформы. Ресурсы корпуса, подлежащие архивированию в хранилище IDS, представлены в самых разнообразных моделях и форматах и поддерживаются с помощью самых разнообразных систем, включающих реляционные базы данных, файловую систему и другие частные решения. Описание проекта организации репозитория [19] включает четыре этапа:

*Выравнивание.* Метаданные и данные выравниваются по отношению друг к другу. Часто метаданные представляются в виде вектора, разделенного запятыми, со ссылками на фактические данные. Этот шаг обеспечивает однозначное соответствие между данными и метаданными, что часто требует нормализации ссылок и идентификаторов.

*Валидация / курирование*. Форматы данных проверяются с помощью специальных валидаторов форматов, а для данных которые недоступны в одном из рекомендуемых форматов генерируются дополнительные представления (обычно на основе XML DocBook или TEI для письменных корпусов).

*Извлечение метаданных*. Дополнительные метаданные, такие как название или дата выпуска, извлекаются из данных, используя преимущества форматов данных на основе XML, созданных на предыдущем шаге.

*Генерация CMDI*. Метаданные преобразуются в подходящий профиль компонента CMDI. Часто этот шаг включает в себя спецификацию нового профиля, частично основанного на существующих компонентах CMDI, чтобы надлежащим образом представить доступные метаданные.

***CMDI в репозиториях на DSpace (Lindat/CLARIN*)**

Страница портала репозитория Lindat/CLARIN[1] содержит описание деятельности по управлению и поддержке метаданных. Приводится список поддерживаемых форматов метаданных[2].

Для каждого формата существует ссылка на пространство имен и на схему. Для поставщиков ЛИР в репозиторий приводится список обязательных и рекомендуемых элементов метаданных. Дается сопоставительная таблица форматов CMDI и OAI DC.

## Стандартизация метаданных

На основе метаданных компонентов CMDI разработан стандарт *ISO 24622 Управление языковыми ресурсами – Инфраструктура метаданных компонентов (CMDI)* [20]. Стандарт включает две части:

ISO 24622–1:2015 Часть 1: Модель метаданных компонентов [20]

ISO 24622–2:2019 Часть 2: Язык спецификации метаданных компонентов [21]

Ниже приводится сокращенное изложение этого стандарта.

Метаданные компонентов (CMD) – это подход к моделированию и созданию метаданных. В наши дни он все чаще используется для обеспечения возможности создания метаданных для различных типов ЛИР, но с учетом синтаксической и семантической совместимости. CMD также является ядром инфраструктуры метаданных компонентов (CMDI): эта инфраструктура содержит спецификации формата, набор реестров и инструментов для моделирования и создания метаданных.

Преимущества наличия унифицированного подхода к описаниям метаданных для ЛИР, который будет использоваться многими проектами и инициативами, очевидны: во-первых, есть больше шансов для достижения согласованности и совместимости между описаниями метаданных из разных источников, а во-вторых – возможность разрабатывать инструменты, которые работают намного более эффективно в этой структуре метаданных.

Ландшафт метаданных для ЛИР продолжает оставаться фрагментарным. До недавнего времени при создании описаний метаданных для ЛИР использовалась практика выбора конкретной схемы метаданных из широко распространенных схем, либо из других дисциплин; например, OLAC – это адаптированная версия DCMI, восходящей к библиотечной сфере. Кроме того, для целей описания метаданных ЛИР существуют специально разрабо-

[1] LINDAT/CLARIAH-CZ Repository Home. – URL: https://lindat.mff.cuni.cz/repository/xmlui/page/metadata (дата обращения: 01.04.2022).

[2] List of Metadata Formats – URL: https://lindat.mff.cuni.cz/repository/oai/request?verb=ListMetadataFormats (дата обращения: 01.04.2022).

танные схемы метаданных определенных типов ЛИР (например IMDI). Существующие противоречия влекут за собой плохую совместимость системы метаданных ЛИР.

Поэтому международные, европейские и национальные инфраструктуры, такие как CLARIN и META-SHARE, теперь разделяют подход CMD к метаданным для ЛИР.

Описание модели – это первая часть инфраструктуры, которая формирует полный пакет для создания схем метаданных. Полный стандарт инфраструктуры содержит один или несколько языков спецификации компонентов метаданных и ряд рекомендуемых компонентов и профилей метаданных (запланировано). Поскольку эта часть ISO 24622 определяет абстрактную модель, для ее описания используется UML.

Эта часть ISO 24622 упрощает разработчикам моделей метаданных (например исследователям и специалистам по описанию ресурсов) создание новых схем метаданных, которые в свою очередь могут использоваться либо для описания новых типов ресурсов, либо для включения в более подходящее описание ресурсов в конкретных обстоятельствах. Схема метаданных воплощается в записи метаданных. Этот вариант требует гибкой структуры, которая позволяет легко создавать новые схемы метаданных для различных целей, но также является структурой (i), в которой экземпляры имеют строго определенный формат, что обеспечивает синтаксическую корректность, и (ii) содержится семантическая характеристика элементов схемы метаданных для интерпретации содержимого записи метаданных.

Описания метаданных, сгенерированные схемами, совместимыми с этой моделью, также будут соответствовать другим международным стандартам ТС 37, например тем, которые требуют, чтобы ссылки на описанные ресурсы и части ресурсов использовали постоянные идентификаторы (PID), совместимые с ISO 24619: 2011.

Определение ресурса в этом контексте очень широкое. В этой части ISO 24622 используется прагматический взгляд: например, изображение может быть ресурсом само по себе, если оно связано с PID и можно ссылаться на него как таковое, или оно может быть частью документа, в котором отсутствует его собственная идентификация. Кроме того, ссылка может указывать на часть этого изображения. Отдельный ресурс может находиться отдельно в одной среде и рассматриваться как часть коллекции в другой. Кроме того, описания метаданных включают ресурсы, но они тоже являются ресурсом в разных контекстах. Эта часть ISO 24622 должна поддерживать все такие случаи, а модель должна предоставлять описания на всех уровнях детализации.

В этой части ISO 24622 учитываются два типа коллекций:

а) сложный ресурс мог быть изначально создан как коллекция, и не считая управления версиями он будет существовать как таковой в статической опубликованной форме. Его спецификация будет рассматриваться как независимый объект ответственным архивным учреждением, которое также предоставляет PID для такой коллекции. В контексте этой части ISO 24622 метаданные для коллекции являются спецификацией коллекции. Архивное учреждение отвечает за поддержание метаданных, представляющих коллекцию;

б) напротив, другой тип коллекции – это тот, который не планировался и не создавался как коллекция его создателями или архивом хранения, но приобрел статус объединенного ресурса на основе исследований, которые должны быть проверены. Такие коллекции, хотя и специально созданы исследователем, могут не иметь никакого значения вне контекста исследования, для которого они были созданы. Ссылка из исследовательских документов на коллекцию также может стать перегруженной, если коллекция содержит сотни отдельных ресурсов. Необходимо фиксировать эти типы коллекций с помощью записи метаданных, которая связана со всеми составляющими ее ресурсами и соответствующими метаданными, но только как воплощение этой коллекции. Однако отсутствует ответственная структура за ведение и поддержку записи метаданных. Маловероятно, что исследователь, создавший «виртуальную» коллекцию, имеет какой-либо способ постоянно поддерживать и курировать эту запись метаданных в долгосрочной перспективе. Цифровые архивы или издатели могут вести специальные реестры, в которых исследователи могут регистрировать такие виртуальные коллекции.

Оба типа коллекции идентифицируются с помощью PID, который относится к метаданным коллекции.

Сфера применения этой части ISO 24622 состоит в описании модели, которая обеспечивает гибкое построение интероперабельных схем метаданных для языковых ресурсов (ЛИР). Схемы метаданных, основанные на этой модели, могут использоваться для описания ресурсов на разных уровнях детализации (например, описания как на уровне коллекции, так и на уровне отдельных ресурсов).

Вторая часть стандарта ISO 24622–2:2019 содержит представление языка спецификации метаданных компонентов.

В CMDI жизненный цикл метаданных начинается с потребности в моделировании метаданных для создания специального профиля метаданных для определенного типа ресурса. Разработчики моделей могут просматривать и искать в реестре компоненты и профили, которые подходят или почти соответствуют их требованиям. Компонент группирует вместе элементы метаданных, которые принадлежат друг другу и потенциально могут быть повторно использованы в другом контексте.

Существующие реестры компонентов могут уже содержать любое количество компонентов. Их можно повторно использовать в том виде, в каком они есть, или адаптировать путем изменения, добавления или удаления некоторых элементов и / или компонентов метаданных. Также могут быть созданы совершенно новые компоненты для моделирования уникальных аспектов рассматриваемых ресурсов. Все необходимые компоненты объединены в один профиль, соответствующий типу ресурсов.

Любой компонент, элемент и значение в таком профиле могут быть связаны с семантическим описанием – концептом, отображающим содержательное значение. Эти семантические описания могут храниться в семантическом реестре, например Регистре понятий CLARIN (см. ниже). Создатели метаданных могут создавать записи для определенных ресурсов, которые

соответствуют профилю, соответствующему типу ресурса, и эти записи могут быть предоставлены в локальные и глобальные каталоги.

Жизненный цикл метаданных компонента требует комплексной инфраструктуры с хорошо взаимодействующими друг с другом системами.

Вторая часть стандарта, кроме вступления и определения терминов, включает следующие разделы:

- Условные обозначения и пространство имен XML
- Структура экземпляров CMDI
  - Общая структура
  - Основная структура
  - Элемент <Header> element
  - Элемент <Resources> element
  - Элемент <IsPartOfList> element
  - Компоненты CMD
- Язык спецификации компонентов CMDI (CCSL)
  - Общая структура CCSL
  - Заголовки CCSL
  - Спецификации CMD
  - Определение элементов CMD
  - Определение атрибутов CMD
  - Схемы значений для CMD элементов и атрибутов
- Метаданные компонентов (CMD)
  - Преобразование CCSL в определение схемы профиля CMD
  - Общие свойства определения схемы профиля CMD
  - Интерпретация спецификаций CMD в CCSL
  - Интерпретация определений элементов CMD в CCSL
  - Интерпретация определений атрибутов CMD в CCSL
  - Модель контента для элементов и атрибутов CMD в определении схемы

### Остинские принципы цитирования лингвистических данных

Остинские принципы цитирования лингвистических данных были разработаны Калифорнийским университетом в Санта-Барбаре. Эти принципы разработаны в рамках проекта по цифровой лингвистике и опираются на формат данных DLx[1], созданный в рамках этого проекта. (См. также главу 9.) В этом документе [22] представлены руководящие принципы, которые позволяют лингвистам принимать обоснованные решения в отношении доступности и прозрачности их научных данных.

Составление, комментирование и редактирование началось летом 2017 года. В настоящее время Остинские принципы прошли несколько раундов открытого комментирования и редактирования и будут по-прежнему открыты для комментариев. Новые версии станут появляться по мере совер-

[1] The Digital Linguistics (DLx) developers site. – URL: https://developer.digitallinguistics.io/ (дата обращения: 01.12.2021).

шенствования инфраструктуры и стандартов цитирования данных. Приведем основное содержание этого документа.

*Значимость*. Данные следует рассматривать как легитимные, цитируемые продукты исследования. Данные, на которых основан лингвистический анализ, имеют фундаментальное значение для данной области и должны рассматриваться как таковые. Лингвистические данные следует цитировать и, по возможности, повторно использовать.

*Заслуги и авторство*. Цитирование данных должно способствовать укреплению научного авторитета, а также нормативной и юридической атрибуции всех авторов лингвистических данных.

*Авторитетность источника*. Лингвистам следует приводить данные, на которых основаны научные утверждения. Чтобы данные можно было цитировать, они должны храниться в доступном и надежном репозитории.

*Уникальная идентификация*. Цитирование данных должно включать в себя постоянную идентификацию, которая должна быть компьютеризованной, уникальной и широко используемой сообществом.

*Доступ*. Ссылки на данные должны облегчить доступ к самим данным и к связанным метаданным. Лингвистические данные должны быть максимально открытыми, чтобы облегчить воспроизводимость, и настолько закрытыми, насколько это необходимо.

*Надежность*. Уникальные идентификаторы и метаданные, описывающие данные и их расположение, должны сохраняться даже по истечении срока жизни данных, которые они описывают.

*Специфичность и проверяемость*. Цитирование данных должно облегчить читателю поиск конкретных данных или подмножества данных в более крупном наборе данных, подтверждающих некое утверждение.

*Интероперабельность и гибкость*. Методы цитирования данных должны быть достаточно гибкими, чтобы приспособиться к разным практикам разных сообществ, и их отличительные признаки не должны влиять на интероперабельность.

## Реестр категорий данных для ЛИР

***Краткая история***

Разработка и реализация систем метаданных для описания ЛИР существенным образом зависит от качества (полноты, точности и однозначности) терминов, используемых в этих метаданных. Еще в начале деятельности по созданию метаданных была поставлена задача формирования реестра категорий лингвистических данных, т.е. словарей терминов (понятий), применяемых для метаописаний ЛИР.

Соответствующий стандарт был впервые выпущен ISO ТК 37 как ISO 12620:1999, который позже был признан устаревшим и появилось второе издание ISO 12620: 2009 [23].

Второе издание было также признано устаревшим в соответствии с ISO 12620:2019 [24]. Однако третье издание больше не предоставляет реестр терминов для языковых технологий и терминологии, теперь оно ог-

раничено терминологическими ресурсами; отсюда обновленное название «Управление терминологическими ресурсами – Спецификации категорий данных».

Второе издание ISO 12620:2009 было переведено на русский язык в форме российского ГОСТ Р ИСО 12620–2012 [25]. Содержание этого ГОСТа будет изложено ниже.

Реестр категорий лингвистических данных (DCR) под эгидой ISO ТК 37 был создан в 2008 году в Институте психолингвистики Общества Макса Планка (MPI) в Неймегене, Нидерланды, под названием ISOcat.

Оригинальный ISOcat был задуман как официальный онлайновый реестр информации о категориях данных для поддержки исследований и разработок в различных лингвистических дисциплинах. Однако некоторым пользователям нужна концептуальная база данных, предназначенная для поиска данных в больших текстовых корпусах, что требует иной модели данных, чем ISOcat. Кроме того, первоначальный мандат на «стандартизацию» категорий данных в рамках ИСО так и не был выполнен, и в конечном итоге был признан ненужным в пользу более открытой и публичной версии.

В результате отпала необходимость в регистрирующем органе, и хранилище ISOcat перестало быть проектом ИСО. Был создан DatCatInfo[1] – это репозиторий категорий данных (DCR), который заменяет ISOcat. Лидеры сообщества пользователей решили создать хранилище определений категорий данных, переименованное в DatCatInfo и поддерживаемое отраслевыми лингвистами и терминологами.

Следующим этапом развития регистра стало создание для пользователей CLARIN нового реестра CLARIN Concept Registry (CCR)[2]. Этот реестр размещен в институте Meertens. В начале февраля 2015 года реестр компонентов был обновлен для использования CCR вместо ISOcat.

Наконец, еще один ресурс, содержащий лингвистическую терминологию, в том числе метаданных, был внесен в качестве модуля метаданных пакета словарей Ontolex Ontology-Lexicon, а затем размещен на облаке LLOD.

Ниже будет представлено описание перечисленных проектов.

### ***ГОСТ Р ИСО 12620–2012 [25]*** (сокращенное изложение)

*Введение*

Идентификация, сбор, администрирование и хранение данных, ассоциируемых с ЛИР, выполняются в многочисленных разнообразных средах. Элементы данных, входящие в описания отдельных ЛИР, рассматриваются в настоящем стандарте как категории данных. Различия в подходах, используемых для разных типов ЛИР, неизбежно приводят к отличиям в определениях и именах категорий данных. Использование единообразных имен и определений категорий данных для ресурсов одной тематической области (например, для терминологических и лексико-графических ресурсов, текстовых аннотаций и т.д.), по крайней мере на уровне обмена, способствует со-

[1] Data Category Repository (DCR). – URL: https://datcatinfo.net/ (дата обращения: 01.12.2021).

[2] CLARIN Concept Registry. – URL: https://www.clarin.eu/ccr (дата обращения: 01.12.2021).

гласованности систем и расширяет возможности повторного использования данных.

*Область применения*

В настоящем стандарте приведены руководящие указания относительно ограничений реализации реестра категорий данных (DCR) для любых типов языковых ресурсов, например, терминологических, лексикографических, основанных на использовании сборников или машинного перевода и т.д. Определены механизмы создания, выбора и ведения категорий данных, а также формат обмена для представления этих категорий.

*Роль категорий данных в управлении языковыми ресурсами*

Спецификации категорий данных описывают отдельные информационные блоки, определяющие схему сбора или аннотации данных для конкретного ЛИР. Каждая спецификация задает формальное представление категории данных и включает конкретные признаки, описывающие эту категорию (например, ее имя, определение, примеры, комментарии и т.д.).

*Выборки категорий данных (DCS)*

Группы категорий данных, которые выделены в качестве подмножеств глобального реестра DCR, образуют выборки категорий данных (DCS). В зависимости от приложения, выборка DCS может быть просто списком категорий данных с обратной ссылкой на полные спецификации в DCR, либо она может быть представлена полным поднабором или даже расширенным набором DCR, состоящим из такого списка с добавленными определениями и ограничениями, связанными с конкретными спецификациями категорий данных.

DCS можно рассматривать как документированный источник для применяемой схемы лингвистической аннотации. Так как в DCS содержится список всех категорий данных, которые могут использоваться вместе со схемой аннотации, то вероятно, это лучший источник информации для потенциальных пользователей или разработчиков.

Некоторые категории данных относятся лишь к одной тематической области или типу ЛИР. Например, категория */conceptIdentifier/ (идентификатор понятия),* вероятно, уникальна для терминологических ресурсов, а категория */senseNumber/ (номер значения)*, возможно, характерна для лексикографических ресурсов. С другой стороны, многие категории данных, часто фиксированной лингвистической природы, например, */partOfSpeech/ (часть речи), /grammaticalGender/ (грамматический род), /grammaticalNumber/ (грамматическое число)* и т.д., являются общими для самых разнообразных ресурсов.

Подмножество категорий данных и их спецификаций, используемое в тематической области, должно составлять специфическую для этой области выборку DCS из реестра DCR. Для конкретного приложения можно затем составить подмножество из выборок DCS одной или нескольких тематических областей. Это подмножество целиком содержится в выборке DCS для терминологических категорий данных и предназначено для терминологического приложения, хотя некоторые содержащиеся в нем спецификации категорий данных являются общими для ЛИР разного рода. Некоторые при-

ложения могут заимствовать категории данных из DCS нескольких тематических областей. Если входящие в DCS категории в настоящее время не являются частью DCR, возможно эта выборка будет надмножеством категорий. В таких случаях разработчикам и пользователям рекомендуется регистрировать новые категории данных в DCR.

*Требования к реализации реестра DCR для ЛИР*

Реестр DCR должен:

- быть справочным хранилищем категорий данных и связанной информации для всех существующих и будущих стандартов;
- быть бесплатно доступным в интерактивном режиме (online);
- обеспечивать регистрацию существующих практических схем путем включения каждой категории данных и способа, которым она реализована в конкретных проектах или инициативах;
- предоставлять имена и определения на различных языках;
- описывать использование каждой категории данных на разнообразных объектных и рабочих языках;
- описывать использование категорий данных в разнообразных ЛИР;
- связывать административную информацию с каждой категорией данных, что позволит отслеживать представление, одобрение или пересмотр категорий данных;
- связывать каждую категорию данных с одним или несколькими профилями, соответствующими тематическим областям, к которым относится эта категория;
- предоставлять механизм для рассмотрения групп категорий данных;
- предоставлять частные рабочие области для создания или загрузки спецификаций и выборок категорий данных пользователями;
- регулярно обновляться путем включения предложений, полученных от экспертов в данной области;
- быть всегда доступным во всем мире благодаря распределению копий на сайтах-зеркалах;
- предусматривать долгосрочные идентификаторы благодаря системе твердой привязки ссылок на отдельные спецификации данных;
- поддерживать безопасные передовые практические методы архивирования;
- предусматривать предоставление периодических снимков текущего состояния.

***ISOcat***

Как было отмечено выше**,** регистр ISOcat в настоящее время не поддерживается. Однако пока эта база данных была доступна, проводились интересные исследования. Например, специалисты Тюбингенского университета разработали интерпретацию дерева зависимостей для регистра ISOcat [26]. Ниже кратко описывается данная работа.

Лингвистическое сообщество создает инфраструктуру на основе метаданных для описания своих исследовательских данных и инструментов. В его основе лежит реестр ISOcat, совместная платформа для хранения (подлежа-

щего стандартизации) набора категорий данных (т.е. дескрипторов). Дескрипторы имеют определения на естественном языке и мало выраженных взаимосвязей. С ростом реестра до многих сотен записей, авторами которых являются многие, становится все более очевидным, что довольно неформальные определения и их конструкция в виде глоссария затрудняют пользователям понимание, использование и управление содержимым реестра.

Если из большого подмножества набора терминов ISOcat восстановить древовидную структуру, то при таком онтологическлм реинжиниринге пользователь получает более полное новое представление о лингвистической терминологии, связанной с метаданными. Полученная иерархия доступна на цитированном сайте ISOcat, а для академического и исследовательского использования приводится XML-файл[1].

Новое представление повышает точность всех определений, делая явной информацию, которая только неявно дана в реестре ISOcat. Она также помогает выявить и устранить потенциальные несоответствия в определениях терминов, а также пробелы и избыточности в общем наборе терминов ISOcat. Новое представление может служить дополнением к существующей модели ISOcat, обеспечивая дополнительную поддержку авторам и пользователям в просмотре, повторном использовании, поддержке и дальнейшем расширении терминологического репертуара метаданных сообщества. Набор категорий данных был взят из метаданных TDG реестра категорий данных ISOcat.

***База Данных DatCatInfo***[2]

Как отмечалось выше, DatCatInfo – это репозиторий категорий данных (DCR), разработанный в соответствии с ISO 12620: 2019, который заменяет ISOcat. DatCatInfo поддерживается LTAC Global/TerminOrgs, которая является связующим звеном с ISO TC37.

Для поиска в массиве данных DatCatInfo разработана база данных TERMWEB[3]. Поисковые параметры этой базы:

- по термину
- по словарю
- по разделу словаря
- по области применения
- по языку
- по полям профиля

***Реестр понятий CLARIN***[4]

Реестр понятий CLARIN (CCR) образует основу семантического слоя совместимости CLARIN, особенно в контексте метаданных, т.е. компонентной инфраструктуры метаданных (CMDI). Реестр предлагает набор понятий с

[1]XML file. – URL: http://www.sfs.uni-tuebingen.de/nalida/images/isocat/isocatOWL.owl (дата обращения: 01.12.2021).

[2] Data Category Repository (DCR). – URL: http://datcatinfo.net/ (дата обращения: 01.12.2021).

[3] Termweb. – URL: https://datcatinfo.termweb.se/termweb/app (дата обращения: 01.12.2021).

[4] CLARIN Concept Registry. – URL: https://www.clarin.eu/ccr (дата обращения: 01.12.2021).

их постоянным идентификатором, имеющим отношение к предметной области ЛИР. CCR содержит 3163 понятий и определений. Пример:

***writing systems*** *The visual representation of spoken language on paper or other media, and the issues involved in writing and creating a writing system. (source: CLARIN)*

(***системы письма*** *Визуальное представление устной речи на бумаге или других носителях, а также вопросы, связанные с написанием и созданием системы письма. (Источник: CLARIN.*))

Доступ к реестру понятий через фасетный браузер может получить любой пользователь в режиме *только для чтения*. Добавление новых понятий или изменение существующих могут быть осуществлены только национальными координаторами CCR.

*Фасетный браузер для поиска в CCR*[1]

В CCR возможен поиск по части термина или по термину целиком. Также возможен поиск по полям БД:

- наименование
- определение
- поля документации по умолчанию

Ниже приводится классификация фасетных фильтров в CCR. В скобках указывается число понятий, включенных в CCR.

- Статус
  - Одобренный (236)
  - Кандидат (2847)
  - Истекший (80)
  - Любой (0)
- Концептуальные схемы
  - Диалоговые акты (1)
  - Код языка (4)
  - Онтология языковых ресурсов (4)
  - Лексические ресурсы (15)
  - Лексическая семантика (4)
  - Лексикография (25)
  - Метаданные (1520)
  - Морфосинтакс (399)
  - Многоязычное управление информацией (1)
  - Представление семантического содержания (210)
  - Язык жестов (115)
  - Синтаксис (95)
  - Терминология (98)
  - Перевод (10)
  - Лингвистика (0)
  - Метаданные (0)
  - Морфология (0)

[1] CLARIN Concept Registry Browser. – URL: https://concepts.clarin.eu/ccr/browser/ (дата обращения: 01.12.2021).

- Не доступен (4)
- Фонология (0)
- Синтаксис (0)
- Не определено (875)

- Коллекции (список)
- Владелец коллекции (список)

***Словарь лингвистических метаданных (LIME)***[1]

LIME (лингвистические метаданные – LInguistic MEtadata) – это словарь лингвистических метаданных о ЛИР. Словарь метаданных был первоначально представлен на конференции LDL, а затем внесен в качестве модуля метаданных пакета словарей Ontolex Ontology-Lexicon.

Первоначально разработанный в рамках арт-группы Римского университета Tor Vergata, LIME послужил основой дискуссии в рамках группы сообщества W3 C OntoLex об общей терминологии метаданных.

Сегодня LIME является модулем метаданных пакета словарей OntoLex для определения интерфейсов онтологий и лексиконов.

Несмотря на то, что он содержит специфику, связанную с моделью OntoLex, словарь LIME подходит для описания актива любого вида лексикализации. Описание этого актива состоит в качественной и количественной информации о лексической реализации описываемого набора данных. Соответствующие метаданные включают список естественных языков, принятых для лексикализации набора данных, лексические модели, принятые для обеспечения лексикализации, а также статистические данные о покрытии элементов набора данных лексическими записями для каждого данного языка.

Словарь LIME (который строится поверх и дополняет другие существующие словари метаданных, такие как Dublin Core или DCAT) рекомендуется использовать для улучшения видимости онтологий и наборов данных, с тем чтобы улучшить их доступность и квалифицировать их лексическую характеристику.

Спецификация словаря LIME в настоящее время доступна в специальном разделе модели OntoLex, описанной в отчете [27].

***LexInfo***

LexInfo – это онтология, которая была разработана для предоставления категорий данных в модели Lemon. С тех пор она была обновлена в виде новой модели OntoLex-Lemon группы сообщества OntoLex. LexInfo теперь представлена на GitHub[2]. В качестве языка разметки LexInfo использует Lexical Markup Framework, версия 1.0.

## Исследование лексики метаданных российских ЛИР

[1] Linguistic Metadata (LIME) vocabulary. – URL: https://old.datahub.io/dataset/lime (дата обращения: 01.12.2021).

[2] Ontolex/lexinfo. – URL: https://github.com/ontolex/lexinfo (дата обращения: 01.12.2021).

В данном разделе описано исследование лексики метаданных российских ЛИР, проведенное с участием автора, целью которого являлось создание лексической и терминологической основы для полноценной онтологии по лингвистике (языкознанию). Продукт, который получил название *Онтология поисковых терминов по лингвистике (ОПТЕЛ*), также может служить для навигации и / или метапоиска в российских ЛИР. Принципы отбора источников для ОПТЕЛ, структура БД, особенности отдельных словарей метаданных описаны в работе [28]. В настоящее время ОПТЕЛ реализована и размещена в Интернете [1].

Реализованная версия ОПТЕЛ включает 55 словарей, использованных в 28 российских ЛИР разных типов. Всего в ОПТЕЛ представлено свыше 430 тыс. уникальных лексических единиц, объем каждого словаря указан в работе [29], а также в приложении 10.

Легко видеть, что словари имеют большой разброс по объему (здесь и далее рассматриваем только русскоязычные словари, если не оговорено особо). Можно условно выделить три класса словарей; большие (св. 1000 ЛЕ) – их в ОПТЕЛ восемь; средние (от 100 до 1000 ЛЕ) – их всего 16, и малые (до 100 ЛЕ), которых в ОПТЕЛ большинство – 29 шт.

Определенная доля ЛЕ является не только принадлежностью естественного языка русской лексики: в ОПТЕЛ представлены два англоязычных словаря, также ряд словарей включает различные обозначения, или коды. Если в исходных словарях проведено разделение на основной термин и его кодовое обозначение, или на дескриптор и аскриптор, то в ОПТЕЛ основной термин, или дескриптор представлен (или выражен) прописными буквами, аскриптор – строчными.

Распределение пересечений лексики ОПТЕЛ по словарям представлено на таблице 2. Всего общая лексика имеется в 14 словарях, но число общей лексики составляет менее 1/4 от числа уникальной лексики.

*Таблица 2*

**Распределение пересечений лексики по всем словарям**

| Число словарей | Число терминов |
|---|---|
| 1 | 338 676 |
| 2 | 55 108 |
| 3 | 16 554 |
| 4 | 13 377 |
| 5 | 6134 |
| 6 | 1344 |
| 7 | 233 |
| 8 | 77 |
| 9 | 23 |
| 10 | 15 |
| 11 | 9 |

[1] Онтология поисковых терминов по лингвистике. – URL: http://db.inion.ru/optel/ (дата обращения: 01.12.2021).

| | |
|---|---|
| 12 | 4 |
| 13 | 2 |
| 14 | 2 |

Парадигматика, представленная в форме словарных статей, имеется не во всех словарях метаданных, в частности ее нет в крупнейших по объему словарях №№ 3, 5, 28, 29, 30, 31, 50, а также в англоязычных словарях №№ 2, 4, которые образуют основной массив лексики ОПТЕЛ (номера и характеристики словарей см. в приложении 10). Всего словарные статьи присутствуют в 39 словарях из 55.

Все словарные статьи независимо от исходной структуры словаря метаданных приведены к классической тезаурусной (дескрипторной) форме, как это описано в работе [30].

Существенное значение имеет распределение лексики ИПЯ с учетом парадигматических отношений (словарные статьи), поскольку именно парадигматика определяет условия и правила включения лексики конкретных ИПЯ в онтологию.

При анализе учитывались родовидовые отношения (выше–ниже), ассоциации и синонимии. Распределение лексики, имеющей парадигматику по словарям, представлено на таблице 3.

*Таблица 3*

**Распределение пересечений лексики, имеющей словарные статьи**

| Общих словарей | Число терминов |
|---|---|
| 1 | 66 966 |
| 2 | 895 |
| 3 | 219 |
| 4 | 56 |
| 5 | 17 |
| 6 | 12 |
| 7 | 5 |
| 8 | 2 |
| 9 | 2 |

Функциональность ОПТЕЛ заключается в возможности исследовать лексику и парадигматику словарей ИПЯ, вошедших в ОПТЕЛ, и что особенно важно, сопоставлять словарные статьи для выработки общей (сводной) словарной статьи, устранения обнаруженных противоречий и исправления ошибок.

## Литература к главе 6

1. Wittenburg P., Broeder D., Sloman B. A Proposal for a Meta Description Standard for Language Resources // EAGLES/ISLE. – 2000. – URL: https://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf (дата обращения: 01.12.2021).

2. Wittenburg P., Gibbon D., Peters W. ISLE Metadata Initiative (IMDI) Metadata Elements for Lexicon Descriptions. Technical Report. – URL: https://www.researchgate.net/publication/240114649_Metadata_Elements_for_Lexicon_Descriptions (дата обращения: 01.04.2022).

3. IMDI Team Vocabulary Taxonomy and Structure, Version 1.1 // MPI Nijmegen. – 2001. – URL: https://www.mpi.nl/isle/documents/draft/IMDI%20Vocabulary%201.1.pdf (дата обращения: 01.04.2022).

4. Arbil. For editing and managing IMDI metadata. Version 2.6. – URL: https://www.mpi.nl/corpus/html/arbil-imdi/index.html (дата обращения: 01.12.2021).

5. Mapping IMDI Session Descriptions with OLAC. – URL: https://www.mpi.nl/ISLE/documents/draft/IMDI%20to%20OLAC%20Mapping%201.04.pdf (дата обращения: 01.04.2022).

6. Arbil 2.4 User Guide. An introduction to editing and managing IMDI metadata. – URL: https://www.mpi.nl/corpus/html/arbil-imdi_ug/index.html (дата обращения: 01.12.2021).

7. OLAC Metadata. – URL: http://www.language-archives.org/OLAC/olacms.html#Attributes (дата обращения: 01.12.2021)

8. OLAC-Usage. – URL: http://www.language-archives.org/NOTE/usage.html (дата обращения: 01.12.2021).

9. Dublin Core XML (DCXML). – URL: https://dcxml.readthedocs.io/en/latest/ (дата обращения: 01.12.2021).

10. OLAC Metadata Usage Guidelines. – URL: http://www.language-archives.org/NOTE/usage.html (дата обращения: 01.12.2021).

11. Documentation and User Manual of the META-SHARE Metadata Model / Elina Desipri, Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis (R.C. Athena/ILSP), Francesca Frontini, Monica Monachini (ILC/CNR), Victoria Arranz, Valérie Mapelli (ELDA), Gil Francopoulo (LIMSI), Thierry Declerck (DFKI) ; Editors: Penny Labropoulou, Elina Desipri. – 2012. – 06.03. – URL: http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf (дата обращения: 01.04.2022).

12. A Metadata Schema for the Description of Language Resources (LRs) / M. Gavrilidou, P. Labropoulou, S. Piperidis, M. Monachini, F. Frontini, G. Francopoulo, V. Arranz, V. Mapelli. – URL: https://aclanthology.org/W11-3311.pdf (дата обращения: 01.12.2021).

13. Technologies for the Multilingual European Information Society. Specification of metadata-based descriptions for language resources and technologies / Penny Labropoulou, Maria Gavrilidou, Elina Desipri, Stelios, Piperidis (R.C. Athena/ILSP), Francesca Frontini, Monica Monachini (ILC/CNR), Victoria Arranz (ELDA), Gil Francopoulo (LIMSI). Final Report. – 2012. – URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf

14. CMDI 1.2 specification Version 1 Date 2016-10-20. – URL: https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf (дата обращения: 01.12.2021).

15. CMDI Best Practices Guide. – URL: https://www.clarin.eu/content/cmdi-best-practices-guide (дата обращения: 01.12.2021).

16. AP3–007-CMDI_and_granularity.pdf. – URL: https://www.clarin.eu/sites/default/files/AP3-007-CMDI_and_granularity.pdf (дата обращения: 01.12.2021).

17. Arbil information, manuals & download. – URL: https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045 (дата обращения: 01.04.2022).

18. Tutorial 1 – Introduction to Fedora. – URL: https://wiki.lyrasis.org/display/FEDORACREATE/Tutorial+1+-+Introduction+to+Fedora (дата обращения: 01.04.2022)

19. IDS Repository Architecture and Ingest Pipelines. – URL: http://repos.ids-mannheim.de/reposdescription.html (дата обращения: 01.12.2021).

20. ISO 24622–1:2015 Language resource management. Component Metadata Infrastructure (CMDI). Part 1. The Component Metadata Model. – URL: https://www.iso.org/ru/standard/37336.html (дата обращения: 01.12.2021).

21. ISO 24622–2:2019(en) Language resource management. Component metadata infrasctructure (CMDI). Part 2. Component metadata specification language. – URL: https://www.iso.org/standard/64579.html (дата обращения: 01.04.2022).

22. The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. The Austin Principles of Data Citation in Linguistics. Version 1.0) / Berez-Kroeker A.L., Andreassen H.N., Gawne L., Holton G., Kung S.S., Pulsifer P., Collister L.B. – 2018. – URL: http://site.uit.no/linguisticsdatacitation/austinprinciples/ (дата обращения: 01.12.2021).

23. ISO 12620:1999. Computer applications in terminology – Data categories. General information. Status : Withdrawn. Publication date : 1999–10. Edition : 1. – URL: https://www.iso.org/standard/2517.html (дата обращения: 01.12.2021).

24. ISO 12620:2019 Management of terminology resources – Data category specifications General information. Status : Withdrawn. Publication date : 1999–10. Edition : 1. – URL: https://www.iso.org/standard/69550.html (дата обращения: 01.12.2021).

25. ГОСТ Р ИСО 12620–2012. Национальный стандарт Российской Федерации. Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов. Terminology and other language and content resources. Specification of data categories and management of a Data Category Registry for language resources. – URL: http://docs.cntd.ru/document/1200104401 (дата обращения: 01.12.2021).

26. Hierarchy of ISOcat data categories. – URL: http://www.sfs.uni-tuebingen.de/nalida/en/docu/isocat-hierarchy.html (дата обращения: 01.12.2021).

27. Cimiano Phillip, McCrae John P., Buitelaar Paul. «Lexicon Model for Ontologies: Community Report, 10 May 2016 Final Community Group Report 10 May 2016». W3 C. – 2019. – 6 December 2019.

28. Антопольский А.Б., Савчук С.О., Тамеев А.А. О разработке онтологии поисковых терминов по лингвистике // Информационные ресурсы России. – 2020. – № 4. – С. 2–7.

29. Антопольский А.Б., Максимов Н.В., Тамеев А.А. Экспериментальная база данных источников для создания онтологии по лингвистике // Информационные ресурсы России. – 2021. – № 3. – С. 24–30. – DOI: 10.46920/0204–3653_2021_03181_24

30. Каленов Н.Е., Белоозеров В.Н. Формирование терминологических словарей по лексике классификационных систем // Научно-техническая информация. Сер. 1: Организация и методика информационной работы. – 2015. – № 3. – С. 60–70.

# ГЛАВА 7. ЛИНГВИСТИЧЕСКАЯ АННОТАЦИЯ

## Общие сведения

Лингвистическая аннотация, также известная как корпусная аннотация, представляет собой маркировку языковых данных в текстовой или устной форме. Лингвистическая аннотация направлена на выявление и маркировку грамматических, фонетических и семантических лингвистических элементов в тексте или аудиозаписи[1]. Лингвистическая аннотация охватывает любые описательные или аналитические обозначения, применяемые к необработанным языковым данным.

Основные данные могут быть динамическими, в виде временных рядов – аудио-, видео- и / либо физиологических записей, или текстовыми. Добавленные обозначения могут включать в себя транскрипции всех видов (от фонетических признаков до дискурсивных структур), метку части речи и смысла, синтаксический анализ, идентификацию «именованной сущности», аннотацию со ссылкой и т.д.

Лингвистическое аннотирование стало одним из важных направлений прикладной лингвистики, поскольку оно стало источником и базой для проведения разнообразных исследований в различных областях лингвистики.

Например, одним из распространенных типов аннотаций является добавление к словам тегов, или меток, указывающих часть речи, к которому принадлежат слова в тексте. Это так называемая частеречная маркировка речи (или POS-маркировка), которая может быть полезна, например для различения слов, имеющих одинаковое написание, но разные значения или произношение. Используется один простой метод представления POS: теги – прикрепление тегов к словам символом подчеркивания. Эти три слова могут быть аннотированы следующим образом:

- present_NN1 (единственное число нарицательное)
- present_VVB (базовая форма лексического глагола)
- present_JJ (общее прилагательное)

Помимо разметки части речи (POS) существуют и другие типы аннотаций, соответствующие различным уровням лингвистического анализа корпуса или текста. Например:

[1] linguistic annotation. – URL: https://exmaralda.org/en/linguistic-annotation-wiki-en/ (дата обращения: 01.12.2021).

- *фонетическая аннотация* – добавление информации о том, как было произнесено слово в устном корпусе;
- *просодическая аннотация* – также в устном корпусе – добавление информации о просодических особенностях, таких как ударение, интонация и паузы;
- *синтаксическая аннотация* – добавление информации о том, как разбирается данное предложение с точки зрения синтаксического анализа в такие единицы, как фразы и предложения;
- *семантическая аннотация* – добавление информации о семантической категории слов или для различения многозначности;
- *прагматическая аннотация* – добавление информации о видах речевого акта (или акта диалога), которые происходят в устном диалоге;
- *дискурсивная аннотация* – добавление информации об анафорических связях в тексте;
- *стилистическая аннотация* – добавление информации о представлении речи и мысли (прямая речь, косвенная речь, свободная косвенная мысль и т.д.);
- *лексическая аннотация* – добавление лексемы к каждой словоформе в тексте.

На самом деле можно изобрести бесчисленные виды аннотаций, которые могут быть полезны для конкретных видов исследований, например изучения различных речевых дисфункций или ошибок при изучении языка.

### Справочник по лингвистической аннотации

Наиболее детальное описание методов и проектов по лингвистическому аннотированию представлено в фундаментальном труде Н. Иде «Справочник по лингвистической аннотации» [1].

В нем указывается, что лингвистическое аннотирование языковых данных первоначально осуществлялось с целью получения информации для разработки и проверки лингвистических теорий при помощи корпусной лингвистики.

За последние три десятилетия прогресс в мощности вычислительной техники как места хранения и развитие надежных методов автоматической аннотации сделали лингвистически аннотированные данные все более доступными в постоянно растущих объемах. В результате эти ресурсы теперь служат не только лингвистическим исследованиям, но и разным задачам NLP (обработки естественного языка).

В последние годы наблюдается заметный подъем лингвистической активности, которая распространилась на широкий спектр языковых явлений. Этот рост сопровождался распространением инструментов аннотирования для поддержки, создания и хранения размеченных данных, средств совместной и распределенной работы с аннотациями.

В справочнике Н. Иде предлагается всесторонний обзор развития и современного состояния лингвистической аннотации языковых ресурсов,

включая методы проектирования схем аннотаций, создания аннотаций, рассмотрения физического формата, инструментов аннотаций, использования аннотаций, оценки и т.д.

Справочник разделен на две части: часть I включает обзорные главы о различных этапах и размышлениях по проектам аннотаций, а часть II состоит из тридцати девяти тематических исследований, описывающих основные проекты аннотирования для широкого круга лингвистических явлений.

Лингвистическая аннотация представляет собой связь описательных или аналитических обозначений с языковыми данными. Исходные данные могут быть текстовыми, взятыми из любого источника или жанра, или – в форме динамических данных (аудио-, видео- и / или физиологических записей).

Сами аннотации могут включать в себя транскрипции всех видов (от фонетических признаков до дискурсивных структур), теги частей речи и смыслов, синтаксический анализ, метки «именованных сущностей», семантические ролевые метки, идентификацию времени и событий, цепочки кореференции, анализ на уровне дискурса и многое другое.

Ресурсы различаются по диапазону типов аннотаций, которые они содержат: некоторые ресурсы содержат только один или два типа, в то время как другие содержат несколько «слоев» аннотаций, или «уровней» лингвистических описаний.

Наиболее важным компонентом проекта лингвистической аннотации является схема аннотации, что определяет метки и связанные с ними объекты, которые должны быть связаны с соответствующей единицей аннотации (например, тип звука, лексемы или слова, фразы, предложения, документа). Метки и единицы измерения должны иметь операционные определения, чтобы люди, глядя на один и тот же фрагмент данных, с большей вероятностью присваивали ему одну и ту же метку.

Схемы, существующие с целью обучения автоматических методов аннотирования, могут идентифицировать признаки (например, орфографические атрибуты, n-граммы или информацию из других аннотаций, таких как часть речи, субъект / объект, семантическая роль и т.д.), которые коррелируют с метками аннотаций.

Проект аннотации может использовать существующую схему или может потребовать разработки новой схемы для явлений, которые ранее не рассматривались. В последнем случае проект может потратить больше времени на разработку схемы, чем на аннотацию, независимо от того, разработан ли он априори или разработан итеративно с циклами аннотации, оценки и пересмотра схемы. Нахождение баланса между достаточно богатым описанием рассматриваемого языкового явления и возможностями человека и / или машины, чтобы надежно и последовательно идентифицировать его, – возможно, самая важная часть проекта аннотации.

Наконец, современные ручные или полуавтоматические методы аннотирования обычно опираются на инструмент аннотации с интерфейсом, который позволяет идентифицировать промежутки символов и / или связи между такими промежутками, а также средства для связывания метки или меток с идентифицированными промежутками и / или связями.

Проект может сопровождаться инструментами для измерения согласованности аннотаторов. Чтобы измерить согласованность, определяют порог ожидаемой производительности с помощью автоматических инструментов аннотации и / или определяют, подходит ли конкретная шкала для измерения рассматриваемого явления и т.д.

## Семинар по лингвистическим аннотациям (LAW)

Фундаментальные компоненты лингвистического аннотирования претерпели значительную эволюцию за последние пять десятилетий. Эту эволюцию можно наблюдать, рассмотрев содержание Семинара по лингвистическим аннотациям (LAW), который проводится ежегодно с 2007 года специальной группой по интересам «Аннотации» (SIGANN)[1] Ассоциации компьютерной лингвистики (ACL). На каждом заседании представляется 20–30 докладов и сообщений по проблемам лингвистического аннотирования.

SIGANN включает исследователей, занимающихся всеми аспектами лингвистического аннотирования ЛИР. Ее деятельность включает в себя:

- обмен и распространение результатов исследований в области аннотирования, манипулирования и эксплуатации аннотированных ЛИР с учетом различных прикладных и теоретических исследований в области языковых технологий и исследований;
- гармонизацию и интероперабельность лингвистических аннотаций с использованием все большего числа инструментов и методов, поддерживающих создание, обработку и использование аннотированных ресурсов;
- работу по достижению консенсуса по всем вопросам, имеющим решающее значение для развития области аннотации ЛИР;
- спонсорство ежегодного семинара по лингвистическим аннотациям.

Материалы семинара LAW доступны по адресу [2]. Кратко опишем тематику нескольких предыдущих семинаров LAW.

### *2007 г. 1-й семинар по лингвистической аннотации (LAW)*

Лингвистически аннотированные корпусы играют важную роль в синтаксическом анализе, извлечении информации, ответе на вопросы, машинном переводе и многих других областях компьютерной лингвистики, и обеспечивают эмпирическую базу для теоретических лингвистических исследований. Это привело к распространению систем аннотаций, методик, форматов и схем. Признание необходимости согласования практики и схем аннотаций приобретает все более важное значение, о чем свидетельствуют многочисленные семинары, посвященные различным аспектам лингвистических аннотаций за последние несколько лет.

LAW рассматривает все аспекты лингвистической аннотации на одном форуме, объединяя две существующие серии семинаров: NLPXML

[1] ACL Special Interest Group for Annotation. – URL: https://www.cs.vassar.edu/sigann/ (дата обращения: 01.12.2021).

(Обработка естественного языка и XML) и FLAC (Границы в лингвистически аннотированных корпусах).

***2009 г. LAW № 3***

Тематика докладов охватывала весь спектр лингвистических фактов и соответствующих им рамок аннотаций, от словесных сетей до банков деревьев, от эмоций до убеждений и от речи до дискурса. В этих работах рассматривается ряд уровней аннотаций, с точки зрения как макроперспективы, т.е. инфраструктуры для международного сотрудничества и интероперабельности, так и микроперспективы, т.е. создания инструментов для борьбы с несоответствиями между аннотациями. Это богатство свидетельствует о растущей зрелости области лингвистического аннотирования, которое представлено в специальном выпуске журнала Language Resources and Evaluation, включающем избранные работы семинара.

***2012 г. LAW № 6***

Тематика семинара – содействие использованию и совместной разработке открытых общих ресурсов, а также выявлению и продвижению лучших практик взаимодействия аннотаций. Критерии оценки включали в себя следующее:

- инновационное использование лингвистической информации из различных слоев аннотаций;
- совместимость по крайней мере с одной-другой схемой аннотаций или форматом;
- качество аннотируемого ресурса с точки зрения проектирования схемы, документации, инструментальной поддержки и т.д.;
- открытая доступность разработанных ресурсов для использования сообществом;
- удобство использования и возможность повторного использования схемы аннотаций или аннотированного ресурса;
- выдающийся вклад в разработку лучших практик аннотации.

Победителем первого состязания по аннотации LAW стало «Контрастивное исследование синтаксико-семантических зависимостей», которое изучает взаимодействие между широким спектром общих схем аннотаций синтаксико-семантических зависимостей в рамках *проекта LinGO Redwoods Treebank*. Сильными сторонами проекта были признаны его ориентация на интероперабельность, которая поможет сообществу разработать более общие и глубокие способы взаимодействия с использованием различных лингвистических инструментов.

***2013 г. LAW № 7***

Наибольшую пользу приносят обсуждения лучших практик синтаксического аннотирования нестандартных языков. Поэтому LAW был особенно заинтересован в сравнении и совместимости различных моделей и методов, используемых для аннотации дискурса, сосредоточив внимание на любой из следующих целей:

- создание новых идей в области дискурса (путем сопоставления двух или более точек зрения, отраженных различными схемами аннотации или методами аннотации);
- содействие взаимодействию прагматических и семантических феноменов в дискурсе, от функциональных категорий (например методы, результаты, гипотезы и т.д.) до традиционных дискурсивных отношенияй (связки, анафоры, метонимии и т.д.);
- соединение синтаксического, семантического и прагматического слоев аннотации;
- разработка структуры, стандартов репрезентации, инструментов и методов, которые позволят интегрировать текущие и будущие схемы аннотаций, связанные с дискурсом, чтобы они могли сосуществовать.

***2015 г. LAW № 9***

Специальная тема: *Синтаксическая аннотация неканонического языка*. На семинаре подробно обсуждаются различные подходы к понятию «неканонический» и влияние различных подходов на методы лингвистического аннотирования

Например, «неканоническое» относится к структурам, которые не могут быть описаны или порождены данной лингвистической структурой.

Структура может быть неканонической, потому что это безграмотно, или она может быть неканонической, потому что на данной платформе невозможно эту структуру проанализировать.

Существует серьезная проблема – что делать с языками, где есть более одного стандарта (например, для английского языка: Британский / Североамериканский / Австралийский), или там, где никакого стандарта вообще не существует? Последнее особенно актуально для устных языков, не имеющих письменности**.**

Такое положение дел подводит нас к вопросу о достоверности аннотаций. Было много обсуждений того, стоит ли использовать экспертов-аннотаторов, учитывая временные требования и высокие затраты, или же можно добиться аналогичных результатов с неподготовленными аннотаторами. Кроме того, и это было в центре внимания LAW прошлого года: «Хорошее, плохое и совершенное: насколько хорошей должна быть аннотация?»

Ответ на этот вопрос тесно связан со следующим: какой тип аннотаторов нужен? Достаточно ли надежен краудсорсинг и можно ли его эффективно использовать?

***2017 г. LAW № 11***

В статьях семинара этого года изучаются методы аннотации в областях эмоций и отношений; беседы и структура дискурса; события и причинность; семантические роли и перевод.

Специальная тема: *Оценка качества аннотаций.*

Эта специальная тема рассматривает текущую практику оценки лингвистических аннотаций, ее успехи и неудачи, задавая следующие вопросы:

1. Как мы, как сообщество, измеряем межаннотаторское соглашение на сегодняшний день, и есть ли более надежные способы его измерения?

2. Как мы можем оценить качество аннотаций существующих ресурсов, и что можно сделать для документирования аннотированных данных, чтобы помочь другим оценить их надежность?

3. Как измеряется согласие в различных (новых или существующих) проектах аннотаций, и что говорят нам различные оценки в каждом конкретном случае?

4. Хорошие пороги приемлемости для различных задач аннотации и метрик и / или способы их определения.

5. Ранее предложенные, но не широко используемые меры по обеспечению качества согласования или аннотации.

6. Предложения по количественным или качественным методам измерения качества согласования или аннотации.

7. Предложения по документированию опубликованных ресурсов для поддержки их оценки, средства и методы достижения общественной оценки лингвистически аннотированных ресурсов и т.д.

***LAW-MWECxG-2018 № 12***

Совместный семинар LAW с группой по многословным выражениям и конструкциям состоялся в августе 2018 года совместно с XXVII Международной конференцией по компьютерной лингвистике (COLING 2018)[3]. Семинар объединил три расходящиеся, но пересекающиеся исследовательские сообщества, изучающие лингвистические аннотации, многословные выражения и грамматические конструкции. Лингвистическая аннотация корпусов естественных языков является основой контролируемых методов статистической обработки естественного языка. Это также важно для оценки как основанных на правилах, так и контролируемых систем, и может помочь формализовать и изучить языковые явления.

Конкретные темы, представленные на семинаре:

- процедуры аннотирования, будь то ручные или автоматические, включая машинное обучение и методы, основанные на знаниях;
- ведение и интерактивное исследование структур аннотаций и аннотированных данных;
- качественная и количественная оценка аннотаций;
- лингвистические размышления, форматы представления и инструменты исследования для слияния аннотации различных явлений;
- стандарты, лучшие практики, документация, интероперабельность и сравнение схем аннотаций;
- разработка, оценка и инновационное использование программных фреймворков аннотаций.

***2019 г. LAW № 13***

Специальная тема: *Маркировка качества информации в дискурсе.*

Эта специальная тема рассматривает маркировку качества информации в дискурсе, т.е. аннотации, которые маркируют качество информации в дискурсе, как говорящий / пишущий выражает оценки. Эти оценки могут быть эксплицитными и / или имплицитными в дискурсе и могут отражать позиции,

убеждения, мнения, оценки и / или оценки письменных или устных предложений, например то как политик показывает в дискурсе степень правдивости одного из своих предвыборных обещаний или как репортер показывает свою степень веры в то, что заявил политик.

***2020 г. LAW № 14***

В предисловии к трудам последнего семинара, проведенного в 2020 году, говорится:

«LAW идет уже четырнадцатый год. Первый семинар состоялся в 2007 г. в Праге. С тех пор LAW проводится ежегодно, последовательно привлекая все большее число участников, которые еше и привлекаются к обсуждению по вопросам условий подачи документов / постеров. Этот факт свидетельствует о том, что общая направленность LAW по-прежнему остается важной областью интереса в сфере лингвистических технологий. Значительная часть обсуждаемых проблем касается контролируемого обучения на основе наборов данных золотого стандарта»[4].

Специальная тема 2020 года: *Разрушение стереотипов*, т.е. продвижение новых или менее изученных типов аннотаций и данных. Это – материалы по менее распространенным аннотациям (например метафора, смысловая анафора, жест, юмор); аннотации в недостаточно изученных настройках или типах данных (новые типы текста или новые методы сбора данных); и аннотации для других языков, кроме английского, особенно за пределами тегов / древовидных банков.

## Стандартизация лингвистического аннотирования

Методы лингвистического аннотирования получили важное развитие и в рамках разработки стандартов ISO. Подробное описание особенностей разработки стандарта LAF ISO 24612:2012 [5] представлено в работе Н. Иде и К. Зудермана [6]. Ниже приводятся основные положения этой работы.

Мотивацией для разработки LAF было формирование архитектуры аннотированных языковых ресурсов, которая удовлетворяла бы потребности всех видов деятельности по аннотированию в области компьютерной лингвистики и обеспечивала полную совместимость между форматами аннотаций.

На момент первоначального развития LAF большинство форматов аннотаций разрабатывались без учета какой-либо базовой модели данных, и выбор часто в первую очередь определялся потребностями конкретного программного обеспечения для обработки.

С самого начала LAF определила набор фундаментальных принципов, которые будут лежать в основе развития архитектуры. Одним из наиболее важных является четкое разделение структуры аннотаций, т.е. физического формата аннотаций, и содержания аннотаций, которое включает в себя категории или метки, используемые в схеме аннотаций для описания языковых явлений. Интересно, что ранее это различие не проводилось в явном виде, и фактически смешение вопросов структуры и содержания при разработке многих ранее существовавших схем аннотаций часто было источником несо-

ответствий и упущений. Другим принципом было требование, чтобы вся информация аннотации была представлена явно. Многие схемы, включая широко используемые, такие как формат *Penn Treebank*, опирались на неявное знание о толковании различных категорий и отношений. Это само по себе было серьезным препятствием для интероперабельности, поскольку обработка аннотаций часто требовала использования специализированного программного обеспечения, в которое были встроены эти знания.

Исходя из этих фундаментальных принципов, архитектура LAF была разработана с двумя различными частями:

(1) структура данных для представления отношений между аннотациями вместе с механизмом ассоциирования лингвистических категорий с соответствующими частями этой структуры данных; и

(2) средство определения лингвистических категорий с точки зрения теории или конкретных соглашений об именах.

Часть (2) обеспечивает семантическую когерентность; с самого начала предполагалось, что это будет обеспечиваться реестром лингвистических категорий и признаков.

Этот план в конечном итоге привел к созданию ISOCat (см. гл. 6), который фактически стал самостоятельной работой.

Модель данных LAF должна была охватывать общие принципы и практику как существующих, так и предполагаемых лингвистических аннотаций, включая аннотации всех типов носителей, таких как текст, аудиовидеоизображение и т.д., чтобы в конечном итоге обеспечить общие механизмы для их обработки. Кроме того, модель должна была позволять варьировать схемы аннотаций, и в то же время позволять сравнивать и оценивать, объединять различные аннотации и разрабатывать общие инструменты для создания и использования аннотированных данных.

Для достижения интероперабельности между форматами при сохранении максимальной гибкости LAF предписывает, чтобы соответствующие форматы аннотаций, как уже существующие, так и недавно разработанные, были или могли быть визуализированы с помощью отображения – изоморфных моделей данных LAF. Таким образом, модель служит в качестве ссылки, или «стержня», в который и из которого аннотации могут быть сопоставлены для обмена, или в который различные аннотации могут быть сопоставлены для сравнения или слияния. Это доказывает, что модель может вместить все типы лингвистических аннотаций.

Сопоставление между пользовательскими форматами и абстрактной моделью данных LAF осуществляется с помощью XML-сериализации модели данных, называемой Graph Annotation Format (GrAF).

LAF и GrAF были разработаны, чтобы обеспечить базовую основу для лингвистических аннотаций. В принципе GrAF не дает никаких указаний ни для именования лингвистических категорий, ни для организации, или связи конкретных категорий. Этот принцип позволил нам определить и сосредоточиться на основных механизмах, необходимых для учета структурных и референциальных свойств аннотаций. В результате получается универсальный механизм, который может быть использован в качестве стержня для обмена и

объединения аннотаций, что доказало свою эффективность для удовлетворения многих требований к структуре лингвистических аннотаций.

Полная стандартизация лингвистических аннотаций, однако, требует гораздо большего, чем некий каркас, который предоставляет GrAF. В дополнение к стандартизации лингвистических семантических категорий, которая сейчас предпринимается ISOCat, необходима еще и разработка перечня и, по меньшей мере, грубая онтология лингвистических объектов и функций.

Даже простой набор стандартных лингвистических объектов еще не получил широкого распространения, но очень важно создать некоторую основу для общения между веб-сервисами и другие инструменты обработки языка для продвижения в этой области.

С этой целью в рамках ISO TK 37 SC4 WG1 был предложен новый рабочий элемент для разработки по крайней мере базового набора лингвистических дескрипторов объектов / признаков с опорой на существующие предложения, разработанные или разрабатываемые в ряде недавних проектов (например European Language Grid[1], AusNC – Австралийский национальный корпус[2]) и с учетом передового опыта в этой области. Учитывая, что в настоящее время существует широкая база рекомендаций, накоплен большой опыт, а также имеется растущая практическая потребность, рабочая группа должна быть в состоянии относительно быстро разработать хотя бы базовую схему, которая может служить бурному развитию модульных веб-сервисов для NLP.

В результате большинство вновь разработанных схем аннотаций и форматов основаны на абстрактной модели данных LAF и, таким образом, сопоставимы с GrAF. Это относится ко всей серии стандартов ИСО ISO 24615 на лингвистические аннотации.

Остановимся только на стандарте ISO 24617–2, ч. 2., посвященном семантической аннотации на диалоговые акты [7], который обладает рядом особенностей. Представление этого стандарта содержится в презентации [8].

Схемы семантических аннотаций предназначены для аннотирования семантической информациеи первичных данных, таких как тексты, речевые транскрипты, изображения или записи мультимодального или невербального коммуникативного поведения.

Основными составляющими коммуникативного акта являются его коммуникативная функция и смысловое содержание. Семантическое содержание определяет объекты, пропозиции, события и т.д., о которых идет речь в диалоге; коммуникативная функция определяет, каким образом адресат должен использовать семантическое содержание для актуализации своего информационного состояния. Аннотация диалогического акта – это деятельность по разметке отрезков диалога с информацией о содержании диалогических актов, которая обычно направлена на разметку коммуникативных функций высказываний.

[1] European Language Grid. – URL: https://www.european-language-grid.eu/ (дата обращения: 01.12.2021).

[2] AusNC – Australian National Corpus. – URL: http://www.ausnc.org.au/ (дата обращения: 01.04.2022).

Самые важные изменения по сравнению с предыдущей версией этого стандарта касаются следующих моментов:

- дополнение теоретических отношений между языковыми единицами;
- разработка полномасштабной композиционной семантики для языка разметки диалоговых актов DiAML (что приводит, как побочный эффект, к различной трактовке отношений функциональной зависимости между диалоговыми актами и обратными зависимостями);
- систематическое применение понятия «идеальный конкретный синтаксис» к основанному на дизайне представлению на основе FXML (Format for DiAML);
- разработка и внедрение метода пошаговой интерпретации диалоговых актов, автоматической аннотации диалогов.

При разработке методов лингвистического аннотирования постоянно шло соревнование ручных и автоматических методов. Сравнительные преимущества и недостатки обоих подходов очевидны, однако для конкретных применений возможны разные ситуации; поэтому сравнению этих методов посвящено достаточно много работ. В данном случае сошлемся на работу [9], где проведено систематическое сравнение этих методов для различных уровней лингвистического аннотирования и для различного использования аннотированных ЛИР.

Важным направлением лингвистического аннотирования является его применение для машинного обучения. Обзор этих подходов можно найти в электронной публикации [10].

Конечно, для практического применения методов лингвистического аннотирования важнейшее значение имеют программные инструменты, которые используются для разных процессов в ходе аннотирования. Существует несколько каталогов этих инструментов, в том числе М. Вайссера [11], который был использован при подготовке общего каталога ЛИР типа инструментов, который вынесен в приложение 7 «Инструменты лингвистических технологий».

## Литература к главе 7

1. Ide N. Introduction : The Handbook of Linguistic Annotation // Handbook of Linguistic Annotation / Ide N., Pustejovsky J. (eds). – Dordrecht : Springer, 2017. – URL: https://doi.org/10.1007/978-94-024-0881-2_1 (дата обращения: 01.12.2021).

2. Special Interest Group for Annotation (SIGANN). Proceedings of the Linguistic Annotation Workshop. – 2007–2020. – URL: https://aclanthology.org/sigs/sigann/ (дата обращения: 01.12.2021).

3. COLING 2018. – Santa Fe ; New Mexico, USA, 2018. – August 20–26. – URL: http://coling2018.org/index.html%3Fp=491.html (дата обращения: 01.12.2021).

4. Proceedings of the 14 th Linguistic Annotation Workshop. – URL: https://www.aclweb.org/anthology/volumes/2020.law-1/ (дата обращения: 01.12.2021).

5. ISO 24612:2012 Language resource management – Linguistic annotation framework (LAF). – URL: https://www.iso.org/standard/37326.html (дата обращения: 01.12.2021).

6. Suderman I., Suderman N., Suderman K. «The Linguistic Annotation Framework: a standard for annotation interchange and merging» // Language Resources and Evaluation 48. – 2014. – P. 395–418.

7. ISO 24617–2:2020 Language resource management. Semantic annotation framework (SemAF). Part 2. Dialogue acts. – URL: https://www.iso.org/standard/76443.html (дата обращения: 01.12.2021).

8. Towards an ISO standard for dialogue act annotation / Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, Claudia Soria, David Traum, Kiyong Lee, Laurent Romary. – URL: https://publications.idiap.ch/downloads/papers/2010/Bunt_LREC2010_2010.pdf (дата обращения 01.04.2022).

9. Communication Methods and Measures. – DOI: 10.1080/19312458.2020.1846695. – URL: https://www.tandfonline.com/doi/pdf/10.1080/19312458.2020.1846695 (дата обращения: 01.12.2021).

10. Natural Language. Annotation for Machine Learning by James Pustejovsky, Amber Stubbs. O' Reilly. – URL: https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/ (дата обращения 01.04.2022).

11. Weisser M. Annotation & Text-Processing Tools. – URL: http://martinweisser.org/corpora_site/annotation_tools.html (дата обращения: 01.12.2021).

# ГЛАВА 8. ЯЗЫКОВАЯ ДОКУМЕНТАЦИЯ

## Введение

Языковая документация (также: документальная лингвистика) – это подраздел лингвистики, целью которого является описание грамматик и использования человеческих языков. Его цель – предоставить исчерпывающий отчет о языковых практиках, характерных для данного речевого сообщества. Документация по языку направлена на создание как можно более подробных записей речевого сообщества, как для потомков, так и для возрождения языка. Эта запись может быть публичной или частной, в зависимости от потребностей сообщества и цели документации. На практике языковая документация может варьироваться от индивидуальных лингвистических антропологических полевых исследований до создания обширных онлайн-архивов, содержащих десятки разных языков, таких как FirstVoices[1] или OLAC[2].

Сфера языковой документации в современном контексте включает в себя сложный и постоянно развивающийся набор инструментов и методов изучения, средств использования. Выявление и продвижение передовых практик можно рассматривать как подобласть языковой документации.

*Принципы и рабочие процессы.* Исследователи языковой документации часто проводят лингвистические полевые исследования для сбора данных, на которых основана их работа, записывая аудиовизуальные файлы, которые документируют использование языка в традиционном контексте. Поскольку условия, в которых часто проводятся полевые лингвистические исследования, могут быть сложными с точки зрения логистики, не все типы записывающих устройств пригодны или идеальны, то часто приходится искать компромисс между качеством, стоимостью и удобством использования. Также важно представить себе полный рабочий процесс и предполагаемые результаты; например, если созданы видеофайлы, может потребоваться некоторая обработка, чтобы подвергнуть аудиокомпонент обработке различными способами с помощью разных пакетов программного обеспечения.

*Форматы данных.* Соблюдение стандартов форматов имеет решающее значение для взаимодействия программных инструментов. Многие индивиду-

[1] FirstVoices – suite of web-based tools and services. – URL: https://www.firstvoices.com/ (дата обращения: 01.12.2021).

[2] Language_documentation. – URL: https://ru.qaz.wiki/wiki/Language_documentation (дата обращения: 01.12.2021).

альные архивы или репозитории данных имеют свои собственные стандарты и требования к данным, размещаемым на их серверах. Знание этих требований должно определять стратегию сбора данных и используемые инструменты, а также должно быть частью плана управления данными, разработанного до начала исследования.

В настоящей главе будет предложен обзор основных проектов, посвященных языковой документации и созданию архивов ЛИР для исчезающих языков. Российские проекты описаны отдельно.

## Международные проекты языковой документации

***Сеть архивов цифровых языков и музыки, находящихся под угрозой исчезновения DELAMAN***[1]

Сеть DELAMAN была создана в 2003 году в качестве международного зонтичного органа для архивов и других инициатив с целью документирования и архивирования языков и культур, находящихся под угрозой исчезновения во всем мире. Цель – стимулировать обмен опытом полевых работников и архивистов и организовать информационный центр для обмена информацией. DELAMAN задумана как открытая организация, в которой может участвовать любая организация, активно занимающаяся архивированием и сохранением исчезающих языков и музыки.

DELAMAN разработала требования по сохранению материалов цифровой языковой документации [1]. Эти требования могут рассматриваться как профессиональный стандарт языковой документации:

- материалы сдаются на хранение в цифровой репозиторий с институциональным обязательством долгосрочного сохранения и доступа (например архив DELAMAN или институциональный репозиторий). Депонирование осуществляется регулярно и по мере создания материалов;
- материалы представлены в цифровом формате, рекомендованном международными архивными стандартами (например IASA). Форматы не являются проприетарными, должны быть хорошо документированы и / или иметь открытый исходный код;
- аудиофайлы, как минимум, 48 khz/16 bit, в идеале 96 khz/24 bit BWF. При оцифровке должны использоваться высококачественные аналого-цифровые преобразователи;
- текстовые файлы: txt,. xml, rtf, pdf;
- видеофайлы в несжатом формате (в идеале JPEG2000);
- материалы должны быть в форматах, которые доступны для загрузки (например, сжатые в формате mp3, mp4);
- материалы описываются стандартизированными метаданными (например OLAC, IMDI, Dublin Core, MODS);
- описание депонированной коллекции включено в коллекцию;

[1] The Digital Endangered Languages and Musics Archives Network (DELAMAN). – URL: https://www.delaman.org/about/ (дата обращения: 01.12.2021).

- значительная часть коллекции находится в открытом доступе, либо указан четкий порядок запроса доступа. Если публичный доступ невозможен, в описани коллекции должно быть объяснение.

***Электронная метаструктура для данных по исчезающим языкам E-MELD***

Наиболее известным проектом, посвященным документированию исчезающих языков, является проект *E-MELD*[1]. Сайт проекта предлагает способы сбора, преобразования и хранения языковых данных в надежных цифровых форматах. Обосновывается важность лучших практик в области цифровой языковой документации. Приводятся примеры оцифровки данных с использованием технологий, представленных на практических занятиях. Размещена база данных ссылок на дополнительные ресурсы. На сайте имеются специальные разделы:

- *Рабочая комната* предлагает пользователям использовать онлайн-инструменты, такие как CharWrite[2] или FIELD[3], для работы со своими данными;
- *Инструментальная комната* перечисляет различные загружаемые инструменты для использования полевыми лингвистами;
- *Спросите экспертов* – это форум, на котором пользователи могут задать вопросы экспертам о создании и сохранении цифровой языковой документации.

Проект E-MELD в сотрудничестве с другими крупными проектами, посвященными развитию цифровой инфраструктуры в лингвистике, например, OLAC и DOBES, поддерживает рекомендации по передовой практике. Эти рекомендации оформлены в виде отдельного раздела портала E-MELD «Школа лучших практик» [2]. Лучшие практики призваны сделать цифровую языковую документацию оптимальной, долговечной, доступной и повторно используемой другими лингвистами и носителями языка. Рекомендации лучших практик охватывают все аспекты оцифровки и архивирования языковой документации, включая способы ее записи, аннотирования, каталогизации, хранения и отображения, чтобы обеспечить соблюдение прав интеллектуальной собственности заинтересованных сторон. В рекомендациях использованы результаты других проектов по оцифровке языка и языковой инженерии. Основные принципы:

*Долгосрочная консервация*. Рекомендации лучшей практики были разработаны в первую очередь для того, чтобы ценная языковая документация была доступна будущим поколениям. Сайт предоставляет информацию об

[1] Electronic Metastructure for Endangered Languages Data (E-MELD). – URL: http://emeld.org (дата обращения: 01.12.2021).

[2] CharWrite©: A Unicode Input Tool for the Web. – URL: http://emeld.org/tools/charwrite.cfm (дата обращения: 01.12.2021).

[3] Field Input Environment for Linguistic Data. – URL: http://emeld.org/tools/fieldinput.cfm (дата обращения: 01.12.2021).

архивных форматах, программном обеспечении и открытых стандартах, которые обеспечат оцифрованным данным сохранность.

*Поиск и доступность*. Подготовка метаданных для ресурсов и экспортирование их в общую поисковую систему обеспечат осведомленность лингвистического сообщества об актуальных исследованиях. Кроме того, соотнесение лингвистической разметки данных с общей онтологией лингвистического описания (GOLD) будет способствовать их долгосрочной устойчивости, и в конечном итоге сделает их доступными для детального межъязыкового поиска.

*Гибкое представление*. Передовая практика рекомендует различать формат архива и форматы представления данных. Например, с помощью разметки XML и таблиц стилей XSL один текстовый файл может быть представлен множеством различных способов. Сайт E-MELD предлагает примеры различных форматов представления метаданных и лексиконов, а также информацию о таблицах стилей.

Участники проекта E-MELD Берд и Саймонс сформулировали семь областей, в которых цифровые ЛИР можно сделать более полезными для лингвистики и речевых сообществ. Приводим эти области и предлагаемые решения [3].

*Терминология*. Одной из наиболее острых проблем является непоследовательное использование терминологии во всех ресурсах. При этом очень сложно сравнивать ресурсы. Решение: терминология должна быть связана с общей онтологией GOLD[1], с помощью которой различная терминология может быть интерпретирована с помощью компьютера.

*Формат.* Ресурсы часто не поддаются интерпретации исследователей. В большинстве случаев это происходит из-за использования нестандартных шрифтов или из-за того, что файлы находятся в проприетарных форматах. Решение: все символы должны быть закодированы с помощью Юникода. Это гарантирует, что символьные коды всегда представляют один и тот же символ, независимо от того, на какой машине они отображаются. Следует использовать открытые или, по крайней мере, опубликованные форматы файлов, чтобы они были доступны на многих платформах программного обеспечения. XML-файлы, документированные схемой или DTD, являются хорошим выбором.

*Открытие данных*. Поиск данных, пожалуй, – одна из самых сложных задач для лингвиста. Универсальные поисковые системы, такие как Google и Yahoo, работают хорошо, но известно, сколько нерелевантных результатов в этих поисках. Решение: ресурсы должны быть отражены в лингвистической поисковой системе, например репозитории OLAC.

*Доступ*. Многие лингвистические источники данных недоступны через Интернет. Действительно, большинство лингвистических данных существует только в виде лент и рукописных карточек. Другие данные, несмотря на доступность в Интернете, – с ограниченным доступом. Решение: материалы в не электронной форме должны быть размещены в Интернете. Материалы в электронном виде, доступные ограниченному кругу пользователей, также

[1] GOLD Community. – URL: http://linguistics-ontology.org/ (дата обращения: 01.12.2021).

могут быть доступны через систему дифференцированного доступа, которую E-MELD разрабатывает совместно с IMDI и AILLA.

*Цитирование*. Цитирование интернет-ресурсов – одна из главных проблем, связанных с ними. URL-адреса могут стать недоступными, или ресурсы могут переместиться, не оставив никаких записей о том, куда они переместились. Решение: архивная копия всего материала должна быть помещена в стабильный онлайновый лингвистический архив, а метаданные должны быть включены в лингвистическую поисковую систему.

*Сохранение.* Сохранение является проблемой как для цифровых архивов, где форматы файлов со временем устаревают, так и для физических архивов, где носители могут со временем ухудшаться, теряться или повреждаться, или находиться в настолько архаичном формате, что никакое современное оборудование не будет их поддерживать. Решение: если важна долгосрочная сохранность, файлы лучше всего помещать в цифровые архивы, которые обязуются переносить форматы по мере их изменения. Если это невозможно, текстовый материал должен быть заархивирован в формате XML. Материал также должен храниться в нескольких копиях, в более чем одном физическом месте.

*Права.* Создатели ресурсов, исследователи и речевые сообщества, предоставляющие первичные данные, имеют разные приоритеты по доступу к языковым ресурсам. Решение: условия использования должны быть хорошо документированы и соблюдаться с помощью шифрования или лицензирования. Однако важно ограничить продолжительность ограничений доступа: ресурс, доступ к которому постоянно ограничен одним пользователем, не имеет долгосрочной ценности.

## Документирование языков, находящихся под угрозой исчезновения DOBES[1]

Еще один известный международный проект, посвященный сохранности исчезающих языков, – это проект DOBES. Портал проекта предназначен для облегчения доступа исследователей, заинтересованных в работе с данными в архиве DOBES.

Архив содержит материалы по 68 языкам, находящимся под угрозой исчезновения, включая видео с переводами / аннотациями. В архиве также содержатся описания грамматик и материалы исследований. Материалы находятся в свободном доступе, просмотр архива возможен с помощью браузера метаданных IMDI.

Существует два основных типа поиска: поиск метаданных и поиск аннотаций. Также можно просмотреть архив.

Для студентов и других исследователей предлагаются руководства для различных сценариев использования архива.

При работе с данными можно использовать набор программных инструментов, а именно программное обеспечение TLA (The Language Archive)[1],

[1] Dobes Documentation De Langues En Danger. – URL: https://dobes.mpi.nl/research/ (дата обращения: 01.12.2021).

специально разработанное с учетом потребностей команд языковой документации и реализующее развивающиеся в ней стандарты.

- *Elan* – это мощный инструмент для выравнивания по времени аннотации видео- или аудиоданных. Большая часть данных в архиве DoBeS была аннотирована с помощью этой программы.
- *Arbil* – новый инструмент для создания хорошо структурированного каталога метаданных, пригодного для архивирования. Он позволяет вводить метаданные отдельных сеансов и размещать эти сеансы в дереве архивов.
- *Lamus* – инструмент для загрузки данных и метаданных в архив DoBes, а также для управления существующими коллекциями.
- *Lexus* – веб-лексикон, который предоставляет гибкую структуру для поддержания лексической структуры и содержания. Это – первая реализация модели LMF. Среди его функционала – возможность создавать структуры лексики, манипулировать содержанием и использовать типизированные отношения.

Все TLA-инструменты поставляются с руководствами, которые можно найти на соответствующих веб-сайтах. В нашей монографии все адреса инструментов размещены в приложении 7. Имеются руководства для конкретных сценариев использования.

На данных DOBES проводятся разнообразные исследования, например:

- кросс-лингвистические паттерны в кодировании событий с тремя участниками;
- демонстративы с экзофорической отсылкой. Функциональное исследование, основанное на дискурсивных данных из пяти языков;
- дискурс и просодия через границы языковой семьи: два корпусных тематических исследования контактно-индуцированной синтаксической и просодической конвергенции в кодировании информационной структуры;
- проект референтности: исследовательский проект по корпусной типологии референтных стратегий на 12 различных языках;
- относительные частоты существительных, местоимений и глаголов в разговорных корпусах семи языков, представляющих широкий спектр ареального и типологического разнообразия**.**

***Сохранение и ревитализация языков FirstVoices***[2]

Интернет-проект, направленный на поддержку преподавания и архивирования языков и культуры коренных народов, находится в ведении Совета по культуре первых народов Британской Колумбии (Канада). FirstVoices был запущен в 2003 году для помощи в сохранении оставшихся 34 языков коренных народов Британской Колумбии. Он предоставляет языковым группам коренных народов пространство для архивации своих языков путем записи и загрузки слов, фраз, песен и рассказов в безопасную централизованную базу данных. FirstVoices размещает 47 языковых архивов (36 государственных и

[1] The Language Archive. – URL: https://archive.mpi.nl/forums/ (дата обращения: 01.12.2021).
[2] FirstVoices. – URL: https://www.firstvoices.com/ (дата обращения: 01.12.2021).

11 частных) в Британской Колумбии, а также поддерживает 70 общин коренных народов в Канаде, США и Австралии. Контент полностью контролируется и управляется администраторами языков сообщества. FirstVoices предоставляет следующие инструменты, чтобы каждый архив можно было настроить для языков, которые он обслуживает:

- *Алфавит* предоставляет набор письменных символов для языка с образцами звуковых файлов для каждого символа;
- *Словарь* предоставляет список слов с переводами, определениями, звуками, изображениями и видео;
- *Разговорник* содержит повседневный разговорный язык с соответствующими текстовыми, звуковыми изображениями и видеофайлами для поддержки изучения языка.

Типичные этапы включают запись, ведение метаданных, расшифровку (часто с использованием Международного фонетического алфавита) и / или «практическую орфографию», составленную для этого языка, аннотацию и анализ, перевод на язык более широкого общения, архивирование и распространение. Важным является создание хороших записей в процессе описания языка.

Новые технологии позволяют делать более качественные записи с лучшими описаниями, которые можно хранить в цифровых архивах, таких как AILLA или PARADISEC.

***Проект «Языки под угрозой исчезновения» ELP***[1]

При помощи этого проекта компания Google предоставляет свои технологии организациям и людям, борющимся с исчезновением языков и занимающимся их документированием, сохранением и преподаванием. На сайте проекта пользователи могут не только найти самый полный и обновленный архив данных и образцов языков, находящихся под угрозой, но также активно поучаствовать в их сохранении: разместить информацию и образцы языков в форме текста, аудио- или видеофайлов. Более того, пользователи могут поделиться практическим опытом и примерами в разделе «Обмен знаниями» и участвовать в связанных с проектом группах Google Groups.

Google курировала разработку и запуск этого проекта с долгосрочной целью, чтобы им руководили авторитетные эксперты в области сохранения языка. В настоящее время проект управляется Первым Народным культурным советом и командой проекта «Каталог исчезающих языков / Исчезающие языки» (ELCat/ELP) Гавайского университета в Маноа в координации с Советом по управлению ELP.

Список языков, входящих в этот проект, и информация о них отражают исходные данные из Каталога исчезающих языков (ELCat)[2], составленного

[1] Endangered Languages Project (ELP). – URL: http://www.endangeredlanguages.com/about/ (дата обращения: 01.12.2021).

[2] Catalogue of Endangered Languages. – URL: http://ling.hawaii.edu/research-current/projects/elcat/ (дата обращения: 01.12.2021).

Гавайским университетом в Маноа (University of Hawai'i at Manoa) и Linguist List (см. главу 2).

В каталоге ELP представлено 7341 ресурса, в том числе по темам:

- Языковые исследования и лингвистика 2936
- Языковая ревитализация 659
- Языковые материалы 3723
- Языковое образование 719
- Языковая пропаганда 758
- Язык, культура и искусство 2683
- Язык и технология1402
- Материалы СМИ 2830

На сайте каталога обсуждается подход проекта по сложным вопросам языковой документации: соотношения «язык / диалект» или понятия «вымершие языки». Во всех случаях предполагается широкое обсуждение со специалистами.

В проекте используется количественный показатель *индекс языковой опасности* (*LEI*), и приводится методика расчета этого показателя[1]. Каждому языку присваивается балл от 0 до 5 (безопасный – находящийся под угрозой исчезновения) – в зависимости от того, насколько корректно он соответствует установленным критериям. Если для данного языка нет информации, относящейся к одной из этих четырех категорий, этот язык не оценивается для этой категории.

***Формат данных для цифровой лингвистики DaFoDiL***[2]

Этот проект направлен на создание стандартизированного, предназначенного для чтения веб-совместимого формата для хранения лингвистических данных, в основном по исчезающим языкам, следуя лучшим практикам управления данными в современной Сети. DaFoDiL – часть более широкого проекта под названием *Цифровая лингвистика* (DLx)[3], целью которого является создание веб-инструментов для управления лингвистическими данными.

Репозиторий содержит спецификацию формата данных для цифровой лингвистики (сокращенно DaFoDiL). Эта спецификация представляет собой рекомендацию, как хранить лингвистические данные стандартизированным, удобочитаемым и совместимым с Интернетом способом, используя популярный в Интернете формат хранения данных JSON. Инструменты этого рекомендуемого формата интероперабельны, что позволит пользователям легко переносить свои данные с одного инструмента на другой. Кроме того, этот формат совместим с современной веб-платформой, что позволяет легко управлять лингвистическими данными онлайн или в браузере. Все проекты DLx используют такой формат данных.

[1] Индекс языковой опасности. – URL: https://www.hmong.press/wiki/ Catalogue_of_ Endangered_Languages (дата обращения: 01.12.2021).

[2] The Data Format for Digital Linguistics. – URL: https://format.digitallinguistics.io/ (дата обращения: 01.12.2021).

[3] Digital Linguistics. – URL: https://digitallinguistics.io/about/ (дата обращения: 01.12.2021).

Этот формат также способствует соблюдению Остинских принципов цитирования данных в лингвистике (см. гл. 6), поддерживая использование постоянных идентификаторов, полей для идентификации участников данных и их ролей, легкость поиска, читабельность (в виде удобочитаемых ключей в дополнение к непрозрачным идентификаторам баз данных) и совместимость между различными инструментами и веб-технологиями в целом.

Проект в основном нацелен на документирование исчезающих и низкоресурсных языков. Многие языковые сообщества и ученые, работающие с ними, пытаются защитить эти языки, разрабатывая для них инструменты, веб-сайты и ресурсы. Однако эти подходы являются фрагментарными, часто сопряженными с большими затратами на разработку и финансирование отдельных, не расширяемых вариантов использования. Чтобы облегчить этот процесс для всех заинтересованных сторон, была создана первая база данных всех инструментов с открытым исходным кодом, связанных с языками с низким уровнем ресурсов[1].

База данных представлена в виде простого списка, размещенного в репозитории GitHub. В настоящее время список включает более 241 проекта с открытым исходным кодом, который содержит специальные разделы для расширяемого кода для 26 различных языков.

Ранее этот каталог относился только к исчезающим языкам, теперь он распространен на языки меньшинств, которые обладают ограниченными ресурсами. Во введении приводятся определения и комментарии к включенным (и не включенным) данным.

*Пример записи*

Каждая запись представляет собой одну строку, содержащую название ресурса, ссылку на ресурс и краткое описание. Если ресурс также является репозиторием GitHub, то включается ссылка на значок, который показывает количество звезд (аналогично лайкам на других сайтах социальных сетей) для этого репозитория. Вот одна из таких записей для языка чичева[2]:

nya :: chicheŵa

Chichewa – NLP resources for Chichewa.

Ссылка «GitHub stars» автоматически включает в себя SVG-файл изображения количества звезд, полученных конкретным репозиторием, что позволяет читателям легко увидеть, насколько популярен репозиторий на GitHub. Обычно это бывает одна строка кода.

Каталог включает данные по следующим категориям (приводится несколько примеров наполнения категорий и количество продуктов по остальным категориям):

- Проекты и утилиты для одноязычной лексикографии
  - проект для бесплатных электронных словарей коренных языков (для мобильных телефонов)

[1] Digital Linguistics. – URL: https://github.com/digitallinguistics (дата обращения: 01.12.2021).

[2] kscanne/Chichewa. – URL: https://github.com/kscanne/chichewa (дата обращения: 01.12.2021).

- Webonary-сайт, на котором размещаются цифровые словари для отдельных языков
- WeSay (The SIL International) – позволяет языковым сообществам создавать свои собственные словари[1]

- Программное обеспечение. Аннотированный алфавитный список ПО для NLP (Около 160 программ)
- Помощники по настройке раскладки клавиатуры (восемь продуктов)
- Аннотации (13 продуктов)
- Спецификации форматов
  - spec – официальная спецификация формата лингвистических данных DLx[2]
  - формат лингвистической аннотации на основе XML для представления ЛИР (включая корпуса) с лингвистическими аннотациями[3]
  - xdxf_makedict – формат словаря XDXF и программа для преобразования словарей «makedict» (официальный репозиторий)
- Репозитории интернационализированных приложений (i18 n) (три продукта)
- Аудиоавтоматизация (28)
- Преобразование текста в речь (TTS) (3)
- Автоматизация текста (4)
- Экспериментирование (6)
- Флеш-карты (3)
- Генерация естественного языка. Библиотека OpenCCG[4] для синтаксического анализа и реализации с помощью CCG. Включает в себя мини-грамматики для инуитского, баскского и других языков
- Вычислительные системы. Общие языковые ресурсы и технологическая инфраструктура (Норвегия)
- Приложения для Android (11)
- Расширения Chrome
  - расширение для изучения языка через браузер[5]
  - словари низкоресурсных языков для веб-сайтов. Расширения для Google Chrome
- FieldDB6 Автономная / онлайн-полевая база данных, которая адаптируется к терминологии своего пользователя и I-языку, имеет плагины для различных процедур автоматизации данных
- Веб-службы FieldDB / Компоненты / Плагины (11)

---

[1] WESAY. LANGUAGE TECHNOLOGY. – URL: https://software.sil.org/wesay (дата обращения: 01.12.2021).

[2] The Data Format for Digital Linguistics (Daffodil). – URL: https://format.digitallinguistics.io/ (дата обращения: 01.04.2022).

[3] FoLiA: Format for Linguistic Annotation version 0.8 – revision 2.2. – URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.221.4523&rep=rep1&type=pdf (дата обращения 01.04.2022).

[4] OpenCCG/openccg. – URL: https://github.com/OpenCCG/openccg (дата обращения: 01.12.2021).

[5] Babelfrog. – URL: https://github.com/dergachev/babelfrog (дата обращения: 01.12.2021).

[6] DictionaryChromeExtension. – URL: https://github.com/FieldDB/DictionaryChromeExtension (дата обращения: 01.12.2021).

- Репозитории научных исследований (8)
- Примеры репозиториев (5)
- Шрифты (3)
- Корпуса (2)
- Организации
  - на GitHub (27)
  - другие (4)
- Руководство «Как написать корректор орфографии» [4]
- Языковые проекты (перечень из 38 языков). По каждому языку приводится подборка инструментов

В качестве примера рассмотрим *Библиотеку программ для уральских языков,* описанную в работе [5].

Эта библиотека включает 26 языков, в том числе 20 урало-алтайских.

UralicNLP может поддерживать следующие функции:

- морфологический анализ
- генерация морфологических форм
- лемматизация слова
- синтаксис – устранение неоднозначности
- многоязычная грамматика ограничений
- словари

UralicNLP позволяет получить лексикографическую информацию из словарей Giella[1]. Информация может содержать такие данные, как переводы, примеры предложений, семантические теги, морфологическую информацию и так далее. Нужно определить языковой код словаря. UralicNLP предоставляет семантические модели для финского (SemFi) и других уральских языков (SemUr) – для коми-зырян, эрзя, мокша и сколт-саамов.

Другие функциональные возможности:

- машинный перевод
- анализ зависимостей в финском языке

***Архив исчезающих языков ELAR***[2]

Это – цифровое хранилище, содержащее и публикующее мультимедийные коллекции исчезающих языков. В архиве собраны коллекции со всего мира с региональными опорными пунктами в Африке, на Ближнем Востоке, в Азии, Австралии и Латинской Америке. На сегодняшний день в ELAR хранятся записи, охватывающие более 450 языков. Коллекции ELAR содержат аудио- и видеозаписи повседневного использования языка, словесного искусства, песен, рассказов, ритуалов и многое другое. Коллекции также содержат словари, педагогические материалы, например, буквари для преподавания языка, транскрипции и переводы записей на основные контактные языки, такие как испанский, мандаринский, английский или русский.

[1] Giella. – URL: https://ru.wiktionary.org/wiki/giella (дата обращения: 01.12.2021).

[2] Endangered Languages Archive. – URL: https://www.elararchive.org/ (дата обращения: 01.12.2021).

ELAR создан в рамках Программы документирования исчезающих языков (ELDP)[1], основанной в 2002 году.

Миссия ELAR состоит в том, чтобы:

- обеспечить безопасное долгосрочное хранение коллекций языковой документации;
- обучать и поддерживать вкладчиков в создании и сохранении коллекций;
- сделать коллекции бесплатными для исследователей, сообществ и общественности;
- поддержать пользователей при поиске записей и доступе к ним.

Коллекции могут быть просмотрены и доступны через онлайн-каталог ELAR. Все материалы являются цифровыми и доступны бесплатно (после регистрации). ELAR сотрудничает с институтами по всему миру, проводя тренинги и представляя свою работу на выездных мероприятиях. Для пользователей предоставляется инструкция по использованию архива[2].

ELAR в настоящее время переходит на новую систему архивирования LAMUS, поддержка прежней платформы была прекращена.

ELAR требует создания метаданных в формате IMDI. Все файлы метаданных должны быть сделаны с помощью нового инструмента создания метаданных ELAR *lameta*. Программу, а также видеоурок и справочный лист можно найти на сайте lameta[3].

С 2017 года ELAR входит в состав Центра знаний CLARIN по лингвистическому разнообразию и языковой документации (CKLD), предлагая свой опыт исследователям и членам сообщества по всей Европе.

***Архив языков коренных народов Латинской Америки AILLA***[4]

AILLA – это цифровой языковой архив записей, текстов и мультимедийных материалов на языках коренных народов Латинской Америки и о них. Миссия AILLA – сохранить эти материалы и сделать их доступными для коренных народов, современных исследователей и других друзей этих языков и для будущих поколений. Имеются разнообразные пользовательские функции, в том числе возможность выполнять поиск по ключевым словам во всех коллекциях, а также возможность потоковой передачи и просмотра некоторых файлов мультимедиа без необходимости их предварительной загрузки. Возможен поиск по коллекциям, языкам, странам, организации и лицам.

---

[1] Endangered Language Documentation Programme. – URL: https://www.eldp.net/en/about s (дата обращения: 01.12.2021).

[2] ELAR_Navigating. – URL: https://www.dropbox.com/s/mrj9vao8rveabgy/ ELAR_ Navigating_20210225. pdf?dl=0 (дата обращения: 01.12.2021).

[3] Lameta. – URL: https://sites.google.com/site/metadatatooldiscussion/home (дата обращения: 01.12.2021).

[4] Archive of the Indigenous Languages of Latin America (AILLA). – URL: https://ailla. utexas.org/ (дата обращения: 01.12.2021).

***Тихоокеанский и региональный архив цифровых источников культур, находящихся под угрозой исчезновения, PARADISEC***[1]

PARADISEC предлагает возможность цифрового сохранения и доступа к находящимся под угрозой исчезновения языкам и материалам со всего мира. Архив может предоставлять доступ заинтересованным сообществам, и соответствует новым международным стандартам цифрового архивирования. Основная мотивация этого проекта – сделать полевые записи доступными для зарегистрированных пользователей и их потомков. Изначально задуманный как архивный проект, ориентированный на Азиатско-Тихоокеанский регион, PARADISEC превратился в центр для деятельности, в том числе в следующих сферах:

- обучение управлению данными (лексикографическое программное обеспечение, транскрипция текстов и подстроечные примечания), технике записи и связыванию данных;
- Предоставление каталога, в котором пользователи могут создавать описания своих коллекций;
- стандартизированные метаданные – создание описаний в формах, которые соответствуют стандартам и собираются поисковыми системами Open Archives Initiative, обеспечивая доступ для более широкого сообщества;
- создание моделей, которые показывают, как создавать повторно используемые данные (например ExSite9), с использованием современных инструментов, таких как Elan и Toolbox;
- создание моделей, показывающих, как повторно использовать данные (например EOPAS, онлайн-словари, репозитории ITunes);
- сохранение культурного наследия – резервное копирование и предоставление данных для культурных агентств в регионе;
- Global Focus – размещение файлов с исследовательской направленностью со всего мира (включая США, Чили, Мексику).

***Рекомендации библиотеки Йельского университета***

Библиотека Йельского университета разработала комплекс рекомендаций по оцифровке языковых материалов, которые пользуются авторитетом и известностью в сообществе. Эти рекомендации собраны на отдельном разделе портала библиотеки[2].

В рекомендациях говорится, что пока не существует единого программного пакета, который был бы способен обрабатывать все аспекты типичного рабочего процесса документирования языковых данных. Вместо этого существует большое и постоянно увеличивающееся количество пакетов, предназначенных для обработки различных аспектов рабочего процесса, многие из которых значительно перекрываются. Некоторые из этих

[1] Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). – URL: https://www.paradisec.org.au/ (дата обращения: 01.12.2021).

[2] Yale University Library. – URL: https://web.library.yale.edu/digitizationguidelines (дата обращения: 01.12.2021).

пакетов используют стандартные форматы и совместимы, тогда как другие гораздо менее совместимы. Далее приводится перечень рекомендуемых пакетов.

*SayMore* – это пакет языковой документации, разработанный SIL International, который в первую очередь ориентирован на начальные этапы языковой документации и нацелен на относительно несложный пользовательский интерфейс. Основными функциями SayMore являются:

- аудиозапись;
- импорт файлов с записывающего устройства (видео и / или аудио);
- организация файлов;
- ввод метаданных на уровне сеанса и файла;
- связь AV-файлов со свидетельством согласия и другими дополнительными объектами (например фотографии);
- сегментация AV-файла;
- транскрипция / перевод;
- выделение «осторожной речи» и устный перевод.

Файлы SayMore можно далее экспортировать для аннотации в FLEx, а метаданные можно экспортировать в форматы csv и IMDI для архивирования.

*ELAN* – инструмент для выравнивания по времени аннотации видео- или аудиоданных.

*FLEx* (*FieldWorks Language Explorer*). FLEx разработан SIL International. FLEx позволяет пользователю создавать «словарный запас» языка, т.е. список слов с определениями и грамматической информацией, а также сохранять тексты определенного языка. В текстах каждое слово или часть слова (например «морфема») связаны с записью в лексиконе. Для новых проектов и для студентов, которые начали учиться, FLEx теперь является лучшим инструментом для подстрочного перевода и создания словарей.

*Toolbox* полевого лингвиста является предшественником FLEx, и в течение нескольких десятилетий он был одним из наиболее широко используемых пакетов языковой документации, ранее известный как *Shoebox*. Основными функциями *Toolbox* являются построение лексической базы данных и интерлинеаризация текстов посредством взаимодействия с лексической базой данных. Как лексическую базу данных, так и тексты можно экспортировать в среду обработки текста, в случае с лексической базой данных – с помощью инструмента преобразования *Multi-Dictionary Formatter* (*MDF*). Также можно использовать *Toolbox* в качестве среды транскрипции. По сравнению с ELAN и FLEx Toolbox имеет относительно ограниченную функциональность, и некоторые считают, что он имеет неинтуитивный дизайн и интерфейс. Тем не менее, большое количество проектов было выполнено в среде *Shoebox*/*Toolbox*, и она продолжает пользоваться поддержкой сообщества. *Toolbox* также имеет преимущество работы напрямую с удобочитаемыми текстовыми файлами, которые можно открывать в любом текстовом редакторе, легко ими манипулировать и архивировать их. Файлы *Toolbox* также можно легко преобразовать для хранения в XML.

### Языковая документация и ресурсы для ревитализации языков Living Languages[1]

Этот ресурс является одновременно руководством «how-to» и простым английским шаблоном для написания руководства по обучению коренным австралийским языкам. Руководство было написано с учетом опыта ревитализации языков пама-нюнган, но может быть полезно и для других языков. Шаблон написан доступным языком, чтобы помочь нелингвистам использовать академические грамматики для создания удобного для сообщества ресурса. Он также предлагает структуру руководства для учащихся и предоставляет шаблонный текст, который может быть использован или адаптирован в соответствии с языком:

- Часть 1: говори на своем языке! Полезные слова и фразы;
- Часть 2: как работает ваш язык – складывание слов и предложений вместе.

Шаблон был разработан как *Google Doc* – простой в использовании формат, который позволяет в режиме реального времени удаленной совместной работы. В то же время можно предоставить шаблон в других форматах (например Word doc или PDF).

В Википедии отмечены как существенные для создания ЛИР исчезающих языков *Рекомендации по набору форматов архива Института психолингвистики Общества Макса Планка*, размещенные на странице портала Института[2]. Эти рекомендации являются приложением к описанию языка управления архивом Lamus.

С известной долей условности к ЛИР этого же класса *Языковая документация* можно отнести порталы, посвященные комплексному описанию конкретного языка. Примером такого комплексного портала является Информационная система *Grammis*[3]. Она включает комплекс баз данных и других ЛИР по разным аспектам изучения немецкого языка. Кратко опишем основные разделы системы.

Раздел *Исследования* системы Grammis представляет результаты текущих и завершенных научных исследований, в том числе:

- систематическая грамматика – иерархически структурированные и мультимедийно обработанные грамматические знания;
- корпусная грамматика – исследование грамматических вариаций на основе корпуса;
- сравнительные исследования языка по морфологии и синтаксису немецкого языка с выбранными европейскими языками;

---

[1] Living Languages. – URL: https://www.livinglanguages.org.au/resources (дата обращения: 01.12.2021).

[2] Lamus – Language Archive Management and Upload System. Appendix A. Accepted file types and formats. – URL: https://www.mpi.nl/corpus/html/lamus/apa.html (дата обращения: 01.12.2021).

[3] GRAMMIS. Grammatisches Informationssystem. – URL: https://grammis.ids-mannheim.de/ (дата обращения: 01.12.2021).

- фонология слов по просодическим свойствам слов и их связей с морфологическими структурами;
- научная терминология – последовательная систематика лингвистических словарей с акцентом на грамматику.

Раздел *Основные знания* включает в себя подготовленную специализированную информацию по выбранным грамматическим темам и сомнительным случаям:

- грамматика в вопросах и ответах;
- каталог основных грамматических терминов;
- пропедевтическая грамматика;
- контрастивное описание выбранных тематических областей немецкой грамматики с французским, итальянским и другими языками;
- немецкая орфография.

Раздел *Ресурсы* этой системы содержит:

- Словарь словесности глаголов
- Словарь предлогов
- Словарь союзов
- Словарь аффиксов
- Маркированная база данных родительного падежа
- База данных синтаксических связей в предложении
- База данных атрибутивных прилагательных
- База данных орфографическая
- Библиография по немецкой грамматике
- Библиография по немецкой орфографии

## Российские ЛИР языковой документации

Проблематика языковой документации привлекает внимание отечественных лингвистов. Например, можно сослаться на материалы Международной конференции «Документирование языков и диалектов коренных малочисленных народов России»» [6], состоявшейся в октябре 2019 года в Институте лингвистических исследований РАН. Одной из обсуждавшихся проблем было создание ЛИР языковой документации, и ряд докладчиков затрагивали эту проблему.

В данном разделе мы приведем описания некоторых российских ЛИР языковой документации.

### *Платформа ЛингвоДок*[1]

Выше было приведено описание архива языковых данных ЛингвоДок, в котором в настоящее время собраны аудиословари и корпуса более чем на

[1] LingvoDoc 3.0. – URL: http://lingvodoc.tsu.ru/ (дата обращения: 01.12.2021).

900 исчезающих диалектах уральских и алтайских языков России. Описание проекта можно также найти по адресу[1].

Помимо места для хранения данных и поиска данных, на этой платформе есть возможность одновременной распределенной обработки языкового материала и его анализа, в частности выявления в онлайн-режиме фонетического сходства языков, употребления тех или иных морфологических параметров в определенном значении, возможности построения карт фонетических, морфологических или лексических изоглосс в синхронии и их изменений в диахронии.

На платформе ЛингвоДок есть возможность размещения данных пользователей из различных организаций с сохранением всех прав создателей словарей и корпусов, работы с данными в режиме, когда материалы открыты только ограниченному числу пользователей, выбранных создателем словаря или корпуса. Но при этом для каждого пользователя ЛингвоДока появляется возможность сравнения данных его словарей по любым параметрам с данными других диалектов с помощью авторских программ сотрудников Лаборатории. Благодаря тому, что на платформе уже сейчас представлены материалы в едином цифровом формате по 900 диалектам уральских и алтайских языков России, суммарный объем которых превышает два миллиона словоформ, то анализ сравнительный – исторический, фонетический, морфологический – проводится методом обработки больших данных, что значительно повышает точность полученного результата.

В настоящее время ведется работа в коллаборации с создателями национальных корпусов языков России. Создается специальная среда для применения парсеров в онлайн-режиме, снятия омонимии, выявления коллокационных сочетаний. Планируется создание специальных программ для корпусного описания морфологии. На основании более полного описания морфологии планируется создание учебных платформ на базе Revita в сотрудничестве с учеными из Хельсинского университета и филиала НИУ «Высшая школа экономики» в Санкт-Петербурге.

***Тематическая сеть языковой документации и языковых технологий для приполярного региона***

В центре внимания планируемой Сети[2] находится разработка устойчивых языковых технологий и создание языковых архивов, которые улучшают качество жизни и позволяют развивать обработку естественного языка на Севере, сохраняя при этом исчезающие языки и сообщества. Общие цели:

- синергия арктического лингвистического образования;
- поддержание исчезающих языковых сообществ в приполярном регионе, наращивая и развивая языковые технологии;

[1] ИСП РАН. Лаборатория «Лингвистические платформы». – URL: https://www.ispras.ru/groups/modis/laboratoriya-lingvisticheskie-platformy/ (дата обращения: 01.12.2021).

[2] Университет Арктики. Языковая документация и технологии изучения языка в циркумполярном регионе. – URL: https://ru.uarctic.org/tematicheskie-seti/yazykovaya-dokumentafiya-i-tekhnologii-izucheniya-yazyka-v-firkumpolyarnom-regione/ (дата обращения: 01.12.2021).

o возрождение и содействие устойчивости исчезающих языков в Приполярном регионе с помощью языковых технологий.

Основное направление деятельности Сети – документирование языковой практики и создание языковых технологий для языков коренных народов Приполярного региона. На основе технологий и инфраструктуры, разработанных в UiT (в сотрудничестве с другими учреждениями), разрабатываются вычислительные модели для языков коренных народов по всей Арктике. Инфраструктура Тромсё уже успешно используется для создания крупномасштабных вычислительных моделей для саамских языков, гренландского, ненецкого, коми и квенского, и еще есть базовые модели для множества других языков, включая бурятский, эвенкийский, равнинный кри и инупиак.

***Проект «Малые языки России»***[1]

Проект реализован в Лаборатории исследования и сохранения малых языков Института языкознания РАН под руководством д-ра филол. наук О.А. Казакевич.

Главной задачей проекта является систематизация имеющихся сведений по конкретным языкам. Каждый язык показывается максимально полно и с разных сторон: дается ареально-генетическая, базовая лингвистическая и детальная социолингвистическая, а также социально-политическая информация о языке (например мероприятия по общественной и административной поддержке языка и наличие нормативных актов, регулирующих юридический статус языка). Кроме того, для каждого языка приводится список основных публикаций.

Благодаря онлайн-формату можно преодолеть ограничения бумажных энциклопедий и собрать в одном месте с удобным доступом разнообразные типы данных: ссылки на основные интернет-ресурсы по малым языкам (электронные корпуса, словари и другие виды ресурсов), иллюстративные материалы (видеозаписи и аудиозаписи текстов на языке с расшифровками, фотографии), актуальные картографические и социолингвистические данные по результатам экспедиций последних лет, контакты основных исследовательских центров и специалистов по языку. Это позволяет любому пользователю быстро найти интересующую его информацию. Особенно важно, что все данные о языках могут регулярно пополняться и оставаться актуальными. Сведения о других проектах лаборатории можно найти на сайте Лаборатории исследования и сохранения малых языков[2].

***Список исчезающих языков в России***[3]

Организация Объединенных Наций по вопросам образования, науки и культуры определяет пять уровней языковой опасности между «безопасным» (не находящимся под угрозой исчезновения) и «вымершим»:

---

[1] Малые языки России. Проект Института языкознания РАН. – URL: https://minlang.site/about (дата обращения: 01.12.2021).

[2] Институт языкознания РАН. Лаборатория исследования и сохранения малых языков. – URL: https://iling-ran.ru/web/ru/departments/minority_languages (дата обращения: 01.12.2021).

[3] Список исчезающих языков в России. – URL: https://en.wikipedia.org/wiki/List_of_endangered_languages_in_Russia (дата обращения 01.04.2022).

- уязвимый – «большинство детей говорят на этом языке, но он может быть ограничен определенными доменами (например, дома)»;
- определенно находится под угрозой исчезновения – «дети больше не изучают язык как родной в семье»;
- находится под серьезной угрозой – «на языке говорят бабушки и дедушки и старшие поколения; хотя родительское поколение может его понимать, они не говорят на нем детям или между собой»;
- находится под угрозой исчезновения – «самые молодые носители являются бабушками и дедушками и старше, и они говорят на этом языке частично и нечасто»;
- вымерший – «говорящих не осталось; включены в Атлас, если предположительно вымерли с 1950-х годов».

В цитируемый список включены результаты третьего издания Атласа языков мира, находящихся под угрозой (2010 г.; ранее – Красная книга языков, находящихся под угрозой исчезновения), а также онлайн-издание вышеупомянутой публикации, оба опубликованные ЮНЕСКО.

***Интерактивная карта «Полшага до немоты»***[1]

Популярный интернет-ресурс, посвященный проблемам исчезающих языков.

Представленный обзор по языковой документации, конечно, не претендует на исчерпание темы, но как надеется автор, содержит достаточно сведений для общего знакомства с темой и представления о технологиях, используемых для создания ЛИР по исчезающим языкам.

## Литература к главе 8


1. Digital Endangered Languages and Music Archiving Network (DELAMAN). 2018. Minimal checklist for the preservation of digital language documentation materials. – URL: http://hdl.handle.net/10125/55829 (дата обращения: 01.12.2021).
2. Electronic Metastructure for Endangered Languages Data E-MELD School of Best Practice. – URL: http://emeld.org/school/index.html (дата обращения: 01.12.2021).
3. Steven Bird, Gary Simons. Seven Dimensions of Portability for Language Documentation and Description // Language. – 2003. – N 79. – P. 557–582.
4. How to Write a Spelling Corrector. – URL: http://norvig.com/spell-correct.html (дата обращения: 01.12.2021).
5. Hämäläinen M. UralicNLP: библиотека НЛП для уральских языков // Журнал программного обеспечения с открытым исходным кодом. – 2019. – N 4(37). – P. [1345]. – URL: https://doi.org/10.21105/joss.01345 (дата обращения: 01.12.2021).
6. Международная конференция «Документирование языков и диалектов коренных малочисленных народов России». Тезисы докладов международной научной конференции, Санкт-Петербург 14–16 октября 2019 г. – Санкт-Петербург : ИЛИ РАН, 2019. – 88 с. – URL: https://iling.spb.ru/events/language_documentation2019/abstracts.pdf (дата обращения: 01.12.2021).


[1] Образовательный центр Сириус. Полшага до немоты. – URL: http://d-storytelling.sochisirius.ru/language#rec233605395

# ГЛАВА 9. КАТАЛОГИ И БИБЛИОТЕКИ ЛИНГВИСТИЧЕСКОГО ИНСТРУМЕНТАРИЯ

Программные инструменты для различных задач обработки текста, создания и поддержания ЛИР представляют собой огромную область, включающую тысячи разнообразных продуктов. Их анализ требует самостоятельного исследования, которое выходит за рамки настоящей работы. Следует также учитывать, что некоторые категории лингвистического ПО, например для обработки корпусов, лексикографии, лингвистического аннотирования и прочего описаны в других главах настоящей книги.

В данной главе мы ограничимся описанием ресурсов, в которых имеются каталоги и обзоры лингвистических программных средств, с учетом их классификации. В этой главе будут описаны также некоторые библиотеки лингвистических программ.

Глава снабжена достаточным количеством ссылок, по которым читатель сможет найти более подробную информацию. Авторский каталог лингвистических инструментов содержится в приложении 7.

## Каталоги лингвистических программ

***Исследовательские инструменты анализа текстов TAPoR 3.0***[1]

Начнем обзор с каталога программного обеспечения для анализа текста TAPoR, разработанного в Университете Альберта (Канада). В настоящее время TAPoR 3.0 включает описания свыше 1600 инструментов, в том числе несколько десятков, созданных непосредственно коллективом TAPoR. Для всех программных инструментов, включенных в каталог, приводится подробное описание, включающее аннотацию, назначение и методы в соответствии с таксономией TADIRAH[2], а также несколько собственных классификаций лингвистических инструментов.

TAPoR в соответствии со своим названием первоначально включал только инструменты для обработки и анализа текста. Но в 2018 году TAPoR

[1] Discover research tools for studying texts. TAPoR 3.0. – URL: https://tapor.ca/pages/about_tapor. (дата обращения: 01.12.2021).

[2] TaDiRAH. – URL: http://tadirah.dariah.eu (дата обращения: 01.12.2021). См. также гл. 20

поглотил каталог проекта DiRT[1] и теперь включает инструменты, которые работают с нетекстовыми данными, и другие инструменты, используемые в цифровой гуманитаристике. Эта эволюция во многом отражает эволюцию цифровой гуманитаристики. Ниже приведены некоторые области инструментов и услуг, кроме анализа текста, которые в настоящее время представлены в TAPoR:

- издательские инструменты
- коммуникационные инструменты
- репозитории и инструменты архивирования
- инструменты ГИС
- расширения браузера
- фото- / видео- / аудиоредактор и сопутствующие инструменты
- инструменты разработки для цифровых гуманитарных наук

***Каталог SIL International***[2]

Это тематически организованный список ресурсов в Интернете, относящихся к компьютерной лингвистике. Приведем в сокращенном виде классификацию этого каталога:

- Общая информация
  - «Использование компьютеров в лингвистике: практическое руководство», под ред. Д. Лоулера и Х. Драй
  - консорциум лингвистических данных
  - коммерческие научно-исследовательские и опытно-конструкторские площадки
  - архивы программного обеспечения
- Программные средства
  - шрифты и многоязычные ресурсы
  - управление данными
  - анализ речи и фонетика
  - фонология и морфология
  - лексические средства
  - анализ текста и корпусная лингвистика
  - перевод
  - историко-сравнительное языкознание, диалектология
  - языки и утилиты обработки текста

***Обработка естественного языка на GitHub***[3]

По нашему мнению, самый полный каталог лингвистического ПО имеется на портале *GitHub* – крупнейшем веб-сервисе для хостинга IT-проектов. В разделе *Обработка естественного языка* представлено 7379 публичных

[1] Digital Research Tools (DiRT). – URL: https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage (дата обращения: 01.12.2021).

[2] SIL. Linguistics Software. – URL: https://www.sil.org/linguistics/linguistics-software (дата обращения: 01.12.2021).

[3] Natural-language-processing. – URL: https://github.com/topics/natural-language-processing (дата обращения: 01.12.2021).

информационных объектов, соответствующих этой теме. К сожалению, на GitHub нет систематического каталога, возможен только поиск по языкам и по степени популярности программных продуктов.

***Каталог LINGUIST List***[1]

Каталог наиболее популярного справочного портала по лингвистике *LINGUIST List* включает свыше 400 программных продуктов. Они распределены по следующим категориям:

- системы автоматизированного перевода
- конкордансеры
- программы визуализации
- инструменты полевой лингвистики
- историческая реконструкция
- словари
- морфологический анализ
- обработка естественного языка
- парсеры
- фонетический анализ
- распознавание и синтез речи
- таггеры
- транскрипция
- анализ речи (включая клинические приложения)
- другие программные средства

***Аннотированный список ресурсов Стэнфордского университета***[2]

Каталог создан на факультете статистической обработки естественного языка и корпусной компьютерной лингвистики. Раздел *Программные инструменты* включает следующие категории:

- системы машинного перевода
- таггеры частей речи
- чанкеры естественного языка
- модель родовых последовательностей
- парсеры
- семантические анализаторы
- распознавание именованных сущностей
- разрешение кореференции (анафоры)
- моделирование языковых инструментариев
- инструменты анализа текста
- инструменты текстового сжатия
- другие
- не классифицировано

[1] The LINGUIST List. Software. – URL: https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=2 (дата обращения: 01.12.2021).

[2] Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources. – URL: https://nlp.stanford.edu/links/statnlp.html (дата обращения: 01.12.2021).

***Универсальный каталог лингвистического и переводческого программного обеспечения LINGTRANSOFT.INFO***[1]

Данный портал собирает программные продукты, используемые для перевода с исчезающих языков и языков с минимальными цифровыми ресурсами. Каталог этого портала представляется наиболее подробным. Приводим его с небольшими сокращениями.

- Основополагающие технологии
  - шрифты и кодировка
  - клавиатура и автоматическое предложение
  - орфографические словари
- Социолингвистика. Изучение социальных и культурных влияний на язык
  - языковые и исторические карты
  - формирование списков слов из аудиозаписей
  - расчет лингвистического расстояния между родственными языками
  - графическое отображение языковых и культурных границ
  - оценка жизнеспособности языка
  - проверка понимания прочитанного
  - определение многоязычия
- Изучение языка
  - частотность слова в корпусе естественного текста
  - мультимедийные флеш-карты для языковой практики
  - замещающие упражнения для практики
  - тестовые аспекты произношения (тон, интонация и т.д.)
  - создание контента для новых языковых курсов
  - обмен курсами / материалами для изучения языков
  - геймификация в изучении языка
  - изучение грамматики путем экспериментов
  - импорт внешних списков слов
  - помощь в изучении иностранных языков
- Грамотность
  - анализ слов и букв
  - производство материалов для чтения
  - макетирование и верстка
  - стандартные шаблоны страниц
  - раннее развитие
  - адаптированное чтение
  - создание обучающих игр или учебные стимулы
- Антропология
  - хранение культурных явлений в базе данных
  - разметка культурных явлений с помощью стандартизованных кодов

[1] LINGTRANSOFT.INFO. – URL: http://lingtransoft.info/your-one-stop-shop-information-linguistic-and-translation-software (дата обращения: 01.12.2021).

    - качественный антропологический анализ текстов
    - количественный антропологический анализ текстов
    - хранение и визуализация родственных отношений
- Этноискусство
    - запись в цифровом виде музыкальных образцов
    - анализ музыкальных образцов
    - публикация музыки в альтернативной нотной записи
    - цифровая запись изобразительного искусства
    - каталогизация изобразительного искусства
- Лингвистика
    - лексикография
    - языковая документация / управление корпусом
    - фонология
    - грамматический анализ
    - публикация лингвистических данных в структурированном виде
- Общий перевод
    - адаптация перевода с родственного языка
    - автоматическое использование памяти переводов
    - нечеткое сопоставление из памяти переводов
    - ведение базы данных глоссария или терминологии
    - поиск в словаре
    - проверка орфографии
    - обмен памятью переводов с несколькими переводчиками
    - обмен комментариями и отзывами среди членов команды
    - управление рабочим процессом перевода
    - интерфейс перевода в документе, при просмотре оригинала в контексте
    - обратный перевод для проверки точности
    - интерфейс для перевода с сохранением форматирования оригинала.
    - параллельное выравнивание текста
- Перевод Священного Писания

***Каталог инструментов для корпусной лингвистики М. Барбера***[1]

Аннотированный каталог включает свыше 200 программных продуктов, представленных в Интернете. Программные продукты упорядочены по алфавиту. В предисловии указано, что основное внимание уделяется корпусно-ориентированному программному обеспечению NLP (особенно таггеры, парсеры, чанкеры, корпусные системы запросов и т.д.), но также и текстовым анализаторам (конкорданты и т.д.), поскольку они также представляют некоторый интерес для NLP, корпоративного обслуживания и запросов. Как правило, не отражены автоматические системы распознавания речи, средства перевода, электронные словари и средства набора текста на экзотических языках.

[1] General Resources.Tools. – URL: http://www.bmanuel.org/clr/clr2_tt.html (дата обращения: 01.12.2021).

***Набор инструментов для лингвистических исследований***[1]

Этот сайт функционирует как виртуальный инструментарий для многоаспектных лингвистических исследований. Сайт разделен на пять уровней, которые варьируются от очень общих (информационная грамотность), дисциплины в целом (лингвистика), до конкретной отрасли лингвистических исследований, а именно: корпусная лингвистика.

***Википедия*** приводит небольшой перечень лингвистических программ, разбитых на следующие разделы[2]:

- Обработка текста на естественном языке
  - электронные словари
  - орфокорректоры или спеллчекеры
  - поисковые системы
  - системы машинного перевода
  - системы автоматизированного перевода, в том числе программы управления памятью переводов
- Системы распознавания символов OCR
- Речевые системы
  - системы анализа речи
  - системы синтеза речи
  - системы голосового перевода (распознавание и синтез)

Приведем список еще некоторых каталогов лингвистического софта:

***Указатель и архив лингвистических программ Мичиганского университета***[3];

***Перечень программ для обработки лингвистических текстов, сформированных с использованием TeX/LaTeX***[4]*;*

***Аннотированный алфавитный список программного обеспечения Ричарда Литтауэра***[5]. Этот каталог ориентирован на NLP языков с ограниченными ресурсами;

***Программы для лингвистов*** **на портале DELAMAN**[6], портал ориентирован на сохранение исчезающих языков.

---

[1] Toolbox for linguistic research. – URL: https://lx.ugent.be/toolbox/ (дата обращения: 01.12.2021).

[2] Лингвистическое_программное_обеспечение. – URL: https://ru.wikipedia.org/wiki/ Лингвистическое_программное_обеспечение (дата обращения: 01.12.2021).

[3] Index of /~archive/linguistics/software/mac. http://websites.umich.edu/~archive/linguistics/software/mac/ (дата обращения 01.04.2022).

[4] TeX/LaTeX Information. – URL: https://www.ling.upenn.edu/advice/latex.html (дата обращения: 01.12.2021).

[5] Low Resource Languages. – URL: https://github.com/RichardLitt/low-resource-languages (дата обращения: 01.12.2021).

[6] Resources Minimal Checklist for the Preservation of Digital Language Documentation Materials. – URL: https://www.delaman.org/resources/#softwarel (дата обращения: 01.12.2021).

## Европейские каталоги ПО

***Обзоры лингвистических программ CLARIN***[1]

Европейская инфраструктура языковых средств и технологий CLARIN не только ведет учет лингвистических программных средств, но и готовит по ним краткие аналитические обзоры. В настоящее время на сайте CLARIN представлено несколько таких обзоров по следующим категориям лингвистических программ:

- инструменты для нормализации
- инструменты для распознавания именованных сущностей
- таггеры и лемматизаторы частей речи
- инструменты для анализа настроений и мнений

***Карта LRE***[2]

Как указывалось выше, одним из наиболее полных и авторитетных европейских ресурсов по учету и идентификации ЛИР является *карта LRE*. Приведем перечень категорий лингвистического ПО согласно этой карте и упорядоченных по числу ЛИР, отнесенных к этой категории. В скобках приводится число продуктов, имеющихся в массиве карты LRE.

- Таггеры и парсеры (400)
- Инструменты аннотации (245)
- Инструменты корпусные (83)
- Оценка программных продуктов (71)
- Распознаватель именованных сущностей (60)
- Инструменты машинного перевода (51)
- Программный инструментарий (41)
- Токенизаторы (35)
- Инструменты машинного обучения (32)
- Средства языка моделирования (29)
- Разрешение лексической многозначности (17)
- Распознаватели речи / Транскриберы (14)
- Обработка сигналов / Извлечение функций (14)
- Веб-сервисы (9)
- Синтезатор текста в речь (9)
- Идентификатор языка (6)
- Распознаватель динамиков (4)
- Инструменты анализа настроений (4)
- Просодические анализаторы (3)
- Анализаторы изображений (3)
- Инструмент устного диалога (1)

[1] Resource Families Introduction. – URL: https://www.clarin.eu/resource-families (дата обращения: 01.12.2021).

[2] LRE Map. – URL: http://lremap.elra.info/ (дата обращения: 01.12.2021).

***ELRC-SHARE Repository***

Приведем также сведения из другого европейского каталога ЛИР, а именно Репозитория Службы обмена Европейской координации языковых ресурсов[1] (в скобках также указано количество инструментов данной категории)

- Таггеры частей речи (46)
- Распознавание именованных сущностей (37)
- Токенизация (32)
- Другое (29)
- Анализ зависимостей (26)
- Лемматизация (22)
- Чанкинг (19)
- Разделение предложений (19)
- Парсинг (18)
- Выравнивание (10)
- Идентификация языка (9)
- Грамматика составляющих (7)
- Маркировки семантических ролей (7)
- Категоризация текста (7)
- Извлечение термина (6)
- Аннотация (5)
- Стемминг (5)
- Сегментация слов (5)
- Семантическая разметка (4)
- Кроулинг по вебу (4)
- Машинный перевод (3)
- Выравнивание предложений (3)
- Проверка орфографии (3)
- Выявления темы (3)
- Индукция двуязычной лексики (2)
- Структурная аннотация (2)
- Аннотация соединений (1)
- Аннотация структуры документа (1)
- Оценка (1)
- Языковое моделирование (1)
- Разделение абзаца (1)
- Семантическая маркировка классов (1)
- Маркировка семантических отношений (1)

***Языковые инструменты и ресурсы для польского языка***[2]

Этот каталог – пример национального каталога лингвистического ПО, созданного в рамках европейских инициатив. В этом каталоге представлено свыше 250 общедоступных ЛИР (данных и инструментов), разделенных на следующие категории (указываем только инструменты):

[1] ELRC-SHARE Repository. – URL: https://elrc-share.eu/ (дата обращения: 01.04.2022).

[2] Language Tools and Resources for Polish. – URL: http://clip.ipipan.waw.pl/LRT?action=show&redirect=lrt (дата обращения: 01.12.2021).

- Морфологические инструменты и ресурсы (19)
- Таггеры (10)
- Парсеры, грамматики, банки деревьев (26)
- Анализ настроений, майнинг мнений (7)
- Анализ кореференции (2)
- Инструменты анализа и синтеза речи (14)
- Демонстрационные программы машинного перевода (8)
- Сумматоры (6)
- Диакритизация (6)
- Распознавание именованных сущностей (3)
- Программное обеспечение для многословного выражения (3)
- Агрегирование услуг (3)

В этом каталоге также имеется список неклассифицированных лингвистических программных продуктов. Приведем его в качестве примера:

*Mobile plWordNet* – бесплатное мобильное приложение для просмотра plWordNet;

*WSDDE* – система для проектирования и проведения экспериментов по устранению неоднозначности смысла слов (R. Młodzki et al.);

*Frazeo* – поисковая система и кластеризатор новостей на польском языке;

*Segment* – основанный на правилах токенизатора предложений поддерживающий стандарт;

*Toki* – токенизатор, поддерживающий стандарт SRX, библиотеку C++ и инструментарий;

*Translatica SRX* – правила сегментации предложений для польского языка (LGPL);

*SyMGIZA++, расширение Giza++,* которое вычисляет симметричные модели выравнивания слов;

*Hipisek* – экспериментальная система ответов на вопросы**;**

*Fextor* – фреймворк для извлечения объектов;

LexCSD – система полуавтоматического устранения смысловых неоднозначностей;

*Суперматрица* – общий инструмент для получения лексико-семантических знаний;

*WordnetLoom* – приложение редактора wordnet;

*Toposław* – инструмент для создания электронных словарей словоизменений многословных единиц;

*CorpCor* – веб-инструмент для коррекции морфосинтактических аннотаций в кодированных TEI XML корпусах (например, NKJP);

*Stylo 2* – демонстрация стилометрии;

*DeepEvents* – извлечение событий на польском языке, основанное на глубоких нейронных сетях;

*Word similarity* – расчет сходства слов на основе вхождений слов, онлайн-сервис;

*Baltoslav* – с несколькими конвертерами скриптов;

*SpaCyPL* – модели польского языка и ресурсы для SpaCy;
*Jasnopis* – анализатор уровня неясности текста.

## Российские каталоги лингвистического ПО

***Портал знаний по компьютерной лингвистике***[1]

Мы уже указывали, что из всех известных каталогов ЛИР данный ресурс предлагает наиболее детальную таксономию, в том числе программных инструментов и методов компьютерной лингвистики. Вызывает большое сожаление, что этот портал не актуализируется. В этом каталоге перечни прикладных систем, технологий и программных средств отделены от классификации методов обработки текстов и речи, т.е. от задач, для которых эти программы применяются. Поэтому приводятся фрагменты разных разделов портала.

- Классификация методов (сокращенный перечень)
  - Методы анализа массива текстов
    - методы обработки речи
  - Методы обработки текста
  - Методы анализа текста
    - методы морфологического анализа
    - методы разрешения анафоры
    - методы разрешения неоднозначности
    - методы сегментации текста
    - методы синтаксического анализа
    - методы генерации текста
- Методы оценки работы алгоритмов и систем
  - Перечни инструментов и систем
  - Прикладные системы (список – 41 шт.)
  - Технологии и программные продукты (список – 50 шт.)

Опишем еще несколько российских каталогов лингвистического ПО с указанием классификаторов, применяемых в этих каталогах.

***NLPub – каталог ресурсов для обработки естественного языка***[2]

- Графематический анализ
- Морфологический анализ
- Синтаксический анализ
- Проверка правописания
- Расстановка переносов
- Построение конкордансов
- Извлечение ключевых слов

[1] Компьютерная лингвистика. Портал знаний. – URL: https://uniserv.iis.nsk.su/cl/index.php?ent=375 (дата обращения: 01.12.2021).

[2] NLPub – каталог ресурсов для обработки естественного языка. – URL: https://nlpub.ru/ (дата обращения: 01.12.2021).

- Автоматическое реферирование
- Тематическая классификация
- Тематическое моделирование
- Извлечение именованных сущностей
- Извлечение отношений
- Анализ тональности
- Информационный поиск
- Машинный перевод
- Обнаружение дубликатов
- Сегментация текста
- Интегрированные пакеты

***Каталог лингвистических программ и ресурсов в Сети***[1]

- Программы анализа и лингвистической обработки текстов
- Программы преобразования текстов
- Психолингвистические программы
- Генераторы текстов
- Системы обработки естественного языка и машинного перевода
- Каталоги и коллекции ресурсов
- Словари и тезаурусы
- Поисковые машины и системы полнотекстового поиска
- Системы синтеза и распознавания речи

***Продукты Центра речевых технологий***[2]

- Программы для распознавания речи в текст
- Синтез речи
- Системы речевого оповещения

***Учебные пособия***. Полноценное изложение проблематики лингвистического ПО с достаточным количеством примеров приведено в учебном пособии Е.И. Большаковой и др. [1].

## Библиотеки лингвистических программ

***Открытая библиотека ПО для NLP на Python spaCy***

Из специализированных библиотек лингвистического ПО наибольшей популярностью пользуется ***spaCy***[3]. Это бесплатная библиотека программных продуктов с открытым исходным кодом для расширенной обработки естественного языка (NLP) в Python.

---

[1] Каталог лингвистических программ и ресурсов в Сети. – URL: https://rvb.ru/soft/catalogue/index.html (дата обращения: 01.12.2021).

[2] Центр речевых технологий. – URL: https://www.speechpro.ru/product/ (дата обращения: 01.12.2021).

[3] spaCy- is a free, open-source library for advanced NLP in Python. – URL: https://spacy.io/usage/spacy-101 (дата обращения: 01.12.2021).

Библиотека spaCy разработана специально для промышленного использования и помогает создавать приложения, обрабатывающие и «понимающие» большие объемы текста. Ее можно использовать для построения систем извлечения информации, или понимания естественного языка, или предварительной обработки текста для глубокого обучения. Функции spaCy:

*токенизация:* сегментирование текста на слова, знаки препинания и т.д.;

*определение части речи: (POS:* присваивание типов слов токенам, например, глагол или существительное.

*анализ зависимостей:* Присваивание меток синтаксических зависимостей, описывающих отношения между отдельными токенами, такими как субъект или объект.

*лемматизация:* определение базовых форм слов. Например, лемма «было» – это «быть», а лемма «крысы» – «крыса»;

*определение границ предложения (SBD):* поиск и разбивка отдельных предложений;

*распознавание именованных сущностей (NER*): маркировка именованных «реальных» объектов, таких как люди, компании или местоположения;

*связывание сущностей:* устранение неоднозначности текстовых сущностей с преобразованием в уникальные идентификаторы в базе знаний;

*сходство:* сравнение слов, фрагментов текста и документов;

*текстовая классификация:* назначение категорий или меток всему документу или его частям;

*соответствие на основе правил:* поиск последовательностей токенов на основе их текстов и лингвистических аннотаций, аналогичных регулярным выражениям;

*подготовка:* обновление и улучшение прогнозов статистической модели;

*сериализация:* сохранение объектов в файлы или байтовые строки.

***Общая архитектура обработки текстов GATE***[1] – система NLP с открытым исходным кодом, использующая наборы компонентов на языке Java. Система изначально была разработана в Университете Шеффилда. С помощью GATE реализуются задачи, где требуется выявить смысловое содержание текста и кодировать его в структурированном виде путем добавления аннотаций к сегментам текста. Система применяется для извлечения информации, ручной и автоматической семантической аннотации, анализа кореферентности, работы с онтологиями (например WordNet), машинного обучения, анализа потока сообщений в блогах (например Twitter).

Семейство инструментов GATE включает:

- *GATE Developer* – среда разработки, предоставляющая богатый набор графических интерактивных инструментов для NLP;

[1] General Architecture for Text Engineering (GATE). – URL: https://gate.ac.uk/ (дата обращения: 01.12.2021).

- *GATE Mímir* – многопарадигмальный репозиторий, который может использоваться для индексации и поиска по тексту, аннотациям, семантическим схемам (онтологиям) и семантическим метаданным;
- *GATE Cloud* – для работы с крупномасштабными лингвистическими проектами;
- *GATE Teamware* – оптимизация работы серверов для совместного аннотирования текстов;
- *GATE Embedded* – библиотека объектов;
- *GATEWiki* – управляемая Вики.

GATE поддерживается обширным сообществом разработчиков, пользователей, преподавателей, студентов и ученых. GATE применяется в самых разных областях научных знаний, относящихся к компьютерной лингвистике, NLP, моделированию языковых процессов. GATE применяется во многих проектах в Великобритании и других странах.

Архитектура GATE состоит из взаимосвязанных компонентов: «кусочков» программного обеспечения с четко определенными интерфейсами, которые могут быть развернуты в различных контекстах. В GATE реализованы готовые решения для токенизации, тегирования, разделение текста на высказывания (сплитер), извлечение именованных сущностей, машинного обучения. Компоненты делятся на три категории по функциям:

- Language Resources (LR) – лингвистические ресурсы (данные);
- Processing Resources (PR) – программы для обработки документов (ресурсы);
- Visual Resources (VR) – графические интерфейсы для LR и PR.

В GATE поддерживаются следующие форматы документов: Plain Text, HTML, SGML, XML, RTF, Email, PDF, Microsoft Office (some formats), OpenOffice, UIMA CAS, CoNLL/IOB.

В GATE встроены различные средства для работы с Unicode. Поддерживаются языки: английский (по умолчанию), испанский, китайский, арабский, болгарский, французский, немецкий, хинди, итальянский, кебуано, румынский, русский.

В GATE разработана детальная технология, которая включает экспертные знания для создания, обучения и руководства многоцелевыми командами, деятельность которых требует сложных рабочих процессов, необходимых для проведения анализа текста экономически эффективным, устойчивым и точным образом. Технология охватывает четыре области:

- интеграция данных и моделирование предметной области (с помощью Ontotext);
- обогащение контента (путем извлечения информации или семантической аннотации);
- поиск, навигация и презентация (с помощью Ontotext);
- системная интеграция.

Процесс задается как набор диаграмм активности UML, которые поддерживаются инструментами управления бизнес-процессами, GATE Teamware, GATE Developer и GATE embedded.

***База данных для полевой лингвистики FieldDB***[1]

Это бесплатный модульный проект с открытым исходным кодом, разработанный совместно полевыми лингвистами и разработчиками программного обеспечения для создания расширяемого удобного приложения, которое можно использовать для сбора, поиска и обмена данными как онлайн, так и офлайн. По сути, это приложение, написанное на 100% Javascript и полностью работающее на стороне клиента, поддерживаемое базой данных NoSQL. У него есть ряд веб-сервисов, к которым он подключается, чтобы позволить пользователям выполнять задачи, требующие Интернета / облака (например синхронизацию данных между устройствами и пользователями, публичный обмен данными, запуск интенсивных процессов процессора для анализа, извлечения или поиска аудио / видео / текста). Несмотря на то, что приложение было разработано для полевых лингвистов, оно может быть использовано любым человеком, собирающим текстовые / аудио- / видеоданные или другие высокоструктурированные данные, где поля в каждой точке данных требуют шифрования или настройки от пользователя к пользователю, и где схема данных, как ожидается, будет развиваться в ходе сбора данных в «поле».

***Цифровая лингвистика DLx***[2]

Проект Калифорнийского университета в Санта-Барбаре. Согласно разработчикам, этот проект направлен на управление цифровыми данными для лингвистики, включая хранение, представление, обработку и распространение лингвистических данных. Этот проект занимается представлением лингвистических данных в цифровой форме, а также передовыми методами работы с этими данными, используя все преимущества современной *открытой веб-платформы* (OWP)[3]. Проект преследует четыре основные цели:

- определить стандартный формат данных для хранения лингвистических данных в машиночитаемой форме[4]. Этот формат должен быть независимым от платформы и программного обеспечения (т.е. не ограничиваться использованием Windows или конкретным программным обеспечением, таким как ELAN) и кодировать лингвистические концепции, сохраняя гибкость и удобочитаемость;
- предоставить сценарии и библиотеки, которые упрощают разработчикам работу с данными в формате DLx и создают инструменты и программное обеспечение, использующие этот формат. Все приложения DLx имеют открытый исходный код;
- создавать разнообразные веб-инструменты, которые позволят лингвистам без особых усилий вводить, искать свои данные и управлять ими.

[1] FieldDB. An offline/online field database which adapts to its user's terminology and I-Language. – URL: http://fielddb.github.io/ (дата обращения: 01.12.2021).

[2] Digital Linguistics (DLx). – URL: https://digitallinguistics.io/about/ (дата обращения: 01.12.2021).

[3] Web platform. – URL: https://en.wikipedia.org/wiki/Web_platform (дата обращения: 01.12.2021).

[4] The Data Format for Digital Linguistics (Daffodil). – URL: https://format.digitallinguistics.io/

Первый из этих инструментов – приложение для управления лексиконами – находится в стадии разработки;

o распространять передовые методы управления лингвистическими данными.

***Библиотека Python для исторической лингвистики LingPy***

Это набор модулей Python с открытым исходным кодом для сравнения последовательностей, анализа расстояний, операций с данными и методов визуализации в количественной исторической лингвистике. Описание LingPy представлено в: [2]. Основная идея LingPy – предоставить программный пакет, который, с одной стороны, объединяет различные методы анализа данных в количественной исторической лингвистике в рамках единой структуры, а с другой – служит интерфейсом для подготовки и анализа лингвистических данных с использованием биологических программных пакетов. С помощью LingPy пользователи могут:

– токенизировать и анализировать последовательности;

– проводить анализ попарного и множественного выравнивания;

– производить автоматический поиск родственных слов на нескольких языках;

– вычислять лексико-статистические расстояния между языками;

– реконструировать языковые филогении, используя базовые кластерные алгоритмы;

– экспортировать результаты этих анализов в различные форматы, которые можно использовать в качестве входных данных для внешних программ или для визуализации результатов.

***Библиотека программ OpenCCG***

Это библиотека программ с открытым исходным кодом для NLP, написанных на Java[1]. OpenCCG обеспечивает лингвистический разбор на основе формализма *Комбинаторной категориальной грамматики* (CCG) М. Стидмана.

В библиотеке используются мультимодальные расширения CCG, разработанные как часть системы Grok (предшественник OpenCCG). Последующие усилия по разработке были сосредоточены на том, чтобы сделать реализацию практичной для использования в диалоговых системах, а в последнее время – на реализации с грамматиками с широким охватом. Начиная с версии 0.9.4, OpenCCG включает широкую поддержку синтаксического анализа и реализации английского языка, что в совокупности позволяет экспериментировать с грамматическим перефразированием открытой области. Версия 0.9.5 добавляет функции для упорядочивания зависимостей и минимизации длины зависимостей, наряду с поддержкой использования 5-граммовых языковых моделей и создания дизъюнктивных логических форм на основе различий между выровненными семантическими графами. Она также включает расши-

[1] OpenCCG: The OpenNLP CCG Library. – URL: http://openccg.sourceforge.net/ (дата обращения: 01.12.2021).

рение, разработанное для грамматики, используемой системой полнотекстового поиска Sphinx[1], которая теперь выпущена с открытым исходным кодом.

OpenCCG используется для ряда диалоговых систем. Список проектов, использующих OpenCCG, размещен на сайте Лаборатории компьютерной лингвистики Техасского университета в Остине[2].

***СТАРЛИНГ***

К библиотекам лингвистического программного обеспечения можно также отнести комплексную программу СТАРЛИНГ, разработанную в рамках проекта Вавилонская башня[3], более подробно описанного в главе 13. СТАРЛИНГ была создана под руководством С. Старостина для работы с лингвистически ориентированными текстами и базами данных, она поддерживает разветвленную систему шрифтов для DOS и Windows. СТАРЛИНГ также включает программы морфологического анализа русского языка и программы для управления лексиконами.

## Российские разработчики лингвистического ПО

В России действует несколько исследовательских центров и коммерческих компаний, имеющих заметные результаты в разработке лингвистического ПО. Приведем несколько примеров, не претендуя на полноту сведений. Более полный перечень российских и зарубежных организаций, работающих в области компьютерной лингвистики (218 названий), приводится на портале знаний «Компьютерная лингвистика»[4].

***Яндекс.*** Компания начала свою деятельность именно с языковых технологий. Даже название компании «**Я**зыковой и**НДЕКС**» она получила от морфологического анализатора, разработанного одним из основателей компании – к сожалению, ушедшим от нас И. Сегаловичем. Этот анализатор *Mystem* является и сейчас одним из лучших для русского языка. В компании разработана и действует система словарей и машинного перевода, а также голосовой помощник *Алиса* со встроенной системой анализа и синтеза речи. Технологии Яндекса, включая языковые, описаны на отдельной странице портала[5].

***ABBYY***. Компания предлагает словари и продукты для переводчиков *ABBYY Lingvo x6* для различных платформ, а также популярный оптический

[1] Open Source Search Server. Sphinx. – URL: https://ru.wikipedia.org/wiki/Sphinx. (дата обращения: 01.12.2021).

[2] The UT Austin Computational Linguistics Lab. Projects using OpenCCG. – URL. http://openccg.sourceforge.net/ : (дата обращения: 01.04.2022).

[3] Вавилонская башня. Проект «Эволюция языка». – URL: https://starling.rinet.ru/program.php?lan=ru (дата обращения: 01.12.2021).

[4] Компьютерная лингвистика. Портал знаний. – URL: https://uniserv.iis.nsk.su/cl/index.php?ent=3 (дата обращения: 01.12.2021).

[5] Яндекс. Технологии. – URL: https://yandex.ru/company/technologies (дата обращения: 01.12.2021).

распознаватель символов *FineReader*. Продукты и решения ABBYY описаны на портале компании[1].

***Лаборатория информационных исследований***[2]. Разработаны собственные технологии:

- тематический анализ текстов (классификация, аннотирование, многоязычный поиск) на основе больших лингвистических онтологий;
- технологии оценки тональности, извлечения фактографической информации из текста;
- технологии кластеризации, классификации и обзорного реферирования новостного потока.

***Лаборатория «Лингвистические платформы ИСП РАН***[3]. Осуществляет разработку платформы Lingvodoc – совместный проект ИСП РАН, Института языкознания РАН и Томского государственного университета. Это система для совместной многопользовательской документации исчезающих языков, создания многослойных словарей и научной работы с полученными звуковыми и текстовыми данными. Среди распространяемых программ – TEXTERRA, технология автоматического построения онтологий и семантического анализа текста.

***Центр речевых технологий***[4]. ЦРТ – российская компания, разработчик инновационных систем в сфере технологий синтеза и распознавания речи.

***PROMT***. Комплекс переводчиков для различных приложений и платформ для 40 языков представлены на сайте компании[5].

***Информатик***[6]. Комплекс программ для проверки орфографии, морфологического анализа и поддержки словарей.

***Школа лингвистики ВШЭ***. Научные группы школы занимаются исследованиями в области типологии, социолингвистики и ареальной лингвистики, корпусной лингвистики и лексикографии, древних языков и истории языка. Кроме того, в школе разрабатываются лингвистические технологии и ресурсы: корпуса, обучающие тренажеры, словари и тезаурусы, технологии для электронного представления текстов культурного наследия. Проекты Школы перечислены на странице сайта[7].

***Кафедра математической лингвистики СПбГУ***. Ученые кафедры работают в области автоматической обработки текстов на разных языках, лингвистической семантики, синтаксиса, теории моделирования, терминоведения, автоматической лексикографии, стилеметрии, автоматической атрибу-

---

[1] ABBYY. Продукты и решения. – URL: https://www.abbyy.com/ru/company/management/ (дата обращения: 01.12.2021).

[2] Лаборатория информационных исследований. – URL: http://www.labinform.ru/ (дата обращения: 01.12.2021).

[3] ИСП РАН. Лаборатория «Лингвистические платформы». – URL: https://www.ispras.ru/groups/modis/laboratoriya-lingvisticheskie-platformy/

[4] Центр речевых технологий. – URL https://www.speechpro.ru/about/

[5] PROMT. Продукты. – URL: https://www.promt.ru/

[6] Продукты компании «ИНФОРМАТИК». – URL: http://www.informatic.ru/

[7] Проекты Школы лингвистики НИУ ВШЭ. – URL: https://linghub.ru/

ции текстов, квантитативной лингвистики. Труды кафедры отражены в коллективной монографии [3], а проекты перечислены на сайте кафедры[1].

***Лаборатория компьютерной лингвистики ИППИ РАН***[2]. Основные научные направления:

- действующая модель языка «Смысл <=> Текст»;
- многоцелевой лингвистический процессор ЭТАП-3 – компьютерная реализация модели «Смысл <=> Текст»;
- глубоко аннотированный корпус – составляющая часть Национального корпуса русского языка.

***Лаборатория «Цифровая документация русского языка» ИППИ РАН***[3]. Направления исследований:

- разработка лингвистических корпусов;
- корпусные и экспериментальные исследования русского языка;
- русская корпусная грамматика.

***Институт прикладной семиотики АН РТ***. Комплексные разработки ЛИР и процессоров для автоматической обработки татарского языка, перечень которых приведен на сайте института[4].

***Лаборатория речевых и многомодальных интерфейсов СПб ФИЦ РАН***[5]. Среди разработок: система распознавания слитной русской речи со сверхбольшим словарем (> 100 тыс. слов); компьютерная система синтеза аудиовизуальной русской речи и жестового языка по тексту; аватар для русского жестового языка, голосовые и диалоговые системы.

***Институт прикладной и математической лингвистики МГЛУ***[6]. Научные исследования и прикладные разработки в области речеведения, экспериментальной фонетики и психология речи, методик идентификации говорящего по голосу и речи.

---

[1] Научно-исследовательские проекты кафедры математической лингвистики. – URL: http://mathling.phil.spbu.ru/node/9 (дата обращения: 01.12.2021).

[2] Лаборатория компьютерной лингвистики ИППИ РАН. – URL: http://iitp.ru/ru/researchlabs/245.htm (дата обращения: 01.12.2021).

[3] Лаборатория «Цифровая документация русского языка» ИППИ РАН. – URL: http://iitp.ru/ru/researchlabs/digital_documentation (дата обращения: 01.12.2021).

[4] НИИ «Прикладная семиотика» АН РТ. Фундаментальные и прикладные разработки. – URL: http://www.antat.ru/ru/ips/science/rnd/ (дата обращения: 01.12.2021).

[5] СПб ФИЦ РАН Лаборатория речевых и многомодальных интерфейсов. – URL: https://spcras.ru/units/laboratory.php?ID=462624&UNITS=468487 (дата обращения: 01.12.2021).

[6] МГЛУ. Институт прикладной и математической лингвистики. – URL: https://linguanet.ru/fakultety-i-instituty/institut-prikladnoy-i-matematicheskoy-lingvistiki/nauchno-issledovatelskaya-rabota/ (дата обращения: 01.12.2021).

## Литература к главе 9

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. // МИЭМ. – Москва, 2011. – 272 с. – ISBN 978–5–94506–294–8.

2. LingPy. Библиотека Python для исторической лингвистики. Версия 2.6.5 / Лист Иоганн-Маттис, Гринхилл Саймон, Тресольди Тьяго, Форкель Роберт. – 2019. – URL: http://lingpy.org. – DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy (дата обращения: 01.12.2021).

3. Прикладная и компьютерная лингвистика [Текст] : коллективная монография /. Бочаров В.В [и др.] ; под ред. Николаева И.С., Митрениной О.В. Ландо Т.М. – Москва : URSS, 2016. – 315 с. : ил., портр., табл.; 21 см. – ISBN 978-5-9710-3472-8.

# ЧАСТЬ 3
# КАТЕГОРИИ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

## ГЛАВА 10. ТЕКСТОВЫЕ КОРПУСА

### Общие замечания

Лингвистический, или языковой, корпус текстов – это большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. Основными чертами современного корпуса являются машиночитаемый формат, репрезентативность, наличие металингвистической информации [1]. Репрезентативность достигается с помощью специальной процедуры отбора текстов.

Обычно лингвистическим корпусом называют совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединенных каким-то общим признаком (языком, жанром, автором, периодом создания текстов).

Целесообразность создания текстовых корпусов объясняется:

- представлением лингвистических данных в реальном контексте;
- достаточно большой представительностью данных (при большом объеме корпуса);
- возможностью многократного использования единожды созданного корпуса для решения различных лингвистических задач, таких как например реализация графематического и лексико-грамматического анализа текста и т.д.

Часто обсуждается терминологический вопрос – в чем разница между «корпусами», «коллекциями» и «архивами языковых данных»? Вот возможные определения [2]:

- архив: хранилище читаемых электронных текстов, не связанных каким-либо скоординированным образом, например Oxford Text Archive;
- электронная текстовая библиотека (ЭБ, фр. «Textothèque»): собрание электронных текстов в стандартизированном формате с определенными соглашениями, относящимися к содержанию и т.д.;
- корпус: подмножество ЭБ, построенное в соответствии с явными критериями проектирования для конкретной цели, например Corpus Révolutionnaire (Bibliothèque Beaubourg, Париж), Cobuild Corpus, корпус Longman / Lancaster, корпус Oxford Pilot.

Полный и содержательный обзор применения корпусов и корпусных технологий в различных областях теоретической и прикладной лингвистик содержится в докладе академика РАН В.А. Плунгяна «Современные применения корпусных технологий», сделанном на Научной сессии «Гуманитарные науки в эпоху цифровизации» Отделения историко-филологических наук РАН 15.04.2021 г. [3].

В данной главе мы будем рассматривать в основном текстовые корпуса. Специфика устных корпусов будет рассмотрена в главе 14.

## Статистика корпусов

Корпуса, особенно текстовые, стали самым распространенным видом ЛИР. Приведем некоторые статистические данные по крупнейшим мировым собраниям ЛИР.

В крупнейшем в мире *Языковом архиве Института психолингвистики Общества Макса Планка*[1] содержится около 150 тыс. ЛИР. Из них корпуса и другие коллекции текста составляют 91,2 тыс. ЛИР. По тематике к корпусной лингвистике и лингвистике текста относятся 22,3% ЛИР.

Еще некоторые данные о количестве корпусов в разных архивах в отношении общего числа ЛИР представлены в таблице 4. Следует, впрочем, иметь в виду, что в некоторых архивах не все ЛИР отнесены к определенному типу.

*Таблица 4*

**Количество ЛИР и корпусов в некоторых мировых архивах**

| Название архива | Количество ЛИР | Количество корпусов (Primary text) |
|---|---|---|
| Lingust list[2] | 13 200 | 5400 |
| Калифорнийский языковой архив[3] | 14 959 | 12 648 |
| Коллекция устных цифровых корпусов (CoCoON ex-CRDO)[4] | 15 515 | 13 102 |
| Кайпулеохоне[5] | 5359 | 3732 |
| Архив языков и культур SIL[6] | 30 177 | 1658 |
| Проект Rosetta: библиотека долговременного хранения языков человечества[7] | 6571 | 1322 |

[1] The Language Archive (TLA). – URL: https://archive.mpi.nl/tla/ (дата обращения: 01.12.2021).

[2] Lingust list. – URL: https://linguistlist.org/home/ (дата обращения: 01.12.2021).

[3] The survey of california and other indian languages. – URL: http://cla.berkeley.edu (дата обращения: 01.12.2021).

[4] COllections de COrpus Oraux Numeriques. – URL: http://cocoon.huma-num.fr/exist/crdo/

[5] Kaipuleohone. – URL: http://scholarspace.manoa.hawaii.edu/handle/10125/4250/

[6] Language & Culture Archives. – URL: http://www.sil.org/resources/language-culture-archives

[7] The Rosetta Project: A Long Now Foundation Library of Human Language. – URL: http://www.rosettaproject.org/

## Классификации корпусов

Существует множество коллекций, перечней и каталогов корпусов – можно привести несколько ссылок[1, 2]. Достаточно полный перечень корпусов различных языков и языковых семейств имеется на сайте Национального корпуса русского языка[3].

Мы приведем классификацию корпусов по цитируемой выше работе В.П. Захарова и С.Ю. Богдановой – наиболее известной работы по корпусной лингвистике на русском языке:

«Несмотря на разнообразие корпусов, можно выделить два основных способа их деления на классы:

1) противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому-либо жанру, стилю, языку определенной возрастной или социальной группы, языку писателя или ученого и т.д.;

2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks). При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц» [1].

Приведем сводную таблицу классификационных признаков и типов корпусов, выделенных по этим признакам, согласно работе В.П. Захарова и С.Ю. Богдановой (таблица 5).

*Таблица 5*

**Классификация корпусов**

| Признак | Тип корпуса |
|---|---|
| Цель | многоцелевые |
| | специализированные |
| Тип языковых данных | письменные |
| | устные (речевые) |
| | смешанные |
| «Литературность» | литературные |
| | диалектные |
| | разговорные |
| | терминологические |
| | Смешанные |

[1] English-Corpora.org: a guided tour. – URL: https://www.english-corpora.org/pdf/english-corpora.pdf

[2] University of Warwick. The Applied Linguistics Repository. – URL: https://warwick.ac.uk/fac/soc/al/repository/resources/

[3] Национальный корпус русского языка. Другие корпуса. – URL: https://ruscorpora.ru/new/corpora-other.html

*Продолжение таблицы*

| Признак | Тип корпуса |
|---|---|
| Жанр | литературные |
| | фольклорные |
| | драматургические |
| | публицистические |
| Назначение | исследовательские |
| | иллюстративные |
| Динамичность | динамические (мониторные) |
| | статические |
| Разметка | размеченные |
| | неразмеченные |
| Характер разметки | морфологические |
| | синтаксические |
| | семантические |
| | анафорические |
| | просодические и т.д. |
| Доступность | свободно доступные |
| | коммерческие |
| | закрытые |
| Объем текстов | полнотекстовые |
| | «фрагментнотекстовые» |
| Параллельность | одноязычные |
| | двуязычные |
| | многоязычные |
| Тип языковых данных | письменные |
| | устные (речевые) |
| | смешанные |

Данный подход к классификации корпусов, очевидно, не единственный. Приведем в сокращенном виде обзор корпусов, сделанный Д. Фишер и Я. Ленардичем для проектов CLARIN-PLUS, PARTHENOS и SSHOC [4]. В этом обзоре корпуса разделяются на 12 семейств:

- интернет-корпуса (корпуса компьютерно-опосредованной коммуникации, CMC)
- корпуса научных текстов
- исторические корпуса
- учебные корпуса второго языка
- литературные корпуса
- аннотированные вручную корпуса
- мультимодальные корпуса
- газетные корпуса
- параллельные корпуса
- парламентские корпуса
- референтные (эталонные) корпуса
- корпуса устной речи

При описании каждого семейства корпусов авторы приводят количество корпусов данного типа в CLARIN, с указанием языков, а также приводят примеры, которые можно увидеть по ссылкам, указанным при описании семейств.

***Интернет-корпуса (CMC)***[1] отражают публичную и частную коммуникацию в режиме онлайн, такую как сообщения в блогах, форумах, комментарии на новостных сайтах онлайн, социальных сетях и сетевых сайтах, таких как Twitter и Facebook, приложениях для мобильных телефонов, таких как WhatsApp, электронная почта и чаты. Поскольку корпуса CMC часто включают в себя неформальные стили письма, они интересны для широкого круга исследовательских областей, таких как языковая вариативность, прагматика, исследования средств массовой информации и коммуникации и т.д. Они также очень важны для развития надежных инструментов, которые могут справиться с нестандартной орфографией, лексикой и грамматикой. Компиляция и распространение таких корпусов затруднены неясным правовым статусом данных CMC при их распространении в качестве ресурса для научного сообщества, что еще более усугубляется быстро меняющимися условиями предоставления услуг контент-провайдерами.

***Корпуса научных текстов***[2] содержат научную литературу, включающую научные статьи, эссе и тезисы, опубликованные в научных журналах, материалы конференций, диссертации на уровне бакалавриата и магистратуры, а также научные монографии.

***Исторические корпуса***[3]. Инфраструктура CLARIN предлагает доступ к 75 историческим корпусам, охватывающим почти все языки, на которых говорят в странах, являющихся членами или наблюдателями CLARIN. В подавляющем большинстве случаев корпуса могут быть непосредственно загружены из национальных хранилищ или запрошены с помощью простых в использовании онлайн-поисковых сред.

***Учебные корпуса иностранного языка***[4]. Учебные корпуса играют решающую роль в исследованиях и педагогике второго языка, позволяя систематически изучать, как учащийся второго языка усваивает новый язык как на лексическом, так и на синтаксическом уровне, и то как на него влияет его родной язык. Особой характеристикой данного типа корпусов являются разметка ошибок и просодические особенности обучающихся.

***Литературные корпуса***[5] включают поэтические и прозаические тексты, такие как романы, рассказы и пьесы. Они объединяют собрание сочинений

---

[1] CMC corpora. – URL: https://www.clarin.eu/content/cmc-corpora (дата обращения: 01.12.2021). Следует отметить, что не все авторы отождествляют интернет-корпуса и CMC.

[2] Corpora of academic texts. – URL: https://www.clarin.eu/resource-families/corpora-academic-texts (дата обращения: 01.12.2021).

[3] Historical-corpora. – URL: https://www.clarin.eu/content/historical-corpora (дата обращения: 01.12.2021).

[4] L2 learner corpora. – URL: https://www.clarin.eu/resource-families/L2-corpora (дата обращения: 01.12.2021).

[5] Literary corpora. – URL: https://www.clarin.eu/resource-families/literary-corpora (дата обращения: 01.12.2021).

одного автора или представителя определенного литературного периода. Поскольку литературные корпуса часто доступны через мощные конкорданты, они особенно хорошо подходят для количественного и качественного подхода к сравнительному литературоведческому анализу внутри или между различными жанрами и историческими периодами.

***Аннотированные вручную корпуса***[1] – это наборы текстов, содержащих назначенную или проверенную вручную лингвистическую информацию, такую как морфосинтаксические теги, леммы, синтаксические разборы, именованные сущности и т.д. Эти корпуса могут быть использованы для обучения новым языковым инструментам аннотации, а также для проверки точности существующих инструментов аннотации. Корпуса подразделяются на шесть категорий в зависимости от типа ручной аннотации:

- PoSтегирование
- лемматизация
- синтаксический анализ
- распознавание именованных сущностей
- анализ настроений
- прочее

***Мультимодальные корпуса***[2] – это наборы данных, используемые для изучения взаимодействия модальностей (т.е. каналов коммуникации) друг с другом в человеческом общении. Мультимодальные корпуса часто представляют собой коллекции видео- и речевых записей, сопровождаемых транскрипциями и жестовыми аннотациями, хотя существуют и мультимодальные корпуса текстовых данных, дополненных изображениями. Такие корпуса могут быть использованы для исследования целого ряда лексических, просодических и жестовых особенностей речи, а также для исследования того, как эти особенности взаимодействуют в реальной, повседневной речи. Эти корпуса богато аннотированы для различных вербальных и невербальных элементов коммуникации, таких как жест тела, направление взгляда, движение головы, глаз и губ.

***Газетные корпуса***[3]. Собрание газетных материалов в цифровом виде является богатым источником информации для исследователей в ряде дисциплин гуманитарных и социальных наук и особенно ценно для синхронических и диахронических исследований, начиная от истории, медиа и коммуникации и заканчивая лексикографией, для которой газеты являются богатым источником неологизмов и других лексикографических явлений.

---

[1] Manually annotated corpora. – URL: https://www.clarin.eu/content/manually-annotated-corpora (дата обращения: 01.12.2021).

[2] Multimodal corpora. – URL: https://www.clarin.eu/content/multimodal-corpora (дата обращения: 01.12.2021).

[3] Newspaper corpora. – URL: https://www.clarin.eu/content/newspaper-corpora (дата обращения: 01.12.2021).

***Параллельные корпуса***[1] занимают центральное место в переводоведении и контрастивной лингвистике. Многие из параллельных корпусов доступны через простые в использовании конкорданты, что значительно облегчает изучение межъязыковых явлений. Такие корпуса также являются богатым источником материалов для преподавания языка. Кроме того, параллельные корпуса служат обучающими данными для статистических систем машинного перевода.

***Парламентские корпуса***[2] – это очень важный междисциплинарный языковой ресурс, к которому можно подходить с разных исследовательских позиций, включая не только политологию, но и социологию, историю, психологию, прикладные подходы к лингвистике, например критический дискурс-анализ. Хорошая доступность парламентских процедур в цифровом виде и предоставление прав доступа к публичной информации в странах ЕС побудили ряд национальных и международных инициатив по составлению, обработке и анализу парламентских корпусов.

***Эталонные (референтные)*** корпуса[3] предназначены для предоставления исчерпывающей информации о языке.

«Это должен быть общий корпус широкого охвата языка, и надеюсь, он будет рассматриваться сообществом пользователей как своего рода “стандарт” для этого языка» [5]. Таким образом, референтные корпуса контрастируют со специализированными семействами корпусов (например парламентские корпусы, CMC-корпусы) в том, что они являются всеобъемлющими в отношении жанров. Такие корпуса также принято называть национальными. Эталонные корпуса тоже хорошо аннотированы, как правило отображая богатые морфосинтаксические аннотации.

В изложенном обзоре CLARIN есть также раздел «Устные корпуса»[4], однако эта категория корпусов будет рассмотрена отдельно, в силу большой специфики.

## Банки деревьев (treebanks)

Во многих каталогах и классификациях корпусов в качестве отдельной категории выделяют корпуса с синтаксической и / или семантической разметкой. Их принято называть банками деревьев, поскольку разметка обычно имеет древовидную структуру. Древовидные структуры часто создаются поверх корпуса, который уже был аннотирован тегами части речи. В свою очередь банки деревьев иногда дополняются семантической или другой лингвистической информацией. Банки деревьев могут быть созданы полностью

[1] Parallel corpora. – URL: https://www.clarin.eu/resource-families/parallel-corpora (дата обращения: 01.12.2021).

[2] Parliamentary corpora. – URL: https://www.clarin.eu/resource-families/parliamentary-corpora (дата обращения: 01.12.2021).

[3] Reference corpora. – URL: https://www.clarin.eu/resource-families/reference-corpora (дата обращения: 01.12.2021).

[4] Spoken corpora. – URL: https://www.clarin.eu/content/spoken-corpora-0 (дата обращения: 01.12.2021).

вручную, когда лингвисты аннотируют каждое предложение синтаксической структурой, или полуавтоматически, когда синтаксический анализатор назначает некоторую синтаксическую структуру, которую лингвисты затем проверяют и при необходимости исправляют.

Первым известным крупномасштабным банком деревьев стал проект *The Penn Treebank*[1], положивший начало множеству таких корпусов. В англоязычной Википедии[2] перечисляется около 300 банков деревьев, с указанием языка, способа представления синтаксической или семантической информации, а также условий и лицензии распространения и доступа.

Среди банков деревьев выделяют синтаксические, использующие грамматику структуры фразы, или деревья зависимостей, семантические, а также банки глубинных деревьев, сочетающие синтаксическую и семантическую разметку. В качестве примера семантического банка деревьев можно привести *Groningen Meaning Bank (GMB)*[3], аннотированный с использованием теории представления дискурса, а в качестве банка глубинных деревьев – проект *Deep-sequoia*[4].

Для банков деревьев используются разнообразные инструментальные средства. Инструменты поиска для проанализированных корпусов обычно зависят от схемы аннотаций, примененной к корпусу. Сложность пользовательских интерфейсов варьируется от систем запросов на основе выражений, предназначенных для программистов, до сред полного исследования, предназначенных для лингвистов общего профиля. Подробный обзор инструментов поиска в банках деревьев можно найти в работе С. Уоллиса [6].

## Инструментальные средства корпусной лингвистики

Для создания и использования лингвистических корпусов в мире разработано множество инструментальных средств. Вероятно, наиболее полный их перечень находится на специализированной сайте *Инструменты корпусной лингвистики*[5]. На нем размещены сведения о более чем 250 программных средствах, используемых для обработки и анализа корпусов. О каждом средстве сообщается название, адрес в Интернете, описание, назначение (категория), платформа и условия поставки (цена, бесплатно). Перечень категорий инструментальных средств весьма обширен – свыше 200, причем конкретный инструмент может относиться к нескольким категориям. Примеры категорий: аннотирование, парсеры, теггеры, конкорданты, ключевые слова, визуализация, поиск, токенизация и прочие.

---

[1] The Penn Treebank Project. – URL: https://web.archive.org/web/ 19970614160127/ http://www.cis.upenn.edu/~treebank/ (дата обращения: 01.12.2021).

[2] Treebank. – URL: https://en.wikipedia.org/wiki/Treebank (дата обращения: 01.12.2021).

[3] The Groningen Meaning Bank (GMB). – URL: https://gmb.let.rug.nl/ (дата обращения: 01.12.2021).

[4] Deep-sequoia: A multilayer French corpus. – URL: http://deep-sequoia.inria.fr/ (дата обращения: 01.12.2021).

[5] Tools for Corpus Linguistics. – URL: https://corpus-analysis.com/ (дата обращения: 01.12.2021).

Наиболее известной специализированной системой для управления корпусами является *Sketch Engine*[1]. Эта система позволяет пользователям создавать и работать с более чем 500 текстовыми корпусами на более чем 90 языках. Sketch Engine содержит ряд уникальных инструментов для анализа больших корпусов – до 30 млрд слов. Каждый пользователь может воспользоваться полностью автоматизированной функцией создания словарей. Доступ к Sketch Engine финансируется ЕС в рамках проекта ELEXIS с 2018 по 2022 год. Доступ предоставляется бесплатно академическим учреждениям и наблюдателям ELEXIS и применяется только для некоммерческого использования. В настоящее время системой пользуются более 350 учреждений.

В качестве примера приведем перечень инструментов, которые Sketch Engine предлагает для работы с корпусами русских текстов.

*Извлечение русских словосочетаний.* Словосочетания отображаются в категоризованных списках, чтобы легко идентифицировать сильные и слабые словосочетания.

*Конкордант* может быть использован для отображения списка примеров (называемых конкордансами) поискового слова или фразы в том виде, в каком они появляются в текстовых корпусах русского языка.

*Извлечение русского термина*. Функция Sketch Engine, которая автоматически идентифицирует однословные и многословные термины в предметном русском тексте путем сравнения его с общим корпусом русского языка.

*Извлечение двуязычных терминов.* Параллельные корпуса используются для извлечения терминов на двух языках одновременно и отображения списка терминологии с переводами на другой язык.

*Русский тезаурус.* Тезаурус – это функция, которая автоматически генерирует список слов, сходных по значению с ключевым словом.

*Списки русских слов.* Инструмент будет генерировать частотный список всех слов, которые появляются в тексте или корпусе. Очень большой корпус может быть использован для создания списка всех слов, существующих в русском языке, или всех слов, которые начинаются или заканчиваются определенными символами, или содержат их.

*N-граммы на русском языке.* Список N-грамм, содержащихся в тексте, позволяет выявлять и изучать закономерности и замечать явления, связанные с многословными единицами в русском языке, которые не могут быть обнаружены другими средствами

*Тренды* – диахронический анализ автоматически выявляет неологизмы и изменения в употреблении.

В заключение раздела отметим, что наиболее популярным в настоящее время является корпусный формат *XCES*[2]. Это стандарт на основе XML для кодирования корпусов текстов, который широко используются лингвистами и исследователями естественного языка. XCES в значительной степени основан на предыдущем стандарте кодирования корпуса EAGLES Corpus Encoding Standard (CES), но использует XML в качестве языка разметки. Он

[1] Sketch Engine. – URL: https://www.sketchengine.eu/ (дата обращения: 01.12.2021).

[2] XML Corpus Encoding Standard (XCES). – URL: https://en.wikipedia.org/wiki/XCES (дата обращения: 01.12.2021).

поддерживает простые корпуса, а также аннотированные корпуса, параллельные корпуса и другие.

## Корпусная лингвистика в России

***Краткая история***

Создание корпусов является важным направлением современной российской лингвистики. Оно имеет свою историю: еще в 1980-х годах была начата работа по созданию *Машинного фонда русского языка.* Сейчас материалы, созданные в рамках проекта Машинного фонда, размещены по адресу[1]. Работы по созданию Машинного фонда русского языка были начаты после состоявшейся в 1983 году специальной Всесоюзной конференции, материалы которой позднее были опубликованы в книге [4]. Тогда же был создан отдел Машинного фонда русского языка в Институте русского языка РАН.

Была разработана «Комплексная программа научных исследований и прикладных разработок по созданию Машинного фонда русского языка на 1996–2000 гг. и информатизации исследований в Институте русского языка АН СССР», в основу которой легли упомянутые материалы. Руководителями Отдела были последовательно член-корреспондент АН СССР Ю.Н. Караулов (1985–1991), доктор филологических наук В.М. Андрющенко (1992–1998), профессор, доктор филологических наук А.Я. Шайкевич (1998–2006). Результаты работ по этому проекту были использованы при создании Национального корпуса русского языка (НКРЯ).

Более полное описание российской корпусной лингвистики имеется в работе В.П. Захарова и С.Ю. Богдановой [1], которая выдержала уже три издания.

В 2012–2014 гг. в Российской академии наук под руководством академиков В.В. Иванова и В.А. Плунгяна действовала программа Президиума РАН «Корпусная лингвистика», в рамках которой было реализовано много проектов по созданию и развитию корпусов русского языка и языков народов России. Приведем темы программы «Корпусная лингвистика»:

- создание корпусов миноритарных тюркских языков России;
- создание корпусов на диалектах языков Поволжья;
- корпуса литературных языков Дагестана: лакский и табасаранский языки;
- Корпус бурятского языка;
- корпуса литературных языков Дагестана: аварский и даргинский языки;
- создание корпусов на языках народов Северной Сибири;
- Электронный корпус древнетюркских текстов;
- Национальный корпус калмыцкого языка;
- сбор материалов для Национального и Устного корпусов осетинского языка;

[1] Машинный фонд русского языка. – URL: http://cfrl.ruslang.ru/ (дата обращения: 01.12.2021).

- Устный корпус основных диалектов современного осетинского языка;
- Национальный корпус осетинского языка: расширение и развитие;
- создание корпуса текстов республиканских газет на башкирском языке;
- Корпус вепсского языка: пополнение и развитие электронного ресурса;
- Электронная библиотека литературных языков Дагестана;
- корпуса новописьменных лезгинских языков: агульский и удинский;
- корпуса языков Дальнего Востока;
- развитие и пополнение электронного корпуса фольклорных текстов на языках малочисленных народов Сибири (на материалах ненецкого, телеутского, шорского и эвенкийского языков);
- Татарский корпус текстов.

В настоящее время в области корпусной лингвистики работает много научных коллективов в Москве, Санкт-Петербурге, Казани, Новосибирске, Петрозаводске, Уфе, Ижевске, других городах. Описания многих проектов и ссылки на них собраны на портале *Лингвистические корпуса и сервисы*[1].

***Корпуса русского языка***

Центральным проектом российской корпусной лингвистики является *Национальный корпус русского языка (НКРЯ)*[2], который разрабатывается с 2004 года при участии ряда академических институтов и университетов во главе с Институтом русского языка им. В.В. Виноградова РАН. НКРЯ представляет собой заметное достижение отечественной прикладной лингвистики. Руководитель проекта – академик В.А. Плунгян.

***НКРЯ*** в настоящее время включает следующие подкорпуса:

- Основной корпус
- Газетный корпус СМИ 2000-х гг.
- Газетный региональный корпус
- Диалектный корпус
- Обучающий корпус
- Параллельный корпус
- Поэтический корпус
- Устный корпус
- Акцентологический корпус
- Мультимедийный корпус
- Древнерусский
- Берестяные грамоты
- Старорусский
- Церковнославянский

[1] Лингвистические корпуса и сервисы. – URL: http://web-corpora.net/ (дата обращения: 01.12.2021).

[2] Национальный корпус русского языка. – URL: https://ruscorpora.ru/new/ (дата обращения: 01.12.2021)

Суммарный объем НКРЯ в марте 2021 года составлял:
Число текстов – 2 419 215
Число предложений – 78 694 781
Число словоупотреблений – 961 081 047

Функциональные возможности НКРЯ, так же как принципы разметки отдельных подкорпусов, описаны на сайте НКРЯ, а также в статье В.А. Плунгяна [8].

НКРЯ является, безусловно, лидером российской корпусной лингвистики, но не единственным корпусом русского языка, созданным в России. Перечислим еще несколько российских проектов текстовых корпусов, но корпуса звучащей речи будут описаны в главе 14.

***Общие корпуса современного русского языка***

*Открытый корпус русского языка*[1]– это проект по созданию силами сообщества размеченного корпуса текстов. Корпус будет доступен бесплатно и в полном объеме (под лицензией CC-BY-SA). Создаётся хранилище текстов, специально предназначенное для текстов с лингвистической разметкой, интерфейс редактирования разметки и исправления ошибок, инструменты для контроля качества и стандарт разметки для русского языка.

*Корпус русского литературного языка*[2] задуман как представленный в электронной форме массив морфологически аннотированных текстов на русском литературном языке. Корпус содержит тексты со сбалансированным жанровым составом (художественная проза – не менее 30%, публицистика – не более 30%, научная литература (аналитика и обзоры, научно-популярная) – не более 20%, а также драматические произведения (как некоторое приближение к разговорному языку) – около 20%), насчитывающие чуть больше 1 млн словоупотреблений. Во всех текстах восстановлена в правах буква «ё», и проставлены словесные ударения. В корпус включаются тексты с начала 50-х годов XX века до настоящего времени. На базе корпуса создан частотный словарь словоформ. Идет подготовка морфологически аннотированного варианта корпуса.

*Генеральный интернет-корпус русского языка* (*ГИКРЯ*)[3] – мегакорпус (более 20 млрд слов), созданный при помощи полностью автоматической технологии сбора и разметки текстов из Рунета и основанный на современных достижениях компьютерной лингвистики. Проект осуществляется при технологической и организационной поддержке компании ABBYY.

[1] Открытый корпус. – URL: http://opencorpora.org/ (дата обращения: 01.12.2021).

[2] Корпус русского литературного языка. – URL: http://narusco.ru/ (дата обращения: 01.12.2021).

[3] Генеральный Интернет-корпус Русского Языка. – URL: http://www.webcorpora.ru/ (дата обращения: 01.12.2021).

***Синтаксические корпуса русского языка***

*Тестовый корпус с параллельной синтаксической разметкой*[1] (RSTB) – банк синтаксических деревьев. В нем представлены результаты разбора 64 800 предложений (1 млн словоупотреблений) тремя автоматическими системами синтаксического анализа: SyntAtom, SemSin, Russian Malt. В корпус вошли предложения из текстов разных жанров, включая научную и художественную литературу, а также тексты новостных сообщений. На сайте также представлено 800 предложений из этого корпуса, выбранных случайным образом и размеченных вручную. Это дает возможность сравнения разборов систем с разборами, представленными в эталонном корпусе.

*RUS-Treebank:* корпус с автоматической разметкой синтаксических зависимостей[2].

*UD-Russian:* синтаксически аннотированные корпуса русского языка. Включают:

- корпус электронной коммуникации UD-Russian-Taiga[3];
- корпус фрагментов википедии UD-Russian-GSD[4] (конвертированный корпус Google Stanford Dependencies);
- UD-Russian-SynTagRus[5] (конвертированный корпус SynTagRus, совместный проект Школы лингвистики и Карлова Университета в Праге).

*Корпус риторических структур* [9]. Это – корпус русских текстов, размеченных в теории риторических структур, и предназначенный для исследователей, заинтересованных в изучении письменного дискурса. Корпус позволяет проводить различные эксперименты по автоматическому анализу текста с привлечением данных о дискурсивных связях внутри него. Возможные области применения: генерация текстов, извлечение фактов, автоматическое реферирование, разрешение анафоры и выявление кореферентных цепочек и т.д. В корпусе возможен поиск по словам, словосочетаниям, риторическим отношениям для эффективного исследования письменного дискурса русского языка.

***Учебные корпуса русского языка***

*Русский учебный корпус*[6]: образцы устной и письменной речи изучающих русский язык. В Русском учебном корпусе содержатся образцы устной и письменной речи двух категорий нестандартных говорящих на русском языке: изучающих русский язык как иностранный и так называемых эритажных говорящих.

---

[1] Russian syntax tree bank. – URL: http://otipl.philol.msu.ru/~soiza/testsynt/ (дата обращения: 01.12.2021).

[2] RUS-Treebank. – URL: http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php (дата обращения: 01.12.2021).

[3] UD Russian Taiga. – URL: https://universaldependencies.org/treebanks/ru_taiga/index.html (дата обращения: 01.12.2021).

[4] Universal Dependencies. – URL: http://universaldependencies.org/ (дата обращения: 01.12.2021).

[5] SynTagRus. – URL: https://universaldependencies.org/treebanks/ru_syntagrus/index.html (дата обращения: 01.12.2021).

[6] RLC Русский учебный корпус. – URL: http://web-corpora.net/RLC (дата обращения: 01.12.2021).

*Корпус русских учебных (академических) текстов (КРУТ)*[1]: коллекция текстов на русском языке, написанных студентами разных вузов. Общий объем корпуса составляет около 3,1 млн слов. Тексты сопровождаются несколькими типами разметки (метатекстовой, морфологической и разметкой по ошибкам), что позволяет осуществлять поиск по корпусу. КРУТ является информационно-справочной системой, предназначенной для исследователей, преподавателей, студентов, а также для всех, кто интересуется проблемами современной русской грамматики, актуальными процессами в области лексики, морфологии и синтаксиса современного русского языка.

*Корпус региональных вариантов русского языка*[2] базируется на социолингвистических интервью, собранных в разных селах Дагестана. Эти тексты, записанные от носителей даргинского, аварского, кумыкского, лакского, арчинского, табасаранского и других языков, расшифровываются и выравниваются со звуком в программе Praat, чтобы пользователю был доступен как звук, так и текст в письменном виде. Кроме того, размечаются типичные грамматические особенности этой *речи.*

*Russian Learner Translator Corpus*[3] (LTC) – это двунаправленный корпус параллельных англо-русских текстов, который содержит предложения исходных текстов, согласованные с их многочисленными целями, созданные студентами-переводчиками в 14 российских университетах. Значительную часть корпуса составляют переводы англоязычных текстов СМИ, выполненные студентами старших курсов переводческих специальностей.

***Диалектные и диахронические корпуса русского языка***

*Электронные базы данных по русским народным говорам*[4]. Лингвистическая информация в базе организована по многоступенчатому принципу. Выделяется девять уровней членения письменного текста; на каждом из них выделяется своя основная (базовая) единица членения.

*Издания Археографической комиссии [10]*. Коллекция Президентской библиотеки им. Б.Н. Ельцина.

*Древнерусские берестяные грамоты*[5]. Основу сайта составляет база данных, включающая фотографии берестяных грамот, их прориси, древнерусские тексты, переводы на современный русский язык и основную информацию о документах. База данных является частью более обширной информационной системы, содержащей полную археологическую информацию о документах и текстовый корпус с морфологической разметкой.

[1] CoRST. Corpus of Russian Student Texts. – URL: http://web-corpora.net/learner_corpus (дата обращения: 01.12.2021).

[2] Корпус региональных вариантов русского языка. – URL: https://ling.hse.ru/regional_corp (дата обращения: 01.12.2021).

[3] Russian Learner Translator Corpus (Russian LTC) или «Корпус несовершенных переводов». – URL: http://rus-ltc.org/search (дата обращения: 01.12.2021).

[4] Электронные базы данных по русским народным говорам. – URL: http://www.ruslang.ru/krylov_dialect (дата обращения: 01.12.2021).

[5] Древнерусские берестяные грамоты. – URL: http://gramoty.ru/birchbark/about-site/ (дата обращения: 01.12.2021).

Лингвистическая составляющая этой системы доступна в составе Национального корпуса русского языка.

*Санкт-Петербургский корпус агиографических текстов* (*СКАТ)*[1]. На данном сайте вы можете ознакомиться с общей информацией о проекте, получить информацию о способе представления текста, загрузить введенные жития в pdf- и xml-формате, ознакомиться со списком публикаций и воспользоваться поиском по электронному словоуказателю. Для представления житийных текстов в корпусе разработан оригинальный компьютерный шрифт, используемый для воспроизведения основных параметров рукописей. Для просмотра и поиска в текстовой базе житий строятся словоуказатели с указанием адресов вхождений словоформ в виде номеров рукописных страниц и строк. К настоящему времени в базу данных введено более 50 рукописей общим объемом около 500 тыс. словоупотреблений.

*Корпус «Манускрипт»*[2] Удмуртского государственного университета. Основными модулями для работы с коллекцией являются:

- сайты и запросные формы, позволяющие познакомиться с текстами, указателями и осуществить выборку данных;
- специализированный редактор для ввода, редактирования и фрагментирования текстов;
- модуль выборок и запросов, позволяющий подготовить данные для лингвистических, палеографических и текстологических исследований;
- морфологический анализатор для автоматического анализа и синтеза словоформ древнерусского языка;
- модуль грамматических словарей для ввода, редактирования и согласования словарных материалов.

*Корпус русских публицистических текстов второй половины XIX века*[3] Петрозаводского государственного университета.

***Прочие корпуса***

- *Симфония текстов*[4]. Параллельный корпус переводов «Слова о полку Игореве»
- *Русский корпус биографических текстов*[5]
- *База данных русской прессы «Integrum»*[6]
- *Компьютерный корпус текстов русских газет конца XX века*[1]

[1] Санкт-Петербургский корпус агиографических текстов. – URL: http://project.phil.spbu.ru/scat/page.php?page=project (дата обращения: 01.12.2021).

[2] Манускриптъ. Славянское Письменное наследие. – URL: http://mns.udsu.ru/ (дата обращения: 01.12.2021).

[3] Статистические методы анализа литературного текста (ИС «СМАЛТ»). – URL: http://smalt.karelia.ru/ (дата обращения: 01.12.2021).

[4] Параллельный корпус переводов «Слова о полку Игореве». – URL: http://cfrl.ruslang.ru/slovo/index.htm (дата обращения: 01.12.2021).

[5] Корпус биографических текстов = Russian Corpus of Biographical Texts. – URL: https://www.sites.google.com/site/utcorpus/ (дата обращения: 01.12.2021).

[6] Integrum World Wide– URL: http://www.integrumworld.com/rus/about.html

- *Русско-французский поэтический корпус первой трети XIX в.*[2]

***Корпуса языков народов России***

Крупнейшим собранием корпусов языков народов России является собрание, представленное на платформе LINGVODOC[3]. Количество корпусов указано в квадратных скобках в списке после названия языка или языковой семьи:

Алтай-кижи [1], Алтайские языки [54], Алтайский [23], Башкирский [5], Казахский [2], Камасинский [5], Карельский [2], Коми [1], Мансийский [4], Марийский [1], Мокшанский [2], Нанийская группа / Южнотунгусские языки [2], Негидальский [1], Прибалтийско-финские [2], Саамские [2], Самодийские [54], Селькупский [48], Татарский [2], Тунгусо-маньчжурские языки [3], Тунгусские языки [2], Тюркские языки [51], Удмуртский [3], Уральские [84], Хакасский [7], Хантыйский [7], Челканский [1], Чувашский [1], Чулымский [8], Энецкий [1], Эрзянский [8].

Кроме того, в различных российских информационных системах имеются следующие корпуса языков народов России:

*Бурятский корпус*[4]. Текстовая база Бурятского корпуса (БК) стала включать образцы общественно-публицистического (тексты интернет-версий СМИ, журнальной периодики) и учебно-научного стилей (научные статьи). Объем корпуса достиг более 2 млн 200 тыс. словоупотреблений, зарегистрированных в письменных текстах в основном художественного стиля с их метаописанием. Метаописание текстов включает их основные библиографические и классификационные характеристики. Корпус обеспечен начальной морфологической разметкой входящих в него слов на основе словоизменительных характеристик.

*Калмыцкий корпус*[5]. Литературные тексты на калмыцком языке – романы, повести, рассказы, очерки, газетные статьи (вторая половина XX – начало XXI в.), включенные в корпус, снабжены морфологической разметкой, позволяющей осуществлять поиск по лексеме, по грамматическим характеристикам, по переводу, а также по сочетаниям словоформ на заданном расстоянии. В настоящее время расширяется текстовая база корпуса, осуществляется пополнение грамматического словаря и морфологическая разметка текстов. Объем корпуса – 800 тыс. словоупотреблений.

*Татарский национальный корпус «Туган тел»*[6]. Объем корпуса на конец 2018 года составлял свыше 180 млн словоупотреблений. Корпус содержит

---

[1] Компьютерный корпус текстов русских газет конца XX века. – URL: https://www.philol.msu.ru/~lex/corpus/ (дата обращения: 01.12.2021).

[2] Русско-французский поэтический корпус первой трети XIX в. – URL: http://www.nevmenandr.net/fr/ (дата обращения: 01.12.2021).

[3] LINGVODOC. – URL: http://lingvodoc.ru/corpora_all (дата обращения: 01.12.2021).

[4] Бурятский корпус. – URL: http://web-corpora.net/BuryatCorpus/ search/index.php?interface_language=ru (дата обращения: 01.12.2021).

[5] Калмыцкий корпус. – URL: http://web-corpora.net/KalmykCorpus/search/ index.php?interface_language=ru (дата обращения: 01.12.2021).

[6] Татарский национальный корпус «Туган тел». – URL: http://www.tugantel.tatar/ (дата обращения: 01.12.2021).

тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.). Каждый документ имеет метаописание (авторы, их пол, выходные данные, даты создания, жанры, части, главы и др.). Тексты, включенные в корпус, снабжены морфологической разметкой (информация о части речи и грамматических характеристиках словоформы). Морфологическая разметка текстов корпуса выполняется автоматически с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO.

*Крымско-татарский корпус*[1]. «Лингвистический корпус крымско-татарского языка» – корпус современного письменного крымско-татарского языка. В состав корпуса входят преимущественно тексты из крымско-татарских газет начала XXI века. В настоящее время корпус содержит 521 тыс. токенов (включая пунктуацию), что составляет около 57 тыс. словоформ.

*Корпус удмуртского языка*[2]. В настоящий момент размер корпуса составляет около 7,3 млн словоупотреблений. Тексты корпуса были размечены с помощью автоматического морфологического анализатора, около 88% словоформ корпуса имеют грамматический разбор.

*Машинный фонд башкирского языка*[3] представляет собой автоматизированную информационную систему, состоящую из генеральной картотеки, лексикографической, экспериментально-фонетической, грамматической, диалектологической базы и каталога рукописных и старопечатных книг.

*Башкирский поэтический корпус*[4]. Объём более 1,8 млн словоупотреблений, около 450 тыс. стихотворных строк, более 17 тыс. стихотворений 101 поэта.

*Корпус вепсского языка*[5]. Создатель – ИЯЛИ КарНЦ. Вепсский корпус позволяет строить конкордансы для лексем и конкретных словоформ. Вепсский корпус может обеспечить частное и сопоставительное языкознание источником лингвистических данных по вепсскому языку – прибалтийско-финскому языку с богатой и своеобразной грамматикой, повысить возможность независимой проверки примеров из вепсского языка, обогатить инструментарий лингвистики.

*Электронный корпус текстов тувинского языка*[6] включает базу данных тувинских текстов современного и советского периода, а также базы данных грамматических форм (аффиксов, аналитических конструкций) и

---

[1] Лингвистический корпус крымско-татарского языка. – URL: https://korpus.sk/QIRIM/ (дата обращения: 01.12.2021).

[2] Корпус удмуртского языка. – URL: http://web-corpora.net/UdmurtCorpus/ search/?interface_language=ru (дата обращения: 01.12.2021).

[3] Машинный фонд башкирского языка. – URL: http://mfbl2.ru/ (дата обращения: 01.12.2021).

[4] Башкирский поэтический корпус. – URL: http://web-corpora.net/bashcorpus/search/?interface_language=ru (дата обращения: 01.12.2021).

[5] Корпус вепсского языка. – URL: http://vepsian.krc.karelia.ru (дата обращения: 01.12.2021).

[6] Электронный корпус текстов тувинского языка. – URL: http://tuvancorpus.ru/?q=content/ korpusy-po-tyurkskim-yazykam (дата обращения: 01.12.2021).

первичных основ (именных и глагольных) на основные типы слогов. На основе этих баз данных будет создан электронный словарь частотных лексем тувинского языка. Предполагается создание и применение на практике компьютерных программ для автоматизации сбора и обработки материала для лингвистических исследований в области тувинского языка.

*Адыгейский корпус*[1]. Веб-интерфейс корпуса обеспечивает поиск по собранию адыгейских текстов с учетом грамматической разметки. Доступ к полным текстам не предоставляется. Корпус дает возможность поиска по последовательностям букв, морфемам и их сочетаниям, грамматическим признакам словоформ, переводам, позволяет учитывать расположение орфографического слова в предложении. Допускается ограничение по жанрам текстов. Грамматический анализ словоформ выполнен автоматически и не выверен вручную; создатели корпуса не несут ответственности за правильность всех разборов. В таблице кратко представлены основные характеристики корпуса. На ноябрь 2018 года корпус включал около 8 млн словоупотреблений.

*Электронный корпус хакасского языка*[2]. Существующий объем материалов по тюркским языкам России нуждается в компьютеризации и обеспечении общего доступа к нему, т.е. в создании открытого корпуса тюркских языков России. Открытость корпуса должна обеспечить не только дальнейшее изучение этих языков, но и внести вклад в дело их сохранения и развития. В рамках проекта предполагается делать параллельные корпуса (все тексты обеспечены русским переводом) с морфологической (в дальнейшем и синтаксической) разметкой. Материалом для корпуса хакасского языка служат в первую очередь параллельные (хакасско-русские) литературные тексты художественного жанра и эпические тексты, оцифрованные и приведенные к стандартному формату.

*Корпуса уральских языков Поволжья*[3]. Стартовая страница корпусов нескольких уральских языков Поволжья объемом от 14 тыс. до 9,5 млн словоупотреблений. Языки, для которых доступны корпуса:

- горный марийский
- коми-зырянский
- коми-пермяцкий
- луговой марийский
- мокшанский
- удмуртский
- эрзянский

Все корпуса содержат автоматическую морфологическую разметку и переводы лемм на русский язык. Для каждого из языков доступны корпус соцсетей и «основной» корпус (все остальное; в основном пресса). Корпуса соцсетей содержат также и тексты на русском языке, написанные теми же авторами или в тех же группах.

---

[1] Адыгейский корпус. – URL: http://adyghe.web-corpora.net/ (дата обращения: 01.12.2021).

[2] Электронный корпус хакасского языка. – URL: https://khakas.altaica.ru/ (дата обращения: 01.12.2021).

[3] Корпуса уральских языков Поволжья. – URL: http://volgakama.web-corpora.net/ (дата обращения: 01.12.2021).

*Языки Обско-Енисейского языкового ареала*[1]. В проект вошли фольклорные тексты, собранные в период 1960–1970 годов и хранящиеся в архивах кафедры-лаборатории языков народов Сибири Томского государственного педагогического университета. Языковой материал представлен максимально полно, как в современной академической транскрипции, так и в оригинальной нотации на основе кириллицы, сделанной исследователями, осуществившими запись данных текстов. Архивные тексты проекта подверглись подробному современному лингвистическому анализу, унифицированной транскрипции, морфемному глоссированию, морфемному и свободному переводу на русский и английский язык, выполненному авторами. Языки проекта:

- восточные диалекты хантыйского языка
- южный, центральный диалект селькупского языка
- северный диалект кетского языка
- авамский диалект нганасанского языка
- средний диалект чулымско-тюркского языка

*Корпуса текстов на языках малочисленных народов Сибири*[2]. Предполагалось включать фольклорные материалы лишь на двух языках: шорском и эвенкийском. Объем шорского и эвенкийского подкорпусов к концу 2011 года пранировалось довести до примерно 120 000 и 6 000 словоупотреблений соответственно. С целью демонстрации возможностей корпуса в плане включения материалов на других языках в нем дополнительно размещены фольклорные тексты на телеутском языке. В настоящий момент «корпусная машина» обслуживает следующие корпуса:

- ненецкий
- телеутский
- шорский
- эвенкийский

В заключение отметим, что российскими корпусными лингвистами созданы также корпуса для других языков, не входящих в число языков народов России. Среди них тайский[3], амхарский[4] и ряд других языков, информация о которых собирается в рамках международного проекта по Универсальным зависимостям[5].

[1] Языки Обско-Енисейского языкового ареала. – URL: https://siblang.tspu.ru/project09/RUSS/#:~:text=Языковой%20и%20культурный%20материал%20проекта,и%20северо-кетский%20диалект%20кетского%20языка (дата обращения: 01.04.2022).

[2] Корпуса текстов на языках малочисленных народов Сибири. – URL: https://corpora.iea.ras.ru/corpora/ (дата обращения: 01.12.2021)

[3] HSE Thai Corpus. – URL: http://web-corpora.net/ThaiCorpus/search/?interface_language=ru

[4] Корпус амхарского языка. – URL: http://web-corpora.net/AmharicCorpus/search/?interface_language=ru (дата обращения: 01.12.2021).

[5] Universal Dependencies (UD). – URL: https://universaldependencies.org/ (дата обращения: 01.12.2021).

## Литература к главе 10


1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Санкт-Петербург : Издательство СПбГУ, 2019. – ISBN 978-5-288-05997-1.
2. Atkins, Clear, Ostler. Corpus design criteria // Literary & Linguistic Computing. – 1992, N 7(1). – P. 1–16.
3. Плунгян В.А. Современные применения корпусных технологий // Видеотрансляция общего собрания Отделения историко-филологических наук РАН. – URL: http://hist-phil.ru/events/427/ (дата обращения: 01.12.2021).
4. Fišer D., Lenardič J. CLARIN, Resource Families // European Union's Horizon 2020. – URL: https://www.clarin.eu/resource-families (дата обращения: 01.12.2021).
5. Leech G. The Importance of Reference Corpora // UZEI Hizkuntza-corpusak. Oraina eta geroa (2002-10-24/25). – URL: https://www.uzei.eus/wp-content/uploads/2017/06/06-Geoffrey-LEECH.pdf (дата обращения: 01.12.2021).
6. Wallis S. Searching treebanks and other structured corpora. Chapter 34 // Lüdeling Corpus Linguistics: An International Handbook. Handbücher zur Sprache und Kommunikationswissenschaft series / A. & Kytö M. (ed.). – Berlin : Mouton de Gruyter, 2008. C. 66–79.
7. Машинный фонд русского языка: идеи и суждения. – Москва : Наука, 1989. – 239 с.
8. Плунгян В.А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. – Москва : Индрик, 2005. – С. 6–20.
9. Kobozeva M. Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // June 2016 Conference: DIALOGUE-2016. 22 nd International Conference on Computational Linguistics and Intellectual Technologies. – URL: https://www.researchgate.net/publication/311618194 (дата обращения: 01.12.2021).
10. Президентская библиотека им. Б.Н. Ельцина. Издания Археографической комиссии. – URL: https://www.prlib.ru/collections/689008 (дата обращения: 01.04.2022).

# ГЛАВА 11. ЛЕКСИЧЕСКИЕ РЕСУРСЫ И КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ

## Определение, классификация, статистика

Создание и применение лексических ресурсов является одной из центральных задач компьютерной лингвистики, для ее решения сформировалось целое направление компьютерной лингвистики, которая называется компьютерной лексикографией.

Основной объект этой дисциплины – электронные (компьютерные, автоматические и др.) словари, которые могут быть загружены в базу данных и обработаны с помощью прикладного программного обеспечения. Электронный словарь может быть словарем с собственной структурой, который поддерживается специальным программным обеспечением (например, онлайн через Интернет), или это может быть словарь, который имеет открытую структуру и доступен для загрузки в компьютерные базы данных и, таким образом, может использоваться с помощью различных программных приложений.

Википедия предлагает такое определение: «Электронный словарь – это любой упорядоченный, относительно конечный массив лингвистической информации, представленный в виде списка, таблицы или перечня, удобного для размещения в памяти ЭВМ и снабженного программами автоматической обработки и пополнения»[1].

Различают электронные словари пользователя-человека и словари для программ обработки текста. Электронные словари, предназначенные для человека, по интерфейсу и структуре словарной статьи существенно отличаются от словарей, включенных в системы машинного перевода, системы автоматического реферирования, информационного поиска и т.д.

Автоматические словари, предназначенные для конечного пользователя, чаще всего являются компьютерными версиями хорошо известных обычных словарей. Они практически повторяют структуру словарной статьи обычных словарей, однако они обладают функциями, недоступными своим прототипам.

Автоматические словари для систем машинного перевода, автоматического реферирования, информационного поиска и т.д. по интерфейсу и структуре словарной статьи существенно отличаются от пользовательских.

[1] Компьютерная лексикография. – URL: https://ru.wikipedia.org/wiki (дата обращения: 01.12.2021).

Особенности их структуры, сфера охвата словарного материала задаются теми программами, которые с ними взаимодействуют [1].

Пока не существует общепринятой классификации лексических ресурсов, хотя попыток такого рода сделано очень много. Укажем, например, на работы [2; 3]. Дело в том, что при классификации электронных словарей часто делается попытка перенести на них типологию традиционных словарей, которая достаточно разнообразна и многомерна. Так, в исследовании Л. Поповой [3] проведен анализ различных типологий словарей, при этом в разных классификациях выявлено свыше 100 дифференцирующих признаков. Однако для электронных словарей такая детальная классификация не является релевантной, поскольку в большинстве случаев различие типов и назначений словарей не влияет на их структуру в электронном виде.

Действительно, если мы посмотрим на классификации лексических ресурсов в крупнейших мировых архивах и справочных системах ЛИР, то мы увидим, что если вынести за скобки языки, к которым относятся словари, классификация лексических ЛИР очень проста.

Так, в каталоге ELRA[1] выделено четыре типа лексических и / или концептуальных ЛИР:

- терминологические ресурсы
- лексиконы
- онтологии
- машиночитаемые ресурсы

В обзоре семейств ЛИР, который предлагает CLARIN[2], выделено пять типов лексических ресурсов, которым предложены следующие толкования.

- *Лексиконы* в основном используются в NLP-приложениях. Они обычно содержат обширный лексический запас с конкретной лингвистической информацией (например, морфосинтаксис).
- *Словари* были созданы в основном для использования человеком (например для изучения языка, перевода, лексикологии) и, как правило, являются семасиологическими, что означает, что они организованы вокруг слов и содержат информацию об их значениях, определениях, произношении и т.д.
- *Концептуальные ресурсы* включают ономасиологические лексические ресурсы, такие как словарные сети, фреймовые сети, тезаурусы и онтологии. Такие ресурсы обычно связаны семантическими отношениями (например гипернимия, гипонимия).
- *Глоссарии* – это специализированные словари, содержащие специфичную для данной предметной области терминологию и / или выражения.
- *Списки слов* – это лексические ресурсы, которые предоставляют только алфавитные или частотные лексические списки.

[1] Catalogue of Language Resources. – URL: http://www.elra.info/en/catalogues/catalogue-language-resources/ (дата обращения: 01.12.2021).

[2] Resource Families. – URL: https://www.clarin.eu/resource-families (дата обращения: 01.12.2021).

Отметим, что в поисковых системах по ЛИР даже такая простая классификация не используется. Для оценки общего количества лексических ресурсов мы провели поиск в трех крупнейших лингвистических поисковых системах: VLO[1], OLAC[2], TLA[3]. Для поиска использовались такие запросы: lexicon, lexical, dictionary, thesaurus, glossary. Результаты поиска представлены в таблице 6.

*Таблица 6*

**Статистика лексических ресурсов по трем крупнейшим мировым архивам**

| Архив | Всего ЛИР | lexicon | lexical | dictionary | thesaurus | glossary |
|---|---|---|---|---|---|---|
| VLO | 1 204 730 | 3429 | 3699 | 9559 | 86 | 114 |
| OLAC | 4 063 28 | 12 547 | 4020 | 4150 | 518 | 233 |
| TLA | 146 648 | 14 937 | 6297 | 2516 | | 35 |

Из данных этой таблицы можно сделать вывод, что лексические ЛИР занимают второе место среди ЛИР после корпусов. Добавим, что на портале LINGUIST List[4] приводится обширный (около 300) перечень словарных сайтов, порталов, а также агрегаторов лексикографических ЛИР. Из этого количества около 200 ЛИР – это двуязычные и многоязычные словари.

## Международное сотрудничество по электронной лексикографии

***Европейская сеть электронной лексикографии*** **(ENEL)**[5]

Целью ENEL является расширение, координация и гармонизация европейских исследований в области электронной лексикографии и обеспечение легкого доступа к авторитетной информации о языках Европы. Сеть будет решать следующие задачи:

- создание европейского портала словарей. Этот портал будет служить центральной точкой отсчета для всех пользователей словарей, которые ищут надежную, авторитетную словарную информацию по европейским языкам и их истории в Интернете;

[1] CLARIN Virtual Language Observatory. https://www.clarin.eu/content/virtual-language-observatory-vlo – URL: (дата обращения: 01.04.2022).

[2] Open Language Archives Community. – URL: http://olac.ldc.upenn.edu/ (дата обращения: 01.12.2021).

[3] The Language Archive Max Planck Institute for Psycholinguistic. – URL: https://archive.mpi.nl/tla/ (дата обращения: 01.12.2021).

[4] LINGUIST List. – URL: https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=16 (дата обращения: 01.12.2021).

[5] European network of e-lexicography. – URL: https://www.elexicography.eu/ (дата обращения: 01.12.2021).

- обеспечение сотрудничества и обмена ресурсами, технологиями и опытом в области электронной лексикографии, а также поддержки словарей, которые еще недоступны в Интернете;
- внедрение стандартов для инновационных электронных словарей, которые полностью используют возможности цифровых носителей;
- установление новых способов представления общего наследия европейских языков путем разработки общих редакционных практик и объединения уже существующей информации.

В составе ENEL действуют следующие рабочие группы.

WG1*: Интегрированный интерфейс с европейским словарным содержанием*. Задачи:

- создание портала европейских словарей;
- возможности связывания содержания европейских словарей;
- требования пользователей в отношении содержания словарей;
- участие пользователей в создании содержания словарей.

WG2: *Оцифровка словарей.* Задачи:

- стандарты кодирования информации для печатных словарей;
- программное обеспечение для преобразования словарной информации;
- повышение доступности и совместимости содержания словарей;
- план работ по оцифровке, включая оценку затрат;
- возможности использования содержания словарей для компьютерных лингвистических приложений;
- организация школы обучения стандартным инструментам и методам оцифровки словарей.

WG3: *Инновационные электронные словари*. Задачи:

- описание рабочего процесса для корпусной лексикографии;
- обзор существующего ПО для корпусной лексикографии;
- словарные системы письма;
- анализ методов автоматического сбора лексических данных;
- анализ интерфейса между словарем и компьютерной лексикой (см. словарные сети) и синтаксически и семантически аннотированными корпусами (см. FrameNet, SemCor, Senseval);
- исследование возможности использования словарного содержания для компьютерных лингвистических приложений;
- организация школы обучения инновационным подходам в электронной лексикографии.

WG4: *Лексикография и лексикология с паневропейской перспективой.* Задачи:

- разработать способы отображения и связывания информации из одноязычных словарей, чтобы более адекватно представить их общее европейское наследие;
- разработать принципы для интеграции европейской информации в более традиционные и инновационные словари;
- найти новые приложения для взаимосвязанной словарной информации на Европейском словарном портале в области цифровых гуманитарных наук.

***Европейская ассоциация лекскографии EURALEX***[1]

Это ведущая профессиональная ассоциация для людей, работающих в области лексикографии и смежных областях. В быстро меняющемся мире языкового анализа и описания языков EURALEX предоставляет форум для обмена идеями. Несмотря на то, что EURALEX базируется в Европе, она имеет всемирный охват и всемирное членство. В ее состав входят лексикографы, издатели справочников, корпусные лингвисты, компьютерные лингвисты, ученые, работающие в соответствующих дисциплинах, разработчики программного обеспечения и все, кто проявляет живой интерес к языку.

Перечислим также некоторые другие международные структуры в данной области по данным Международного проекта Globalex[2]:

AFRILEX – Африканская ассоциация лексикографии
ASIALEX – Азиатская ассоциация лексикографии
AustraLex – Австралийская ассоциация лексикографии
DSNA – Словарное общество Северной Америки
NFL – Скандинавская ассоциация лексикографии
Pangaealex – Пангейская ассоциация лексикографии
Sealex – Лексикография Юго-Восточной Азии

Укажем, что наиболее полный перечень интернет-ресурсов, связанных с лексикографией, составлен Р. Гартманом в 2010 году [4].

***Электронный журнал словарей Dictionaria***[3]

Это журнал открытого доступа, который публикует высококачественные словари языков со всего мира, особенно языков, которые не имеют большого количества носителей. Список словарей, опубликованных в Dictionaria, доступен по отдельному адресу[4]. Dictionaria издает электронные словари в электронном формате, которые могут быть связаны путем сравнения их значений с другими словарями и словарными коллекциями. Словари представлены в виде баз данных, состоящих из различных таблиц реляционной базы данных (записи, смыслы, примеры). Словари легко доступны для поиска (по лемме, значению, семантической области и другими способами), они легко экспортируются и в них может быть включен медиаконтент (изображение и звук). Dictionaria предлагает пользователям веб-приложение и руководство для представления лексикографической информации, которое будет рассмотрено ниже.

## Функциональность электронных словарей

При обсуждении проблем компьютерной лексикографии большую известность получила публикация В. Селегея «Электронные словари и ком-

[1] European Association for Lexicography EURALEX. – URL: https://euralex.org/ (дата обращения: 01.12.2021).
[2] Globalex. – URL: https://globalex.link/ (дата обращения: 01.12.2021).
[3] Dictionaria. – URL: https://dictionaria.clld.org/ (дата обращения: 01.12.2021).
[4] Dictionaries. – URL: https://dictionaria.clld.org/contributions (дата обращения: 01.12.2021).

пьютерная лексикография» [5], в которой он обсуждает перспективы этой дисциплины. Приведем выдержку из этой работы.

«К новым возможностям электронного словаря относятся:

1. Существенно более изощренные возможности показа содержания словарной статьи, включая возможность частичного показа по разным критериям (различные “проекции” словаря), разнообразные графические средства, которые не используются в обычных словарях.

2. Использование для доступа к содержанию различных лингвистических технологий, таких как морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.

С точки зрения пользователя смысл реализации в электронном словаре всех этих технологий состоит в том, что становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме. При традиционном подходе минимальной единицей доступа является лексема (имя словарной статьи): мы должны прочитать всю статью, чтобы определить, содержится ли в ней ответ на наш запрос. Для таких словарей, как оксфордский, это представляет серьезную проблему. Например, глагол *set* имеет там 400 только основных значений (и у многих из них имеются подзначения). Пользователь хотел бы, чтобы словарь максимально локализовал релевантную информацию. При этом речь не идет об автоматическом выборе переводного эквивалента (если мы говорим о переводном словаре). Специфика словарного ответа в том, что он дает весьма разнообразную информацию о слове или словосочетании, а не просто переводное соответствие, предполагает активный выбор пользователя из нескольких возможных хорошо обоснованных альтернатив. Однако попытка решить проблему адекватной реакции словаря на запрос неизбежно наталкивается на сопротивление самого словарного материала, перенесенного из бумажного словаря».

И далее:

«Отрыв лексикографической теории от лексикографической практики велик. Это должно быть особенно обидно для российской лингвистической науки, в которой лексическая семантика занимает особое место. Достаточно назвать такие имена, как Мельчук, Апресян, Падучева и многие другие (…). При этом в массовых бумажных словарях никаких следов этих идей вы не обнаружите. И именно в развитии этих идей мы видим будущее практической компьютерной лексикографии».

Очевидно, что развитие компьютерной лексикографии подтвердило выводы В. Селегея – функциональные возможности современных электронных словарей постоянно растут, и в них находят реализацию лучшие достижения лексической семантики. Процитируем, например, один из интернет-курсов[1]:

«По сравнению с печатными аналогами компьютерные словари предоставляют пользователю множество дополнительных возможностей:

[1] Компьютерная лингводидактика. – URL: http://bookzooka.com/book/179-kompyuternaya-lingvodidaktika-avtor-neizvesten/18-32-prikladnye-i-instrumentalnye-programmy.html (дата обращения: 01.12.2021).

- многократное увеличение скорости поиска;
- множество входов в словарь: словник, алфавитный индекс, ввод слова и словосочетания с клавиатуры, из текстового редактора;
- поиск слов с недостаточно точным правописанием;
- полнотекстовый поиск (не только в словнике, но и в текстах всех словарных статей);
- применение средств мультимедиа для семантизации лексики;
- наличие системы гиперссылок;
- наличие перекрестных ссылок ко всем словам, имеющимся в словаре;
- возможность хранения большого объема информации;
- в двуязычных словарях – возможность прямого и обратного перевода;
- включение в структуру компьютерного словаря нескольких словарей разных типов и жанров;
- одновременный поиск сразу в нескольких словарях;
- ограничение области поиска ключевыми словами, тематическими группами, частями речи и т.п.;
- пополнение словаря пользователем, или создание пользовательского словаря;
- сохранение последовательности поиска в течение сеанса работы (так называемая хронология / история поиска);
- сохранение «закладок» в словаре;
- совместимость с текстовыми редакторами, возможность копирования словарных статей и обращения к словарю из редактора;
- совместимость с программами машинного перевода;
- совместимость с веб-браузерами и другими типами программ (прикладными, обучающими, игровыми);
- предоставление дополнительной справочной информации по фонетике, грамматике, стилю и другим аспектам языка;
- возможность использования словарей в локальной и глобальной сетях и др.».

Конечно, далеко не все словарные системы предоставляют подобный набор возможностей. Обычно их функциональность гораздо скромнее. В качестве примера приведем перечень функций отечественной информационной системы «Словари», разработанной в ИРЯ РАН[1]:

- одновременный поиск по сотням тысяч словарных статей во всех представленных словарях;
- возможность задавать сложные поисковые запросы с использованием масок и логических операторов;
- запросы с использованием транслитерации;
- отображение результатов в графическом режиме;
- возможность задать запрос в современном написании и получить результат словарной статьи в ее оригинальном написании;

[1] SLOVARI.RU. – URL: http://slovari.ru/start.aspx?s=0&p=3050 (дата обращения: 01.12.2021).

- постоянное обновление и пополнение словарных баз данных и справочных материалов по русскому языку;
- возможность формирования поисковых запросов в библиотеке справочной литературы;
- поиск слова в многолетних архивах Службы русского языка и форума сайта.

### Концептуальные лексико-семантические ЛИР

Одним из перспективных направлений развития компьютерной лексикографии является формирование словарей, отражающих семантику естественного языка. Действительно, общедоступный семантический словарь в электронном виде сегодня входит в стандартный набор необходимых инструментов и ресурсов для автоматической обработки текстов на конкретном языке (наряду с морфологическим анализатором, синтаксическим парсером, большим аннотированным корпусом и т.п.).

Наиболее известным и общепринятым для широкого класса лексико-семантических ЛИР образцом является проект *Princeton WordNet (PWN),* работа над которым началась в 1986 году. Принципы разработки PWN, методы автоматического пополнения тезауруса, а также некоторые приложения на основе данного ресурса описаны в книге [6]. На сегодняшний день WordNet-подобные ресурсы созданы для многих языков. Их обзор можно найти в монографии Н.В. Лукашевич [7].

В настоящее время *WordNet*-подобными словарями называют лексические базы, построенные по базовым принципам проекта PWN. В частности, они состоят из синсетов (synset, от synonym set) – «смыслов», которые выражаются набором квазисинонимов. В свою очередь, синсеты связаны между собой различными семантическими отношениями: гипероним – гипоним, мероним – холоним и др. В PWN входят значения существительных, глаголов, прилагательных и наречий. Текущая версия PWN содержит более 117 тыс. синсетов, которым соответствуют примерно 150 тыс. различных словарных входов (отдельных слов и фраз). PWN успешно используется для решения широкого круга задач: снятия лексической неоднозначности, автоматического реферирования, семантического поиска, классификации и кластеризации документов, обработки поисковых запросов, машинного перевода и т.д.

Конечно, WordNet-подобные словари являются далеко не единственным типом концептуальных ЛИР. К ним также относятся разнообразные таксономии, классификации и – как наиболее общий тип лексико-семантических ЛИР – онтологии.

В качестве примера лексико-семантического ЛИР укажем на базу данных *FrameNet*[1]. Эта база данных содержит более 1200 семантических фреймов, 13 000 лексических единиц (соединение слова со значением; многозначные слова представлены несколькими лексическими единицами) и 202 000 примеров предложений. *FrameNet* – в значительной степени создание Чарльза

[1] FrameNet. – URL: https://framenet.icsi.berkeley.edu/fndrupal/ (дата обращения: 01.12.2021).

Дж. Филмора, который разработал теорию семантики фреймов, на которой основан проект, и первоначально был руководителем проекта, когда проект начался в 1997 году. Проект *FrameNet* оказал влияние как на лингвистику, так и на обработку естественного языка, где он привел к задаче автоматической семантической разметки ролей.

Для широкого класса лексико-семантических ЛИР разработан международный стандарт ISO 22274: 2013 [8]. Он устанавливает основные принципы и требования для обеспечения того, чтобы эти ЛИР были пригодны для применения во всем мире, учитывая такие аспекты, как культурное и языковое разнообразие, а также требования рынка. Стандарт содержит информацию о разработке, развитии и использовании лексико-семантических ЛИР, которые адаптированы к различным языковым, культурным и рыночным условиям.

ISO 22274:2013 в первую очередь определяет факторы, которые необходимо учитывать при создании и заполнении для использования в различных лингвистических средах. Эти факторы включают в себя способы интернационализации ЛИР, а также использование этих способов для структурирования информационных потоков.

В сферу применения стандарта ISO 22274:2013 входят следующие положения:

а) руководящие принципы по информационному содержанию для поддержки интернационализации ЛИР;

б) терминологические принципы, применимые к ЛИР;

в) требования к интернационализации ЛИР;

г) требования к документообороту и администрированию содержания ЛИР для поддержки использования во всем мире.

Стандарт ISO 22274:2013 предназначен для разработчиков ЛИР, включая терминологов и контент-менеджеров. Это также актуально для людей, которые разрабатывают и моделируют соответствующие ИТ-инструменты.

Как было указано, наиболее общим и универсальным средством представления лексико-семантических и, шире, понятийных систем являются онтологии. Этому направлению прикладной лингвистики посвящена обширная литература, анализ которой выходит за рамки данной работы. На русском языке наиболее полно проблема онтологий представлена в монографиях В.Ш. Рубашкина [9] и Н.В. Лукашевич с соавторами [10]. Там же приводятся описания российских и зарубежных проектов онтологий различного типа.

Для современной компьютерной лингвистики основополагающим документом по созданию и применению онтологий являются рекомендации Консорциума W3 C *Семантика и абстрактный синтаксис языка веб-онтологий OWL* [11] (последняя редакция 2009 г.).

В последние годы наиболее перспективным международным проектом в области концептуальных ресурсов представляется проект *Ontology-Lexicon*, или *Ontolex*[1], который также развивается в рамках Консорциума W3.

[1] Ontology-Lexica Community Group. – URL: https://www.w3.org/community/ontolex/ (дата обращения: 01.12.2021).

Миссия группы сообщества *Ontology-Lexicon* состоит в том, чтобы:

- разработать модели представления ЛИР относительно онтологий. Эти лексические модели предназначены для представления лексических записей, содержащих информацию о том, как элементы онтологии (классы, свойства, индивиды и т.д.) реализуются в нескольких языках. Кроме того, лексические записи содержат соответствующую лингвистическую (синтаксическую, морфологическую, семантическую и прагматическую) информацию, которая ограничивает использование записи;
- продемонстрировать дополнительную ценность представления лексики в семантической сети, уделяя особое внимание тому, как использование принципов связанных данных может позволить повторно использовать существующую лингвистическую информацию из таких ресурсов, как WordNet;
- обеспечить наилучшую практику использования лингвистических категорий данных в сочетании с лексикой;
- продемонстрировать, что создание такой лексики в сочетании с семантикой, содержащейся в онтологиях, может улучшить производительность инструментов NLP.
- объединить людей, работающих над стандартами представления лингвистической информации (синтаксической, морфологической, семантической и прагматической), опираясь на существующие инициативы и определяя направления сотрудничества на будущее.
- обеспечивать взаимодействие между существующими моделями для представления и структурирования лингвистической информации.
- продемонстрировать дополнительную ценность приложений, основанных на использовании комбинации лексики и онтологий.

Спецификация модели *Ontology-Lexicon* представлена в: [12].

## Представление электронных словарей

Проблема формализованного представления текстовой информации в целом и лексикографической в частности имеет большую историю и обширную литературу. Мы кратко опишем несколько подходов к решению этой проблемы, имеющих наибольшее распространение и авторитет в компьютерной лингвистике.

### *Инициатива текстового кодирования TEI*

Первой универсальной системой формализованного представления текстовой информации была Инициатива текстового кодирования (TEI)[1]. TEI – это консорциум, который коллективно разрабатывает и поддерживает стандарты представления текстов в цифровой форме. Его главным результатом является набор руководящих принципов, которые определяют методы кодирования для машиночитаемых текстов, главным образом в гуманитарных, социальных науках и в лингвистике. С 1994 года руководство TEI широко используется библиотеками, музеями, издателями и отдельными учеными для

[1] Text Encoding Initiative. – URL: https://tei-c.org/ (дата обращения: 01.12.2021).

представления текстов для онлайн-исследований, преподавания и сохранения. Учебники по TEI вместе с примерами, тестами, упражнениями и описанием инструментария можно найти на сайте консорциума[1].

Правила применения TEI для электронных словарей на русском языке представлены в пособии В.П. Захарова [13].

***Структура лексической разметки LMF***

Международный стандарт для представления лексикографической информации (ISO 24613: 2008) [14], известный под акронимом LMF (Lexical Markup Framework), был разработан с учетом модели TEI.

Стандарт направлен на оптимизацию производства, обслуживания и расширения электронных лексических ЛИР и их интеграцию в прикладные программы. LMF – это абстрактная метамодель, которая обеспечивает общую стандартизированную структуру для построения лексических ЛИР. LMF обеспечивает кодирование лингвистической информации таким образом, чтобы ее можно было повторно использовать в различных приложениях и для различных задач. LMF обеспечивает общее, разделяемое представление лексических объектов, включая морфологические, синтаксические и семантические аспекты.

Цели LMF заключаются в том, чтобы обеспечить общую модель создания и использования электронных ЛИР в диапазоне от малых до больших, управлять обменом данными между ними, а также способствовать объединению большого числа различных отдельных ЛИР для формирования обширных глобальных ЛИР. Конечная цель LMF – создать модульную структуру, которая будет способствовать истинной совместимости контента во всех аспектах электронных ЛИР.

LMF описывает базовую иерархию информации лексической записи, включая информацию о форме. Основной стандарт дополняется различными ресурсами, которые являются частью LMF. Эти ресурсы включают в себя:

– конкретные категории данных, используемые различными типами лексикографических ЛИР;

– ограничения, регулирующие отношение этих категорий данных к метамодели и ее расширениям;

– стандартные процедуры для выражения этих категорий и, таким образом, для закрепления их на структурном каркасе LMF и соотнесения их с соответствующими моделями расширения;

– словари, используемые LMF для выражения связанных информационных объектов для описания того, как расширить LMF через связь с различными конкретными ресурсами (расширениями) и методами анализа и проектирования таких связанных систем.

В приложениях к ISO 24613: 2008 рассмотрено применение стандарта в двух сферах:

а) машиночитаемые словари;

б) обработка лексических ресурсов естественного языка.

[1] TEI by Example. – URL: https://teibyexample.org/TBE.htm (дата обращения: 01.12.2021).

Расширения LMF выражаются в структуре, которая описывает повторное использование основных компонентов LMF (таких как структуры, категории данных и словари) в сочетании с дополнительными компонентами, необходимыми для конкретного ресурса.

Типы отдельных экземпляров LMF могут включать такие электронные лексические ресурсы, как достаточно простые лексические базы данных, лексиконы NLP и машинного перевода, а также электронные одноязычные, двуязычные и многоязычные лексические базы данных. LMF предоставляет общие структуры и механизмы для анализа и проектирования новых электронных лексических ресурсов, но LMF не определяет структуры, ограничения данных и словари, которые будут использоваться при проектировании конкретных электронных лексических ресурсов. LMF также предоставляет механизмы для анализа и описания существующих ресурсов с использованием общей описательной структуры. Для целей проектирования новых ЛИР и описания существующих LMF определяет условия, позволяющие сопоставлять данные, выраженные в любом лексическом ресурсе, со схемой LMF и, таким образом, обеспечивает промежуточный формат для обмена лексическими данными.

В 2013 году опубликована книга [15], полностью посвященная LMF. Первая глава посвящена истории моделей лексики, вторая глава представляет собой формальное представление модели данных, а третья посвящена связи с категориями данных ISO-DCR. Остальные 14 глав посвящены лексическим ЛИР научно-исследовательских лабораторий или промышленных приложений.

***Простая система организации знаний SKOS***

Данная модель разработана Консорциумом W3 C и принята в качестве рекомендаций [16]. Различные семейства систем организации знаний, включая тезаурусы, схемы классификации, системы предметных рубрик и таксономии, широко признаны и применяются как в современных, так и в традиционных информационных системах. На практике бывает трудно провести абсолютное различие между тезаурусами, классификационными схемами или таксономиями.

В этом документе определяется *Простая система организации знаний* (SKOS), общая модель данных для обмена и связывания систем организации знаний через Интернет. Многие системы организации знаний, такие как тезаурусы, таксономии, схемы классификации и системы предметных рубрик, имеют схожую структуру и используются в аналогичных приложениях. SKOS выявляет это сходство и делает его явным, чтобы обеспечить обмен данными и технологиями между различными приложениями.

Модель данных SKOS обеспечивает стандартный и недорогой путь миграции для переноса существующих систем организации знаний в Семантическую сеть. SKOS также предоставляет легкий, интуитивно понятный язык для разработки и обмена новыми системами организации знаний. SKOS может использоваться сам по себе или в сочетании с официальными языками представления знаний, такими как язык веб-онтологий (OWL).

Модель данных SKOS формально определена в этой спецификации как полная онтология OWL[1]. Данные SKOS выражаются в виде троек RDF и могут быть закодированы с использованием любого конкретного синтаксиса RDF.

Модель данных SKOS рассматривает систему организации знаний как концептуальную схему, содержащую набор понятий (концепций). Концептуальные схемы SKOS и понятия SKOS идентифицируются с помощью URI, что позволяет любому однозначно ссылаться на них из любого контекста и делает их частью Всемирной паутины. Понятия SKOS могут быть помечены любым количеством символьных строк. Одна из этих меток на любом заданном языке может быть указана как предпочтительная метка для этого языка, а другие – как альтернативные метки.

Понятиям SKOS может быть присвоено одно или несколько обозначений. Хотя URI являются предпочтительным средством идентификации понятий SKOS в компьютерных системах, обозначения служат переходом к другим уже используемым системам идентификации, таким как классификационные коды, используемые в каталогах библиотек.

Понятия SKOS могут быть задокументированы с помощью примечаний различного типа. Они могут быть связаны с другими понятиями SKOS через свойства семантических отношений. Понятия SKOS можно сгруппировать в коллекции, которые можно пометить и / или определить.

Понятия SKOS могут быть сопоставлены с другими концепциями SKOS в различных концептуальных схемах. Модель данных SKOS обеспечивает поддержку четырех основных типов сопоставления ссылок: иерархических, ассоциативных, близких эквивалентов и точных эквивалентов, но список отношений может быть расширен для поддержки более конкретных потребностей.

***Общие рекомендации по представлению словарей и рекомендации по передовой практике для словарных статей***

Выше мы упоминали о рекомендациях, которые были разработаны для поставщиков ЛИР для журнала *Dictionaria*. Согласно версии рекомендаций 9–1-2016 [17], представление словаря должно состоять из вводного текста и двух-четырех файлов, содержащих следующие наборы данных: записи, смыслы, примеры и ссылки. Эти наборы данных могут быть представлены в табличной форме и будут функционировать как таблицы нашей базы данных; поэтому файлы должны быть связаны с помощью идентификаторов, как описано в разделах ниже.

Представление словаря может также содержать звуковые файлы, видеофайлы и графические файлы. Dictionaria предоставляет веб-приложение для просмотра и поиска. Каждый словарь следует рассматривать как структурированную реляционную базу данных, состоящую из (до) четырех таблиц данных плюс (необязательно) мультимедийный контент.

[1] OWLWeb Ontology Language Semantics and Abstract Syntax. – URL: http://www.w3.org/TR/owl-semantics/ (дата обращения: 01.12.2021).

***Проект по объединению словарей***[1]

Это открытый проект по объединению всех существующих словарных форматов на основе универсального XML-формата, поддерживающего возможность структурно-семантической разметки словарных статей. Проект включает в себя открытый формат XDXF и open-source конвертер словарей различных форматов. Формат позволяет создавать как обычные пользовательские словари, так и тезаурусы и онтологии. Проект закрыт в 2014 году. Список существующих XDXF-словарей проекта представлен на сайте проекта, в нем их более 300, в том числе много русско-иноязычных.

## Средства программной поддержки электронных словарей

В электронном учебном курсе [18] предлагается классификация электронных словарей по программным и технологическим характеристикам, которую мы здесь воспроизводим с сокращениями.

1. *По используемой операционной системе*. Электронные словари могут работать под управлением различных операционных систем; версии существуют для всех современных операционных систем.

2. *По способу загрузки*. Можно подразделить на нерезидентные и резидентные. К первым относятся простейшие программы (например, подстрочечный словарь DIC), которые работают только в собственной среде и не вызываются из других оболочек, например из текстовых редакторов. В большинстве случаев они функционируют в режиме автоматического («пакетного») перевода. Вторые загружают свое ядро в оперативную память компьютера (например «LINGVO») и могут вызываться в любой момент работы компьютера, например из любого текстового редактора,

3. *По количеству подключаемых словарных баз* (словарей). Ранние версии ЭС позволяли подключать только один словарь. Современные программы, независимо от того в какой операционной системе они работают, позволяют подключать до нескольких десятков словарных баз и устанавливать приоритет последних.

4. *По возможностям расширения словарной базы*. Устаревшие ЭС не имели возможности расширения словарных баз пользователем. Современные версии, например LINGVO 4.6 и выше, имеют специальные утилиты для создания пользователем собственных и расширения существующих словарей.

5. *По режиму перевода*. Известны два основных режима перевода: автоматический пакетный и интерактивный (режим «запрос – ответ»).

Это конечно, далеко не единственная классификация программ электронной лексикографии.

Иногда выделяется тип *Словарная система письма* (DWS) – это программное обеспечение для написания и создания словаря, глоссария или тезауруса. Он может включать в себя редактор, базу данных, веб-интерфейс для

[1] Список существующих XDXF словарей проекта XML Dictionary eXchange Format. – URL: http://dicto.org.ru/xdxf.html (дата обращения: 01.12.2021).

совместной работы и различные инструменты управления. В WG3 ENEL был подготовлен обзор[1] программных средств *Словарных систем письма*.

Часто в качестве отдельного типа лексических ЛИР выделяют конкордансы. Конкорданс – это список всех употреблений заданного языкового выражения (например слова) в контексте, возможно со ссылками на источник. В этом значении данный термин широко используется в корпусной лингвистике. Поиск в корпусе данных позволяет по любому слову построить конкорданс – список всех употреблений данного слова в контексте со ссылками на источник. Иногда конкордансом называют список примеров, полученных в результате поиска по корпусу интересующего пользователя языкового выражения со ссылками на источник. Конкордансы используются для решения следующих лингвистических задач:

- сравнение различных использований одного слова;
- анализ ключевых слов;
- анализ частотности слов и словосочетаний;
- поиск и исследование фраз и идиом;
- поиск перевода, например терминологии;
- создание списков слов (что используется при публикации).

Существуют специальные программы составления конкордансов по некоторому корпусу текстов, так называемые конкорданты. Они позволяют получать частоту той или иной языковой единицы по произвольному корпусу текстов, список контекстов, в которых данная единица встретилась. Многие из них позволяют также сортировать контексты по ключевому слову или по словоформе, по ближайшему контексту. Пример – конкордансы произведений Ф.М. Достоевского[2].

Инструментальные средства создания электронных словарей включают средства извлечения из текста слов, словосочетаний, терминов и именованных сущностей, средства сегментирования текста, статистические инструменты и другое. Приведем примеры доступных инструментов для лексикографов согласно перечню на портале ELEXIS[3]. Все эти инструменты распространяются бесплатно. Их могут использовать терминологи, лингвисты, переводчики, преподаватели и другие.

***Sketch Engine***[4] содержит ряд уникальных инструментов для анализа больших корпусов, до 30 миллиардов слов. Каждый пользователь может воспользоваться полностью автоматизированной функцией построения словаря.

---

[1] Survey – WG3 ENeL Dictionary Writing Systems& Corpus Query Systems. https://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf – URL: (дата обращения: 01.04.2022).

[2] Конкордансы произведений Ф.М. Достоевского. – URL: https://philolog.petrsu.ru/fmdost/concordance/user_new/index.php (дата обращения: 01.04.2022).

[3] ELEXIS. Tools and services. – URL: https://elex.is/tools-and-services/ (дата обращения: 01.12.2021).

[4] Sketch Engine. – URL: https://www.sketchengine.eu/ (дата обращения: 01.12.2021).

***Lexonomy***[1] – это облачная система для написания словарей, а также для публикации онлайн-словарей, которая хорошо масштабируется для адаптации к крупным словарным проектам, а также к небольшим лексикографическим работам, таким как редактирование и онлайн-публикация тематических глоссариев или терминологических ресурсов. *Lexonom*у уже взаимодействует со *Sketch Engine*, который может помещать лексикографические данные в *Lexonomy* для создания автоматически сгенерированных черновиков словарей, а *Lexonomy* может извлекать данные из корпусов *Sketch Engine* во время процесса редактирования записи.

***OneClick Dictionary (OCD)***[2] – это модуль для создания словаря. Он связывает систему управления корпусом (например *Sketch Engine*) или даже листы Excel с системой написания словарей и онлайн-публикации словарей *Lexonomy,* и обеспечивает автоматически созданный черновик словаря (например заглавные слова, словосочетания, словосочетания, примеры) для последующего редактирования лексикографом. Словарь *OneClick* позволяет лексикографам перенести всю интеллектуальную работу на этап постредактирования, вместо того чтобы вручную анализировать входные данные перед созданием черновика словаря.

Следовательно, этот инструмент предназначен не только для профессионалов, но и для спонтанной лексикографии – небольших проектов лексикографического характера, таких как глоссарии, тематические словари и словари, часто подготовленные учителями или другими профессионалами без формального обучения лексикографии.

***Elexifier***[3] – это облачный сервис преобразования словарей. Он использует передовые методы синтаксического анализа XML и машинного обучения, чтобы преобразовать словари PDF и XML в стандартизированный машиночитаемый формат. Пользователи могут загружать свои словари PDF и пользовательские словари XML в *Elexifier*, определять правила сопоставления для преобразования XML или создавать обучающий набор машинного обучения для преобразования PDF и загружать преобразованный словарь XML или PDF в формате файла, совместимого с TEI, на основе модели данных Elexis.

В заключение раздела приведем немного статистики по инструментальным средствам компьютерной лексикографии.

Каталог META-SHARE[4] содержит 35 инструментов для поддержки лексических ЛИР.

Каталог LRE[5] включает 60 инструментов для извлечения именованных сущностей.

---

[1] Lexonomy. – URL: https://lexonomy.elex.is/ (дата обращения: 01.12.2021).

[2] Dictionary Drafting Module (One-Click Dictionary) Lexonomy. – URL: https://github.com/elexis-eu/ocd (дата обращения: 01.12.2021).

[3] Elexifier. – URL: https://elexifier.elex.is/ (дата обращения: 01.12.2021).

[4] Search & exchange language resources. – URL: http://www.meta-share.org/ (дата обращения: 01.12.2021).

[5] LRE Map. – URL: https://lremap.elra.info/ (дата обращения: 01.04.2022).

Каталог LINGUIST List[1] дает следующие ссылки на инструментальные лексикографические средства:

- конкордансеры – 10
- средства для поддержки лексиконов – 19
- комплексные инструменты NLP, включая создание и поддержку словарей – 25

Каталог NL-pub[2] включает следующие инструментальные средства компьютерной лексикографии:

- построение конкордансов – 3
- извлечение ключевых слов – 12
- извлечение именованных сущностей – 30
- извлечение отношений – 9
- обнаружение дубликатов – 1
- редакторы тезауруса – 12

## Электронная лексикография в России

***Словарные службы и их обзоры***

В настоящее время российским пользователям Интернета доступны сотни как оцифрованных традиционных словарей на различных языках, так и специализированных лексикографических баз данных различного назначения. Назовем наиболее популярные российские агрегаторы лексикографических ЛИР: *Грамота.ру*[3]*, Словари*[4]*, Lingvo*[5]*, Словари Онлайн*[6]*, Мультитран*[7]*, Академик*[8]*, Словари. СС*[9].

Крупная база словарей языков народов России собрана на платформе LingvoDoc[10]. Создатели платформы ориентировались на исчезающие и слаборазвитые языки, но в списке присутствует много словарей и активно функционирующих языков. Количество словарей по отдельным языкам указано в списке в квадратных скобках после названия языка или языковой семьи.

Азербайджанский [7], Алтай-кижи [1], Алтайские языки [341], Алтайский [46], Башкирский [44], Бурятский [1], Венгерский [4], Вепсский [1], Водский [1], Долганский [1], Ижорский [1], Казахский [9], Камасинский [4], Карачаево-балкарский [1], Карельский [152], Коми [21], Корейский [1],

[1] LINGUIST List. – URL: https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=2 (дата обращения: 01.12.2021).

[2] NL-pub. – URL: https://nlpub.ru/Обработка_текста (дата обращения: 01.12.2021).

[3] Грамота ру. – URL: http://gramota.ru/slovari/ (дата обращения: 01.12.2021).

[4] Словари. – URL: http://slovari.ru/start.aspx?s=0&p=3050 (дата обращения: 01.12.2021).

[5] Lingvo. – URL: https://www.lingvolive.com/ru-ru (дата обращения: 01.12.2021).

[6] Словари Онлайн. – URL: https://slovaronline.com/ (дата обращения: 01.12.2021).

[7] Мультитран. – URL: https://www.multitran.com/ (дата обращения: 01.12.2021).

[8] Академик. – URL: https://dic.academic.ru/ (дата обращения: 01.12.2021).

[9] Словари. СС. – URL: https://slovar.cc/ (дата обращения: 01.12.2021).

[10] LingvoDoc. – URL: http://lingvodoc.ru/dashboard/dictionaries_all (дата обращения: 01.12.2021).

Крымско-татарский [1], Кумыкский [1], Мансийский [16], Марийский [6], Мокшанский [16], Монгольские [1], Нанийская группа / Южнотунгусские языки [24], Нганасанский [6], Негидальский [2], Ненецкие [16], Ногайский [3], Огузские [29], Орокский [1], Орочский [1], Прибалтийско-финские [157], Саамские [4], Самодийские [66], Северно-монгольские [1], Селькупский [30], Солонский [1], Татарский [16], Теленгитский [1], Тунгусо-маньчжурские языки [86], Тунгусские языки [24], Туркменский [4], Тюркские языки [252], Удмуртский [12], Удэгейский [5], Узбекский [4], Уйгурский [2], Ульчский [8], Уральские [355], Финский [1], Хакасский [11], Хантыйский [30], Челканский [1], Чувашский [9], Чулымский [4], Эвенкийский [36], Энецкий [6], Эрзянский [22], Якутский [21], Японско-корейская группа [1].

Сведения о российских лексикографических ЛИР для научных исследований собраны в Навигаторе информационных ресурсах по языкознанию[1]. Всего там описано около 100 лексикографических ЛИР, разработанных в учреждениях РАН, в том числе различные словари на основе НКРЯ.

Обзор словарных баз данных и обсуждение различных проблем их создания можно найти в материалах конференции «Слово и Словарь» [19], в том числе в статье В.П. Захарова [20]. Перечень российских электронных словарей и лексических баз данных можно найти в каталоге NL-Pub в разделах «Словари» и «Тезаурусы»[2].

Большая подборка электронных словарей, как созданных в России, так и зарубежных, но включающих русский язык, имеется на портале *Лексиколог*[3].

Еще один обзор российских лексикографических ресурсов представлен в работе [21]. Были проанализированы 12 лексикографических ЛИР, и по результатам анализа были сделаны следующие выводы.

1. *Перекрытия между списками словарных слов*. Развитая традиция и преемственность различных проектов создания словарей объясняют значительные совпадения между списками слов традиционных словарей. Словарь USH[4] выделяется в этом отношении, что можно объяснить тем, что проект USH больше не развивается. Низкое перекрытие между RWN[5] и другими словарями, напротив, косвенно подтверждает идею о том, что прямой перевод тезауруса на другой язык значительно сокращает лексику.

2. *Количество уникальных слов в словарях*. Как мы выяснили, существует относительно мало уникальных слов и фраз (т.е. содержащихся только в одном словаре). Этот факт отчасти обусловлен выбором репрезентации производных – в виде отдельного заглавного слова или внутри записи. В то же

[1] Навигатор информационных ресурсах по языкознанию. – URL: http://niryaz.inion.ru/ (дата обращения: 01.12.2021).

[2] Ресурсы. Словари. – URL: https://nlpub.ru/Ресурсы#.D0.A1.D0.BB.D0. BE.D0. B2. D0. B0. D1.80. D1.8 C (дата обращения: 01.12.2021).

[3] Лексиколог. – URL: https://www.lexilogos.com/english/russian_dictionary.htm (дата обращения: 01.12.2021).

[4] Толковый словарь русского языка Д.Н. Ушакова (USH). – URL: https://gufo.me/dict/ushakov (дата обращения: 01.04.2022).

[5] RWN Русский Wordnet. – URL: http://wordnet.ru (дата обращения: 01.12.2021).

время в этих словарях (даже в EFR[1]) наблюдается существенный недостаток многословных выражений, которые представлены в электронных ресурсах гораздо лучше. Кроме того, большое количество уникальных терминов в WIKT[2] можно объяснить тем, что его список слов включает имена собственные (35 000 слов из 193 500).

3. *Корпоративное покрытие*. Судя по доле уникальных слов, можно было бы предположить, что традиционные словари не имеют хорошего корпусного охвата. Однако это не так – особенно в отношении BTS[3], EFR и MAS[4]. Заметный «дефицит» словарей синонимов вполне понятен и ожидаем. Полученные результаты позволяют дать сырую оценку русских лемм, вовлеченных в синонимические отношения, – около 60–70%.

4. *Количественный анализ определений*. Доля моносемантических слов, содержащихся в традиционных словарях, была значительно выше, чем в электронных ресурсах. Этот факт указывает на ориентацию последних на фактическое словоупотребление и тенденцию к репрезентации конкретных значений.

5. *Анализ охвата современной лексики*. Наконец, сравнение словарей по наличию неологизмов показывает большой потенциал современных электронных ресурсов, которые могут быть динамически модифицированы. Это не значит, что традиционные словари устарели. Отставание от изменений в языке дает возможность отразить в словаре не просто случайные, а устоявшиеся языковые явления: слова, значения, вариации и т.д. Современная ситуация в современной российской лексикографии отражает переходный период от традиционных печатных изданий к масштабным проектам, основанным на крупных корпорациях и краудсорсинге. Традиционные словари, основанные на ручной выборке и обработке данных, считаются высококачественными источниками, но они явно отстают от таких ресурсов, как WIKT, учитывая их объем и охват современной лексикой. В то же время специфика электронных проектов часто подвергается критике за их качество.

### *Концептуальные ЛИР в России и проблемы их интеграции*

В России имеется обширный опыт создания лексико-семантических ресурсов, в том числе информационно-поисковых и лингвистических тезаурусов. Их обзоры можно найти в монографиях Н.В. Лукашевич [7], автора этой книги [22], а также в фундаментальной дилогии Вал. А. и Вл. А. Луковых [23]. Заметим, что для информационно-поисковых тезаурусов разработаны международные и национальные стандарты [24; 25], регламентирующие их структуру и оформление.

---

[1] Толковый словарь русского языка Ефремовой (EFR). – URL: https://www.efremova.info/ (дата обращения: 01.12.2021).

[2] Русский Викисловарь (WIKT). – URL: http://ru.wiktionary.org (дата обращения: 01.12.2021).

[3] Большой словарь русского языка (BTS). – URL: https://gufo.me/dict/kuznetsov (дата обращения: 01.12.2021).

[4] Малый академический словарь (MAS). – URL: https://gufo.me/dict/mas (дата обращения: 01.12.2021).

Всего в России были разработаны сотни лексико-семантических ЛИР, в основном информационно-поисковых тезаурусов для отдельных предметных отраслей, но большинство из них было рассчитано на интеллектуальное индексирование и в современных технологиях информационного поиска не используются.

Традиционные информационно-поисковые тезаурусы созданы и продолжают функционировать в ИНИОН РАН[1], ЦНСХБ[2], ООО «Интегрум Медиа»[3] и еще нескольких организациях. Отметим также, что в Центральной научной медицинской библиотеке, где используется международный медицинский тезаурус *MESH*[4], разрабатывается представление тезауруса в виде связанных открытых данных.

Появление и распространение онтологий как универсального средства представления понятийной структуры предметной области снова привлекли внимание к тезаурусам, которые естественно рассматривать как промежуточный этап создания полноценных онтологий.

В России наиболее полно и глубоко онтологическая проблематика реализована в Лаборатории информационных исследований[5], которая известна своими разработками в области создания тезаурусов и автоматической обработки текста. Созданный в лаборатории тезаурус *RuThes*[6] используется в информационной системе УИС Россия[7] и в других проектах с государственными и коммерческими организациями.

Тезаурус *RuThes* представляет собой лингвистический ресурс концептуального типа, т.е. представляет собой иерархическую сеть понятий, к которым приписаны текстовые выражения. И в этом смысле *RuThes* относится к тому же классу, что и тезаурус *WordNet*. При этом, в отличие от *WordNet*, который создавался как модель человеческой памяти (раздельное описание частей речи, специальные типы отношений и др.), тезаурус *RuThes* создавался именно как ресурс для автоматической обработки текстов.

Текущий объем тезауруса *RuThe*s составляет 158 000 слов и выражений, уложенных в сеть 55 000 понятий, между которыми вручную установлено более 210 000 отношений. Особенностью тезауруса является то, что в течение многих лет он тестировался в реальных проектах.

---

[1] Библиографические базы данных. Рубрикатор и тезаурусы. – URL: http://inion.ru/ru/resources/bazy-dannykh-inion-ran/ (дата обращения: 01.12.2021).

[2] ЦНСХБ, Тезаурус AGROVOC. – URL: http://www.cnshb.ru/AGROVOC.shtm (дата обращения: 01.12.2021).

[3] Интегрум. – URL: https://integrum.ru/ (дата обращения: 01.12.2021).

[4] Медицинские предметные рубрики MESH. – URL: https://rucml.ru/pages/mesh (дата обращения: 01.04.2022).

[5] Лаборатория информационных исследований. – URL: http://www.labinform.ru/ (дата обращения: 01.12.2021).

[6] О лингвистической онтологии «Тезаурус РуТез». – URL: http://www.labinform.ru/pub/ruthes/index.htm (дата обращения: 01.12.2021).

[7] УИС Россия. – URL: https://www.uisrussia.msu.ru/ (дата обращения: 01.12.2021).

Авторами *RuThes* создана также версия этого тезауруса в формате *WordNet – RuWordNet (RWN)*[1]. Тезаурус *RuWordNet* содержит синсеты трех частей речи: существительные (отдельные существительные, группы существительного, предложные группы), глаголы (отдельные глаголы и глагольные группы), прилагательные (отдельные прилагательные и группы прилагательного).

Между синсетами, относящимися к разным частям речи, но выражающими один и тот же смысл, установлены отношения частеречной синонимии, соединяющие разделенные синсеты.

Между синсетами каждой части речи установлены связи гипоним – гипероним (род – вид). Кроме того, между синсетами установлены отношения: часть – целое, экземпляр – класс и отношение антонимии, предметной области. Для глаголов указаны отношения причины и следования. Поисковый интерфейс *RuWordNet* расположен на сайте *Тезаурус русского языка RuWordNet*[2].

Следует отметить также разработки компьютерных лингвистов из СПбГУ. Разработанный этим коллективом тезаурус *RussNet* стал одним из заметных российских ЛИР[3]. Целью проекта является построение лексико-семантического ресурса, отражающего организацию лексической системы русского языка в целом (в противоположность терминологическим или частным словарям); покрывающего ядро общеупотребительной лексики русского языка; фиксирующего все семантические, семантико-грамматические и семантико-деривационные отношения, характерные для русского языка.

*RussNet* унаследовал основные особенности Принстонского *WordNet, EuroWordNet* и других подобных ресурсов:

Тезаурус состоит из четырех взаимосвязанных файлов, содержащих слова основных частей речи: существительные, глаголы, прилагательные и наречия.

Базовой единицей *RussNet* является синонимический ряд (синсет), объединяющий слова со сходным значением.

Синсеты связаны различными парадигматическими и синтагматическими отношениями.

Ценным источником онтологической и лексико-семантической информации является русский *Викисловарь (WIKT)*[4]. Словарь был открыт 1 мая 2004 года, и сейчас в нем содержится 1 114 852 статьи о словах, словообразовательных единицах и словосочетаниях более 500 языков мира. Русский *Викисловарь* является восьмым по величине викисловарем, состоящим из более чем 520 000 статей – одна статья представляет собой лексическую запись, написанную более чем 120 000 пользователей с 2004 года.

---

[1] Тезаурус русского языка в формате WordNet – RuWordNet. – URL: http://www.labinform.ru/pub/ruwordnet/index.htm (дата обращения: 01.12.2021).

[2] Тезаурус русского языка RuWordNet. – URL: https://ruwordnet.ru/ru (дата обращения: 01.12.2021).

[3] RussNet. – URL: http://project.phil.spbu.ru/RussNet/index_ru.shtml (дата обращения: 01.12.2021).

[4] Русский Викисловарь. – URL: https://ru.wiktionary.org/wiki/Заглавная_страница (дата обращения: 01.12.2021).

Предметом описания в *Викисловаре* являются все лексические единицы, зафиксированные во всех письменных языках и диалектах мира. Словник *Викисловаря* формируется из лексем, морфем и других словообразующих единиц, а также устойчивых словосочетаний этих языков, с использованием тех графических основ (систем письменности), которые традиционно применяются в соответствующих языках.

*Викисловарь* сочетает функции нескольких видов традиционных словарей, включая толковые, орфографические, орфоэпические, грамматические, переводные, фразеологические и этимологические словари, а также тезаурусы.

Структура описания языковых единиц в *Викисловаре* основывается на том, как с точки зрения лингвистики описывается язык вообще. Лингвисты выделяют в языке следующие структурные уровни:

- фонетический (звучание)
- орфографический (правильное написание)
- морфологический (состав слов)
- синтаксический (взаимодействие между словами в речи, тексте)
- семантический (значение, смысл)
- этимологический (происхождение слов и выражений)

Все эти уровни должны находить, по возможности, полноценное отражение в каждой статье словаря.

Относительно новым проектом является *Yet Another RussNet* (сокр. *YARN*)[1] – проект создания нового открытого электронного тезауруса русского языка. Основная идея проекта – эксперимент по комбинированию традиционных принципов создания *WordNet* и вики-подхода к наполнению и редактированию ЛИР. Тезаурус доступен на условиях лицензии CC BY-SA, разрабатывается в Уральском федеральном университете с 2013 года для задач автоматической обработки текста и информационного поиска. *YARN* включает в себя около 120 тыс. словарных единиц, 46 тыс. синсетов и 30 тыс. иерархических отношений. Лексикографический подход *YARN* предполагает сочетание краудсорсинга с автоматическими методами построения тезаурусов. В рамках проекта разработаны онлайн-инструменты для коллективного редактирования тезауруса, а также автоматические методы подготовки «сырья» для построения ресурса на основе данных словарей и анализа корпусов текстов.

Идею интеграции нескольких тезаурусов русского языка в семантическую сеть в облаке открытых лингвистических связанных данных реализовал Д.А. Усталов [26].

В проекте интеграции рассмотрены различные тезаурусы русского языка. Существуют четыре известных электронных тезауруса для русского языка, которые находятся в открытом доступе по открытым лицензиям: 1) *RuThes-lite*; 2) Русский Викисловарь; 3) Универсальный сетевой язык и 4) *YARN*.

*RuThes-lite* – это подмножество лексической онтологии *RuThes*, описанной выше. *RuThes-lite* доступен на условиях лицензии CC BY-NC-SA в

[1] Yet Another RussNet – URL: https://russianword.net/ (д ата обращения: 01.12.2021).

виде квазиструктурированных HTML-страниц в Интернете, представляющих примерно 26 000 концептов и 100 000 отношений между ними.

Универсальный сетевой язык (*UNL*)[1] – это проект, возглавляемый Организацией Объединенных Наций и посвященный разработке компьютерного языка, который воспроизводит функции естественных языков. Русская версия его семантической сети – *UNLDC* – распространяется по лицензии CC BY-SA. Он содержит примерно 62 000 универсальных слов (UWS) и 90 000 ссылок между ними.

Русский Викисловарь описан выше. Родной формат страниц викисловаря – это квазиструктурированный синтаксис вики. Кроме того, существует *Wikokit*[2] – проект, который анализирует русские и английские викисловари и выводит их в машиночитаемую форму реляционной базы данных, доступной на условиях лицензии CC BY-SA.

YARN включает в себя лексикон и синсеты русского Викисловаря среди нескольких других ресурсов, лицензированных по той же лицензии. Поэтому в проекте интеграции участвуют только три ресурса: *RuThes-lite, UNLDC*[3], *YARN*. Интегрированный тезаурус получил название *Russian Thesauri as Linked Open Data (сокр. RTLOD)*[4].

Результирующие ресурсы представляются при помощи следующих RDF-словарей: *SKOS* для представления понятий, *OntoLex-Lemon* для лексических входов, значений, определений и примеров употребления, *LexInfo* для записи морфосинтаксических помет, а также *RDFS, OWL* и *Dublin Core* для выражения описания онтологии.

Процедуры интеграции тезаурусов для данного проекта описаны также в работе [27]. Авторы указывают, что «для достижения результата необходимо выполнить интеграцию гетерогенных лексикографических данных: RussNet построен путем лексико-статистического лингвистического подхода, YARN создан путем краудсорсинга с дополнительным применением автоматических методов построения тезаурусов. Интеграция включает в себя согласование концептуальных оснований двух ресурсов, схем данных, разработку автоматических методов выравнивания и сравнения единиц тезаурусов; методики, сценариев и инструментов редактирования и пополнения объединенного ресурса (…) Задача интеграции семантических ресурсов решается в рамках направления создания открытых связанных данных. Специфика конкретных ресурсов и особенности конкретного языка определяют оригинальность и научную новизну используемых методов».

Создаваемый тезаурус может быть использован в системах NLP; потенциальные пользователи – это широкий круг российских и зарубежных исследователей и разработчиков-практиков, а также студентов. Ресурс распро-

[1] UNLWEB. – URL: http://www.unlweb.net/unlweb/ (дата обращения: 01.12.2021).

[2] Wikokit – Machine-readable Wiktionary. – URL: https://github.com/componavt/wikokit (дата обращения: 01.12.2021).

[3] Universal Networking Language System. – URL: http://www.unl.ru/ (дата обращения: 01.12.2021).

[4] Russian Thesauri as Linked Open Data. – URL: https://nlpub.ru/RTLOD (дата обращения: 01.12.2021).

страняется по лицензии CC BY-SA, которая допускает модификацию и применение продукта как в некоммерческих, так и в коммерческих проектах. Опыт использования тезаурусов других языков показывает, что у подобных ресурсов очень широкая область использования. Тезаурусы используются в большинстве современных систем NLP: в задачах классификации текстов, автоматического реферирования, генерации текстов, в информационном поиске, машинном переводе и другом.

Таким образом, можно утверждать, что российская электронная лексикография находится на достаточно высоком уровне: главное, чего не хватает – это организованной инфраструктуры для повышения степени повторного использования электронных словарей различных типов. Решение этой задачи должно быть возложено на словарную службу России, которая должна дополнить Национальный корпус русского языка.

При создании словарной службы России весьма желательна также более глубокая интеграция с зарубежными, прежде всего европейскими, коллегами, использование общепринятых стандартов и форматов представления лекскографической информации.

## Литература к главе 11


1. Щипицина Л.Ю. Информационные технологии в лингвистике : учеб. пособие. – Москва : ФЛИНТА : Наука, 2013. – 128 с. – URL: https://narfu.ru/university/library/books/1580.pdf (дата обращения: 13.11.2021).
2. Фесенко О.П., Лаухина С.С. Электронные словари как продукт современной лексикографии // ОНВ. – 2015. – № 4(141). – URL: https://cyberleninka.ru/article/n/elektronnye-slovari-kak-produkt-sovremennoy-leksikografii (дата обращения: 13.11.2021).
3. Попова Л.В. Типологии и классификации словарей // Вестник ЧелГУ. – 2012. – № 20(274). – URL: https://cyberleninka.ru/article/n/tipologii-i-klassifikatsii-slovarey (дата обращения: 13.11.2021). Hartmann R.R.K. Reference Portals to Internet Sources Relevant to Lexicography and Terminology. – 2010. – URL: http://euralex.pbworks.com/f/Reference+Portals+aug+2010.pdf (дата обращения: 01.04.2022).
5. Селегей В. Электронные словари и компьютерная лексикография. – URL: https://www.elibrary.ru/item.asp?id=42849760 (дата обращения: 01.04.2022).
6. Fellbaum C. WordNet: An Electronic Lexical Database. – Cambridge : MIT Press, 1998. – 422 с.
7. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – Москва : МГУ, 2011. – 512 с.
8. ISO 22274:2013 Systems to manage terminology, knowledge and content – Concept-related aspects for developing and internationalizing classification systems [Электронный ресурс]. – URL: https://www.iso.org/ru/standard/36173.html (дата обращения 12.2021).
9. Рубашкин В.Ш. Онтологическая семантика. – Москва : Физматлит, 2013. – 348 с. – ISBN 978-5-9221-1439-4.
10. Онтологии и тезаурусы: модели, инструменты, приложения / Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. – Москва : Изд-во ИНТУИТ, 2008. – 176 с.
11. OWL Web Ontology Language Semantics and Abstract Syntax. – URL: https://www.w3.org/TR/2004/REC-owl-semantics-20040210/ (дата обращения: 01.12.2021).

12. Final Model Specification. – URL : https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Metadata_.28lime.29 (дата обращения: 01.12.2021).

13. Захаров В.П. Электронный обменный формат для словарей проекта TEI (Text Encoding Initiative) : учебное пособие. – Санкт-Петербург : СПбГУ, РИО, Фил. ф-т, 2013. – 80 с.

14. ISO 24613:2008 Language resource management – Lexical markup framework (LMF). – URL: https://www.iso.org/standard/37327.html

15. LMF Lexical Markup Framework, ISTE / Gil Francopoulo (edited by). – Wiley, 2013. – ISBN 978-1-84821-430-9

16. SKOS Simple Knowledge Organization System Reference. – URL: https://www.w3.org/TR/skos-reference/ (дата обращения: 01.12.2021).

17. General Dictionaria Submission Guidelines and Best Practice Recommendations for Dictionary Entries. – URL: https://dictionaria.clld.org/submit (дата обращения: 01.12.2021).

18. Классификация электронных словарей. – URL: https://studbooks.net/2108752/literatura/predposylki_poyavleniya_elektronnyh_slovarey (дата обращения: 01.12.2021).

19. Слово и словарь = Vocabulum et vocabularium : сборник научных статей / ред. О.Н. Крылова, С.А. Мызников. – Санкт-Петербург : Нестор-История, 2016. – Т. 14. – 768 с.

20. Захаров В.П. Электронная лексикография XXI века // Слово и словарь = Vocabulum et vocabularium. : сборник научных статей / ред. О.Н. Крылова, С.А. Мызников. – Санкт-Петербург : Нестор-История, 2016. – Т. 14. – С. 304–324.

21. Русский лексикографический ландшафт: история о 12 словарях // Computational Linguistics and Intellectual Technologies: «Dialogue». – Moscow : RGGU, 2015. – Iss. 14(21). – P. 254–271. – URL: https://www.dialog-21.ru/digests/dialog2015/materials/pdf/kiselevyabraslavskipietal.pdf (дата обращения: 01.04.2022).

22. Антопольский А.Б. Лингвистическое обеспечение электронных библиотек / НТЦ «Информрегистр». – Москва, 2003. – 302 с.

23. Луков Вал. А., Луков Вл. А. Тезаурусы: субъектная организация гуманитарного знания. – Москва : Изд-во Нац. ин-та бизнеса, 2008. – 784 с. – 1000 экз. – ISBN 978-5-8309-0272-4.

24. Луков Вал. А., Луков Вл. А. Тезаурусы II: тезаурусный подход к пониманию человека и его мира. – Москва : Изд-во Нац. ин-та бизнеса, 2013. – 640 с. – 700 экз. – ISBN 978-5-8309-0391-2.

25. ГОСТ 7.25–2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления (Система стандартов по информации, библиотечному и издательскому делу). – Москва : Стандартинформ, 2007.

26. ГОСТ 7.24–2007. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению: межгосударственный стандарт (Система стандартов по информации, библиотечному и издательскому делу). – Москва : Стандартинформ, 2007.

27. Усталов Д.А. Семантические сети и обработка естественного языка. Открытые системы // СУБД, открытые системы. – Москва, 2017. – № 2. – С. 46–47.

28. Интеграция тезаурусов RussNet и YARN. Компьютерная лингвистика и вычислительные онтологии : сборник научных статей XIX Объединенной конференции «Интернет и современное общество» IMS-2016, Санкт-Петербург, 22–24 июня 2016 года / И.В. Азарова, П.И. Браславский, В.П. Захаров, Ю.А. Киселев, Д.А. Усталов, М.В. Хохлова. – URL: https://openbooks.itmo.ru/ru/file/4100/4100.pdf (дата обращения: 01.12.2021).

# ГЛАВА 12. ТЕРМИНОЛОГИЧЕСКИЕ БАЗЫ ДАННЫХ

## Общие сведения

Особым типом лексикографических ЛИР являются терминологические базы данных (ТБД). Большинством специалистов они выделяются, поскольку создание и применение ТВД обычно выходит за пределы лингвистических технологий (автоматизированной обработки текста и речи, преподавания языков, машинного перевода и др.) и предназначено для отраслевых специалистов, переводчиков, редакторов.

Согласно ISO 30042:2008 [1], ТБД – это «база данных, содержащая информацию о специальных языковых понятиях и терминах, предназначенных для представления этих понятий, а также связанную с ними концептуальную, связанную с терминами и административную информацию».

Назначение ТБД – это обеспечение надежной и качественной коммуникации в различных сферах деловой, политической и социальной жизни, особенно в условиях многоязычия. Поэтому создание ТБД прежде всего входит в задачи международных организаций, особенно профессиональных.

Созданием ТБД активно занимаются также органы стандартизации, поскольку терминологические стандарты традиционно являются одним из направлений стандартизации.

Некоторые авторы отличают ТБД от других видов лексикографии еще и по содержательному критерию. Так, в методическом материале международного консорциума TerminOrg указывается:

«Концептуальный подход к управлению терминологией требует, чтобы вся терминологическая информация, относящаяся к одному понятию, обрабатывалась как единая терминологическая запись. В концептуальной системе данные организованы вокруг значения, а не языковой формы (термина). Все термины, варианты и переводы, обозначающие одно понятие, а также все описательные и административные данные, относящиеся к этому же понятию, хранятся в одной терминологической записи.

Именно этот подход отличает терминологию и терминологический менеджмент от лексикологии и лексикографии, которые представляют собой практику разработки словарей (управление терминологией иногда называют “терминологией”, расширяя термин “лексикография”). Словари “основаны на словах”, в отличие от терминологических баз. Каждая запись в словаре сосредоточена вокруг слова и описывает все различные значения этого слова» [2].

Это утверждение можно признать дискуссионным**,** поскольку в лексикографическую традицию часто включают и создание различных концептуальных или семантических словарей, в том числе тезаурусы и онтологии, которые мы затронули в главе 11. Тем не менее, мы сочли целесообразным описание ТБД вынести в отдельную главу, руководствуясь соображением специфики ТБД с точки зрения назначения и организации терминологической деятельности, по сравнению с другими языковыми технологиями. В этой же главе описаны ЛИР, близкие к ТБД, – базы данных «Память перевода», а также таксономии и номенклатуры.

## Европейский опыт

Особенно важен европейский опыт управления терминологией, ведь для Евросоюза преодоление языкового барьера при сохранении равенства языков является магистральной политической задачей. Поэтому ТБД обслуживают огромную сеть сотрудников европейских структур и переводчиков. ТБД используются для подготовки и рассылки документации, проведения разнообразных встреч и многих других целей.

Созданию системы управления терминологической деятельностью в Евросоюзе предшествовало фундаментальное исследование, проведенное в 1994–1995 гг. в рамках проекта ***POINTER***[1]. Одним из исполнителей проекта был университет Суррея в Великобритании. Здесь кратко излагаются основные выводы, сделанные в отчете по проекту [2].

*Отсутствие терминологических ресурсов*

Несмотря на важность терминологии, существует общая нехватка точных, современных, структурированных, легко (повторно) используемых терминологических ресурсов и литературы и, что еще более важно, легкодоступных информационных и распределительных каналов. Это относится ко всем областям применения терминологии, и особенно к двум ее основным рынкам: продуктам языковой индустрии и лингвистическим услугам, в частности переводу. Немногие существующие ресурсы являются надежными, и еще меньше их доступно в режиме онлайн. Кроме того, является проблемой сохранение существующих терминологий, особенно в инновационных и, следовательно, быстро развивающихся областях.

Недостаток ресурсов вызывает проблемы, как с точки зрения данного момента, так и с учетом развития. Особенно остро ситуация обстоит в отношении менее используемых языков, таких как греческий, итальянский и восточноевропейские языки; в таких случаях английский часто используется в качестве языка-посредника.

*Отсутствие повторного использования*

На сегодняшний день большая часть терминологической работы выполняется в государственном и частном секторах в качестве дополнения

[1] POINTER: proposals for an operational infrastructure for terminology in Europe. – URL: https://cordis.europa.eu/project/id/LRE63090/results 1. (дата обращения: 01.12.2021).

или побочного продукта и часто не предназначена для повторного использования и представления в открытый доступ.

*Качество*

Другой ключевой проблемой, влияющей на создание и распространение ресурсов, является качество. На практике качество тех коллекций, которые имеются в наличии, сильно варьируется и во многих случаях просто неадекватно (особенно это касается многоязычных коллекций). Кроме того, не существует общих стандартов / процедур валидации, отсутствует методология качества любых эквивалентов, добавленных на разных языках.

В более общем плане можно сказать, что официальные стандарты (как стандарты предметной области, так и, в частности, стандарты терминологической работы) не всегда используются или даже известны. Кроме того, до сих пор не существует широкого и последовательного использования инструментов или носителей информации, хотя использование компьютеров постепенно растет.

*Неологизмы*

Неологизмы неразрывно связаны с эволюцией языка, т.е. предметная область включает в себя социолингвистические и культурные аспекты. Адаптация исследования неологизмов для создания терминологии и последующего ее использования является ключевой проблемой. Однако лишь немногие европейские страны систематически обращаются с неологизмами.

*Многоязычные ресурсы*

Количество терминологических и лексических ресурсов, доступных на двух, трех или более языках, относительно невелико, а те, которые существуют, часто являются бедными. Как правило, можно сказать, что чем больше корпус словаря, тем ниже стандарт – отражение, возможно, значительных денежных средств, затраченных на создание многоязычных ресурсов. Качество особенно низкое в языках, отличных от исходного. Во многих случаях наблюдается тенденция просто переводить термины при добавлении эквивалентов, даже если особое внимание было уделено организации и качеству определений в корпусе исходного языка. В результате может быть включен высокий процент нерелевантных терминов.

*Рекомендации*

- Поощрение и более широкое использование соответствующих методологий одноязычной и многоязычной терминологии.
- Дальнейшая разработка, принятие и продвижение методологии оценки качества многоязычных терминологических ресурсов.
- Обеспечение международных организаций, занимающихся разработкой многоязычных терминологических ресурсов, современной методологией.
- Эквиваленты должны быть основаны на терминологической, а не лексикографической точке зрения.
- Создание и распространение высококачественных машиночитаемых многоязычных ресурсов.

- Сосредоточение государственного финансирования на улучшении существующих качественных (одноязычных) ресурсов в роли прелюдии к многоязычным ресурсам.
- Информирование пользователей и авторов о проблемах, связанных с созданием многоязычных справочных работ на трех и более языках (потребность в более чем простых списках слов).
- Содействие широкому включению неологизмов.

В данном исследовании рассмотрены и другие проблемы, в том числе:

- менее используемые языки и языки меньшинств
- ситуация в отдельных тематических областях
- создание ТБД для конкретных целей или задач
- вопросы координации ТБД на национальном и европейском уровне
- этапы процесса создания и использования специализированных словарей
- возможность повторного использования ТБД
- фразеологизмы в текстах предметного поля как ресурс для ТБД
- распространение ТБД
- качество и валидация ТБД

*Выводы проекта*

Проект предлагает создать европейскую инфраструктуру для максимального использования терминологии теми, кто владеет такими ресурсами, теми, кто их создает, теми, кто их распространяет и теми, кто желает их использовать. Среди тех видов деятельности, которые, вероятно, выиграют от улучшенной организации ТБД, относятся: документация и индексация; письменный и устный перевод; техническое письмо; маркетинг; документация на продукцию; научные, технические и медицинские исследования; обучение и образование.

Проект содержит набор рекомендаций, определяющих ключевые характеристики общеевропейской инфраструктуры сотрудничества по терминологии для поддержки пользователей в создании, распространении и (повторном) использовании многоязычной терминологии, принимая во внимание необходимость решения проблем, связанных с техническими и методологическими инструментами, а также по вопросам экономики и авторского права. Проект *POINTER* призывает к инициативам по сотрудничеству и внедрению рекомендованной инфраструктуры с учетом появления общеевропейских инициатив, таких как Европейская ассоциация языковых ресурсов (ELRA).

Судя по современному состоянию терминологической деятельности в ЕС, большая часть рекомендаций проекта *POINTER* была принята и реализована.

### Терминологические структуры Еврокомиссии

В составе Еврокомиссии существует несколько структур, которые создают и обслуживают терминологические ресурсы ЕС.

Одна из них ***TermCoord***[1] – группа по координации терминологии при Европарламенте. Основная роль команды *TermCoord* заключается в оказании помощи переводчикам в выполнении их повседневных задач и содействии терминологическим исследованиям и управлению терминологией в подразделениях перевода, а также в увеличении вклада Европарламента в терминологическую базу данных ЕС IATE. *TermCoord* поддерживает обширный каталог ТБД, веб-сайтов, блогов и других терминологических ресурсов[2].

Другая структура – ***Центр знаний по интерпретации***[3], координирует переводческую деятельность в структурах Еврокомиссии. Практически эту деятельность осуществляет компания *DG Interpretation*.

Одно из направлений ее деятельности – предоставление в доступ разнообразных ТБД, число которых в европейских структурах весьма велико; они включают собственные ТБД ЕС, ТБД отдельных организаций ЕС, в том числе Евростата, ТБД стран-членов, ТБД международных организаций и т.д. Глоссарии составлялись на протяжении многих лет терминологами и переводчиками *DG Interpretation*. Эти глоссарии основаны на двух основных источниках: информации с совещаний, на которых работают устные переводчики, и документах Еврокомиссии или внешних документах, которые существуют на нескольких языках.

Кроме того, существует множество глоссариев от внешних организаций (Организации ООН, ОЭСР и т.д.). В общей сложности база данных DG Interpretation в настоящее время содержит 1089 глоссариев, содержащих более 200 тыс. терминов.

Приведем список основных европейских ТБД на основе Glossary Links[4]. Там же имеются ссылки на все перечисленные ТБД.

- IATE I (Interactive Terminology for Europe) – межведомственная терминологическая база данных ЕС. ИАТЕ используется в учреждениях и агентствах ЕС с лета 2004 года для сбора, распространения и совместного управления терминологией, относящейся к конкретному ЕС.
- EuroVoc – это многоязычный, междисциплинарный тезаурус, охватывающий деятельность ЕС. Он содержит термины на 23 языках ЕС (болгарском, хорватском, чешском, датском, голландском, английском, эстонском, финском, французском, немецком, греческом, венгерском, итальянском, латышском, литовском, мальтийском, польском, португальском, румынском, словацком, словенском, испанском и шведском), а также на трех языках стран-кандидатов на вступление в ЕС: македонском, албанском и сербском.
- Межинституциональное руководство по стилю содержит единые стилистические правила и конвенции, которые должны использоваться всеми учреждениями, органами и ведомствами Европейского союза.

---

[1] TermCoord. – URL: https://termcoord.eu / 1. (дата обращения: 01.12.2021).

[2] Terminology websites & blogs. – URL: https://termcoord.eu/terminology-websites (дата обращения: 01.12.2021).

[3] Knowledge Centre on Interpretation EU Terminology Sources. – URL: https://ec.europa.eu/education/knowledge-centre-interpretation/knowledge-centre-interpretation_en (дата обращения: 01.12.2021).

[4] Glossary Links – URL: https://termcoord.eu/glossarylinks (дата обращения 01.04.2022).

Глоссарии генерального директората ЕС и Евростата:

- Сельское хозяйство и развитие сельских районов
- Бюджет
- Коммуникации, контент и технологии
- Европейская политика соседства и переговоры о расширении
- Миграция и внутренние дела
- Региональная политика
- Налогообложение и Таможенный союз:
  - Таможенный глоссарий
  - Налоговый глоссарий
- Аббревиатуры
- Словарь PRADO[1]
- Комбинированная номенклатура
- Европейское агентство по окружающей среде
- Европейское управление по безопасности пищевых продуктов
- Агентство Европейского союза по сетевой и информационной безопасности
- Европейская информационная система о природе – база данных по биоразнообразию
- Многоязычная терминология Европейского химического агентства

Для переводчиков разработана специальная поисковая система ***Lithos***[2]. Она позволяет переводчикам искать термины во всех глоссариях *DG Interpretation* на основе их собственных заранее определенных профилей. Созданная с этой целью база данных обновляется каждый месяц, чтобы включить в нее все новые записи и любые новые глоссарии.

Сбор ТБД в странах – членах ЕС происходил также в проекте *ELRC*, в рамках которого собирались ЛИР, которые можно использовать в системах машинного перевода, разрабатываемого в ЕС. В рамках этого проекта было собрано 225 ЛИР, в том числе 50 ТБД. Все эти ЛИР размещены в репозитории *ELRC-SHARE*[3].

Вообще в структурах ЕС разработано или используется множество разнообразных глоссариев, терминологических словарей и тезаурусов. На портале открытых данных ЕС[4] размещено около 600 наборов данных, относящихся к этой категории.

[1] PRADO – публичный реестр подлинных проездных и удостоверяющих личность документов онлайн.

[2] Lithos. – URL: https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/terminology-tools-and-resources/terminology-dg-interpretation_en#lithos

[3] ELRC-SHARE repository. – URL: https://elrc-share.eu/ (дата обращения: 01.12.2021).

[4] Datasets. – URL: https://data.europa.eu/euodp/en/data/ dataset?q=Glossary+terms+terminology&ext_boolean=any&sort= (дата обращения: 01.12.2021).

## Мировые терминологические структуры

Важную роль в международной организации терминологической деятельности играют ***ООН и ЮНЕСКО.*** В таблицах 7 и 8 приведены перечни терминологических ресурсов этих организаций[1].

*Таблица 7*

**Тезаурусы ООН и ЮНЕСКО**

| Наименование | Описание |
|---|---|
| Тезаурус Продовольственной и сельскохозяйственной организации | AGROVOC – Многоязычный тезаурус ФАО, Продовольственной и сельскохозяйственной организации Объединенных Наций, на 14 языках |
| Общий многоязычный экологический тезаурус | GEMET – Общий многоязычный экологический тезаурус на 19 европейских языках, разработанный Европейским агентством по окружающей среде и Европейским тематическим центром по каталогу источников данных |
| Многоязычный тезаурус населения | POPIN – Многоязычный тезаурус по народонаселению |
| Тезаурус по беженцам | ITRT – Международный тезаурус терминологии беженцев (английский, французский, испанский) |
| Архивный тезаурус Великобритании | |
| Тезаурус библиографической информационной системы ООН | UKAT – контролируемый словарь для индексирования архивных коллекций и каталогов |
| Тезаурус Всемирного банка | UNBIS Тезаурус библиографической информационной системы Организации Объединенных Наций (английский, арабский, испанский, китайский, русский и французский) |

*Таблица 8*

**Глоссарии / Базы данных терминологии**

| Наименование | Описание |
|---|---|
| Глоссарий по безопасности Международного агентства по атомной энергии | В Глоссарии МАГАТЭ по безопасности (IAEA Safety Glossary) перечислены термины, обычно используемые в публикациях, связанных с безопасностью (только на английском языке) |
| Терминологическая база данных Международной организации труда | ILOTERM – терминологическая база данных Международной организации труда (английский, французский, немецкий и испанский) |
| Терминология Международного валютного фонда | Более 4500 записей терминов, полезных для переводчиков, работающих с материалами Международного валютного фонда (МВФ) (английский, французский, немецкий, португальский и испанский) |
| Глоссарии ООН | Страница ресурсов устных переводчиков Организации Объединенных Наций – двуязычные и многоязычные глоссарии (китайский, английский, французский, русский и испанский) |

[1] UNESCO thesaurus. Other thesauri, glossaries and terminology databases. – URL: https://web.archive.org/web/20110728143742/http://ftp.unesco.org/thesaurus/other.html

*Продолжение таблицы*

| Наименование | Описание |
|---|---|
| База данных многоязычной терминологии Организации Объединенных Наций | UNTERM – многоязычная терминологическая база данных Организации Объединенных Наций (арабский, китайский, английский, французский, испанский и русский) |
| Глоссарий Всемирной торговой организации | Этот глоссарий призван помочь понять некоторые термины, используемые в ВТО и в международной торговле |

Центральной ТБД ООН является база данных *Многоязычная терминология (UNTERM)*[1] – лингвистический инструмент, который преобразует терминологию, используемую в ООН на шести официальных языках ООН (арабском, китайском, английском, французском, русском и испанском языках). База данных содержит более 85 000 терминов и обновляется ежедневно. База данных ведется Справочно-терминологической секцией Отдела документации Департамента Генеральной Ассамблеи и конференционного управления со штаб-квартирой в Нью-Йорке.

Возможен общий поиск по терминологическим ресурсам ЮНЕСКО. Для этой цели создана специальная поисковая система *ЮНЕСКОТЕРМ*[2].

Еще одной известной международной организацией, работающей в сфере ТБД, является Международный консорциум ***TerminOrgs***[3]. Задачи этой организации:

- распространение лучших практик для управления терминологией;
- пропаганда роли терминологии для эффективных внутренних и внешних коммуникаций, передачи знаний, образования, особенно в крупных организациях;
- представление терминологических стандартов и инструментов;
- определение и пропаганда экономической ценности управленческой терминологии.

*TerminOrgs* разработано *Руководство по терминологической деятельности для корпораций*, а также спецификация упрощенного варианта формата TBX-Basic.

TBX-Basic – это XML-формат для записи терминологических данных, таких как термины, определения и примечания к использованию. Это более легкая версия TermBase eXchange[4], которая оформлена стандартом ISO 30042:2019. TBX-Basic отражает лучшие отраслевые практики для терминологических баз данных.

[1] UNTERM The United Nations Terminology Database. – URL: https://unterm.un.org/unterm/portal/welcome (дата обращения: 01.12.2021).

[2] UNESCOTERM Search. – URL: https://web.archive.org/web/ 20110728143741/ http://termweb.unesco.org/ (дата обращения: 01.12.2021).

[3] TerminOrgs. – URL: http://www.terminorgs.net/ (дата обращения: 01.12.2021).

[4] TBX Overview. – URL: https://www.tbxinfo.net/ (дата обращения: 01.12.2021).

Из публикаций на русском языке, посвященных зарубежному опыту создания ТБД, отметим работу Л.Г. Федюченко [4]. В ней сделан анализ структуры и функциональности семи известных ТБД.

В приложении 8 приводится аннотированный перечень с адресами 27 наиболее известных и крупных мировых ТБД, включая единственный российский ТБД Ростерм.

## Российские ТБД

Наиболее известным российским ТБД является ***Банк данных Российская терминология (терминологические словари) Ростерм***[1], содержащий свыше 140 тыс. терминологических статей из ГОСТ, ГОСТ Р, стандартов ИСО и МЭК, а также терминологических приложений к ним. Кроме того, в БД РОСТЕРМ введены наиболее актуальные термины из словарей Комитета научной терминологии в области фундаментальных наук (КНТ РАН) и из тематических словарей отечественных и международных научных обществ и ассоциаций. Термины и определения даны на русском языке, а также представлены эквиваленты терминов на английском. По желанию заказчика возможно представление эквивалентов терминов на немецком и французском языках.

В рамках Терминологического центра Института русского языка РАН[2] функционирует ***Терминологическая база знаний «Научная терминология» (ТБЗ НТ)***, включающая 12 компьютерных терминологических баз знаний, охватывающих различные дисциплины и области знаний – как гуманитарные (литературоведение, языкознание, философия), так и технические (робототехника, гидромеханика, гироскопия, информатика, персональные телекоммуникации и т.п.). Восемь ТБЗ соответствуют источникам на русском языке и представляют информацию на русском языке, остальные – источникам на английском языке и представляют информацию на английском языке, общий суммарный объем всех источников – около 2500 терминов.

ТБЗ НТ представляет собой терминологический банк данных, обладающий дополнительным уникальными возможностями анализа понятийной структуры терминологии и структуры знания той или иной области. Эти дополнительные возможности АИСНТ связаны с анализом определений терминов и делают возможным:

– получить и графически представить понятийную структуру терминологии соответствующей области знания, в которой каждое понятие характеризуется отношением к другим понятиям и своим уровнем в иерархии этих понятий (имеющим числовую характеристику);

– получить и графически представить родовидовую и цело-частную структуры терминологических понятий данной области знания;

[1] Каталог Банк данных Российская терминология (Терминологические словари). – URL: http://www.gostinfo.ru/catalog/terminlist (дата обращения: 01.12.2021).

[2] Терминологический центр ИРЯ РАН. – URL: http://www.ruslang.ru/terminol_results (дата обращения: 01.12.2021).

– получить разнообразные списки терминов, называющих специфические понятия данной области, анализировать структуру данной области знания и производить некоторые логические выводы.

Кроме органов стандартизации и профессиональных лингвистов ТБД, в России создают библиотеки, информационные центры, службы переводов, службы технического письма и другие организации. Учета ТБД в России нет, нам также неизвестны и проекты по интеграции ТБД.

## Память перевода

Разновидностью ТБД являются базы данных, называемые «памятью перевода» (translation memory, TM). Одна запись в такой базе данных соответствует сегменту, или «единице перевода», за которую обычно принимается одно предложение (реже – фразеологический оборот либо абзац). Помимо ускорения процесса перевода повторяющихся фрагментов и изменений, внесенных в уже переведенные тексты (например, новых версий программных продуктов или изменений в законодательстве), системы TM также обеспечивают единообразие перевода терминологии в одинаковых фрагментах, что особенно важно при техническом переводе.

В каждой конкретной системе TM-данные могут храниться в своем собственном формате (текстовый формат в Wordfast, база данных Access в Deja Vu), но существует международный стандарт TMX (Translation Memory eXchange format)[1], который основан на XML и может генерироваться практически всеми системами TM.

Большинство систем TM поддерживает создание и использование словарей пользователя, создание новых баз данных на основе параллельных текстов, а также полуавтоматическое извлечение терминологии из оригинальных и параллельных текстов.

### *Популярные программные системы TM*[2]

Существует более различных 50 программ памяти переводов. Ниже перечислены программы, которые популярны среди крупных русскоязычных бюро переводов – участников *translationrating.ru*.

Продукт и число компаний-пользователей на 03.2017

- Программы SDL – 122
- Memsource – 59
- Smartcat – 48
- memoQ – Kilgray* – 33
- STAR Transit – 20
- Across – 17
- Lionbridge Translation Workspace – 17
- WordFast – 17
- Atril DejaVu – 16
- XTM – 14

[1] Расширение файла TMX. – URL: https://www.file-extension.info/ru/format/tmx

[2] Топ-10 программ памяти переводов. – URL: https://translationrating.ru/top-10-cat-tools-2017/

***Стандарты и форматы памяти переводов***

- *TMX* – ссылка выше.
- *TBX* (Termbase Exchange format – обмен терминологическими базами). Это принятый ассоциацией LISA[1] формат пересмотрен согласно ISO 30042. Этот стандарт позволяет проводить обмен терминологией, в том числе детальной лексической информацией. Основная база TBX определяется стандартами: ISO 12620, ISO 12200 и ISO 16642.
- *SRX* создан для улучшения формата TMX и большей эффективности обмена памятью переводов между программами. Возможность указывать правила сегментации, которые использовались в предыдущем переводе, повышает эффективность отождествления сегментов в текущем тексте с содержимым ПП.
- *GMX GILT* означает Globalization, Internationalization, Localization and Translation (Глобализация, интернационализация, локализация, перевод). Стандарт GILT Metrics состоит из трех частей: GMX-V для показателей объема, GMX-C для показателей сложности, GMX-Q для показателей качества. Стандарт GILT Metrics направлен на квантификацию объема работ и требований качества переводов.
- *OLIF* – открытый стандарт, совместимый с XML, который используется для обмена терминологическими и лексическими данными. Хотя изначально он применялся в качестве способа обмена лексическими данными между частными лексиконами машинного перевода, постепенно этот формат превратился в более общий стандарт терминологического обмена.
- *XLIFF* (XML Localisation Interchange File Format – XML формат для взаимного обмена при локализации) создан как единый формат файлов для взаимного обмена, который распознается всеми программными средствами локализации. XLIFF – это наилучший в современной индустрии переводов способ обмена информацией в формате XML. Некоторые инструменты используют проприетарные форматы XLIFF, не позволяющие открывать созданные в них файлы в других программах.
- *TransWS* (Translation Web Services – переводческие веб-сервисы) определяет требуемые параметры вызова веб-сервисов при отправлении и получении файлов и сообщений, имеющих отношение к проектам локализации. Задумывался как развернутая система автоматизации процесса локализации с использованием сервисов в сети Интернет.
- *xml:tm* – этот подход к памяти переводов основан на концепции текстовой памяти, которая позволяет совмещать авторскую память и память переводов. Формат xml:tm был передан Lisa OSCAR компанией XML-INTL.

## Номенклатуры, классификации, таксономии

Еще одной разновидностью ЛИР, которые мы с определенной степенью условности относим к терминологическим ЛИР, являются различные

[1] The Localization Industry Standards Association (LISA). – URL: https://www.translationdirectory.com/article371.htm (дата обращения: 01.12.2021).

перечни наименований объектов реального мира, иногда систематизированные, иногда снабженные разнообразными идентификаторами. Лексические единицы таких перечней иногда называют номенами, а сами перечни могут называться номенклатурами, классификаторами, таксономиями. Частично номены являются именами собственными (персоны, организации, географические наименования), иные – нарицательными (наименования продукции, химические вещества, биологические объекты), а часть представляют имена таксономических группировок, обычно в виде фразеологических оборотов.

Особенностью этой категории ЛИР является то, что разные типы ЛИР создавались в различных традициях и, соответственно, оформлялись по различным моделям. Часть из них имеет официальный характер и утверждается в качестве стандартов или других нормативных документов – некоторые на международном уровне, другие на национальном.

Некоторые из них создавались в лексикографической традиции (топонимика, ономастика), часть – в библиотечной (авторитетные файлы авторов, мест изданий, издательств), другие – в научно-информационной (тезаурусы, рубрикаторы, коды стран, языков, словари метаданных). Некоторые ЛИР формировались в соответствующей научной дисциплине еще до появления информационных систем и имели международный характер (химическая, биологическая, минералогическая номенклатуры, классификация болезней), но большинство – в различных информационно-управляющих системах (продукция, профессии, услуги, виды деятельности, валюты и прочее).

В России многие перечни объектов были включены в состав Общесоюзных классификаторов технико-экономической информации, которых в настоящее время насчитывается 32. Термины, именующие подотрасли или дисциплины различных сфер деятельности, являются основой рубрикаторов научной информации. Таких рубрикаторов зарегистрировано в рамках учета информационных языков в Государственной системе научно-технической информации свыше 300. Перечни общесоюзных классификаторов и локальных рубрикаторов приводятся в работе [5].

Самостоятельным крупным ЛИР является Государственный каталог географических названий[1]. В нем размещены реестры наименований географических объектов по каждому субъекту Российской Федерации в алфавитной последовательности наименований всех географических объектов, а также наименований географических объектов континентального шельфа и исключительной экономической зоны Российской Федерации, географических объектов, открытых или выделенных российскими исследователями в пределах Открытого моря и Антарктики.

Наибольшее разнообразие демонстрирует лексика, отражающая разделы и дисциплины науки и образования. Она организована во множестве классификаций – библиотечных, статистических, управленческих и других. Анализ этих классификаций, как международных, так и национальных, можно также найти в цитируемой выше книге Р.С. Гиляревского и соавторов.

[1] Росреестр. – URL: https://cgkipd.ru/science/names/reestry-gkgn.php (дата обращения: 01.12.2021).

Приведем перечень номенклатур еще некоторых классов объектов, образующих самостоятельные ЛИР, но не вошедших в число классификаторов или стандартов. Некоторые из них утверждены нормативными актвми, другие являются стандартами де-факто.

***Номенклатуры, утвержденные нормативные актами***

1. Регистр химических соединений CAS. Обязательность применения установлена в НПА «Предельно допустимые концентрации (ПДК) загрязняющих веществ в атмосферном воздухе населенных мест» ГН 2.1.6.13 1338–03"#, утвержденные Главным государственным санитарным врачом Российской Федерации 21 мая 2003 г.
2. «Номенклатура продукции и услуг (работ), в отношении которых законодательными актами Российской Федерации предусмотрена их обязательная сертификация», Постановление Правительства РФ от 10.02.2004 г. № 72, Приказ Ростехрегулирование от 22.07.2004 г. № 7.
3. «Номенклатура продукции, соответствие которой может быть подтверждено декларацией соответствия». Постановление Правительства РФ от 10.02.2004 г. № 72, Приказ Ростехрегулирование от 22.07.2004 г. № 7.
4. Номенклатура клинических лабораторных исследований. Минздрав России, Приказ от 21 февраля 2000 г. № 64.
5. Номенклатура учреждений (отделений) социального обслуживания граждан пожилого возраста и инвалидов. Письмо Минтруда РФ от 5 января 2003 г. № 30-ГК.
6. Единая номенклатура государственных и муниципальных учреждений здравоохранения. Приложение к Приказу Минздрава России от 3 июня 2003 г. № 229.
7. Единая тарифно-статистическая номенклатура грузов (ЕТСНГ). Указания МПС России от 2 декабря 1996 г. № К-1058 у и от 3 апреля 1997 года № 404 у.
8. Номенклатура военно-медицинских учреждений внутренних войск Министерства внутренних дел Российской Федерации. Приложение к приказу МВД РФ от 20 марта 2001 г. № 298.
9. Номенклатура объектов сертификации в строительстве // Основные положения сертификации продукции в строительстве РДС 10–231–93* Приложение Б (обязательное).
10. Номенклатура основной подкарантинной продукции. Утверждена Заместителем Министра сельского хозяйства Российской Федерации В.И. Алгиниым 25 декабря 2001 г.
11. Номенклатура специальностей научных работников. Министерство образования и науки Российской Федерации, Приказ от 23 октября 2017 г. № 1027.
12. Единая тарифно-статистическая номенклатура грузов (Утв. МПС России как Приложение к прейскуранту № 10–01).
13. «Номенклатура и объемы важнейших лекарственных средств, по которым целесообразно осуществить импортные закупки», утв. Минздравом РФ 26.10.98.

14. Номенклатура врачебных и провизорских специальностей в учреждениях здравоохранения РФ, Приказ Минздрава России от 16.02.95 № 33.

***Научные номенклатуры, фактически являющиеся официальными, но не утвержденные нормативно-правовыми актами***

15. Регистр минералов. Международная Минералогическая Ассоциация, при которой существует Комиссия по Новым Минералам и Названиям Минералов. Порядок ведения определен в документе [6].

16. Международная ветеринарная анатомическая номенклатура на латинском и русском языках [7].

17. Международный кодекс зоологической номенклатуры [8].

18. Международный кодекс ботанической номенклатуры [9].

19. Международный кодекс номенклатуры бактерий [10]. – М.: Наука, 1978.

20. Международная номенклатура нарушений, ограничений жизнедеятельности и социальной недостаточности [11]. Руководство по классификации последствий болезней и причин инвалидности / Рос. акад. мед. наук, НИИ социал. гигиены, экономики и управления здравоохранением им. Н.А. Семашко; [Подгот. В.К. Овчаров и др.]. – М.: НИИСГЭИУЗ, 1994. 100 с.

Этот перечень не является исчерпывающим, эти номенклатуры приведены в качестве примера. Однако при создании словарной службы России будет стоять задача создания именно исчерпывающего перечня всех объектов реального мира, сведения о которых понадобятся в информационных системах обработки естественного языка.

*Инструментарий терминологических ЛИР*. Для создания ТБД используется разнообразный программный инструментарий. В руководстве TerminOrgs [2] предлагается следующая классификация.

***Электронные таблицы и приложения для обработки текстов***

Часто используются для сбора списков важных терминов до создания терминологических БД. Преимущество*:* легко доступны. Недостатки: не способны правильно управлять терминологией на основе концептуального подхода; не интегрированы в инструменты перевода.

***Сопутствующий модуль в более крупном программном решении***

Инструменты перевода, авторские инструменты, корпоративный поиск и инструменты управления таксономией требуют хранения одноязычной или многоязычной терминологии. Поэтому они часто включают в себя некоторые функции управления терминологией (TMS). Доступные функции и уровень сложности варьируются от очень простых (плоский список терминов) до очень сложных (многофункциональный, концептуальный). Преимущество: интегрированы в существующее используемое программное обеспечение. Недостатки: часто не в состоянии обеспечить надлежащее управление терминологией, так как функции управляются родительским программным обеспечением (функции управления терминологией в инструментах перевода или авторских инструментах часто имеют меньший объем, чем собственно TMS).

***Специальный терминологический инструмент***

Полнофункциональное программное обеспечение с основной целью хранения, поддержания и распространения терминологических данных. Это может быть серверное и / или веб-приложение, а также автономное программное обеспечение. Преимущество: адекватно для терминологических нужд. Недостаток: интеграция с переводческими или авторскими решениями может потребовать отдельного программного обеспечения «интегратора» или процесса экспорта / импорта.

Управление терминологией, согласно руководству специализированной организации[1], включает следующие функции:

- оценка объема необходимых действий по управлению терминологией;
- проверка существующей терминологии, глоссариев, руководств по стилю и т.д., понимание основной терминологии клиента;
- извлечение и перевод терминологии. Использование средств управления терминологией;
- создание или обновление существующих глоссариев;
- постоянная поддержка баз данных терминологии;
- создание, проверка и обновление руководств по стилю;
- управление вопросами и ответами;
- поддержка переводческой памяти;
- доступ к базам данных терминологии.

***Статистический и лингвистический подходы***

Некоторые инструменты извлечения терминов находят термины только с помощью статистики (какие слова или словосочетания встречаются чаще всего). Поскольку подход является чисто статистическим, эти инструменты часто одинаково хорошо работают на любом языке.

Другие инструменты интегрируют лингвистические правила, чтобы помочь определить, какие слова или словосочетания скорее всего будут терминами – обычно это именные сочетания, а не функциональные слова, такие как предлоги и артикли, фразы или случайные строки слов. Поскольку эти инструменты опираются на лингвистические правила, они обычно работают только на одном языке или ограниченном количестве языков. Однако результаты, как правило, лучше чем чисто статистические инструменты.

Другой тип инструмента может обрабатывать параллельные тексты (пара исходного текста и его перевода, выровненные) и извлекать термины на обоих языках: термины исходного языка и один или несколько вероятных переводов для каждого. Некоторые из этих инструментов требуют, чтобы человек проверил предложенный перевод исходного термина, в то время как другие могут быть настроены на выбор того, что статистически является «наиболее вероятным» совпадением.

[1] Управление терминологией. – URL: http://www.ptb-localization.com/ru/terminology.html (дата обращения: 01.12.2021).

***Учет терминов в контексте***

Большинство инструментов извлечения терминологии позволяют просматривать все или несколько вхождений термина в контексте. Возможность просматривать некоторые предложения, в которых встречается этот термин, облегчает сначала проверку того, что термин существует и в случае многословных терминов правильно разграничен, а также поиск другой информации, такой как определения, синонимы и примеры использования.

***Другие особенности***

Другие важные функции, которые следует искать в инструменте извлечения терминов:

- извлечение как однословных терминов, так и многословных;
- указание частоты встречаемости термина;
- указание части речи термина;
- имя файла, из которого был извлечен термин;
- наличие списка исключений.

***Система управления терминологией (TMS*)** – это программный инструмент, специально разработанный для сбора, поддержки и доступа к терминологическим данным.

Далее приводятся примеры TMS, согласно специализированному порталу[1] (там же имеются ссылки на все перечисленные продукты):

- **Acrolinx** – платформа оптимизации контента, которая интегрируется с инструментами для письма, указывает на языковые проблемы и дает предложения по улучшению;
- **Anylexic** – новое поколение программ управления терминологией, не привязанных к какой-либо конкретной терминологии. Он может помочь вам на каждом этапе процесса управления терминологией перевода: создание, редактирование, поиск и обмен;
- **ApSIC Xbench** – обеспечивает простое и эффективное управление качеством и терминологией в одном пакете. Просто загрузите файлы в любом из десятков поддерживаемых форматов CAT и поднимите качество своего перевода на новый уровень;
- **evoTerm** – централизованно хранимая терминология, доступная через Интернет. Демоверсия доступна для тестирования платформы;
- **flashterm** – управление терминологией и управление знаниями;
- **InterpretBank** – инструмент управления терминологией, специально разработанный для переводчиков. Это помогает создавать, изучать и искать глоссарии даже в кабине;
- **Intragloss Home** – Intragloss – это профессиональный глоссарий, который дает вам все необходимое для подготовки к заданиям;
- **i-Term** – современный инструмент управления терминологией и знаниями, который позволяет хранить, структурировать и искать знания о концепциях в Интернете;

[1] Terminology Management Systems. – URL: http://recremisi.blogspot.com/p/acrolinxterminology-lifecycle.html (дата обращения: 01.12.2021).

- **LogiTerm Web** имеет удобный веб-интерфейс, который обеспечивает доступ к четырем базам данных: «Терминология», «Битексты», «Полный текст» и «Справочник». Записи терминологии для базы данных терминологии можно создавать, изменять и просматривать в веб-интерфейсе или в Microsoft Word;
- **qTerm ™** – программное обеспечение для управления терминологией через Интернет. Он определяет и переводит важную терминологию. Он также предоставляет подробное объяснение использования каждого термина, включая контекст, язык и историю использования;
- **quickTerm** – система управления жизненным циклом терминологии, основанная на базе данных SDL MultiTerm или Acrolinx. Расширяет охват терминобазы для многих различных пользователей, делая ее более доступной. Кроме того, quickTerm позволяет команде терминологов разрабатывать сложные рабочие процессы терминологии в масштабах всей компании на основе многолетних данных SDL и эффективно управлять жизненным циклом терминологии;
- **SDL MultiTerm Desktop** – инструмент управления терминологией рабочего стола от SDL. Он может быть использован в качестве автономного настольного инструмента для управления всей корпоративной терминологией, или его мощность может быть увеличена в среде перевода за счет интеграции с SDL Trados Studio;
- **Termbases** – мощное веб-программное обеспечение для создания многоязычных терминологических ресурсов и управления ими;
- **TermWeb –** обеспечивает согласованность языкового перевода и капитала бренда во всех точках соприкосновения, от страны к стране и по всему миру;
- **TermWikiPro** – защищенная облачная система управления терминологией, разработанная для того чтобы помочь глобальным предприятиям ускорить бизнес; предоставляет полные готовые решения для улучшения качества контента при одновременном снижении затрат на создание и перевод;
- **Taas** – облачные сервисы для работы с терминологией: предоставляет многоязычные и совместные терминологические услуги;
- **TippyTerm**, разработанный для всех систем MS Windows и обеспечивающий последовательное использование терминологии, легкую доступность, простоту в обращении, простое обслуживание;
- **Interplex** – программное обеспечение для глоссария для устных и письменных переводчиков;
- **Interpreters Help** – инструменты для устных переводчиков конференций.

Еще один пример описания инструмента управления терминологией, который называется *TermX Terminology Tool*, сделан его автором [12]. База данных TermX, состоящая из четырех различных таблиц, основана на очень простой модели данных, так что ее можно быстро и эффективно использовать для создания расширяемых баз или автономных хранилищ терминологии для переводчиков. Хотя TermX не предназначен для полноценного управления

терминологией или лексикографической системой для профессионалов, TermX все же может быть использован для создания выходных файлов, которыми можно дополнительно манипулировать для онлайн- или печатного глоссария. TermX – это программный пакет, который может использоваться как независимыми переводчиками, так и переводческими агентствами для создания обширных хранилищ терминологии для конкретной предметной области.

Среди инструментов управления терминологией важное место занимают программы извлечения из текстов терминов и номенов, которые имеют общее название ***именованные сущности.***

Достаточно полный перечень таких программ с их главными характеристиками приводится в отечественном каталоге[1].

Добавим в этот перечень указание на отечественную разработку, сделанную в *Лаборатории информационных исследований*[2]. Этот инструмент обеспечивает автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода. При этом используются размеченные коллекции для извлечения именованных сущностей, которые доступны для использования.

## Литература к главе 12


1. ISO 30042:2008 Systems to manage terminology, knowledge and content – TermBase eXchange (TBX). – URL: https://www.iso.org/standard/45797.html (дата обращения: 01.12.2021).
2. Terminology Starter Guide/Terminology for Large Organizations 2016. – URL: https://terminorgs.net/downloads/TerminOrgs_StarterGuide_V2.pdf (дата обращения: 01.12.2021).
3. Terminology resources. Scope and quality of resources. – URL: http://www.computing.surrey.ac.uk/AI/pointer/report/section4 (дата обращения: 01.12.2021).
4. Федюченко Л.Г. Структурно-функциональные параметры терминологической базы данных // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. – 2017. – Т. 3, № 2. – С. 45–58. – DOI: 10.21684/2411–197 X-2017–3-2–45–58
5. Гиляревский Р.С., Шапкин А.В., Белоозеров В.Н. Рубрикатор как инструмент информационной навигации. – Санкт-Петербург : Профессия, 2008. – 352 с.
6. Никель Э.Х., Грайс Дж. Д. Комиссия по новым минералам и названиям минералов Международной минералогической ассоциации: правила и руководства по номенклатуре минералов, 1998 // Зап. Всеросс. минерал. о-ва. – 1999. – Ч. 128, вып. 2. – С. 51–65.
7. Зеленевский Н.В. Nomina Anatomica Veterinaria : учебное пособие. – Санкт-Петербург : Лань, 2013. – 400 с. – ISBN 978-5-8114-1492-5. – URL: https://e.lanbook.com/book/5706 (дата обращения: 01.12.2021).
8. Международный кодекс зоологической номенклатуры : пер. с англ. и фр. Второе, исправленное издание русского перевода. – Издание четвертое. Принят Международным союзом биологических наук. – Москва : Т-во научных изданий КМК, 2004. – 223 с. – URL: https://www.iczn.org/assets/d5f35966d3/Code_Russian-Edition-2004_0.pdf (дата обращения: 01.12.2021).


[1] NL-PUB. – URL: https://nlpub.ru (дата обращения: 01.12.2021).

[2] Размеченные коллекции для извлечения именованных сущностей. – URL: http://labinform.ru/pub/named_entities/descr_ne.htm (дата обращения: 01.12.2021).

9. Международный кодекс ботанической номенклатуры (Венский кодекс). Принят Семнадцатым международным ботаническим конгрессом, Вена, Австрия, июль 2005 года / пер. с англ. Т.В. Егоровой [и др.]. – Москва ; Санкт-Петербург : Т-во науч. изд. КМК, 2009. – 282 с. – 800 экз. – ISBN 978-5-87317-588-8.

10. Международный кодекс номенклатуры бактерий. – Москва : Наука, 1978.

11. Международная номенклатура нарушений, ограничений жизнедеятельности и социальной недостаточности. Руководство по классификации последствий болезней и причин инвалидности / Рос. акад. мед. наук, НИИ социал. гигиены, экономики и управления здравоохранением им. Н.А. Семашко ; подгот. В.К. Овчаров [и др.]. – Москва : НИИСГЭиУЗ, 1994. – 100 с. – 29 см.

12. Graham P. TermX Terminology Tool. – URL: https://www.translatorscafe.com/Cafe/images/Articles/TermX_Terminology_Tool.pdf (дата обращения: 01.12.2021).

# ГЛАВА 13. ТИПОЛОГИЧЕСКИЕ ЛИР

## Общие сведения

Важным типом ЛИР являются типологические БД (ТИПБД), обеспечивающие структурированное представление данных о свойствах языков, в том числе содержащие классификации языков в различных аспектах: генеалогическом, ареальном, эволюционном. ТИПБД создаются для задач контрастивной лингвистики, лингвистической типологии и языковой систематики и других лингвистических дисциплин.

«Среди ТИПБД выделяются ТИПБД широкого профиля и специализированные ТИПБД. Типологические базы данных широкого профиля (или «характеристические» ТИПБД) представляют собой универсальные хранилища данных о широком спектре характеристик различных естественных языков; они могут использоваться как справочные инструменты, а также при решении задач, связанных с классификацией языков. Специализированные ТИПБД используются при детальном исследовании отдельных языковых явлений на примере ограниченных наборов языков: они тщательно моделируют структуру этих явлений и включают примеры, иллюстрирующие их функционирование в рассматриваемых языках»[1].

ТИПБД содержат описания в формализованном виде фонологических, грамматических и лексических свойств языков и сопровождаются инструментальными средствами интерфейса и статистических расчетов.

Анализ функциональности ТИПБД можно найти в работе В.Д. Соловьева [2].

«ТИПБД исходно создавались как справочники с удобным пользовательским интерфейсом, позволяющим быстро находить интересующую информацию. Однако быстро выяснилось, что ТИПБД предоставляют принципиально новые возможности изучения грамматик языков с применением математических (в том числе, статистических) и компьютерных методов. Многие явления, которые до сих пор рассматривались лишь на качественном уровне и на основе отдельных примеров, теперь могут изучаться количественными методами с использованием огромных массивов информации. Важным аспектом подобных исследований является их объективный характер, основанный на применении строгих математических методов. Вот несколько примеров вопросов, на которые появилась возможность дать ответ с помощью ТИПБД.

1. Насколько однородным является тот или иной языковой ареал? Можно ли считать его языковым союзом? ТИПБД позволяют применить количественные методы в ареальной лингвистике для оценки степени близости языков.

2. Как происходило распределение языковых признаков в связи с распространением человечества и собственно языковой эволюцией? Пионерские исследования Дж. Николс в этом направлении проводились на очень ограниченной выборке данных. Современные ТИПБД могут помочь уточнить многие аспекты расселения человечества.

3. Установление (дальнего) языкового родства. Как принято считать, сравнительно-исторический метод позволяет реконструировать историю развития языков не более чем на 8–10 тыс. лет. Есть надежда, что тщательный анализ грамматических свойств (многие из которых, вероятно, более стабильны, чем лексические) позволит выявить сверхдальнее родство.

4. Языковая динамика: какие разделы грамматики языков меняются быстрее? С какой скоростью?»

Координацию исследований в области лингвистической типологии осуществляет профильная Ассоциации лингвистической типологии (ALT)[1], в настоящее время насчитывающая более 600 членов. Целью ALT является продвижение научного исследования типологии, т.е. межъязыкового разнообразия и лежащих в его основе закономерностей. С этой целью ALT стремится способствовать взаимному осознанию, диалогу и сотрудничеству в международном сообществе типологов; придать типологии более высокий статус, как в лингвистике, так и за ее пределами.

Наиболее полным собранием ТИПБД является портал Ассоциации лингвистической типологии, на котором есть специальный раздел *Базы данных*[2]. К сожалению, там не отражены российские ТИПБД. Достаточно полный обзор типологических интернет-ЛИР имеется также в работе Моравчика «Введение в лингвистическую типологию» [3].

ТИПБД являются основным инструментом для контрастивной лингвистики. В этой связи отметим проект *Кросс-лингвистические связанные данные (CLLD)*[3], описанный в статье Р. Форкеля [4]. В этой статье содержится также обзор существующих типологических БД. Проект CLLD более подробно описан в главе 19.

В сферу лингвистической типологии обычно включают и вопросы языковой систематики. *Языкова́я систе́матика* – вспомогательная дисциплина, помогающая упорядочивать изучаемые лингвистикой объекты – языки, диалекты и группы языков. Результат такого упорядочивания также называется систематикой языков.

[1] Association for Linguistic Typology. – URL: https://linguistic-typology.org/ (дата обращения: 01.12.2021).

[2] ALT. Databases. – URL: https://linguistic-typology.org/databases/ (дата обращения: 01.12.2021).

[3] Cross-Linguistic Linked Data project. – URL: https://clld.org/ (дата обращения: 01.12.2021).

В основе систематики языков лежит их генетическая классификация: эволюционно-генетическая группировка является естественной, а не искусственной, она достаточно объективна и устойчива (в отличие от зачастую быстро меняющейся ареальной принадлежности). Целью языковой систематики является создание единой стройной системы языков мира на основе выделения системы лингвистических таксономических уровней и соответствующих названий, выстроенных по определенным правилам (лингвистическая номенклатура). Термины «систематика» и «таксономия» часто используют как синонимы.

Для систематики характерны следующие принципы:

- единая иерархически организованная система;
- единая система таксонов;
- единая система номинации.

### Зарубежные типологические ЛИР

В мире насчитывается несколько тысяч языков. Наиболее известные справочники языков – *Этнолог* и *Реестр Лингвосферы*. Подробное описание и сопоставление этих ресурсов можно найти в работе Ю. Б Корякова и Т. Майсака [5]. Это описание дано по единой схеме, включающей:

- наличие единой иерархически организованной системы и принципы организации;
- принципы выделения языков / диалектов;
- наличие единой системы таксонов;
- наличие единой системы номинации языков (и более мелких единиц);
- наличие единой системы номинации групп языков;
- дополнительные названия на языке описания и на других языках;
- объем сведений для каждого языка;
- общий объем работы: количество языков; количество глоттонимов;
- включение мертвых, искусственных языков, пиджинов, языков глухонемых;
- наличие приложений: указатели, карты, библиографии;
- доступность для пользования и для исправления.

Видимо, наиболее полным справочным ЛИР по языкам мира является ***Этнолог***[1], который используется для идентификации языков наряду с их генетическим и ареальным происхождением. Он включает свыше 7 тыс. языков, предоставляет возможность поиска по наименованиям и кодам языков, странам, языковым семьям. Данный ресурс также представляет возможности формировать различные указатели, в том числе на коммерческой основе.

Из зарубежных типологических ДИР известность получил также ***Реестр Лингвосферы***[2]. Подробное его описание представлено на портале Лингвариум, фрагмент которого мы приводим:

[1] Ethnologue. – URL: http://www.ethnologue.com (дата обращения: 01.12.2021).

[2] Linguasphere. – URL: http://www.linguasphere.net/ (дата обращения: 01.12.2021).

«В Реестр включена информация о всех живых языках Земли. Из мертвых языков учтены: во-первых, те, которые в письменной форме продолжают использоваться в настоящее время (например, латынь, санскрит, церковнославянский и др.), причем предполагается учесть и все языки, от которых остались какие-либо письменные памятники (этрусский, древнекитайский, хеттский и пр.); во-вторых, языки, исчезнувшие прежде всего в течение XX века (убыхский, айнский и др.), а по возможности и за последние пять столетий (например, полабский, готский, многие языки Америки, Австралии и других регионов, исчезнувшие в процессе экспансии европейских языков), – поскольку языки как первой, так и второй групп непосредственно влияли и влияют на современное состояние лингвосферы.

Система классификации языков, принятая в Реестре, является оригинальной разработкой Дэвида Долби. В ее основу положены принципы в чем-то традиционные, но в чем-то весьма отличные от других указателей языков (прежде всего (link) "Этнолога"). Хотелось бы подчеркнуть, что Долби не ставит своей задачей создать принципиально новую (генетическую, ареальную) классификацию языков. Его задача – разработать достаточно простой и удобный принцип каталогизации языков мира в подробном Реестре.

Этот каталог языков можно сравнить с каталогом библиотеки: он устроен так, чтобы объединить языки и группы языков как можно более наглядным образом, облегчающим работу с указателем. Наиболее крупными референциальными единицами (так сказать, "стеллажами библиотеки") являются 10 секторов, каждый из которых подразделяется на 10 зон (которые можно сравнить с полками на стеллаже). Как сектора, так и зоны выделяются либо по генетическому принципу (это, соответственно, "филосектора" и "филозоны"), так и по ареальному ("геосектора", "геозоны") в случае, если последнее основание выделения предпочтительнее» [6].

***Всемирный атлас языковых структур WALS***.[1]

Из собственно типологических ЛИР наиболее известным является *WALS*. По мнению ряда исследователей [7], создание *WALS* ознаменовало новый этап развития лингвистической типологии.

*WALS* создан в 2005 году объединенным коллективом типологов разных стран в рамках проекта отдела лингвистики Института эволюционной антропологии Общества Макса Планка в Лейпциге. С 2008 года *WALS* представлен в Интернете. В базе данных представлены 192 различные языковые характеристики. В основном это грамматические и синтаксические характеристики, но присутствуют также фонетические и даже лексические. В качестве отдельного параметра есть и тип системы письма. Обратившись к *WALS*, можно выяснить даже, насколько верно предположение, что в языках местоимения первого лица часто содержат звук [м], а местоимения второго лица – звук [т]. База снабжена краткими статьями работавших над ней лингвистов, где описываются и комментируются включенные в WALS языковые явления.

[1] The World Atlas of Language Structures (WALS). – URL: https://wals.info/ (дата обращения: 01.12.2021).

Возможно составление пользовательских карт, где комбинируются данные по нескольким. Используемые параметры выбраны очень удачно и продуманно, и в целом дают представление о варьировании важнейших типологических переменных по языкам мира. Авторы проекта преследовали цель отразить все существующие семьи и ареалы, поэтому охарактеризованные языки равномерно распределяются по территории планеты. Выборка языков различна для каждого параметра. В отдельных случаях она относительно невелика, но для некоторых параметров (например «базовый порядок слов») достигает 2 тыс. языков, что дает картину, по богатству сопоставимую с языковым разнообразием в целом. Однако далеко не каждый язык представлен по всем параметрам, фактически лишь немногие описаны более чем по половине из них.

***Система типологических баз данных (TDS)***[1]

В мире создано уже несколько систем, включающих целые наборы ТИПБД. Так, сразу много независимо созданных типологических баз данных объединены в Систему типологических баз данных (***TDS***), расположенную на сайте Университета Утрехта.

TDS – это веб-сервис, обеспечивающий интегрированный доступ через единый интерфейс к коллекции независимо созданных типологических баз данных (описание некоторых из них представлено ниже):

- Anaphora Typology Database
- Free Personal Pronoun System
- Graz Database on Reduplication
- Person Agreement Database
- Smith's Phoneme Inventories
- Stress Typology Database
- Syllable Typology Database
- Typological Database Amsterdam
- Typological Database of Intensifiers and Reflexives
- Typological Database Nijmegen
- Topic Focus Database
- UCLA Phonological Segment Inventory Database
- World Color Survey

Дополнительную информацию можно найти в документации TDS [8].

***Программа типологических исследований AUTOTYP***[2]

Другая система типологических баз совместно поддерживается учеными из Лейпцигского университета и Калифорнийского университета в Беркли.

AUTOTYP – это крупномасштабная исследовательская программа в области как количественной, так и качественной типологии. В количественной типологии задача заключалась в обнаружении и объяснении географического распределения типологических особенностей и в получении статистических оценок универсальных предпочтений, а также генеалогической

[1] Typological Database System Curator. – URL: https://doi.org/10.17026/dans-xc9-mnrf (дата обращения: 01.12.2021).

[2] AUTOTYP. – URL: https://www.autotyp.uzh.ch/theory.html (дата обращения: 01.12.2021).

наследственности и потенциалов ареальной диффузии. В отношении качественной типологии проводился систематический анализ разновидностей вариаций, обнаруженных в различных типологических областях.

AUTOTYP был разработан в ответ на две проблемы, с которыми сталкиваются традиционные ТИПБД. Эти ТИПБД обычно полагаются на статический и заранее определенный список категорий, который может конфликтовать с данными по мере ввода большего количества языков. Традиционные базы данных обычно объединяются в один файл, содержащий широкий спектр информации, что затрудняет, если не делает невозможным, повторное использование любой части этой информации в других базах данных или для поиска типологических корреляций между ними.

Программа AUTOTYP устраняет эти недостатки и предлагает общие принципы разработки типологических баз данных.

*Автотипология.* Базы данных AUTOTYP автоматизируют типологию. Вместо того, чтобы начинать с заранее определенного списка категорий, базы данных AUTOTYP полагаются на автоматическое создание списков категорий во время ввода данных.

*Точность*. Базы данных AUTOTYP стремятся к как можно более детальной разбивке описательных понятий на однозначные термины.

*Примерный метод*. Парадигмы часто неоднородны, языки часто имеют конкурирующие конструкции в одной и той же структурной области,. Чтобы выбрать на каждом языке сопоставимые данные, мы следуем так называемому методу, основанному на образцах: мы выбрали один конкретный образец парадигм или структурных областей, и этот образец идентифицируется в соответствии со стандартным алгоритмическим определением.

*Модульность*. Чтобы достичь максимальной гибкости в формулировании запросов, базы данных AUTOTYP распределяют информацию по Сети из нескольких отдельных, тематически определенных файлов, связанных вместе в реляционной сети с помощью стандартизованных языковых идентификационных кодов.

*Связь*. Модули базы данных AUTOTYP связаны друг с другом посредством числовых языковых идентификационных кодов, которые могут быть сопоставлены с другими кодами, такими как идентификационный код Ethnologue.

Принципы автотипологии требуют, чтобы базы данных различали файлы данных, которые содержат записи по конкретным вопросам по языку, и файлы определений, что содержат записи понятий и их определений, которые оказываются необходимыми в файлах данных. Два типа файлов позволяют двойное использование базы данных в исследованиях: файлы данных позволяют количественно типологически исследовать статистические корреляции между структурными, генеалогическими или географическими объектами, в то время как файлы определений вносят вклад в качественную типологию, поскольку они содержат все необходимые понятия, которые являются кросслингвистически актуальными и жизнеспособными.

***Атлас структур пиджина и креольского языка APiCS***[1]

Этот атлас, который создан в Оксфордском университете, содержит грамматическую и лексическую информацию о большом количестве пиджинских и креольских языков по всему миру.

*APiCS Online* позволяет выяснить, какие лингвистические характеристики имеют различные языки контактного происхождения, т.е. пиджины, креольские и смешанные языки. В базе на данный момент описано 76 языков, каждый из которых охарактеризован по 130 параметрам. Среди этих параметров есть фонетические (наличие носовых гласных, тонов, губно-зубных щелевых согласных), лексические (различаются ли обозначения синего и зеленого цвета, одним словом или разными обозначаются рука и палец, синий и зеленый цвет, как обозначается различие пола у животных), грамматические (порядок слов, есть или нет двойственное число у местоимений, противопоставлены ли формы инклюзива и эксклюзива, какая система вида и времени глаголов). Можно даже узнать, где распространены языки, в которых слезы называются сложным словом со структурой типа «глаз + вода». Карты, которыми снабжена база *APiCS Online*, позволяют оценить географическое распространение интересующих пользователя явлений.

Есть в числе параметров *APiCS Online* такие, которые специфичны именно для пиджинов и креольских языков. Например, во многих таких языках слово со значением ‘ребенок’ или ‘маленький’ восходит к *savoir*. Распространение этих слов в контактных языках мира помогает восстановить историю формирования этих языков.

К сожалению, в *APiCS Online* пока очень мало данных из пиджинов и креольских языков на основе русского. В базу вошло описание лишь одного – дальневосточного китайско-русского пиджина. Его подготовила Е.В. Перехвальская, работающая в Институте лингвистический исследований РАН.

***Типология и универсалии профессора Крофта***

Сравнение грамматик человеческих языков выявляет систематические закономерности вариации. Исследование типологии и универсалий раскрывает эти закономерности, чтобы сформулировать универсальные ограничения языка и найти их объяснение. Книга профессора Крофта [10] представляет собой всестороннее введение в метод и теорию, используемые при изучении типологии и универсалий. Обсуждаемые теоретические вопросы варьируются от самых фундаментальных – на каком основании можно сравнивать грамматики различных языков? – к самому абстрактному – какова роль функционально-исторического объяснения языковых универсалий? Книга предоставляет студентам и исследователям множество примеров языковых универсалий в фонологии, морфологии, синтаксисе и семантике. Книга сопровождается несколькими ресурсами:

- Веб-сайт Cambridge University Press: типология и универсалии
- Ошибки для типологии и универсалий

[1] The Atlas of Pidgin and Creole Language Structures (APiCS). – URL: https://apics-online.info/ (дата обращения: 01.12.2021). При описании этой и некоторых других БД были использованы материалы М. Руссо [9].

- Глаголы: аспект и каузальная структура
- Наборы задач из различных языков различной степени сложности, обновленные в августе 2014 года

***Типологические инструменты для полевой лингвистики***

Совместно с Группой порядка слов проекта *Eurotyp* Дик Баккер и Анна Северска разработали общую схему базы данных, которая будет использоваться вместе с лингвистическими вопросниками. Как пишут авторы:

«Предлагаемая схема предназначена для облегчения компьютерного ввода данных и анализа лингвистических данных с помощью общедоступных и специально разработанных программ. Она представляет собой мощный аналитический инструмент, который можно легко применить в любой лингвистической области. Для использования системы все, что требуется, – это структурирование исследуемой лингвистической области в определенном формате переменных и значений и преобразование в обычный текстовый файл, читаемый компьютером. Как только это будет выполнено, все аналитические процедуры могут применяться автоматически без дополнительных технических требований к лингвисту. Лингвисту “нужно всего лишь” применить свои лингвистические знания и опыт для уменьшения количества возможностей, открытых для компьютерных программ, или, альтернативно, для выбора из совокупности потенциально релевантных лингвистических фактов, которые представляют реальный интерес» [11].

Анкета, предназначенная в первую очередь для написания грамматики, но с полезными структурными вопросами, которые следует решать в полевых условиях, Анкета Lingua лежит в основе серии описательной грамматики *North Holland/Croom Helm/Routledge*. В этой анкете представлены основные вопросы для описания многих конструкций, встречающихся в человеческом языке. На основе анкеты написано большое количество грамматик.

***Всемирная служба аффиксных заимствований AfBo*** [12]

AfBo содержит описание 101 случая заимствования аффиксов, т.е. случаев, когда один язык заимствовал по крайней мере один аффикс из другого языка, включая в общей сложности 657 заимствованных. База данных включает в себя онлайн-интерфейс с описаниями заимствованных аффиксов с точки зрения их форм и функций, примеры комбинаций заимствованных аффиксов с родными основами, функции поиска, карты и более 230 библиографических ссылок. Можно выгрузить всю базу данных, лежащую в основе AfBo.

***Мировая база данных заимствований WOLD***[1]

Если речь идет о языковых заимствованиях, нельзя не упомянуть о проекте *WOLD***,** существующем с 2009 года. Он посвящен как раз лексическим заимствованиям и охватывает данные по 369 языкам – источникам заимствований и 41 языку, принимающему заимствования. Возможен поиск по

[1] The World Loanword Database (WOLD). – URL: https://wold.clld.org/ (дата обращения: 01.12.2021).

значениям, например мы узнаем, какие языки заимствовали слово «свадьба», а какие – слово «развод». Можно только пожалеть, что эта интересная база данных не пополняется.

***1000 языков в Интернете***[1]

ЛИР представляет собой «генеалогическое древо» из 1000 языков, где близость в дереве измеряется простым статистическим сравнением письменных систем. То, что автор называет «языками», на самом деле является системами письма; например, отдельные узлы создаются для bo (Тибетский) и bo-Latn (Тибетский, написанный латинским шрифтом). В небольшом числе случаев отслеживаются макроязыки, региональные варианты (например, en, en-IE, en-ZA) и некоторые диалекты. В общей сложности существует 919 различных кодов ISO 639–3 среди 1000 представленных систем письма.

Эти данные используются для разработки основных ресурсов, которые помогают людям использовать свой язык в Интернете: методы ввода с клавиатуры, проверки орфографии, онлайн-словари и так далее. Эта работа также лежит в основе проектов «твиты коренных народов» и «блоги коренных народов», направленных на укрепление языков с помощью социальных сетей.

Все основано на анализе последовательностей трех символов («3-граммов») в разных языках. Анализируя набор текстов на языке, можно вычислить частоты всех 3-граммов, которые появляются в коллекции, определив (разреженный) вектор в V, «представляющий» язык. Затем определяется расстояние между двумя языками как угол между их репрезентативными векторами в V. Узнав расстояние между каждой парой языков, можно восстановить филогенетическое дерево, используя любой из известных алгоритмов.

Полученный ресурс визуализирован на карте мира с цветовыми определителями для языковых семейств.

***Всемирная база данных по фонотактике***[2]

База данных позволяет пользователям исследовать типологические особенности более чем 3000 языков со всего мира. Эта база данных была собрана Марком Донохью и командой специализированных научных сотрудников Австралийского национального университета.

Тихоокеанский регион широко представлен в этой базе данных, где кодируется более 1300 языков. Это отражает тот факт, что Тихоокеанский регион является мировым лидером по языковому разнообразию, а также огромное количество данных о тихоокеанских языках, которые были накоплены за последние пять десятилетий.

***База данных инвентаризации фонологических сегментов UPSID***

Данные о фонологических системах 451 языка с программами для доступа к ним подготовлены Яном Мэддисоном и Кристин Прекода в Универси-

[1] Indigenous Tweets. 1000 Languages on the Web. – URL: http://indigenoustweets.blogspot.com/2011/12/1000-languages-on-web.html (дата обращения: 01.12.2021).

[2] World Phonotactics Database (WPD). В момент подготовки книги к печати книги БД была временно отключена.

тете Калифорнии, Лос-Анджелес. Это сборник веб-страниц с перечислением фонологических свойств большого числа языков мира. БД доступна для загрузки в виде одного zip-файла (и устанавливается и запускается на вашем собственном компьютере). Веб-интерфейс Хеннинга Ритца, который предоставляет несколько способов просмотра и поиска информации в *UPSID* и не требует наличия *UPSID* на вашем собственном компьютере, находится в библиотеке программ фонетической лаборатории UCLA[1].

***Диахронический атлас компаративной лингвистики DiACL***[2]

Стабильность признаков, время и темп изменений, а также роль генеалогии в сравнении с ареальными признаками в формировании языкового разнообразия являются важными вопросами в современных вычислительных исследованиях по лингвистической типологии. Цель проекта *DiACL Typology* – предоставить ресурс для решения этих вопросов с учетом специфики расширенного индоевропейского языкового ареала Евразии, региона с наиболее документированной лингвистической историей. База данных предварительно подготовлена для статистического и филогенетического анализа и содержит как лингвистические типологические данные из языков, охватывающих более четырех тысячелетий, так и лингвистические метаданные, касающиеся географического положения, периода времени и надежности источников [13].

Задача типологии *DiACL*, базы данных для сравнительной и филогенетической лингвистики, также содержащей лексические данные, состоит в том чтобы предоставить набор исследовательских данных для исследования языкового разнообразия, особенно для изучения диахронической типологии.

DiACL – это база данных открытого доступа с лексическими и типологическими / морфосинтаксическими данными для исторической, сравнительной и филогенетической лингвистики. Содержит данные из 500 языков 18 семейств, разделенных на три макрорегиона: Евразию, Тихий океан и Амазонку. База данных имеет следующее содержание:

- лексические наборы данных с базовыми словарями (списки Swadesh);
- лексические наборы данных с лексикой культуры, ориентированные на лексику системы жизнеобеспечения;
- типологические / морфосинтаксические наборы данных, включающие основные типы порядка слов, выравнивания и именной / вербальной морфологии.

DiACL содержит данные из современных и исторических языков, а также по возможности реконструированных языков. Данные получены из словарей, грамматик или с помощью новых полевых работ (в частности, данные с Кавказа и Амазонки).

Приведем далее краткие описания еще нескольких типологических БД.

[1] Interface to the UPSID database. – URL: http://web.phonetik.uni-frankfurt.de/upsid.html (дата обращения: 01.12.2021).

[2] Diachronic Atlas of Comparative Linguistics Online DiACL. – URL: https://diacl.ht.lu.se (дата обращения: 01.12.2021).

***Архив языковых универсалий***[1]

Архив, доступный на сайте университета Констанца (Германия), является важнейшим ресурсом для поиска языковых типологических универсалий. Инициированный Франсом Планком, он содержит список из 2029 универсалий, которые свойственны всем или по крайней мере многим языкам мира, каждый из которых содержит ссылки на литературу, и контрпримеры, если таковые имеются. С ним связан Кабинет грамматических особенностей, который будет рассмотрен в разделе, посвященном грамматическим ЛИР.

***Группа морфологических исследований университета Суррея SMG***[2]

Центр лингвистических исследований, посвященный изучению языкового разнообразия и его теоретических последствий с использованием явных формальных и статистических рамок для выражения типологических и теоретических обобщений. SMG представляет несколько типологических БД: лексических расщеплений, визуализации глагольных парадигм нескольких языков и другого.

***Типологическая база данных Павии***[3]

База данных с различными типологическими данными о структурах евросредиземноморских языков.

***База данных по редупликации Граца***[4]

На этом сайте представлены описания редупликативных структур большого числа языков. В базе возможен поиск по языкам, семействам, географическим областям, разнообразным свойствам языков, лингвистическим дисциплинам и т.д. Предлагается удобный интерфейс с развитой навигацией.

***Типологическая база данных МЭТЬЮ ДРАЙЕРА***[5]

Эта большая база данных включает в себя множество языков и множество параметров; она также включает в себя список языковых семейств мира и карты.

***База данных кросс-лингвистических колексификаций (CLICS)*** [14]

Достижения в области компьютерной лингвистики оказали большое влияние на содержание лингвистических исследований. С увеличением дос-

[1] The Universals Archive. – URL: http://typo.uni-konstanz.de/archive (дата обращения: 01.12.2021).

[2] Surrey Morphology Group (SMG). – URL: https://www.smg.surrey.ac.uk/ (дата обращения: 01.12.2021).

[3] The Pavia Typological Database. – URL: http://www-3.unipv.it/paviatyp (дата обращения: 01.12.2021).

[4] GRAZ Database on Reduplication. – URL: http://reduplication.uni-graz.at/redup (дата обращения: 01.12.2021).

[5] The University at Buffalo/ Department of Linguistics. – URL: http://linguistics.buffalo.edu/people/faculty/dryer/dryer/database (дата обращения: 01.12.2021).

тупности взаимосвязанных наборов данных, создаваемых и контролируемых исследователями, теперь можно исследовать все больше и больше взаимосвязанных вопросов. Однако такие достижения предъявляют высокие требования к строгости подготовки и обработки наборов данных. CLICS решает взаимосвязанные междисциплинарные исследовательские вопросы о совокупности слов по семантическим категориям в мировых языках и демонстрирует передовые методы подготовки данных для кросс-лингвистических исследований.

В завершение обзора зарубежных типологических ЛИР снова процитируем М. Руссо:

«Свои типологические базы ведутся также лингвистами из университетов города Кан (Франция) и Павии (Италия). Несколько небольших баз доступны на сайте группы по изучению морфологии университета Суррея. Среди них есть посвященные, например, согласованию, синкретизму в выражении категории лица, супплетивизму, дефектным парадигмам. Ученые из Утрехта и Берлина собрали базу данных по показателям реципрока («взаимного залога»).

Также будут полезны база данных Лейденского университета, посвященная типам ударения, или база тональных систем, которую создали в Беркли.

Есть и базы данных, относящиеся к лексической типологии. Тот же Калифорнийский университет в Беркли поддерживает базу данных, посвященную цветообозначениям в различных языках. В университете Граца (Австрия) создана база данных, посвященная интересному явлению – редупликации. В специализированной базе можно даже сравнить числительные от одного до десяти в более чем пяти тысячах языков» [9].

## Российские типологические ЛИР

В России исследования по лингвистической типологии и созданию типологических ЛИР ведутся достаточно активно.

***Вавилонская башня***[1]

Это проект международной этимологической базы данных, душой которого был, к сожалению рано ушедший от нас, замечательный лингвист С. Старостин. Главная задача проекта – объединение усилий по исследованию дальнего родства между языковыми семьями мира, с целью создать общедоступную базу данных корней самых разнообразных языковых семей мира. На самом деле, этот проект стал одним из оригинальнейших и богатых российских лингвистических ресурсов. Приведем перечень разделов сайта *Вавилонская башня.*

[1] Вавилонская башня Проект международной этимологической базы данных. – URL: https://starling.rinet.ru/babel.php?lan=ru (дата обращения: 01.12.2021).

*Этимология.* В настоящее время на портале представлено свыше 20 этимологических баз данных по различным языкам и группам языков, а также типологическая БД М Рулена и библиографическая БД.

*Русские словари и морфология.* Компьютерные базы данных по словарям Ожегова, Зализняка и Мюллера и программы морфологического анализа русских слов. В базах данных каждое заглавное слово имеет отсылку к программе автоматического морфологического анализа. Эту программу можно вызвать и в качестве отдельного окна. В последнем случае введено может быть любое русское или английское слово в произвольной грамматической форме.

*Аналитический каталог мифологических мотивов.* Тематическая классификация и распределение фольклорно-мифологических мотивов по ареалам Ю.Е. Березкина.

*База данных по русским народным диалектам.*

*База данных «Квантитативно-реализационный грамматический словарь современного монгольского языка» С.А. Крылова.*

*СТАРЛИНГ* – программа, разработанная С. Старостиным для работы с лингвистически ориентированными текстами и базами данных, поддерживающая разветвленную систему шрифтов для DOS и Windows.

***БД «Языки мира»***

Наибольшим российским достижением в области типологических БД, по общему мнению, является многотомное энциклопедическое издание и БД «Языки мира». Подробные данные о полиграфической и цифровой версиях этого издания представлены на портале [15]. Изложим кратко его историю.

В Институте языкознания АН СССР в середине 1970-х годов под руководством В.Н. Ярцевой началась работа над энциклопедией «Языки мира», которая предоставляла единый формат для единообразного описания любых языков, включая их общую характеристику, описание фонетики, грамматики и особенностей лексикона. Тома энциклопедии выходят с начала 1990-х годов.

Статьи издания создаются по типовым схемам – шаблонам, или вопросникам, которые могут быть применены к любому естественному языку. Это – схема для группы языков, основная схема для хорошо изученных языков, краткая схема для малоизученных языков и схема для описания диалектов. Схемы базируются на знаниях из области лингвистической типологии – одной из центральных областей лингвистики, исследующей пределы и параметры языкового разнообразия.

В настоящее время опубликован 21 том энциклопедии. На разных этапах подготовки находятся:

- палеоевропейские языки;
- семитские языки. Арабский язык. Языки южной Аравии;
- андаманские языки. Никобарские языки. Язык кусунда;
- реликтовые индоевропейские языки Европы;
- мон-кхмерские языки;
- языки мунда;
- синитские языки.

Издание «Языки мира» и БД подготавливаются одноименной рабочей группой (ныне – Сектор ареальной лингвистики) в Институте языкознания РАН. Руководитель рабочей группы – А.А. Кибрик.

Работа с инвентарем признаков, используемых в статьях энциклопедии, требует постоянного решения сложной задачи. Нужно выяснить, применимо ли то или иное устоявшееся грамматическое понятие к определенному свойству некоторого языка; иными словами, являются ли два похожих свойства в двух разных языках одним и тем же свойством. Например, подлежащее в русском и английском языках – это одна и та же грамматическая категория, несмотря на ряд различий в свойствах (в русском подлежащее маркируется в первую очередь именительным падежом, в английском – позицией перед глаголом-сказуемым). А вот следует ли подлежащеподобную категорию в филиппинских языках, которая обычно именуется «топик», отождествлять с подлежащим, или же должна использоваться отдельная категория «топик», предназначенная специально для этих языков? Сам термин «топик» для разных языков используется по-разному и может означать совершенно разные явления, когда речь идет о языках Юго-Восточной Азии, Африки или Европы. И наоборот, похожие синтаксические конструкции именуются в разных традициях «изафет», «сопряженное состояние», «вершинное маркирование в посессивной группе», что создает видимость различий при фактическом сходстве. Окончательные ответы на подобного рода вопросы появятся еще не скоро, но определенные решения приняты, и впервые получены унифицированные описания сотен разноструктурных языков.

Цифровой аналог издания не заставил себя долго ждать. Благодаря усилиям сначала М.А. Журинской, А.И. Новикова и Е.И. Ярославцевой, позже Ю.П. Скокана и затем В.Н. Полякова появилась база данных «Языки мира» (БД ЯМ). Отметим, что пока далеко не вся информация, отраженная в одноименном издании, включена в эту базу. При создании БД ЯМ пришлось решать целый ряд новых проблем. В частности, при переносе текстовых описаний из книг в цифровую форму выявилось значительное число не предусмотренных в исходной схеме признаков, которые были добавлены в БД. Приведем признаки из раздела «Морфологический тип языка» (количество точек перед названием маркирует уровень иерархии в системе признаков):

. способ соединения морфем в слове
..агглютинативные языки
..агглютинативные языки с элементами флексии
...только в имени
...только в глаголе
. в основообразовании
. только в словоизменении
..флективные языки
..флективные языки с элементами агглютинации
...только в имени. только в глаголе
. только в словоизменении
...только в словообразовании
...только в отыменном словообразовании

БД существует в нескольких формах: наиболее современная – программа для ОС Windows, Web-версия, Excel-версия. Кроме описания грамматик языков, БД ЯМ содержит обширный справочный материал – географический и генеалогический указатели, перевод названий языков и признаков на английский и т.д.

***Лингвариум***[1]

Из крупных отечественных ЛИР, содержащих разнообразную типологическую информацию, отметим портал одного из создателей БД «Языки мира» Ю.Б. Корякова и Т. Майсака *Лингвариум.*

В настоящее время основными разделами данного проекта являются:

- Реестр языков мира – создание русскоязычного справочника по всем языкам мира
- Языки России – создание базы данных населенных пунктов и атласа
- Языковые карты и их создание
- Статистические документы по этноязыковому составу разных стран
- Некоторые справочные сведения по лингвистике:
  - письменности, орфографии, фонологии
  - лингвогеография / языки по странам мира
  - вспомогательный глоссарий
  - словари
  - язык или диалект
  - личные страницы участников

***Портал «Языки мира» И. Гаршина***[2]

Из справочных ресурсов по языкам мира следует указать на портал И. Гаршина, где представлены следующие разделы (в скобках указан объем раздела):

- Лингвистика и семиотика (783 страницы)
- Общая лингвистика (229 страницы и 14 статей), включая интерлингвистику (45) и другие разделы языкознания
- Графика (161 страниц и 10 статей)
- Ностратические языки (208 страниц и 2 статьи)
- Языки других макросемей (214 страниц, включая общеязыковые)

Страницы по всем языкам находятся в подкаталоге garshin.ru/linguistics/languages /

***Московская лексико-типологическая группа***[3]

Отметим также эту группу, созданную под руководством Е.В. Рахилиной в МГУ и РГГУ. Основные направления исследований этой группы представлены на ее сайте. В ходе исследований группы создаются различные типоло-

[1] LINGVARIUM project. – URL: http://www.lingvarium.org/ (дата обращения: 01.12.2021).

[2] Языки мира. – URL: http://www.garshin.ru/linguistics/languages/index.html (дата обращения: 01.12.2021).

[3] Московская лексико-типологическая группа– URL: http://lextyp.org/about/ (дата обращения: 01.12.2021).

гические компьютерные БД и инструменты. Некоторые из них перечислены ниже, как и сведения о других отечественных работах, целью которых было создание типологических ЛИР – данных или инструментов. Очевидно, что этот перечень не является исчерпывающим.

***Типологическая база данных адъективной лексики***

Данная БД, описанная в работе [17], представляет новый инструмент для исследований по лексической типологии – *Типологическую Базу данных адъективной лексики*. База включает в себя информацию о лексикализации в различных языках ряда признаковых полей (‘острый’ – ‘тупой’, ‘пустой’ – ‘полный’, ‘твердый’ – ‘мягкий’, ‘ровный’, ‘гладкий’, ‘шершавый’ и др.). В статье обсуждаются вопросы, касающиеся структуры БД (в частности, выбор единицы информации, которая обеспечила бы сопоставимые описания лексем разных языков). Особое внимание уделяется представлению в БД переносных значений признаковых слов. Описываются основные прикладные и теоретические задачи, которые призвана решать БД. К первым относится возможность применения Базы в качестве мультиязычного словаря, ко вторым – разнообразные типологические исследования в области семантики признаковой лексики, в том числе изучение моделей полисемии.

***Атлас многоязычия Дагестана***[1]

Сайт содержит базу данных по дагестанскому многоязычию с поисковым интерфейсом. Для поиска можно использовать различные параметры (конкретные деревни, районы, годы рождения, пол, родные языки и вторые языки) и строить свои собственные графики и диаграммы. Онлайн-база данных постоянно обновляется. По состоянию на апрель 2021 года на сайте размещалась информация о многоязычном репертуаре жителей 60 сел**.**

***Мировая БД чередований согласных***[2]

База данных, включающая материал синхронных чередований в языках мира.

***Типологическая БД глаголов плавания***[3]

На данном сайте представлены основные результаты, полученные в ходе работы над проектом по типологическому исследованию одной из групп глаголов способа движения – а именно, глаголов плавания.

***Звуки Му***[4]

Типологическая база данных по семантической зоне звуков, издаваемых животными. Представлены данные более 20 языков разных языковых семейств.

---

[1] Atlas of multilingualism in Dagestan. – URL: https://multidagestan.com/ (дата обращения: 01.12.2021).

[2] The World Consonant Alternation Database. – URL: https://agricolamz.github.io/wcad/ (дата обращения: 01.12.2021).

[3] AQUA-motion. – URL: https://linghub.ru/aquamotion/ (дата обращения: 01.12.2021).

[4] Звуки Му. – URL: http://www.web-corpora.net/zvukimu/ (дата обращения 01.04.2022).

***База лингвистических данных (на материале оценочной лексики*)**

В работе [17] описаны принципы формирования и структуры лингвистической базы данных отрицательно-оценочных лексико-семантических единиц. Охарактеризовано наполнение столбцов таблицы базы данных. В качестве основного источника наполнения базы данных использован Русский семантический словарь (первый том). Проектируемая база данных предназначена для автоматического поиска представленной в ней лексики при экспертировании конфликтных текстов.

***Сервис lingtypology***

Данный сервис, описанный в [18], связывает пользователя с базой данных Glottolog и предоставляет дополнительную функциональность для лингвистической типологии. База данных Glottolog содержит каталог мировых языков. Этот сервис помогает исследователям создавать лингвистические карты, используя философию проекта CLLD[1], который создает единый доступ к лингвистическим данным в различных ЛИР. Кроме того, сервис предоставляет возможность загрузки данных из типологических баз данных, таких как WALS, AUTOTYP и других.

## Литература к главе 13


1. Кружков М.Г. Информационные ресурсы контрастивных лингвистических исследований: типологические базы данных // Системы и средства информатики. – 2015. – Т. 25, № 1. – С. 198–212. – DOI: 10.14357/08696527150113
2. Соловьев В.Д. Типологические базы данных: перспективы использования // Вопросы языкознания. – 2010. – № 1. – С. 94–110.
3. Edith A. Moravcsik. Introducing Language Typology. – Cambridge : Cambridge University Press, 2013. – ISBN:9780511978876. – DOI: https://doi.org/10.1017/CBO9780511978876 (дата обращения: 01.12.2021).
4. Forkel R. The Cross-Linguistic Linked Data project. – URL: https://clld.org/docs/ldl2014/main.pdf (дата обращения: 01.12.2021).
5. Коряков Ю.Б., Майсак Т.А. Систематика языков мира и базы данных в интернете. – URL: http://lingvarium.org/sys.shtml (дата обращения: 01.04.2022).
6. Обсерватория Лингвосферы. – URL: http://www.lingvarium.org/paedia/linguasphere/sphera_ru.shtml (дата обращения: 01.04.2022).
7. Кибрик А.А., Соловьев В.Д. Чем компьютерные технологии могут помочь лингвистической типологии? // Вестник Российской академии наук. – 2015. – Т. 85, № 1. – С. 32–38. – ISSN: 0869–5873.
8. The Typological Database System (TDS) Curator documentation. DANS / Alexis Dimitriadis, Utrecht Institute of Linguistics OTS (project leader), Menzo Windhouwer, Max Planck Institute for Psycholinguistics, Marc Kemps-Snijders, Meertens Instituut, Rob Zeeman, Meertens Instituut, Vesa Åkerman, DANS, Marjan Grootveld // DANS CLARIN-NL. – 2012. – URL: https://doi.org/10.17026/dans-xkf-qjmg (дата обращения: 01.12.2021).


[1] Описание CLLD см. в гл. 19.

9. Руссо М. Лингвистические базы данных. – URL: https://polit.ru/article/2013/12/12/ps_databases/ (дата обращения: 01.12.2021).

10. Croft W. Verbs: Aspect and Causal Structure. – 2012. – DOI: 10.1093/acprof:oso/9780199248582.001.0001

11. Bakker D., Siewierska A. Typological tools for field linguistics. – URL: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/database-system_description.php (дата обращения: 01.04.2022).

12. Seifart F. AfBo: A world-wide survey of affix borrowing. – Leipzig : Max Planck Institute for Evolutionary Anthropology, 2020. – DOI: 10.5281/zenodo. 3610155

13. Diachronic Atlas of Comparative Linguistics (DiACL) / Gerd Carling, Filip Larsson, Chundra A. Cathcart, Niklas Johansson, Arthur Holmer, Erich Round, Rob Verhoeven. – 2018. – URL: https://doi.org/10.1371/journal.pone.0205313 (дата обращения: 01.12.2021).

14. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. Sci Data 7, 13 / Rzymski C., Tresoldi T., Greenhill S.J. [et al.]. – 2020. – URL: https://pubmed.ncbi.nlm.nih.gov/31932593/ (дата обращения: 01.04.2022).

15. Многотомное энциклопедическое издание Языки мира. – URL: https://iling-ran.ru/langworld/ (дата обращения: 01.12.2021).

16. Резникова Т.И., Кюсева М.В., Рыжова Д.А. Типологическая база данных адъективной лексики. Компьютерная лингвистика и интеллектуальные технологии // по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.) : в 2 т. / под редакцией: В.П. Селегей. – Москва : РГГУ, 2013. – Т. 1 : Основная программа конференции, вып. 12(19). – С. 407–419.

17. Кочергина К.С. Проектирование базы лингвистических данных (на материале оценочной лексики). – URL: http://sfk.spbu.ru/ru/issues/2385/3215/1657 (дата обращения 01.04.2022).

18. Moroz G. lingtypology: easy mapping for Linguistic Typology. – 2017. – URL: https://CRAN.R-project.org/package=lingtypology (дата обращения: 01.12.2021).

# ГЛАВА 14. РЕСУРСЫ ЗВУЧАЩЕЙ РЕЧИ

## Общие сведения

Настоящая глава посвящена в основном корпусам звучащей речи, которые представляют собой важнейший тип языковых ресурсов.

Википедия предлагает следующее определение:

«Речевой корпус – база данных аудиофайлов и транскрипций текстов, разновидность корпуса текстов. В речевых технологиях речевые корпуса используются, среди прочего, для создания акустических моделей. В лингвистике речевые корпуса используются для исследований фонетики, диалектологии, конверсационного анализа и в других областях. Существует два типа речевых корпусов:

1. Базы начитанных текстов.

2. Базы аудиозаписей спонтанной речи.

Особый вид речевых корпусов – это базы данных текстов, наговоренных людьми, не являющимися носителями языка, которые содержат речь с иностранным акцентом»[1].

Далее обратимся к мнению специалистов, профессионально занимающихся созданием звучащих ЛИР.

«Интерес к созданию корпусов звучащей речи был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с огромной акустической вариативностью звуковых единиц языка, которая имеет весьма разнообразные источники – от системной контекстной вариативности, обусловленной коартикуляцией, до психофизиологического состояния говорящего или технических характеристик микрофона, который используется при записи речевого материала. Современные системы распознавания речи, которые дают наиболее высокие показатели надежности, базируются преимущественно на методах статистического (вероятностного) моделирования речевых и языковых явлений.

Такие системы обычно обучаются на очень больших массивах звучащей речи, записанной от многих дикторов (не менее 100 человек). Звуковые файлы, входящие в обучающий речевой корпус, специальным образом аннотируются, т.е. снабжаются акустико-фонетической и лингвистической информацией, необходимой для построения статистических моделей. В послед-

[1]Речевой корпус. – URL: https://ru.wikipedia.org/wiki/Речевой_корпус (дата обращения: 01.12.2021).

нее десятилетие заметен переход от “ручных” правил и алгоритмов к корпусному моделированию и в области автоматического синтеза речи. (…)

Было бы неправильно думать, что речевые корпусы представляют интерес только для развития речевых технологий. Проблема описания и моделирования звуковой стороны речевых сообщений с учетом ее акустической вариативности в разнообразных речевых ситуациях представляет самостоятельный научный интерес и возникает во многих задачах, связанных с анализом звучащей речи. Несмотря на достаточно большой опыт исследования этой проблемы в акустической фонетике, нельзя сказать, что она решена в полном объеме даже для такого относительно простого речевого режима, как чтение связных текстов, нейтральных в экспрессивно-эмоциональном отношении. Использование представительных речевых корпусов, снабженных специальной аннотированной информацией, уровень развития современных речевых технологий и постоянно возрастающие мощности компьютерной техники дают недоступную ранее возможность для проведения крупномасштабных и статистически достоверных фонетических исследований, обычно весьма трудоемких» [1].

## Классификация речевых корпусов

Классификация речевых корпусов предложена в другой работе О.Ф. Кривновой [2].

Аннотированные речевые корпуса – важнейший компонент исследований в области звучащей речи. Сегодня они созданы и создаются для большого количества языков, научных дисциплин и технологий. Опыт, накопленный в области их разработки и использования, позволяет выделить ряд признаков, которые могут быть положены в основу классификации речевых баз данных и учитываться при проектировании нового РК. Укажем наиболее важные характеристики:

- целевое использование корпуса: специализированные, технологические, общие (репрезентативные), учебно-иллюстративные;
- тип речевого материала: дискретная речь, непрерывная речь-чтение, спонтанная речь, специальные и естественные диалоги;
- тип текстового материала: списки слов / слогов, наборы отдельных предложений, связные тексты; монотематические или политематические;
- тип речевого сигнала: лабораторная речь, офисная речь, публичная речь, телефонная речь (обычная или через мобильный телефон); радио&, телеречь, речь в условиях естественной внешней среды, иноязычная (акцентная) речь и т.д.;
- тип информации, ассоциированной с речевым сигналом (аннотации): орфографическая запись, фонемная / фонетическая транскрипция, просодическая транскрипция, акустико-фонетическая разметка сигнала: «событийная», сегментная, просодическая, включение других типов лингвистических аннотаций и комментариев, например об индивидуальных особенностях произношения говорящего или эмоциональной окраске речевых фрагментов;
- тип статистической балансировки звуковых единиц языка: равномерная, репрезентативная, по специальной статистической схеме;

o наличие и типы дополнительной сигнальной информации, включенной в корпус наряду с речевым сигналом: простые, мультимодальные и специальные корпуса.

### Статистика ЛИР звучащей речи

ЛИР звучащей речи, которые называют также речевыми базами данных, а также устными ресурсами, в количественном отношении занимают второе место среди ЛИР, после текстовых корпусов.

Приведем данные о количестве ЛИР звучащей речи по ведущим мировым каталогам и поисковым системам для ЛИР. Заметим, однако, что такой класс ЛИР выделяется не во всех каталогах, поэтому для каждого архива указывается критерий, по которому выделялись эти ЛИР.

В крупнейшем в мире архиве The Language Archive Института психолингвистики Общества Макса Планка содержатся ЛИР в виде аудио- и видеофайлов следующих форматов (количество ЛИР указано в скобках):

аудиофайлы
audio/x-wav (67 681)
audio/x-mpeg (22 529)
audio/mp4 (12 785)
application/mediatagger (4521)
audio/mpeg (1920)
audio/x-aiff (581)
видеофайлы
video/x-mpeg1 (34 632)
video/x-mpeg2 (17 510)
video/mp4 (5628)
video/quicktime (3924)

При поиске через The Virtual Language Observator инфраструктуры CLARIN мы имеем следующие результаты:

audio (141 045)
video (24 608)

При поиске в архиве OLAC выдается 8670 ЛИР по запросу «аудио» и 2670 – при поиске по запросу «видео», но возможны пересечения.

В архиве LDC находим 367 ЛИР, представленных в виде аудиофайлов и 18 ЛИР – видеофайлов. Возможны пересечения.

В каталоге ELRA представлено 551 аудио ЛИР и 21 видео.

### Краткая история создания ЛИР звучащей речи

Первый корпус устной речи Лондон-Лунд (The London-Lund Corpus) был разработан в рамках проекта «Исследование функционирования английского языка» (The Survey of English Usage). Цель проекта заключалась в том, чтобы по возможности полно зафиксировать особенности грамматической системы английского языка в речи взрослого образованного носителя. Проект разраба-

тывался с 1959 года под руководством Р. Квирка (R. Quirk) в Лондонском университетском колледже. Объем корпуса – 1 млн словоупотреблений. Текстами устной речи были записи радиопередач, заседаний официальных структур, а также неформальных бесед. Машинный вариант корпуса создавался в Лундском университете (Швеция) и был готов к использованию в 1979 году [3].

Следующие речевые корпусы появились в середине 1980-х годов в США, где их разработка финансировалась прежде всего Министерством обороны. При поддержке этого ведомства были созданы: TI-DIGITS – корпус (1984) для тестирования систем распознавания изолированных цифр и цифровых последовательностей; Road Rally для анализа и распознавания ключевых слов и King Corpus для систем идентификации говорящего.

В рамках программы развития лингвистических технологий, реализованной агентством ARPA / DARPA (*the Advanced Research Projects Agency*), это же Министерство финансировало создание корпуса TIMIT, который послужил прототипом для многих других речевых баз данных.

При этой же финансовой поддержке были разработаны специализированные речевые корпусы *Resourse Management* (*RM*) и *Wall Street Journal* (*WSJ*) для исследований в области распознавания слитной речи, а также *Air Travel Information Service* (*ATIS*) – для исследования спонтанной речи и понимания естественного языка в диалоговых системах [4].

Накопленный к концу 1980-х годов опыт показал, что создание представительных речевых корпусов требует совместных усилий исследовательских институтов, индустриальных компаний и государственных спонсоров. Финансовые и временные затраты на разработку высококачественных ресурсов оказались очень велики. Разработчики пришли к выводу, что ресурсы должны обеспечивать возможность их многократного использования разными пользователями, т.е. быть общедоступными и использоваться более чем для одной цели, значит – многофункциональными.

В связи с этими требованиями возникла проблема стандартизации лингвистических описаний, согласования форматов представления информации в разных видах лингвистических ресурсов и их типологии. Обзор стандартов, в том числе используемых для аннотирования звучащих корпусов, представлен выше. Упомянем также фонетический алфавит SAMPA (*Speech Assessment Methods Phonetic Alphabet*). Он представляет собой Международный фонетический алфавит, записанный символами ASCII, с рядом изменений под конкретный язык[1].

## Обзоры ЛИР звучащей речи

В настоящее время количество речевых ЛИР очень велико – количественные данные приведены выше. Их обзоры содержатся во многих работах: кроме цитированных выше публикаций [2,3,4], нужно отметить коллективную монографию под редакцией Р.К. Потаповой «Речевая коммуникация в

[1] SAMPA – computer readable phonetic alphabet). – URL: https://www.phon.ucl.ac.uk/home/sampa/index.html (дата обращения: 01.12.2021).

информационном пространстве» [5]. К числу наиболее перспективных исследований, проводимых коллективом авторов, относятся следующие: соотношение прагмафонетики и прагмалингвистики; поиск перцептивно-слуховых и акустических коррелятов эмоций в речевой коммуникации; психолингвистический подход к оценке результатов автоматической обработки текстов.

Также из российских авторов упомянем диссертацию Е.В. Лосевой [6] и публикации [7,8, 9].

Из зарубежных подборок ЛИР звучащей речи, кроме универсальных каталогов ЛИР отметим портал М. Вайссера[1], который содержит описания 29 устных корпусов английского языка. Автор включает в этот класс не только корпуса, содержащие аудио- и видеофайлы, но также ЛИР, представляющие текстовые орфографические представления / транскрипции устных данных.

Для получения перечня фонематических / акустических / артикуляционных банков данных (в основном включающих изолированные слова, фонемы или предложения) М. Вайссер советует обратиться к отдельному списку ссылок Института фонетики и обработки цифровой речи Кильского университета[2], или к универсальным каталогам ELRA и LDC.

М. Вайссер указывает, что некоторые авторы проводят различие между «речевыми корпусами» (пригодными для акустических / фонетических исследований) и «устными (разговорными) корпусами» (содержащими транскрипцию любого типа разговорного языка). М. Вайссер использует «устные корпуса» как обобщающий термин для обоих типов.

Он отмечает, что LDC также содержит различные ресурсы, которые не являются корпусами как таковыми, но могут представлять интерес.

Пример: *Словарь американского английского устного языка*[3]. Этот словарь содержит произношения 53 602 наиболее распространенных слов на английском языке, представленных в виде отдельных аудиофайлов (слова взяты из ленты новостей и записей телефонных разговоров).

Одна из крупнейших БД звучащей речи это *Аудиосэмплы языков мира*[4]. Запись библейских историй и других текстов можно слушать почти на 6000 языков мира.

Википедия также рекомендует подборку корпусов звучащей речи Гамбургского университета[5].

В обзоре CLARIN[6] описывается 90 корпусов устной речи, которые входят в инфраструктуру CLARIN. 79 из них содержат как транскрипции устной

[1] Spoken Corpora. – URL: http://martinweisser.org/corpora_site/ spoken_corpora.html#corpus_section (дата обращения: 01.12.2021).

[2] A collection of databases for phonetic purposes. – URL: http://www.ipds.uni-kiel.de/links/datenmaterial.en.html (дата обращения: 01.12.2021).

[3] American English Spoken Lexicon. – URL: https://catalog.ldc.upenn.edu/LDC99L23 (дата обращения: 01.12.2021).

[4] GRN. Global Recordings Network Audio samples from the languages of the world. – URL: http://globalrecordings.net/en/languages (дата обращения: 01.12.2021).

[5] Linguistic Corpora at the HZSK Repository. – URL: https://corpora.uni-hamburg.de/hzsk/en/repository-search (дата обращения 01.04.2022)

[6] Spoken corpora. – URL: https://www.clarin.eu/resource-families/spoken-corpora (дата обращения: 01.12.2021).

или спонтанной речи, так и соответствующие аудиозаписи, а 11 – только транскрипции. Корпуса содержат транскрипции, например, новостей в эфире или повествований и диалогов. Они являются бесценным ресурсом для различных лингвистических исследований, таких как фонология, анализ устной коммуникации и диалектология. Эти корпуса тщательно отбираются и содержат множество социально-демографических метаданных.

Большинство корпусов являются одноязычными, на них приходится 15 языков: арабский, чешский, голландский, эстонский, финский, французский, немецкий, венгерский, итальянский, непальский, норвежский, польский, шотландский, саамский, словенский, испанский и шведский. В подавляющем большинстве случаев корпуса можно напрямую загрузить из национальных репозиториев или запросить через простые в использовании среды онлайн-поиска. Они также имеют богатые теги, многие из которых имеют разметку, специфичную для речевых корпусов, например фонематическую и просодическую аннотацию.

### Проектирование речевых ЛИР

Общее описание проблем, с которым сталкивается разработчик речевых ресурсов, содержится в работе [1], которую мы процитируем с сокращениями.

Условно всю совокупность возникающих вопросов можно разделить на четыре группы: технические, содержательные, структурные и инструментальные (исполнительские).

К техническим относятся вопросы, связанные с акустическими и техническими условиями записи речевого материала (выбор типа и количества микрофонов, звуковой карты компьютера, режима цифрового кодирования и формата звуковых файлов, акустическая среда записи, тип канала связи и пр.).

Содержательные вопросы более разнообразны и принципиально существенны. Перечислим основные проблемы, которые приходится здесь решать.

- Выбор дикторов (количество, пол, возраст, диалектные различия, образование, социальное положение, профессия и пр.).
- Подбор текстового материала (специализированный / репрезентативный, тип произносимых речевых образцов (слова, отдельные предложения, тексты, образцы спонтанной речи), фонетически сбалансированный / не сбалансированный, тип балансировки, статистическая представительность звуковых единиц и т.п.).
- Распределение текстового материала по дикторам, включая количество подходов для каждого диктора.
- Распределение речевого материала на тренировочную и тестовую части.
- Выбор типов информации, ассоциированной с каждым звуковым файлом (орфографическая запись, фонемная запись / фонетическая транскрипция реального произнесения, акустико-фонетическая разметка звукового сигнала, прочие типы аннотаций и комментариев).

Структурные вопросы касаются организации информации, содержащейся в корпусе, в формат, удобный для размещения, хранения, поиска

и использования нужной информации (структура директорий и файлов, создание протоколов и пр.).

«Инструментальные», или исполнительские вопросы возникают в связи с автоматизацией и стандартизацией разных этапов создания речевого корпуса. Для репрезентативных или общих корпусов главная проблема связана с разработкой стандартов для транскрипции речевых сигналов на разных уровнях их представления и для разных языков, с установлением набора транскрипционных символов, соглашений о разметке сигналов, задающих уровни транскрипции – акустический, фонетический, фонемный, словесный, просодический и прочие.

Один из наиболее сложных вопросов для формирования речевых ЛИР – это вопрос сегментации устной речи. Из российских исследователей наиболее подробно этот вопрос рассмотрен в монографии Р.К. Потаповой и В.В. Потапова [10].

Отсутствие баланса в доступности устного и письменного материала в машиночитаемом формате продлится еще очень долго. В силу различных причин построение корпусов устной речи продвигается намного медленнее, чем построение корпусов письменной речи. В первую очередь устную речь нужно как-то зафиксировать – например, с помощью магнитной ленты, цифровой записи или видеокассеты. Затем ее нужно записать буквами, что является утомительной и дорогой работой, качество которой зависит в большой степени от качества записи и степени шума внешней среды в естественных условиях.

Как правило, для создания речевых корпусов, содержащих транскрипционную информацию, привлекаются фонетические эксперты, но даже в этом случае для получения согласованных экспертных транскрипций и разметочных файлов необходимо разрабатывать специальные рабочие инструкции, в которых приходится предусматривать не только типовые, но и трудные случаи фонетической интерпретации речевых сигналов. Кроме того, при разработке таких корпусов необходим специальный компьютерный инструментарий для обеспечения удобной, быстрой и надежной работы эксперта. Хотя сейчас существует довольно большое количество компьютерных программ, которые позволяют анализировать, размечать, транскрибировать и аннотировать речевые сигналы, каждая из них имеет свои особенности, которые не всегда удобны для решения конкретных задач. В особенности это относится к программам, которые относительно дешевы или находятся в свободном доступе. Специальных программ требует также организация записи и файлирования речевого материала. Как правило, используется так называемый метод суфлера (prompt-method), который позволяет создавать звуковые файлы, соответствующие отдельным объектам речевого корпуса, непосредственно в процессе его записи.

Процитируем В.П. Захарова и С.Ю. Богданову:

«Главная сложность создания фонетических лингвистических ресурсов связана с необходимостью транскрибирования устной речи. При этом возникают следующие проблемы:

1) выбор алгоритма для транскрибирования;

2) учет индивидуальных особенностей произношения;

3) учет всего устного текста или его фрагментов;

4) учет диалектных вариантов произношения слов;

5) учет ударений в словах;

6) учет просодических признаков произносимых фраз;

7) маркирование слов, которые при прослушивании не распознавались;

8) маркирование в записи для фонетического корпуса паралингвистических явлений, сопутствующих речи (паузы, смех, бормотание, кашель, и т.д.).

В настоящее время общепринято, что для создания машиночитаемых фонетических корпусов используется транскрипция на основе орфографического представления звуков речи с дополнительными знаками, передающими (при необходимости) просодические, паралингвистические и другие особенности произношения» [3].

Качественная орфографическая расшифровка является основой для всех видов анализа данных устной речи. Однако расшифровка речи – трудоемкая и утомительная задача. Но автоматическое распознавание речи, а также инструменты NLP и текстовых аннотаций могут значительно ускорить эту задачу, сэкономить много времени и избежать разочарований.

Очень часто исследователи получают многие часы ценных устных аудиоданных в процессе своей работы, но у них очень мало времени для их обработки и анализа. Чтобы решить эту проблему, исследовательская группа CLARIN по устной истории и технологиям разработала удобный портал транскрипции, который помогает ученым эффективно работать с речевыми аудиоданными. Этот портал обсуждался в веб-семинаре «Практическое руководство CLARIN по расшифровке данных интервью», организованном SSHOC и CLARIN ERIC в марте 2020 года[1].

## Разработки ресурсов устной речи в России

Одним из первых российских ЛИР для звучащей речи была база данных, разработанная еще в 1990-х годах в Институте системного анализа под названием ISABASE [11]. Авторы этой разработки в более поздней публикации описывают ее так:

«К сожалению, к настоящему времени в некоторых, технологически важных отношениях корпус ISABASE морально устарел: слишком мало дикторов (36), дискретное чтение предложений и др. При подготовке корпуса много усилий и времени было потрачено на ручную сегментацию и транскрипцию звуковых файлов. Современные технологии построения систем автоматического распознавания речи не требуют наличия большого массива размеченных звуковых файлов. Это дает возможность существенно увеличить количество дикторов и размеры текстового материала при разработке речевого корпуса. При этом особое значение приобретают статистические характеристики фонетического содержания текстового материала (балансировка, представитель-

[1] SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data. – URL: https://sshopencloud.eu/news/sshoc-webinar-clarin-hands-tutorial-transcribing-interview-data (дата обращения: 01.12.2021).

ность и разнообразие контекстов) и наличие фонетической транскрипции, отражающей реальное дикторское произнесение текстового материала» [1].

Далее в этом докладе говорится:

«В настоящее время в ИСА РАН осуществляется разработка нового корпуса русской речи с нашим участием. Надо заметить, что подбор текстовых массивов с заранее оговоренными статистическими требованиями на контекстное употребление фонем представляет собой очень трудоемкую задачу. В целях ее автоматизации мы разработали специализированный компьютерный инструментарий, который включает автоматический транскриптор русских письменных текстов и программу статистической обработки транскрипционных записей. Статистическая программа, сопровождающая работу транскриптора, используется не только для окончательного подсчета частоты встречаемости фонем в разных контекстах, но также как фильтр-накопитель, который позволяет накапливать текстовый массив, удовлетворяющий априорным статистическим требованиям на частоту встречаемости тех или иных звуковых объектов.

Работа с записанным речевым материалом требует верификации канонической транскрипционной записи, которая была построена для текстовых массивов с помощью автоматического транскриптора. Цель верификации состоит в том, чтобы учесть реальное произнесение предложенных материалов дикторами. Временные затраты на эту работу можно существенно сократить, если использовать каноническую автоматическую транскрипцию как своего рода «подстрочник», который может исправляться экспертами-фонетистами в интерактивном режиме работы со звуковыми сигналами. Для облегчения и унификации этой деятельности была разработана специальная инструкция и компьютерная программа, обеспечивающая удобный режим работы фонетиста-эксперта.

Несколько иной тип русского речевого корпуса разрабатывается нами в рамках исследовательского проекта, который посвящен моделированию акустической вариативности звуковых единиц в связной речи. Проект поддерживается РФФИ. При подборе и верификации речевого материала для этого корпуса мы также активно использовали созданный нами автоматический транскриптор, а также статистический и верификационный компьютерный инструментарий. Для выполнения этого проекта необходима сегментация звуковых файлов на фрагменты фонемной размерности и их фонетическая аннотация (phonetic labelling)».

В настоящее время в России создано значительное количество ЛИР звучащей речи. Официальным центральным местом хранения этих ЛИР является Национальный электронный звуковой депозитарий (НЭЗД)[1]. Хотя этот депозитарий не специально лингвистический – в нем, кроме текстов, хранится большое количество музыкальных и фольклорных записей – тем не менее он является крупнейшим в стране архивом звучащей речи на десятках языков

[1] Официальное открытие НЭЗД состоялось в 2009 г. на базе оцифрованных ресурсов фонограммархива ИРЛИ (Пушкинского дома). В 2021 г. на сайте ИРЛИ сведения о НЭЗД отсутствуют, есть только ресурсы фонограммархива. – URL: http://pushkinskijdom.ru/ nauchnye-otdely/fonogrammarhiv/ (дата обращения: 01.12.2021).

народов России. В НЭЗД имеется поисковая система, где можно выбрать язык, жанр, отличить музыку от текста и использоваать другие полезные метаданные. Формат машиночитаемого описания аудиозаписей отражает основные параметры:

- технические характеристики носителя;
- технические и технологические характеристики записи;
- лингвистические и жанровые характеристики;
- характеристики содержания;
- привязку к территории распространения этноса, бытования собранного фольклорного материала, географии проведения исследовательских работ;
- персональную информацию об исполнителях, собирателях, исследователях;
- библиографические ссылки;
- информацию об экспедициях различных научных учреждений, связанных проблематикой исследований с тематикой данного проекта;
- учетные данные оригинала;
- данные о наличии копий;
- служебная и технологическая информация.

Среди специальных ЛИР звучащей речи России ведущее место занимает ***Национальный корпус русского языка***[1], в который входят устный, акцентологический и мультимедийный подкорпуса.

***Устный корпус***. Корпус устной речи (как самостоятельный корпус существует с 2007 г.) включает в себя расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов. Использована русская стандартная орфография (при этом приводятся наиболее частотные и общепринятые стяженные формы). Возможен лексический, морфологический и семантический поиск, а также формирование пользовательских подкорпусов, в том числе и по социологическим параметрам. Включены тексты самых разных жанров и типов, разного происхождения с точки зрения географии (Москва, Санкт-Петербург, Саратов, Ульяновск, Таганрог, Екатеринбург, Норильск, Воронеж, Новосибирск и мн. др.). Хронологический охват корпуса – 1900–2000 годы. Устный корпус НКРЯ подробно описан в работе [12].

***Акцентологический корпус*** (корпус истории русского ударения) (открылся в 2008 г.) включает тексты, несущие информацию об истории русского ударения. Во-первых, это все тексты поэтического корпуса, где в силлабо-тонических, а отчасти и в чисто тонических текстах содержится информация о месте ударения в слове (требующая дополнительной интерпретации). Во-вторых, это акцентуированные (в соответствии с реально звучащим ударением) записи устной речи, в том числе кинофильмов. Эти тексты доступны для поиска по месту ударения и просодической структуре слова. В-третьих, это подкорпус наивной поэзии. Наивная поэзия – это стихотворные тексты, написанные

[1] НКРЯ. Устный корпус. – URL: https://ruscorpora.ru/new/search-spoken.html (дата обращения: 01.12.2021).

поэтами-любителями, не публикующимися в признанной литературной периодике. Эстетические достоинства этих текстов не важны для исследования русской акцентологии, но регулярность чередований ударных и безударных слогов в русском стихе дает бесценный материал для уяснения множества вопросов, связанных с расстановкой ударений в словах современными носителями русского языка. Подробное описание акцентологического корпуса имеется в работе [13].

***Мультимедийный корпус*** (открылся в декабре 2010 г.) включает фрагменты кинофильмов 1930–2000-х годов. Они представлены в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также наблюдаемых в кадре жестов. В мультимедийном корпусе возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т.п.) и типу речевого действия (согласие, ирония и т.п.). В поисковой выдаче видеофрагменты доступны для просмотра и прослушивания. Описание корпуса представлено в работе [14].

***Русский учебный корпус*** включает образцы устной и письменной речи изучающих русский язык как иностранный. Учебный корпус доступен по адресу[1].

Далее опишем другие российские проекты ЛИР звучащей речи.

***Устный подкорпус Национального корпуса калмыцкого языка***[2]

Наилучшим способом презентации материала текстов является фонетическая транскрипция звучащей речи, наряду с орфографической записью. Именно она позволит сохранить все особенности диалекта, хотя это, конечно, не отменяет необходимости записывать нетекстовый материал на основе вопросников для сбора диалектологического материала. Звуковой корпус по калмыцкому языку состоит из трех модулей: базы данных «KalmykSpeech», звуковых файлов и расшифровок в формате .eaf. Общая метаинформация, собранная в процессе заполнения анкеты, фиксировалась в специально созданных таблицах в программе Access MS Office. Таблицы Informants, SoundFiles, Epizods связаны друг с другом при помощи общих полей. На сайте корпуса имеется подробное описание технологии записей, анкеты информанта, модели метаданных и транскрипции.

***Корпус русской устной речи СПбГУ***[3]. В данном корпусе собраны результаты многолетних исследований звучащей речи в виде текстов с орфографической и акустико-фонетической аннотацией и соответствующим звуковым материалом. Из-за огромной трудоемкости процесса аннотирования, выполняемого экспертами вручную, в настоящее время мы располагаем сравнительно небольшим по объему материалом: около 22 тыс. словоупотреблений и, соответственно, словарем более 12 тыс. словоформ. Отличительной особен-

[1] RLC Русский учебный корпус. – URL: http://web-corpora.net/RLC (дата обращения: 01.12.2021).

[2] Калмыцкий корпус. Kalmyk Corpus. – URL: http://web-corpora.net/KalmykCorpus/search/?interface_language=ru (дата обращения: 01.04.2022).

[3] Корпус русской устной речи. – URL: http://russpeech.spbu.ru/project.htm (дата обращения: 01.12.2021).

ностью корпуса является то, что пользователю доступны не только акустический сигнал и орфографическая расшифровка, но и сплошная акустико-фонетическая транскрипция.

Корпус русской устной речи отражает употребление словоформ, грамматических конструкций и словосочетаний в русской речи, начиная с середины XX века и до настоящего времени. В Корпус включены звучащие тексты разных стилей: чтение профессиональными дикторами, чтение рядовыми носителями языка, монологическая речь (с включением диалогов как последовательности минидиалогов) и детская речь.

Цели и задачи Корпуса:

- служить источником фактического материала для функционального компьютерного моделирования речевой деятельности;
- обеспечивать возможность проверки лингвистических гипотез на достаточно представительном материале;
- выполнять функции справочного пособия для выяснения особенностей современного русского произношения, т.е. служить эффективным помощником для всех, работающих со словом (лингвисты, преподаватели русского языка, разработчики систем автоматического распознавания речи и др.).

***Один речевой день. СПбГУ***[1]

Корпус «Один речевой день» (ОРД) создается с использованием методики 24-часовой записи, что позволяет получить максимально естественную речь человека в условиях повседневного общения. Запись проводится с использованием диктофона, который, предварительно настроив, информант закрепляет на себе стационарно. Статистические характеристики корпуса – 1250 часов звукозаписей, полученных от 128 информантов и более 1000 их коммуникантов, представляющих разные социальные группы современного российского города, 2800 макроэпизодов речевой коммуникации, 1 млн словоупотреблений в текстовых расшифровках.

***Проекты речевых ЛИР СПИИРАН***

На сайте Института[2] представлены описания этих ЛИР*:*

- Аудиокорпус речевых и неречевых акустических событий (САРГАС-БД)
- Корпус слитной русской речи для систем автоматического распознавания речи
- Аудиовизуальный корпус слитной русской речи с высокоскоростными видеозаписями HAVRUS

***«Рассказы о сновидениях» и другие корпуса звучащей речи***[3]

Данный проект стал возможным благодаря сотрудничеству исследователей из Института языкознания РАН, Российского государственного гума-

[1] СПбГУ. Один речевой день. – URL: https://ord.spbu.ru/ (дата обращения: 01.12.2021).

[2] Санкт-Петербургский институт информатики и автоматизации РАН. – URL: http://www.spiiras.nw.ru/ru/ (дата обращения: 01.12.2021).

[3] Рассказы о сновидениях и другие корпуса звучащей речи. – URL: http://spokencorpora.ru/ (дата обращения: 01.12.2021).

нитарного университета, Московского государственного университета и Новосибирского государственного технического университета. В основе проекта – разрабатывавшаяся в течение нескольких лет концепция описания и дискурсивного аннотирования данных живой устной речи. Ее практическим воплощением стала система дискурсивной транскрипции – графической записи звукового сигнала, призванной в последовательной и наглядной форме отразить наиболее значимые дискурсивные феномены, сопутствующие порождению спонтанной устной речи, и во многом определяющие форму языковых выражений. На сайте проекта имеется подробное описание системы транскрипции, а также функциональных возможностях ресурса. В состав данного ЛИР входит ряд перечисленных ниже корпусов.

*Рассказы о сновидениях*

Корпус состоит из 129 рассказов детей и подростков от семи до 17 лет об увиденном ими во сне. Рассказы записывались непосредственно после пробуждения. Общая длительность звучания – около двух часов; объем корпуса – около 14 тыс. словоупотреблений.

*Рассказы сибиряков о жизни*

Корпус состоит из 17 рассказов взрослых жителей Новосибирска (от 19 до 70 лет) о каких-либо ярких событиях в их жизни. Общая длительность – около 40 минут звучания; объем – приблизительно 5 тыс. словоупотреблений.

*Веселые истории из жизни*

Корпус состоит из 40 пар рассказов взрослых людей (от 18 до 60 лет) о смешных происшествиях в их жизни. От каждого рассказчика было получено по два рассказа об одном и том же событии, устный и письменный. Общая длительность устной части корпуса – около 1 часа 10 минут; объем – около 7 тыс. словоупотреблений. Объем письменной части – около 10 тыс. словоупотреблений.

*Группа корпусов «Истории о подарках и катании на лыжах»*

Корпуса устных текстов, полученных от носителей различных языков в ходе двухступенчатого эксперимента. На первом этапе испытуемым последовательно предъявлялись два набора картинок: «Подарки» и «Катание на лыжах». Для каждого набора давалось несколько секунд на ознакомление с общим сюжетом, после чего информанты составляли рассказы по картинкам, имея их перед глазами. На втором этапе, производившемся через 6 – 8 часов после первого, испытуемым предлагалось пересказать те же истории по памяти, не глядя на картинки.

*Корпус на русском языке*

Корпус состоит из 20 рассказов и 20 пересказов десяти носителей русского языка. Записи производились в Москве в 2003–2004 гг. Все информанты являются москвичами, возраст информантов на момент записи – от 20 до 30 лет. Общая длительность звучания – около 35 минут; объем корпуса – около 4,5 тыс. словоупотреблений. Транскрипты доступны в трех формах: полной, упрощенной и минимальной; доступно проигрывание определяемых пользователем отрывков.

*Корпус на армянском языке*

Корпус состоит из 20 рассказов и 20 пересказов десяти носителей армянского языка. Записи производились в Ереване в 2004–2005 гг. Все информанты являются ереванцами, возраст информантов на момент записи – от 17 до 25 лет. Общая длительность звучания – около 42 минут; объем корпуса – около 4,5 тыс. словоупотреблений. Для корпуса выполнена модифицированная минимальная транскрипция, включающая в себя армянскую графику, латинизированную запись, глоссинг и русский перевод.

*Корпус на японском языке*

Корпус состоит из 20 рассказов и 20 пересказов носителей японского языка. Записи производились в Москве в 2005–2006 гг. Информантами являются японские студенты и аспиранты, проходившие стажировку в Российском гуманитарном государственном университете. Для всех говорящих японский язык является родным. Возраст говорящих – от 21 до 30 лет. Общая длительность звучания – около 45 минут; объем корпуса – около 4,1 тыс. словоформ. Для корпуса выполнена модифицированная минимальная транскрипция, включающая в себя латинизированную транскрипцию, глоссинг, запись в оригинальной японской графике и русский перевод.

***Проекты Школы лингвистики ВШЭ***

*Устные корпуса*[1]

Все корпуса, размещенные на этой странице, представляют устную речь определенного региона России и содержат аудиофайлы и их расшифровки, сделанные в стандартизованной орфографии. Поисковая система позволяет прослушивать фрагменты текстов, содержащие искомое слово или сочетание слов. Для многих корпусов доступны полные тексты. Все корпуса созданы при активном участии Международной лаборатории языковой конвергенции НИУ ВШЭ.

*Учебные корпуса для русского языка как второго*

- Корпус русской речи в Дагестане
- Корпус русской речи в Чувашии
- Корпус русской речи Башкирии

*Русские диалекты*

- Корпус бассейна реки Устья
- Корпус говора села Роговатка
- Корпус говора села Спиридонова Буда
- Корпус говора села Малинино
- Корпус опочецких говоров

*Языки народов России*

- Устный корпус башкирского языка деревни Рахметово и села Баимово
- Устный корпус диалектов хакасского языка
- Устный корпус абазинского языка

[1] Устные корпуса. – URL: https://ilcl.hse.ru/corpora (дата обращения: 01. 12.2021).

- Устный корпус адыгейского языка (темиргоевский диалект)
- Устный корпус бесленеевского диалекта кабардино-черкесского языка

*Увулярные согласные в языках Кавказа*[1]

База данных увулярных согласных подсистем коренных языков Кавказа из всех ветвей (северокавказская, восточнокавказская и картвельская). Данные для базы данных были собраны из существующих языковых описаний и полевых работ. Всего было проанализировано 39 языков. Эта база данных позволяет систематически сравнивать инвентари увулярных согласных в кавказских языках и создавать прогностическую модель сосуществования тех или иных увулярных согласных.

*Всемирная база данных чередования согласных*[2]

База данных, включающая материал синхронных чередований в языках мира.

*Всемирная база данных систем письма*[3]

Эта база данных содержит корреляции между графемами и фонологией. Основная цель этой базы данных – исследовать вариабельность омографии и гомофонии и найти некоторые корреляции между ними и некоторыми другими лингвистическими и социолингвистическими переменными.

*Акцентуатор*[4]

Веб-сервис предлагает воспользоваться двумя системами, проставляющими ударение в тексте на русском языке. Акцентуатор на нейросетях быстро и с высокой точностью обрабатывает большие объемы текста. Вторая система основана на правилах, помимо самого ударения она выдаст обоснование выбора ударения. Нейронный акцентуатор доступен также в виде пакета для Python.

*Мультимедийный корпус идиш*[5]

Корпус устного современного идиша; латиница, аудиозаписи предложений, возможность поиска по словоформам, леммам или грамматической информации.

*Сентинет*[6]

База данных тональности русских прилагательных.

---

[1] On Uvular Consonants in Languages of the Caucasus. – URL: https://agricolamz.github.io/uvular_database/ (дата обращения: 01.12.2021).

[2] The World Consonant Alternation Database. – URL: https://agricolamz.github.io/wcad/ (дата обращения: 01.12.2021).

[3] The World Writing System Database. – URL: https://agricolamz.github.io/wwsd/ (дата обращения: 01.12.2021).

[4] sStress. – URL: https://www.hse.ru/data/2018/01/30/1163745908/akcent.png (дата обращения: 01.12.2021).

[5] Мультимедийный корпус языка идиш. – URL: http://web-corpora.net/YiddishMultimediaCorpus/search/ (дата обращения: 01.12.2021).

[6] Сентинет. – URL: http://web-corpora.net/~adaneyko/index.html (дата обращения: 01.12.2021).

***Прочие российские мультимедийные и аудио ЛИР***

*Аудиословарь «Русский устный»*[1]

Интернет-версия «Русского устного» снабжена рубрикатором, который поможет найти интересующие записи. Запись для прослушивания можно выбрать двумя способами: с помощью алфавитного перечня ключевых слов или обратившись к полному списку передач. Красным цветом выделено правильное ударение во всех словах перечня.

*Электронная библиотека русских народных говоров*[2]. На текущий момент в электронной библиотеке выставлены фонограммы, которые были оцифрованы в 2008–2009 гг. Часть фонограмм прошла первичную лингвистическую обработку, и пользователь имеет возможность параллельно с прослушиванием ознакомиться с содержанием электронной библиотеки. Файлы для прослушивания выставлены в формате *ogg*, также можно использовать программу *RealPlayer*, возможность прослушивания формата *ogg* в эту программу включена.

*Фонотека записей звучащей речи*[3] Кабинета русской диалектологии филологического факультета МГУ насчитывает приблизительно от 1600 до 1800 часов звучания.

*Аудиовизуальный фонд эвенского языка*[4]

Лингвистические и фольклорные материалы, собранные от эвенов, населяющих районы компактного проживания эвенского этноса Республики Саха (Якутия), Камчатки, Магаданской области, Хабаровского края и Чукотки, размещаются в соответствии с территориальной локализацией групп. При этом учитывается специфика функционирования говора или диалекта в условиях меняющегося мира и особенности этнокультурной ситуации. Лингвистическое представление включает расшифровку материала в принятой в эвенской диалектологии фонетической транскрипции и перевод.

*Мультимедийный лингвострановедческий словарь «Россия»*[5]

Словарь создан в Государственном институте русского языка им. А.С. Пушкина на базе книжных версий Большого лингвострановедческого словаря «Россия» (2007–2009) и дополнен мультимедийным контентом: вербальными текстами различных жанров, репродукциями, фотографиями, аудиозаписями, видеофрагментами, панорамами, караоке, интерактивными заданиями.

---

[1] Русский устный. – URL: http://gramota.ru/slovari/radiosafonova/ (дата обращения: 01.12.2021).

[2] Ибрагимов Тавзих Ибрагимович, Кульшарипова Равза Экзамовна. Электронная библиотека русских народных говоров Казанского университета: возможности применения, информационный потенциал //Международный журнал экспериментального образования.- М., 2013.- С. 95–96.

[3] Кабинет русской диалектологии. – URL: http://www.philol.msu.ru/~dialectology/ (дата обращения: 01.12.2021).

[4] Аудиовизуальный фонд эвенского языка. – URL: http://igi.ysn.ru/indexsakha.php?page=karta (дата обращения: 01.12.2021).

[5] Лингвострановедческий словарь Россия. – URL: https://ls.pushkininstitute.ru/lsslovar/index.php (дата обращения: 01.12.2021).

*Аудиословарь русского языка*[1]

Словарь состоит из 20 частей и mp3-файлов в нем тоже двадцать. Словарь составлен на основе передачи по радио, носящей название «уроки русского». Словарь будет полезен не только школьникам и студентам, но и всем тем, кто желает улучшить свои познания в русском языке.

## Литература к главе 14


1. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпусы (опыт разработки и использование). – URL: http://www.dialog-21.ru/digest/2001/articles/krivnova (дата обращения: 01.12.2021).
2. Кривнова О.Ф. Речевые корпуса на новом технологическом витке // Речевые технологии = Speech technology. – 2008. – № 2. – С. 14–23. – URL: http://speechtechnology.ru/files/2-2008.pdf (дата обращения: 01.12.2021).
3. Захаров В.П., Богданова С.Ю. Корпусная лингвистика : учебник для студентов направления «Лингвистика». – 2-е изд., перераб. и дополн. – Санкт-Петербург : СПбГУ, РИО, Филологический факультет, 2013. – 148 с.
4. Потапова Р.К., Потапов В.В. Речевые базы данных как часть мультимодальных корпусов в Интернете // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2018. – № 6(797). – URL: https://cyberleninka.ru/article/n/rechevye-bazy-dannyh-kak-chast-multimodalnyh-korpusov-v-internete (дата обращения: 30.03.2022).
5. Речевая коммуникация в информационном пространстве : коллективная монография / Потапова Р.К., Потапов В.В., Долинский В.А., Хитина М.В., Харламов А.А., Баженова И.Ю., Комалова Л.Р., Бобров Н.В., Гордеев Д.И., Оськина К.А. ; отв. ред. Потапова Р.К. – Москва : ЛЕНАНД, 2017. – 112 с. – ISBN 978-5-9710-3940-2
6. Лосева Е.В. Формирование многоязычной фонетической базы данных (применительно к речевой реализации вибрантов) : дис. ... канд. филол. наук : 10.02.21. – Москва, 2006. – 177 с. – РГБ ОД, 61:07–10/468.
7. Речевые базы данных / Викторов А.Б., Викторова К.О., Воронцова А.В. [и др.] // Современные речевые технологии : сб. трудов IX сессии Российского акустического общества. – Москва, 1999.
8. Бабин Д.Н., Мазуренко И.Л., Холоденко А.Б. О перспективах создания системы автоматического распознавания слитной устной русской речи. – URL: http://intsys.msu.ru/magazine/archive/v8(1-4)/babin-045-070.pdf (дата обращения: 01.04.2022).
9. Продеус А.Н. Речевые корпуса: создание и проблемы // Электротехнические и компьютерные системы. – 2013. – № 09(85). – С. 118–126.
10. Потапова Р.К., Потапов В.В. Речевая коммуникация: от звука к высказыванию. – Москва : Языки славянских культур, 2012. – 464 с. – (Studia Philologica).
11. База речевых фрагментов русского языка «ISABASE» / Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В. // Сборник : Интеллектуальные технологии ввода и обработки информации. – Москва : Эдиторил УРСС, 1998. – С. 74–85.


[1] Аудио словарь русского языка. – URL: http://klassikaknigi.info/audio-slovar-russkogo-yazyka/ (дата обращения: 01.12.2021).

12. Гришина Е.А., Савчук С.О. Корпус устных текстов в Национальном корпусе русского языка : состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – Санкт-Петербург : Нестор-История, 2009. – С. 129–149.

13. Гришина Е.А. Корпус «История русского ударения» // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – Санкт-Петербург : Нестор-История, 2009. – С. 150–174.

14. Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – Санкт-Петербург : Нестор-История, 2009. – С. 175–214.

# ГЛАВА 15. ЛИНГВИСТИЧЕСКИЕ КАРТЫ И АТЛАСЫ

## Общие сведения

ЛИР, опирающиеся на географические информационные системы (ГИС) и предназначенные для изучения пространственных аспектов языков, диалектов и языковых свойств, называются лингвистическими картами. Здесь и далее мы будем говорить о цифровых лингвистических картах, хотя лингвистические карты появились давно и активно применялись и в докомпьютерную эру. Подробно об истории лингвистического картографирования, а также основных методах этой дисциплины главным образом применительно к русской диалектологии, можно прочитать в работе Н.Н. Пшеничной [1].

Лингвистические карты являются инструментом и результатом исследований ряда лингвистических дисциплин, которые различные авторы поразному соотносят между собой. Здесь мы приведем общепринятые определения этих дисциплин.

*Лингвистическая география* – раздел лингвистики, изучающий вопросы территориального размещения языков и распространения языковых явлений. Изучает географии языковых явлений (так называемые «изоглоссы») различного уровня. Многофункциональна, является как одним из разделов общей лингвистики, так и более частных – диалектографии и диалектологии. Основная задача лингвистической географии – составление подробных языковых карт.

*Арéальная лингви́стика*[1] (пространственная лингвистика) (от лат. area – площадь, пространство) – раздел языкознания, изучающий распространение языковых явлений в пространственной протяженности и межъязыковом (междиалектном) взаимодействии на основе методов лингвистической географии. Ареальная лингвистика выявляет ареалы взаимодействия диалектов, языков, языковых союзов (ареальных общностей) в результате изучения территориального распространения языковых особенностей и интерпретации изоглосс языковых явлений.

*«Геолингвистика* – одна из отраслей лингвистики, содержательное наполнение которой существенно разнится в различных лингвистических школах. Некоторые лингвисты рассматривают ее как раздел лингвистической географии, другие рассматривают ее более широко, как разновидность соци-

[1] Арéальная лингви́стика. – URL: https://ru.wikipedia.org/wiki (дата обращения: 01.12.2021).

ально-экономической географии. Одни исследователи стремятся минимизировать область геолингвистики, другие, наоборот, придают ей значение своего рода планетарной модели, описывающей общемировую лингвистическую ситуацию» [2].

*Языковая география*[1] это дисциплина географии, которая изучает географическое распространение языка (языков) или его составляющих. Лингвистическая география также может относиться к исследованиям того, как люди говорят о ландшафте. Например, топонимия – это изучение географических названий.

Лингвистическая карта – это тематическая карта, показывающая географическое распределение говорящих на языке, или изоглоссы из диалектного континуума одного и того же языка. Коллекция таких карт представляет собой лингвистический атлас.

Отметим, что Ю.Б. Коряков предлагает различать лингвистические и языковые карты:

«Лингвистические карты демонстрируют распространение некоторых языковых явлений (слов, фонем, рефлексов, грамматических признаков и т.д.) с помощью точек, изоглосс или областей

- в пределах одного языка (например, на диалектологических картах; ALF 1902–10; ЕННА 2008–2010 и др.)
- в одном регионе / семье (например, OLA 1965–2011, ALFE 2004; Atlas Linguarum Europae 1975–2002)
- по всему миру (WALS 2005)

Собственно языковые карты демонстрируют распространение языков или их диалектов, сгруппированных по общей территории (языковой Атлас Китая 1987), общей генеалогической принадлежности (ACL 2006), еще одной общей черте (ALICPAA 1996).

Эти два различных типа карт связаны с различными отраслями языковой географии. Лингвистические карты служат для лингвистической географии, которая имеет дело с региональной лингвистикой вариации внутри языков. А собственно языковые карты – это инструмент географии языков, которая имеет дело с распространением языков в пространстве (и истории)» [3].

Применение ГИС в лингвистике было предметом исследований многих ученых. В качестве примера приведем работу [4]. В ней говорится:

«Доказано, что географические информационные системы (ГИС) и связанная с ними геопространственная аналитика являются комплексными и высокоэффективными инструментами. Они широко применяются в лингвистических исследованиях с 1990-х годов и дают плодотворные результаты. К ним относятся управление, хранение и анализ данных, пространственный и временной анализ, а также картирование и визуализация лингвистических данных. Даже при таких достижениях все еще остается много возможностей для дальнейшего применения ГИС в лингвистических исследованиях. Будущие

[1] Языковая география. – URL: https://wiki2.wiki/wiki/Language_geography (дата обращения: 01.12.2021).

направления включают интеграцию ГИС с пространственно-временным анализом, с технологией GPS, с веб-картографической технологией, визуализацией и анализом неопределенности лингвистических данных и другими. Кроме того, моделирование полученных лингвистических данных для поиска встроенных признаков или проверки ожидаемых пространственных и даже временных паттернов может помочь вывести лингвистические исследования на новый путь, выходящий за рамки простого построения изоглосс или статистических сводок. Моделирование также позволило лингвистическим исследователям изучать пространственные процессы изменения языковых тенденций и особенностей в пространстве и во времени».

И еще цитата: «В то время как языковое разнообразие является неотъемлемым компонентом культурных ландшафтов, пространственное изображение языков не отражает всех членов сообщества. Язык трудно сопоставить, а установленные руководящие принципы отсутствуют. Восприятие передаваемой силы, возможно, является наиболее значимой проблемой дизайна в языковом картографировании, поскольку большинство языковых карт неточно показывают один язык на место» [5].

## Крупнейшие международные проекты лингвистических карт и атласов

***Атлас языков Европы, ALE***

Первым компьютеризированным лингвистическим атласом был Атлас языков Европы, ALE[1] – исследовательский проект, посвященный картированию лексических и грамматических черт всех языков, распространенных в Европе. Атлас покрывает территорию 51 страны от Исландии до России (до Урала), лингвистические данные собираются для 2631 пункта. Исследование охватывает распространенные на данной территории языки индоевропейской, уральской, алтайской, баскской и семитской семей, а также кавказские языки; в общей сложности, 22 языковые группы с 90 языками и диалектами. Это – крупнейший международный исследовательский проект в истории лингвистической географии, как по величине обследуемой территории, так и по количеству привлекаемых языков. Проект начат в 1970 году с помощью ЮНЕСКО и издавался с 1975 по 2007 год. ALE использовала свою собственную фонетическую систему транскрипции, в основе – Международный фонетический алфавит с некоторыми изменениями. Он охватывает шесть языковых семей, присутствующих на европейском континенте: алтайскую, баскскую, индоевропейскую, каузасскую, семитскую и уральскую; эти семьи разделены на 22 языковые группы, состоящие из 90 языков и диалектов.

***Всемирный атлас языковых структур WALS***[2]– одна из крупнейших открытых баз данных в области лингвистической типологии, упомянутая

[1] Atlas Linguarum Europae (ALE). – URL: http://www.lingv.ro/ALE.html (дата обращения: 01.12.2021).

[2] The World Atlas of Language Structures (WALS). – URL: https://wals.info/ (дата обращения: 01.12.2021).

выше. Карты распространения языковых явлений WALS были изготовлены на основе Google Maps. Атлас содержит информацию о географическом распределении важнейших структурных языковых признаков. Он включает 144 главы, каждая из которых посвящена одному признаку, и 160 карт с изображением географического распределения. Количество записей в базе данных (т.е. индивидуальных пар «язык – реализуемое в языке значение признака») в версии 2005 года составило 58 000[3], в дополненной онлайн-версии – 76 492. Всего языков – 2679.

Ядро WALS составляют главы, посвященные грамматическим признакам и составляющие восемь основных разделов: фонология, морфология, категории имени, синтаксис имени, категории глагола, порядок слов, простое предложение, сложное предложение. Один раздел представляет лексическую типологию, однако не систематическим образом. Также имеются разделы, посвященные жестовым языкам, паралингвистическим звукам и системам письма.

Информация о каждом из рассматриваемых признаков представлена на сайте проекта в нескольких видах:

- каждому признаку посвящена отдельная страница, на которой указаны значения признака и приводится список языков, реализующих каждое из значений;
- каждому признаку посвящена описательная глава, в которой о признаке, его значениях и их распределении по языкам рассказывается в форме текста (с примерами из языков);
- распределение значений признаков можно также посмотреть на карте, где каждое значение изображается специальным знаком.

Также имеются разделы, посвященные:

- языкам: для каждого языка приводится основная информация о нем, а также список значений всех признаков для данного языка;
- авторам: приводится список авторов с указанием тех признаков, которые были исследованы ими для проекта;
- источникам: приводится общий библиографический список с возможностью экспорта библиографических описаний литературных источников в различные форматы (BibTeX, RIS, !EndNote, XML и пр.);
- новостям и обновлениям.

Для WALS разработан специальный *Интерактивный справочный инструмент*[1]. Он позволяет пользователю атласа просматривать карты в различных формах, а кроме того комбинировать объекты, т.е. генерировать составные объекты и также их отображать. Интерактивная база данных тоже содержит дополнительную информацию о языках (генеалогическая классификация, альтернативные названия) и о каждой паре «язык – признак» (библиографическая ссылка, пример предложения). Интерактивные карты можно масштабировать и панорамировать, настраивать цвета и формы точек, пере-

[1] Interactive Reference Tool. – URL: https://www.eva.mpg.de/lingua/research/tool.php (дата обращения: 01.12.2021).

ключать некоторые свойства карты (реки, названия стран и т.д.), а также искать языки по названию языка, названию семьи и рода, стране и региону внутри страны. При наведении курсора мыши сразу же отображается соответствующее название языка, а при щелчке мыши языковой профиль появляется в отдельном окне. Генерация сложных признаков будет очень полезна для типологических исследований. Например, пользователь сможет соотнести существование правила вопросительного слова с определенными типами порядка слов, существование тона с размером инвентаря согласных или тип выравнивания (винительный, эргативный, активный-неактивный) с типом маркировки, зависящей от головы. Кроме того, может быть включена географическая и генеалогическая информация.

Основные возможности *Интерактивного справочного инструмента*:

- автономная версия для MacOSX, MacOS9; Windows 98, 2000, XP со встроенным сервером баз данных – все карты являются масштабируемой векторной графикой (лучшее разрешение для масштабирования);
- интерактивные карты можно масштабировать, панорамировать, настраивать символы, свойства карты можно переключать (реки, границы, названия стран, топологические данные и т.д.);
- эффект наведения курсора мыши для отображения соответствующего названия языка;
- языковой профиль с географической, генеалогической информацией, альтернативными названиями и обзором типологических особенностей;
- 142 предопределенные настраиваемые карты объектов (символы – цвет и форма, слияние и / или скрытие значений объектов);
- каждая пара язык-функция с источниками данных;
- генерация сложных признаков типологических, генеалогических, географических данных вручную или комбинаторно;
- функции импорта / экспорта;
- инструмент для создания пользовательских карт на основе собственных данных;
- карты можно сохранять, распечатывать и копировать в буфер обмена.

***Интерактивный атлас исчезающих языков мира ЮНЕСКО***[1]

Данный ресурс предназначен для повышения осведомленности о языковых угрозах и необходимости защиты языкового разнообразия в мире среди политиков, сообществ носителей языка и широкой общественности, а также в качестве инструмента для мониторинга состояния языков, находящихся под угрозой исчезновения, и тенденции языкового разнообразия на глобальном уровне.

Последнее издание Атласа (2010 г., доступно на английском, французском и испанском языках в издательстве ЮНЕСКО) стало возможным благодаря поддержке правительства Норвегии. В нем перечислены около 2500 языков (из которых 230 языков исчезли с 1950 г.), что приближается к

[1] Endangered languages. – URL: http://www.unesco.org/new/en/culture/themes/endangered-languages/atlas-of-languages-in-danger/ (дата обращения: 01.12.2021).

общепринятой оценке около 3000 языков мира, находящихся под угрозой исчезновения. Для каждого языка в печатном Атласе указывается его название, степень опасности (см. ниже) и страна или страны, в которых на нем говорят.

Онлайн-издание предоставляет дополнительную информацию о количестве выступающих, соответствующих политиках и проектах, источниках, кодах ISO и географических координатах. Эта бесплатная интернет-версия Атласа впервые обеспечивает широкую доступность, интерактивность и своевременное обновление информации на основе отзывов пользователей.

***Язык и местоположение – проект интеграции лингвистических карт LL-MAP***[1]

LL-MAP – это проект, предназначенный для интеграции языковой информации с данными из физических и социальных наук с помощью географической информационной системы (ГИС). Самая важная часть проекта – это языковая подсистема, которая связывает географическую информацию о местности, в которой говорят или говорили на каком-либо языке, с данными о ресурсах, относящихся к этому языку. В рамках проекта информация обо всех предполагаемых генетических связях языков доступна, включая просмотр в географическом контексте. Система также включает дополнительную информацию о топографии, политических границах, демографии, климате, растительности и дикой природе, обеспечивая таким образом основу для построения гипотез о перемещении языков по территории. Также включена некоторая культурная информация, например о религии, этнической принадлежности и экономике.

Система LL-MAP поощряет сотрудничество между лингвистами, историками, археологами, этнографами и генетиками, поскольку они исследуют взаимосвязь между языком и культурной адаптацией и изменениями.

Разработчики считают, что проект поможет выявить новые идеи и гипотезы, а также послужит образовательным ресурсом. Как ГИС, LL-MAP может стать увлекательным учебным инструментом, представляющим сложные данные в доступной форме для всех уровней образования. Наконец, в качестве бесплатной онлайн-услуги LL-MAP расширяет общественные знания о менее известных языках и культурах, подчеркивая важность языка и языкового разнообразия для культурного понимания и научных исследований.

Проект LL-MAP начался как совместный проект Университета Восточного Мичигана (EMU) и Стокгольмского университета в сотрудничестве с несколькими проектами и архивами в США, Европе и Австралии. Техническая поддержка обеспечена Институтом геопространственных исследований и образования (IGRE) в EMU.

Проект LL-MAP в настоящее время ведется в Департаменте лингвистики Университета штата Индиана.

[1] LL-MAP Language and Location – A Map Annotation Project. – URL: http://www.llmap.org/about (дата обращения: 01.12.2021).

***Система картографирования мировых языков (WLMS)***[1] – это наиболее полный, актуальный и надежный набор географических данных о точках и областях (многоугольниках) почти 7 100 живых языков мира.

WLMS является результатом более чем 15-летней совместной работы компании *Global Mapping International* (*GMI*) и SIL по составлению карт более 6 800 языков, описанных в 14-м издании SIL Ethnologue.

Данные предоставляются в формате shape-файла Esri (.shp) и файловой базы геоданных для ГИС-систем. Источником названий и кодов языков является стандарт ISO 639–3 19-го издания (2016 г.). Файлы слоев ArcGIS включены для облегчения работы с символами.

Лингвисты и другие исследователи найдут эти данные ценными для понимания местоположения и распространения языков по всему миру. Новаторским является включение экологических (исследования взаимосвязи между культурным разнообразием людей и биоразнообразием), политологических (исследования взаимосвязи между вооруженным конфликтом и языковыми границами) и других данных.

Основные особенности:

- расположение языков в виде точек и многоугольников (родственных языков) в формате shape-файла Esri (. shp) и файловой базы геоданных;
- введены полигоны для смешанных языковых областей;
- данные накладываются на базовые карты Digital Chart of the World / VMAP0;
- используются файлы слоев и файлы mxd для ArcView 10.1 или выше.

Новые функции в версии 19:

- данные точек и полигонов были обновлены в соответствии со стандартом ISO 639–3 19-го издания (2016 г.). Страны, в которых произошло много изменений, включают Южную Африку, Гану, Малайзию, Перу, Парагвай, Грузию;
- более 60 языков были добавлены в качестве точек, поэтому все живые языки в стандарте ISO 639–3 19-го издания (2016 г.) будут иметь как минимум точку;
- обновлен список иммигрантских, вымерших языков и языков жестов.

Системные требования:

- программное обеспечение ГИС, способное читать файлы формата Esri shape (.shp);
- shape -файлы были протестированы на Esri ArcView 10.1–10.5;
- для полноценного использования файлов проекта и слоев требуется ArcView 10.1 или более поздней версии;
- система Windows для запуска самораспаковывающихся установочных файлов .exe. Компьютер должен иметь возможность записывать данные в место, доступное для ГИС-системы.

[1]World Language Mapping System. – URL: https://www.worldgeodatasets.com/language/ (дата обращения: 01.12.2021).

***Проект лингвистического картографирования Беркли BeLMaP***[1]

Проект BeLMaP был задуман как попытка объединить цифровые инструменты для совместного изучения роли пространства и географических объектов в распространении лингвистических признаков между языками через процессы лингвистической диффузии и заимствования. В частности, цель состояла в том, чтобы в полной мере использовать возможности географических информационных систем (ГИС-платформ).

BeLMaP использует интерфейс браузера для ввода лингвистических и пространственных данных в базу данных, доступную ГИС, и веб-сервер, обеспечивающий визуальные представления, создаваемые платформой ГИС (т.е. карты) в режиме онлайн. Конечная цель состояла в том, чтобы интегрировать эти инструменты, чтобы облегчить удаленное сотрудничество по созданию и анализу общих наборов данных пространственно-индексированных лингвистических данных. Кроме того, BeLMaP стремился протестировать эти инструменты, используя данные из амазонских и кавказских языков как испытательный полигон.

***Языковые карты в БД Этнолог***[2]

Выше мы приводили основные сведения о *БД Этнолог*. Здесь добавим особенности картографической составляющей этой БД. Карты, на которых показаны места функционирования языков, доступны для большинства стран мира. Отображенные на карте языковые области показывают распределение носителей языков на первом языке (L1), если это явно не указано на карте. Распределения второго языка (L2) находятся на отдельных вставках, отличных от основной карты. Там, где данные доступны, используеися условное обозначение языка в районе, если по крайней мере 25% населения в этом районе свободно говорят на этом языке как L1. Тем не менее, язык меньшинства будет отображаться на карте, даже если говорящие не составляют более 25% населения в любом данном районе.

На большинстве карт используются многоугольники, чтобы показать приблизительные границы языковых групп. Справочные номера используются на картах, где свободное место не позволяет разместить названия языков. Для некоторых карт, где языковые границы не известны, имена или числа отображаются отдельно.

Текущие карты нарисованы с помощью программного обеспечения ArcGIS®[3]. Уровень географической детализации карт постоянно повышается, чтобы помочь в поиске языковых групп внутри страны.

Кроме геоданных от ArcGIS, другие карты нарисованы на фоне бесшовной цифровой карты мира (*SDCW*), опубликованной GMI. Геоданные Esri имеют более высокий уровень разрешения, чем базовая карта, используемая для карт в более ранних версиях. В 2020 году все языковые многоугольники

[1] Berkeley Linguistic Mapping Project (BeLMaP). – URL: file:///C:/Users/%D0%90%D0%BB%D0%B5%D0%BA%D1%81%D0%B0%D0%BD%D0%B4%D1%80/Downloads/BeLMaP%20Final%20 Report%20(1). pdf (дата обращения: 01.04.2022).

[2] Ethnologue. – URL: https://store.ethnologue.com/ (дата обращения: 01.04.2022).

[3] ArcGIS Online. – URL: https://www.arcgis.com/index.html (дата обращения: 01.12.2021).

были перемещены, чтобы соответствовать более подробной информации о географических объектах, предлагаемых этим новым набором данных. Эти улучшенные данные будут включены в остальные карты в следующих нескольких выпусках. Точность построения языковых областей постоянно улучшается, особенно за счет более широкого использования данных о местоположении, собранных с помощью устройств GPS и спутниковых изображений, таких как Google Maps и *Esri® Living Atlas*.

Внедрение данных GPS и более широкое использование данных переписи также означает, что данные становятся все более подробными. Гладкие обобщенные кривые заменяются более сложными функциями, и в некоторых странах теперь мы можем более точно представить сложность, когда носители нескольких языковых групп живут в одном районе.

Топонимическая информация основана на базе данных географических названий, содержащей официальные стандартные названия, утвержденные Советом США по географическим названиям и поддерживаемые Национальным агентством геопространственной разведки.

Карты используют различные картографические проекции: африканские экваториальные страны используют синусоидальную проекцию. Другие экваториальные страны используют проекцию Меркатора (цилиндрическую). Карты стран, расположенных в более высоких широтах, используют проекцию конической формы Ламберта.

Как и в случае со всем содержимым «Этнолога», в базе не делается никаких политических заявлений путем выделения какой-либо территории отдельно на карте или в списках языков, а также путем размещения каких-либо линий границ для любых языков или стран на любой карте.

***Отображение языкового разнообразия в глобализирующемся мире с помощью цифровых инструментов с открытым исходным кодом***[1]

Данный проект реализует новое совместное партнерство *Университета Британской Колумбии* и *Альянса исчезающих языков*.

Языковое разнообразие по-прежнему находится под чрезвычайным давлением. Сегодня сообщества, говорящие на все более исчезающих языках, весьма мобильны, часто по необходимости. Поскольку языки движутся вместе с людьми, сопоставление языков затруднено, будь то в печатной или цифровой форме. Методы, которые определяют языки как точки на карте, ошибочны (где бы вы нашли точку для английского?), в то время как полигоны неточно представляют полиязычные реальности. Принимая этот вызов, базирующийся в Нью-Йорке *Альянс исчезающих языков* выпускает популярные печатные карты самого лингвистически разнообразного города мира. Муниципальное руководство в Нью-Йорке использует эти знания для формирования государственной политики. Предлагаемый проект превратит эти карты в универсальные, интерактивные цифровые инструменты, которые поддерживают отображение языков на основе сообщества в любой точке мира.

[1] Mapping Linguistic Diversity in a Globalizing World through Open Source Digital Tools. – URL: https://languagemapping.org/ (дата обращения: 01.12.2021).

***Интерактивная карта языков и диалектов для путешествий LocaLingual***[1]

Проект LocaLingual демонстрирует языковое разнообразие нашей планеты. Он будет полезен не только путешественникам, которые готовятся к зарубежной поездке, но и людям, изучающим новые языки, а также рядовым пользователям, которые хотят оценить мелодику иностранной речи. Проект имеет открытую архитектуру, пользователи постоянно дополняют его новыми словами и фразами, по образу и подобию универсальной интернет-энциклопедии Википедии. За эту особенность автор называет свою карту «Википедией языков».

Путешественники могут воспользоваться языковой картой как на этапе подготовки к зарубежной поездке, так и во время живого общения с иностранцами. Пользоваться картой довольно просто. На ней показаны разные страны мира, с указанием регионов, провинций, областей и штатов. Выбрав на карте страну и местность, пользователь может прослушать аудиозаписи со словами и выражениями, записанными на местном диалекте. Работа с картой поможет усовершенствовать свое произношение.

## Собрания цифровых лингвистических карт

***Сокровищница лингвистических карт TLM***[2]

TLM представляет лингвистические карты, выбранные из различных атласов, которые были опубликованы на протяжении многих лет De Gruyter Mouton и другими изданиями De Gruyter. Впервые этот материал доступен, и доступен для поиска в одном месте и в новом, улучшенном формате. Все карты были отсканированы в высоком разрешении для максимального качества и обогащены подробными метаданными, которые можно использовать для поиска. Например: категория / субдисциплина, язык, географическое положение, год или период, заголовок / описание карты, автор карты, автор книги, название книги, год публикации, полный текст.

Карты – это уникальный и незаменимый ресурс для визуализированной информации по различным темам лингвистики. Многие из этих карт ранее не были доступны в цифровом виде, а некоторые трудно найти даже в печатном виде. Интерфейс карты позволяет увеличивать детализацию, печатать и экспортировать в PDF.

TLM обновляется два раза в год, добавляя около 600 новых карт в каждом обновлении. Полное содержание 8000 карт будет завершено в 2021 году. Обновления включают как заново открывшееся из архивов Де Грёйтера, так и материалы из новых публикаций.

[1] LocaLingual. – URL: https://zen.yandex.ru/media/biletikaero/sozdana-iazykovaia-karta-mira-dlia-puteshestvii-i-izucheniia-dialektov-5ef526befab32a2ddf9eaad6 (дата обращения: 01.12.2021).

[2] Treasury of Linguistic Maps (TLM). – URL: https://www.degruyter.com/database/tlm/html l (дата обращения: 01.04.2022).

***Портал Альфонса Эйленбурга***[1]

На данном портале представлена карта язывковых семей, в которой каждой семье сопутствует следующая информация:

- языковая семья
- предлагаемая макросемья
- количество говорящих
- количество языков в семье

Отдельно показаны языки, имеющие более 10 млн говорящих.

***Карты Мэтью Драйера***[2]

На этом портале представлены следующие карты:

- Карта языковых семей Европы (уменьшенная версия)
- Карта языковых семей Европы (увеличенная версия)
- Карта подсемейств индоевропейского языка в Европе
- Карта языковых семей континентальной части США (уменьшенная версия)
- Карта языковых семей континентальной части США (увеличенная версия)
- Карта языковых семей Калифорнии
- Карта языковых семей Тихоокеанского Северо-Запада
- Карта языковых семей Канады
- Карта языковых семей Южной Азии
- Карта языковых семей Кавказского региона
- Карта языковых семей Евразии
- Карта германских языков
- Карта романских языков: три ответвления
- Карта романских языков: итало-западно-романская ветвь
- Карта романских языков: иберо-романская ветвь итало-западно-романской
- Карта романских языков: западно-иберо-романская ветвь иберо-романской ветви итало-западно-романской
- Карта языковых семей на юго-западе Папуа-Новой Гвинеи

***Проект LangMap***[3]

LangMap – это инструмент визуализации данных для носителей языка, студентов и энтузиастов. В настоящее время на LangMap можно отобразить 100 языков. В этот список входят двадцать наиболее распространенных языков, а также смесь индоевропейских, семитских, индоарийских, уральских, дравидийских, нигеро-конголезских и сино-тибетских языков. Определенные языковые семьи имеют точечный охват или настолько локализованы, что визуализировать их в глобальном масштабе не удается: например, навахо, на котором говорят только в Соединенных Штатах, или кечуа, диалекты которого

[1] Карта мира языковых семей. – URL: https://eylenburg.github.io/languages_map.htm (дата обращения: 01.04.2022).

[2] https://www.acsu.buffalo.edu/~dryer/family_maps.htm

[3] LangMap.me. – URL: https://langmap.me/about (дата обращения: 01.12.2021).

настолько специфичны для страны, что их будет трудно понять, передать правильную карту. Некоторые языки также более документированы, чем другие, и трудно собрать данные в политически нестабильных или бедных регионах. Таким образом, пустая область на карте для языка не означает, что нет говорящих, – это означает, что нет данных.

***Лингвистическая картография Эстремадуры***[1]

Большая коллекция ссылок на лингвистические карты в Интернете имеется на сайте *Лингвистическая картография Эстремадуры*. Они сведены в Интерактивный атлас с дополнительной информацией, а также классифицированы по континентам и странам.

***Диалекты немецкого языка REDE***[2]

Картографические языковые данные в этом проекте интегрируются с современными диалектами немецкого языка, связываются с другими существующими классами данных (звукозаписи опросов, библиографическая информация, социодемографические или историко-административные и политические средства интерпретации) и предоставляются пользователю для систематического сравнительного анализа. В системе создана коллекция свыше 40 лингвистических карт и атласов. Система обладает рядом функциональных возможностей:

*Исследование и просмотр контента*

- наложение языковых карт различных атласов
- сравнение языковых карт с картами диалектов и другими интерпретациями
- фильтрация контента с использованием опций дифференцированного поиска (включая пространственный поиск, поиск с использованием исторических, фонологических и морфоло-гических регистров)

*Создание карт и визуализация данных*

- дизайн базовых карт
- создание тематических карт
- импорт наборов данных (файлы *.csv / *.Kml)
- визуализация наборов пространственных данных, таких как
  - карты круговой диаграммы
  - карты столбчатой диаграммы
  - карты со значками очков
  - карточки оригинальной формы
- создание карт из локальных сетей (преобразование точечных карт в карты областей)
- редактирование данных
- сравнение наборов данных

*Экспорт и публикация*

- сохранение и публикация карт / данных в системе
- экспорт записей данных
- экспорт карты в виде файла изображения с высоким разрешением

[1] Cartografía Lingüística De Extremadura. – URL: http://www.geolectos.com/atlas.htm

[2] Regionalsprache.de (REDE). – URL: https://regionalsprache.de/

***Лингвистический атлас Франции***[1]

Лингвистический атлас Франции является результатом полевого исследования, проведенного Жюлем Жильероном и Эдмоном Эдмоном между 1897 и 1900 гг. Он касается языков langue d'oïl, langue d'oc и francoprovençale, на которых говорят во Франции, Бельгии и франкоговорящей Швейцарии.

## Российские проекты

***Общеславя́нский лингвисти́ческий а́тлас (ОЛА)***[2]

ОЛА – это международный исследовательский проект по исследованию и лингвистическому картографированию фонетических, лексических и грамматических черт всех славянских языков. Атлас покрывает территорию всех славянских стран. Объем сетки – около 850 населенных пунктов. Это один из крупнейших международных исследовательских проектов в истории диалектологии и лингвистической географии, как по величине обследуемой территории, так и по количеству привлекаемых славянских языков.

Объектом ОЛА является группа славянских языков в их совокупности, а не отдельный язык. Это определяет качественное своеобразие Атласа, так как при таком подходе меняется объект картографирования: если национальные атласы изучают диалектные различия в пределах данного языка, имеющие национальное значение, то в ОЛА картографируются различия в пределах всей славянской группы языков, имеющие общеславянское значение.

***ЛингвоДок***

Важную роль цифровые лингвистические карты занимают в российском проекте ***ЛингвоДок,*** описанном выше. Это совместный проект Института системного программирования им. В.П. Иванникова РАН, Института языкознания РАН и Томского государственного университета. На сайте проекта организовано хранилище карт[3], на которых представлены различные типологические характеристики уральских языков. Примером могут служить:

- Карта по системам ударения, гласным фонемам и их изменениям, многозначности в уральских языках
- Общетипологические фонетические особенности в уральских языках
- Ареальные морфологические особенности в уральских языках
- Дифтонги в уральских языках

Всего в хранилище в настоящее время представлено свыше 40 карт.

***Лингвариум***

Из российских цифровых проектов, посвященных лингвистическому картографированию, следует отметить сайт Ю.Б. Корякова *Лингвариум.*

[1] Atlas linguistique de la France. – URL: – https://www.lexilogos.com/atlas_linguistique_france.htm (дата обращения 01.04.2022).

[2] Общеславянский лингвистический атлас (ОЛА). – URL: http://www.slavatlas.org/ (дата обращения: 01.12.2021).

[3] ISPRAS. LINGVODOC-REACT. Хранилище карт. – https://github.com/ispras/lingvodoc-react/wiki/%D0%A5%D1%80%D0%B0%D0%BD%D0%B8%D0%BB%D0%B8%D1%89%D0%B5-%D0%BA%D0%B0%D1%80%D1%82 (дата обращений 01.04.2021).

В разделе *Языковые карты*[1], на котором представлены разнообразные лингвистические карты, обсуждаются принципы их создания, а также имеется библиография по этим вопросам. Карты классифицированы по странам и континентам. Отдельно выделены исторические лингвистические карты.

Перечислим далее еще несколько российских проектов в этой области.

***Научно-учебная группа «Прикладная гуманитарная геоинформатика» в Санкт-Петербургской ВШЭ***[2]

Деятельность группы будет осуществляться по трем направлениям: лингвистическому, макро- и микроисторическому. В рамках первого направления будут разработаны критерии и стандарты лингвистического картографирования, а также исследованы возможности создания картографической базы данных разных языков. База данных позволит накапливать результаты исследований по разным областям лингвистики в качестве языковых параметров, что позволит впоследствии находить корреляции в рамках разных лингвистических поддисциплин. В рамках второго направления будет осуществляться картографирование политического пространства федерализма, регионализма и автономизма в Российской империи и Советском Союзе. Деятельность в рамках третьего направления будет посвящена созданию мультимедийной геоинформационной системы района расположения нового кампуса НИУ ВШЭ СПб на Васильевском острове. Все три направления будут объединены методологически через единую программную среду – Google Earth и Quantum GIS и теоретические подходы цифровых гуманитарных наук. Разработка методологии гуманитарной геоинформатики является ключевой задачей группы.

***Лингвистические карты от Мутурзикина***[3]

На сайте собрано достаточно много лингвистических карт – свыше 220 карт по отдельным странам, а также коллекции лингвистических карт по континентам:

- Лингвистическая карта Африки
- Лингвистическая карта Европы
- Лингвистическая карта Северной Америки
- Лингвистическая карта Латинской Америки
- Лингвистическая карта Азии
- Лингвистическая карта Океания
- Карта юго-восточных азиатских языков

Например, коллекция *Лингвистическая карта Африки* включает следующие карты:

- Языки в Северо-западной и Западной Африке

[1] Языковые карты. – URL: http://www.lingvarium.org/maps.shtml (дата обращения: 01.12.2021).

[2] Научно-учебная группа «Прикладная гуманитарная геоинформатика». – URL: https://spb.hse.ru/soc/gis/project (дата обращения: 01.12.2021).

[3] Лингвистические карты от мутурзикин. ком. – URL: https://www.muturzikin.com/countryru.htm (дата обращения: 01.12.2021).

- Языки Восточной и Центральной Африки
- Языки в Южной Африке
- Языки в Северной Африке
- Языки в Нигерии и Камеруне
- Африканские языковые семьи
- Разговорные языки в каждой стране
- Африканская карта на баскском языке

Кроме того, коллекция включает список языков, на которых говорят в каждой стране; в списке более 1500 африканских языков.

***Разработка инструментов лингвокартографирования и поисковой системы, создающей карту на основании заданных параметров***[1]

Авторы проекта Г.А. Мороз, И.В. Саблин, К.В. Дубова

***Карта языковых ситуаций на территории России***[2]

В описании проекта говорится:

«На территории России распространены десятки и сотни различных языков, при этом все они имеют разную витальность – часть из них находится во вполне благополучном состоянии, в то время как другие имеют статус «находящихся под угрозой» (endangered) или даже «умирающих» (moribund). Описание языковых ситуаций – классическая задача макросоциолингвистики, а современные технические методы позволяют создавать базы данных и карты языковых ситуаций – как для исследовательских целей, так и для привлечения внимания широкой публики к проблемам языкового сдвига и вымирания языков. При этом подобного проекта, посвященного исключительно территории России, никогда не проводилось – а и научная, и практическая необходимость в нем очевидна. Предлагаемый проект ставит задачу заполнить этот пробел: мы планируем создать полноценную (в рамках доступных данных) карту языковых ситуаций на территории России, которая будет (1) обобщать социолингвистические данные по витальности языков народов России и наглядно демонстрировать степень такой витальности в различных регионах и стадию языкового сдвига для разных языков; (2) показывать степень социолингвистической изученности различных языков и территорий. Проект подразумевает как исследовательскую составляющую – сбор и анализ данных (не только в научной литературе, но и в медиа, соцсетях и статистике), так и прикладную (собственно картографирование)».

***Методы компьютерной лингвогеографии в исследовании границ между близкородственными языками***

Данное исследование Н.Г. Горлова и его коллег можно назвать примером применения методов компьютерной лингвогеографии для решения конкретных лингвистических задач. В описании проекта говорится:

[1] Лингвистическое картографирование. – URL: https://ling.hse.ru/Projects_LingMaps (дата обращения: 01.12.2021).

[2] Карта языковых ситуаций на территории России. – URL: https://hum.hse.ru/proj/map (дата обращения: 01.12.2021).

«Цель исследования – применить новейшие компьютерные методы кластеризации и визуализации к достоверным и количественно релевантным данным по смежным диалектам двух близкородственных южнославянских языков, сербского и болгарского [SAOSWB]. Задачи исследования состоят в разработке цифрового инструментария обработки и кластеризации первичных данных и в генерировании пробных лингвистических карт. Гипотезой теоретического исследования является предположение о том, что в результате применения методов систематизации, анализа, синтеза и визуализации географической дистрибуции кластеров языковых и экстралингвистических данных станет возможной картографическая экспликация объективных границ между близкородственными языками, в частности – между сербским и болгарским. Практическим результатом исследования станет преодоление относительного неудобства и несовершенства печатных лингвистических атласов, в частности – очевидных ограничений, накладываемых самим их форматом. Это и невозможность масштабирования карт и добавления в них новой информации (от новых пунктов до новых языковых данных), и трудоемкость сопоставления символов на лингвистических картах с данными из прилагающихся таблиц, и неудобства ручного наложения сетки пунктов или фоновой физико-географической карты на карты лингвистические, и невозможность создания большого количества комбинированных и диалектометрических карт и т.д.; в целом, такой формат лишен динамичности, и работа с ним представляется излишне трудоемкой и крайне ограниченной в плане интерактивности. Эти проблемы решены в результате разработки нового цифрового лингвогеографического инструментария» [6].

В работе *Русская диалектология* справедливо отмечается:

«При картографировании осуществляется последовательное различение разных уровней языка. Каждая карта посвящена явлению одного уровня. При этом на каждом следующем, более высоком уровне не учитываются (объединяются) те языковые различия, которые связаны с варьированием единиц более низкого уровня» [7].

Еще примеры применения методов компьютерной лингвогеографии:

***Электронные карты пинежских говоров*** (по материалам Словаря пинежских говоров) [8]

***Картографирование диалектов удмуртского языка*** [9]

В заключение укажем, что в России цифровые лингвистические карты создаются также в образовательных и просветительских целях, например, ***Языковые карты мира***[1].

[1] ИНФОКАРТ. Все карты мира. – URL: https://www.infokart.ru/yazykovaya-karta-mira/

## Литература к главе 15

1. Пшеничнова Н.Н. Лингвистическая география (по материалам русских говоров). – Москва : Азбуковник, 2008. – 222 с.

2. Кузнецов С.Н. Геолингвистика. Спецкурс. – 2007. – Sep. 5. – URL: http://genhis.philol.msu.ru/article_197.shtml (дата обращения: 01.12.2021).

3. Создание языковых карт. – URL: http://www.lingvarium.org/koryakov/Map-creating.shtml (дата обращения 01.04.2022).

4. Jay Lee, Jiajun Qiao, Dong Han. GIS in Linguistic Research. – 2018. – DOI: 10.1016/B978–0-12–409548–9.09662–7. – URL: https://zh.booksc.eu/book/71851422/8f71b9

5. Candice R. Luebbering, Korine N. Kolivras, Stephen P. Prisley. Visualizing Linguistic Diversity Through Cartography and GIS // The Professional Geographer. – 2013. – November, N 65(4). – DOI: 10.1080/00330124.2013.825517

6. Горлов Никита Геннадьевич, Кочановская Анна Вячеславовна, Соболев Андрей Николаевич. Методы компьютерной лингвогеографии в исследовании границ между близкородственными языками (на примере диалектов Восточной Сербии и Западной Болгарии) // Вестн. Том. гос. ун-та. Филология. – 2020. – № 64. – URL: https://cyberleninka.ru/article/n/metody-kompyuternoy-lingvogeografii-v-issledovanii-granits-mezhdu-blizkorodstvennymi-yazykami-na-primere-dialektov-vostochnoy-serbii-i (дата обращения: 30.03.2022).

7. Русская диалектология / Бромлей С.В. , Булатова Л.Н., Гецова О.Г. [и др.] ; под ред. Касаткина Л.Л. – Москва, 2005. – 288 с.

8. Левичкин А.Н., Крылова О.Н., Гайдамашко Р.В. Электронные карты пинежских говоров (по материалам Словаря пинежских говоров) // Лексический атлас русских народных говоров (Материалы и исследования). – Санкт-Петербург : ИЛИ РАН, 2018. – С. 212–225. – ISSN 2658–6150. – DOI: 10.30842/26586150201819–

9. Рублева Е.А., Саранча М.А. Геоинформационные технологии в картографировании диалектов удмуртского языка / Удмуртский государственный университет. – URL: http://www.geogr.msu.ru/cafedra/karta/anniversary/docs/rubleva.pdf (дата обращения: 01.12.2021).

# ГЛАВА 16. РЕСУРСЫ ЖЕСТОВЫХ ЯЗЫКОВ

## Общие сведения

Жестовые языки являются языками, которые используют визуально-ручную модальность, чтобы передать смысл. Языки жестов выражаются посредством ручной артикуляции в сочетании с элементами, не являющимися ручными. Языки жестов – это полноценные естественные языки со своей грамматикой и лексикой. Языки жестов не универсальны и не взаимно понятны друг другу, хотя между языками жестов также есть поразительное сходство. Лингвисты считают, что и устное, и жестовое общение являются типами естественного языка, а это означает, что оба они возникли в результате абстрактного, длительного процесса старения и развивались с течением времени без тщательного планирования. Язык жестов не следует путать с языком тела, типом невербального общения. Же́стовый язы́к – самостоятельный язык, состоящий из жестов, каждый из которых производится руками в сочетании с мимикой, формой или движением рта и губ, а также в сочетании с положением корпуса тела. Эти языки в основном используются в культуре глухих и слабослышащих с целью коммуникации. Использование жестовых языков людьми без нарушения слуха вторично, однако довольно распространено: часто возникает потребность в общении с людьми с нарушениями слуха, являющимися пользователями жестового языка.

Одним из главных неправильных представлений о жестовых языках является представление, что они каким-то образом зависят от словесных (звуковых и письменных) языков или произошли от них, что эти языки были придуманы слышащими, однако это не так.

Так же часто за жестовые языки принимается дактилирование букв (на самом деле оно используется в жестовых языках в основном для произнесения имен собственных, географических названий, а также специфичных терминов, взятых из словесных языков), калькирующая жестовая речь, или жестовое артикулирование, используемая слышащими для передачи информации жестами грамматически идентично словесному языку. На самом же деле жестовые языки почти полностью независимы от словесных, и они продолжают развиваться: появляются новые жесты, отмирают старые – и чаще всего это мало связано с развитием словесных языков. Количество жестовых языков в стране не связано с количеством в ней словесных языков. Даже в одной стране, где присутствует несколько словесных языков, может

быть единственный общий жестовый язык, и в некоторых странах даже с одним словесным языком могут сосуществовать несколько жестовых[1].

## Список жестовых языков[2]

Сегодня в мире используется около трехсот жестовых языков. Число точно не известно; новые жестовые языки часто появляются в результате креолизации и de novo (а иногда и в результате языкового планирования).

В некоторых странах, таких как Шри-Ланка и Танзания, в каждой школе для глухих может быть отдельный язык, известный только учащимся и иногда запрещенный школой; с другой стороны, разные страны могут использовать один и тот же язык жестов, хотя иногда и под разными названиями (хорватский и сербский, индийский и пакистанский).

Глухие жестовые языки также возникают за пределами учебных заведений, особенно в деревенских общинах.

Предлагаемый список сгруппирован в три раздела.

*Языки жестов глухих*, которые являются предпочтительными языками сообществ глухих во всем мире; к ним относятся деревенские жестовые языки, распространенные среди слышащего сообщества, и жестовые языки глухих.

Наряду с разговорным языком используются *вспомогательные жестовые языки*, которые не являются родными языками, а являются жестовыми системами различной сложности. Простые жесты не включены, поскольку они не составляют языка.

*Знаковые режимы разговорных языков*, также известные как языки с ручным кодированием, которые являются мостами между жестовыми и устными языками.

Список жестовых языков глухих содержит около 250 названий. Выделены региональные категории:

- Африка – 44
- Америка – 50
- Азиатско-Тихоокеанский регион – 55
- Европа – 47
- Средний Восток – 22

Другие виды жестовых языков

- Исторические глухие жестовые языки – 3
- Вспомогательные жестовые языки – 8
- Ручные режимы разговорных языков – 8

Каждый включенный в список язык содержит отсылку к статье Википедии, где имеется его описание. Региональная привязка не связана с генетической классификацией жестовых языков, которая приводится отдельно[1].

---

[1] Жестовые языки. – URL: https://ru.wikipedia.org/wiki/%D0%96%D0%B5%D1%81%D1%82%D0%BE%D0%B2%D1%8B%D0%B5_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B8 (дата обращения: 01.12.2021).

[2] Список жестовых языков – List of sign languages. – URL: https://ru.abcdef.wiki/wiki/List_of_sign_languages (дата обращения: 01.12.2021).

## Обследование ЛИР жестовых языков

Было проведено комплексное обследование доступных в Интернете словарей жестовых языков, которое выполнил российский исследователь А.Е. Харламенков из НИИ «Русского жестового языка» [1].

Ниже будут кратко изложены результаты этого обследования.

Методика сравнительного анализа наиболее известных русских и зарубежных словарей национальных жестовых языков имела следующие цели:

- сбор и обработка информации по существующим словарям жестового языка;
- выявление проектов, которые можно использовать в качестве примера;
- выявление условий создания успешных проектов online-словарей жестового языка.

В ходе обследования были выявлены ключевые особенности словарей:

- статус развития словаря – завершен или пополняется;
- назначение словарей;
- функционал и технологические возможности словарей;
- уровень доступа к словарям.

В результате были сформулированы общее описание словаря и общий вывод.

При анализе использовались следующие критерии оценки:

- авторитетность
- объем
- непротиворечивость
- точность
- описательность
- качество исполнения
- качество отображения

На основе данных критериев была разработана критериально-балльная система, призванная привести к «общему знаменателю» чрезвычайно разнородные сложные объекты.

В ходе обследования было изучено 66 словарей, представленных в сводной таблице на сайте «Лаборатории лингвистики жестового языка (ЛЛЖЯ)»[2]. Также описаны два словаря, известных руководству НИИ РЖЯ, но не указанных в перечне ЛЛЖЯ. Всего обследовано 68 словарей. В перечне в том числе оказалось:

- четыре – дактильной азбуки (анимированные и статичные);
- 12 словарей прекратили свое существование;

---

[1] Язык знаков.– URL: https://ru.abcdef.wiki/wiki/Sign_language#Classification (дата обращения: 01.12.2021).

[2] Лаборатория лингвистики жестового языка. – URL: https://signlang.ru/links/slovari (дата обращения: 01.12.2021).

- 25 являются либо примитивным рисуночным кратким справочником, либо дублем уже представленных в списке словарей;
- по 27 online-словарям составлены детальные отчеты.

На сайте ЛЛЖЯ[1] представлены отчеты по нескольким словарям, которые содержат описание и экспертные оценки каждого словаря в соответствии с критериально-балльной системой:

- online-словарь «Jestov.Net»
- online-словарь «DigitGestus»
- online-словарь «DeafNet»
- online-словарь «Сурдосервер»
- online-словарь «Сурдопортал» МГТУ им. Баумана
- online-словарь «Spreadthesign», мультинациональный словарь жестовых языков

На сайте ЛЛЖЯ представлены следующие ЛИР жестовых языков:

- мультиязыковые жестовые словари (4)
- русский жестовый язык (РЖЯ, Russian Sign Language, RSL) (7)
- австралийский жестовый язык (Australian Sign Language, Auslan) (2)
- австрийский жестовый язык (Austrian Sign Language, Österreichische Gebärdensprache, OEGS, ÖGS) (1)
- американский жестовый язык (American Sign Language, ASL, Ameslan) (7)
- арабский жестовый язык (Arabic Sign Language, ArSL) (2)
- аргентинский жестовый язык (Argentine Sign Language, Lengua de Señas Argentina, LSA)
- бразильский жестовый язык (Brazilian Sign Language, Língua Brasileira de Sinais, Libras, LSB, LGB, Brazilian Cities Sign Language, LSCB) (2)
- британский жестовый язык (British Sign Language, BSL) (5)
- брунейский жестовый язык (Bruneian Sign Language, Bahasa Isyarat Brunei)
- венгерский жестовый язык (Hungarian Sign Language, Magyar Jelvnyelv, Magyar Jelnyelv, HSL)
- датский жестовый язык (Danish Sign Language, Dansk tegnspråk, DSL)
- израильский жестовый язык (Israeli Sign Language, Sfàt ha-simaním ha-israelít, ISL)
- индо-пакистанский жестовый язык (Indo-Pakistani Sign Language, IPSL)
- ирландский жестовый язык (Irish Sign Language, ISL)
- испанский жестовый язык (Spanish Sign Language, Lengua de Signos Española, LSE) (2)
- каталонский жестовый язык (Catalan Sign Language, Llengua de Signes Catalana, LSC)
- китайский жестовый язык (Chinese Sign Language, 中国手语, CSL)

[1] Обзоры словарей жестовых языков. – URL: https://www.surdocentr.ru/publikatsii/obzory-slovarej-zhestovykh-yazykov (дата обращения: 01.12.2021).

- корейский жестовый язык (Korean Sign Language, 手話 □□, KSL)
- кхмерский жестовый язык (Khmer Sign Language, Langue des Signes Khmère, KSL)
- латышский жестовый язык (Latvian Sign Language, Latviešu Zīmju Valoda, LSL)
- литовский жестовый язык (Lithuanian Sign Language, LtSL) (2)
- немецкий жестовый язык (German Sign Language, Deutsche Gebärdensprache, DGS) (2)
- нидерландский жестовый язык (Dutch Sign Language, Nederlandse Gebarentaal, NGT, Sign Language of the Netherlands, SLN) (2)
- новозеландский жестовый язык (New Zealand Sign Language, NZSL)
- польский жестовый язык (Polish Sign Language, Polski Języka Migowy, PJM)
- португальский жестовый язык (Portuguese Sign language, Língua Gestual Portuguesa, LGP)
- североамериканских индейцев жестовый язык (Plains Indian Sign Language, PISL)
- тайваньский жестовый язык (Taiwanese Sign Language, 台灣手語, Taiwan Shouyu, TSL)
- турецкий жестовый язык (Turkish Sign Language, Türk İşaret Dili, TID) (2)
- финский жестовый язык (Finnish Sign Language, Suomalainen viittomakieli, SVK) (3)
- фламандский жестовый язык (Flemish Sign Language, Vlaamse Gebarentaal, Belgian Sign Language, VGT)
- французский жестовый язык (French Sign Language, la langue des signes française, LSF) (3)
- чешский жестовый язык (Czech Sign Language, Český znakový jazyk, CZJ)
- шведский жестовый язык (Swedish Sign Language, Svenskt teckenspråk, SSL) (2)
- швейцарский жестовый язык делится на три диалекта: швейцарский немецкий (Deutschschweizer Gebärdensprache, DSGS), швейцарский французский (Langue des Signes Française, LSF-CH) и швейцарский итальянский (Lingua Italiana dei Segni, LIS-CH)
- шриланкийский жестовый язык (Sri Lankan Sign Language, SLSL)
- японский жестовый язык (Japanese Sign Language, Nihon Shuwa, JSL)

Также на сайте ЛЛЖЯ имеется ссылка на электронную справочно-аналитическую систему *Толковый лексикографический словарь русского жестового языка*[1].

Далее приведем описание некоторых ЛИР жестовых языков, которые не вошли в цитированное обследование.

[1]Электронная справочно-аналитическая система «Толковый лексикографический словарь русского жестового языка». – URL: https://slovar.surdocentr.ru/ (дата обращения: 01.12.2021).

## Мировые жестовые (знаковые) языки

***Указатель словарей жестовых языков***[1]

Наиболее полный каталог всех онлайн-словарей жестовых языков (на французском и английском языках). Цель каталога – индексировать все (бесплатные) онлайн-словари жестовых языков. Имеется вход по странам (68 стран + раздел «Прочее»). Для каждой страны приводится список жестовых языков, применяемых в стране с отсылками к соответствующим ресурсам. Например:

Australie Australian Sign Language (AUSLAN):

1. Auslan SignBank
    - Vidéos

-- Videos

2. Auslan.net (Sign Planet)
    - Dessins, animations 3 D et descriptions

-- Drawings, 3 D animations and descriptions

3. The Auslan Tuition System
    - Logiciel d'apprentissage, dont la version démo gratuite contient 60 phrases signées

-- Learning software, the free demo version contains 60 signed sentences

4. http://www.dictionaryofsign.com
    - <50 signes

-- <50 signs

    - Vidéos

-- Videos

    - Dictionnaire collaboratif

На сайте имеется список знаков (слов и выражений), которые можно искать в разных языках. Также имеется список категорий знаков.

Сайт построен по принципу Вики, т.е. это сайт, на котором каждая страница создается и изменяется ее посетителями. Любой пользователь может просматривать свои страницы, и если у него есть знакомые, которых нет на сайте, он может свободно добавлять их. Это позволяет улучшить и обогатить сайт гораздо быстрее и эффективнее, чем если бы это сделал один человек. На сайте предоставляется возможность пополнять все имеющиеся перечни знаков, категорий знаков и прочее.

***Каталог ресурсов американского языка жестов ASL***[2]

ASL содержит краткие описания и интернет-адреса 35 ресурсов, представляющих американский язык жестов (ASL). Ресурсы разделены на три категории:

---

[1] SL Dictionary directory. – URL: http://lsf.wikisign.org/wiki/ Langue:Signes_du_Monde/English_TOC (дата обращения: 01.12.2021)

[2] American Sign Language Resources. – URL: https://www.fcps.edu/sites/default/files/media/forms/AmericanSignResources.pdf

- ресурсы с печатными изображениями (2)
- ресурсы с видеоклипами или анимацией (19)
- ресурсы ASL для обучения глухих (14)

В описаниях отражаются особенности каждого ресурса: способ организации знаков, порядок просмотра или доступа, область применения (например, математический язык жестов), назначение (например, для дошкольных учреждений), пользовательские возможности, наличие мобильного приложения и другое.

***Жестовый язык – проект Школы лингвистики ВШЭ***[1]

На сайте представлено описание содержания проекта, приведены списки курсовых и дипломных работ, а также докладов на семинарах, выполненных студентами и в рамках этого проекта. Приводятся публикации по проблемам жестового языка, а также учебные курсы. Для онлайн-корпуса русского жестового языка выполняется разметка жестов, которая состоит в установлении соответствий между конкретным жестом и его конкретным значением.

***Онлайн-корпус русского жестового языка***[2]

Аннотированный корпус текстов на русском жестовом языке создан в ходе выполнения работ по проекту «Корпусное исследование морфосинтаксиса и лексики русского жестового языка». Основные единицы корпуса – видеофрагменты, к которым применяется несколько видов разметки: семантическая, морфологическая, орфоэпическая, метатекстовая, социологическая, а также выполняемые вручную разметка жестов и типов речевых актов. Представленные в корпусе тексты на двух разных локальных вариантах русского жестового языка – «сибирском» и «московском» – открывают возможности для изучения не только внутренней структуры и функционирования, но и территориального варьирования этого языка. Кроме того, онлайн-корпус может использоваться в процессе обучения русскому жестовому языку, поскольку дает студентам практический материал, с которым они столкнутся при использовании языка в реальных ситуациях общения.

В онлайн-корпусе реализованы несколько подсистем. Основными из них являются:

- подсистема визуализации материалов, включающая:
  - синтаксический анализатор транскрипций ELAN
  - HTML5 видеоплеер
  - система представления транскрипций
- подсистема управления данными
- подсистема поиска данных по заданным параметрам

[1] Жестовые языки. – URL: https://ling.hse.ru/Projects_SignLang

[2] Корпус русского жестового языка. – URL: http://rsl.nstu.ru/

***Каталог словарей жестового языка***

Данный каталог размещен на информационно-справочном портале Веб-библиотеки для глухих (Web Deaf Library)[1]. Он содержит 19 словарей, в том числе дактильные азбуки – 2, жестовые языки – 17.

## Литература к главе 16

1. Методика сравнительного анализа наиболее известных русских и зарубежных словарей национальных жестовых языков. – URL: https://www.surdocentr.ru/publikatsii/obzory-slovarej-zhestovykh-yazykov/190-metodika-sravnitelnogo-analiza-naibolee-izvestnykh-russkikh-i-zarubezhnykh-slovarej-natsionalnykh-zhestovykh-yazykov (дата обращения: 01.12.2021).

[1]Web Deaf Library. Словари. – URL: http://wdl.ru/directory/dictionary

# ГЛАВА 17. ОБРАЗОВАТЕЛЬНЫЕ ЛИР

## Общие сведения

Применение цифровых ЛИР в образовательных целях, в основном для изучения иностранных языков, представляет собой огромное поле разнообразных ресурсов. Соответственно существует множество обзоров, классификаций, рейтингов, методических материалов по их использованию. Поэтому в настоящей главе будет сделан лишь краткий обзор источников, обеспечивающих первичную ориентацию в этом информационном пространстве, в основном каталогов, рейтингов и рекомендательных сервисов.

Приведем некоторые подходы к классификации электронных образовательных ресурсов (ЭОР). Заметим при этом, что детальная классификация ЭОР, разработанная например в работе [1], включает сотни рубрик по десяткам фасетов. Приведем примеры, заимствованные из работы [2].

С точки зрения организации учебного процесса основными параметрами для оценки являются:

- тип электронного издания (ресурса)
- предметная образовательная область
- рекомендуемый уровень образования
- рекомендуемая форма образовательного процесса
- специфика аудитории

По типу можно выделить следующие основные группы ЭОР:

- компьютерный учебник (учебное пособие, текст лекций и т.д.)
- электронный справочник
- компьютерный задачник
- компьютерный лабораторный практикум (модели, тренажеры и т.д.)
- компьютерная тестирующая система

По формату основной информации выделяются следующие типы ЭОР:

- текстовой – электронное издание, содержащее преимущественно текстовую информацию, представленную в форме, допускающей посимвольную обработку;
- графический – электронное издание, содержащее преимущественно графические сущности, представленные в форме, допускающей просмотр и печатное воспроизведение, но не допускающей посимвольной обработки;

- звуковой – электронное издание, содержащее цифровое представление звуковой информации в форме, допускающей ее прослушивание, но не предназначенной для печатного воспроизведения;
- программный – автономный программный продукт, представляющий собой публикацию текста в некоторой автономной программной среде;
- мультимедийный – электронное издание, в котором информация различной природы присутствует взаимосвязанно для достижения заданных разработчиком дидактических целей.

## Каталоги лингвистических ЭОР

***Единое окно доступа к образовательным ресурсам***[1]

Обзор ресурсов, посвященных лингвистическим ЭОР, лучше начать с профессиональных каталогов. В России это прежде всего *Единое окно доступа к образовательным ресурсам*. Каталог ЭОР в *Едином окне* содержит более 30 тыс. ресурсов, из которых к языкознанию относится свыше 1 тыс. При поиске возможны следующие поисковые фильтры:

Аудитория
- абитуриент
- исследователь
- менеджер
- преподаватель
- учащийся

Тип ресурса
- справочные материалы
- образовательные сайты
- электронные библиотеки
- периодические издания
- программные продукты
- учебные материалы
- учебно-методические материалы
- нормативные документы
- научные материалы
- дополнительные информационные материалы

Уровень образования
- общее
- профессиональное
- дополнительное

***Единая коллекция цифровых образовательных ресурсов***[2]

Еще один официальный каталог ЭОР. Целью создания Коллекции является сосредоточение в одном месте и предоставление доступа к полному

[1] Единое окно доступа к образовательным ресурсам. – URL: http://window.edu.ru/ (дата обращения: 01.12.2021).

[2] Единая коллекция цифровых образовательных ресурсов. – URL: http://school-collection.edu.ru/ (дата обращения: 01.12.2021).

набору современных обучающих средств, предназначенных для преподавания и изучения различных учебных дисциплин в соответствии с федеральным компонентом государственных образовательных стандартов начального общего, основного общего и среднего (полного) общего образования.

В настоящее время в Коллекции размещено более 111 000 цифровых образовательных ресурсов практически по всем предметам базисного учебного плана. В Коллекции представлены наборы цифровых ресурсов к большому количеству учебников, рекомендованных Минобрнауки РФ к использованию в школах России, инновационные учебно-методические разработки, разнообразные тематические и предметные коллекции, а также другие учебные, культурно-просветительские и познавательные материалы.

Поиск в каталоге возможен с помощью следующих поисковых фильтров:

- предмет (русский язык, иностранный язык)
- класс
- назначение (для учителей, для учащихся)

В данном каталоге также имеется сервис формирования тематических подборок ЭОР в виде комплектов учебно-методических ресурсов (комплексных ЭОР) по предметам на основе Федерального базисного учебного плана, примерных программ среднего (полного) общего образования. Комплексные ЭОР строятся как тематические образовательные траектории с возможностью индивидуальных подборок ресурсов по темам учебных дисциплин на базе содержания Единой коллекции.

***Общий каталог учебных электронных ресурсов по РКИ***[1]

Этот каталог предназначен для преподавателей русского языка как иностранного. Он содержит ссылки на существующие электронные ресурсы по РКИ и их краткое описание. Ресурсы созданы специалистами из российских и зарубежных университетов и включают учебные материалы для развития навыков речевой деятельности у учащихся с разным уровнем владения русским языком. Возможен поиск ресурсов по уровню образования и по типу учебных материалов, причем выделяются следующие типы:

- банк учебных материалов
- блог
- видеокурс
- грамматические таблицы
- интерактивные задания
- медиаматериалы
- словарь
- справочник
- страноведение
- тесты
- тренажер
- упражнения

[1] Каталоги Института русского языка и культуры. – URL: https://www.catalogue.irlc.msu.ru/ (дата обращения 01.04.2022).

- учебник
- учебное пособие
- форум

Кроме официальных каталогов ЭОР, создаваемых государственными образовательными учреждениями, существует множество других каталогов. Приведем перечень некоторых наиболее популярных и полных каталогов ЭОР по русскому языку.

- Перечень электронных образовательных ресурсов для учителя русского языка и литературы[1]
- Интернет-ресурсы по русскому языку и литературе[2]
- Образовательные интернет-ресурсы по русскому языку[3]
- Образовательные интернет-ресурсы, направленные на поддержку и продвижение русского языка[4]
- Обзор интернет-ресурсов по предмету «Русский язык»[5]
- Русский язык[6]
- ЭОР для учителя русского языка и литературы[7]
- Интернет-ресурсы по русскому языку и литературе[8]
- Русский язык (интернет-ресурсы)[9]
- ЭОР по русскому языку[10]
- Каталог интернет-ресурсов в помощь учителю русского языка и литературы[11]
- Интернет-ресурсы учителя русского языка и литературы[12]

---

[1] ИНФОУРОК. – URL: https://infourok.ru/perechen-elektronnyh-obrazovatelnyh-resursov-dlya-uchitelya-russkogo-yazyka-i-literatury-4288134.html

[2] Интернет-ресурсы по русскому языку и литературе. – URL:. https://nsportal.ru/user/1016409/page/internet-resursy-uchitelya-russkogo-yazyka-i-literatury (дата обращения: 01.04.2022).

[3] Каталог образовательных программ. – URL: https://www.kop.ru/handbook/v-pomoshch-uchitelyu/obrazovatelnye-internet-resursy-po-russkomu-yazyku/ (дата обращения: 01.12.2021).

[4] Образовательные интернет-ресурсы, направленные на поддержку и продвижение русского языка. – URL: https://nsuem.ru/library/resources/rus-lang-resouces/ (дата обращения: 01.12.2021).

[5] Обзор интернет-ресурсов по предмету Русский язык. – URL: http://yarikov.vspu.ru/home/studentam/materialy-dla-studentov/russkij (дата обращения: 01.12.2021).

[6] Русский язык. – URL: http://edu-top.ru/katalog/?cat=34 (дата обращения: 01.12.2021).

[7] ЭОР для учителя русского языка и литературы. – URL: http://kykyshkina.semenovschool3-nn.edusite.ru/p11aa1.html (дата обращения: 01.12.2021).

[8] Интернет-ресурсы по русскому языку и литературе. – URL: https://ifiyak.sfu-kras.ru/poleznye-ssylki/item/183/ (дата обращения: 01.12.2021).

[9] Русский язык (интернет-ресурсы). – URL: http://yspu.org/%D0%A0%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9 _%D1%8 F%D0%B7%D1%8 B%D0%BA (%D0%B8% D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%82-%D1%80%D0%B5%D1% 81%D1%83%D1%80%D1%81%D1%8 B) (дата обращения: 01.12.2021).

[10] ЭОР по русскому языку. – URL: https://rosuchebnik.ru/material/eor-po-russkomu-yazyku/ (дата обращения: 01.12.2021).

[11]- Каталог интернет-ресурсов в помощь учителю русского языка и литературы. – URL: https://ludmilka.okis.ru/internet-resursi.html (дата обращения: 01.12.2021).

[12]Интернет-ресурсы учителя русского языка и литературы. – URL: https://nsportal.ru/user/1016409/page/internet-resursy-uchitelya-russkogo-yazyka-i-literatury (дата обращения: 01.12.2021).

### Рекомендательные сервисы

В Интернете популярны рекомендательные сервисы, предлагающие выбор лучших, с точки зрения составителя, ресурсов. Такие сервисы распространены как в мире, так и в России. Некоторые рекомендательные сервисы используют рейтинги. Приведем примеры.

- Как выбрать цифровые ресурсы для использования в вашем языковом классе[1]
- 60 сайтов и приложений для изучения иностранных языков[2]
- ТОП 25 сайтов для практики английского[3]
- 10 лучших ресурсов по изучению английской грамматики[4]
- Топ-20 сайтов для изучения иностранных языков[5]
- Восемь лучших цифровых ресурсов для изучающих английский язык[6]

Некоторые ЭОР ориентированы на использование на уроках иностранного языка на разных этапах обучения школьников[7], другие рассчитаны на различные формы самообразования и онлайн-обучения[8].

Обширный каталог инструментов и ресурсов для онлайн-обучения представлен, например, на сайте Университета Южного Квинсленда, Австралия[9]. Там говорится, что интернет-обучение языку можно определить как обучение языку, проводимое в Интернете с использованием интернет-инструментов и ресурсов. Оно имеет две основные формы: компьютерно-опосредованная коммуникация (CMC) и веб-обучение языку (WBLL). В обоих случаях онлайн-инструменты играют ключевую роль в реализации CMC и WBLL.

Список онлайн-инструментов для обучения языку (WBLL) включает двенадцать категорий:

---

[1] How to select digital resources for use in your language classroom / Pete Sharma. – URL: https://www.cambridge.org/elt/blog/2020/06/29/how-to-select-digital-resources-for-use-in-your-language-classroom /

[2] 60 сайтов и приложений для изучения иностранных языков. – URL: https://blog.mann-ivanov-ferber.ru/2019/02/21/60-sajtov-i-prilozhenij-dlya-izucheniya-inostrannyx-yazykov/ (дата обращения: 01.12.2021).

[3] ТОП 25 сайтов для практики английского: подробно о каждом. – URL: https://www.englishdom.com/blog/top-25-sajtov-dlya-praktiki-anglijskogo-podrobno-o-kazhdom/ (дата обращения: 01.12.2021).

[4] 10 лучших ресурсов по изучению английской грамматики. – URL: https://proglib.io/p/eng-grammar-free (дата обращения: 01.12.2021).

[5] Топ-20 сайтов для изучения иностранных языков. – URL: https://top100lingua.ru/blog/uroki/top-20-sajtov-dlja-izuchenija-inostrannyh-jazykov (дата обращения: 01.12.2021).

[6] Eight excellent digital resources for English language learners. – URL: https://www.ef.com/wwen/blog/teacherzone/self-study-resources-for-students/ (дата обращения: 01.04.2022).

[7] Использование ЭОР на уроках английского языка. – URL: http://ext.spb.ru/index.php/11530 (дата обращения: 01.12.2021).

[8] Online Teaching Tools and Resources. – URL: https://cls.yale.edu/faculty/resources/online-teaching-tools-and-resources (дата обращения: 01.12.2021).

[9] Online Tools for Language Teaching. – URL: http://www.tesl-ej.org/wordpress/issues/volume15/ej57/ej57 int/ (дата обращения: 01.12.2021).

- системы обучения / управления контентом
- коммуникация
- живые и виртуальные миры
- социальные сети и закладки
- блоги и вики
- презентация
- совместное использование ресурсов
- создание веб-сайтов
- создание веб-упражнений
- веб-поисковые системы
- словари и конкордансы
- утилиты

Еще один каталог называется *Выучить язык онлайн*[1]. Это большой список веб-сайтов, посвященных изучению языка онлайн – в основном бесплатно. Некоторые сайты предоставляют контент; другие предлагают войти в сообщества; у ряда сайтов есть и то, и другое. Список включает 72 ресурса для многоязычного обучения (предлагающих более двух языков), а также несколько сотен сайтов для африкаанс, арабского, китайского, хорватского, датского, английского, эсперанто, французского, немецкого, иврита, хинди, исландского, ирландского, гэльского, итальянского, греческого, японского, португальского, русского, испанского, тагальского (филиппинского), урду и валлийского. Например, для английского языка предлагается 32 ресурса, для русского – 10 и т.д.

Естественно, существуют специальные ЭОР для продолжающегося обучения[2] и для высшего образования. Примером анализа последней категории может служить работа [3].

Практически все ведущие университеты мира составляют и предоставляют пользователям каталоги и рекомендательные списки онлайн-ресурсов для изучения иностранных языков. Выше мы привели пример каталога Университета Южного Квинсленда, Австралия. Подобные каталоги можно найти на сайтах университетов Лос-Анжелеса, Манчестера, Миннесоты, Йеля и многих других.

Среди разнообразных ресурсов, направленных на изучение иностранных языков, несколько особняком стоит несомненно образовательный ресурс *Языкознание.ру*[3], созданный для изучающих различные лингвистические дисциплины. Информация, представленная на сайте, имеет прежде всего справочный характер. Данная информация может быть полезна не только студентам-лингвистам, но и преподавателям лингвистики. Особенностью данного сайта является структурирование информации не по уровням языка,

[1] Centre for Learning & Performance Technologies. – URL: https://c4lpt.co.uk/directory-of-learning-performance-tools/learn-a-language-online/ (дата обращения: 01.12.2021).

[2] Linguistics for Continuing Education: Free online resources. – URL: https://www.semanticscholar.org/paper/Oxford-LibGuides%3A-Linguistics-for-Continuing-Free-Wilkin/27d85fbc200f994529b0282c10b415f995122448 (дата обращения: 01.04.2022).

[3] Языкознание. ру. – URL: http://yazykoznanie.ru/ (дата обращения: 01.12.2021).

а по изучаемым дисциплинам на лингвистических специальностях в вузах России. Таким образом, сайт может помочь студентам в подготовке к экзаменам и семинарским занятиям, а также преподавателям, когда нужно быстро ознакомиться с дисциплиной, составить план лекции. На сайте в настоящее время представлены все самые основные предметы специальности «Теоретическая и прикладная лингвистика», с которыми у студентов могут возникнуть сложности.

Следует упомянуть ресурсы, на которых предлагаются аналитические системы для выбора программных продуктов для изучения иностранных языков. Это, например, Language Learning Software[1] или Language Software Reviews[2]. На этом последнем можно выбрать аналитические обзоры для разных языков, разных авторов, использовать существующие рейтинги, а также проголосовать за определенную программу.

Например, для изучения русского языка сравниваются 5 программ с указанием цены, рейтинга и скорости освоения языка:

- Transparent Russian
- Pimsleur Russian
- Rosetta Stone Russian
- Berlitz Russian
- BYKI Russian

## Литература к главе 17

1. Башмаков А.И., Старых В.А. Систематизация информационных ресурсов для сферы образования: классификация и метаданные. – Москва : Европейский центр по качеству, 2003. – 383 с. : ил., табл.

2. Есенина Н.Е. Обзор электронных образовательных и информационных ресурсов для обучения иностранному языку // Информатика и образование. – 2012. – № 12. – С. 103–105.

3. Salimzanova D.A., Gilfanova G.T. Electronic (Digital) Educational Resource as a Tool of Teaching a Foreign Language in the System of Higher Education // Conference: “New Silk Road: Business Cooperation and Prospective of Economic Development” (NSRBCPED 2019). – 2020. – January. – DOI: 10.2991/aebmr.k. 200324.035

[1] Language Learning Software. – URL: https://www.languagelearningsoftware.com/ (дата обращения: 01.12.2021).

[2] Language Software Reviews. – URL: https://www.languagesoftware.net/ (дата обращения: 01.12.2021).

# ГЛАВА 18. РЕСУРСЫ ПО РУССКОМУ ЯЗЫКУ В ЗАРУБЕЖНЫХ СОБРАНИЯХ

Общее исследование ЛИР, как оно представлено в настоящей книге, очевидно должно включать анализ состояния ЛИР по русскому языку. Для российской прикладной лингвистики эти ЛИР являются магистральным направлением, хотя бы потому, что наиболее востребованные языковые технологии в нашей стране связаны с обработкой русских письменных текстов и устной речи. Очевидно, что самыми востребованными являются и образовательные ресурсы по русскому языку.

В предыдущих главах, посвященных различным классам ЛИР, мы старались показать состояние российских ЛИР, в том числе, конечно, и относящихся к русскому языку.

В данной главе мы предложим краткий обзор ЛИР по русскому языку, представленных в различных зарубежных собраниях.

В этот обзор, в отличие от других глав, где рассматривались только специальные ЛИР, мы включили также сведения о тематических ЛИР по русскому языку, прежде всего перечень сайтов зарубежных центров русистики. Этот перечень представлен в приложении 9. Интерес к состоянию зарубежной русистики обусловлен, в частности, проектом создания комплексной академической информационно-справочной системы по русистике, который обсуждается в настоящее время в нескольких институтах РАН.

Приложение 9 содержит сведения о свыше 300 центрах изучения русского языка и культуры в различных странах. Оно состоит из четырех разделов:

- Центры русистики в США и Канаде – 165 наименований и адресов
- Центры русистики в Европе – 99 наименований
- Центры русистики в других регионах – 20 наименований
- Центры русистики без аффилиации и каталоги образовательных ресурсов по русскому языку – 23 наименования

Для полноты картины по центрам русистики приведем ссылки на перечни зарубежных филиалов российских организаций, также занимающихся продвижением русского языка. Это зарубежные представительства Россотрудничества[1]

[1] Зарубежные представительства Россотрудничества. – URL. https://rwp.agency/russkie-doma/ (дата обращения: 01.04.2022).

(в настоящее время 91 представительство в 78 странах) и русские центры фонда «Русский мир»[1] (центры в 52 странах).

Анализ собственно ЛИР по русскому языку проводился на основе поиска в основных зарубежных собраниях и системах поиска ЛИР, которые мы рассматривали в главе 2.

***OLAC***

Крупнейшим собранием ЛИР, как уже говорилось, является собрание архивов *OLAC*, поисковая система которого доступна по адресу[2]. При поиске по запросу *Русский язык* в фасете *Язык* выдается 5 440 наименований. В этой системе возможно получить сведения, в каких архивах хранятся найденные ЛИР. Список этих архивов приводятся ниже, в скобках указывается число ЛИР в данном архиве:

- Endangered Languages Archive (4849)
- The Language Archive (282)
- Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (149)
- Graduate Institute of Applied Linguistics Library (106)
- WALS Online RefDB (23)
- CHILDES Data repository (6)
- The Rosetta Project: A Long Now Foundation Library of Human Language (5)
- Slovenian language resource repository CLARIN.SI (4)
- LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University (2)
- The Crúbadán Project (2)
- AfBo: A world-wide survey of affix borrowing (1)
- COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO) (1)
- Ethnologue: Languages of the World (1)
- Glottolog 4.3 (1)
- LAPSyD (1)
- PHOIBLE 2.0 (1)
- SAILS Online (1)
- SIL Language and Culture Archives (1)
- The LINGUIST List Language Resources (1)
- WALS Online (1)
- WOLD (1)

Видно, что львиная доля ЛИР, отнесенных к русскому языку (почти 90%), находятся в архиве исчезающих языков. Как выяснилось – это как раз

[1] Русские центры фонда «Русский Мир». – URL: https://russkiymir.ru/rucenter (дата обращения: 01.04.2022).

[2] OLAC Language Resource Catalog. Search for language resources. – URL: http://dla.library.upenn.edu/dla/olac/index.html (дата обращения: 01.12.2021).

файлы исчезающих языков, а на русском языке представлены аннотации, переводы и сопровождающая информация.

Кроме того, в OLAC можно получить сведения, к какому типу ЛИР относятся найденные ЛИР по русскому языку. К сожалению, лишь небольшая часть найденных ЛИР заиндексирована по этому признаку. поэтому OLAC выдает такие данные:

- Language description (34)
- Lexicon (32)
- Primary text (9)

В OLAC можно также получит распределение, к каким областям лингвистики относятся найденные ЛИР по русскому языку. Вот это распределение:

- Computational linguistics (2)
- General linguistics (24)
- Language acquisition (6)
- Language documentation (150)
- Lexicography (2)
- Morphology (7)
- Phonetics (11)
- Phonology (8)
- Psycholinguistics (1)
- Semantics (10)
- Sociolinguistics (3)
- Syntax (20)
- Text and corpus linguistics (2)
- Typology (25)
- Writing systems (4)

***Виртуальная языковая обсерватория VLO***[1]

Всего результатов по запросу *Русский язык* – 6,618 ЛИР. Как и в предыдущем случае, большинство ЛИР – это тексты или аудиофайлы различных исчезающих языков, где на русском языке сопровождающий текст или перевод.

Распределение ЛИР по форме:

- местные (985)
- монологи (118)
- письменный (28)
- транскрибированные (2)

Представляет также интерес распределение ЛИР по русскому языку по типам (приводим 5 типов):

- текст (238)
- аудио (182)
- корпус (98)
- лексический ресурс (81)

[1]Virtual Language Observatory. – URL: https://vlo.clarin.eu/search?2&q=russian (дата обращения: 01.12.2021).

- наборы данных (17)

Параллельных ЛИР, включающих русский язык, найдено 17.

***Языковой архив Института психолингвистики Общества Макса Планка***[1]

*Архив TLA*, крупнейший в Европе, содержит 287 ЛИР, найденных по запросу *Русский язык*. Приведем пример описания ЛИР из этого архива.

*Name* Bochilikan Etiken
*Title* Stepanova_ZA_Bochilikan_Etiken
*Collection* DoBeS archive : Even
*Contributor* Brigitte Pakendorf
*Language* Even, Russian
*Country* Russian Federation
*Genre* Discourse
*Format* video/x-mpeg2, image/jpe

Покажем некоторые распределения русскоязычных ЛИР в данном архиве.

По жанрам:

- Unspecified (136)-
- Discourse (110)
- Stimuli (17)
- Fiction (7)
- Personal notes (5)
- Ethnografical recording (3)-
- Singing (3)-
- Poetry (2)-
- Biography (1)
- Dictionary (1)

По форматам:

- audio/x-wav (227)+ -
- text/x-eaf+xml (116)+ -
- video/x-mpeg2 (61)+ -
- application/pdf (60)+ -
- text/plain (54)+ -
- text/x-pfsx+xml (47)+ -
- text/html (31)+ -
- text/x-chat (31)+ -
- text/x-trs (28)+ -
- video/x-mpeg1 (23)

***META-SHARE***[2]

[1] The Language Archive. – URL: https://archive.mpi.nl/tla/islandora/search /%2 A%3 A% 2 A?f%5 B0%5 D=cmd.Language%3 A%22 Russian%22 (дата обращения: 01.12.2021).

[2] META-SHARE. URL: – http://www.meta-share.org/ (дата обращения: 01.04.2022).

Одним из наиболее развитых европейских языковых архивов является *META-SHARE*. В нем представлено 118 ЛИР, выдаваемых по запросу *Русский язык* по фасету *Язык*. Приведем распределение этих ЛИР по разным основаниям (в основном приводятся верхние уровни).

По назначению ЛИР:
- лексические и концептуальные ресурсы (61)
- корпуса (55)
- инструменты и сервисы (2)

По семиотическому типу;
- текст (111)
- аудио (13)
- видео (3)
- изображение (1)

По количеству языков:
- многоязычные (75)
- моноязычные (23)
- двуязычные (20)

По способу многоязычности:
- параллельные (26)
- другие (8)
- один текст на разных языках (3)
- сопоставления (2)

По модальности:
- письменный язык (74)
- разговорный язык (7)
- голос (7)
- жестовый язык (3)
- выражение лица (2)

По типу MIME:
- Text/plain (2)
- Text/xml (2)
- Plain text (1)
- Application/x-tmx+xml (1)
- Audio/wav (1)

Соответствие стандартам и лучшим практикам
- TBX (48)
- TEI (2)
- TEI_P5 (2)
- TMX (2)

Доступность:
- доступно с ограничениями (65)
- доступно без ограничений (52)
- в стадии переговоров (1)

Предусмотренное использование:
- приложение NLP (61)
- ручная обработка (59)

Способ использования для NLP:
- информационный поиск (48)
- машинный перевод (10)
- лингвистические исследования (7)
- идентификация языка (1)

Область знаний:
- естественные прикладные науки (14)
- техника (6)
- прочие отрасли (5)
- экономика (4)
- информатика и обработка данных (4)

Приведем обобщенные сведения по русскоязычным ЛИР еще в некоторых зарубежных каталогах.

***Каталог ELRA***[1]. Всего 20 ЛИР по русскому языку, в том числе:
- корпуса текстов (3)
- устные корпуса (12)
- словарные БД (2)
- тестовые (оценочные) массивы (3)

***Каталог LDC***[2]. Всего 22 ЛИР по русскому языку, в том числе:
- параллельные письменные корпуса (3)
- многоязычные устные корпуса (1)
- корпус телефонных разговоров (1)
- тестовые оценочные ЛИР (13)

***The Linguee App. English-Russian Dictionary***[3] содержит 1 млрд англо-русских и русско-английских соответствий.

В заключение и в качестве примера приведем список ЛИР, включающих русский язык, из поисковой системе ***Linghub***[4]. Каждый ЛИР описан по следующей схеме:
- Адрес (Contact Point)
- Описание (Description)
- Язык(и) (Language)
- Права доступа (Rights)
- Источник (Source)
- Наименование (Title)
- Тип ЛИР (Type)

1. EuroTermBank.
2. PIARC Multilingual Terminology Database.
3. Latvian-Russian Personal Names Glossary.

[1] ELRA. – URL: http://www.elra.info/en/catalogues/catalogue-language-resources/ (дата обращения: 01.04.2022).

[2] LDC. – URL: https://www.ldc.upenn.edu/language-resources (дата обращения: 01.12.2021).

[3] The Linguee App. – URL: https://www.linguee.com/ (дата обращения: 01.12.2021).

[4] Linghub. – URL: http://linghub.org (дата обращения: 01.12.2021).

4. Latvian-Russian Geographical Names Glossary.
5. German and Russian gold standard for knowledge-rich context extraction.
6. Khanty Corpus (North Khanty, Corpora and Translations) (UHLCS).
7. Bilingual term pairs extracted from comparable Web resources using the TaaS Bilingual Term Extraction System.
8. Nganasan Corpus.
9. DIALUKI – Diagnosing reading and writing in a second or foreign language.
10. Names of Countries in Seven Languages.
11. Finnish Government Termbank Valter.
12. MultiJur: Multilingual Parallel Corpus of Legal Texts.
13. Corpus of Early Modern Finnish.
14. ProoF – Pronunciation of Finnish by Immigrants in Finland\n.
15. Multilingual Resource Collection of the University of Helsinki Language Corpus Server.
16. The Helsinki Annotated Corpus of Russian Texts (HANCO) Database.
17. MULCOLD.
18. Nenets Corpus (Tundra Nenets) (UHLCS).
19. ParRus: Russian-Finnish Parallel Corpus of Literary Texts.
20. The National Certificates Corpus.
21. FiRuLex: Finnish-Russian Comparable Corpus of Legal Texts.
22. FiRuLex: Finnish-Russian Comparable Corpus of Legal Texts.
23. ParFin: Finnish-Russian Parallel Corpus of Literary Texts.
24. The Corpus of Ingrian Finnish.
25. AddictionLink in Finnish Sign Language.
26. PELCRA mutlilingual parallel corpora (CC-BY).
27. Bibliography of Bulgarian Lexicology, Phraseology and Lexicography.
28. Polish-Russian Parallel Corpus.
29. Bulgarian-X language Parallel Corpus.
30. Replication Data for: Ditransitive constructions in Russian and Ukrainian.
31. Replication Data for: Rival forms of Comparatives in Russian.

Итак, можно констатировать, что в зарубежных архивах, репозиториях и других собраниях ЛИР представлено достаточно много ЛИР, либо русскоязычных, либо многоязычных, включающих русский в качестве одного из языков. Однако по сравнению с ЛИР по русскому языку, разработанных в России, зарубежные собрания достаточно бедны. Более или менее представлены ЛИР в нескольких архивах по исчезающим языкам, где русский язык используется для аннотаций или перевода, а также в образовательных ресурсах.

В то же время качественные академические ЛИР по русскому языку представлены в зарубежных собраниях совершенно недостаточно. Это означает, что российские разработчики ЛИР должны озаботиться участием в современных международных проектах. В частности, необходимо депонировать свои ресурсы в различных репозиториях типа CLARIN, тем более что в большинстве случаев это бесплатно. Это будет способствовать продвижению

научных результатов российской прикладной лингвистики, и в конечном счете повышению статуса российской науки.

# ЧАСТЬ 4
# ПЕРСПЕКТИВЫ РАЗВИТИЯ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

## ГЛАВА 19. ЛИНГВИСТИЧЕСКИЕ СВЯЗАННЫЕ ОТКРЫТЫЕ ДАННЫЕ (LLOD)

### Общие сведения

Наиболее перспективным направлением международной коллаборации в области ЛИР является, по мнению многих зарубежных специалистов а также автора, проект *Лингвистические связанные открытые данные LLOD*[1]. Это направление опирается на идеологию и технологию Семантической сети, разрабатываемой многими исследовательскими группами под общим руководством Фонда открытого знания[2], председателем которого является создатель Всемирной паутины Тим Бернерс-Ли.

Краткое описание проекта представлено по адресу[3]. Наиболее полное описание современного состояния LLOD имеется в работе [1].

Реализацию проекта LLOD осуществляет международная рабочая группа OWLG[4], созданная в рамках деятельности Фонда открытого знания.

Проект LLOD представляет собой международную коллаборацию по созданию, обеспечению совместимости и повторному использованию ЛИР в соответствии с принципами связанных открытых данных:

- данные должны быть открытыми и лицензированы с использованием таких лицензий, как лицензии Creative Commons;
- элементы в наборе данных должны быть однозначно идентифицированы с помощью универсального идентификатора ресурса (URI);
- URI должен разрешаться, чтобы пользователи могли получить доступ к дополнительной информации с помощью веб-браузеров;
- разрешение ресурса LLOD должно возвращать результаты с использованием веб-стандартов, таких как Resource Description Framework (RDF);
- ссылки на другие ресурсы должны быть включены, чтобы помочь пользователям открывать новые ресурсы и обеспечивать точность семантики запросов и других транзакций.

---

[1] Linguistic Linked Open Data. – URL: https://linguistic-lod.org/llod-cloud (дата обращения: 01.12.2021).

[2] The Open Knowledge Foundation. – URL: https://okfn.org/ (дата обращения: 01.12.2021).

[3] Linguistic_Linked_Open_Data. – URL: https://ru.qaz.wiki/wiki/Linguistic_Linked_Open_Data (дата обращения: 01.12.2021).

[4] The-Open-Linguistics-Working-Group. – URL: https://linguistics.okfn.org/index-42.html (дата обращения: 01.04.2022).

Реализация LLOD обеспечивает следующие основные преимущества:

- *представление*: связанные графы – более гибкий формат представления лингвистических данных;
- *совместимость*: общие модели RDF могут быть легко интегрированы;
- *федеративность*: данные из нескольких источников можно легко объединить;
- *открытость*: инструменты для RDF и связанных данных широко доступны по лицензиям с открытым исходным кодом;
- *выразительность*: существующие словари помогают представить все необходимые ЛИР;
- *семантика*: ссылки точно определяют понятия;
- *динамичность*: веб-данные можно постоянно улучшать.

## Облако LLOD

### *Стандарты LLOD*

Помимо сбора метаданных и создания облачной диаграммы LLOD, сообщество OWLG стимулирует разработку стандартов сообщества в отношении словарей, метаданных и рекомендаций передового опыта. Согласно цитируемому выше обзору Джимиано и соавторов, к ним относятся:

- для моделирования лексических ресурсов – OntoLex-Lemon, стандарт сообщества для лексических ресурсов (машиночитаемые словари, многоязычная терминология, лексикализация онтологий);
- для моделирования лингвистических аннотаций (в корпусах или для NLP):
  - Web Annotation – стандарт W3 C для аннотации веб-ресурсов (текстовых или иных);
  - Формат обмена NIF – стандарт сообщества для грамматической аннотации текста;
  - CoNLL-RDF – словарь на основе NIF для представления RDF корпусов в традиционных форматах («CoNLL»);
  - POWLA – словарь для общих лингвистических структур данных, которые можно использовать для дополнения NIF, CoNLL-RDF или Web Annotation;
- для категорий лингвистических данных:
  - Онтологии лингвистической аннотации (OLiA);
  - LEXINFO для грамматических и других функций в лексических ресурсах;
- для идентификация языка:
  - в виде строк с языковыми тегами IETF BCP 47;
  - ISO 639–3 (URI, предоставленными lexvo.org);
  - Glottolog (URI для языковых разновидностей, не охваченных ISO 639);
- для метаданных:
  - Dublin Core;

- Словарь каталогов данных (DCAT), стандарт W3 C для каталогов данных, опубликованных в Интернете;
- METASHARE-OWL, словарь для метаданных языковых ресурсов.

***Избранные ресурсы***

Облако LLOD развивается достаточно быстро, и в нем уже размещены сотни ЛИР. Ниже приводится список из десяти ЛИР с наибольшим количеством связей, представленных в LLOD по состоянию на октябрь 2018 года (в порядке количества связанных наборов данных):

- Онтологии лингвистической аннотации (*OLiA*, связанные с 74 наборами данных) предоставляют справочную терминологию для лингвистических аннотаций и грамматических метаданных;
- *WordNet* (связан с 51 набором данных) – лексическая база данных для английского языка и сводная база для разработки аналогичных баз данных для других языков, в нескольких редакциях (редакция Princeton связана с 36 наборами данных; редакция W3 C связана с восемью наборами данных; версия VU связана с семью наборами данных);
- *DBpedia* (связан с 50 наборами данных) – многоязычная база знаний общего мира, основанная на Википедии;
- Онтология *LEXINFO* (связан с 36 наборами данных) предоставляет справочную терминологию для лексических ресурсов;
- *BabelNet* (связан с 33 наборами данных) – многоязычная лексикализованная семантическая сеть, основанная на агрегировании различных других ресурсов, в первую очередь WordNet и Wikipedia;
- *LEXVO* (связан с 26 наборами данных) предоставляет идентификаторы языка и другие данные, связанные с языком. LEXVO обеспечивает представление RDF ISO 639–3 трехбуквенных кодов для идентификаторов языков и информации об этих языках;
- *ISO 12620 Реестр категорий данных* (ISOcat; версия RDF, связанная с 10 наборами данных) предоставляет частично структурированный репозиторий для различной терминологии, связанной с языком. Хостинг ISOcat находится в The Language Archive, соответственно в проекте DOBES, в Институте психолингвистики им. Макса Планка, но в настоящее время осуществляется переход на CLARIN;
- *UBY* (RDF версия, связан с девятью наборами данных), лексическая сеть для английского языка, собранная из различных лексических ресурсов;
- *Glottolog* (связан с семью наборами данных) предоставляет детализированные идентификаторы языков для языков с низким уровнем ресурсов, в частности многие из них не охвачены lexvo.org;
- *Викисловарь* – ссылки на DBpedia (wiktionary.dbpedia.org, связанный с семью наборами данных), лексикализация концепций DBpedia на основе Викисловаря.

***Приложения LLOD***

LLOD применяется для решения ряда проблем научных исследований:

- во всех областях прикладной лингвистики, компьютерной филологии и обработки естественного языка; лингвистическая аннотация и лингвистическая разметка представляют собой центральные элементы анализа. Однако прогрессу в этой области препятствуют проблемы совместимости, в первую очередь различия в словарях и схемах аннотаций, используемых для разных ресурсов и инструментов. Использование связанных данных для соединения языковых ресурсов и онтологий / терминологии облегчает повторное использование общих словарей;
- в корпусной лингвистике перекрывающаяся разметка представляет собой известную проблему для обычных форматов XML. Модели данных на основе графов были предложены с конца 1990-х годов. Они традиционно представлены в виде множества взаимосвязанных файлов XML, которые плохо поддерживаются стандартной технологией XML. Моделирование таких сложных аннотаций, как связанные данные, представляет собой формализм, семантически эквивалентный XML, но устраняет необходимость в специальной технологии и вместо этого опирается на существующую экосистему RDF;
- многоязычные проблемы, включая связывание ЛИР, таких как WordNet, как это выполнено в Межъязыковом указателе WordNet, и взаимосвязанные разнородные ресурсы, такие как WordNet и Wikipedia, как это было сделано в BabelNet;
- информационное обеспечение стандартизации лингвистических ресурсов.

LLOD тесно связаны с разработкой:

- передовых методов связывания лексических данных в Интернете;
- лучших методов создания аннотаций (например, с использованием стандарта Web Annotation);
- лучших практик моделирования и совместного использования текстовых ресурсов с перекрывающейся разметкой.

***Классификация лингвистических данных***

При формировании LLOD существенным был вопрос о релевантности включаемых в LLOD понятию ЛИР. Для решения этой задачи были разработаны критерии «лингвистической релевантности». Следствием этого подхода явилась классификация лингвистически релевантных наборов данных (которые мы в главе 1 назвали *специальными* ЛИР). OWLG разработала следующую классификацию ЛИР, размещаемых в LLOD:

- корпус: лингвистически проанализированный набор языковых данных
- лексиконы:
  - лексико-концептуальные данные
  - лексические ресурсы: лексиконы и словари
  - базы терминов: терминология, тезаурусы и базы знаний
- метаданные:
  - метаданные лингвистических ресурсов (метаданные о языковых ресурсах, включая цифровые языковые ресурсы и печатные книги)

- категории лингвистических данных (метаданные о лингвистической терминологии, включая лингвистические категории, языковые идентификаторы)
  - типологические базы данных (метаданные об отдельных языках, лингвистических особенностях этих языков)
  - другое (множество ресурсов, которые (пока) не классифицированы)

В этой классификации терминологические базы находятся на грани лингвистической релевантности, поскольку они обычно создаются для целей, отличных от языковых технологий или лингвистических исследований.

***Открытые данные и доступность***

LLOD должны соответствовать лицензиям в соответствии с определением открытых данных. Однако для генерации облака LLOD это требование выполняется не всегда, поэтому техническим критерием является доступность через Интернет и наличие метаданных. В OWLG неоднократно обсуждалось, могут ли быть включены некоммерческие (академические) ресурсы. В 2015 году было достигнуто согласие принять их, но с последующим введением более строгих требований вместе с ростом облака LLOD. По состоянию на январь 2020 года машиночитаемые метаданные лицензий были доступны для 86 ресурсов LLOD, из них 82 приняли открытые лицензии, четыре приняли некоммерческие лицензии.

***Трактовка понятия LLOD***

Аббревиатура «LLOD» может использоваться для обозначения либо технологии LLOD (использование связанных данных независимо от их правового статуса), либо ресурсов LLOD (открытых данных). Для устранения неоднозначности могут использоваться термины «ресурсы LLOD» и «технология LLOD». Чтобы подчеркнуть применение или применимость к закрытым ресурсам, также использовались «LLD» (лингвистические связанные данные). Возможный компромисс – это аббревиатура технологии LL (O) D. Однако «лицензированных лингвистических связанных данных» в облаке LLOD в июне 2020 года не существовало.

***Форматы связанных данных***

Для реализации связанных данных требуется приложение RDF или соответствующие стандарты. Сюда входят рекомендации W3 C: SPARQL, Turtle, JSON-LD, RDF-XML, RDFa и т.д. Однако в языковых технологиях и языковых науках в настоящее время более популярны другие формализмы, поэтому вопрос о включении таких данных в облако LLOD возникает регулярно. Для нескольких таких языков существуют стандартизированные W3 C механизмы упаковки (например для XML, CSV или реляционных баз данных), и такие данные могут быть интегрированы при условии, что соответствующее отображение предоставляется вместе с исходными данными.

### Семинар по LLOD (LDL)

С момента своего создания в 2012 году серия семинаров *Linked Data in Linguistics (LDL)* стала основным форумом для представления, обсуждения и распространения технологий, словарей, ресурсов и опыта применения семантических технологий и парадигмы *Linked Open Data (LOD*) к ЛИР для облегчения их видимости, доступности, взаимодействия, возможности повторного использования, обогащения, комбинированной оценки и интеграции.

Серия семинаров LDL организована OWLG, а труды 2-го, 4-го и 7-го семинаров представлены в публикации [2], труды 5-го семинара в [3], а 6-го семинара – в [4].

Семинары LDL способствуют обсуждению, распространению и установлению стандартов сообщества, в первую очередь модель Lemon/OntoLex для лексических ресурсов, а также стандарты для других типов языковых ресурсов, которые все еще находятся в стадии разработки.

Большинство семинаров носит тематический характер. Например, на 2-м семинаре обсуждались следующие темы:

- примеры использования для создания, ведения и публикации коллекций лингвистических данных, связанных с другими ресурсами;
- моделирование лингвистических данных и метаданных с помощью OWL и / или RDF;
- онтологии для коллекций лингвистических данных и метаданных;
- применение других онтологий или связанных данных из любой субдисциплины лингвистики;
- описания наборов данных, в идеале следующие принципам связанных данных;
- правовые и социальные аспекты лингвистически связанных открытых данных.

На 4-м семинаре обсуждалось применение парадигмы связанных открытых данных к лингвистическим данным в различных областях лингвистики, обработки естественного языка, управления знаниями и информационные технологии, принципы, тематические исследования и лучшие практики представления, публикации и связывания моноязычных и многоязычных коллекций лингвистических данных и знаний, включая корпусы, грамматики, словари, словосочетания, память переводов, предметно-ориентированные онтологии и т.д.

5-й семинар был посвящен созданию, использованию и управлению связанных ЛИР. В центре обсуждения были вопросы применения парадигмы связанных открытых данных к лингвистическим данным, поскольку это может стать важным шагом на пути к тому, чтобы сделать лингвистические данные:

- легко и единообразно запрашиваемыми;
- совместимыми;
- общими с использованием открытых стандартов, таких как протокол HTTP и модель данных RDF.

Хотя было показано, что связанные данные имеют большое значение для управления ЛИР в Сети, эта практика все еще далека от общепринятого стандарта. Таким образом, важно, чтобы продолжалась разработка и внедрение технологий связанных данных среди создателей ЛИР. В частности, способность связанных данных повышать качество, совместимость и доступность данных в Сети побудила OWLG сосредоточить внимание на управлении, улучшении и использовании там ЛИР – в качестве ключевого направления семинара.

6-й семинар по связанным данным (LDL-2018), проведенный совместно с LREC 2018, был посвящен формированию лингвистической науки о данных, т.е. исследовательским методам и приложениям, основанным на LLOD и существующих технологиях и интеграции ЛИР для лингвистических исследований, NLP и цифровой гуманитаристики**.**

Наконец, на последнем, 7-м семинаре в 2020 году, также совместно с LREC 2020, рассматривались вопросы создания инструментов и инфраструктуры для LLOD.

В последние годы наблюдается рост интереса к применению графов знаний и технологий Семантической сети к ЛИР и их публикации в виде связанных данных в Сети. На сегодняшний день большое количество ЛИР либо преобразовано, либо создано изначально как связанные данные на основе моделей данных, специально разработанных для представления лингвистического контента. Примерами являются сети слов, словари, корпуса – исследовательские работы, описывающие создание этих ресурсов, были представлены в предыдущих изданиях LREC и LDL.

Однако несмотря на то, что критическая масса LLOD уже существует, по-прежнему существует острая потребность в надежной экосистеме инструментов, которые работают с LLOD. Недавно начатые исследовательские сети и европейские проекты, такие как *Nexus Linguarum*, *ELEXIS и Prêt-à-LLOD*, направлены на создание устойчивых инфраструктур вокруг ЛИР с использованием LLOD в качестве одной из основных технологий. Этим проектам, прежде всего проекту *Prêt-à-LLOD,* был посвящен один из центральных докладов на семинаре [5], который мы кратко изложим далее.

## Проекты по развитию LLOD

***Кросс-лингвистические связанные данные CLLD***[1]

Проект CLLD координирует более десятка лингвистических баз данных, охватывающих языки мира. Он проводится в отделении лингвистической и культурной эволюции Института истории человечества Общества Макса Планка в Йене, Тюрингия, Германия.

Цель проекта CLLD – разработать и сопровождать методы совместимости лингвистических данных с использованием принципов LOD в качестве механизма интеграции для распределенных ресурсов.

[1] Cross-Linguistic Linked Data (CLLD). – URL: https://clld.org/(дата обращения: 01.12.2021).

Этот подход позволяет с минимальными затратами публиковать отдельные ЛИР, такие как WALS (Всемирный атлас языковых структур) или WOLD (Всемирная база данных заимствованных слов), сохраняя бренды этих проектов и в то же время обеспечивая унифицированный пользовательский интерфейс для всех ЛИР.

CLLD включает ЛИР, уже скомпилированные в MPI-EVA и других местах. Это привело к созданию программной среды, которую можно использовать для разработки отредактированных коллекций баз данных, представленных лингвистами со всего мира.

Описание методологии CLLD можно найти в работе [3]. Список баз данных, реализованных как приложения CLLD и опубликованных на платформе CLLD, доступен по ссылке[1].

*Dictionaria*[2] – журнал для размещения электронных словарей малоизучаемых языков, который работает на платформе CLLD, уже опубликовал 10 словарей.

Для целей однозначной привязки лингвистических данных к языкам и каждой разновидности проект CLLD также включает *Glottolog* (каталог всех языков, семейств и диалектов с исчерпывающей справочной информацией).

***Формат кросс-лингвистических данных CLDF***[3]

Возможно, наиболее важным результатом проекта CLLD была спецификация стандарта Cross-Linguistic Data Formats (CLDF). CLDF предоставляет стандарт и рекомендации по хранению наборов лингвистических данных в виде взаимосвязанных текстовых файлов, упрощая долгосрочное архивирование и справедливый доступ к этим наборам данных через репозитории, такие как Zenodo, стандартизированный формат подачи заявок для таких журналов, как Dictionaria, упрощенное создание приложений CLLD из схем элементов, специфичных для модуля CLDF.

Использование наборов данных CLDF в качестве «входных» для приложений CLLD также решает одну из самых больших проблем публикации данных в веб-приложении: как обрабатывать несколько версий данных? С помощью CLDF наборы данных могут быть версионными, и несколько версий могут быть опубликованы в репозитории, в то время как веб-приложение переведено в доступный для просмотра интерфейс последней версии.

Стандартизированные форматы данных могут стать основой не только для инструментов, но и для учебного материала по исторической лингвистике и лингвистической типологии.

Основными типами кросс-лингвистических данных являются любые табличные данные, которые обычно анализируются с использованием количественных (автоматизированных) методов или становятся доступными с помощью программных средств, таких как например фреймворк CLLD:

[1] CLLD. Published datasets. – URL: https://clld.org/datasets.html (дата обращения: 01.12.2021).

[2] Dictionaria. – URL: https://dictionaria.clld.org/ (дата обращения: 01.12.2021).

[3] Cross-Linguistic Data Formats. – URL: https://cldf.clld.org/ (дата обращения: 01.12.2021).

- списки слов (или более сложные лексические данные)
- структурные наборы данных (например функции WALS)
- простые словари

*Принципы проектирования.* Данные должны быть как редактируемыми «от руки», так и поддающимися чтению и записи с помощью программного обеспечения. Данные должны быть закодированы в виде текстовых файлов UTF-8.

Если на сущности можно ссылаться, например – на языки через их код в Glottolog, то это следует делать, а не дублировать информацию, такую как имена языков.

Автоматическое повторное использование требует, чтобы стандарт определял не только структуру, но и семантику хранимых данных. Конечно, новые типы данных не могут быть немедленно совместимы с независимо разработанными инструментами; поэтому стандарт CLDF должен также предоставлять механизмы, позволяющие типам данных развивать хорошо понятную семантику, будучи синтаксически совместимыми с самого начала.

*Технология.* CLDF построен на модели W3 C для табличных данных и метаданных в Интернете и словаре метаданных для табличных данных. Эта модель, в силу того что она является диалектом JSON-LD, идеально подходит для объединения с онтологией, для чего необходимо указать синтаксис и семантику формата сериализации данных. CLDF структурирует кросс-лингвистические данные, чтобы сделать возможным автоматическое повторное использование.

Одной из основных целей спецификации CLDF является разграничение данных и инструментов. Использование формата на основе CSV упрощает использование этих данных в процедуре преобразования данных в стиле UNIX. Этот конвейерный стиль преобразования и анализа данных, по-видимому, лежит в основе типичных рабочих процессов, например в исторической лингвистике с использованием LingPy[1].

*История*. В то время как форматы для обмена лингвистическими данными существуют уже некоторое время, например SFM для видео или стандартный формат, используемый Toolbox, новые разработки в области исследований языкового разнообразия мотивировали новый интерес к стандартизации табличных данных в Интернете с особым акцентом на CSV.

Структура CLLD показала, что на основе одной и той же базовой модели данных может быть построено множество различных кросс-лингвистических баз данных. Проект предложил идею очень простого формата CSV для обмена очень простыми кросс-лингвистическими данными. Простота была главной целью дизайна с самого начала, поэтому рассматриваемые форматы будут развиваться, начиная с максимально простых. CLDF 1.0 обеспечивает стабильную основу для дальнейшей эволюции.

[1] LingPy. Библиотека Python для исторической лингвистики. Версия 2.6.5. – URL: http://lingpy.org. – DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy (дата обращения: 01.12.2021).

***Связанные данные для аналитики LIDER:FP***[1]

Проект предполагает определение эталонной архитектуры, которая включает в себя общие и частные задачи, требующие услуг NLP и бесплатных, открытых и совместимых ресурсов для анализа многоязычного и мультимедийного контента. Эталонная архитектура LIDER основана на открытых стандартах, обеспечивая две вещи:

- эталонную модель, которая идентифицирует те задачи, в которых лингвистические связанные данные могут поддерживать контент-аналитику, и обеспечивает стандартную декомпозицию этих задач на элементы, которые совместно могли бы решать эти задачи вместе с потоками данных между элементами;
- каталог архитектурных шаблонов, описывающий типы элементов, которые могут быть использованы в вышеуказанных задачах, типы связей между этими элементами, и ограничения на то, как они могут использоваться.

В разделах 1 и 2 проекта описываются различные источники входных данных для этой эталонной архитектуры, а также проблемы и барьеры, которые решаются лишь частично. Эталонная архитектура LIDER помещена в контекст связанных данных. В остальных разделах консорциум собрал и упорядочил обширную коллекцию задач NLP и архитектурных паттернов. Эта коллекция используется в качестве основы для разработки согласованной эталонной архитектуры.

***Модель OntoLex-Lemon***[2]

Эта модель, возникшая в результате работы Группы сообщества W3 C, первоначально была разработана с целью обеспечить полное лингвистическое обоснование онтологий. Это означает, что выражения естественного языка, используемые в метках, определениях или комментариях элементов онтологии, снабжены подробным лингвистическим описанием.

Онтологии являются важным компонентом семантической сети, но современные языки онтологий, такие как OWL и RDF (S), не поддерживают обогащение онтологий лингвистической информацией, в частности информацией о том, как объекты онтологии, т.е. свойства, классы, индивиды и т.д., могут быть реализованы на естественном языке. Модель *OntoLex-Lemon* направлена на то, чтобы закрыть этот пробел, предоставив словарь, который позволяет онтологиям обогащаться информацией о том, как описанные в них элементы словаря реализуются лингвистически, в частности в естественных языках.

OWL и RDF (S) полагаются на свойство *RDFS:lable* для фиксации связи между словарным элементом и его (предпочтительной) лексикализацией в

[1] LIDER: FP Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe. – URL: https://docplayer.net/139432478-Lider-fp-linked-data-as-an-enabler-of-cross-media-and-multilingual-content-analytics-for-enterprises-across-europe.html (дата обращения: 01.12.2021).

[2] Lexicon Model for Ontologies: Community Report. – URL: https://www.w3.org/2016/05/ontolex/ (дата обращения: 01.12.2021). См. также главу 6.

данном языке. Эта лексикализация обеспечивает лексический якорь, который делает класс, свойство, индивидуум и т.д. понятными для пользователя – человека. Использование простой метки для лингвистического обоснования, доступной в OWL и RDF (S), далеко не способно нести необходимую лингвистическую и лексическую информацию, в которой нуждаются приложения NLP, работающие с конкретной онтологией.

Цель проекта – обеспечить богатое лингвистическое обоснование онтологий. Богатое лингвистическое обоснование включает в себя представление морфологических и синтаксических свойств лексических единиц, а также синтаксически-семантический интерфейс, т.е. значение этих лексических единиц по отношению к онтологии или лексике.

Основной организующей единицей для этих лингвистических описаний является лексический класс, который обеспечивает представление морфологических паттернов для каждой записи (многозначное выражение, слово или аффикс).

Связь лексического входа с онтологической сущностью маркируется в основном свойстве денотата или опосредуется классами лексического смысла или лексико-концепта. *OntoLex-Lemon* включает в себя явный способ кодирования концептуальных иерархий, опирающийся на стандарт SKOS[1]. Лексические записи могут быть связаны через *The ontolex:evokes property* с такими концептами SKOS, которые могут представлять собой синсеты WordNet. Эта структура распараллеливает отношение между лексическими записями и онтологическими источниками.

Помимо своей первоначальной области применения, *OntoLex-Lemon* стал де-факто стандартом в области цифровой лексикографии, и применяется например в европейском инфраструктурном проекте ELEXIS[2].

Расширение сферы применения модели также нашло отражение в разработке новых модулей для лексики *OntoLex-Lemon*. Это касается, например, расширения для лексикографии [4], спецификаций для морфологии [5], а также представления частотной и корпусной информации [6].

Морфологическая модель особенно важна для кросс-лингвистической применимости *OntoLex-Lemon*, поскольку она направлена на поддержку языков с большим количеством внутренних чередований.

В настоящее время обсуждаются технические характеристики фонологических процессов и морфологические и синтаксические комбинаторные ограничения, такие как ограничения на компаундирование и деривацию.

### *Проект PRET-a-LLOD*[3]

Языковые технологии все больше полагаются на большие объемы данных, лучший доступ и повторное использование ЛИР; они позволят предоставлять многоязычные решения, которые будут поддерживать формирующийся

[1] Simple Knowledge Organization System (SKOS) Primer. – URL: https://www.w3.org/TR/skos-primer (дата обращения: 01.12.2021). См. также главу 8.

[2] European Lexico-Graphic Infrastructure. – URL: http://www.elex.is/ (дата обращения: 01.12.2021).

[3] The Prêt-à-LLOD Project. – URL: https://pret-a-llod.github.io/ (дата обращения: 01.12.2021).

единый цифровой рынок в Европе. Однако данные редко бывают «готовыми к использованию», и специалисты по языковым технологиям тратят более 80% своего времени на очистку, организацию и сбор наборов данных. Снижение этих усилий обещает огромную экономию средств для всех секторов, где требуются языковые технологии. Важная часть процесса извлечения – преобразования – загрузки включает связывание наборов данных с существующими схемами, но лишь немногие специалисты используют преимущества технологий связанных данных для выполнения этой задачи.

Цель проекта – увеличить применение языковых технологий, используя комбинацию связанных данных и языковых технологий, т.е. LLOD, для создания готовых многоязычных данных. *PRET-a-LLOD* стремится достичь этого путем создания новой методологии и технологии создания данных, применимых к широкому спектру секторов и приложений и основанных на ЛИР, которые могут быть интегрированы с помощью семантических технологий, в частности использования LLOD.

В рамках проекта будут разработаны новые инструменты для преобразования и связывания наборов данных, которые будут применяться как к данным, так и к метаданным, чтобы обеспечить доступ к разнородным репозиториям данных.

Проект предполагает автоматический анализ лицензий, чтобы определить, как данные могут быть законно использованы и проданы поставщиками ЛИР.

Будут созданы инструменты для объединения языковых сервисов и ресурсов в сложные конвейеры с использованием семантических технологий. Это приведет к появлению устойчивых предложений данных и услуг, которые можно будет развернуть на многих платформах, включая еще неизвестные, и которые могут быть самоописаны с помощью связанной семантики данных. Этот инструментарий будет апробирован в четырех пилотных проектах. Инструментарий увеличит распространение языковых технологий за счет устранения препятствий на пути их использования, и обеспечит экономию средств, которая принесет пользу пользователям как государственного, так и частного секторов.

Основная цель проекта заключается в обеспечении многоязычного междисциплинарного доступа к ЛИР, используемых в многоязычных трансграничных ситуациях. Это достигается за счет предоставления инструментов обнаружения данных, основанных на метаданных, агрегированных из нескольких источников, методологий описания свойств данных и услуг, а также инструментов для вывода возможных значений ресурса, полученного после сложного конвейера. С этим связана разработка трансформационной платформы, которая отображает наборы данных в форматы и схемы, которые могут быть использованы LLOD. Наконец, проект развивает экосистему для поддержки разработки языковых технологий, основанных на LOD, – от базовых инструментов, таких как теггеры, до полноценных приложений, таких как системы машинного перевода, основанных на семантических технологиях.

Существующие технологии семантического связывания применяются, чтобы обеспечить полуавтоматическую интеграцию услуг.

Устойчивость языковых технологий и ресурсов является серьезной проблемой, поэтому необходимо повысить устойчивость ресурсов, предоставляя услуги в виде данных и используя программное обеспечение с открытым исходным кодом.

Создаются также вспомогательные инструменты для измерения и анализа достоверности, ремонтопригодности и лицензирования данных и услуг. Это повышает качество и охват языковых ресурсов и технологий, гарантируя, что услуги легче архивировать и повторно использовать и, таким образом, им дольше оставаться доступными.

В проекте реализуются методы для обнаружения, преобразования и связывания лингвистических данных таким образом, чтобы они могли быть опубликованы как LLOD.

*Обнаружение*

PRET-a-LLOD предоставляет гибкий метод обнаружения, который может выполнять поиск как ЛИР, так и сервисов. Поскольку многие реальные проблемы могут быть решены только комбинацией нескольких наборов данных и сервисов, проект разрабатывает новую систему workflow, которая поддерживает цепочку нескольких сервисов с использованием семантических описаний сервисов и контейнеризации.

Полученная в результате платформа обнаружения и поиска состоит из единого и удобного для пользователя портала. Эта платформа построена поверх платформы *Linghub*, которая теперь импортирована на платформу CKAN (портал открытых данных), обеспечивая устойчивость и масштабируемость.

*Трансформация*

Существующие ЛИР используют различные форматы. Для того чтобы их (повторно) использовать, необходимо преодолеть структурные и концептуальные различия. PRET-a-LLOD решает эту проблему с помощью интегрированной методологии, которая преобразует языковые ресурсы. Модель преобразования – это OntoLex-Lemon (кратко представленные выше) для лексических данных, поддерживающих представление языковых данных в RDF.

PRET-a-LLOD объединяет множество компонентов для трансформации, обогащение и манипулирование языковыми ресурсами в рамках гибкой интегрированной платформы RDF-преобразований, получившей название *Fintan* [7].

Помимо преобразования корпусных данных, *Fintan* был расширен для преобразования лексических наборов данных в представлении RDF с использованием OntoLex-Lemon. *Fintan* в настоящее время поддерживает 16 разных корпусных форматов.

Выбранные наборы данных, преобразованные в рамках проекта, включают в себя:

- RDF – преобразование полных данных Apertium: 55 двуязычных данных;
- RDF – преобразование базы данных PanLex: 2500 словарей, скомпилированных в 1651 двуязычных словарей (т.е. те, которые содержат более 10 000 записей в языковой паре);

- RDF – конверсия других словарных коллекций: 252 двуязычных словаря (FreeDict, XDXF);
- RDF – конверсия инвентаризаций морфем: 110 моноязычных инвентаризаций морфем из UniMorph и семь крупномасштабных морфологических ресурсов для семи языков ЕС;
- RDF – преобразование WordNet: три для романских языков и один для немецкого языка;
- преобразование пяти терминологических ресурсов из TBX в RDF.

*Связывание*

Проект разрабатывает (полу)автоматизированные механизмы связывания. Это касается как концептуального уровня языковых описаний, так и лексических данных.

В контексте межъязыкового сопоставления концептов уже существующий инструмент сопоставления онтологий CIDER-CL[1] дополняется современными технологиями, основанными на межъязыковых встраиваниях слов. Лексикализация онтологий направлена на разработку методов, которые могут связать существующие онтологии с лексиконами в более широком масштабе.

Другие работы по связыванию выполняются при поддержке «Naisc», инструмента, разработанного в Национальном университете Ирландии в Голуэе и используемого в рамках проекта Европейской лексикографической инфраструктуры (ELEXIS)[2].

*Обнаружение охраняемых ЛИР и доступ к ним*

В рамках PRET-a-LLOD решалась проблема обнаружения и исполнения лицензионных условий для ЛИР, объединенных в сложные конвейеры. Разрабатывались методы автоматизированного исполнения лицензионной политики для операций с ЛИР. Эта работа основана на спецификациях ODRL[3]. Поскольку все эти шаги должны быть тщательно разработаны и интегрированы в рабочий процесс, PRET-a-LLOD разрабатывает протокол, основанный на семантической разметке, который направлен на то, чтобы позволить языковым сервисам легко подключаться к многосерверной рабочей среде.

Практические результаты PRET-a-LLOD включают в себя четыре отраслевые пилотные проекта, которые призваны продемонстрировать актуальность, переносимость и применимость методов и методик к практическим проблемам в индустрии языковых технологий.

*Извлечение терминов и сопоставление понятий*

В пилоте I для компании Semantic Web было необходимо улучшить извлечение терминов и сопоставление концепций, предлагаемые их флагманским

[1] CIDER-CL A system for monolingual and cross-lingual ontology alignment. – URL: https://oeg.fi.upm.es/files/cider-cl/ (дата обращения: 01.12.2021).

[2] ELEXIS. – URL: https://elex.is/ (дата обращения: 01.12.2021).

[3] Open Digital Rights Language. W3 C specification. – URL: https://www.w3.org/TR/odrl-model (дата обращения: 01.12.2021).

продуктом PoolParty. Кроме того, требовалась замена некоторых проприетарных ЛИР, используемых в настоящее время в PoolParty, на ЛИР с открытым исходным кодом.

*Связывание лексических знаний для облегчения интеграции и более широкого применения лексикографических ресурсов для технологических компаний*

В пилотном проекте II для Издательства Оксфордского университета, разрабатывалась методология связывания лексических данных с языковыми сервисами, таким образом непосредственно решая одну из ключевых проблем, которыми занимается PRET-a-LLOD.

Два варианта задачи связывания будут решаться в соответствующих предметах, а именно: связывание различных словарей (моно- или двуязычных) на уровне значения (т.е. на уровне смысла) и связывание корпусных данных со словарными смыслами посредством устранения смысловой неоднозначности слов.

*Поддержка развития государственных услуг в рамках Открытого правительства как внутри страны, так и за ее пределами*

В рамках пилотного проекта III для сервиса *Derilinx*[1] поставлена задача предложить инструменты и интерфейсы для интуитивного и трансграничного доступа к открытым данным с использованием естественного языка, т.е. веб-приложение, отвечающее на запросы касающиеся информации о государственных услугах, через информационную панель; это включает в себя анализ пользовательских запросов на естественном языке, преобразование их в формальные запросы к порталу здравоохранения и разработка чат-бота, обеспечивающего устные ответы на запросы.

*Многоязычная текстовая аналитика для извлечения реальных данных в фармацевтическом секторе*

В пилоте IV для компании Semalytix[2] разработана система многоязычного обучения, поиска и анализа текстов для фармацевтической промышленности. Извлечение реальных данных требует анализа больших объемов разнородного контента, включая субъективные оценки пациентов и медицинских экспертов, которые обычно доступны в виде неструктурированного текста на нескольких языках. При разработке специфичных для предметной области многоязычных текстовых аналитических приложений необходимы методы, поддерживающие генерацию фактических данных, взаимодействие между ресурсами LLOD и архитектурами глубокого машинного обучения.

***Сотрудничество и технологическое взаимодействие с проектом Европейской лексикографической инфраструктуры ELEXIS***[3]

*ELEXIS* актуален из-за того, что связанные данные играют все большую роль в цифровой лексикографии. Эта связь с *ELEXIS* важна для PRET-a-LLOD,

[1] Drive Decision-Making. Inspire Change. – URL: https://derilinx.com
[2] SEMALYTIX. – URL: https://www.semalytix.com/ (дата обращения: 01.12.2021).
[3] ELEXIS. – URL: https://elex.is (дата обращения: 01.12.2021).

поскольку все большее сообщество лексикографов использует OntoLex-Lemon и другие технологии LLOD, обеспечивая тем самым устойчивость методов, разработанных в рамках PRET-a-LLOD.

Сотрудничество с *ELEXIS* также связано с повышением совместимости стандартов, например установлением мостов между OntoLex-Lemon как результатом работы Группы сообщества W3 C и руководящими принципами кодирования TEI Lex-0 в рамках развития сообщества TEI. Важно определить, какой (де-факто) стандарт лучше всего подходит для разных аспектов цифровой лексикографии.

Еще одна важная связь была установлена с проектом *European Language Grid* (*ELG*)[1], который стартовал одновременно с PRET-a-LLOD в январе 2019 года. Отношения и сотрудничество с ELG заключаются в том, чтобы использовать услуги LLOD на платформе ELG.

Первое и успешное испытание было реализовано для выполнения преобразования из набора TBX в RDF на основе OntoLex-Lemon.

Это важное достижение, поскольку оно поддерживает устойчивость результатов проекта PRET-a-LLOD, позволяя развертывать свои данные и сервисы на различных платформах помимо основной инфраструктуры LLOD.

Наконец, упомянем о влиятельной роли, которую PRET-a-LLOD сыграл в недавно созданной Европейской сети для веб-ориентированной лингвистической науки о данных (*NexusLinguarum*[2]) проекта, направленного на взаимодействие между лингвистами и компьютерщиками по всему ЕС. В *NexusLinguarum* технологии LLOD будут играть центральную роль, и результаты PRET-a-LLOD будут необходимы для построения целостной экосистемы многоязычных и семантически совместимых лингвистических данных, которые использует *NexusLinguarum*.

Текущее состояние проекта PRET-a-LLOD свидетельствует о дальнейшем расширении облачной инфраструктуры LLOD и повышении устойчивости LLOD-совместимых сервисов и наборов данных. Мы считаем, что технологии и ресурсы LLOD означают развитие устойчивой экосистемы интерактивных, веб-языковых технологических сервисов и языковых ресурсов в соответствии с целями всего направления связанных открытых данных. Благодаря проекту PRET-a-LLOD и связанным с ним инфраструктурным инициативам, эффект ожидаемый от использования связанных данных будет достигнут в течение ближайших лет.

Более подробная информация о современных инструментах и ресурсах PRET-a-LLOD представлена на веб-сайте проекта[3].

Использование и развитие LLOD было и есть предмет еще нескольких крупномасштабных исследовательских проектов, включая:

- LOD2. Создание знаний из взаимосвязанных данных (11 стран ЕС + Корея, 2010–2014 гг.)

[1] European Language Grid (ELG). – URL: https://www.european-language-grid.eu/ (дата обращения: 01.04.2022).

[2] NexusLinguarum. – URL: https://nexuslinguarum.eu/ (дата обращения: 01.12.2021).

[3] The Prêt-à-LLOD Project. – URL: https://pret-a-llod.github.io/

- MONNET. Многоязычные онтологии сетевых знаний (пять стран ЕС, 2010–2013 гг.)
- QTLeap. Качественный перевод с использованием подходов глубинной инженерии (шесть стран ЕС, 2013–2016 гг.)
- LiODi. Связанные открытые словари (BMBF eHumanities Early Career Research Group, Университет Гете, Франкфурт, Германия, 2015–2020 гг.)
- FREME. Открытая структура электронных услуг для многоязычного и семантического обогащения цифрового контента (шесть стран ЕС, 2015–2017 гг.)
- POSTDATA. Стандартизация поэзии и связанные открытые данные (стартовый грант ERC, UNED, Испания, 2016–2021 гг.)
- LiLa (Linking Latin). Создание базы знаний ЛИР для латинского языка (ERC Consolidator Grant, Universita Cattolica del Sacro Cuore, Италия, 2018–2023 гг.)

Обобщение проектов по применению связанных открытых данных в коллаборативных лингвистических исследованиях и разработках, предполагающих интенсивное использование лингвистических данных, представлено в работе [11].

### Литература к главе 19


1. Linguistic Linked Data: Representation, Generation and Applications / Cimiano P., Chiarcos C., McCrae John P., Gracia J. – Springer International Publishing, 2020.
2. Workshop on Linked Data in Linguistics (LDL). – URL: https://www.aclweb.org/anthology/venues/ldl/ (дата обращения: 01.12.2021).
3. 5 th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked. – URL: https://www.science-community.org/en/node/163285 (дата обращения: 01.12.2021).
4. 6 th Workshop on Linked Data In Linguistics. – URL: https://elex.is/6th-workshop-on-linked-data-in-linguistics/ (дата обращения: 01.12.2021).
5. Recent Developments for the Linguistic Linked Open Data Infrastructure // Proceedings of the 12 th Conference on Language Resources and Evaluation (LREC 2020). – Marseille, 2020. – 11–16 May. – P. 5660–5667. – URL: https://zenodo.org/record/3934626 (дата обращения: 01.12.2021).
6. Reconciling Heterogeneous Descriptions of Language Resources / John P. McCrae, Philipp Cimiano CIT-EC, Bielefeld University Bielefeld, Germany, Victor Rodr´ıguez Doncel, Daniel Vila-Suero, Jorge Gracia, Universidad Politecnica de Madrid, Madrid, Spain @fi.upm.es Luca Matteis, Roberto Navigli, University of Rome, La Sapienza, Rome, Italy, Andrejs Abele, Gabriela Vulcu, Paul Buitelaar Insight Centre, National University of Ireland Galway, Ireland. – URL: https://www.aclweb.org/anthology/W15-4205.pdf (дата обращения: 01.12.2021).
7. Bosque-Gil J., Gracia J. The OntoLex Lemon lexicography module : Technical report // W3 C Commu-nity Group Ontology-Lexica. Final Community GroupRepor. – 2019.
8. Challenges forthe representation of morphology in ontology lexicons / Klimek B., McCrae J.P., Bosque-Gil J., Ionov M., Tauber J.K., Chiarcos C. // Proceedings of eLex 2019. Electronic lexicography in the 21 st century: Smart lexicography. – 2019.

9. Chiarcos C., Ionov M. The OntoLex Lemonmodule for frequency, attestation and corpus infor-mation : Technical report / W3 C Community GroupOntology-Lexica, draft version. – 2019. – Mar 3.

10. F¨ath C., Chiarcos C., Ebbrecht B. Fintan – Flexible, Integrated Transformation and Anno-tation eNgineering // Proceedings of the Twelfth Inter-national Conference on Language Resources and Evalu-ation (LREC 2020), Marseille, France, May 11–16, 2020 / European Language Resources Association (ELRA). – Marseille, 2020.

11. Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences / Edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, Christian Chiarcos. – The MIT Press, 2020. – 247 p. – DOI: https://doi.org/10.7551/mitpress/10990.001.0001 ; ISBN electronic: 9780262357210

# ГЛАВА 20. ЛИР В КОНТЕКСТЕ ЦИФРОВОЙ ГУМАНИТАРИСТИКИ

## Общие сведения

Назначение значительной части создаваемых ЛИР выходит за рамки узкопрофессиональных задач поддержки лингвистических исследований. Одна важная сфера применения за этими рамками – прикладные задачи, т.е. применение языковых технологий в различных компьютерных системах обработки языка и речи, о чем было достаточно сказано в предыдущих главах книги.

В этой главе мы рассмотрим другую сферу применения ЛИР – это различные задачи использования цифровых ЛИР и лингвистических инструментов в комплексных и междисциплинарных задачах социальных и гуманитарных наук. Сферу применения ИТ в этих науках в последние годы принято называть цифровой гуманитаристикой.

Обсудим эту область подробнее. Цифровая гуманитаристика (DH)[1] – это область научной деятельности на стыке компьютерных или цифровых технологий и гуманитарных дисциплин. Она включает в себя систематическое использование цифровых ресурсов в гуманитарных науках, а также анализ их применения.

В конкретных терминах DH охватывает целый ряд тем: от создания онлайн-коллекций первичных источников (в первую очередь текстовых) до интеллектуального анализа больших наборов культурных данных и тематического моделирования. DH включает в себя как оцифрованные, так и рожденные в цифровом виде материалы и сочетает методологии из традиционных гуманитарных и социальных дисциплин (таких как риторика, история, философия, лингвистика, литература, искусство, археология, музыка и культурология), а также разнообразные методы и инструменты, предоставляемые современным уровнем развития информатики. В качестве примера обзорной работы по проблемам DH приведем работу [1].

DH занимает важное место в структуре современных гуманитарных и социальных исследований. Выше была кратко описана деятельность Европейской ассоциации цифровой гуманитаристики. Особенно важно, что для DH создается специальная инфраструктура, поддержку которой реализует специальный консорциум типа ERIC, получивший название DARIAH.

[1] Для цифровой гуманитаристики в зарубежной литературе распространена аббревиатура DH (Digital humanities), которую мы тоже будем использовать.

## Консорциум DARIAH

Цифровая исследовательская инфраструктура для искусств и гуманитарных наук (DARIAH)[1] направлена на расширение и поддержку цифровых исследований и преподавания в области искусств и гуманитарных наук. DARIAH – это сеть людей, экспертных знаний, информации, знаний, контента, методов, инструментов и технологий из стран – членов организации. Она разрабатывает, поддерживает и эксплуатирует инфраструктуру в поддержку научно-исследовательской практики, основанной на ИКТ, и поддерживает исследователей в использовании ее для создания, анализа и интерпретации цифровых ресурсов. Поскольку подавляющая часть информации, создаваемой и используемой в гуманитарной науке, представлена в виде текстов и речи, большинство создаваемых в сети ресурсов и инструментов в той или иной степени используют ЛИР.

DARIAH была создана как Европейский исследовательский инфраструктурный консорциум (ERIC) в августе 2014 года. В настоящее время DARIAH имеет 19 членов, одного наблюдателя и несколько сотрудничающих партнеров в семи странах, не являющихся членами организации.

Структурно DARIAH работает через четыре общеевропейские сети виртуальных центров компетенций (VCC):

VCC1 – электронная инфраструктура

VCC2 – связь науки и образования

VCC3 – управление контентом

VCC4 – пропаганда и расширение влияния

В составе DARIAH функционирует 24 рабочие группы (РГ), причем некоторые из них полностью занимаются проблематикой ЛИР, в том числе:

***Лексические ресурсы***. Целями этой РГ являются: исследование, оценка и рекомендации по стандартным инструментам и методам для создания, применения и распространения цифровых лексических ресурсов и других видов структурированных данных; поощрение, развитие и публикации цифровых лексикографических исследований.

***Аналитика текста и данных***. Обработка естественного языка является ключевой технологической областью для анализа и извлечения знаний из неструктурированных данных в текстовой форме. Существует широкий спектр задач и подходов: от статистических инструментов, таких как конкорданты, до аннотаций для анализа тональности и до парсинга для извлечения информации. РГ разрабатывает методологию применения современных средств NLP к вопросам гуманитарных исследований.

***Разработка и сопровождение тезауруса***. Цель РГ – создание тезауруса для гуманитарных наук, прежде всего выявление понятий верхнего уровня, отвечающих требованиям объективности и междисциплинарности. РГ использует методы категориальной семантики, чтобы определить существенные свойства общих понятий, независимо от научной области, в которой они применяются, что позволяет проводить классификацию последовательно и объективно.

[1] The Digital Research Infrastructure for the Arts and Humanities (DARIAH). – URL: https://www.dariah.eu (дата обращения: 01.12.2021).

Другие РГ также занимаются проблемами ЛИР, в том числе в рамках общих задач по управлению контентом или образовательных программ.

Среди инструментов и сервисов, созданных в рамках DARIAH, важное место занимают ЛИР. Среди них можно отметить следующие:

***Vocabs***. Во многих областях научной деятельности контролируемые словари (справочники, тезаурусы и т.д.) обеспечивают качество ресурсов и взаимодействие между ними. Сервис *Vocabs* предоставляет услуги и инструменты, которые позволяют совместно создавать, поддерживать и публиковать словари и таксономии любого рода. Система основана на программном обеспечении с открытым исходным кодом *Skosmos*, которое использует SKOS[1] в качестве базовой модели данных. Словари могут быть найдены с помощью интерфейса поиска или через алфавитный или тематический индекс, а также могут быть представлены в виде связанных данных. *Vocabs* используется также в качестве сервиса для участников консорциума CLARIAH-AT, для рабочей группы по ведению тезаурусов в DARIAH и для проекта PARTHENOS.

***Generic Search.*** Общий поиск представляет собой распределенный мета-поиск по коллекциям и ресурсам, хранящимся в реестрах DARIAH, независимо от их схем данных и метаданных. Он устанавливает и отслеживает семантические связи между структурно различными коллекциями и их конкретными ресурсами. Это достигается с помощью переходов, которые соединяют не только элементы данных из различных форматов и стандартов данных, но и отображают типы данных и значения.

***TaDiRAH*** (Taxonomy of Digital Research Activities in the Humanities). Таксономия цифровой исследовательской деятельности в гуманитарных науках была разработана для использования общественными сайтами и проектами, направленными на структурирование информации относящейся к цифровым гуманитарным наукам. Ожидается, что таксономия будет особенно полезна для усилий, направленных на сбор информации о цифровых гуманитарных инструментах, методах, проектах. Отдельные фасеты таксономии – это виды деятельности, научные объекты, научные методики.

***NeMO.*** Онтология методов *NeMO* – это комплексная онтологическая модель научной практики в области социальных и гуманитарных наук, разработка которой осуществляется через исследовательскую сеть ESF NeDiMAH[2]. NeMO совместима с известным проектом CIDOC CRM[3]. NeMO опирается на категории агентов (акторов), процессов (деятельности и методов) и ресурсов (информационных ресурсов, инструментов, концепций), проявляющихся в научном процессе. Онтология основана на результатах обширных эмпирических исследований и моделирования научных практик, выполненных в проектах DARIAH. NeMO включает в себя многие существующие таксономии научных методов и инструментов, посредством ото-

[1] SKOS. Simple Knowledge Organization System Reference. – URL: https://www.w3.org/TR/skos-reference/ (дата обращения: 01.12.2021).

[2] NeDiMAH (Network for Digital Methods in the Arts and Humanities). – URL: https://www.dariah.eu/activities/projects-and-affiliations/nedimah/ (дата обращения: 01.12.2021).

[3] CIDOC Conceptual Reference Model (CRM). – URL: http://www.cidoc-crm.org/ (дата обращения: 01.12.2021).

бражений определенных в них терминов на семантическую структуру понятий NeMo, что позволяет сочетать подходы разных научных практик и использовать разную лексику.

***DARIAH-DE Topics Explorer*** – метод анализа распределения семантических кластеров слов, так называемых «тем» в текстовой коллекции. Он может быть использован для изучения содержимого корпуса, а также для генерирования связанных с содержимым признаков для классификации цифрового текста. Тематическое моделирование полностью опирается на сами анализируемые тексты; оно не использует дополнительных источников информации, таких как словари или внешние обучающие данные, что делает его в значительной степени независимым от языка и орфографических условностей. Метод основан исключительно на статистическом анализе встречаемости символов (на уровне слов), который переводится в вероятные семантические отношения.

***DKPro Wrapper*** – оболочка для программного продукта компании DKPro, ориентированная на извлечение лингвистической информации из текстов. В руководстве пользователя объясняется простое управление этим инструментом при помощи командной строки на Java.

Многие другие инструменты и сервисы, разработанные и распространяемые через DARIAH, предназначены в том числе для разработчиков ЛИР.

В качестве примера рассмотрим проект ***PARTHENOS***[1], действовавший в 2016–2019 гг. *PARTHENOS* нацелен на взаимодействие исследований в широком круге дисциплин: лингвистики, культурного наследия, истории, археологии и смежных областях. *PARTHENOS* добивается этого посредством объединения европейских исследовательских инфраструктур, интеграции инициатив, электронных ресурсов, поддержки общих стандартов, координации совместной деятельности, согласования понятий, осуществления политики, а также разработки объединенных услуг и общих решений одних и тех же проблем.

*PARTHENOS* вырабатывает общие решения для реализации совместных стратегий и решений для жизненного цикла гуманитарных и лингвистических данных. В это понятие входят: положения о междисциплинарном и повторном использовании данных, реализации общих политик; стандартизация и интероперабельность; общие инструменты для сервисов, ориентированных на данные, таких как обнаружение ресурсов, поисковые сервисы, оценка качества метаданных, аннотирование источников; коммуникационная деятельность, а также совместные учебные мероприятия.

В заключение обзора деятельности DARIAH отметим, что рабочие группы и проекты DARIAH реализуются в тесной координации со специализированными коллективами в области ЛИР, прежде всего входящих в специализированную инфраструктуру CLARIN, деятельность которого была описана выше. Это относится и к проекту PARTHENOS.

[1] Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS). – URL: https://www.dariah.eu/activities/projects-and-affiliations/parthenos/ (дата обращения: 01.12.2021).

## Huma-Num[1]. Программа цифровой гуманитаристики

Хорошим примером включения проблематики ЛИР в общий контекст цифровой гуманитаристики является французская научная инфраструктура цифровой гуманитаристики, именуемая *Huma-Num.*

Huma-Num нацелена на поддержку исследовательских сообществ, предоставляя услуги, оценку и инструменты для цифровых исследовательских данных. Для выполнения своих задач Huma-Num создает консорциумы, которые представляют собой финансируемые Huma-Num группы исследователей и инженеров, работающих в общих областях.

Вместе с консорциумами Huma-Num координирует производство цифровых данных, предлагая при этом различные платформы и инструменты для обработки, сохранения, распространения и долгосрочного хранения цифровых научных данных. Одна из целей – способствовать обмену данными, чтобы другие исследователи, сообщества или дисциплины могли повторно использовать их, в том числе с междисциплинарной точки зрения и разными способами. Принципы и методы Семантической сети (LOD, RDF, SPARQL, SKOS, OWL), на которых основаны сервисы Huma-Num, позволяют документировать или повторно использовать данные для различных целей.

Основная задача Huma-Num заключается в сохранении научных результатов лабораторий, в частности, ресурсов, полученных в ходе исследовательской деятельности: корпусов, баз данных, документальных баз, информационных систем, обследований, данных наблюдений, подготовленных или находящихся в процессе производства.

С 2013 года услуги Huma-Num ориентированы на использование цифровых данных, т.е. предлагаются инструменты в виде программных средств и / или услуг для применения к данным методов преобразования, анализа, вывода визуализации и взаимодействия. Эти методы включают открытие данных и метаданных, взаимодействие метаданных (стандартизация, API) и предоставление доступа к данным и метаданным с помощью протоколов взаимодействия (OAI-PMH, APIs, Семантической сети и т.д.). О масштабах Huma-Num можно судить по таким цифрам: эта инфраструктура поддерживает 320 специализированных сайтов[2], в ее проектах участвует свыше 200 организаций, и не только французских.

Один из основных сервисов Huma-Num, в котором реализованы эти принципы, это – ***ISIDORE***[3], поисковая система, предоставляющая доступ к цифровым данным из области гуманитарных и социальных наук (SSH). ISIDORE – сервис, который собирает, обогащает и выделяет цифровые данные и документы по гуманитарным и социальным наукам, обеспечивая при этом единый доступ к ним. ISIDORE собирает метаданные и полные тексты

[1] Huma-Num est une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales. – URL: https://documentation.huma-num.fr (дата обращения: 01.12.2021)

[2] Liste des sites web hébergés par Huma-Num. – URL: https://www.huma-num.fr/annuaire-des-sites-web/ (дата обращения: 01.12.2021).

[3] ISIDORE. – URL: https://isidore.science/about (дата обращения: 01.12.2021).

из электронных изданий, корпусов, баз данных и научных новостей, доступных в сети Интернет.

ISIDORE собирает данные на французском (произведенном во Франции или во франкоговорящем мире), английском и испанском языках. После сбора эта информация обогащается на трех языках (английском, испанском и французском) с использованием словарных списков или тезаурусов, и устанавливаются ссылки к научным репозиториям, которые создаются либо научными сообществами (GeoEthno, Pactols и т.д.), либо крупными институтами (Rameau, LCSH, BNE, Gemet, Lexvo, GeoNames и т.д.). Многоязычное обогащение позволяет связывать данные воедино. Эта информация является точками входа в полный текст, который также индексируется, когда это возможно.

Поисковая платформа ISIDORE отличается от традиционных поисковых систем во многом:

- целенаправленный сбор метаданных и научных данных, структурированных в соответствии с международными стандартами и свободно доступных в Интернете;
- индексация неструктурированных данных (например, полный текст научной статьи) и структурированных данных (например, документальные метаданные);
- стандартизация метаданных и обогащение данных на основе стандартов, признанных в сообществе;
- графический интерфейс поиска, использующий богатство структурированных данных и словарей, чтобы превратить пользователя в актера своего исследования;
- выделение индексированных источников данных (каталогов источников);
- предоставление, в соответствии с возможностями и соглашениями, с производителями данных, метаданных, обогащенных движком, следующим принципам связанных данных.

Из приведенных примеров организаций, действующих в сфере DH, очевидно, что ЛИР являются либо инструментом этого научно-прикладного направления, либо научным результатом организаций DH. Причем ЛИР могут создаваться как самостоятельные ресурсы, пригодные для многократного использования, а могут входить в состав различных информационных систем.

Еще одним доказательством тезиса, что ЛИР являются основной частью DH, могут служить материалы научной сессии Отделения историко-филологических наук РАН 15 апреля 2021 г. «Гуманитарные науки в эпоху цифровизации» [2]. Из 12 докладов на этой сессии восемь было посвящено проектам по созданию и использованию ЛИР и филологических ресурсов в целом.

**Гуманитарные ресурсы с лингвистическим компонентом**

В этом разделе мы приведем примеры многоцелевых ресурсов, включающих существенные лингвистические данные или инструменты лингвистического анализа и предназначенные для применения в смежных гуманитарных науках. Некоторые из этих проектов упоминались выше.

***Информационная система немецких граффити INGRID***[1] – это проект сотрудничества между кафедрой лингвистики университета Падерборна и кафедрой истории искусств Технологического института Карлсруэ (KIT). В рамках совместного проекта создается коллекция изображений граффити, которые будут храниться в базе данных изображений и станут доступными для научного использования. В настоящее время зарегистрировано более 100 000 граффити с 1983 по 2018 год из крупных городов Германии, включая Кельн, Мангейм и Мюнхен.

***База данных о местах, языке, культуре и окружающей среде D-PLACE***[2]**.** D-PLACE содержит культурную, лингвистическую, экологическую и географическую информацию для более чем 1400 человеческих «обществ». «Общество» представляет собой группу людей в конкретной местности, которые часто имеют общую языковую и культурную идентичность. Все культурные описания помечены датой, к которой они относятся, и этнографическими источниками, которые предоставили описания. Большинство культурных описаний в D-PLACE основаны на этнографических работах, выполненных в XIX и первой половине XX века (до 1950 года).

***Тихоокеанский и региональный архив цифровых источников по исчезающим культурам PARADISEC***[3] предлагает средства для цифрового сохранения и доступа к находящимся под угрозой исчезновения материалам со всего мира. Создана система доступа, каталогизации и оцифровки аудио-, текстовых и визуальных материалов, а также сохранения цифровых копий. Основное внимание уделяется сохранности материалов, которые в противном случае были бы потеряны, особенно полевых записей 1950-х и 1960-х годов.

***Немецкий текстовый архив DTA***[4] представляет онлайн-подборку немецкоязычных работ по различным дисциплинам примерно с 1650 по 1900 год. Электронные полные тексты сопровождаются лингвистическими аннотациями. Средства поиска допускают варианты правописания. DTA представляет немецкоязычные печатные работы в виде полного текста и цифрового факсимиле. Отбор текстов производился на основе лексикографических критериев и включал в себя научные тексты, тексты из повседневной жизни и литературные произведения. Оцифровка производилась с первого издания каждой работы. Используя цифровые изображения этих изданий, текст сначала дважды набирался вручную («двойной ключ»). Электронный полнотекстовый текст для представления его структуры был закодирован в соответствии со стандартом XML TEI P5: текст маркируется, лемматизируется, а части речи аннотируются. Таким образом, ДТА представляет собой лингвистически обработанный исторический полнотекстовый корпус. Благодаря

[1] Informationssystem Graffiti in Deutschland. – URL: https://www.uni-paderborn.de/forschungsprojekte/ingrid/ (дата обращения: 01.12.2021).

[2] Database of Places, Language, Culture and Environment (D-PLACE). – URL: https://d-place.org (дата обращения: 01.12.2021).

[3] Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). – URL: http://www.paradisec.org.au/ (дата обращения: 01.12.2021).

[4] Deutsches Textarchiv. – URL: http://www.deutschestextarchiv.de/ (дата обращения: 01.12.2021).

междисциплинарному характеру корпуса ДТА он также предлагает ценные исходные тексты для смежных гуманитарных дисциплин, а также для правоведов и экономистов.

***Система обмена данными детской речи CHILDES***[1]. Система предназначена для изучения детской речи, прежде всего для педагогических, а также медицинских исследований. CHILDES – это детский компонент известного проекта TalkBank, который создан в Карнеги Университет Меллона при поддержке и сотрудничестве сотен участников и десятков сотрудников. Цель TalkBank – способствовать фундаментальным исследованиям в области изучения человеческой коммуникации с акцентом на устную речь.

***Центр данных и услуг для гуманитарных наук DaSCH***[2] – учреждение Швейцарской академии гуманитарных и социальных наук. Основными целями DaSCH являются: сохранение данных исследований в области гуманитарных наук и долгосрочное управление данными, обеспечение постоянного доступа к результатам исследований, чтобы сделать их доступными для дальнейших исследований и таким образом облегчить повторное использование существующих данных исследований в будущей работе, поощрение создания цифровых сетей. Включает данные по литературоведению, лингвистике, изобразительному искусству, музыке, театру, истории.

***Репозиторий исследовательских данных открытого доступа для историков LAUDATIO***[3]. Для доступа и (повторного) использования исторических корпусов репозиторий LAUDATIO использует гибкую схему представления документов с использованием TEI. Обширная схема метаданных содержит информацию о методах подготовки и проверки данных, инструментах, форматах и рекомендациях по аннотации, используемых в проекте. Репозиторий предназначен для хранения данных широкого круга исторических, текстологических и лингвистических исследований.

***Репозиторий TextGrid***[4] – цифровой архив для хранения данных исследований в области гуманитарных наук, включая корпус XML / TEI закодированных текстов, изображений и баз данных. Среди постоянно пополняющихся корпусов находится цифровая библиотека TextGrid, которая состоит из произведений более чем 600 авторов немецкой художественной литературы (проза, стихи и драма), а также нон-фикшн с начала печатного станка до начала XX века. Файлы сохраняются в различных выходных форматах (XML, ePub, PDF), публикуются и становятся доступными для поиска. Различные инструменты, например для просмотра или количественного анализа текста, могут быть использованы для визуализации или дальнейшего текстового исследования.

---

[1] Child Language Data Exchange System (CHILDES. – URL: https://childes.talkbank.org/ (дата обращения: 01.12.2021).

[2] Data and Service Center for the Humanities (DaSCH). – URL: http://data.dasch.swiss (дата обращения: 01.12.2021).

[3] Long-term Access and Usage of Deeply Annotated Information (LAUDATIO). – URL: https://www.laudatio-repository.org/ (дата обращения: 01.12.2021).

[4] The TextGrid Repository. – URL: https://www.textgridrep.org/ (дата обращения: 01.12.2021).

***Фонд института Гюйгенса Huygens ING***[1] ориентируется на старые и недоступные источники, в целях их углубленного изучения. Основное внимание фонд уделяет цифровым гуманитарным наукам, лингвистике, истории, текстологии, литературоведению. Фонд стремится использовать разнообразные инновационные инструменты для публикации цифровых источников и данных.

***Синайский Кодекс***[2]**.** Синайский кодекс – одна из самых важных книг в мире. Codex Sinaiticus – это международный проект, направленный на то чтобы объединить всю рукопись в цифровом виде и впервые сделать ее доступной для глобальной аудитории. Опираясь на опыт ведущих ученых, проект дает специалистам различных дисциплин и всем желающим возможность напрямую работать с этой знаменитой рукописью.

***Кельнский языковой архив LAC***[3] – это хранилище научных данных для лингвистики и всех гуманитарных дисциплин, работающих с аудиовизуальными данными. Архив образует кластер информационного центра гуманитарных наук в сотрудничестве с Институтом лингвистики Кельнского университета.

***Архив языков коренных народов Аляски***[4]**.** Архив хранит в цифровом виде документацию о языках народов Аляски и помогает сохранить и культивировать это уникальное наследие для будущих поколений. Архив служит исследователям различных дисциплин, преподавателям и студентам, а также членам более широкого сообщества. Эта коллекция имеет непреходящую культурную, историческую и интеллектуальную ценность, особенно для носителей аляскинских языков и их потомков.

***Южноафриканский центр цифровых языковых ресурсов SADILAR***[5] – это национальный центр, поддерживаемый Департаментом науки и техники. SADILAR уделяет особое внимание всем официальным языкам Южной Африки. Он поддерживает исследования и разработки в области языковых технологий и связанных с ними исследований в области гуманитарных и социальных наук.

***Средиземноморский центр наук о человеке. Фонотека***[6]**.** Являясь местом живой памяти, фонотека стремится объединить записи звукового наследия, которые имеют ценность с этнологической, лингвистической, исторической, музыковедческой или литературной точки зрения. Коллекция содержит более 8 000 часов аудиоархивов, записанных с конца 1950-х годов по всем гуманитарным наукам.

---

[1] Huygens Instituut voor Nederlandse Geschiedenis. – URL: https://www.huygens.knaw.nl/?lang=en (дата обращения: 01.12.2021).

[2] Codex Sinaiticus. – URL: http://www.codexsinaiticus.org/en/ (дата обращения: 01.12.2021).

[3] Language Archive Cologne. – URL: https://lac.uni-koeln.de (дата обращения: 01.12.2021).

[4] Alaska Native Language Archive. – URL: http://www.uaf.edu/anla/ (дата обращения: 01.12.2021).

[5] SADILAR. – URL: https://www.sadilar.org/ (дата обращения: 01.12.2021).

[6] Phonothèque MMSH – Maison méditerranéenne de sciences de l'homme. – URL: http://phonotheque.mmsh.huma-num.fr/ (дата обращения: 01.12.2021).

***PolMine***[1]. Проект создает хранилище текстовых данных для научных задач политологии. Основное внимание PolMine уделяется текстам, опубликованным государственными учреждениями Германии. В основе проекта лежат корпуса парламентских протоколов: парламентские материалы доступны в течение длительного периода времени, охватывают широкий спектр государственной политики и являются общественным достоянием, что делает их ценным текстовым ресурсом для политологии. Документы, издаваемые государственными учреждениями, преобразуются в устойчивый цифровой формат (TEI/XML).

***Письма Буркхардта***[2]. На платформе размещается критическое издание писем Якоба Буркхардта, реконструирующее в открытом доступе одну из важнейших европейских корреспонденций XIX века. За несколькими исключениями все эти письма не опубликованы. На более позднем этапе проект нацелен также на публикацию писем Якоба Буркхардта. Процесс редактирования осуществлялся с использованием методики семантической цифровой библиотеки *Muruca*[3]. Эта методика была изменена в ходе реализации проекта, поскольку уточнились требования филологических исследователей.

***Цифровая Библиотека Южной Азии***[4] предоставляет ученым, государственным служащим и другим пользователям цифровые материалы, в том числе на различных языках коренных народов, для справочных и исследовательских работ по Южной Азии.

***Проект GerManC***[5]. Репрезентативный исторический корпус немецкого языка 1650–1800 гг. Это – первый подобный корпус, и он предназначен в качестве основного исследовательского ресурса в ряде дисциплин. Его структура намеренно параллельна структуре существующих исторических корпусов английского языка, чтобы облегчить систематические сравнительные исследования. Региональный аспект, который был существенной чертой этих проектов, также дает информацию о связи между языком и изменениями в культурных и политических областях в Германии.

***Графическая база данных схоластических отношений в Вавилонском Талмуде***[6]. Рассматривая схоластические взаимодействия Вавилонского Талмуда как социальную сеть, состоящую из нескольких поколений, в базе построен график социальной сети Талмуда. На основе распознавания именованных сущностей в выровненных по утверждениям текстах на иврите и английском языке, были получены графики раввинов (узлов) и их взаимодействий (ребер). Новая графическая база данных, доступная в настоящее время в Интернете, предоставляет визуальную и цветовую информацию о поколении

---

[1] PolMine Project. – URL: https://polmine.github.io / (дата обращения: 01.12.2021).

[2] J.BURCKHARDT. – URL: https://burckhardtsource.org/ (дата обращения: 01.12.2021).

[3] Muruca. – URL: https://www.muruca.org/ (дата обращения: 01.12.2021).

[4] Digital South Asia Library. – URL: http://dsal.uchicago.edu/ (дата обращения: 01.12.2021).

[5] GerManC: A representative historical corpus of German 1650–1800. – URL: https://www.alc.manchester.ac.uk/modern-languages/research/german-studies/germanc/ (дата обращения: 01.12.2021).

[6] A graph database of scholastic relationships in the Babylonian Talmud. – URL: https://academic.oup.com/dsh/advance-article-abstract/doi/ 10.1093/llc/fqab015/6146162?redirectedFrom=fulltext (дата обращения: 01.12.2021).

схоластов, отношениях учителя и ученика, а также о локальных (на уровне страниц) и глобальных (на уровне Талмуда) взаимодействиях сложной динамики талмудического дискурса.

***Онлайн-корпус для изучения исторической диалектологии ОДА***[1] – новый цифровой ресурс для изучения исторической диалектологии, который сочетает текстовую и лингвистическую аннотацию в рамках одного XML-представления на основе TEI. ОДА – это диахронический корпус испанских документов, написанных между 1492 годом и концом XIX века. Включает приблизительно 600 000 лексем. Корпус специально предназначен для анализа исторических диалектных исследований, сочетает в себе филологическую / текстологическую науку и подход корпусной лингвистики. Он позволяет работать с документами, визуализированными пользователями в различных форматах. Кроме того, он позволяет осуществлять независимое управление данными, поскольку ученые могут загружать и редактировать свои материалы.

***Гуманистическая виртуальная библиотека BVH***[2] содержит материалы культурного наследия и проводит исследования, объединяющие гуманитарные и компьютерные методики. Объединяет несколько типов цифровых документов:

- факсимиле книг эпохи Возрождения;
- текстовую базу Epistemon, которая предлагает цифровые издания в XML-TEI;
- цифровые издания литературных рукописей и архивных документов XVI века.

***Морфо-синтаксическая база данных цыганского языка RMS***[3] . Манчестерский проект является частью международной сети научных проектов, посвященных исследованиям цыганского языка. В рамках проекта исследуются лингвистические особенности диалектов цыганского языка и их распространение в географическом пространстве. Разрабатывается интерактивное веб-приложение, которое позволит пользователям искать и находить на карте различные диалектные варианты, а также исследовать, как они группируются в конкретных регионах. Также будут доступны примеры предложений и слов со звуковыми файлами, чтобы дать представление о диалектных вариациях в цыганском языке. Из распределения языковых форм между диалектами можно будет сделать выводы о социально-исторических контактах между цыганскими общинами и о характере миграции.

Из представленных примеров различных ресурсов видно, что многоцелевые ресурсы цифровой гуманитаристики с лингвистическим компонентом имеют достаточно разнообразный характер и относятся к различным типам ЛИР. Это корпуса текстов, относящихся к конкретным народам, регионам

[1] Real Academia Española – Corpus de Referencia del Español Actual (CREA). – URL: http://corpus.rae.es/creanet.html (дата обращения: 01.12.2021).

[2] Bibliothèques Virtuelles Humanistes (BVH). – URL: http://www.bvh.univ-tours.fr / (дата обращения: 01.12.2021).

[3] Romani Morpho-Syntax Database (RMS). – URL: https://romani.humanities.manchester.ac.uk//rms/ (дата обращения: 01.12.2021).

или периодам, собрания памятников письменности, структурированные базы данных, фонотеки и прочее. ЛИР используют разнообразный лингвистический инструментарий: разметку, в том числе по принципам TEI, лингвистические аннотации, статистические методы, сопоставительный анализ, графовое представление, извлечение именованных сущностей и другое.

Однако все рассмотренные и многие другие ЛИР имеют общую черту – они используют лингвистические ресурсы и методы для решения задач из разных гуманитарных и социальных дисциплин – истории, этнографии, коммуникативистики, социологии, политологии.

## Российские гуманитарные ресурсы с лингвистическим компонентом

### *Проекты НИУ ВШЭ*

Одним из лидеров российской цифровой гуманитаристики в настоящее время является Центр цифровых гуманитарных исследований НИУ ВШЭ, который делает множество проектов DH. Приведем перечень некоторых из проектов, имеющих прямое отношение к ЛИР и к электронной филологии. Подробности можно найти по адресу[1]. Некоторые проекты доступны, для них указаны адреса.

*Персонажи «Игры Престолов»* (*2016*)[2]

*Совместная встречаемость в русской и английской Википедиях* (*русские классические композиторы*)[3]

*Открытый гуманитарный репозиторий* предназначен для хранения описаний исследования / научной работы и самими данными, и их метаданными (таблицы, текст, картинки), которые прошли отбор редакционной коллегии и получили рейтинг качества.

*Дерево кириллических алфавитов* – интерактивная AR-выставка кириллических алфавитов.

*Цифровая эпиграфика* – разработка базы данных для хранения, презентации и поиска описаний надписей и их 3 D-моделей.

*Цифровой корпус русской драмы в TEI*. Корпус содержит свыше 200 русских пьес, снабженных тщательно выверенной машиночитаемой разметкой в формате Text Encoding Initiative (TEI), а также визуализации пьес в виде социальных сетей (графов) и различные количественные метрики активности персонажей. Корпус RusDracor – часть международной инициативы DraCor по созданию «программируемых» корпусов драмы с машиночитаемыми данными и богатым API.

---

[1] ВШЭ проекты DH. – URL: https://hum.hse.ru/digital/projects (дата обращения: 01.12.2021).

[2] Тюториал по "Игре Престолов" в Ясной Поляне. – URL: https://github.com/nevmenandr/Martin_tutorial (дата обращения: 01.12.2021).

[3] Совместная встречаемость в русской и английской википедиях. – URL: https://github.com/nevmenandr/DigitalHumanitiesMinorFeatures/blob/master/Composers_in_wiki.pdf (дата обращения: 01.12.2021).

*Instagram Л.Н. Толстого.* Проект направлен на цифровизацию нетекстового наследия русского писателя, а именно фотографий.

*Автоматический поиск формульных конструкций в древнеисландских сагах* – алгоритмы поиска и квалификации неколлокационных повторов в корпусе древнеисландских саг.

*Цифровое издание сочинений Елены Шварц* – создание сайта, на котором опубликованы размеченные тексты Елены Шварц со ссылками к: ее личному архиву; «библиотеке» с опубликованными изданиями; комментариям к ее текстам; ее черновикам.

*Осип Мандельштам Digital.* Проект посвящен цифровизации поэтического наследия Осипа Мандельштама и одновременно – научных работ о нем. На сайте стихотворения и их варианты будут связаны с посвященными им публикациями.

*Визуализация «Войны и мира»*. Визуальный и лингвистический анализ романа Льва Толстого «Война и мир», интерактивные графы связей героев романа.

*Tolstoy Digital.* Семантическая разметка электронного издания полного собрания сочинений Л.Н. Толстого с воспроизведением на новом технологическом уровне метатекстовой информации, критического аппарата, комментариев и указателей. Семантическая разметка должна соответствовать стандарту TEI [2]. Для реализации семантической разметки был разработан специальный инструментарий по названием *Тестограф*[1].

***Памятники письменности***

Одним из важных направлений российской цифровой гуманитаристики филологического характера являются оцифровка и представление рукописных и печатных памятников письменности. Одним из лидеров этого направления является проект *Манускриптъ*[2], возглавляемый В.А. Барановым. Сайт проекта содержит коллекции древнейших и средневековых славянских и русских текстов, информацию о разработках в области их электронной публикации и методах создания полнотекстовых баз данных сложных по структуре и составу рукописных памятников. Проект *Манускриптъ* имеет богатый лингвистический инструментарий для работы с древнерусскими и старославянскими текстами. Основными модулями для работы с коллекцией являются:

- сайты и запросные формы, позволяющие познакомиться с текстами, указателями и осуществить выборку данных;
- специализированный редактор для ввода, редактирования и фрагментирования текстов;
- модуль выборок и запросов, позволяющий подготовить данные для лингвистических, палеографических и текстологических исследований;
- морфологический анализатор для автоматического анализа и синтеза словоформ древнерусского языка;
- модуль грамматических словарей для ввода, редактирования и согласования словарных материалов.

[1] Testograf. – URL: https://www.testograf.ru/ru/provedenie/obzor-funkcij/ (дата обращения: 01.12.2021).

[2] Манускриптъ. – URL: http://mns.udsu.ru/ (дата обращения: 01.12.2021).

В России имеется еще несколько информационных систем, специализирующихся на электронном представлении памятников письменности. Все они созданы с активным участием лингвистов и используют лингвистический инструментарий, и поэтому с полным основанием могут быть отнесены к ЛИР. Перечислим некоторые из них:

- Manuscripta Islamica Rossica[1]
- Санкт-Петербургский корпус агиографических текстов[2]
- Древнерусский подкорпус Национального корпуса русского языка[3]
- Памятники древнеславянской письменности[4]
- Древнерусские берестяные грамоты[5]
- Рукописная книга[6]
- Полное собрание русских летописей[7]
- Славяно-русский Пролог по древнейшим рукописям[8]
- Эпиграфическая письменность Древней Руси (XI–XV вв.)[9]

К этому же направлению цифровой гуманитаристики относятся также три проекта, изложенные на научной сессии ОИФН РАН «Гуманитарные науки в эпоху цифровизации» 15.04. 2021 г.[10]:

- создание информационной системы по древнегреческим памятникам в северном Причерноморье, описанное в докладе «Цифровые методы в издании и изучении древних и средневековых надписей» – член-корреспондент РАН А.И. Иванчик;
- проект «История письма европейской цивилизации»: коллекции памятников письменности академических институтов Санкт-Петербурга, оцифровка и изучение – член-корреспондент РАН А.В. Сиренов;
- «Собираемый пазл: реконструкция рукописи до прочтения текста» – член-корреспондент РАН И.Ф. Попова.

[1] Manuscripta Islamica Rossica. – URL: http://manuscriptaislamica.ru/ru (дата обращения: 01.12.2021).

[2] СКАТ. – URL: http://project.phil.spbu.ru/scat/page.php?page=project (дата обращения: 01.12.2021).

[3] Древнерусский корпус. – URL: http://ruscorpora.ru/search-old_rus.html (дата обращения: 01.12.2021).

[4] Памятники древнеславянской письменности. – URL: https://ksana-k.ru/ (дата обращения: 01.12.2021).

[5] Древнерусские берестяные грамоты. – URL: http://gramoty.ru/birchbark/ (дата обращения: 01.12.2021).

[6] Рукописная книга. – URL: http://www.lrc-lib.ru/?id=3 (дата обращения: 01.12.2021).

[7] Полное собрание русских летописей. – URL: http://www.lrc-lib.ru/rus_letopisi (дата обращения: 01.12.2021).

[8] Славяно-русский Пролог по древнейшим рукописям. – URL: http://prolog-manuscript.org/index.php (дата обращения: 01.12.2021).

[9] Эпиграфическая письменность Древней Руси (XI–XV вв.). – URL: https://ilingran.ru/official/2012-2014_oifn_langlit_5.pdf (дата обращения: 01.12.2021).

[10] Гуманитарные науки в эпоху цифровизации. – URL: http://hist-phil.ru/events/427/ (дата обращения: 01.12.2021).

***Филологические ресурсы***

Еще одним направлением создания гуманитарных ресурсов со значительным лингвистическим компонентом являются информационные системы, созданные для решения широкого круга филологических задач. Примером такого рода проектов являются информационные системы, созданные под руководством И.А. Пильщикова и К.В. Вигурского.

Первая и самая значительная по объему – Фундаментальная электронная библиотека «Русская литература и фольклор» (ФЭБ)[1]. Это полнотекстовая информационная система по произведениям русской словесности, библиографии, научным исследованиям и историко-биографическим работам. Основное содержание ФЭБ представляется в электронных научных изданиях (ЭНИ), каждое из которых посвящено отдельному автору (Пушкин, Лермонтов и т.д.), жанру (былины, песни и т.д.) или произведению (например «Слово о полку Игореве»). Кроме произведений русской литературы и литературоведения, ФЭБ включает большое количество справочной информации, библиографии, энциклопедии и словари, разнообразные указатели. ФЭБ стала заметным явлением в области электронной филологии, получила широкую, в том числе международную известность. К сожалению, с 2015 года этот проект заморожен.

В новом проекте тех же авторов «Сравнительная поэтика и сравнительное литературоведение»[2] представлены переводные стихотворные произведения на русском языке, их иноязычные источники, научная литература по сопоставительной поэтике и сравнительному литературоведению, а также словарь-тезаурус терминов этих дисциплин. В рамках текущего проекта материал ограничен русско-романскими литературными связями, а формальное описание текстов – метрикой и строфикой. Информационная система состоит из четырех взаимосвязанных разделов (подсистем): «Корпус» (включает стихотворные переводы, их оригиналы и переводы-посредники), «Библиотека» (издания переводов и их оригиналов, а также научная литература), «Энциклопедия» (биобиблиографические справки о поэтах, переводчиках и исследователях) и «Тезаурус» (термины, встречающиеся в исследовательской литературе).

В этом жанре – научных филологических ресурсов – находится также специализированная информационная система для сравнения переводов «Слова о полку Игореве»[3], которую автор назвал *Электронный инструмент сравнительного изучения текстов*.

В России также начато создание новых видов цифровых ресурсов, которые получили название *Семантические издания* (англ. semantic publishing), или *публикация в семантическом вебе* – размещение информации в Интернете документов, сопровождаемых семантической разметкой. Семантическая публикация (или издание) дает возможность поисковым машинам более точно интерпретировать структуру и смысл опубликованной информации, что делает поиск информации в Интернете и интеграции данных более эф-

[1] Русская литература и фольклор. – URL: http://feb-web.ru/ (дата обращения: 01.12.2021).

[2] Сравнительная поэтика и сравнительное литературоведение. – URL: http://cpcl.feb-web.ru (дата обращения: 01.12.2021).

[3] «Слово о полку Игореве»: Параллельный корпус переводов. – URL: http://nevmenandr.net/slovo/ (дата обращения: 01.12.2021).

фективным. Существенно, что идея и методика семантической разметки заимствована из опыта ЛИР. Стандарты семантической разметки рассмотрены в главе «Стандартизация ЛИР». Подробное описание семантических изданий можно найти в работе [3]. Примером российского семантического издания является упомянутый выше проект *Tolstoy Digital*[1].

Семантические сетевые академические издания стали также центральным направлением работ по цифровизации, проводимых в ИРЛИ РАН [4]. Подробно об этих проектах также говорится в докладе член-корреспондента РАН М.Н. Виролайнен «Pushkin Digital и цифровые проекты Пушкинского дома» на цитированной выше сессии ОИФН [2].

Еще один интересный пример применения цифровой филологии – проект «Цифровой комментарий к древнегреческой комедии: проблемы и перспективы», представленный Б.М. Никольским и А.А. Бонч-Осмоловской (Институт мировой литературы им. А.М. Горького РАН, Высшая школа экономики), также изложенный на этой научной сессии. Проект предполагает создание развитой концептуально-лексикографической БД по лексике древнегреческой комедии с набором переводов, комментариев, толкований и другой вспомогательной информацией.

В данной главе не ставится задача рассмотреть все направления и подходы электронной филологии, и тем более цифровой гуманитаристики в целом. Мы хотели лишь привести примеры того, как идеи и методы цифровых ЛИР используются в проектах смежных гуманитарных наук и становятся общими решениями для цифровой гуманитаристики. Эти примеры служат бесспорным доказательством, что ЛИР являются одним из центральных элементов цифровой гуманитаристики, и их развитие неразрывно связано с развитием этой дисциплины.

## Литература к главе 20


1. Цифровые гуманитарные науки : хрестоматия. – URL: http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf (дата обращения: 01.12.2021).

2. Гуманитарные науки в эпоху цифровизации. Видеотрансляция Общего собрания Отделения историко-филологических наук РАН 15 апреля 2021 года. – URL: http://hist-phil.ru/events/427/ (дата обращения: 01.12.2021).

3. Семантическое издание текстов Л.Н. Толстого: от текста к онтологии / Бонч-Осмоловская А., Колбасов М., Орехов Б., Павлова И., Скоринкин Д. – Москва, 2018. – URL: https://publications.hse.ru/mirror/pubs/share/direct/307083397.pdf (дата обращения: 01.12.2021).

4. Гронас М., Орехов Б. Что такое семантическое издание и почему в будущем все издания станут семантическими. – URL: https://publications.hse.ru/mirror/pubs/share/direct/307083240.pdf (дата обращения: 01.12.2021).

5. Гуськов С.Н. Академические собрания: новый сетевой инструмент филологических исследований / ИРЛИ РАН (Пушкинский Дом) // Единое цифровое пространство научных знаний: проблемы и решения : сборник научных трудов / под ред. Каленова Н.Е., Сотникова А.Н. – Москва ; Берлин : Директмедиа Паблишинг, 2021. – 503 с. – DOI: 10.23681/610687


[1] Tolstoy Digital. Семантическое издание. – URL: http://tolstoy.ru/projects/tolstoy-digital/

# ГЛАВА 21. ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ РОССИЙСКОЙ ИНФРАСТРУКТУРЫ ЛИР

## Вводные замечания

Из проведенного исследования ЛИР с очевидностью следует, что деятельность по созданию и использованию ЛИР в России развернута достаточно широко. В стране создано множество ЛИР по всем направлениям как прикладной, так и теоретической лингвистики и языковой индустрии, а также образовательных ЛИР. Только в учреждениях РАН выявлено свыше тысячи ЛИР различных типов и назначения, которые внесены в Навигатор информационных ресурсов по языкознанию[1].

Напомним, что в этой системе под ЛИР понимались как обычные информационные ресурсы, относящиеся к лингвистике лишь по тематике (библиотечные и архивные фонды, периодика, материалы конференций, энциклопедические данные, диссертации, отчеты, каталоги, сайты и др.), так и специальные ЛИР, к которым отнесены:

- корпуса текстов и звучащей речи
- словарные БД и электронные лексиконы
- памятники письменности (кодикологические БД)
- лингвистические процессоры
- грамматические ресурсы, описания языков
- типологические БД, реестры языков
- лингвистические атласы, ГИС
- этно- и социолингвистические БД
- комплексные лингвистические АИС
- информационные языки

Нужно отметить, что в этой классификации могут отсутствовать ЛИР, которых нет в учреждениях РАН, например учебные ЛИР. Если же оценивать общее количество российских ЛИР, создаваемых в вузах, школах, в информационной индустрии, на любительских сайтах, то их число, вероятно, превысит 10 тыс. Несмотря на широкий фронт работ по созданию ЛИР, координация в этой области развита совершенно недостаточно и требует определенных управляющих усилий.

[1] Навигатор информационных ресурсов по языкознанию. – URL: http://niryaz2.alexo.beget.tech/ (дата обращения: 01.12.2021).

В мировом масштабе координация в сфере ЛИР осуществляется при помощи профессиональных сообществ, которые достаточно эффективно организуют процессы сотрудничества и коллаборации. Особенно это очевидно на примере Рабочей группы по открытой лингвистике Фонда открытого знания по созданию облака LLOD, которая была описана в главе 19. Также очень эффективно идет работа по созданию документации языков, находящихся под угрозой уничтожения, которую поддерживают несколько фондов и международных консорциумов. Авторитетом пользуются языковые архивы и репозитории, особенно интегратор языковых архивов OLAC (см. главу 2).

Особенно большое внимание уделяется координации деятельности по ЛИР в Европе, где созданы специальные координирующие и сервисные структуры в форме инфраструктурных консорциумов. Напомним, что в системе Еврокомиссии создано 18 таких консорциумов по различным направлениям, связанным с цифровизацией науки. Для нас, конечно, особый интерес представляет специализированный инфраструктурный консорциум языковых технологий CLARIN (глава 4).

## Российская ситуация

В России существует множество различных органов по управлению наукой и высокими технологиями (Министерство науки и высшего образования, Министерство цифрового развития, связи и массовых коммуникаций, Российская академия наук, Российский научный фонд, разнообразные научные советы, в том числе при Президенте). В то же время в стране нет ясного представления о необходимой инфраструктуре для поддержки цифровизации науки.

Напомним, что среди материалов по цифровизации науки, которые были разработаны в 2018–2019 гг. и размещены на сайте Минобрнауки[1], есть *Концепция цифровой автоматизированной системы предоставления сервисов научной инфраструктуры коллективного пользования* (АС УСНИКП) и *Концепция создания Единой цифровой платформы научного и научно-технического взаимодействия, организации и проведения совместных исследований в удаленном доступе, в том числе с участием зарубежных ученых* (ЦПСИ).

Однако эти концепции, во-первых плохо отражают существующее положение вещей в России, а во-вторых, кажется, вообще не будут реализовываться, поскольку новое руководство Министерства науки и высшего образования имеет иной взгляд на перспективы развития науки.

Эти вопросы достаточно подробно рассмотрены в предыдущей монографии автора [1] применительно к информационной инфраструктуре в целом. В этой книге автор исходил из представления, что имеется реальная перспектива создания Единого российского электронного пространства знаний, поскольку имеется ряд нормативных документов, регламентирующих создание этого пространства. Все эти документы достаточно подробно прокоммен-

[1] Совет по цифровому развитию и ИТ. – URL: https://minobrnauki.gov.ru/colleges_councils/kollegialnye-organy/digitalcouncil/ (дата обращения: 01.04.2022).

тированы в цитированной книге. Однако события последних лет заставили усомниться в реальности такой перспективы.

В этой же книге изложены некоторые соображения об информационной инфраструктуре языкознания, которая должна быть основой системы координации отрасли языковых технологий. В частности, была предложена модель системы мониторинга ЛИР в форме Навигатора информационных ресурсов по языкознанию, программа создания репозитория лингвистических данных (Центр лингвистических ресурсов), методика валидации (оценки лингвистических ресурсов), а также план создания онтологии для представления содержания и метаданных российских ЛИР.

Исследования мирового опыта индустрии ЛИР, отраженные в настоящей монографии, укрепили автора в мысли о необходимости информационной инфраструктуры для отрасли. Действительно, несмотря на излишнюю, на его взгляд, бюрократизацию процессов управления, инфраструктура отрасли языковых технологий в Европе развита очень хорошо, количество и качество создаваемых ресурсов и сервисов для поддержки отрасли производит сильное впечатление, и этот опыт в значительной степени можно применить в России.

Однако ситуация в России в части инфраструктуры ЛИР далека от идеала.

Приведем несколько примеров.

Действовавшая в 2012–2014 гг. Программа Президиума РАН *Корпусная лингвистика* весьма эффективно способствовала реализации многих проектов: активно развивается НКРЯ, создано множество других корпусов, как для русского, так и для других языков народов России, разработан сайт *Лингвистические корпуса и сервисы*, объединяющий корпусных лингвистов. Таким образом, очевиден позитивный эффект от этой программы.

Теперь противоположный пример. В российском Техническом комитете по стандартизации ТК 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле» разработаны, т.е. переведены на русский язык, несколько стандартов по управлению ЛИР из числа разработанных в ИСО, и их утвердили в качестве национальных стандартов. Выбор стандартов для перевода производит впечатление случайного, а качество переводов при этом чрезвычайно низкое: кажется, что результаты автоматического перевода вообще не редактировались.

Главный же недостаток деятельности российского ТК 55 заключается в том, что разработанные им стандарты вообще не применяются при разработке российских ЛИР. Это неудивительно, ведь в составе этого ТК практически нет разработчиков ЛИР. Исключение только одно – возглавляет ТК 55 институт Стандартинформ, который поддерживает известный российский банк терминологических данных Ростерм. Однако этот банк данных не замечен в активном сотрудничестве с другими разработчиками отечественных ЛИР, а также в разработке открытого доступа и интеграции терминологических данных.

Как мне представляется, деятельность этого комитета не оказывает вообще никакого влияния на российскую индустрию в области ЛИР, хотя в мире и особенно в Европе соответствующие стандарты активно применяются и оказывают значительное влияние на совместимость и повторное использование ЛИР.

Еще один пример. Имеется большой зарубежный опыт сбора, депонирования и архивирования ЛИР. В этих процессах большое внимание уделяется валидации ЛИР, т.е. проверке ЛИР на соответствие стандартам, форматам представления и другим формальным и качественным критериям.

В то же время существующих в России каталогах и порталах ЛИР, описанных в главе 2, отсутствуют какие бы то ни было оценки качества. Единственное исключение – автоматическое измерение Индекса качества сайта при помощи известного сервиса компании Яндекс в Навигаторе информационных ресурсов по языкознанию. Хотя самый беглый анализ показывает, что в числе российских ЛИР – множество дублирующих, неиспользуемых, вообще некачественных.

Отсутствует в России и правовое обеспечение доступности ЛИР, в том числе возможности их коммерческого и некоммерческого повторного использования. Отчасти это объясняется неопределенной официальной позицией российского руководства науки по отношению к концепции открытой науки и проблеме повторного использования научных данных, в том числе полученных в исследованиях, проведенных за счет бюджета.

Все сказанное приводит нас к выводу, что в России настоятельно необходима программа создания информационной инфраструктуры в целом и инфраструктуры для языковых технологий в частности. Далее излагаются некоторые соображения о содержании подобной программы.

### Справочно-информационная система по языкознанию

Одним из центральных элементов информационной инфраструктуры отрасли должна быть система информационного обеспечения принятия решений при создании ЛИР. Далее формулируются предложения по созданию такой системы.

Действующими нормативными документами [2;3] научно-методическое руководство научными исследованиями в России возложено на Российскую академию наук.

Основные задачи, которые должна решать РАН при осуществлении научно-методического руководства, могут быть сформулированы следующим образом:

- разработка программ научных исследований;
- экспертиза заявок на проекты и темы научно-исследовательских работ;
- оценка результатов научно-исследовательских работ, а также научных учреждений и научных подразделений;
- согласование планов проведения научных мероприятий;
- определение перспективных направлений научных исследований;
- выявление дублирования и лакун научных исследований.

Актуальной задачей для лингвистов является создание и поддержка ЛИР. Эта работа, которая ведется в большинстве научных учреждений РАН, в настоящее время никак не координируется. По нашему мнению, научно-методическое руководство в этой области должно решать следующие задачи:

- мониторинг и учет создаваемых ЛИР;
- координация создания ЛИР, включая координацию работ по оцифровке публикаций и созданию научной электронной библиотеки по языкознанию;
- экспертиза и оценка ЛИР;
- подготовка рекомендаций по интеграции и агрегации ЛИР;
- поддержка созданных ЛИР и обеспечение свободного доступа к ним;
- обеспечение архивирования и сохранности ЛИР, в том числе для повторного использования.

Задачи научно-методического руководства и координации должны решаться на профессиональном уровне, что можно обеспечить, только опираясь на специалистов. При этом специалистам, осуществляющим научно-методическое руководство, необходим специальный информационный инструмент. Этот инструмент должен быть тематическим, т.е. учитывать особенности данной области науки. Универсальные общероссийские информационные системы, такие как ЕГИСУ НИОКТР[1], или Система управления НИР[2], не очень подходят для этой цели. К тому же многие данные, необходимые для качественного научно-методического руководства, разбросаны по разным информационным системам или находятся только на сайтах научных организаций.

Поэтому предлагается создать Справочно-информационную систему (СИС) под условным названием *Языкознание в РАН*. Эта система должна дать полное комплексное представление обо всех участниках и результатах научной деятельности в данной области, причем как в традиционных формах – публикации, отчеты, диссертации, так и в электронной – сайты, корпуса, базы данных, другие ЛИР, аккаунты и другое.

Очевидно, эта система должна охватывать все научные исследования в России в данной тематической области, включая вузовскую и прикладную науку. Однако на первом этапе РАН должна навести порядок в собственном доме и обеспечить качественное научно-методическое руководство исследованиями в академическом секторе, отработав соответствующие методики и инструментарий. Очевидно, что для области языкознания организационной формой научно-методического руководства должно быть Отделение историко-филологических наук РАН. Нет сомнения, однако, что в создании и эксплуатации СИС должны принимать участие и институты РАН, прежде всего ИНИОН РАН и ИРЯ РАН, создавшие существенный задел для СИС.

Кроме информационной поддержки научно-методического руководства, предлагаемая СИС, конечно, должна использоваться и для справочно-информационного обслуживания ученых, работающих в области языкознания, например при составлении заявок на гранты или при подготовке проектов программ. В перспективе СИС поможет существенно сократить затраты ученых на подготовку отчетности.

[1] Единая государственная информационная система учета результатов научно-исследовательских и опытно-конструкторских работ гражданского назначения. – URL: https://www.rosrid.ru/ (дата обращения: 01.12.2021).

[2] Система управления научно-исследовательскими работами. – URL: http://wnir.minobrnauki.gov.ru/ (дата обращения: 01.12.2021).

В основу создания СИС должны быть положены результаты, полученные специалистами ИНИОН РАН И ИРЯ РАН в ходе исследований по гранту РФФИ КОМФИ 18–00–00298 «Интеграция научно-информационных ресурсов учреждений РАН по гуманитарным наукам (на примере языкознания) как части единого цифрового информационного пространства РАН».

Одним из результатов этих исследований было создание информационной системы *Навигатор информационных ресурсов по языкознанию* (*НИРЯЗ 2*)[1]. Описание этой системы содержится также в работе и в главе 2. Опыт создания НИРЯЗ 2, а также накопленные в нем данные могут быть положены в основу создаваемой СИС.

Далее формулируются основные требования к создаваемой СИС, ее наполнению и функциональности.

***Общие требования к СИС***

СИС должна создаваться на основе открытых источников. При создании СИС должен быть исключен сбор обязательной дополнительной отчетности от учреждений и научных работников. Администратор СИС, однако, должен иметь право просить учреждения и научных работников проверить и дополнить имеющуюся у него информацию.

Учреждения и научные работники РАН должны иметь техническую возможность с минимальными затратами корректировать и дополнять информацию, размещенную в СИС.

Корректировка и дополнение информации в СИС должна осуществляться в электронном виде через модератора – администратора СИС.

СИС должна быть размещена в Интернете и быть доступна для свободного и бесплатного доступа без регистрации. Информация в СИС должна размещаться под лицензией Creative Commons Attribution (CC-BY) и быть доступной для скачивания и некоммерческого использования со ссылкой на источник.

Актуализация информации, размещенной в СИС, должна осуществляться не реже, чем два раза в год.

Источниками информации для наполнения СИС являются:

- интернет-сайты институций, персон, периодических изданий, мероприятий;
- государственные учетные информационные системы Минобрнауки, ВАК, Рособрнадзора, государственных научных фондов;
- информационные системы открытого доступа;
- данные, предоставляемые добровольно учреждениями и научными работниками.

При наличии противоречий в источниках приоритет отдается данным, предоставляемым учреждениями и научными работниками.

[1] Навигатор информационных ресурсов по языкознанию. – URL: http://niryaz.inion.ru/ (дата обращения: 01.12.2021).

***Состав контента СИС***

В СИС включаются две категории акторов научных исследований (институции и персоны) и десять категорий информационных объектов, отражающих результаты научных исследований.

*Акторы*

1. *Институции*:
   - управленческие структуры
   - учреждения (институты)
   - научные подразделения языковедческого профиля
   - аффилированные структуры
   - научные советы и комиссии,
   - научно-методические советы
   - диссертационные советы
   - ученые советы
   - советы молодых ученых
2. *Персоналии*

Включаются открытые (размещенные в Интернете) сведения о научных работниках, работающих или работавших ранее в учреждениях РАН.

*Информационные объекты:*

- отчеты и сведения о результатах научных исследований
- библиографии
- полнотекстовые книжные издания электронные библиотеки, репозитории, архивы
- периодика
- научные мероприятия (конгрессы, конференции, семинары и т.д.)
- материалы диссертационных советов
- архивная и музейная информация
- медиафонды фото-, аудио-, видео-, кинодокументов
- материалы популяризации науки, учебные материалы
- лингвистические ресурсы и системы

Данная категория включает следующие виды информационных объектов:

- корпуса текстов
- словарные БД и электронные картотеки
- лингвистические процессоры
- грамматические ресурсы
- описания языков, реестры языков
- справочники, энциклопедии
- лингвистические атласы и ГИС
- этно- и социолингвистические БД
- информационные языки
- каталоги интернет-ссылок
- комплексные лингвистические АИС
- сайты лингвистических институций
- аффилированные лингвистические сайты

В заключение укажем, что исходя из существующего распределения функций научных учреждений задача создания и поддержания СИС может быть возложена на ИНИОН РАН.

## Стратегия информационной инфраструктуры языковых технологий и ресурсов

Прежде всего нужна стратегия развития ЛИР, но только не финансовая, а именно ориентирующая, рассчитанная на руководителей учреждений, формирующих различные научные и образовательные программы.

В этой программе желательно определить, какие ЛИР и сервисы имеет смысл централизовать, а какие должны формироваться и поддерживаться на местах. При этом очевидно, что централизация может быть реализована на различных уровнях, например только на уровне метаданных. Централизованные сервисы также желательно распределить по разным учреждениям и городам, как это сделано в CLARIN.

Для тех ЛИР и сервисов, централизация которых кажется предпочтительной, необходимо определить, имеет ли смысл делать это на национальном уровне, или разумней присоединиться к мировому или европейскому сервису.

Например, если речь идет о лингвистических связанных открытых данных, очевидно, что уже созданное облако LLOD является необходимым и достаточным инструментом, и создавать ему альтернативу нет никакой необходимости.

Или другой пример. В Европе многие лингвистические структуры, (например Ontolex) которые создают концептуальные и / или энциклопедические данные, формируют специальные зоны в Википедии, где размещаются сведения, которые данное сообщество считает правильными. Думаю, что для русскоязычных лингвистических терминов можно то же самое сделать с русской Википедией.

Однако коллаборация при создании и поддержке централизованных ресурсов и сервисов в российских условиях эффективна, когда ресурс делается с подключением ведущих академических учреждений и университетов, как это делалось например при создании НКРЯ. А это, в свою очередь, требует разработки и реализации системы мотивации и вознаграждения (причем, далеко не всегда финансового) участия в коллаборации отдельных ученых и научных учреждений. Конечно, в наших условиях, когда фактически единственным инструментом оценки качества и эффективности научной деятельности стал пресловутый Комплексный балл публикационной активности, такой подход выглядит утопией. Напомню, однако, что все современные декларации по развитию науки и ее инфосферы, начиная от декларации DORA и вплоть до последних рекомендаций ЮНЕСКО по открытой науке, единодушно призывают изменить систему оценки научной деятельности. При этом особое внимание нужно обратить на учет научных результатов в форме открытых научных данных, ориентированных на обмен и повторное использование. Очевидно, что к области ЛИР это относится в полной мере.

Вероятно, лучшей современной формой для реализации ЛИР как открытых научных данных было бы размещение их в облаке LLOD.

Необходимо пересмотреть свое отношение к стандартизации ЛИР. С одной стороны, стандарты должны соответствовать реальным потребностям отрасли (сейчас это совершенно не так). С другой – следует потребовать от разработчиков ЛИР реального соблюдения этих стандартов, что должно быть зафиксировано в проектах, заявках на грант, экспертных заключениях, в общем во всей документации, связанной с разработкой ЛИР.

Очень важными для реализации стратегии являются финансовые аспекты. Дело в том, что большинство ЛИР, создаваемых научными организациями России, создается за счет научных фондов. Однако необходимо поставить вопрос о специальной форме грантов, направленных на поддержку открытых ЛИР, которые могут использоваться как наукой, так и промышленностью. Эти ЛИР требуют финансовой поддержки не только на этапе создания, но и для постоянного пополнения и развития. К сожалению большое число ЛИР, созданных за счет грантов, далее не поддерживаются, что во многих случаях означает их гибель. Соответственно нужно менять и правила экспертизы заявок на получение грантов для создания или модернизации ЛИР.

Следует определить состав организационных структур, образующих информационную инфраструктуру отрасли, которые должны функционировать за счет бюджета и сервисы которых должны быть бесплатно доступны для всех участников. Центральным элементом информационной инфраструктуры отрасли должна быть СИС, функционирующая за счет бюджета.

Конечно, нужен централизованный архив ЛИР, который тоже должен получать отдельное и постоянное финансирование, обеспечивая доступ к ранее созданным ЛИР и прием вновь создаваемых. Зарубежный опыт таких архивов весьма велик, достаточно посмотреть на архивы, входящие в консорциум OLAC[1].

Таким образом, проведенное исследование свидетельствует о необходимости организовать в России создание и устойчивую поддержку инфраструктуры ЛИР. Причем эта инфраструктура должна создаваться на основе коллаборации, как внутри России, так и с зарубежными коллегами и проектами.

## Литература к главе 21

1. Научная информация и электронное пространство знаний : монография / Антопольский А.Б. ; под науч. ред. Ефременко Д.В. – Москва : ИНИОН РАН, 2020. – 252 с.

2. Федеральный закон № 253-ФЗ «О Российской академии наук, реорганизации государственных академий наук и внесении изменений в отдельные законодательные акты Российской Федерации» (в ред. Федеральных законов от 29.07.2017 N 219-ФЗ и от 19.07.2018 N 218-ФЗ). – URL: http://www.consultant.ru/document/cons_doc_LAW_152351/

3. ПОЛОЖЕНИЕ о научно-методическом руководстве РАН научными организациями и образовательными организациями высшего образования. Приложение к постановлению президиума РАН от 17 марта 2015 г. № 45). – URL: https://base.garant.ru/71422984/

[1] Open Language Archives Community. – URL : http://olac.ldc.upenn.edu/ (дата обращения 01.04.2022).

## УКАЗАТЕЛЬ АКРОНИМОВ

Цифра после акронима указывает № приложения.

## РУССКИЕ (КИРИЛЛИЧЕСКИЕ) СОКРАЩЕНИЯ

БД ЯМ База данных «Языки мира»

ГИС Географические информационные системы

ЕГИСУ НИОКТР Единая государственная информационная система учета результатов научно-исследовательских и опытно-конструкторских работ гражданского назначения

ИПЯ Информационно-поисковый язык

КРУТ Корпус русских учебных текстов

ЛЕ Лексическая единица

ЛИР Лингвистический информационный ресурс

ЛЛЖЯ Лаборатории лингвистики жестового языка

МКС Международный классификатор стандартов

МСБО Международное стандартное библиографическое описание

НКРЯ Национальный корпус русского языка

НЭЗД Национальный электронный звуковой депозитарий

ОЛА Общеславя́нский лингвисти́ческий а́тлас

ОПТЕЛ Онтология поисковых терминов по лингвистике

ОЭСР Организация экономического сотрудничества и развития

РЖЯ Русский жестовый язык

РКИ Русский как иностранный

Ростерм Банк данных «Российская терминология»

САРГАС-БД Аудиокорпус речевых и неречевых акустических событий

СКАТ – Санкт-Петербургский корпус агиографических текстов

ТБД Терминологические базы данных

ТБЗ НТ Терминологическая база знаний «Научная терминология»

ТИПБД Типологические БД

УИС РОССИЯ Университетская информационная система РОССИЯ

ФЭБ Фундаментальная электронная библиотека «Русская литература и фольклор»

ЭОР Электронные образовательные ресурсы

# ПРИЛОЖЕНИЯ

## ПРИЛОЖЕНИЕ 1. КАТАЛОГИ, АРХИВЫ И РЕПОЗИТАРИИ ЛИР

(дата обращеия – 01.04.2022)

Academia Sinica Collections https://ndaip.sinica.edu.tw/en_3-1-2-7.html

ACL Anthology – A Digital Archive of Research Papers in Computational Linguistics https://www.semanticscholar.org/paper/ACL-Anthology-A-Digital-Archive-of-Research-Papers-Kan/52356c0107738e6849a3a4e0f81b81baaff941d9

AfBo (A world-wide survey of affix borrowing) http://afbo.info/

AFLAT (African Language Technologies) /http://www.alt-i.org/

ALMA (African Language Materials Archive) http://alma.matrix.msu.edu/

AILLA (Archive of the Indigenous Languages of Latin America) http://www.ailla.utexas.org/

Alaska Native Language Archive http://www.uaf.edu/anla

ALORA (CERDOTOLA – Centre international de recherche et de documentation sur les traditions et les langues africaines) https://en.cerdotola.org/alora

ANPERSANA bibliothèque numérique https://anpersana.iker.univ-pau.fr

APiCS Online (Atlas of Pidgin and Creole Language Structures) http://apics-online.info/

APS Library (The Library of the American Philosophical Society) https://www.amphilsoc.org/library#:~:text=The%20American%20Philosophical%20Society%20Library,and%20linguistics%2C%20and%20digital%20innovation

ARCHE (A Resource Centre for the HumanitiEs) https://arche.acdh.oeaw.ac.at/browser/

Arquivo.pt – Arquivo da Web Portuguesa http://www.arquivo.pt

ASEDA (Aboriginal Studies Electronic Data Archive) https://aseda.aiatsis.gov.au/aseda_main.php

AusNC (Australian National Corpus) https://www.ausnc.org.au/

BAS Repository (Bavarian Archive for Speech Signals, BAS CLARIN Repository, Bayerisches Archiv für Sprachsignale) https://clarin.phonetik.uni-muenchen.de/BASRepository

BathSPAdata (Bath Spa University figshare) https://bathspa.figshare.com/

Best Language Websites https://sites.uni.edu/becker/

BFO (Basic Formal Ontology) https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecBFO

BowPed TRPS Data http://repository.edition-topoi.org/collection/TRPS

Buckeye Speech Corpus http://buckeyecorpus.osu.edu/

Burckhardt Source (The European correspondence to Jacob Burckhardt) https://burckhardtsource.org/

BVH (Les Bibliothèques Virtuelles Humanistes, Humanistic Virtual Libraries) http://www.bvh.univ-tours.fr/

California Language Archive http://cla.berkeley.edu

CEDIFOR (Repository CLARIN-D Centre) https://www.cedifor.de/repository-clarin-d-centre-cedifor-2

C'ek'aedi Hwnax, the Ahtna Regional Linguistic and Ethnographic Archive http://hdl.handle.net/10125/4538

CELR (Center of Estonian Language Resources, Eesti Keeleressursside Keskus) META-SHARE https://keeleressursid.ee/en/resources

Central Institute of Indian Languages: Publications https://www.ciil.org/pubbook.aspx

CHILDES (Child Language Data Exchange System) Data repository http://childes.talkbank.org/

CLAPOP (The Dutch CLARIN Portal Pages) http://portal.clarin.nl

CLARIN Center BBAW (The CLARIN service center of the Zentrum Sprache at the BBAW) https://clarin.bbaw.de

CLARIN Center Tübingen repository (The CLARIN repository at the University of Tübingen) https://uni-tuebingen.de/en/134314

CLARIN Centre Vienna (CCV, Language Resources Portal, LRP, CLARIN-AT) https://clarin.oeaw.ac.at/ccv/.

CLARIN Portal INT Center (CLARIN IvdNT-portaal, CLARIN portal of the Dutch Language Institute) https://portal.clarin.inl.nl/

CLARIN Resource Families https://www.clarin.eu/resource-families

CLARIN.SI (Slovenian language resource repository) http://www.clarin.si/

CLARIN: el inventory of language resources and services https://inventory.clarin.gr/

CLARIN-DK-UCPH Repository (The CLARIN Centre at the University of Copenhagen) https://repository.clarin.dk/repository/xmlui/

CLARIN-ERIC (Common Language Resources and Technology Infrastructure – European Research Infrastructure Consortium) https://www.clarin.eu/

CLARIN-LT (Lithuania) http://clarin-lt.lt/?lang=en

CLARINO Bergen Centre repository https://repo.clarino.uib.no/xmlui/

CLARIN-PL (Language Technology Centre, Centrum Technologii Jezykowiych) https://clarin-pl.eu/dspace/

CLARIN-UK (United Kingdom) https://www.clarin.ac.uk/

CNRTL (Centre National de Ressources Textuelles et Lexicales) https://www.cnrtl.fr/

CoCoON (COllections de COrpus Oraux Numériques, ex-CRDO) http://cocoon.huma-num.fr/exist/crdo/

Codex Sinaiticus Experience the oldest Bible http://www.codexsinaiticus.org/en/

Comparative Corpus of Spoken Portuguese Projeto Tycho Brahe https://www.tycho.iel.unicamp.br/home

CoRSAL (The Computational Resource for South Asian Languages) https://corsal.unt.edu/

Dades DD Dipòsit Digital (Dipòsit Digital de la Universitat de Barcelona) https://www.uab.cat/web/our-collections/uab-digital-repository-of-documents-1345777080660.html

DAIS (Digital Archive of the Serbian Academy of Sciences and Arts, Digitalni arhiv izdanja SANU) https://dais.sanu.ac.rs

DaSCH (Data and Service Center for humanities) http://data.dasch.swiss

DeReKo (Das Deutsche Referenzkorpus, Das Portal für die Korpusrecherche in Textkorpora des Instituts für Deutsche Sprache, The Mannheim German Reference Corpus) https://www.ids-mannheim.de/digspra/kl/projekte/korpora/

DGD (Datenbank Gesprochenes Deutsch, DGD2 (formerly), FDZ AGD, Forschungsdatenzentrum Archiv für Gesprochenes Deutsch am Institut für Deutsche Sprache Database for Spoken German) https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome

Dictionaria http://dictionaria.clld.org/

Digital Himalaya project http://www.digitalhimalaya.com/

D-PLACE (Database of Places, Language, Culture and Environment) https://d-place.org

DSAL (Digital South Asia Library) http://dsal.uchicago.edu/

DTA (Deutsches Textarchiv) http://www.deutschestextarchiv.de/

ELAR (The Endangered Languages Archive) https://www.elararchive.org/

ELP (English Lexicon Project) https://elexicon.wustl.edu/

ELRA Catalogue of Language Resources http://catalogue.elra.info/

ELRC (European Language Resource Coordination) -SHARE Language Resources (LR) Repository https://elrc-share.eu/

Endangered Languages Archive https://www.elararchive.org/

Ethnologue: Languages of the World http://www.ethnologue.com

Eurac Research CLARIN Centre http://clarin.eurac.edu/

GAMS (Humanities' Asset Management System, Geisteswissenschaftliches Asset Management System) http://gams.uni-graz.at/context:gams

GerManC project (A representative historical corpus of German 1650–1800) https://www.alc.manchester.ac.uk/modern-languages/research/german-studies/germanc/

Glottolog 4.5 http://glottolog.org/

HDC (Humanities Data Centre) http://humanities-data-centre.de/

HunCLARIN (A HunCLARIN tagjai) http://clarin.hu/content/hunclarin-tagjai

Huygens ING https://www.huygens.knaw.nl/?lang=en

HZSK Repository (Hamburger Zentrum für Sprachkorpora Repositorium, Hamburg Centre for Speech Corpora Digital Repository) https://corpora.uni-hamburg.de/hzsk/en/repository-search

IANUS (Datenportal Digitaler Forschungsdaten aus Archäologie & Altertumswissenschaften) https://www.ianus-fdz.de/datenportal/

IDR (Informatics Research Data Repository) https://www.nii.ac.jp/dsc/idr/en/index.html

IDS Repository (IDS-Mannheim Repository, Leibniz-Institut für Deutsche Sprache Repository) http://repos.ids-mannheim.de/

ILC-CNR for CLARIN-IT repository http://www.clarin-it.it/

Ilovelanguages https://ilovelanguages.org/

IMS Universität Stuttgart Repository (IMS Fedora Repository, Repository of the CLARIN- Islandora http://clarin04.ims.uni-stuttgart.de/repo/

IULA UPF OAI Archive http://www.language-archives.org/archive/iula.upf.edu

Kaipuleohone http://scholarspace.manoa.hawaii.edu/handle/10125/4250/

Kielipankki (The Language Bank of Finland) https://www.kielipankki.fi/language-bank/

LAC (Language Archive Cologne) https://lac.uni-koeln.de

Language Commons Language Corpora http://www.archive.org/details/Language Commons

Language Documentation and Conservation http://scholarspace.manoa.hawaii.edu/handle/10125/310/

Language Resource Inventory LINDAT/CLARIAH-CZ https://lindat.cz/

Language resources at the Text Laboratory https://www.hf.uio.no/iln/om/organisasjon/tekstlab/

LAPSyD (Lyon-Albuquerque Phonological Systems Database) http://www.lapsyd.ddl.cnrs.fr/lapsyd/

LAUDATIO Long-term Access and Usage of Deeply Annotated Information https://www.laudatio-repository.org/

LDC (Linguistic Data Consortium) https://www.ldc.upenn.edu/

Leipzig Corpora Collection http://clarin.informatik.uni-leipzig.de/repo/

Linghub A comprehensive location for finding information about language resources http://linghub.org/

LINGUIST List https://linguistlist.org/

Linguistic Linked Open Data http://linguistic-lod.org/

Linguistics, Natural Language, and Computational Linguistics Meta-index https://nlp.stanford.edu/links/linguistics.html

List of resources by language https://aclweb.org/aclwiki/List_of_resources_by_language

Living Archive of Aboriginal Languages (LAAL) http://laal.cdu.edu.au/

Lund University Humanities Lab Archive https://corpora.humlab.lu.se

Magoria Books Carib and Romani Archive http://archive.magoriabooks.com/

Meertens Instituut Collecties (Meertens Institute Collections, De Digitale Koepel) http://www.meertens.knaw.nl/cms/en/

University of Melbourne data repository https://melbourne.figshare.com/

**MeSH** Медицинский тезаурус Medical Subject Headings https://www.nlm.nih.gov/mesh/meshhome.html

META-SHARE The open language resource exchange facility http://metashare.elda.org/

MICASE (The Michigan Corpus of Academic Spoken English) http://quod.lib.umich.edu/m/micase/

MMSH (Maison méditerranéenne des sciences de l'homme, Mediterranean Research Centre for the Humanities) Phonothèque http://phonotheque.mmsh.huma-num.fr/

MULCE (Multimodal Learning and Teaching Corpus (LETEC) Exchange, MUltimodal contextualized Learner Corpus Exchange) http://lrl-diffusion.univ-bpclermont.fr/mulce2/
accesCorpus/accesCorpusMulce.php

Native American Languages Collection https://samnoblemuseum.ou.edu/collections-and-research/native-american-languages/

OAI (Open Archives Initiative) https://www.openarchives.org/

ODIN (The Online Database of Interlinear Text) https://odin.linguistlist.org/

OLAC Coverage http://www.language-archives.org/documents/coverage.html

OLAC. Participating Archives http://www.language-archives.org/archives

ORDO (Open Research Data Online, The Open University Data Repository) https://ordo.open.ac.uk/

ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue, Open Resources and TOols for LanGuage) https://www.ortolang.fr/

OTA (The University of Oxford Text Archive) https://ota.bodleian.ox.ac.uk/repository/xmlui/

Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library https://manoa.hawaii.edu/library/research/collections/hawaiian-pacific/

Pangloss Collection https://pangloss.cnrs.fr/?lang=en

PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) Catalog http://catalog.paradisec.org.au

PELIC (The University of Pittsburgh English Language Institute Corpus) -dataset https://github.com/ELI-Data-Mining-Group/PELIC-dataset

PhA (Phonogrammarchiv) https://www.oeaw.ac.at/phonogrammarchiv/

PHOIBLE 2.0 (Repository of cross-linguistic phonological inventory data) http://phoible.org/

POLLEX-Online (Polynesian Lexicon Project Online) http://pollex.org.nz

PolMine Project https://polmine.github.io/

PORTULAN CLARIN repository https://portulanclarin.net/

QMU eData Repository (Queen Margaret University eData Repository) https://eresearch.qmu.ac.uk/handle/20.500.12289/4

RE3 Registry of research data repositories https://www.re3data.org/

Repository CLARIN-D Centre Leipzig (CLARIN-D repository at the University of Leipzig), https://repo.clarin.informatik.uni-leipzig.de/en

RMS (Romani Morpho-Syntax Database) https://romani.humanities.manchester.ac.uk/rms/

RWAAI The Raoul Wallenberg Institute of Human Rights at Lund University https://rwi.lu.se

SADiLaR (South African Centre for Digital Language Resources) https://www.sadilar.org/

SAILS Online (South American Indigenous Language Structures) http://sails.clld.org/

SIL Language and Culture Archives https://www.sil.org/resources/language-culture-archives

SinMin (Texts of different genres and styles of the modern and old Sinhala language) https://osf.io/a5quv/

SLAAP (The Sociolinguistic Archive and Analysis Project) https://slaap.chass.ncsu.edu/

SLDR (Speech and Language Data Repository) https://portal.issn.org/resource/ISSN/2429-6252 https://www.re3data.org/repository/r3d100010828

Spanish CLARIN Knowledge Centre http://www.clarin-es-lab.org/index-es.html

Sprachatlas Baden-Württemberg https://escience-center.uni-tuebingen.de/escience/sprachatlas/#8/48.676/8.992

Språkbanken (The Swedish Language Bank) https://spraakbanken.gu.se/eng

Språkbanken (Speech & Language Data Bank, Språkbankens ressurskatalog)

St. Edward's University institutional repository, St. Edward’s University Figshare https://stedwards.figshare.com/

Standing Rock Sioux Tribe Language and Culture Institute https://www.delaman.org/members/standing-rock-sioux-tribe-language-and-culture-institute/

Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources https://nlp.stanford.edu/links/statnlp.html

SWE-CLARIN https://sweclarin.se/eng/home

SWELANG (CLARIN Knowledge Centre for the Languages of Sweden, Clarin kunskapscentrum) http://www.sprakochfolkminnen.se/om-oss/forskning/sprakbanken-sam/clarin-kunskapscentrum/swelang.html

TalkBank Data repository http://talkbank.org/

Tekstlaboratoriet (tekstlab, The Text Laboratory) http://www.hf.uio.no/iln/english/about/organization/text-laboratory/

TextGrid Repository (Virtuelle Forschungsumgebung für die Geisteswissenschaften, Virtual research environment for the Humanities) https://www.textgridrep.org/

The Crúbadán Project (Corpus Building for Minority Languages) https://cs.slu.edu/~scannell/pub/wac3.pdf

The Eclectic Company Language & Linguistics http://www-personal.umich.edu/~jlawler/lingmarks.html

The LDC Corpus Catalog http://catalog.ldc.upenn.edu/

The LINGUIST List Language Resources http://linguistlist https://linguistlist.org/

The Natural Language Software Registry http://www.lrec-conf.org/proceedings/lrec2000/pdf/267.pdf

The Polinsky Language Sciences Lab Dataverse https://dataverse.harvard.edu/dataverse/polinsky

The Rosetta Project (A Long Now Foundation Library of Human Language) http://www.rosettaproject.org/

The speech-language resources http://www.speechlanguage-resources.com/

The Typological Database System https://portal.clarin.nl/node/1920

Tibetan and Himalayan Digital Library https://www.thlib.org/

TLA (The Language Archive) https://tla.mpi.nl/

TRACTOR https://www.slideserve.com/Gabriel/tractor

TransNewGuinea – database of the languages of New Guinea http://transnewguinea.org/

TROLLing (Tromsø Repository of Language and Linguistics) https://trolling.uit.no

TST-Centrale (De Centrale voor Taal- en Spraaktechnologie or TST-Centrale) http://www.tst-centrale.org

U Bielefeld Language Archive http://www.spectrum.uni-bielefeld.de/langdoc/

UCLA Phonetics Lab Archive http://archive.phonetics.ucla.edu/

UdS Fedora Commons Repository http://fedora.clarin-d.uni-saarland.de/index.en.html

UK RED (UK Reading Experience Database) http://www.open.ac.uk/Arts/reading/UK/

UniLang Online (Многоязычный веб-сайт для работы с ЛИР) https://forum.unilang.org/

University of Guelph Dataverse (University of Guelph Research Data Repository Dataverse) https://dataverse.scholarsportal.info/dataverse/ugrdr

VLO CLARIN (CLARIN Virtual Language Observatory, Виртуальная поисковая система CLARIN) https://vlo.clarin.eu

WALS Online (World Atlas of Language Structures Online) https://wals.info/

Webonary Sites https://www.webonary.org

WOLD (The World Loanword Database) http://wold.clld.org/

World Oral Literature Project http://www.oralliterature.org/

YAGO (Еще одна большая онтология, Yet Another Great Ontology) https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

## ПРИЛОЖЕНИЕ 2. РОССИЙСКИЕ КАТАЛОГИ ЛИР

Lexilogos. Русские словари для онлайн-перевода https://www.lexilogos.com/english/russian_dictionary.htm

Linguists – Ресурсы для переводчиков и лингвистов http://linguists.narod.ru/catalogue.html

NLPub – каталог ресурсов для обработки естественного языка https://nlpub.ru/

Архив петербургской русистики www.ruthenia.ru/apr/index.htm

Ассоциация лингвистов-экспертов Юга России http://www.ling-expert.ru/links.html

Веб-сайты филологической и лингвистической тематики http://it.lang-study.com/veb-sajty-filologicheskoj-i-lingvisticheskoj-tematiki/

Всё о языках, лингвистике, переводе... http://linguistic.ru

Информационные ресурсы по лингвистике http://homepages.tversu.ru/%7Eips/InfoSeek.htm

КАТАЛОГ «НАУКА В РУНЕТЕ» / Лингвистика https://elementy.ru/catalog/t123/Lingvistika

Каталог интернет-ресурсов по РКИ http://www.russischlehrer.at/fileadmin/Veranstaltungen/internetquellen_rki.pdf

Каталог лингвистических программ и ресурсов в Сети https://rvb.ru/soft/catalogue/index.html

Каталог лингвистических программ и ресурсов в Сети / Каталог программ для писателей и журналистов http://ru-writer.blogspot.com/2010/06/c.html

Каталог лингвистических программ и ресурсов в Сети. Программы анализа и лингвистической обработки текстов https://helpiks.org/1-109079.html

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА https://elementy.ru/catalog?type=38

Компьютерная лингвистика. Портал знаний https://uniserv.iis.nsk.su/cl/

Лингвистика http://filologia.su/lingvistika/

Лингвистика в России: ресурсы для исследователей https://archive.vn/9Fu4X#selection-183.2-185.27

Лингвистические интернет-ресурсы https://didacts.ru/termin/lingvisticheskie-internet-resursy.html

Лингвистические корпуса и сервисы http://web-corpora.net/

Лингвистические ресурсы https://www.sites.google.com/site/lingvistika586/lingvisticeskie-resursy

Лингвистические ресурсы Интернета https://poisk-ru.ru/s71285t1.html

Лингвистические ресурсы. Каталог лингвистических программ и ресурсов в Сети / Linguistics Software http://rykov-ling.narod.ru/resurs.html

Лингвистический энциклопедический словарь http://tapemark.narod.ru/les/

Многоязычные сайты и универсальные списки ссылок http://linguodiversity.narod.ru/Links/multlang.htm

Навигатор информационных ресурсов по языкознанию http://niryaz2.alexo.beget.tech/

Общая филология (интернет-ресурсы) http://yspu.org/

Общие ресурсы по лингвистике и филологии http://www.garshin.ru/linguistics/linguistic-portals.html

Онлайн-ресурсы для работы переводчика https://lingvadiary.ru/?p=199

Продукты Центра речевых технологий https://www.speechpro.ru/product/

РОССИЙСКАЯ ЛИНГВИСТИКА (RUSLING) http://rusling.narod.ru/index.htm

Русский язык http://www.philology.ru/linguistics2.htm

Русский язык: Энциклопедия русского языка https://russkiyyazik.ru/

Специализированные ресурсы по лингвистике https://studopedia.ru/13_129557_spetsializirovannie-resursi-po-lingvistike.html

Справочные интернет-ресурсы http://www.oshibok-net.ru/for-all/sites/210/

Тематические порталы, сайты / Языкознание http://library.altspu.ru/lang.phtml

Филологические ссылки. Лингвистика, языкознание http://konfcsu.narod.ru/ze/links.html

Филологический портал Philology.ru http://www.philology.ru/

Электронные ресурсы http://www.ling-theory.ru/links

Электронные ресурсы. Языкознание www.lib.tsu.ru

Энциклопедия русского языка https://ksana-k.ru/?p=1941

Языковые порталы и сайты http://library.mrsu.ru/content/tags/linguistics.html

# ПРИЛОЖЕНИЕ 3. МЕЖДУНАРОДНЫЕ ОРГАНИЗАЦИИ В ОБЛАСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ЦИФРОВОЙ ГУМАНИТАРИСТИКИ

**AADH** Австралийская ассоциация цифровых гуманитарных наук Australasian Association for Digital Humanities https://aa-dh.org/

**AAMT** Азиатско-Тихоокеанская ассоциация машинного перевода Asia-Pacific Association for Machine Translation https://aamt.info/

**ACH** Ассоциация компьютеров и гуманитарных наук Association for Computers and the Humanities https://ach.org/about-ach

**ACL** Ассоциация компьютерной лингвистики Association for Computational Linguistics https://www.aclweb.org/portal

**ADHO** Альянс организаций цифровой гуманитаристики Alliance of Digital Humanities Organizations https://wiki2.org/en/Alliance_of_Digital_Humanities_Organizations

**AFNLP** Азиатская федерация по обработке естественного языка Asian Federation of Natural Language Processing http://www.afnlp.org/wp/

**AILA** Международная Ассоциация прикладной лингвистики Association Internationale de Linguistique Appliquée. https://aila.info/

**AMTA** Ассоциация машинного перевода в Северной и Южной Америке Association in the Americas https://amtaweb.org/

**CEF** Фонд поддержки Европейской инфраструктуры Connecting Europe Facility https://ec.europa.eu/inea/en/connecting-europe-facility

**CenterNet** Международная сеть центров цифровых гуманитарных наук An international network of digital humanities centers https://dhcenternet.org/

**CHAIN** Коалиция гуманитарных и художественных инфраструктур и сетей Coalition of Humanities and Arts Infrastructures and Networks https://mith.umd.edu/news/chain/

**CHCI** Консорциум гуманитарных центров и институтов The Consortium of Humanities Centers and Institutes http://chcinetwork.org/

**CLARIN** Общеевропейская исследовательская инфраструктура для языковых ресурсов и технологий Common European Research Infrastructure for Language Resources and Technology https://www.clarin.eu/

**COCOSDA** Международный Комитет по координации и стандартизации речевых баз данных и методов оценки International Committee for the Coordination & Standardisation of Speech Databases and Assessment Techniques http://www.cocosda.org/

**COLING** Международная конференция по компьютерной лингвистике International Conference on Computational Linguistics https://www.aclweb.org/anthology/venues/coling/

**CSDH/SCHN** Канадское общество цифровых гуманитарных наук Canadian Society for Digital Humanities/société canadienne des humanités numériques https://csdh-schn.org/

**DAISY Consortium** Консорциумпо созданию цифровых аудиокниг https://daisy.org/

**DARIAH** Цифровая инфраструктура искусства и гуманитарных исследований The Digital Research Infrastructure for the Arts and Humanities (DARIAH) https://www.dariah.eu/

**DBpedia Association** Ассоциация для поддержки проекта и сообщества DBpedia https://www.dbpedia.org/about-the-association/

**DELAMAN** Сеть архивов цифровых языков и музыки, находящихся под угрозой исчезновения The Digital Endangered Languages and Musics Archives Network https://www.delaman.org/about/

**EACL** Европейское отделение Ассоциации компьютерной лингвистики Association for Computational Linguistics, European Chapter: http://eacl.org/

**EADH** Европейская ассоциация цифровой гуманитаристики European Association for Digital Humanities https://eadh.org /

**EAGLES** Консультативная группа экспертов по стандартам языковых технологий Expert Advisory Group on Language Engineering Standards http://www.ilc.cnr.it/EAGLES/home.html

**EAMT** Европейская ассоциация машинного перевода European Association for Machine Translation http://eamt.org/

**EIROforum** Соглашение о сотрудничестве для объединения ресурсов, возможностей и опыта организаций-членов для поддержки европейской науки Collaboration agreement to combine resources, facilities and expertise of its member organisations
to support European science https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eiroforum_en

**ELDA** Агентство по оценке и распространению ЛИР The Evaluations and Language resources Distribution Agency http://www.elra.info/en/about/elda/

**ELF** Фонд исчезающих языков The Endangered Language Fund https://ogmios.org/

**ELG** Европейская языковая сеть European Language Grid http://www.elra.info/en/projects/current-projects/european-language-grid/

**ELRA** Европейская ассоциация лингвистических ресурсов European Language Resources Association http://www.elra.info/en/

**EOSC** Европейское открытое научное облако European Open Science Cloud https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en

**ERA** Европейское исследовательское пространство European Research Area, https://ec.europa.eu/info/research-and-innovation_en

**ERF-AISBL** Ассоциация исследовательских инфраструктур Европы The Association of European-Level Research Infrastructures Facilities https://erf-aisbl.eu/

**ERIC** Европейский консорциум научной инфраструктуры European Research Infrastructure Consortium https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en

**ESFRI** Европейский стратегический форум по научным инфраструктурам European Strategy Forum on Research Infrastructures https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/esfri_en

**FLaReNe**t Сеть содействия языковым ресурсам Fostering Language Resources Network https://www.yumpu.com/en/document/read/30185549/fostering-language-resources-network-the-flarenet-european-

**GALA** Ассоциация глобализации и локализации The Globalization and Localization Association https://www.gala-global.org/about/about-gala

**GSF** Глобальный научный форум ОЭСР OECD Global Science Forum https://www.oecd.org/sti/inno/global-science-forum.htm

**H-Net** Гуманитарные и социальные науки онлайн Humanities and Social Sciences Online https://networks.h-net.org/node/513/pages/59033/mission-statem

**Humanistica** Франкоязычная ассоциация цифровой гуманитаристики L'association francophone des humanités numériques/digitales http://www.humanisti.ca/

**IAM**T Международная ассоциация машинного перевода International Association for Machine Translation http://eamt.org/international-association-for-machine-translation/

**IASA** Международная ассоциациция аудио- и видеоархивов International Association of Sound and Audiovisual Archives https://www.iasa-web.org/

**ICCL (COLING**) Международный комитет компьютерной лингвистики International Committee on Computational Linguistics (ICCL) https://wiki2.org/en/International_Committee_on_Computational_Linguistics

**IEEE** Институт инженеров электротехники и электроники Institute of Electrical and Electronics Engineers https://www.ieee.org/

**IFLA** Международная федерация библиотечных ассоциаций и учреждений The International Federation of Library Associations and Institutions http://www.ifla.org/

**IQLA** Международная ассоциация количественной лингвистики The International Quantitative Linguistics Association http://www.iqla.org

**ISCA** Международная ассоциация речевых коммуникаций International Speech Communication Association https://www.isca-speech.org/iscaweb/index.php/about-isca

**ISO** Международная организация по стандартизации International Organization for Standardization https://www.iso.org/ru/home.html

**JADH** Японская ассоциация цифровых гуманитарных наук Japanese Association for Digital Humanities https://www.jadh.org

**JRC** Объединенный исследовательский центр Joint Research Centre https://ec.europa.eu/info/departments/joint-research-centre_en

**LAW** Семинар по лингвистическим аннотациям The LAW. Proceedings of The Linguistic Annotation Workshop https://www.aclweb.org/anthology/W07-1500.pdf

**LDC** Консорциум лингвистических данных Linguistic Data Consortium (LDC) https://www.ldc.upenn.edu

**Linguistic diversity** Программа ЮНЕСКО по поддержке языкового разнообразия и многоязычия в Интернете Linguistic diversity and multilingualism on Internet. http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/linguistic-diversity-and-multilingualism-on-internet/

**LISA** Ассоциация отраслевых стандартов локализации Localization Industry Standards Association https://en.wikipedia.org/wiki/Localization_Industry_Standards_Association

**LSA** Лингвистическое общество Америки Linguistic Society of America https://www.linguisticsociety.org/

**LTAC** Консорциум терминологии и переводов Language Terminology/ Translation and Acquisition Consortium http://ltacglobal.org/about.html

**META-NET** Сеть передового опыта Multilingual Europe Technology Alliance http://www.meta-net.eu/

**NINCH** Национальная инициатива по сетевому культурному наследию The National Initiative for a Networked Cultural Heritage http://www.ninch.org/

**NISO** Национальная организация стандартов по информации National Information Standards Organization http://www.niso.org/

**OASIS** Организация по развитию стандартов структурированной информации Organization for the Advancement of Structured Information Standards http://www.oasis-open.org/

**OLAC** Сообщество открытых лингвистических архивов OLAC, the Open Language Archives Community, http://olac.ldc.upenn.edu/

**OpenAIRE** Программа информационной поддержки открытой науки Open Access Infrastructure for Research in Europe https://www.openaire.eu/

**OWLG** Рабочая группа по открытой лингвистике The Open Linguistics Working Group https://linguistics.okfn.org/index-42.html

**SIL International** – международная некоммерческая организация – бывший Летний институт лингвистики Summer Institute of Linguistics https://www.sil.org/

**SLE** Европейское лингвистическое общество The Linguistic Society of Europe https://societaslinguistica.eu

**TC ISO 37** Технический комитет ИСО 37 Лингвистика и терминология ISO / TC ISO 37 Language and terminology https://www.iso.org/ru/committee/48104.html

**TEI** Консорциум TEI Text Encoding Initiative Consortium https://tei-c.org/

**TELRI** Трансъевропейская инфраструктура лингвистических ресурсов Trans-European Language Resources Infrastructure http://telri.nytud.hu/

T**erminOrg** Терминология для крупных организаций Terminology for Large Organizations http://www.terminorgs.net

**UNGEGN**, Группа экспертов ООН по географическим названиям United Nations Group of Experts on Geographical Names https://unstats.un.org/unsd/ungegn

**W3 C** Консорциум W3 World Wide Web Consortium https://www.w3.org/Consortium/

## ПРИЛОЖЕНИЕ 4. ПРОЕКТЫ, СТАНДАРТЫ, ФОРМАТЫ, РЕСУРСЫ

**ACE** Автоматическое извлечение содержания Automatic Content Extraction https://www.ldc.upenn.edu/collaborations/past-projects/ace

**Agrovoc** Тезаурус по сельскому хозяйству ФАО A portmanteau of agriculture and vocabulary https://agrovoc.fao.org/browse/agrovoc/en/

**AIDA** Инструмент альтернативных семантических интерпретаций (Active Interpretation of Disparate Alternatives) https://www.ldc.upenn.edu/collaborations/current-projects

**ALE** Атлас языков Европы Atlas Linguarum Europae http://www.lingv.ro/ALE.html

**ALTO** Анализируемый макет и текст Analyzed Layout and Text Object https://github.com/altoxml/documentation/wiki

**ANC** Американский национальный корпус American National Corpus https://catalog.ldc.upenn.edu/LDC2005T35

**AQUA-motion** Типологическая БД глаголов плавания https://linghub.ru/aquamotion/

**AUTOTYP** Исследовательская программа по количественной и качественной типологии https://www.autotyp.uzh.ch/theory.html

**Bamboo** Инфраструктурный проект для цифровой гуманитаристики https://www.projectbamboo.org/

**BeLMaP** Проект лингвистического картографирования Беркли file:///C:/Users/%D0%90%D0%BB%D0%B5%D0%BA%D1%81%D0%B0%D0%BD%D0%B4%D1%80/Downloads/BeLMaP%20 Final%20 Report.pdf

**BOLT** [Интеграция независимых систем машинного перевода] Broad Operational Language Translation https://www.ldc.upenn.edu/collaborations/current-projects/bolt

**BTS** Большой словарь русского языка https://gufo.me/dict/kuznetsov

**CAT** Компьютерная поддержка перевода Computer-Aided Translation https://ru.smartcat.com/blog/cat-tools-programma-dlya-perevodchikov/

**CCR** Регистр понятий CLARIN Concept Registry https://www.clarin.eu/ccr

**CCSL** Язык спецификации компонентов CMDI Component metadata specification language https://www.iso.org/obp/ui/#iso:std:iso:24622:-2:ed-1:v1:en

**CES/XCES** Стандарт кодирования корпусов Corpus Encoding Standard http://www.cs.vassar.edu/CES/

**CHAT** Коды для интеллектуального анализа транскрипции Codes for the Human Analysis of Transcripts https://www.allacronyms.com/CHAT/Codes_for_the_Human_Analysis_of_Transcripts

**CHC** Контролируемая человеческая коммуникация Controlled human communication https://www.iso.org/standard/74581.html

**CHILDES** Система обмена данными детской речи Child Language Data Exchange System https://childes.talkbank.org/

**CIDER-CL** Система одноязычного и межъязыкового согласования онтологий System to perform monolingual and cross-lingual ontology alignment. https://oeg.fi.upm.es/files/cider-cl/

**CLDF** Формат кросс-лингвистических связанных данных Cross-Linguistic Data Formats https://cldf.clld.org/

**CLICS** База данных кросс-лингвистических колексификаций The Database of Cross-Linguistic Colexifications https://doi.org/10.1038/s41597-019-0341-x

**CLLD** Кросс-лингвистические связанные данные Cross-Linguistic Linked Data https://clld.org/

**CMC** Сетевые компьютерные коммуникации Computer-mediated communication https://www.clarin.eu/resource-families/cmc-corpora

**CMDI** Инфраструктура компонентов метаданных Component MetaData Infrastructure https://www.clarin.eu/content/component-metadata

**CMDI-to-RDF** Преобразователь записей CLARIN CMD в связанные открытые данные http://cmdi2rdf.meertens.knaw.nl/cmd2rdf/

**CNC** Чешский национальный корпус Czech National Corpus, CNK, Český národní korpus https://korpus.cz/

**CNL** Контролируемый естественный язык Controlled natural language https://www.iso.org/obp/ui/#iso:std:iso:ts:24620:-1:ed-1:v1:en

**COCONUT** Интегрированные методы генерации и интерпретации Cooperative, coordinated natural language utterances http://www.pitt.edu/~coconut/

**CoNLL- RDF** Словарь для представления корпусов в RDF Linked Corpora Done in an NLP-Friendly Way https://www.researchgate.net/publication/318134320_CoNLL-RDF_Linked_Corpora_Done_in_an_NLP-Friendly_Way

**CQLF** Корпусные запросы Lingua Franca Corpus Query Lingua Franca https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecCQLF#SpecCQLF-1-2011

**DaFoDiL** Формат данных для цифровой лингвистики The Data Format for Digital Linguistics https://format.digitallinguistics.io/

**DAML+OIL** Язык разметки агентов, основанный на RDF и расширение для онтологий DARPA Agent Markup Language +Ontology Inference Layer https://en.wikipedia.org/wiki/DARPA_Agent_Markup_Language

**DatCatInfo** Репозиторий категорий данных Data Category Repository https://datcatinfo.net/

**DBpedi**a Краудсорсинговый проект извлечения структурированной информации из данных Википедии https://www.dbpedia.org/

**DC** Дублинское ядро Dublin Core https://dublincore.org/

**DCAM** Абстрактная модель Дублинского ядра DCMI Abstract Model http://dublincore.org/documents/2005/03/07/abstract-model/

**DCAT** Словарь каталога данных Data Catalog Vocabulary https://www.w3.org/TR/vocab-dcat-2/https://www.w3.org/TR/vocab-dcat-2/

**DCEP** Цифровой корпус Европарламента Digital Corpus of the European Parliament https://ec.europa.eu/jrc/en/language-technologies/dcep

**DCMES** Набор элементов метаданных Дублинского ядра Dublin Core Metadata Element Set http://dublincore.org/

**DCMI** Инициатива по метаданным Дублинского ядра Dublin Core Metadata Initiative http://dublincore.org/

**DCR** Репозиторий категорий данных Data Category Repository http://datcatinfo.net/

**DCS** Выборки категорий данных Data Category Selection https://standartgost.ru/g/%D0%93%D0%9E%D0%A1%D0%A2_%D0%A0_%D0%98%D0%A1%D0%9E_12620-2012

**DERILINX [**Проект преобразования публичных ресурсов в открытые связанные данные] Drive Decision-Making. Inspire Change https://derilinx.com/

**DGT** Память перевода Генеральной дирекции переводов Directorate General for Translation Translation Memory https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

**DGT-Acquis** Многоязычный параллельный корпус официального журнала ЕС https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis

**DiACL** Диахронический атлас компаративной лингвистики. БД типологии древних языков Diachronic Atlas of Comparative Linguistics. A database for ancient language typology https://doi.org/10.1371/journal.pone.0205313

**DiAML** Язык разметки диалога Dialogue Act Markup Language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecDiAML

**DITA** Дарвинская архитектура печатания информации Darwin Information Typing Architecture – https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecDita

**DLx** Цифровая лингвистика Digital Linguistics https://digitallinguistics.io/about/

**DOBES** Документирование исчезающих языков Documentation of endangered languages https://dobes.mpi.nl/research/

**DOI** Цифровой идентификатор объекта Digital object identifier https://www.doi.org/

**DOL** Язык распределенных онтологий Distributed Ontology Language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecDOL

**DOLCE** Дескриптивная онтология для языковой инженерии и когнитологии Descriptive Ontology for Linguistic and Cognitive Engineering https://digitaltwinhub.co.uk/top-level-ontologies/descriptive-ontology-for-linguistic-and-cognitive-engineering-r8/#:~:text=Descriptive%20Ontology%20for%20Linguistic%20and%20Cognitive%20Engineering%20(DOLCE)%20is%20a,for%20Applied%20Ontology%20(LOA).

**DRI** (R) Инициатива по ресурсам дискурса Discourse Resource Initiative http://www.pitt.edu/~coconut /, https://people.ict.usc.edu/~traum/Md/DSD/

**DSIs** Инфраструктура цифровых сервисов Digital Service Infrastructures https://digital-strategy.ec.europa.eu/en/library/connecting-europe-facility-cef-digital-service-infrastructures

**DSSSL** Язык семантики и спецификаций стилей документа Document Style Semantics and Specification Language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecDSSSL

**DWS** Словарная система письма Dictionary writing system https://en.wikipedia.org/wiki/Dictionary_writing_system

**EAC** Память перевода Генеральной дирекции образования и культуры EAC-Translation Memory https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory

**EAF** Формат аннотаций ELAN. ELAN Annotation Format http://www.mpi.nl/tools/elan/EAF_Annotation_Format_2.8_and_ELAN.pdf

**EAN** Европейский номер товара European Article Number https://www.ean-search.org/

**ECDC** Память перевода Европейского центра профилактики и контроля заболеваний ECDC-Translation Memory https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory

**EFR** Толковый словарь русского языка Ефремовой https://www.efremova.info/

**ELCat** Каталог исчезающих языков Catalogue of Endangered Languages http://ling.hawaii.edu/research-current/projects/elcat/

**ELDP** Программа документирования исчезающих языков The Endangered Languages Documentation Programme https://www.eldp.net/en/about

**ELEXIS** Европейская лексикографическая инфраструктура European Lexicographic Infrastructure https://elex.is

**ELRC3** Европейская координация ЛИР European Language Resources Coordination http://www.elra.info/en/projects/current-projects/elrc3/

**E-MELD.** Электронная метаструктура для данных об исчезающих языках Electronic Metastructure for Endangered Languages Data http://emeld.org/

**Ethnologue** Этнолог: языки мира Ethnologue Languages of the World https://www.ethnologue.com/about

**EURALEX** Европейская ассоциация лекскографии European association for lexicography https://web.archive.org/web/20130313080818/http://www.euralex.org/

**Europarl** Параллельный корпус трудов Европарламента для статистического МП European Parliament Proceedings Parallel Corpus for Statistical Machine Translation https://www.google.ru/search?q=European+Parliament+Proceedings+Parallel+Corpus+for+Statistical+Machine+Translation+&newwindow=1&sxsrf=APq-WBsupM5qudYF2vhhYROwKTM55S82BQ%3A1648711706977&ei=GlhFYougO7OW9u8P3 u63 wAg&ved=0ahUKEwiLlYbd6 e_2 AhUzi_0 HHV73 DYgQ4 dUDCA4&oq=European+Parliament+Proceeings+Parallel+Corpus+for+Statistical+Machine+Translation+&gs_lcp=Cgdnd3 Mtd2l6 EAw6 BwgjEOoCECdKBAhBGABKBAhGGABQihZYihZg7 CRoAXABeACAAYwBiAGMAZIBAzAuMZgBAKABAaABArABCsABAQ&sclient=gws-wiz

**Eurotyp** Типологические инструменты для полевой лингвистики Typological tools for field linguistics https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/database-system_description.php

**EuroWordNet** Многоязычная БД на wordnets для европейских языков A multilingual database with wordnets for several European languages https://archive.illc.uva.nl/
EuroWordNet/

**EXMARaLDA** Расширяемый язык разметки для аннотации дискурса Extensible Markup Language for Discourse Annotation https://exmaralda.org/de/2015/11/09/exmaralda_schulung_27_11_2015/

**FIELD**. Среда ввода лингвистических данных Field Input Environment for Linguistic Data http://emeld.org/tools/fieldinput.cfm

**FirstVoices** [Инструменты и сервисы поддержки исчезающих языков] https://www.firstvoices.com/

**FLORA-2** Объектно-ориентированный язык базы знаний Object-oriented knowledge base language http://flora.sourceforge.net/

**FoLiA** Формат лингвистической аннотации на основе XML https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.221.4523&rep=rep1&type=pdf

**Forenames_histHub** База данных личных имен https://histhub.ch/tools/histhub-forenames/

**FRAMENET** Лексическая БД английского языка lexical database of English https://framenet.icsi.berkeley.edu/fndrupal/

**FreeDict** БД бесплатных двуязычных словарей https://freedict.org/

**FREME**. Открытая структура электронных услуг для многоязычного и семантического обогащения цифрового контента Multilingual Semantic Enrichment with Linked Data and Language Technologies https://www.aclweb.org/anthology/L16-1660/

**FXML**. Формат для разметки диалогов Format for DiAML; https://www.researchgate.net/publication/267205522_ISO_24617-2_A_semantically-based_standard_for_dialogue_annotation

**GeM** Жанровая и мультимодальная структура Genre and Multimodality framework https://www.aclweb.org/anthology/W16-2109.pdf

**GEMET** Общий многоязычный экологический тезаурус GEneral Multilingual Environmental Thesaurus https://www.eionet.europa.eu/gemet/en/about/

**GEvTerm** Глобальные терминологические вызовы Global Event Terminology http://gevterm.net/gevterm/index.php

**GitHub** Веб-сервис для хостинга IT-проектов https://github.com/

**GMB** Семантический банк деревьев Groningen Meaning Bank https://gmb.let.rug.nl/

**GMX GILT** Глобализация, интернационализация, локализация, перевод Globalization, Internationalization, Localization, and Translation https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BC%D1%8F%D1%82%D1%8C_%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4%D0%BE%D0%B2

**GOLD** Общая онтология лингвистического описания General Ontology for Linguistic Description http://linguistics-ontology.org/info/about

**GrAF** Формат аннотаций в виде графа Graph Annotation Format. https://www.cs.vassar.edu/~ide/papers/LAW.pdf

**Grammis** Портал по грамматике немецкого языка https://grammis.ids-mannheim.de/

**HAVIC** Гетерогенная аудиовизуальная интернет-коллекция Heterogeneous Audio Visual Internet Collection https://www.ldc.upenn.edu/collaborations/past-projects/havic

**HAVRUS Corpus** Высокоскоростное распознавание аудиовизуальной русской речи High-Speed Recordings of Audio-Visual Russian Speech https://www.researchgate.net/

publication/306064591_HAVRUS_Corpus_High-Speed_Recordings_of_Audio-Visual_Russian_Speech

**Hearables Challenge** Распознавание речи в условиях помех https://www.ldc.upenn.edu/collaborations/current-projects/hearables-challenge

**HTML** Язык гипертекстовой разметки HyperText Markup Language http://www.w3.org/MarkUp/draft-ietf-iiir-html-01.txt

**Huma-Num** Инфраструктура цифровой гуманитаристики L'infrastructure humanities numeric https://www.huma-num.fr/

**HyTime** Язык структурирования на основе времени Hypermedia/Time-based Structuring Language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecHyTime

**IAEA Safety Glossary** Глоссарий по безопасности International Atomic Energy Agency Safety Glossary https://www.iaea.org/resources/safety-standards/safety-glossary

**ILOTERM ТБД** Международной организации труда Terminology database of the International Labour Organization https://www.ilo.org/inform/online-information-resources/databases/terminology/lang--en/index.htm

**IMDI** Инициатива метаданных ISLE ISLE Metadata Initiative http://tla.mpi.nl/imdi-metadata/

**IPA** Международный фонетический алфавит International Phonetic Alphabet https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecIPA

**ISABASE** Корпус русской звучащей речи http://www.dialog-21.ru/digest/2001/articles/krivnova

**ISBD** Международное стандартное библиографическое описание International Standard Bibliographic Description https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecISBD

**ISBN** Международный стандартный номер книги International Standard Book Number https://www.isbn-international.org/

**Islandora** Управление цифровыми активами с открытым исходным кодом Open source digital asset management http://clarin04.ims.uni-stuttgart.de/repo

**ISLE** Международные стандарты для языковых технологий International Standard for Language Engineering https://www.mpi.nl/ISLE/

**ISLRN** Международный стандартный номер ЛИР International Standard Language Resource Number http://www.elra.info/en/islrn/

**ISO DCR** Реестр категорий данных The Data Category Registr https://www.aclweb.org/anthology/L08-1034/

**ISOcat** Репозиторий категорий данных The Data Category Repository http://datcatinfo.net/

**ISO-Thesauri** Тезаурусы и взаимодействие с другими словарями Thesauri and interoperability with other vocabularies https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecThesauri

**ITRT** Международный тезаурус терминологии беженцев International Thesaurus of Refugee Terminology

**JATS** Набор тегов журнальных статей Journal Article Tag Suite https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecJATS

**JRC Eurovoc** Тезаурус Eurovoc https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer

**JRC-Names** Имена и варианты их написания JRC-Names https://ec.europa.eu/jrc/en/language-technologies/jrc-names

**JSON** Текстовый формат обмена данными, основанный на JavaScript JavaScript Object Notation https://ru.wikipedia.org/wiki/JSON

**KAF** Формат аннотаций знаний Knowledge Annotation Format https://www.researchgate.net/publication/228922488_KAF_a_generic_semantic_annotation_format

**KAIROS** Система искусственного интеллекта по идентификации сложных событий Knowledge-directed Artificial Intelligence Reasoning Over Schemas https://www.ldc.upenn.edu/collaborations/current-projects

**KYOTO** [формат, представляющий аннотацию документов через изолированную многослойную структуру] Knowledge Yielding Ontologies for Transition-based Organization https://cordis.europa.eu/project/id/211423

**LACITO** Проект по архивированию лингвистических данных Linguistic Data Archiving Project http://xml.coverpages.org/lacitoAR-desc-english.html

**LAF** Структура лингвистической аннотации The Linguistic Annotation Framework https://www.iso.org/obp/ui/#iso:std:iso:24612:ed-1:v1:en

**LCCN** Контрольный номер Библиотеки Конгресса Library of Congress Control Number https://id.loc.gov/authorities/names.html

**LDL** Связанные данные в лингвистике Linked Data in Linguistics https://www.aclweb.org/anthology/venues/ldl/

**LEGO** Расширение лексикона с помощью онтологии GOLD Lexicon Enhancement via the GOLD Ontology http://lego.linguistlist.org/

**LEI** Индекс языковой опасности The Language Endangerment Index http://www.endangeredlanguages.com/about_catalogue/

**LexInfo** Онтология категорий данных для модели Lemon Ontology that was defined during the Monnet Project to provide data categories for the Lemon model https://lexinfo.net/

**LEXVO** Информация о языках в формате связанных данных http://www.lexvo.org/

**LIDER: FP** Связанные данные как средство кросс-медиа и многоязычной аналитики контента для предприятий разных стран Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe https://docplayer.net/139432478-Lider-fp-linked-data-as-an-enabler-of-cross-media-and-multilingual-content-analytics-for-enterprises-across-europe.html

**LILA Linking Latin** Создание базы знаний ЛИР для латинского языка Building a Knowledge Base of Linguistic Resources for Latin https://lila-erc.eu/#page-top

**LIME** Словарь лингвистических метаданных LInguistic MEtadata https://lod-cloud.net/dataset/lime

**Linghub** Поисковая система ЛИР A comprehensive location for finding information about language resources http://linghub.org/

**LingPy** Библиотека программ для исторической лингвистики http://lingpy.org

**LINGUIST List** Международное сообщество лингвистов онлайн The Linguist List Internationai Linguistics Community Online https://linguistlist.org

**LINGVO** Электронный словарь ABBYY https://www.lingvo.ru/

**LINGVODOC** Платформа корпусов http://lingvodoc.ru/corpora_all

**Linport** Портфолио проекта языковой совместимости The Language Interoperability Portfolio Project www.linport.org

**LiODi**. Связанные открытые словари Linked Open Dictionaries http://ionov.me/liodi/

**LIRICS** Лингвистическая инфраструктура для совместимых ресурсов и систем Linguistic Infrastructure for Interoperable Resources and Systems http://lirics.loria.fr

**Lithos** Поисковая система, разработанная компанией DG Interpretation https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/terminology-tools-and-resources/terminology-dg-interpretation_en#lithos

**LL-MAP**. [Лингвистические геоданные] Language and Location – A Map Annotation Project http://llmap.org/

**LLOD** Лингвистические связанные открытые данные Linguistic Linked Open Data https://linguistic-lod.org/

**LMF** Структура лексической разметки Lexical Markup Framework https://en.wikipedia.org/wiki/Lexical_Markup_Framework

**LocaLingual** Интерактивная карта языков и диалектов для путешествий https://zen.yandex.ru/media/biletikaero/sozdana-iazykovaia-karta-mira-dlia-puteshestvii-i-izucheniia-dialektov-5ef526befab32a2ddf9eaad6

**LOD2**. Создание знаний из взаимосвязанных данных Creating Knowledge out of Interlinked Data https://lod2.eu/

**LRE** Оценка распознавания языков Language Recognition Evaluation https://www.nist.gov/itl/iad/mig/language-recognition

**LRE Map** Карта оценочного описания лингвистических ресурсов Linguistic resources evaluation http://www.elra.info/en/catalogues/lre-map/

**LREC** Международная конференция по ЛИР и их оценке The International Conference on Language Resources and Evaluation http://www.lrec-conf.org/

**LTC** Русский учебный корпус переводов Russian Learner Translator Corpus https://rus-ltc.org/search

**MADCAT** Многоязычный автоматический перевод, классификация и анализ текста Multilingual Automatic Document Classification, Analysis and Translation https://www.ldc.upenn.edu/collaborations/current-projects/madcat

**MAF** Морфосинтаксическая аннотационная система Morpho-syntactic annotation framework https://standartgost.ru/g/ISO_24611:2012

**Mailing Lists** Архив лингвистических сайтов Mailing Lists https://old.linguistlist.org/lists/

**MAPA** Многоязычный инструментарий анонимизации для государственных администраций The Multilingual Anonymization Toolkit for Public Administrations https://mapa-project.eu/#:~:text=The%20MAPA%20Project%20is%20an,the%20medical%20and%20legal%20fields.

**MAS** Малый академический словарь https://gufo.me/dict/mas

**META-SHARE** Метамодель META-SHARE Metadata Model http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.4-Final.pdf

**METS** Стандарт кодирования и передачи данных Metadata Encoding and Transmission Standard http://www.loc.gov/standards/mets/

**MHATLEX** Лексические ресурсы моделирования французского произношения Markovian Harmonic Adaptation and Transduction http://www.lrec-conf.org/proceedings/lrec2000/pdf/37.pdf

**MICASE** Мичиганский корпус академического разговорного английского языка Michigan Corpus of Academic Spoken English https://quod.lib.umich.edu/m/micase/

**MIME** Многоцелевые расширения почты Интернета Multipurpose Internet Mail Extension, https://ru.wikipedia.org/wiki/MIME

**MLIA** Многоязычный доступ к информации по COVID-19. Multilingual Information Access initiative http://www.elra.info/en/projects/current-projects/covid-19-mlia-init/

**MODS** Схема метаданных описания объекта Metadata Object Description Schema http://www.loc.gov/standards/mods/

**Monnet.** Многоязычные онтологии сетевых знаний Multilingual Ontologies for Networked Knowledge ID: 248458 https://cordis.europa.eu/project/id/248458

**MULTEXT-East**. Многоязычные текстовые инструменты и корпуса для языков Центральной и Восточной Европы Multilingual Text Tools and Corpora for Central and Eastern European Languages http://nl.ijs.si/ME/

**Multitext project** [Электронная библиотека гомеровских текстов с инструментарием] http://www.homermultitext.org/

**MultiTree** Генеалогические деревья MultiTree https://old.linguistlist.org/projects/multi-tree.cfm

**NER** Распознавание именованных сущностей Named Entity Recognition https://sysblok.ru/glossary/named-entity-recognition-ner/

**NexusLinguarum**. Европейская сеть лингвистических данных, ориентированная на Интернет NexusLinguarum https://nexuslinguarum.eu/

**NISO MIX** Метаданные NISO для изображений в XML-схеме NISO Metadata for Images in XML Schema http://www.loc.gov/standards/mix///

**NLM JATS** Набор меток архивирования и обмена журнальной информацией. NLM Journal Archiving and Interchange Tag Suite https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecNLMJATS

**OAI PMH** Протокол сбора метаданных Инициативы открытх архивов Open Archives Initiative Protocol for Metadata Harvesting https://openarchives.org/pmh/

**OCD** Словарь в один клик OneClick Dictionary https://github.com/elexis-eu/ocd

**ODIN** Онлайновая база данных межлинейных примечаний The Online Database of Interlinear Text http://odin.linguistlist.org/

**ODRL** Открытый язык цифровых прав Open Digital Rights Language. https://www.w3.org/TR/odrl-model

**OLAC Metadata** Метаданные открытого языкового архива Open Language Archive Metadata http://www.language-archives.org/OLAC/metadata-20080531.html

**OLiA** Онтология лингвистических аннотаций Ontologies of Linguistic Annotation http://semantic-web-journal.net/system/files/swj518_0.pdf

**OLIF** Открытый стандарт для обмена терминологическими и лексическими данными Open Lexicon Interchange Format http://www.olif.net/

**OntoIOp** Интеграция и совместимость онтологий Ontology Integration and Interoperability https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecOntoIOp

**OpenCyc** Онтологическая база знаний Open source version of the Cyc technology https://www.opennet.ru/prog/info/3341.shtml

**OWL** Язык веб-онтологий Web Ontology Language http://www.w3.org/TR/2009/REC-owl2-overview-20091027/

**OWP** Открытая веб-платформа Open web platform https://en.wikipedia.org/wiki/Web_platform

**PANLEX** Многоязычная база данных лексических переводов https://panlex.org/

**PDF** Формат переносимого документа Portable Document Format https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecPDF

**PDF/A** Форма долговременного хранения электронного документа Electronic document file format for long-term preservation http://www.pdf-tools.com/public/downloads/whitepapers/Whitepaper-PDFA-Standard-ISO-19005-US.pdf

**Penn Treebank** Банк деревьев синтаксических структур Phrase Structure Treebank https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecPennTB

**PHONOLEX** БД форм произношения словосочетаний, содержащихся в речевых корпусах Phonetik und Sprachverarbeitung https://www.phonetik.uni-muenchen.de/forschung/abgeschlossene_projekte/phonolex.html

**PID** Постоянный идентификатор Persistent identificator https://docs.cntd.ru/document/1200110804

**PISA** Постоянная идентификация и устойчивый доступ Persistent identification and sustainable access https://docs.cntd.ru/document/1200110804

**POINTER** Предложения по оргнизации терминологической инфраструктуры в Европе Proposals for an operational infrastructure for terminology in Europe https://cordis.europa.eu/project/id/LRE63090/results

**POPIN** Многоязычный тезаурус по народонаселению Population multilingual thesaurus https://www.amazon.com/POPIN-thesaurus-Population-multilingual/dp/B0007C9W6E

**POS- tags** Разметка частей речи Part-of-speech tagging https://universaldependencies.org/u/pos/

**POSTDATA**. Стандартизация поэзии и связанные открытые данные Poetry Standardization and Linked Open Data https://cordis.europa.eu/project/id/679528

**POWLA** Словарь для общих лингвистических структур данных

**Prêt-à-LLOD** [Преобразование лингвистических данных в связанные открытые данные] https://pret-a-llod.github.io/

**PWN** Принстонский WordNet Princeton WordNet https://wordnet.princeton.edu/

**QTLeap** Качественный перевод с использованием подходов глубинной инженерии Quality translation by deep language engineering approaches http://qtleap.eu/

**R2 ML** Язык разметки правил REWERSE I1 REWERSE I1 Rule Markup Language https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.540.3053&rep=rep1&type=pdf

**RDF** Структура описания ресурса Resource Description Framework http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

**RDF/XML** Спецификация синтаксиса RDF / XML RDF/XML Syntax Specification http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/

**RDFS** RDF-схема языка описания словарей Vocabulary Description Language: RDF Schema https://www.w3.org/TR/rdf-schema/

**RDQL** Язык запросов данных RDF Data Query Language http://www.w3.org/Submission/RDQL/

**RELAX NG** Регулярный язык для следующего поколения XML REgular LAnguage for XML Next Generation https://www.oasis-

open.org/committees/tc_home.php?wg_
abbrev=relax-ng

**RELISH** Обеспечение функциональной совместимости лексиконов языков, находящихся под угрозой исчезновения, посредством гармонизации стандартов Rendering Endangered Language Lexicons Interoperable through Standards Harmonization https://old.linguistlist.org/projects/relish.cfm

**RIF** Формат обмена правилами Rule Interchange Format https://en.wikipedia.org/wiki/Rule_Interchange_Format

**RIFF** [Формат файлов-контейнеров для хранения потоковых мультимедиа-данных] Resource Interchange File Format https://ru.wikipedia.org/wiki/RIFF

**Rosetta** [Фонд библиотеки долговременного хранения человеческих языков] Rosetta https://rosettaproject.org

**RQL** Язык запросов RDF RDF Query Language https://www.w3.org/wiki/RDF_Query

**RSTB** Тестовый корпус с параллельной синтаксической разметкой http://otipl.philol.msu.ru/~soiza/testsynt/

**RTF** Обогащенный текстовый формат Rich Text Format https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecRTF

**RTLOD** Русский тезаурус в формате открытых связанных данных Russian Thesauri as Linked Open Data. https://nlpub.ru/RTLOD

**RuleML** Инициатива языка разметки правил The Rule Markup Initiative https://ruleml.org/index.html

**RussNet** http://project.phil.spbu.ru/RussNet/index_ru.shtml

**RUS-Treebank** Синтаксически аннотированные корпуса русского языка http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php

**RuThes** Русский тезаурус http://www.labinform.ru/pub/ruthes/index.htm

**RWN** Русский Wordnet http://wordnet.ru

**SBD** Определение границ предложения Sentence Boundary Detector https://www.aclweb.org/anthology/C12-2096.pdf

**SemAF** Структура семантической аннотации Semantic annotation framework https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecSemAF

**SemCor** Семантически аннотированный корпус английского языка Semantically annotated English corpus https://www.sketchengine.eu/semcor-annotated-corpus/

**SemRoleML** Язык разметки семантических ролей Semantic role markup language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecSemRoleML

**Senseval** Инструмент оценки семантического анализа текста Evaluation Exercises for the Semantic Analysis of Text http://web.eecs.umich.edu/~mihalcea/senseval/

**SeRQL** Язык запросов Sesame RDF Query Language http://www.csee.umbc.edu/courses/graduate/691/spring14/01/examples/sesame/openrdf-sesame-2.6.10/docs/users/ch09.html

**SGML** Стандартный обобщенный язык разметки Standard Generalized Markup Languag https://ru.wikipedia.org/wiki/SGML

**SimpL-1** Упрощенный естественный язык Simplified natural language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecSimpL1

**SIMPLE** Ядерная онтология SIMPLE Core Ontology https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecSimple

**SKOS** Простая система организации знания Simple Knowledge Organization System https://www.w3.org/2004/02/skos/

**SMDL** Стандартный язык описания музыки Standard music description language http://xml.coverpages.org/smdlover.html.

**SMG** Морфологическая группа Суррея Surrey Morphology Group https://www.smg.surrey.ac.uk/

**SpaCy** Библиотека открытых программ NLP Free open-source library for NLP in Python https://spacy.io/

**SPARQL** Протокол и язык запросов Protocol and RDF Query Language http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

**SRE** Оценка распознавания речи Speaker Recognition Evaluation https://www.ldc.upenn.edu/collaborations/current-projects

**SRX** Обмен правилами сегментации Segmentation Rules eXchange http://www.gala-global.org/oscarStandards/srx/srx20.html

**SUMO** Онтология верхнего уровня Suggested Upper Merged Ontology http://www.ontologyportal.org/

**SV** Структурированные словари для информационного поиска Structured vocabularies for information retrieval.https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecSV#SpecSV1-BS

**SWRL**Язык правил семантического веба, комбинирующий OWL и RuleML Semantic Web Rule Language Combining OWL and RuleML

**SynAF** Система синтаксического аннотирования Sintactic annotation framework https://nd.gostinfo.ru/document/6262982.

**TAC KBP** Конференция по текстовому анализу и формированию базы знаний Text Analysis Conference Knowledge Base Population https://www.ldc.upenn.edu/collaborations/past-projects/tac-kbp

**TalkBank [**Проект по исследованию устной коммуникации**]** https://www.talkbank.org/

**TBX** Стандарт представления терминологических БД TermBase eXchange https://en.wikipedia.org/wiki/TermBase_eXchange

**TBX-Basic** Облегченная версия TBX https://www.tbxinfo.net/

**TDS** Система типологических баз данных The Typological Database System https://doi.org/10.17026/dans-xc9-mnrf

**TEI Guidelines** Рекомендации по кодированию и обмену электронным текстом Guidelines for Electronic Text Encoding and Interchange http://www.tei-c.org/Vault/GL/P3/index.htm

**TERMWEB** Терминологическая БД в составе DatCatInfo https://datcatinfo.termweb.se/termweb/app

**TeX/LaTeX** Набор макрорасширений системы компьютерной верстки TeX, https://ru.wikipedia.org/wiki/LaTeX

**TextMD** Технические метаданные текста Technical Metadata for Text https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTextMD

**TGIR** Очень большая научная инфраструктура Les très grandes infrastructures de recherche https://www.data.gouv.fr/fr/datasets/les-tres-grandes-infrastructures-de-recherche-tgir/

**TimeML** Язык разметки для событий и временных выражений на естественном языке Markup Language for events and temporal expressions in natural language http://timeml.org/site/publications/timeMLdocs/timeml_1.1b.htm

**TIMIT** Корпус фонетически и лексически транскрибированной речи на американском английском corpus of phonemically and lexically transcribed speech of American English speakers https://catalog.ldc.upenn.edu/LDC93S1

**TLM** Сокровищница лингвистических карт Treasury of Linguistic Maps https://www.degruyter.com/document/doi/10.1515/tlm/html

**TM** Память перевода Translation memory https://en.wikipedia.org/wiki/Translation_memory

**TMF** Структура терминологической разметки Terminological Markup Framework. https://www.iso.org/standard/56063.html

**TML** Язык терминологической разметки Terminological Markup Language https://www.iso.org/standard/56063.html

**TMS** Система управления терминологией Terminology Menagement System https://www.iso.org/standard/81917.html

**TMX** Обмен памяти переводовTranslation Memory eXchange http://www.gala-global.org/oscarStandards/tmx/

**Topic Maps** Тематические карты https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTopicMaps https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTopicMaps

**Tranquality** Качество перевода Translation quality https://www.tranquality.info/

**TransWS** Переводческие веб-сервисы Translation Web Services https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=trans-ws

**TRIPLE** Язык запросов, выводов и преобразований RDF для семантического веба https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTriple

**TSL** Транскрипция устной речи Transcription of Spoken Language https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTranScript

**Turtle** Краткий язык RDF для TRIPLE Terse RDF Triple Language http://www.w3.org/TR/2012/WD-turtle-20120710/ Serialization

**TUSNELDA** Ресурсы языковых данных и аннотаций https://www.lingexp.uni-tuebingen.de/sfb441/abstracts.html

**UCS** Универсальный набор кодированных символов Universal Coded Character Set https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecUCS

**UD-Russian** Синтаксически аннотированные корпуса русского языка http://universaldependencies.org/

**UKAT** Контролируемый словарь для архивных коллекций Controlled vocabulary which archives can use when indexing their collections and catalogues https://web.archive.org/web/20110809235737/http://www.ukat.org.uk/index.html

**UMBEL** (Upper Mapping и Binding Exchange Layer) https://triplydb.com/structured-dynamics/umbel

**UNBIS** Тезаурус библиографической системы ООН United Nations Bibliographic Information System Thesaurus https://digitallibrary.un.org/record/667104?ln=ru

**UNESCO Atlas** Атлас ЮНЕСКО UNESCO Register of Good Practices in Language Preservation http://www.unesco.org/languages-atlas/

**UniMorph** Универсальный морфологический репозиторий The Universal Morphology https://unimorph.github.io/

**UNL** Универсальный сетевой язык (российская версия) Universal Networking Language http://www.unl.ru/

**UNTERM** Многоязычная терминологическая база данных ООН United Nations Multilingual Terminology Database https://unterm.un.org/unterm/portal/welcome

**UPSID** База данных инвентаризации фонологических сегментов UCLA Phonetics Lab Software http://www.linguistics.ucla.edu/faciliti/sales/software.htm

**URI** Универсальный идентификатор ресурса Universal Resource Identifier https://web.archive.org/web/20050730073732/http://gbiv.com/protocols/uri/

**USH** Толковый словарь русского языка Д.Н. Ушакова https://ushakovdictionary.ru/

**VR** Графические интерфейсы для ЛИР и языковых технологий Visual Resources

**WBLL** Веб-обучение языку Web-Based Language Learning http://www.journals.aiac.org.au/index.php/alls/article/view/4638#:~:text=The%20findings%20indicated%20that%20overall,tools%20and%20learning%20management%20systems.

**WIKT** Русский Викисловарь http://ru.wiktionary.org

**WLMS** Система картографирования мировых языков World Language Mapping System https://www.worldgeodatasets.com/language/

**WordNet** Лексическая база данных английского языка A Lexical Database for English https://wordnet.princeton.edu/

**WordSeg** Пословная сегментация письменного текста Word segmentation of written texts https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecWordSeg

**XCES** Стандарт кодирования корпусов на XML Corpus Encoding Standard for XML / http://xml.coverpages.org/xces.html

**XDXF** Обменный формат словарей XML Dictionary eXchange Format https://github.com/soshial/xdxf_makedict/

**XHTML** Расширяемый язык разметки гипертекста Extensible HyperText Markup Language http://www.w3.org/TR/2002/REC-xhtml1-20020801

**XLIFF** Расширяемый стандарт обмена локализуемыми данными Localization Interchange File Format https://ru.wikipedia.org/wiki/XLIFF

**XML** Расширяемый язык разметки Extensible Markup Language http://www.w3.org/TR/1998/REC-xml-19980210

**Xml:tm** Подход к памяти переводов, основан на концепции текстовой памяти https://en.wikipedia.org/wiki/Xml:tm

**XMLNS** Расширяемое пространство имен языка разметки eXtensible Markup Language Name Space http://www.w3.org/TR/2009/REC-xml-names-20091208/

**XPath** Язык навигации XML Path Language http://www.w3.org/TR/1999/REC-xpath-19991116/

**XQuery** Язык запросов XML XML Query Language http://www.w3.org/TR/2014/REC-xqueryx-30-20140408/

**X-SAMPA** Расширенный фонетический алфавит методов оценки речи Speech Assessment Methods Phonetic Alphabet. https://ru.wikipedia.org/wiki/X-SAMPA

**XSD** Схема XML Schema http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/

**XSL-FO** Объекты форматирования расширяемого языка таблиц стилей Extensible Stylesheet Language Formatting Objects http://www.w3.org/TR/2006/REC-xsl11-20061205/

**XSLT** Преобразования расширяемого языка таблиц стилей Extensible Stylesheet Language Transformations http://www.w3.org/TR/1999/REC-xslt-19991116

**YARN** Открытый тезаурус Yet Another RussNet https://russianword.net/

**Z39.87** Словарь данных – технические метаданные для цифровых неподвижных изображений Data Dictionary – Technical Metadata for Digital Still Images http://www.niso.org/apps/group_.public/download.php/6502/Data%20Dictionary%20-%20Technical%20Metadata%20for%20Digital%20Still%20Images.pdf

## ПРИЛОЖЕНИЕ 5. ЦЕНТРЫ ЗНАНИЙ CLARIN

**ACE** Центр знаний CLARIN по экспертизе атипичных коммуникаций http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111650118.

*Сферы компетенции.* Атипичное общение включает язык и речь, которые встречаются во время приобретения и развития (второго) языка, а также при языковых расстройствах, но ещё и – в более широком смысле – при развитии двуязычного языка и языка жестов. ACE специализируется на исследованиях этого типа и сопутствующих инфраструктурных проблемах, связанных со сбором, обработкой и совместным использованием данных, для которых обычно характерны проблемы с конфиденциальностью. Для хранения данных и доступа центр сотрудничает с MPI TLA (The Language Archive), который является центром CLARIN B и также находится в Неймегене.

**CKLD** Центр знаний CLARIN по языковому разнообразию и языковой документации http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228110746110

*Сферы компетенции.* Центр знаний CLARIN по языковому разнообразию и языковой документации предлагает опыт в области данных и связанных с ними методов, технологий и справочную информацию о языковых ресурсах и инструментах для исследователей, включая студентов и носителей языка. CKLD предоставляет информацию и помощь, относящуюся к полевым исследованиям и методологическим аспектам, связанным с данными, и в частности к оборудованию, цифровым инструментам, методам, тому, где искать данные и информацию, к кому обращаться за специальными сведениями по конкретным регионам или языковым семьям.

**CLARIN-HUMLAB** Центр знаний CLARIN гуманитарной лаборатории Лундского университета http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111449108

*Сферы компетенции.* Консультации по мультимодальным и сенсорным методам, включая ЭЭГ, отслеживание глаз, артикулографию, виртуальную реальность, захват движения, AV-запись, обработку естественного языка.

**CLARIN-Learn** Центр знаний CLARIN по анализу изучения языков http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228110824106

*Сферы компетенции.* Центр предоставляет консультации по инструментам, корпусам и методам изучения первого и второго языка, разговорного взаимодействия и различных языковых нарушений и нарушений развития, включая афазию, заикание, черепно-мозговые травмы, деменцию и другие нарушения.

**CLARIN-SMS** Центр знаний CLARIN по шведскому языку в многоязычной среде http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226161900114

*Сферы компетенции.* Языковые технологии и ресурсы для шведского языка, шведского жестового языка и многоязычных языков. Опыт в обработке параллельных корпусов, включая согласование и машинный перевод, предварительно обученные языковые модели, кросс-лингвистически согласованные аннотации в рамках универсальных зависимостей, а также вычисление и оценку показателей сложности текста.

**CLARIN-SPEECH** Центр знаний CLARIN по анализу речи http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111130104

*Сферы компетенции.* Технические рекомендации по анализу речи, относящиеся ко всем аспектам речевой технологии, включая речевую науку, речевые приложения и коммуникации.

**CLASSLA** Центр знаний южнославянских языков CLARIN http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228110955117

*Сферы компетенции.* Предлагает экспертизу языковых ресурсов и технологий для южнославянских языков.

**CORLI-K-center** Французский центр знаний CLARIN для корпусов, языков и взаимодействия http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0511090211123

*Сферы компетенции.* Корпусная лингвистика с особым вниманием к французскому языку и языкам Франции.

**CorpLingCz** Чешский центр знаний CLARIN по корпусной лингвистике http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111414113

*Сферы компетенции.* Предоставляет информацию, консультации и техническую помощь по всем темам, связанным с корпусной лингвистикой. Сюда входят форматы данных, аннотации, запросы к корпусу, методология корпусной лингвистики, статистические методы и т.д. Другой специализацией центра являются эмпирические исследования чешского языка.

**DANSK** К-ЦЕНТР КЛАРИН ДАНСК – Датская служба помощи http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226160500105

*Сферы компетенции.* Датский язык и датский язык жестов; ресурсы датского языка; языковые инструменты для датского языка; методы NLP.

**DiaRes** CLARIN K-центр диахронических языковых ресурсов http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226162722115

*Сферы компетенции.* Коллекции диахронических текстов, исторические тексты, а также инструменты и ресурсы для их обработки и анализа.

**IMPACT-CKC** Центр компетенций IMPACT – CLARIN K-center по оцифровке http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111053112

*Сферы компетенции.* IMPACT-CKC (центр компетенции IMPACT – CLARIN K-центр в области оцифровки) как центр знаний, предлагающий экспертные знания и

ресурсы учреждениям и исследователям, которым нужен совет по оцифровке и смежным областям. Ресурсы IMPACT-CKC включают демонстрационную платформу для инструментов онлайн-тестирования, коллекцию высококачественных изображений с соответствующими фактами, историческую лексику для десяти языков, а также учебные материалы и реестры по инструментам, инициативам, наборам данных и соревнованиям.

**K-BLP** Центр знаний CLARIN по обработке белорусского текста и речи
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111053112
*Сферы компетенции.* Знание обработки текста и речи на белорусском и других языках; знания об изучении белорусского языка; инструменты и ресурсы для обработки текста и речи на белорусском и других языках.

**NLP:EL** CLARIN K-центр обработки естественного языка в Греции
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0330124100122
*Сферы компетенции.* Исследование NLP для греческого языка. Состояние оцифровки греческого языка.

**NSD-K-centre** Центр знаний CLARIN по управлению данными в NLP
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226162209111
*Сферы компетенции.* Предоставляет экспертные знания в области управления данными, включая юридические и этические вопросы, связанные с конфиденциальностью и правами интеллектуальной собственности.

**PhA-OeAW** Фонограммархив / Австрийская академия наук – CLARIN K-Center
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228110614102
*Сферы компетенции.* В качестве аудио- и аудиовизуального архива с многочисленными коллекциями уникальных исследовательских записей со всего мира, охватывающих период в 120 лет Фонограммархив предлагает различные услуги. Помимо предоставления доступа к своим обширным ресурсам данных и метаданных (удаленным и локальным), он консультирует ученых по методологии аудиовизуальных исследований в области социальных и гуманитарных наук, а также по технологиям аудио- и аудиовизуальной документации, поддерживая их необходимым записывающим оборудованием. Кроме того, он широко делится своим обширным опытом по таким темам, как восстановление, оцифровка, устаревание формата, каталогизация, метаданные, долгосрочное сохранение и хранение.

**PolLinguaTec** Центр знаний CLARIN по технологиям польского языка
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111731109
*Сферы компетенции*. Предоставляет обширные знания о методах анализа естественного языка с особым акцентом на анализе польского языка. Предлагает поддержку всех типов приложений лингвистических технологий для польского языка, как одно-, так и многоязычных.

**PORTULAN** Центр знаний по исследованиям и технологиям португальского языка http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226163322120

*Сферы компетенции.* Наука и технология португальского языка – это тематическая область Центра знаний CLARIN. Связанный с португальским языком, он охватывает все темы от фонетики до дискурса и диалога с учетом всех языковых функций, от коммуникативной деятельности до культурного выражения, применяется во всех дисциплинах – от теоретической лингвистики до языковых технологий, охватывает все языковые варианты.

**SAFMORIL** Системы и структуры для языков с богатой морфологией
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228110910119

*Сферы компетенции.* SAFMORIL объединяет исследователей и разработчиков в области вычислительной морфологии и ее приложений в NLP. В центре внимания SAFMORIL находятся актуальные рабочие системы и структуры, основанные на лингвистических принципах, обеспечивающие лингвистически мотивированный анализ и генерирование результатов. Такие системы актуальны, в частности, для языков с богатой морфологией. SAFMORIL предлагает онлайн-курсы для разработки морфологий, токенизаторов и средств проверки орфографии, а также репозиторий для хранения морфологий.

**Spanish-K-centre** Испанский CLARIN K-center
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111241101

*Сферы компетенции.* Испанский CLARIN K-Center стремится предоставлять знания, услуги, консультации и специализированные веб-услуги исследовательским сообществам в области гуманитарных и социальных наук. Наши веб-услуги и консультации посвящены тому, как использовать и исследовать базовые инструменты, которые могут обрабатывать и использовать текстовые данные, по крайней мере на четырех официальных языках Испании (испанский, каталонский, галисийский, баскский) и на английском, который является одним из наиболее распространенных и важных источников информации по многим дисциплинам гуманитарных исследований.

**SWELANG** Центр знаний языков Швеции CLARIN
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0228111325107

*Сферы компетенции.* Информационная служба, предлагающая советы по использованию цифровых языковых ресурсов и инструментов для шведского языка, языков меньшинств в Швеции, шведского языка жестов, шведских диалектов, а также других частей нематериального культурного наследия Швеции в тексте и речи, таких как языковая политика и планирование.

**Treebanking** Центр знаний CLARIN по банкам деревьев
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0227225216103

*Сферы компетенции.* Мы можем помочь сделать банки деревьев доступными, и доступными для поиска на наших двух веб-сайтах. Мы предоставляем документацию, инструкции и поддержку пользователей для онлайн-изучения доступных групп деревьев. Мы можем помочь в онлайн-построении древовидных банков LFG как проанализированных корпусов и в онлайн-редактировании древовидных банков универсальных зависимостей. Мы распространяем наши знания с помощью периодических обучающих программ и семинаров по банкам деревьев.

**TRTC** Ресурсы по терминологии и корпусы переводов
http://vonweber.elsnet.org/cgi/displayall_db.cgi?kcentres&0226163038116

*Сферы компетенции.* К-центр предоставляет информацию и обучение для пользователей по подготовке и документации ресурсов, связанных с переводом, в частности ресурсов терминологии и корпусов переводов. Сюда входят запросы, отправленные в службу поддержки, касающиеся инструментов, методов, данных и рекомендаций по поиску дополнительной экспертной поддержки. Сервис не фокусируется на языковых ресурсах на определенных языках, и не зависит от языка.

## ПРИЛОЖЕНИЕ 6. РОССИЙСКИЕ СТАНДАРТЫ НА ЛИР И СМЕЖНЫЕ ВОПРОСЫ

***Технический комитет 55 (ИСО ТК 37) Терминология, элементы данных и документация в бизнес-процессах и электронной торговле***

ГОСТ Р ИСО 704–2010 Терминологическая работа. Принципы и методы

ГОСТ Р ИСО 10241–1–2013 Терминологические статьи в стандартах. Часть 1: Общие требования и примеры оформления

ГОСТ Р ИСО 12615–2013 Библиографические ссылки и идентификаторы источников для терминологической работы

ГОСТ Р ИСО 12620–2012 Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов

ГОСТ Р ИСО 1951–2012 Представление и изложение словарных статей. Требования, рекомендации и информация

ГОСТ Р ИСО 23185–2013 Оценка и сравнительный анализ терминологических ресурсов. Общие концепции, принципы и требования

ГОСТ Р ИСО 24614–1-2013 Менеджмент языковых ресурсов. Пословная сегментация письменных текстов. Ч. 1. Основные концепции и общие принципы

ГОСТ Р ИСО 24615–2016 Управление языковыми ресурсами. Система синтаксического аннотирования (SynAF). Ч. 1. Синтаксическая модель

ГОСТ Р ИСО 24616–2013 Менеджмент языковых ресурсов. Многоязычная информационная система

ГОСТ Р ИСО 24619–2013 Менеджмент языковых ресурсов. Постоянная идентификация и устойчивый доступ

ГОСТ Р ИСО 30042–2016 Системы управления терминологией, базами знаний и контентом. Обмен терминологическими базами

***Технический комитет 191 (ИСО/ТК 46) Научно-техническая информация, библиотечное и издательское дело***

ГОСТ 7.0–99 Система стандартов по информации, библиотечному и издательскому делу. Информационно-библиотечная деятельность, библиография. Термины и определения

ГОСТ 7.1–2003 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Библиографическое описание. Общие требования и правила составления

ГОСТ 7.11–2004 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Сокращение слов и словосочетаний на иностранных европейских языках

ГОСТ 7.12–93 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Сокращение слов на русском языке. Общие требования и правила

ГОСТ 7.14–98 Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена информацией. Структура записи

ГОСТ 7.19–2001 Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена данными. Содержание записи

ГОСТ 7.24–2007 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению

ГОСТ 7.25–2001 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления

ГОСТ 7.26–80 Система стандартов по информации, библиотечному и издательскому делу. Библиотечное дело. Основные термины и определения

ГОСТ 7.27–80 Система стандартов по информации, библиотечному и издательскому делу. Научно-информационная деятельность. Основные термины и определения

ГОСТ 7.28–2002 Система стандартов по информации, библиотечному и издательскому делу. Расширенный набор символов латинского алфавита для обмена информацией

ГОСТ 7.29–80 Система стандартов по информации, библиотечному и издательскому делу. Представление расширенного кириллического алфавита для обмена информацией на магнитных лентах

ГОСТ 7.30–80 Система стандартов по информации, библиотечному и издательскому делу. Представление греческого алфавита для обмена информацией на магнитных лентах

ГОСТ 7.47–84 Система стандартов по информации, библиотечному и издательскому делу. Коммуникативный формат для словарей информационных языков и терминологических данных. Содержание записи

ГОСТ 7.49–84 Система стандартов по информации, библиотечному и издательскому делу. Рубрикатор ГАСНТИ. Структура, правила использования и ведения

ГОСТ 7.52–85 Система стандартов по информации, библиотечному и издательскому делу. Коммуникативный формат для обмена библиографическими данными на магнитной ленте. Поисковый образ документа

ГОСТ 7.53–2001 Система стандартов по информации, библиотечному и издательскому делу. Издания. Международная стандартная нумерация книг

ГОСТ 7.54–88 Система стандартов по информации, библиотечному и издательскому делу. Представление численных данных о свойствах веществ и материалов в научно-технических документах. Общие требования

ГОСТ 7.56–2002 Система стандартов по информации, библиотечному и издательскому делу. Издания. Международная стандартная нумерация сериальных изданий

ГОСТ 7.59–2003 Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к систематизации и предметизации

ГОСТ 7.60–2003 Система стандартов по информации, библиотечному и издательскому делу. Издания. Основные виды. Термины и определения

ГОСТ 7.64–90 Система стандартов по информации, библиотечному и издательскому делу. Представление дат и времени дня. Общие требования

ГОСТ 7.66–92 Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к координатному индексированию

ГОСТ 7.67–2003 (ИСО 3166–88) Коды названий стран

ГОСТ 7.67–2003 Система стандартов по информации, библиотечному и издательскому делу. Коды названий стран

ГОСТ 7.69–95 Система стандартов по информации, библиотечному и издательскому делу. Аудиовизуальные документы. Основные термины и определения

ГОСТ 7.70–2003 Система стандартов по информации, библиотечному и издательскому делу. Описание баз данных и машиночитаемых информационных массивов. Состав и обозначение характеристик

ГОСТ 7.72–96 Система стандартов по информации, библиотечному и издательскому делу. Коды физической формы документов

ГОСТ 7.73–96 Система стандартов по информации, библиотечному и издательскому делу. Поиск и распространение информации. Термины и определения

ГОСТ 7.74–96 Система стандартов по информации, библиотечному и издательскому делу. Информационно-поисковые языки. Термины и определения

ГОСТ 7.75–97 Система стандартов по информации, библиотечному и издательскому делу. Коды наименований языков

ГОСТ 7.76–96 Система стандартов по информации, библиотечному и издательскому делу. Комплектование фонда документов. Библиографирование. Каталогизация. Термины и определения

ГОСТ 7.77–98 Система стандартов по информации, библиотечному и издательскому делу. Межгосударственный рубрикатор научно-технической информации. Структура, правила использования и ведения

ГОСТ 7.79–2000 Система стандартов по информации, библиотечному и издательскому делу. Правила транслитерации кирилловского письма латинским алфавитом

ГОСТ 7.82–2001 Библиографическая запись. Библиографическое описание электронных ресурсов. Общие требования и правила составления

ГОСТ 7.83–2001 Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Основные виды и выходные сведения

ГОСТ 7.88–2003 Система стандартов по информации, библиотечному и издательскому делу. Правила сокращения заглавий и слов в заглавиях публикаций

ГОСТ 7.90–2007 Система стандартов по информации, библиотечному и издательскому делу. Универсальная десятичная классификация. Структура, правила ведения и индексирования

ГОСТ ИСО 8601–2001 Система стандартов по информации, библиотечному и издательскому делу. Представление дат и времени. Общие требования

ГОСТ Р 7.047–2008 Формат для представления на машиночитаемых носителях словарей информационных языков и терминологических данных. Содержание записи

ГОСТ Р ИСО 15489–1-2019 Система стандартов по информации, библиотечному и издательскому делу. Информация и документация. Управление документами. Часть 1. Понятия и принципы

ГОСТ Р ИСО 23081–1-2008 Система стандартов по информации, библиотечному и издательскому делу. Процессы управления документами. Метаданные для документов. Часть 1. Принципы

ГОСТ Р ИСО 26324–2015 Система стандартов по информации, библиотечному и издательскому делу. Система дискретных идентификаторов объекта

ГОСТ Р ИСО 30300–2015 Система стандартов по информации, библиотечному и издательскому делу. Информация и документация. Системы управления документами. Основные положения и словарь

ГОСТ ЭД1 7.4–90 Система стандартов по информации, библиотечному и издательскому делу. Издания. Выходные сведения

ГОСТ Р 7.0… –2020 (ИСО 25964–2:2013) Взаимодействие тезаурусов и других словарей

ГОСТ Р 7.0.100–2018 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Библиографическое описание. Общие требования и правила составления

ГОСТ Р 7.0.101–2018 Система стандартов по информации, библиотечному и издательскому делу. Информация и документация. Системы управления документами. Требования

ГОСТ Р 7.0.102–2018 Система стандартов по информации, библиотечному и издательскому делу. Профиль комплектования фондов научных библиотек. Структура. Индикаторы комплектования

ГОСТ Р 7.0.103–2018 Система стандартов по информации, библиотечному и издательскому делу. Библиотечно-информационное обслуживание. Термины и определения

ГОСТ Р 7.0.105–2020 Система стандартов по информации, библиотечному и издательскому делу. Номер государственной регистрации обязательного экземпляра печатного издания. Структура, оформление, использование

ГОСТ Р 7.0.60–2020 Система стандартов по информации, библиотечному и издательскому делу. Издания. Основные виды. Термины и определения

ГОСТ Р 7.0.61–2011 Система стандартов по информации, библиотечному и издательскому делу. Текущие государственные библиографические указатели. Общие требования и издательское оформление

ГОСТ Р 7.0.64–2018 Система стандартов по информации, библиотечному и издательскому делу. Представление дат и времени. Общие требования

ГОСТ Р 7.0.66–2010 Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к координатному индексированию

ГОСТ Р 7.0.83–2013 Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Основные виды и выходные сведения

ГОСТ Р 7.0.90–2016 Система стандартов по информации, библиотечному и издательскому делу. Универсальная десятичная классификация. Структура, правила ведения и индексирования

ГОСТ Р 7.0.91–2015 Система стандартов по информации, библиотечному и издательскому делу. Тезаурусы для информационного поиска

ГОСТ Р 7.0.92–2015 Система стандартов по информации, библиотечному и издательскому делу. Формат электронного обмена данными в книжном деле ONIX XML

ГОСТ Р 7.0.94–2015 Система стандартов по информации, библиотечному и издательскому делу. Комплектование библиотеки документами. Термины и определения

ГОСТ Р 7.0.95–2015 Система стандартов по информации, библиотечному и издательскому делу. Электронные документы. Основные виды, выходные сведения, технологические характеристики

ГОСТ Р 7.0.96–2016 Система стандартов по информации, библиотечному и издательскому делу. Электронные библиотеки. Основные виды. Структура. Технология формирования

ГОСТ Р 7.0.98–2018 Система стандартов по информации, библиотечному и издательскому делу. Международный стандартный идентификатор для библиотек и родственных организаций (ISIL)

ГОСТ Р 7.0.99–2018 Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования

ГОСТ Р 7.0.3–2006 Система стандартов по информации, библиотечному и издательскому делу. Издания. Основные элементы. Термины и определения

ГОСТ Р 7.0.5–2008 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая ссылка. Общие требования и правила составления

ГОСТ Р 7.0.6–2008 Система стандартов по информации, библиотечному и издательскому делу. Международный стандартный номер издания музыкального произведения (ISMN). Издательское оформление и использование

ГОСТ Р 7.0.8–2013 Система стандартов по информации, библиотечному и издательскому делу. Делопроизводство и архивное дело. Термины и определения

ГОСТ Р 7.0.9–2009 Система стандартов по информации, библиотечному и издательскому делу. Библиографическое обеспечение издательских и книготорговых процессов. Общие требования

ГОСТ Р 7.0.10–2019 Система стандартов по информации, библиотечному и издательскому делу. Набор элементов метаданных «Дублинское ядро». Основные (ядерные) элементы

ГОСТ Р 7.0.12–2011 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая запись. Сокращение слов и словосочетаний на русском языке. Общие требования и правила

ГОСТ Р 7.0.13–2011 Система стандартов по информации, библиотечному и издательскому делу. Карточки для каталогов и картотек, макет аннотированной карточки в издании. Общие требования и издательское оформление

ГОСТ Р 7.0.14–2011 Система стандартов по информации, библиотечному и издательскому делу. Справочные издания. Основные виды, структура и издательско-полиграфическое оформление

ГОСТ Р 7.0.15–2013 (ИСО 15924:2004) Система стандартов по информации, библиотечному и издательскому делу. Коды для представления наименований письменностей

ГОСТ Р 7.0.29–2010 Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Представление расширенного кириллического алфавита для обмена информацией

ГОСТ Р 7.0.30–2010 Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Представление греческого алфавита для обмена информацией

ГОСТ Р 7.0.34–2014 Система стандартов по информации, библиотечному и издательскому делу. Правила упрощенной транслитерации русского письма латинским алфавитом

ГОСТ Р 7.0.47–2008 Система стандартов по информации, библиотечному и издательскому делу. Формат для представления на машиночитаемых носителях словарей информационных языков и терминологических данных. Содержание записи

ГОСТ Р 7.0.49–2007 Система стандартов по информации, библиотечному и издательскому делу. Государственный рубрикатор научно-технической информации. Структура, правила использования и ведения

ГОСТ Р 7.0.52–2010 Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена библиографическими данными. Поисковый образ документа

ГОСТ Р 7.0.53–2007 Система стандартов по информации, библиотечному и издательскому делу. Издания. Международный стандартный книжный номер. Использование и издательское оформление

ГОСТ Р 7.0.56–2017 Система стандартов по информации, библиотечному и издательскому делу. Международный стандартный сериальный номер (ISSN). Издательское оформление и использование

### *Технический комитет 22 (ИСО JTC 1)*
### *Информационные технологии*

ГОСТ 33244–2015 Информационные технологии. Обучение, образование и подготовка. Концептуальная эталонная модель компетенции и связанных объектов

ГОСТ 33245–2015 Информационные технологии. Эталонная модель распределенного объекта контента (SCORM®) 2004, 3-я редакция. Часть 1. Обзор. Версия 1.1

ГОСТ 33246–2015 Информационные технологии. Обучение, образование и подготовка. Упаковка контента. Часть 1. Информационная модель

ГОСТ 33247–2015 (ISO/IEC 19788–1:2011) Межгосударственный стандарт «Информационные технологии». Обучение, образование и подготовка. Метаданные для образовательных ресурсов. Часть 1. Структура

ГОСТ 33707–2016 Информационные технологии. Словарь

ГОСТ 34.320–96 Информационные технологии. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной базы

ГОСТ 34.321–96 Информационные технологии. Система стандартов по базам данных. Эталонная модель управления данными

ГОСТ ISO/IEC 12785–2-2015 Информационные технологии. Обучение, образование и подготовка. Упаковка контента. Часть 2. XML привязка

ГОСТ ISO/IEC 19788–2-2015 Информационные технологии. Обучение, образование и подготовка. Метаданные для образовательных ресурсов. Часть 2. Элементы дублинского ядра

ГОСТ ISO/IEC 19788–3-2015 Информационные технологии. Обучение, образование и подготовка. Метаданные для образовательных ресурсов. Часть 3. Основной профиль применения

ГОСТ ISO/IEC 19788–5-2015 Информационные технологии. Обучение, образование и подготовка. Метаданные для образовательных ресурсов. Часть 5. Образовательные элементы

ГОСТ Р 52653–2006 Информационно-коммуникационные технологии в образовании. Термины и определения

ГОСТ Р 52656–2006 Информационно-коммуникационные технологии в образовании. Образовательные интернет-порталы федерального уровня. Общие требования

ГОСТ Р 52657–2006 Информационно-коммуникационные технологии в образовании. Образовательные интернет-порталы федерального уровня. Рубрикация информационных ресурсов

ГОСТ Р 53620–2009 Информационно-коммуникационные технологии в образовании. Электронные образовательные ресурсы. Общие положения

ГОСТ Р 55750–2013 Информационно-коммуникационные технологии в образовании. Метаданные электронных образовательных ресурсов. Общие положения

ГОСТ Р 57720–2017 Информационно-коммуникационные технологии в образовании. Структура информации электронного портфолио базовая

ГОСТ Р 57723–2017 Информационно-коммуникационные технологии в образовании. Системы электронно-библиотечные. Общие положения

ГОСТ Р 57724–2017 Информационно-коммуникационные технологии в образовании. Учебник электронный. Общие положения

ГОСТ Р 59168–2020 Информационные технологии. Стандарт базовой деловой лексики

## ПРИЛОЖЕНИЕ 7. ИНСТРУМЕНТЫ ЛИНГВИСТИЧЕСКИХ ТЕХНОЛОГИЙ

Настоящее приложение содержит аннотированный алфавитный указатель программных инструментов, применяемых в лингвистических технологиях, главным образом в аннотировании, а также компании – разработчики программ. Приложение составлено с использованием материалов сайтов [1] и [2], а также включает инструменты, упомянутые в настоящей монографии. После названия инструмента в скобках приводится ключ, в котором сообщается об уровне или стадии разработки. Ключи отсутствуют, если описание инструментов взято из других источников. Ссылка указывает на веб-страницу, в которой описывается инструмент, если такого описания нет – приводится ссылка на страницу энциклопедии, где представлено описание этого инструмента. Ссылки проверены по состоянию на февраль 2022 года.

**Ключи**

F: систематически документированный формат аннотаций

T: доступный инструмент для создания, отображения или поиска (W=Windows, U=Unix, M=macOS)

D: инструмент можно загрузить

P: цитируется статья, которая описывает инструмент или метод

P: другие виды ресурсов, такие как книги и ассоциации

C: методы и стандарты транскрибирования контента

**ABBYY https://www.abbyy.com/ru/** Компания-разработчик ресурсов и программ электронной лексикографии, а также OCR.

**Acrolinx** https://www.acrolinx.com/ платформа оптимизации контента, которая интегрируется с инструментами для письма, указывает на языковые проблемы и дает предложения по улучшению.

**Across** https://techtrans24.com/rus/cat-tools/across/ Программа автоматизированного перевода позволяет работать как с локальными файлами, так и с документами, расположенными на сервере заказчика. Есть возможность обработки многочисленных форматов.

**Advanced Glossing** (P) https://dobes.mpi.nl/documents/Advanced-Glossing1.pdf Схема лингвистической аннотации на всех основных лингвистических уровнях: фонетике и фонологии (включая интонацию), орфографических представлениях предложе-

[1] https://exmaralda.org/en/linguistic-annotation-wiki-en/

[2] http://martinweisser.org/corpora_site/annotation_tools.html

ний, словоформ и морфем, морфологических и синтаксических аннотациях единиц, категорий, конституентности и реляционной структуры, а также семантике. Она была разработана и предложена в проекте DOBES (Documentation of Endangered Languages).

**AGTK: Annotation Graph Toolkit** http://agtk.sourceforge.net/ Формальная структура для представления лингвистических аннотаций динамических данных. В этот инструментарий включены следующие приложения: MultiTrans: расшифровка многостороннего разговора; TableTrans: кодирование звука с наблюдением; TreeTrans: синтаксическая аннотация; InterTrans: подстрочная транскрипция текста.

**Alembic Workbench** (DT/U,W) https://nlp.stanford.edu/nlp/manual/AWB-overview.html Система аннотаций на основе SGML. Помимо обычных видов текстовых аннотаций, Workbench позволяет использовать различные виды специализированных аннотаций, включая аннотации со ссылками, различные типы определяемых пользователем указателей между тегами и общие шаблоны аннотаций – отношения, фреймы или события. Программное обеспечение имеет сложную визуализацию, работает на рабочих станциях Sun и свободно распространялось, сейчас недоступно.

**ANNIS – annotate** (TD) https://corpus-tools.org/annis/ Веб-приложение с открытым исходным кодом, которое обеспечивает доступ к многослойным, богато аннотированным корпусам. Оно обеспечивает функциональность поиска и визуализации сложных констелляций аннотаций на основе токенов и интервалов токенов, иерархических графовых структур, таких как синтаксические деревья или риторические аннотации, произвольных маркированных отношений, используемых например в аннотациях кореференции или синтаксисе зависимостей, а также для метаданных. Мультимодальные данные также могут быть выровнены с корпусами, а параллельные тексты могут быть выровнены на уровне слов, фраз или предложений.

**Annotation Graph Toolkit (AGTK)** (TDP) http://agtk.sourceforge.net/ Графы аннотаций – это формальная структура для представления лингвистических аннотаций динамических данных (временные ряды). Графы аннотаций абстрагируются от форматов файлов, схем кодирования и пользовательских интерфейсов, обеспечивая логический уровень для систем аннотаций.

**Annotation Pro** (TD) http://annotationpro.org/ Инструмент для аннотирования аудио- и текстовых файлов. Он позволяет пользователям создавать множество выровненных по времени слоев аннотаций, выбирать фрагменты записей, динамически масштабировать выделение или точку с помощью прокрутки мыши, воспроизводить, повторно воспроизводить и циклировать звук.

**Anvil (TP)** http://www.anvil-software.org/ Инструмент на базе Java, который позволяет многослойно аннотировать видео с помощью жестов, поз и дискурсивной информации. Используемые теги могут быть свободно определены и легко иерархически упорядочены. Anvil требует Java Media Framework и должен работать на Solaris, Windows и (возможно) Линукс. Формат видео – quicktime и AVI. Хранение и обмен данными – это XML. Он находится в свободном доступе для исследовательского сообщества.

**Anylexic** : https://anylexic.com/ Новое поколение программ управления терминологией, не привязанных к какой-либо конкретной терминологии. ПО может помочь вам на каждом этапе процесса управления терминологией перевода: создание, редактирование, поиск и обмен.

**ApSIC Xbench** https://www.xbench.net/?rd=y Обеспечивает простое и эффективное управление качеством и терминологией в одном пакете. Нужно загрузить файлы в любом из десятков поддерживаемых форматов CAT, и качество перевода перейдет на новый уровень.

**Arbil** https://archive.mpi.nl/forums/c/legacy-software/arbil/16 Инструмент для создания хорошо структурированного каталога метаданных, пригодного для архивирования.

**Arbil** https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045 Редактор метаданных, браузер и органайзер для IMDI, CMDI и аналогичных форматов метаданных.

**Ariadne** (FT) http://annotation.exmaralda.org/index.php?title=Ariadne Ariadne Вебсистема управления корпусом данных с сильным акцентом на многопользовательское управление ресурсами и обработку мультимодальных данных. Он разрабатывается в рамках совместного исследовательского центра (Sonderforschungsbereich) «Выравнивание в коммуникации» при Университете Билефельда. Основные функции Ariadne включают хранение и управление файлами, а также создание общей модели лингвистических форматов данных для упрощения задач, работающих с этими наборами данных. До сих пор были разработаны процедуры преобразования между документами Praat и ELAN и внутренним форматом данных, некоторые из них находятся в активной разработке.

**ATLAS** (FP) https://sourceforge.net/projects/jatlas/ Архитектура и инструменты для лингвистического анализа систем. ATLAS – это совместная инициатива группы организаций по созданию архитектуры аннотаций общего назначения и формата обмена данными. Отправной точкой является графовая модель аннотации с некоторыми значительными обобщениями. Проект ATLAS больше не поддерживается активно – Java-реализация API ATLAS существует под именем jATLAS. Одна из программных реализаций, поддерживающих аннотацию ATLAS, это Calisto.

**ATLAS (2)** (TD) https://www.researchgate.net/publication/261036948 Инструмент аннотаций, разработанный в Университете Ульма.

**ATOMIC** (TD) https://corpus-tools.org/atomic/ Многоуровневый инструмент аннотирования корпусов, разработанный в рамках исследовательского проекта LinkType в Йенском университете имени Фридриха Шиллера.

**Atril DejaVu** https://atril.com/ Проприетарная система автоматизированного перевода, разработанная испанской компанией Atril Language Engineering

**Audiamus** (TDP) https://www.nthieberger.net/audiamus.htm Инструмент для создания корпусов, связанных транскриптов и оцифрованных носителей. Он разработан с учетом ключевых принципов повторного использования и доступности данных. Основная предпосылка в том, что каждый пример, цитируемый в грамматике, должен быть подтвержден архивным источником, если это возможно.

**BibTeX** https://ru.wikipedia.org/wiki/BibTeX Программное обеспечение для создания форматированных списков библиографии.

**BlitzScribe** (T) https://www.media.mit.edu/projects/blitzscribe-speech-transcription-for-the-human-speechome-project/overview / Новый подход к транскрипции речи, основанный на требованиях современных мультимедийных корпусов.

**Bonito** (TD) https://ucnk.ff.cuni.cz/bonito/ Программное обеспечение для управления корпусом и запросов для чешского Национального корпуса.

**Callisto** (TD/W,U,M) https://www.mitre.org/research/technology-transfer/open-source-software /callisto-0 Инструмент аннотации Callisto был разработан для поддержки лингвистической аннотации текстовых источников для любого языка, поддерживающего Unicode. Поддержка Standoff-аннотаций, предоставляемая компанией jATLAS, позволяет решать практически любую задачу аннотации. Модульная конструкция Callisto позволяет расширить ее с помощью компонентов пользовательского интерфейса, специфичных для конкретной области. Возможности редактирования тегов по умолчанию предоставляются с помощью выделенного текстового дисплея и таблиц атрибутов тегов. По мере разработки доменных компонентов расширения они могут быть интегрированы в ядро Callisto, чтобы стать частью стандартного набора доступных компонентов. Callisto написан на Java.

**CasualTranscriber** (TD/M) https://sites.google.com/site/casualconcj/yutiriti-puroguramu/casualtranscriber Утилита, разработанная компанией Yasu Imao для облегчения транскрипции аудио- и видеотекста в распространенные текстовые форматы для Mac OS X. ПО может обрабатывать большинство типов аудио- и видеофайлов QuickTime, доступно для бесплатной загрузки с сайта разработчика.

**CATMA** (TD) https://catma.de/webarchive/catma-4.0/home.html Computer Aided Textual Markup and Analysis Программа с открытым исходным кодом, ориентированная на текстовую разметку и анализ. Разработана в Гамбургском университете как инструмент для литературоведов, студентов и других лиц, интересующихся цифровой гуманитаристикой.

**C-BAS** (T/W) https://eller.arizona.edu/search/node/C-BAS Система поведенческих аннотаций C# (C-BAS), разработанная в Аризонском университете – это компьютерное приложение для поведенческих аннотаций, разработанное для помощи исследователям в кодировании событий на видео или звуковых дорожках.

**CCA** (P) https://www.sscnet.ucla.edu/soc/faculty/schegloff/prosody Программа иллюстрирует традиционный стиль транскрипции среди исследователей, работающих над анализом устной речи.

**CES** (FC) https://www.cs.vassar.edu/CES/ Этот документ является первой версией стандарта Corpus Encoding Standard (CES), входящего в состав EAGLES. CES разработан для использования в исследованиях и приложениях языковой инженерии, чтобы служить широко принятым набором стандартов кодирования для корпусной работы в приложениях обработки естественного языка. CES – это приложение SGML, совместимое со спецификациями TEI.

**CharWrite** http://emeld.org/tools/charwrite.cfm Средство ввода Unicode для Интернета.

**CHILDES** (FTDPRC/W,M) https://childes.talkbank.org/ Программы для Windows и Macintosh, которые позволяют анализировать базу данных проекта CHILDES, а также выравнивать текст в речь и видео.

**CLaRK** http://bultreebank.org/bg/clark/ Основанная на XML система для разработки корпусов и включает в себя редактор Unicode XML, язык XPath для навигации в документах XML, механизм XSLT для преобразования документов XML, каскадные регулярные грамматики, ограничения для документов XML, токенизаторы, инструмент согласования, извлечение, удаление и другие инструменты. Система реализована на JAVA.

**Classical Text Editor** (TD) https://cte.oeaw.ac.at Классический текстовый редактор был разработан для того, чтобы ученый, работающий над критическим изданием или текстом с комментариями или переводом, мог подготовить готовый к работе экземпляр или электронную публикацию, не затрудняясь оформлением и корректурой страниц.

**CMDI Maker** http://cmdi-maker.uni-koeln.de/ Простое в использовании веб-приложение HTML5 для быстрого создания научных метаданных для ЛИР.

**Coala** https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Coala Инструмент для преобразования электронных таблиц метаданных для различных наборов речевых данных в стандартизированные файлы CMDI.

**COMEDI** https://clarino.uib.no/comedi/page Веб-редактор для метаданных, соответствующих правилам CMDI.

**CSAE** (TDC/W) http://transcription.projects.linguistics.ucsb.edu/tools.html В проекте Corpus of Spoken American English было разработано несколько инструментов, включая VoiceWalker, инструмент транскрипции для аудио и видео, а также SoundWriter, который позволяет выравнивать части транскриптов со звуковыми файлами с помощью временных кодов SMPTE. Также разработан свой собственный набор транскрипционных конвенций.

**CSLU Toolkit** (TDPRC/W) https://web.archive.org/web/20110817012415/ http://www.cslu.ogi. edu/toolkit/ Разработанный в Центре понимания устной речи Орегонского Университета (CSLU) – полный набор бесплатных инструментов для сбора и транскрибирования речи. Включает интерактивную программу отображения речи (Speech View), которая позволяет пользователю выравнивать транскрипты со звуковыми файлами. Инструментарий также содержит средства чтения спектрограмм и акустической фонетики, механизм распознавания речи, синтезатор речи, компонент лицевой анимации и интеграционный инструмент для создания собственной системы разговорного языка. CSLU Toolkit также имеет кодировку ASCII для фонетической транскрипции.

**CSTR** https://www.cstr.ed.ac.uk Центр исследования речевых технологий английского языка Эдинбургского университета, разработчик широкого класса программ по анализу и синтезу устной речи. The Centre for Speech Technology Research.

**CuPED** (TD) https://exp-platform.com/cuped/ Настраиваемая презентация документов. ELAN – инструмент для преобразования выровненных по времени транскриптов, которые производятся компанией ELAN, в различные форматы презентаций. В то время как ELAN и аналогичные инструменты обычно концентрируются на производстве транскриптов, пригодных для долгосрочного архивного хранения, CuPED стремится обеспечить простое и удобное для пользователя средство преобразования этих архивных источников в форматы, более доступные для широкой аудитории.

**CWB/CQP** (TP/U) http://cwb.sourceforge.net/ IMS Open Corpus Workbench (CWB) Был разработан для поддержки полнотекстового поиска в больших текстовых ресурсах в области лексикографии и терминологии. Центральным компонентом является «Corpus Query Processor» (CQP), специализированная поисковая система для лингвистических исследований.

**DAISY** (FTP/U,W) https://daisy.org/ DAISY Consortium Всемирная коалиция библиотек и учреждений, обслуживающих лиц с ограниченными возможностями,

разрабатывающая открытые стандарты, инструменты и методы для следующего поколения «цифровых говорящих книг» (DTB).

**DAMSL** (FTRC/U,W) https://staff.fnwi.uva.nl/r.fernandezrovira/teaching/upf-materials/Summary_DAMSL.pdf DAMSL – (Dialog Act Markup in Several Layers) Разметка диалогового акта в нескольких слоях, определяет набор примитивных коммуникативных действий, которые могут быть использованы для анализа диалогов.

**DART** http://martinweisser.org/ling_soft.html#DART Инструмент аннотации и анализа, разработанный для полуавтоматического аннотирования устных (транскрибированных) диалогов на уровнях синтаксиса, прагматики (речевые акты), (поверхностной) полярности, семантики (темы) и семантико-прагматики.

**Delta (TP/U,W)** http://annotation.exmaralda.org/index.php?title=Delta Eloquent Technologies Разработала инструментарий преобразования текста в речь, который синтезирует речь с помощью многоуровневого представления текста, называемого Дельтой.

**Dexter** (T) http://www.dextercoder.org/ Бесплатный набор кроссплатформенных программных средств на основе Java, позволяющих выполнять качественное кодирование корпусных текстов. Dexter написан специально для трех вещей: данные разговорного языка, данные, собранные исследователями, и анализ явлений на уровне дискурса. Dexter Coder позволяет определять и добавлять аннотации к документу. С помощью Coder вы можете выполнять сложный поиск текста и кодов, а также определенные количественные анализы. Все аннотации сохраняются в отдельном файле standoff XML. Входные данные могут быть в различных форматах; он преобразуется в XML, чтобы включить автономную разметку, которая в свою очередь позволяет проводить неограниченное количество анализов, не затрагивая исходные данные.

**DOLMEN** (T) http://julieneychenne.info/dolmen/ Кроссплатформенный инструментарий с открытым исходным кодом для корпусной лингвистики.

**DRS** (TD) http://thedrs.sourceforge.net/ DRS, или цифровая система воспроизведения – это инструмент для компьютерного качественного анализа данных (CAQDAS) следующего поколения.

**DSpace** https://ru.wikipedia.org/wiki/DSpace Открытое, свободное (лицензия BSD) кроссплатформенное J2 EE-приложение, платформа институционального репозитория для долгосрочного хранения цифровых материалов, используемых в академических исследованиях.

**EasyAlign** (TD) http://latlcui.unige.ch/phonetique/easyalign.php EasyAlign Автоматический инструмент фонетического выравнивания для непрерывной речи в режиме Praat. Речь можно выровнять по орфографической или фонетической транскрипции. Это требует нескольких незначительных ручных шагов, и в результате получается многоуровневая аннотация внутри текстовой сетки, состоящей из фонетического, слогового, лексического и речевого уровней.

**eHumanities Desktop** (TPC) https://www.texttechnologylab.org/applications/ehumanities-desktop/ Кафедра цифровой гуманитаристики Университета Гете во Франкфурте (Германия) поддерживает набор лингвистических инструментов и функций для поддержки исследователей в области гуманитарных наук. Интерфейс инструментов основан на браузере и, несмотря на свою мощь, довольно прост в использовании.

**ELAN** (FTD) https://www.mpi.nl/corpus/html/elan/ ELAN (Eudico Linguistic Annotator) Инструмент для многоуровневой аннотации видео и / или аудио, разработанный Институтом Макса Планка (MPI) в Неймегене. ELAN написан на JAVA и работает на операционных системах Windows, Macintosh и Linux. ELAN использует временную модель данных и формат XML. Одним из важных сообществ пользователей является сообщество вымирающих языков (см. проекты DOBES и E-MELD).

**Elexifier** https://elexifier.elex.is/ Облачный сервис преобразования словарей PDF и XML в стандартизированный машиночитаемый формат.

**EMU-SDMS** (FTDP/UW) http://ips-lmu.github.io/EMU.html Система EMU обеспечивает последовательный доступ к разнообразным речевым базам данных, а также средства для легкого извлечения статистических данных и поддержки создания баз данных. EMU допускает сложные многоуровневые и иерархические структуры, которые могут быть построены с использованием комбинации ручного и автоматического аннотирования. Сценарий EMU был написан, чтобы импортировать некоторые файлы; структура EMU способна выражать информационное содержание аннотаций Partitur.

**evoTerm** http://www.evoterm.net/en/Details.aspx Система централизованного управления и хранения терминологии, доступная через Интернет. Демо-версия доступна для тестирования платформы.

**F4** (TD) https://www.audiotranskription.de/f4 Комбинация текстового редактора / медиаплеера для транскрипции аудио или видео.

**Fedora** https://ru.wikipedia.org/wiki/Fedora Дистрибутив Linux, спонсируемый фирмой Red Hat с целью построения целостной свободной операционной системы.

**Feeltrace** (TD) https://www.allaboutux.org/feeltrace Инструмент маркировки для двух измерений эмоций, разработанный в Королевском университете Белфаста Родди Коуи и его коллегами. Он позволяет отслеживать воспринимаемое эмоциональное состояние непрерывно во времени, по двум основным измерениям эмоций – активации и оценке.

**Feldpartitur** (T) http://www.feldpartitur.de/en Программное обеспечение, предназначенное для помощи качественным социальным исследователям в расшифровке видеоданных. Видеоинформация собирается диаграммно по временной шкале с помощью письменных знаков и визуальных символов.

**Festival** (TD/U) https://www.cstr.ed.ac.uk/projects/festival/ Программное обеспечение CSTR в Эдинбурге, первоначально разработанное для использования в синтезе речи, было обобщено и применено также к анализу баз данных. Festival использует гетерогенные графы отношений для представления лингвистической информации.

**flashterm** https://www.flashterm.eu/en/ Система управления терминологией с богатыми фунуциональными возможностями, рассчитанная на использование с 200 языками.

**FLEX** (Fieldworks Language Explorer) http://lingtransoft.info/apps/flex-fieldworks-language-explorer Компонент лексических и текстовых инструментов SIL FieldWorks. Это настольное приложение с открытым исходным кодом, предназначенное для того чтобы помочь полевым лингвистам выполнять многие общие задачи.

**Flora** https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecFlora-2 Объектно-ориентированный язык базы знаний и среда разработки приложений.

**FOLKER** (TD) http://agd.ids-mannheim.de/folker.shtml Инструмент транскрипции, предназначенный для поддержки транскрипции в фольклорном корпусе, исследо-

вательском и учебном корпусе разговорного немецкого языка в Институте немецкого языка. FOLKER построен на технологии EXMARaLDA и оптимизирован для транскрибирования многопартийного взаимодействия в соответствии с соглашениями GAT.

**FORM** (C) http://shachi.org/resources/1412 FORM Схема аннотации жестов, предназначенная для захвата кинематической информации в жесте из видеозаписей выступающих. В настоящее время проект FORM создает подробную базу данных видео с комментариями жестов, хранящихся в формате графика аннотаций. Это позволит дополнить жестовую информацию другой лингвистической информацией, такой как синтаксический анализ предложений, сопровождающих жесты, структура дискурса, интонационная информация и т.д. В настоящее время FORM использует Anvil, однако разрабатывается FORMTool, свой собственный инструмент аннотации жестов с открытым исходным кодом. Этот инструмент будет полезен всем, кто хочет кодировать лингвистическую информацию из видео в формате аннотационного графа.

**FSA's** (TD) http://annotation.exmaralda.org/index.php?title=FSA%27s Концепция и инструментарий для n-арных конечных автоматов, которые обеспечивают полезную модель для выражения, создания и поиска многомерных лингвистических аннотаций.

**GATE** (FTDP/U) https://gate.ac.uk Система обработки естественного языка с открытым исходным кодом, использующая наборы компонентов на языке Java. GATE имеет реализацию архитектуры Tipster, а также графические инструменты для визуализации данных, аннотации, оценки и управления технологическими процессами. Распространяется бесплатно для исследований.

**Glossa** (T) https://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-2008/pdf/159_paper.pdf – Веб-система корпусных запросов, которая сочетает в себе выразительность обычных языков запросов с удобством для пользователя графического интерфейса.

**Grammar explorer** https://eltngl.com/search/productOverview.do?N=200++4294918606&Ntk=NGL%7CP_EPI&Ntt=PRO0000000538%7C104546375515073348147547348231402151049&Ntx=mode%2Bmatchallpartial&homePage=false Инструмент для изучения охвата грамматик. Инструмент работает, по сути, как кодировщик: нужно выбрать предложение или другую грамматическую единицу и попытаться «закодировать» эту единицу, используя термины грамматики. Инструмент проводит пользователя через грамматику, представляя доступные варианты; он также сообщает синтагматические последствия этого выбора (т.е. то, какая структура создается).

**Gsearch** (T/U) https://macsecurity.net/view/398-gsearch-extension-1-0-mac Инструмент для поиска помеченных корпусов. Запросы формулируются в два этапа. Во-первых, пользователь задает контекстно-свободную грамматику, которая используется для разбора данного корпуса и преобразования его тегов в стандартный набор. Во-вторых, выражение поиска использует слова, терминалы и нетерминалы, предусмотренные корпусом и грамматикой. Структурированный вывод может быть визуализирован с помощью программы видового дерева.

**HeidelTime** (TDP) https://code.google.com/archive/p/heideltime Многоязычный междисциплинарный темпоральный таггер, разработанный исследовательской группой по системам баз данных Гейдельбергского университета. Он извлекает временные выражения из документов и нормализует их в соответствии со стандартом аннотаций TimeML.

**HIAT** (FTDPRC/W), **HIAT-DOS** (T) (Review) https://www.ehlich-berlin.de/HIAT/HIAT.htm Философия HIAT включает в себя понятие литературной транскрипции, которая включает систематические отклонения от стандартного орфографического перевода предмета, но таким образом, который имеет смысл для человека знакомого с орфографической системой в целом. Предусмотрены методы аннотирования просодии, невербальной коммуникации и т.д. HIAT использует EXMARaLDA, в которую интегрирован импортный фильтр для преобразования данных HIAT-DOS в более устойчивый формат XML.

**Hyperlex** (TP/U) https://arxiv.org/pdf/1608.02117.pdf Hyperlex Разработанная в поддержку полевого проекта, обеспечивает HTML-опосредованный доступ к лексике, записям речи и парадигматическим каталогам для нескольких языков.

**ikannotate** (TD) http://www.iikt.ovgu.de/Lehrstühle+und+Fachgebiete/KS/Forschung/ikannotate.html#screenshots Программа транскрипции, аннотации и маркировки, разработанная на кафедре когнитивных систем Университета Магдебурга имени Отто фон Герике.

**Interplex** http://www.fourwillows.com/interplex.html Программное обеспечение глоссария для устных и письменных переводчиков.

**InterpretBank** http://www.interpretbank.com/ Инструмент управления терминологией, специально разработанный для переводчиков. Это помогает создавать, изучать и искать глоссарии даже в кабине.

**Interpreters Help** https://interpretershelp.com/ Инструменты для устных переводчиков конференций.

**INTEX** (F/U,W,M) http://intex.univ-fcomte.fr/INTEX.htm Лингвистическая среда разработки, которая включает в себя инструменты для создания и поддержания обширных лексических ресурсов, а также морфологических и синтаксических грамматик. INTEX может строить лемматизированные конкордации и индексы больших текстов по отношению ко всем типам паттернов конечных состояний. INTEX используется как информационно-поисковая система для анализа художественных текстов, для количественной оценки языковых вариаций, для обучения вторым языкам, как терминологический экстрактор и т.д.

**ISIP** (TDP/U) https://www.researchgate.net/publication/271849557_Speech_Acoustic_Unit_Segmentation_Using_Hierarchical_Dirichlet_Processes Сегментатор и транскрибер. В Институте обработки сигналов и информации (ISIP) в штате Миссисипи созданы бесплатные инструменты, изначально оптимизированные для сегментации, транскрибирования и аннотирования телефонных разговоров.

**i-Term** http://www.danterm.dk/ Современный инструмент управления терминологией и знаниями, который позволяет хранить, структурировать и искать знания о концепциях в Интернете.

**J-Safran** (TD) http://rapsodis.loria.fr/jsafran/demo.html J-Safran (Java Syntaxico-semantic French Analyser) Программная платформа, включающая следующие функциональные возможности: синтаксический анализ и анализ зависимостей; аннотирование текстов; анализ структуры зависимостей во французских текстах; обучение синтаксического анализатора на наборе аннотированных файлов; оценка результатов синтаксического анализа по отношению к эталонному корпусу; автоматическая аннотация глаголов с семантическими ролями.

**JSON-LD** https://en.wikipedia.org/wiki/JSON-LD Нотация объектов JavaScript для связанных данных – метод кодирования связанных данных с использованием JSON. Разработана на основе концепции «контекста» для предоставления дополнительных сопоставлений из JSON в модель RDF. Контекст связывает свойства объекта в документе JSON с концепциями онтологии.

**JTrans** (TD) https://hal.inria.fr/inria-00431398 Программное обеспечение с открытым исходным кодом на JAVA для полуавтоматического выравнивания текста и речи. Разработан, чтобы быть интуитивно понятным и простым в использовании, с акцентом на дизайн GUI. Цель JTrans состоит в том, чтобы позволить пользователю контролировать и проверять на лету алгоритмы автоматического выравнивания.

**Kinoath** https://archive.mpi.nl/forums/t/kinoath-software-info/1048 Приложение, основной целью которого является соединение данных родства с архивными данными, такими как аудио, видео или письменные ресурсы, а также тесная интеграция с архивным программным обеспечением, таким как Arbil.

**Knowtator** (TDPC/W,U,M) http://knowtator.sourceforge.net Универсальный инструмент текстовых аннотаций, который интегрирован с системой представления знаний Protégé. Knowtator облегчает ручное создание учебных и оценочных корпусов для различных задач обработки биомедицинского языка. Knowtator разработан в качестве плагина Protégé, который использует возможности представления знаний Protégé для указания схем аннотаций. Уникальное преимущество Knowtator перед другими инструментами аннотаций – это простота, с которой сложные схемы аннотаций (например схемы, которые имеют ограниченные отношения между типами сущностей) могут быть определены и включены в использование. Кроме того, поскольку схемы аннотаций определяются с использованием онтологии Protégé, легко включить знания предметной области в схему аннотаций для семантической аннотации.

**LACITO** (FTD) https://pangloss.cnrs.fr/index.html Проект нацелен на предоставление инструментов и форматов для лингвистических и антропологических полевых данных. Интересной особенностью является использование XML-разметки с DTD, который поддерживает транскрипцию, фразовые и пословные межстрочные переводы, а также аудиоссылки. Приведены некоторые таблицы стилей XSL, которые иллюстрируют потенциальную мощь XML-разметки для поддержки просмотра вебстраниц для материалов этого типа, предоставляя доступ к тексту и звуку.

**Lamus** https://archive.mpi.nl/forums/c/legacy-software/lamus/18 Инструмент для загрузки данных и метаданных в архив DoBes, а также для управления существующими коллекциями.

**Lexonomy** https://www.lexonomy.eu/ Облачная система для написания, а также для публикации онлайн-словарей, которая хорошо масштабируется для адаптации к крупным словарным проектам, а также к небольшим лексикографическим работам, таким как редактирование и онлайн-публикация тематических глоссариев или терминологических ресурсов.

**LEXUS** https://www.aclweb.org/anthology/L06-1078/ Инструмент предоставляет гибкую структуру для поддержания лексической структуры и содержания. Это первая реализация модели Lexical Markup Framework, которая в настоящее время разрабатывается в ISO TC37 / SC4. Среди его возможностей – возможность создавать структуры лексики, манипулировать содержанием и использовать типизированные отношения.

**LinguaLinks-** https://www.sil.org/resources/publications/lingualinks электронная система SIL поддержки производительности труда языковых работников, основанная на «объектно-ориентированной» вычислительной среде для лингвистических, литературоведческих и антропологических исследований.

**LogiTerm Web** https://terminotix.com/index.asp?content=category&cat=4&lang=en Система имеет удобный веб-интерфейс, который обеспечивает доступ к четырем базам данных: Терминология, Битексты, Полный текст и Справочник. Записи терминологии для Терминологии можно создавать, изменять и просматривать в веб-интерфейсе или Microsoft Word.

**LRS** (Language Resource Switchboard) https://www.aclweb.org/anthology/J18-4002.pdf Коммутатор языковых ресурсов. Инструмент поиска программ обработки языковых данных по их свойствам, реализован в VLO CLARIN.

**LT** (T/U,W), [LT-XML, LT-XML2, LT-TTT2] https://www.ltg.ed.ac.uk/software/ Программное обеспечение на основе XML для неглубокой лингвистической обработки текста предоставляет компоненты NLP для различных задач обработки текста, таких как маркировка и разделение предложений, разделение фрагментов и распознавание именованных сущностей, таггер и лемматизатор.

**LTS** (Lionbridge Translation Workspace**)** https://geoworkz.com/Support/Public/Introduction_to_the_Translation_Workspace.pdf Программа автоматизированного перевода, ориентированная на большие компании и большие объемы перевода.

**MacShapa** (TP) http://acs.ist.psu.edu/projects/dismal/macshapa.html Гибкий инструмент видеоаннотации, который поддерживает пользовательские схемы кодирования и поставляется с хорошим поиском и статистикой. Кроме того, он позволяет пользователю выводить данные непосредственно в электронную таблицу. MacSHAPA управляет высококачественными видеомагнитофонами Panasonic JVC либо Quicktime.

**MacVissta** (TD) https://sourceforge.net/projects/macvissta/ Система с открытым исходным кодом, разработанная для Mac OS X. Это программное обеспечение позволяет визуализировать несколько синхронизированных по времени видео в сочетании с одним или несколькими графическими диаграммами / аннотациями.

**MATE** (FT) https://www.aclweb.org/anthology/W00-17.pdf Многосторонний проект направлен на разработку основанного на SGML стандарта аннотирования корпусов устных диалогов и инструментов для «повышения эффективности процессов приобретения и извлечения знаний». MATE реализует концепцию разметки противостояния, которая связана с предложениями CES и TEI. Проект MATE был предшественником проекта NITE.

**MDF** (Multi-Dictionary Formatter) https://software.sil.org/shoebox/mdf/#:~:text=The%20Multi%2DDictionary%20Formatter%20(MDF,dictionary%2Dmaking%20process%20improves%20quality Инструмент преобразования лексической базы данных.

**MediaStreams** (P) https://developer.mozilla.org/ru/docs/Web/API/MediaStream Позволяет пользователям создавать многослойные, знаковые аннотации видеоконтента. Используя более 2500 знаковых примитивов, пользователь комбинирует значки для создания каскадной иерархической структуры.

**MediaTagger** (P) https://www.portablefreeware.com/?id=141 Инструмент для транскрипции и анализа оцифрованных видеофильмов на Apple Macintosh. Основная идея MediaTagger состоит в том, чтобы выбрать временной срез видео- (+аудио) фильма и пометить его транскрипционным текстом или кодом. Теги могут содержать

свободный текст или фиксированный код из разработанной пользователем системы классификации. Эта маркировка может быть выполнена на любом количестве уровней, причем уровни могут быть абсолютно независимыми.

**memoQ** https://www.memoq.com/ Проприетарный пакет программ автоматизированного перевода, работающий в операционной системе Microsoft Windows. Разработан венгерской компанией memoQ Translation Technologies (до середины 2018 года называлась Kilgray Translation Technologies).

**Memsource** https://www.memsource.com/ Проприетарное облачное окружение для автоматизированного перевода, разработанное чешской компанией Memsource.

**MMAX** (TD) http://mmax2.net/index.html Инструмент для мультимодальных аннотаций в XML. Он поддерживает аннотацию противостояния.

**Morphisto** (TD) https://code.google.com/archive/p/morphisto/ Морфологический анализатор и генератор немецких словоформ. Лексикон Morphisto был первоначально разработан в Институте немецкого языка Мангейма (Германия) в рамках проекта TextGrid.

**MPEG** (FPR) https://ru.wikipedia.org/wiki/MPEG-4 Установленный в качестве стандарта ISO/IEC в начале 1999 г. MPEG-4 предоставляет стандарты для потоковой передачи интерактивных мультимедийных данных. MPEG-4 был частично интегрирован с форматом файлов Apple Quicktime, используя QuickTime в качестве отправной точки для разработки единого цифрового формата хранения мультимедиа для спецификации MPEG-4.

**MPI** (FT/UWM) https://www.mpi.nl/world/tg/lapp/lapp.html Группа MPI Language and Cognition Group разработала целый ряд инструментов для работы с аннотированной речью и видеоданными. EUDICO – это универсальный верстак для корпусной лингвистики, написанный на Java. CAVA – это набор программ, предназначенных для ученых-гуманитариев, включая редактор транскрипции для цифровой транскрипции аналогового видео на ПК и программы Macintosh под названием MediaTagger для создания и поиска многоуровневых аннотаций цифрового видео в формате QuickTime. Доступен новый мультиплатформенный инструмент для управления транскрипциями, подобными чату, и выровненными речевыми данными-Spoken Childes Tool. Многие из этих инструментов доступны бесплатно на академических сайтах.

**MTRANS** (TD) http://www.laslab.org/conference_paper/mtrans-a-multi-channel-multi-tier-speech-annotation-tool/ Свободно доступный инструмент для аннотирования многоканальной речи. Программное обеспечение обеспечивает визуальную и слуховую гибкость в представлении информации, чтобы помочь в транскрипции разговоров с многими участниками.

**Multitool** (TD) https://www.researchgate.net/publication/2925619_Annotations_and_Tools_ for_an_Activity_Based_Spoken_Language_Corpus Multitool (университет Гетеборга) Кроссплатформенное мультимодальное программное обеспечение для транскрипции и анализа. Он может быть использован для создания синхронизированных по времени транскрипций с использованием аудио и видео. Это облегчает импорт и экспорт таких транскрипций.

**NEGRA** (FTPC/U) http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html Корпус NEGRA представляет собой тип treebank,но с новой схемой аннотации для разрывных составляющих. Визуальный формат, формат аннотации и экви-

валент Treebank доступны. Инструмент, который поддерживает взаимодействие человека и машины при построении синтаксических деревьев, – Annotate.

**NIEUW** https://www.ldc.upenn.edu/collaborations/current-projects/nieuw Novel Incentives and Workflows in Linguistic Data Collection and Annotation Новые инструменты и технологии по сбору и аннотированию лингвистических данных.

**NITE** https://sourceforge.net/projects/nite/ В проекте NITE (Natural Interactivity Tools Engineering) создано несколько наборов инструментов для многоуровневого, межуровневого и кросс-модального аннотирования, поиска и использования многопартийных естественных интерактивных диалоговых данных человек – человек и человеко-машина. Наборы инструментов NITE становятся способными удовлетворять потребности пользователей, которые хотят кодировать (или аннотировать) и исследовать естественную интерактивную и мультимодальную коммуникацию, включая межуровневое и кросс-модальное кодирование. Формальная основа для обработки данных в NITE определяется в объектной модели NITE (NOM).

**NooJ** http://www.nooj-association.org Инструмент для обработки корпуса и платформа / среда разработки лингвистической инженерии. Может использоваться как процессор корпуса, система извлечения информации, средство извлечения терминологии, инструмент разработки машинного перевода, инструмент для обучения лингвистике и компьютерной лингвистике**.** Позволяет формализовать несколько уровней языковых явлений: орфография, лексика для простых слов, многословные единицы и замороженные выражения, флективная, деривационная и продуктивная морфология, локальный, структурный синтаксис и трансформационный синтаксис**.**

**Observer** (T/W) https://www.noldus.com/observer-xt Коммерческий инструмент для классификации и регистрации событий. При помощи видеоверсии инструмента можно создавать выровненные по времени аннотации видеозаписей, используя регистратор событий. Временной паттерн наблюдений может быть отображен с помощью временной диаграммы событий, и могут быть созданы различные сводные статистические данные. Программное обеспечение было разработано компанией Noldus IT и работает на платформах MS Windows.

**Ontotex**t https://en.wikipedia.org/wiki/Ontotext Компания, которая ведет разработку программного обеспечения на основе языков и стандартов Семантической сети, в частности RDF, OWL и SPARQL. Ontotext наиболее известен благодаря движку базы данных семантического графа Ontotext GraphDB.

**OpenCCG** https://github.com/OpenCCG/openccg Библиотека для синтаксического анализа и реализации с помощью CCG. Включает в себя мини-грамматики для инуитского, баскского и других языков.

**Pacx** (D) http://pacx.sourceforge.net/ Платформа для аннотированных корпусов в интегрированном инструменте XML для корпусных лингвистов, построенных на Eclipse, Vex, Subversive и т.д. для создания и редактирования транскрипций и аннотаций, запросов, управления версиями контролируемых данных, а также построения переадресуемого корпуса.

**PALinkA** (TD) http://rgcl.wlv.ac.uk/?s=PALinkA PALinkA Универсальный (ранее известный как cLinkA) инструмент аннотации, который может использоваться для различных задач аннотирования.

**Paraconc** (TD) http://www.athel.com/para.html Коммерческий двуязычный или многоязычный конкордансер, который может быть использован в контрастивном анализе, изучении языка и обучении переводу.

**Partitur** (FT) https://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html# Partitur Баварский архив речевых сигналов создал формат Partitur, основанный на их опыте работы с различными речевыми базами данных. Цель состояла в том, чтобы создать открытый (т.е. расширяемый), надежный формат для представления результатов многих различных исследовательских лабораторий в общем источнике.

**PAULA** (F) https://www.uni-due.de/imperia/md/content/computerlinguistik/sustainability.pdf Лингвистическая база данных: аннотация и поиск SFB 632. Это формат представления standoff на основе XML, который был разработан для представления данных, аннотированных на нескольких уровнях. Для визуализации и запроса данных PAULA можно использовать базу данных ANNIS.

**PC-KIMMO** https://software.sil.org/pc-kimmo/ Программа предназначена для генерации (продуцирования) и / или распознавания (синтаксического анализа) слов с использованием двухуровневой модели структуры слова, в которой слово представлено как соответствие между его формой лексического уровня и формой поверхностного уровня. Не поддерживается.

**PFC Platform** (TD) https://www.projet-pfc.net/ Инструмент для обработки и анализа корпуса современной французской фонологии (где PFC расшифровывается как Phonologie du Français Contemporain). Он интегрирует Praat и работает на Windows, MAC OS X и Linux.

**Phon** (FTD) https://phonbank.talkbank.org/ Общая база данных для изучения фонологического развития. Эта новая база данных будет опираться на программу PHON, предназначенную для фонологического и фонетического анализа данных, транскрибируемых в формате CHILDES.

**Praat** (TD/U,W,M) https://www.fon.hum.uva.nl/praat/ Бесплатная компьютерная программа, с помощью которой можно анализировать, синтезировать речь и манипулировать речью, а также создавать высококачественные языковые образы для статей и диссертаций.

**PROMT** https://www.promt.ru/ Компания-разработчик систем МП.

**qTerm ™** : https://www.memoq.com/extensions#qterm Программное обеспечение для управления терминологией через Интернет, которое определяет и переводит важную терминологию, а также предоставляет подробное пояснение использования каждого термина, включая контекст, язык и историю использования.

**quickTerm** http://www.kaleidoscope.at/en/terminology/quickterm/ Система управления жизненным циклом терминологии на основе базы данных SDL. Она расширяет охват терминобазы для многих различных пользователей, делая ее более доступной. Кроме того, quickTerm позволяет команде терминологов разрабатывать сложные рабочие процессы терминологии в масштабах всей компании на основе многолетних данных SDL и эффективно управлять жизненным циклом терминологии.

**RSTTool** (TD) http://www.wagsoft.com/RSTTool/ Инструмент – графический интерфейс для разметки структуры текста.

**SABLE** (FP) https://www.w3.org/TR/speech-synthesis11/ Стандарт SABLE для аннотирования лингвистических свойств входных данных синтеза речи обязательно

имеет много общих характеристик с системами лингвистического аннотирования естественной речи.

**SACODEYL** https://www.um.es/sacodeyl/en/pages/what.htm Редактор транскрипции и инструмент для добавления аннотаций к языковым корпусам в проекте, направленном на разработку системы для содействия составлению и открытому распространению европейской подростковой речи в контексте языкового образования. Проект включает в себя сбор и распространение английских, французских, немецких, итальянских, литовских, румынских и испанских подростковых разговоров.

**SAMPA** (C) https://www.phon.ucl.ac.uk/home/sampa/index.html SAMPA (Speech Assessment Methods Phonetic Alphabet) Машиночитаемая транслитерация ASCII Международного фонетического алфавита (IPA). Первоначально разработанный фонетиками для кодирования шести европейских языков, в настоящее время он расширяется и охватывает еще много языков. SAMPROSA – это расширение для транскрибирования просодической информации, а XSAMPA – расширение, которое охватывает каждый символ на диаграмме IPA, в принципе позволяя транскрибировать все языки мира.

**SayMore** (TD) https://documentation.help/SayMore/documentation.pdf Программа предназначена для создания хорошо аннотированных корпусов текстовых ресурсов.

**SBCSAE Surfer** (TD) https://www.linguistics.ucsb.edu/research/santa-barbara-corpus Инструмент для анализа и редактирования корпуса разговорного американского английского. ПО позволяет загружать транскрипции из корпуса вместе с его аудиозаписью.

**SDL MultiTerm Desktop** https://www.rws.com/translation/software/multiterm/ Инструмент управления терминологией рабочего стола от SDL. Он может быть использован в качестве автономного настольного инструмента для управления всей корпоративной терминологией, или его мощность может быть увеличена в среде перевода за счет интеграции с SDL Trados Studio.

**SDL** (Simple DirectMedia Laye) https://www.libsdl.org/ Свободная кросс-платформенная мультимедийная библиотека, реализующая единый программный интерфейс к графической подсистеме, звуковым устройствам и средствам ввода для широкого спектра платформ.

**Semantic Turkey** http://semanticturkey.uniroma2.it/ Сервисная платформа RDF для управления и приобретения знаний, реализованная Исследовательской группой ART Research Group в Римском университете Тор Вергата.

**Serengeti Annotator** (T) http://anawiki.essex.ac.uk/serengeti/ Веб-клиент-сервер-приложение, используемое для аннотирования семантических отношений в текстовых документах.

**SGREP** (TDP/U,W) https://www.cs.helsinki.fi/u/jjaakkol/sgrep.html Инструмент для поиска и индексации текстовых, SGML -, XML-и HTML-файлов, а также фильтрации текстовых потоков с использованием структурных критериев. Модель данных SGREP основана на сегментах, которые являются непустыми подстроками текста. Сегменты обычно представляют собой вхождения постоянных строк, SGML-тегов или значимых текстовых элементов, которые распознаются с помощью некоторых разделительных строк или встроенного синтаксического анализатора SGML, XML и HTML. Сегменты могут быть произвольно длинными, произвольно перекрывающимися и произвольно вложенными.

**Shoebox** https://software.sil.org/shoebox Компьютерная программа SIL, которая помогает полевым лингвистам и антропологам интегрировать различные типы текстовых данных: лексические, культурные, грамматические и т.д.

**SignStream** (TDP/M) http://www.bu.edu/asllrp/SignStream/ Инструмент базы данных для транскрипции и анализа языковых данных на основе видео (в частности, данных на языке жестов). SignStream позволяет пользователю вводить данные в любое число определяемых пользователем полей, где каждый датум связан с начальным и конечным кадром видео. Хотя база данных SignStream хранится в нечитаемом двоичном формате, программа включает функцию экспорта текста. Однако функция импорта отсутствует. В настоящее время эта программа распространяется по стоимости среди исследователей, преподавателей и студентов.

**SIL** (TDPF/W,M) https://www.sil.org/resources/archives/5744 Продукты консорциума SIL : LinguaLinks, Shoebox, WeSay, а также анализатор речи и менеджер речи, программы Windows для маркировки речевых файлов и поиска в базе данных помеченных речевых файлов. SIL также имеет формат аннотаций на основе SGML с именем PTEXT («синтаксический анализ текста»).

**SLAM** (TDP/W) https://www.researchgate.net/publication/221486520_SLAM_ segmentation_ and_labelling_automatic_module Автоматический модуль сегментации и маркировки (SLAM) – это инструмент для полуавтоматической сегментации и маркировки речевых сигналов. Инструмент был разработан в Институте фонетики и диалектологии.

**SMARTCAT** https://ru.smartcat.com/ Система автоматизированного перевода, включающая память переводов, машинный перевод, управление глоссариями, функцию совместной работы переводчиков над одним документом. Предназначена для компаний (в том числе переводческих) и отдельных переводчиков и их клиентов.

**SNACK** (TDP/U,W,M) http://www.speech.kth.se/snack/ Инструментарий для обработки акустических данных с акцентом на речь. Он имеет визуализацию в реальном времени, поддерживает множество форматов файлов и является расширяемым. В комплект поставки входят средство просмотра речевых сигналов и редактор фонетических меток. Существует также Snack «плагин» для веб-браузеров. Wavesurfer – это инструмент, основанный на Snack.

**SpeechIndexer** (TD) http://www.iiis.org/CDs2011/CD2011IDI/ICEIC_2011/PapersPdf/ EI931 JR.pdf Программное обеспечение, которое используется для сегментации и транскрибирования записанной речи. В частности, оно позволяет соотносить фрагменты аудио с соответствующей текстовой транскрипцией. Процесс как транскрибирования, так и индексации поддерживается эффективным устройством поиска пауз, которое обнаруживает интонационные единицы. Первоначальная цель программного обеспечения заключалась в документировании исчезающих языков аборигенов Формозы. Эти языки имеют чистую устную традицию, и только недавно появились архивы записей речи на этих языках. SpeechIndexer можно использовать также для обучения и изучения языков.

**Sphinx –** https://ru.wikipedia.org/wiki/Sphinx. (англ. SQL Phrase Index) Система полнотекстового поиска, распространяемая по лицензии GNU GPL либо для версий 3.0+ без исходных кодов. Отличительной особенностью является высокая скорость индексации и поиска, а также интеграция с существующими СУБД (MySQL, PostgreSQL) и API для распространенных языков веб-программирования (официаль-

но поддерживаются PHP, Python, Java; существуют реализованные сообществом API для Perl, Ruby,. NET[1] и C++).

**SpLaSH** (TD) https://www.aclweb.org/anthology/L10-1335/ Инструментарий, используемый для выполнения сложных запросов к корпусам разговорного языка. В SpLaSH предусмотрены инструменты для интеграции аннотаций, выровненных по времени, с помощью графиков аннотаций с выровненными текстами, с помощью универсальных XML-файлов. SpLaSH накладывает очень небольшое число ограничений на дизайн модели данных, позволяя интегрировать аннотации, разработанные отдельно в пределах одного набора данных и без какой-либо относительной зависимости. Он также предоставляет графический интерфейс, позволяющий выполнять разные типы запросов.

**SPPAS** (TD) http://www.sppas.org/ Инструмент для автоматического создания фонетических аннотаций из записанного речевого звука и его транскрипции. Вся процедура представляет собой последовательность автоматических шагов. Результатом является набор файлов TextGrid. SPPAS – это программное обеспечение с открытым исходным кодом, выпущенное по публичной лицензии GNU. SPPAS в настоящее время разработан для французского, английского, итальянского и китайского языков, и есть способ добавить другие языки. Операционные системы: Linux, MacOS и Windows.

**SSI** (Social Signal Interpretation) (TD) https://www.informatik.uni-augsburg.de/ lehrstuehle /hcm/projects/tools/ssi/ Система интерпретации социального сигнала. SSI предлагает инструменты для записи, анализа и распознавания человеческого поведения в режиме реального времени, такие как жесты, мимика, кивки головы и эмоциональная речь. SSI поддерживает потоковую передачу с нескольких датчиков и включает в себя механизмы их синхронизации. В частности, SSI поддерживает конвейер машинного обучения во всей его полноте и предлагает графический интерфейс, который помогает пользователю собирать собственные учебные корпуса и получать персонализированные модели. SSI также подходит для слияния мультимодальной информации на различных стадиях, включая раннее и позднее слияние.

**STAR Transit** http://cyclowiki.org/wiki/STAR_Transit_NXT Программа, предназначенная для создания, просмотра и редактирования проектов для проведения переводов и других связанных с переводом и локализацией операций.

**SUSANNE** (CP) https://www.grsampson.net/RSue.html Схема аннотаций обеспечивает детальное кодирование логической и поверхностной грамматики английского языка. SUSANNE corpus / treebank, находящийся в свободном доступе, содержит подмножество коричневого корпуса, которое было отмечено в соответствии с этой схемой. Проект CHRISTINE нацелен на расширение аналитической схемы и корпуса SUSANNE для охвата разговорного английского языка, и в частности спонтанного, неофициального разговорного английского языка.

**SyncWriter** (T) https://www.sign-lang.uni-hamburg.de/software/syncwriter/info. english.html Коммерческое программное обеспечение для Mac, которое позволяет синхронизировать различные «треки» – например, несколько треков транскрипции, видеотрек и трек комментариев. Авторы называют коллекцию этих треков «партитурой», а формат напоминает музыкальную партитуру. Видео хранится в треке в виде серии кадров с отметкой времени. Пользователь, заинтересованный в определенной части видео, может поместить нужные кадры на партитуру и использовать другие

треки для их описания. Хотя воспроизведение видео разрешено, оно не синхронизируется с другими дорожками во время воспроизведения. Программное обеспечение больше не распространяется поставщиком, копия по-прежнему доступна в Интернете через сайт EXMARaLDA. Фильтр импорта для преобразования данных syncWriter в более устойчивый формат XML интегрирован в EXMARaLDA.

**Systemic Coder** http://www.wagsoft.com/Coder/index.html Инструмент, который облегчает лингвистическое кодирование материала корпуса посредством подсказки пользователю соответствующих категорий. Лингвистические функции организованы в виде системной сети – иерархии наследования – для уменьшения объема кодирования. Сначала определяется иерархия функций, а затем кодируются сегменты текста в соответствии с иерархией. Затем эти кодировки можно подвергнуть статистическому анализу.

**Taas** https://term-extraction.tilde.com/ Облачные сервисы для работы с терминологией: предоставляет многоязычные и совместные терминологические услуги.

**TASX annotator** (TD/U,W,M) http://annotation.exmaralda.org/index.php?title=TASX Удобная программа для многоуровневой аннотации и транскрипции (многоканальной) видео- и аудиоданных. TASX-аннотатор работает под управлением WinXX (98, XP, 2000), Linux, Solaris и MacOS и распространяется под лицензией с открытым исходным кодом.

**Termbases**: https://www.termbases.eu/ Мощное веб-программное обеспечение для создания многоязычных терминологических ресурсов и управления ими.

**TermWeb:** https://termweb.store/ Система управления терминологией TermWeb обеспечивает согласованность языкового перевода от страны к стране и по всему миру. Предлагается линейка продуктов.

**TermWikiPro**: https://pro.termwiki.com/ Защищенная облачная система управления терминологией, разработанная для того чтобы помочь глобальным предприятиям ускорить бизнес, TermWiki Pro предоставляет полные готовые решения для улучшения качества контента при одновременном снижении затрат на создание и перевод.

**TFA** http://martinweisser.org/ling_soft.html#TFA Инструмент для исследования текстовых функций, который может помочь в выявлении и измерении проблем, связанных со сложностью текста.

**TippyTerm** https://www.tippyterm.de/en/ Инструмент, разработанный для всех систем MS Windows и обеспечивающий последовательное использование терминологии, легкую доступность, простоту в обращении, простое обслуживание.

**Tipster** (F) https://www.aclweb.org/anthology/X98-1001.pdf Инструмент для формирования аннотации текста.

**ToBI (**Tones and Break Indices**)** http://liceu.uab.cat/publicacions/MATE_D1_1_6_Prosody/annex2/Tobi.html Система для расшифровки интонационных паттернов и других аспектов просодии английских высказываний.

**Toolbox** (TD) https://software.sil.org/toolbox/ Toolbox Инструмент управления и анализа данными для полевых лингвистов. Он особенно полезен для хранения лексических данных, а также для синтаксического анализа и интерлинеаризации текста, но его можно использовать для управления практически любыми данными.

**Transana** (T) (Review) https://www.transana.com/ Это программное обеспечение облегчает транскрипцию и анализ видеоданных. Позволяет пользователям просмат-

ривать видео, создавать транскрипт и связывать места в транскрипте с соответствующими кадрами видео. Аналитически интересные видео и части видео могут быть идентифицированы, организованы и легко доступны с помощью предоставляемых инструментов. Организационная структура является центральной для Transana. В этой системе видеофайлы называются эпизодами. Связанные эпизоды можно сгруппировать в серию. Каждый эпизод также может быть разбит на сегменты, называемые клипами. Связанные клипы можно сгруппировать вместе в так называемую коллекцию. Ключевые слова могут быть назначены как клипам, так и эпизодам для описания контента. Затем связанные ключевые слова можно сгруппировать вместе, а также выполнить поиск. Окно базы данных, расположенное в виде древовидной структуры, предназначено для управления и организации этих группировок. Эта организационная система позволяет легко хранить большие коллекции видеофайлов.

**Transcriber** (TDP/U,W,M) http://trans.sourceforge.net/en/presentation.php Бесплатное программное обеспечение для транскрибирования и аннотирования цифрового аудио, изначально предназначенное для транскрипции данных широковещательных новостей. Его пользовательский интерфейс написан на языке Tcl/Tk. Он использует те же форматы транскрипции, что и широковещательные новостные данные HPC, а также был адаптирован для ввода–вывода XML. С июля 2011 г. доступен преемник Transcriber под названием TranscriberAG.

**TranscriberAG** (TD) http://transag.sourceforge.net/ Официальный преемник Transcriber. TranscriberAG предназначен для оказания помощи в ручном аннотировании речевых сигналов. Он предоставляет удобный графический интерфейс пользователя для сегментации длительных речевых записей, их транскрибирования, маркировки речевых оборотов, изменений темы и акустических условий. TranscriberAG ориентирован на потребности сообщества исследователей речи, но его функции могут быть полезны и для других приложений. Он использует формат «графа аннотаций» в качестве собственного формата, но может читать ряд других форматов аннотаций.

**Transformer** (TDP) http://www.romanistik.uni-freiburg.de/ehmer/transformer/ Программный инструмент для ученых, работающих с транскрибированными лингвистическими данными. Она адресована разговорным аналитикам, фонетикам, антропологам и другим социологам, которые хотят анализировать цифровые аудио- или видеоданные и язык. Transformer – это программа для быстрого, безопасного и простого управления и преобразования транскрибированных лингвистических и выровненных данных.

**TransTool** (TD/U,W) http://annotation.exmaralda.org/index.php?title=TransTool Корпус шведского разговорного языка, разработанный на кафедре лингвистики Гетеборгского университета, имеет несколько интересных инструментов: Transtool, чтобы помочь в транскрипции; Synchtool для синхронизации транскрипций с аудио- и видеофайлами; TRASA, инструмент для автоматического анализа корпуса; и TRACTOR, инструмент для поддержки кодирования.

**trAVis** (T) http://www.travis-analysis.org/no_firefox_error.html – Веб-инструмент для ориентированной на музыку транскрипции аудиовизуальных носителей.

**Treebank** (C) https://en.wikipedia.org/wiki/Treebank Проект Penn Treebank выпустил семантические и синтаксические аннотации естественно возникающего текста для The Wall Street Journal, Brown, ATIS и Switchboard Corpora. Аннотации, подготовленные в рамках проекта Treebank, были опубликованы [#LDC LDC]. Treebank имеет два языка запросов: tgrep (в LDC-Online) и CorpusSearch. Разрабатываются

банки деревьев для других языков, в том числе: немецкий, турецкий, польский, чешский, португальский, болгарский, китайский.

**TSNLP** (FT) https://www1.essex.ac.uk/linguistics/external/clmt/group/projects/tsnlp/ Европейский консорциум, предоставляющий технологию NLP test suite и фрагменты test suite для немецкого, французского и английского языков. Схема аннотаций включает в себя сентенциальную информацию (такую как оценка правильности формулировки), а также аннотацию слов и подстрок, которая хранится в табличной форме (для аналитической информации, включая синтаксические составляющие и описания ошибок). Есть веб-интерфейс в тестовые наборы. Схема базы данных описана в руководстве пользователя.

**UAM Corpus Tool** (T) http://corpustool.com/ Среда для аннотирования текстовых корпусов с использованием автономной разметки XML. Особенности: (1) Аннотирование нескольких текстов с использованием одинаковых схем аннотации вашего дизайна; (2) Аннотации к каждому тексту на нескольких уровнях; (3) Поиск экземпляров на разных уровнях, например, конечное предложение, содержащее company-np, или future-clause во введении; (4) Сравнительная статистика по подмножествам.

**Unicode** (RC) https://home.unicode.org/ Консорциум Unicode объединяет корпорации по разработке программного обеспечения и исследователей, находящихся на переднем крае стандартизации международной кодировки символов. Результатом этого сотрудничества стал стандарт Unicode, который обеспечивает основу для интернационализации и локализации программного обеспечения. Unicode имеет серию конференций и часто задаваемые вопросы. Существуют диаграммы символов, в том числе одна для расширений.

**Vakyartha** (T) http://annotation.exmaralda.org/index.php?title=Vakyartha Веб-среда для совместного аннотирования корпуса зависимостей.

**Verbmobil** (FC) http://verbmobil.dfki.de/overview-us.html Крупный немецкий проект перевода речи в текст для таких областей, как переговоры о встрече, планирование поездок и бронирование гостиниц. Проект аннотации Verbmobil включает в себя орфографию, сегментарную аннотацию (с BAS Partitur), просодию (немецкий ToBI), морфологическую и POS-маркировку, семантическую аннотацию и аннотацию диалогового акта. Есть модель совместного использования лексических баз данных в Verbmobil, получившая название HyprLex.

**VideoAnnex** (T) https://www.researchgate.net/publication/228779422_Video_Collaborative_Annotation_Forum_Establishing_Ground-Truth_Labels_on_Large_Multimedia_Datasets Инструмент аннотирования, разработанный IBM, помогает авторам в задаче аннотирования видеопоследовательностей с помощью метаданных MPEG-7. Каждый снимок в видеоряде может быть аннотирован статическими описаниями сцен, описаниями ключевых объектов, описаниями событий и другими наборами лексики. Аннотированные описания связаны с каждым снимком видео и хранятся в виде mpeg-7 описаний в выходном XML-файле. VideoAnnEx также может открывать файлы MPEG-7 для отображения аннотаций для соответствующего видеоряда. Инструмент аннотации также позволяет создавать, сохранять, загружать и обновлять настроенные лексиконы.

**VideoGraph** (T) http://www.ipn.uni-kiel.de/aktuell/videograph/htmStart. htm Мультимедийный инструмент для Windows XP/VISTA/7. Он позволяет пользователю воспроизводить и анализировать видеоданные.

**ViPER** (T) http://viper-toolkit.sourceforge.net/ Инструментарий скриптов и Java-программ, которые позволяют разметку визуальных данных, а также систем для оценки качества результирующих данных.

**VisLab** (TDP) https://books.google.ru/books?id=Cp0pAQAAMAAJ&q=VISLab Кросс-модальный анализ сигнала и смысла. Цель проекта состоит в том, чтобы создать большую базу данных видео, аннотированную информацией о жестах, речи и взгляде. Это будет использовано для эмпирической проверки теорий мультимодальной коммуникации. Имеются описания процессов, а также загружаемые инструменты и данные.

**VOCALE** https://www.ime.usp.br/~tycho/prosody/vocale/project.html Инструмент для автоматического аннотирования вокальных и согласных интервалов на основе вероятностного измерения относительной энтропии и ряда фонетических измерений. Vocale принимает файл wav в качестве входных данных, затем автоматически вызывает некоторые функции Praat, такие как создание спектрограммы, и выдает файл меток Praat в качестве вывода.

**VocBench** http://vocbench.uniroma2.it/ Многоязычная веб-платформа совместной разработки для управления онтологиями, тезаурусами, лексиконами и данными RDF.

**VoiceScribe** (TD) https://www.univie.ac.at/voice/ Небольшой и простой в использовании редактор подсветки с прикрепленным аудиоплеером. Он был разработан для Венско-Оксфордского Международного корпуса английского языка.

**vPrism** (T/W) http://annotation.exmaralda.org/index.php?title=VPrism vPrism Коммерческое программное обеспечение macintosh для временной аннотации и кодирования видео, предназначенное для использования в образовательных и поведенческих исследованиях.

**WaveSurfer** http://www.speech.kth.se/wavesurfer/ Бесплатный инструмент с открытым исходным кодом для визуализации и обработки звука. Работает практически на всех платформах (Windows, Macintosh, Unix / Linux и др.).

**WebAnno** (T) https://webanno.github.io/webanno/ Универсальный веб-инструмент аннотаций для широкого спектра лингвистических аннотаций. WebAnno предлагает управление проектами аннотаций, свободно настраиваемые наборы тегов и управление пользователями в различных ролях. WebAnno использует технологию инструмента быстрого аннотирования для визуализации и редактирования аннотаций в веб-браузере. Он поддерживает аннотацию и визуализацию произвольно больших документов, подключаемые фильтры импорта / экспорта, настройку аннотаций для различных пользователей и выведение аннотаций на платформу краудсорсинга.

**WeSay** (TD/U,W) https://software.sil.org/wesay/ WeSay Помогает неязыковедам построить словарь на своем родном языке. Предлагаются различные способы, чтобы помочь носителям языка ввести некоторые основные данные о словах своего языка. Программа настраивается и ориентирована на задачи, предоставляя консультанту возможность включать / выключать задачи по мере необходимости и по мере того, как пользователь проходит обучение для выполнения этих задач. WeSay использует стандартный формат XML, поэтому данные могут быть использоваться в лингвистических инструментах, таких как FieldWorks. WeSay требует запуска. Net-Framework и доступен для Windows и Linux (с mono). WeSay – это совместное производство Payap Language SIL PNG и SIL International. Software с открытым исходным кодом,

**WinPitch** (TD) http://www.winpitch.com/ Инструмент анализа речи и аннотаций на базе Windows с отображением основной частоты и спектрографическим отображением. Возможность просодического морфинга посредством повторного синтеза естественной речи.

**WordFas**t https://www.wordfast.net/ Автономный инструмент памяти переводов (TM) нового поколения, предназначенный как для корпораций, так и для бюро переводов и переводчиков. Имеет ряд версий разной функциональности.

**WordSmith** (TD) https://lexically.net/wordsmith/ Коммерческий интегрированный набор программ для изучения того, как слова ведут себя в текстах. Инструмент *Список слов* позволяет просмотреть список всех слов или кластеров слов в тексте, расположенных в алфавитном или частотном порядке. Конкордансер Concord дает возможность увидеть любое слово или фразу в контексте. Можно найти ключевые слова в тексте. Эти инструменты были использованы издательством Оксфордского университета для собственной лексикографической работы по подготовке словарей преподавателями иностранных языков и студентами, а также исследователями, изучающими языковые модели на множестве различных языков.

**XTM** https://xtm.cloud/ Система управления переводами со встроенной памятью переводов, терминологией и CAT-инструментами.

**XTran**s (TDPF), https://www.ldc.upenn.edu/language-resources/tools/xtrans Разработка LDC представляет собой мультиплатформенный, многоязычный, многоканальный инструмент транскрипции «следующего поколения» для поддержки ручной транскрипции и аннотации аудиозаписей.

**YouTube Comments and Nicks Collecto**r (T/U,W,M) https://herbst.ist.org/YouTubeCommentsAndNicksCollector.html Инструмент для сбора комментариев и ников со страниц YouTube и сохранения их в простых, чистых и небольших текстовых файлах для дальнейшей обработки.

## ПРИЛОЖЕНИЕ 8. МИРОВЫЕ ТЕРМИНОЛОГИЧЕСКИЕ БАНКИ ДАННЫХ

1. ***BabelNet*** https://babelnet.org Многоязычный энциклопедический словарь с лексикографическим и энциклопедическим охватом терминов на 50 языках и онтологией, которая объединяет понятия и именованные объекты в очень большую сеть семантических отношений, состоящую из более чем 9 млн записей.

2. ***EuroTermBank*** http://www.eurotermbank.com/ Многоязычный банк терминов, созданный для гармонизации терминоведческой деятельности в странах Европейского союза. В настоящее время EuroTermBank включает термины на 33 европейских языках. В основе данного лексикографического ресурса находятся электронные документы, а также оцифрованные версии печатных словарей. Особое внимание при создании данной терминологической базы было уделено языкам с небольшим или недостаточным количеством ресурсов. К наиболее хорошо обеспеченным в плане терминологии языкам в EuroTermBank относятся английский, русский, немецкий, латышский и польский. EuroTermBank позволяет обмениваться данными с существующими национальными и международными базами данных, устанавливая отношения сотрудничества, согласовывая методологии и стандарты, разрабатывая и внедряя механизмы и процедуры обмена данными. Объем – 15 млн терминов.

3. ***EuroVoc*** https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer Многоязычный междисциплинарный тезаурус, охватывающий информацию ЕС, в частности Европейского парламента. Содержит термины на 22 языках.

4. ***IATE*** Интерактивная терминология для Европы https://iate.europa.eu/about Межинституциональная терминологическая база данных Европейского Союза. IATE объединила все существующие терминологические базы данных переводческих служб ЕС в одну межведомственную базу данных, содержащую примерно 1,4 млн многоязычных записей и 9 млн терминов на 24 официальных языках ЕС. Текущая новая версия IATE была выпущена в ноябре 2018 г.

5. ***IMF Terminology, a multilingual directory*** (терминологическая база данных Международного валютного фонда) http://www.imf.org/external/np/term/eng/index.htm IMF terminology Электронный ресурс, разработанный в рамках деятельности Международного валютного фонда. На данный момент содержит более 150 000 терминов и терминологических словосочетаний по банковскому делу и финансам. Данная терминологическая база работает с 9 языками – английским, французским, русским, испанским, арабским, китайским, португальским, немецким и японским.

6. ***ISO Concept Database*** Терминологическая база данных ISO http://cdb.iso.org/ Многоязычная концептуально-терминологическая база данных, была создана Международной организацией по стандартизации ISO для поиска, разработки и поддержки терминологии во всех стандартах ISO. Данная ТБД позволяет осуществлять поиск на английском, французском, русском, испанском и немецком языках по таким категориям, как термины и их определения, графические символы и коды (языков, стран, валют и т.д.).

7. ***ITU Terms and Definitions*** (терминология Международного союза электросвязи (МСЭ)) http://www.itu.int/net/ITUR/index. ITU Terms and Definitions представляет собой электронный ресурс, который содержит аббревиатуры, термины и их определения в сфере коммуникации. Данная система позволяет осуществлять поиск на шести языках (английском, арабском, китайском, испанском, русском и французском) и предоставляет ссылки на источники терминов – публикации ITU. В настоящее время в ITU Terms and Definitions содержится около 130 000 терминов и аббревиатур.

8. ***LIND-We*b** Веб-платформа языковой индустрии https://ec.europa.eu/info/departments/translation/language-industry-platform-lind Содержит факты и цифры по языковой индустрии ЕС. Они предоставляются языковыми специалистами, заинтересованными сторонами отрасли и институтами ЕС.

9. ***METEOTERM*** https://public.wmo.int/en/resources/language%20resources/meteoterm Терминологическая база данных Всемирной метеорологической организации, содержит специальные термины метеорологии и климатологии на шести языках (английском, французском, русском, китайском, арабском и испанском). Источниками данного ресурса являются Международный метеорологический словарь, Международный словарь по гидрологии, а также термины, которые встречаются в официальных документах и отчетах Всемирной метеорологической организации.

10. ***Microsoft Language Portal*** (языковой портал Microsoft) https://www.microsoft.com/en-us/language Электронный ресурс для поиска терминов компании Miscrosoft и общей терминологии в сфере информационных технологий. Данный портал содержит порядка 25 000 терминов на 100 языках и их определения на английском языке. Microsoft Language Portal используется для унификации и стандартизации терминов, связанных с разработкой и интеграцией приложений в среде Miscrosoft, а также может служить основой для создания многоязычных глоссариев в сфере IT.

*11.* ***MINÉFITERM*** https://www.minefiterm.finances.gouv.fr/termino.php Терминологическая база данных Центра переводов при Министерстве экономики и финансов и Министерстве бюджета, государственных счетов и гражданской администрации Франции. В настоящее время содержит порядка 60 000 терминов на французском, английском, испанском, немецком, итальянском, португальском, русском, голландском, китайском, японском и арабском языках. Данная система является открытой и позволяет добавлять и редактировать термины и их определения.

12. ***Multilingual Archival Terminology*** (многоязычная терминология архивных документов) http://www.ciscra.org/mat/ Электронный ресурс многоязычных терминов и их определений в области архивного дела. Основная задача этой терминологической базы данных – максимально широко представить многообразие терминов архивного дела и документоведения и отразить специфику данной отрасли в разных странах. В настоящее время Multilingual Archival Terminology позволяет осуществлять поиск на 23 языках, в том числе на английском, французском, русском, китай-

ском, финском, арабском, немецком и греческом. Данный ресурс отражает актуальную информацию о терминах архивного дела, благодаря возможности добавлять и редактировать термины и их определения.

13. ***TERMIUM*** *Plus* https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng Один из крупнейших в мире банков терминологических и лингвистических данных, предоставляет доступ к миллионам терминов на английском, французском, испанском и португальском языках.

14. ***TermNet,*** https://www.termnet.org/. Международная сеть терминологии является международным форумом сотрудничества для компаний, университетов, учреждений и ассоциаций, которые занимаются дальнейшим развитием глобального рынка терминологии.

15. ***TermScience*** http://www.termsciences.fr/termsciences/?lang=en. Термины науки: это портал терминологии, разработанный INIST совместно с LORIA и ATILF. Его цель – продвигать, объединять и делиться терминологическими ресурсами (словарями специалистов, тезаурусами) государственного сектора и образовательных учреждений для создания общего терминологического справочного ресурса.

16. ***TermWiki.com*** http://ru.termwiki.com/TermWiki_Home Онлайн-портал терминологии, который позволяет пользователям искать, загружать, переводить и делиться терминами и определениями с другими пользователями (более 100 языков).

17. ***UNESCOTERM*** http://uis.unesco.org/en/glossary. База данных терминологии ЮНЕСКО на английском, арабском, испанском, китайском, русском и французском языках.

18. ***UNTERM*** https://unterm.un.org/UNTERM/portal/welcome Терминологическая база данных ООН. Содержит техническую и специализированную терминологию на каждом из шести официальных языков ООН (английский, французский, испанский, русский, мандаринский и арабский).

19. ***WIPO Pearl*** http://www.wipo.int/reference/en/wipopearl Многоязычный терминологический портал Всемирной организации интеллектуальной собственности, который включает специальные термины патентов на 10 языках (арабском, китайском, английском, французском, немецком, японском, корейском, португальском, русском и испанском). Данный ресурс позволяет осуществлять поиск в двух режимах – поиск по термину и поиск по понятию (concept map search, концептуальный поиск). Концептуальный поиск представляет собой поиск по графу, который отображает семантические отношения между терминами. На данный момент WIPO Pearl содержит около 16 000 понятий и 105 000 терминов.

20. ***Банк данных РОСТЕРМ*** http://www.gostinfo.ru/catalog/terminlist Единственный российский терминологический банк, представленный в международных каталогах. Содержит свыше 140 тысяч терминологических статей из ГОСТ, ГОСТ Р, стандартов ИСО и МЭК.

21. ***ВТОТЕРМ*** https://wto.sdlproducts.com/multiterm Терминологическая база данных, созданная Всемирной торговой организацией (ВТО). Поиск терминов и эквивалентов может быть выполнен на французском, английском и испанском языках.

22. ***Портал лексикографических ресурсов НАТО*** https://www.lexicool.com/online-dictionary.asp?FSP=C04&FKW=nato Содержит 36 ЛИР.

23. ***Терминологический портал ФАО*** http://www.fao.org/faoterm/en/ Продовольственная и сельскохозяйственная организация ООН. Содержит более 80 000 за-

писей на шести официальных языках Организации Объединенных Наций (английском, французском, испанском, русском, арабском и мандаринском).

24. ***Терминология Международной организации труда (МОТ)*** https://www.ilo.org/inform/online-information-resources/databases/terminology/lang--en/index.htm Включает терминологическую базу (ILOterm), предметную классификацию и два тезауруса.

25. ***Терминология ОЭСР*** https://www.oecd-ilibrary.org/glossaries Содержит 15 ЛИР.

26. ***Терминология Тильды*** https://term.tilde.com/. Услуги: извлечение терминов и поиск в облаке – около 5 млн стандартизированных и надежных терминов; облачные средства для управления терминологией и совместного использования; встроенное распознавание и поиск терминологии в инструментах перевода.

27. ***Центр по координации терминологии Евросоюза*** (TermCoord). https://termcoord.eu/terminology-websites/ Основная цель TermCoord заключается в оказании помощи переводчикам в выполнении их повседневных задач и содействии исследованию терминологии и управлению терминологией в подразделениях перевода, а также в увеличении вклада в терминологическую базу данных ЕС IATE. Поддерживаются каталоги терминологических ресурсов и организаций.

## ПРИЛОЖЕНИЕ 9. ЗАРУБЕЖНЫЕ ЦЕНТРЫ И РЕСУРСЫ ПО РУСИСТИКЕ

Дата обращения для всех ссылок 01.04.2022

***США и Канада***

1. American Association for the Advancement of Slavic Studies http://worldcat.org/identities/lccn-n79139496/ h
2. American Association of Teachers of Slavic and East European Languages (AATSEEL) http://aatseel.org/
3. American Council of Teachers of Russian/American Council for Collaboration in Education and Language Study (ACTR/ACCELS) https://www.actr.org/
4. American University Russian Studies Русский язык, мировые языки и культуры https://www.american.edu/cas/wlc/languages/russian.cfm
5. Amherst College, Massachusetts Russian Department http://www.amherst.edu/~russian/menu.html
6. Appalachian State University Department of Foreign Languages and Literatures http://www.fll.appstate.edu/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=10&MMN_position=70:70
7. Arizona State University Slavic Languages Section http://www.public.asu.edu/~iclbc/
8. Bates College Department of German, Russian, and East Asian Languages and Literatures http://abacus.bates.edu/pubs/Dept.Letters/greall.html
9. Baylor University Russian http://www.baylor.edu/Russian/
10. Boston College Department of Slavic and Eastern Languages https://www.bc.edu/bc-web/schools/mcas/departments/eastern-slavic-german.html
11. Boston University Institute for the Study of Conflict, Ideology & Policy (ISCIP) (Studying the political, international, and security affairs of Russia and the NIS) https://open.bu.edu/handle/2144/1353
12. Boston University The School of Russian and Asian Studies http://www.alinga.com/
13. Bowdoin College Russian Department https://www.bowdoin.edu/russian/index.html
14. Bowling Green SU Russian Program https://www.bgsu.edu/arts-and-sciences/world-languages-and-cultures/languages/russian.html
15. Brigham Young University Russian Program https://russian.byu.edu/
16. Brown University Department of Slavic Languages http://www.brown.edu/Departments/Slavic_Languages/

17. Bryn Mawr College Russian Department http://www.brynmawr.edu/russian/
18. Bucknell University Russian Studies Department http://www.bucknell.edu/Russian/
19. Carleton College Russian Department http://apps.carleton.edu/curricular/russ/
20. Carleton University The School of Linguistics and Applied Language Studies https://carleton.ca/slals/
21. Case Western Reserve University https://mll.case.edu/undergraduate/russian/
22. Center for Russian and East European Studies https://crees.ku.edu/
23. Colgate University Russian Studies https://www.colgate.edu/academics/departments-programs/russian-and-eurasian-studies-program
24. College of William and Mary Russian Program http://www.wm.edu/modlang/russian/index.php
25. Colorado College Russian and Eurasian Studies https://www.coloradocollege.edu/academics/dept/russianeurasianstudies/
26. Colorado State University Chinese, Japanese, and Russian http://www.colostate.edu/Depts/FLL/ch_ja.html
27. Columbia University Department of Slavic Languages http://www.columbia.edu/cu/slavic/
28. Concordia Language Villages Lesnoe Ozero http://www.concordialanguagevillages.org/youth-languages/russian-language-village
29. Connecticut College Slavic Studies https://www.conncoll.edu/academics/majors-departments-programs/departments/slavic-studies/
30. Cornell University Department of Russian Literature https://complit.cornell.edu/russian-minor
31. CUNY Lehman https://www.lehman.edu/academics/arts-humanities/languages-literatures/russian.php
32. Dalhousie University Russian Studies http://www.registrar.dal.ca/calendar/ug/RUSN.htm
33. Dartmouth College http://www.dartmouth.edu/~russian/
34. Davidson College https://www.davidson.edu/academic-departments/russian-studies
35. Depaul University Department of Modern Languages. Russian https://las.depaul.edu/academics/modern-languages/faculty/russian/Pages/default.aspx
Dickinson College Russian Language and Literature https://www.dickinson.edu/homepage/131/russian
36. Drake University Modern Languages and Literatures https://www.drake.edu/wlc/
37. Drew University Russian Department http://catalog.drew.edu/preview_program.php?catoid=29&poid=1643
38. Duke University https://slaviceurasian.duke.edu/
39. Eckerd College Russian Studies https://www.eckerd.edu/internationalstudies/modern-languages/
40. Emory College. Russian and East Asian Languages and Cultures http://www.realc.emory.edu/

41. Emory University Center for Russian & East European Studies https://rees.emory.edu/#:~:text=The%20Russian%2C%20East%20European%2C%20and,institutions%2C%20and%20the%20general%20public.

42. Ferrum College http://ferrum.catalog.acalog.com/preview_program.php?catoid=9&poid=1082&returnto=529

43. Fordham University Russian Language Program https://bulletin.fordham.edu/courses/russ/

44. George Mason University https://russianstudies.gmu.edu/

45. George Washington University http://www.gwu.edu/~slavic/

46. Georgetown University Center for Eurasian, Russian and East European Studies https://ceres.georgetown.edu/

47. Grinnell College Russian Department http://web.grinnell.edu/russian/

48. Gustavus-Adolphus College Russian Language and Area Studies https://gustavus.edu/general_catalog/12_13/russian

49. Hamilton College Russian Studies http://academics.hamilton.edu/russian/home/default.html

50. Harvard University Department of Slavic Languages and Literatures https://slavic.fas.harvard.edu

51. https://cultures.rice.edu/emeriti-faculty/ewa-m-thompson

52. https://mlli.umbc.edu/

53. https://www.stonybrook.edu/commcms/eurolangs/people/

54. Hunter College The Division of Russian and Slavic Languages http://www.hunter.cuny.edu/classics/russian/

55. Illinois Wesleyan University http://titan.iwu.edu/~mcll/courses/russian.html

56. Indiana University Department of Slavic Languages and Literatures http://www.indiana.edu/~iuslavic/

57. Iowa Центр изучения языка и культуры https://lmc.uiowa.edu/

58. James Madison University https://www.jmu.edu/forlang/people/FacultyRus.shtml

59. Marist College https://www.marist.edu/study-abroad/programs/russia-moscow

60. McGill University Department of Russian and Slavic Studies http://www.mcgill.ca/arts-internships/departments/russian-slavicstudies/

61. Michigan State University Russian and Other Slavic Languages http://www.msu.edu/~linglang/russian/index.htm

62. Middlebury College (Vermont) Russian Department http://www.middlebury.edu/academics/ls/russian/

63. Mount Holyoke College The Department of Russian and Eurasian Studies / https://www.mtholyoke.edu/acad/russian

64. Nazareth College https://www2.naz.edu/academics/foreign-languages-degree-program

65. New York University Department of Russian and Slavic Studies http://www.nyu.edu/cas/dept/slav.htm

66. Northwestern University Department of Slavic Languages and Literatures http://www.wcas.northwestern.edu/slavic/

67. Ohio State University Department of Slavic and East European Languages and Literatures http://slavic.osu.edu/

68. Old Dominion University Department of Foreign Languages and Literatures http://www.odu.edu/al/lang/otherlang.htm#russian ·

69. Penn State University Slavic and East European Languages and Literatures https://german.la.psu.edu/

70. Princeton University Department of Slavic Languages and Literatures http://www.princeton.edu/~slavic/

71. Purdue University Russian Program https://cla.purdue.edu/academic/slc/l/Russian/ /

72. Queens College Russian Program https://www.qc.cuny.edu/academics/degrees/dah/ell/russian/Pages/default.aspx

73. Reed College Russian Department http://academic.reed.edu/russian/

74. REESWeb: Russian and East European Studies Internet Resources http://www.ucis.pitt.edu/reesweb/

75. Rice University. Department of Modern and Classical Literatures and Cultures

76. Russian and East European Institute http://www.indiana.edu/~reeiweb/

77. Russian and East European Studies Consortium http://www.asu.edu/ipo/reesc/

78. Russian, East European, and Eurasian Center at Illinois http://www.reec.uiuc.edu/

79. Rutgers, The State University of New Jersey Program in Slavic and East European Languages and Literatures http://seell.rutgers.edu

80. San Diego State University Department of German and Russian Languages and Literatures https://esdepartment.sdsu.edu/academic-programs/russian

81. Seton Hall University (South Orange, New Jersey) Department of Modern Languages. Russian Language and Literature http://artsci.shu.edu/russian/

82. Southern Methodist University (Dallas, Texas) Foreign Languages Learning Center https://www.smu.edu/Dedman/Academics/Departments/world-languages#:~:text=The%20Department%20of%20World%20Languages,internships%20and%20study%20abroad%20opportunities

83. St. Lawrence University College Russian https://www.lawrence.edu/academics/college/russian

84. Stanford University Center for Russian and East European Studies https://bulletin.stanford.edu/departments/REES/overview

85. Stanford University. Division of literatures, cultures, and languages https://dlcl.stanford.edu/ Stetson University Russian Studies Program https://www.stetson.edu/other/academics/undergraduate/russian-east-european-and-eurasion-studies.php

86. SUNY Albany Slavic and Eurasian Studies https://www.albany.edu/undergraduate_bulletin/department_languages_slavic.html

87. Swarthmore College, Pennsylvania Russian Language, Literature and Culture http://www.swarthmore.edu/Humanities/mll/russian/index.html

88. Syracuse University Russian Language, Literature and Culture https://thecollege.syr.edu/languages-literatures-and-linguistics/russian-language-literature-and-culture

89. The Association for Women in Slavic Studies https://awsshome.org/ http://www.cofc.edu/languages/russian/ https://www.cofc.edu/academics/majorsandminors/russian-studies.php

90. The College of Wooster Russian Studies Department http://www.wooster.edu/russia/

91. The Russian Circle of Grand Valley State University http://www2.gvsu.edu/~russianc/

92. The Slavic Department at the University of Chicago https://slavic.uchicago.edu/

93. The University of Kansas Libraries Slavic Dept. https://slavic.ku.edu/

94. The University of Toledo. The Department of Foreign Languages. Russian Courses https://www.utoledo.edu/al/world-languages-and-cultures/

95. The University of Victoria (Canada) The Department of Germanic and Russian Studies https://www.uvic.ca/humanities/germanicslavic/index.php

96. UC Berkeley Slavic Languages and Literatures Department https://slavic.berkeley.edu/

97. UC Davis Department of German and Russian https://german.ucdavis.edu/people

98. UC, Riverside Comparative Literature and Foreign Languages. Russian Program https://complitlang.ucr.edu/

99. UC, Santa Barbara Department of Germanic, Slavic, and Semitic Studies https://www.gss.ucsb.edu/

100. UC, Santa Cruz Russian Language https://catalog.ucsc.edu/en/Current/General-Catalog/Courses/RUSS-Russian

101. UCLA Department of Slavic Languages and Literatures https://slavic.ucla.edu /

102. University of Alabama Dept. of Modern Languages & Classics -The Russian Program

103. University of Alaska, Anchorage Department of Languages. Russian Program https://www.uaa.alaska.edu/academics/college-of-arts-and-sciences/departments/languages/russian.cshtml

104. University of Alaska, Fairbanks Department of Foreign Languages and Literatures https://uaf.edu/language/index.php

105. University of Alberta in Edmonton Slavic Languages https://www.ualberta.ca/modern-languages-and-cultural-studies/undergraduate-program-information/current-undergraduate-students/slavics-courses/index.html

106. University of Arizona Department of Russian and Slavic Languages http://russian.arizona.edu/

107. University of Chicago Department of Slavic Languages and Literatures https://slavic.uchicago.edu/

108. University of Colorado Boulder The Department of Germanic and Slavic Languages and Literatures. Russian http://www.colorado.edu/germslav/russian/

109. University of Connecticut http://www.catalog.uconn.edu/russ.htm

110. University of Delaware The Russian Web Page https://www.dllc.udel.edu/undergrad-study/languages/russian1

111. University of Denver http://www.du.edu/langlit/russian/

112. University of Florida Department of German and Slavic Studies https://languages.ufl.edu/academics/llc-languages/german-studies/

113. University of Georgia Russian Program https://www.gsstudies.uga.edu/russian-flagship-program#:~:text=This%20federally%2Dsponsored%20initiative%20provides,and%20earn%20prestigious%20Flagship%20certification

114. University of Hawaii Russian Division http://www.hawaii.edu/llea/russian/

115. University of Illinois, Urbana-Champaign Department of Slavic Languages and Literature http://slavic.lang.uiuc.edu/

116. University of Iowa Russian Department https://asian-slavic.uiowa.edu/

117. University of Kansas Department of Slavic Languages and Literatures https://slavic.ku.edu/

118. University of KentuckyRussian Studies https://www.uky.edu/experts/sub-field/russian-studies

119. University of Maryland, Baltimore County Modern Languages and Linguistics https://mlli.umbc.edu

120. University of Michigan Department of Slavic Languages and Literatures http://www.lsa.umich.edu/slavic/

121. University of Minnesota Slavic and Central Asian Languages and Literatures http://www.iles.umn.edu/russian.htm

122. University of Mississippi Modern Languages, Linguistics & Intercultural Communication

123. University of Missouri Department of German and Russian Studies https://sllc.missouri.edu/ University of Nebraska-Lincoln Russian Program http://www.unl.edu/modlang/content/under/russian/index.htm

124. University of New Brunswick Department of Culture and Language Studies. Russian and Eurasian Studies http://www.unbf.ca/arts/Culture_Lang/programs_russian.html

125. University of New Mexico Russian Program https://fll.unm.edu/languages/russian.html

126. University of North Carolina at Chapel Hill The Department of Slavic Languages and Literatures https://gsll.unc.edu/

127. University of North Carolina at Greensboro Department of German and Russian, including Japanese Studies https://llc.uncg.edu/japanese/

128. University of Northern Iowa Russian Program https://catalog.uni.edu/collegeofhumanitiesartsandsciences/languagesandliteratures/

129. University of Oklahoma Department of Modern Languages, Literatures and Linguistics http://www.ou.edu/cas/modlang/

130. University of Oregon Department of Russian http://darkwing.uoregon.edu/~russian/

131. University of Pittsburgh Department of Slavic Languages and Literatures https://www.slavic.pitt.edu/

132. University of Rochester Modern Languages and Cultures: Undergrad Program: Russian https://www.sas.rochester.edu/mlc/undergraduate/russian.html

133. University of Southern California Department of Slavic Languages and Literatures https://dornsife.usc.edu/sll/

134. University of Tennessee, Knoxville Germanic, Slavic and Asian Languages https://mfll.utk.edu/friends.php

135. University of Texas Department of Slavic and Eurasian Studies http://www.utexas.edu/cola/depts/slavic/

136. University of Toronto Centre for Russian and East European Studies https://www.facebook.com/CERESMunk/

137. University of Toronto Slavic Department
http://sites.utoronto.ca/slavic/people/faculty/faculty_index.shtml

138. University of Virginia Department of Slavic Languages and Literatures
https://slavic.as.virginia.edu/

139. University of Washington Department of Slavic Languages and Literatures
https://slavic.washington.edu/

140. University of WashingtonThe Department of Slavic Languages & Literature
https://slavic.washington.edu/

141. University of Waterloo Department of Germanic and Slavic Studies
http://www.germanicandslavic.uwaterloo.ca/

142. University of Wisconsin-Madison Department of Slavic Languages and Literatures https://gns.wisc.edu/slavic/

143. University of WyomingModern & Classical Languages
http://www.uwyo.edu/modlang/

144. Virginia Polytechnic Institute and State University Department of Foreign Languages and Literatures. Russian Section https://liberalarts.vt.edu/departments-and-schools/
department-of-modern-and-classical-languages-and-literatures.html

145. Wellesley College Russian Department https://www.wellesley.edu/russian

146. Wesleyan University Russian and East European Studies
https://www.wesleyan.edu/admission/academics/splashpages/rees.html#:~:text=Wesleyan's%20interdisciplinary%20Russian%2C%20East%20European,A%20minor%20is%20also%20offered.

147. Wichita State University Russian program
https://www.wichita.edu/academics/majors/russian_minor_258.php

148. Yale University Department of Slavic Languages and Literatures
https://slavic.yale.edu/

149. Yale University Russian and East European Studies Program
http://catalog.yale.edu/ycps/subjects-of-instruction/russian-east-european-studies/

150. Yale University Russian Archive Project https://world.yale.edu/russian-archive-project-yale-arts-and-collections

151. York University https://www.york.ac.uk/lfa/courses/russian/

152. Zembla, The Nabokov Butterfly Net. Penn State University
http://www.libraries.psu.edu/nabokov/zembla.htm

***Европа***

1. Association Culturelle Russe de Strasbourg (Франция) www.artradouga.fr

2. Bakhtin Centre (Великобритания) https://www.sheffield.ac.uk/russian/bakhtin

3. Bamberg University (Германия) Slavistik https://www.uni-bamberg.de/slavistik/

4. Bergen University (Норвегия) Department of Russian Studies l
https://www.uib.no/en

5. Berlin Humboldt-University (Германия) Institut für Slawistik
https://www.slawistik.hu-berlin.de/de

6. Brussels University (Бельгия) Departement de Russe
https://www.vub.be/en/home

7. Cambridge russian academy (Великобритания) https://www.camrusacademy.org.uk/ru/

8. Catholic University of Leuven (Бельгия) Slavic Department https://www.arts.kuleuven.be/crs/english

9. Centre for Russian, Soviet and Central and Eastern European Studies (Великобритания) https://crscees.wp.st-andrews.ac.uk/

10. Corpus Cyrillo-Methodianum Helsingiens (Финляндия) https://www.kielipankki.fi/corpora/ccmh/

11. Department of the Russian Languages and Literature (Украина) http://www.knlu.kiev.ua/en/faculties/faculty-of-slavic-philology/department-of-russian-language/

12. Dublin Trinity College (Ирландия) Department of Russian http://www.tcd.ie/Russian/

13. Durham University (Великобритания) Department of Russian https://www.durham.ac.uk/departments/academic/modern-languages-cultures/undergraduate-study/language-areas/russian-studies/

14. ELTE Eötvös Loránd University (Венгрия https://www.elte.hu/en/russian-language-and-literature-ma

15. Georg-August-Universität Gottingen Seminar für Slavische Philologie (Германия) http://wwwuser.gwdg.de/~slavist/welcome.html

16. Ghent University (Бельгия) GCSEES – Ghent Centre for Slavic and East European Studies https://research.flw.ugent.be/en/gcsees

17. Glasgow University (Великобритания) Department of Slavonic Languages and Literatures https://www.gla.ac.uk/postgraduate/research/slavonic/

18. Instituto Ruso Pushkin (Испания) https://institutorusopushkin.com/

19. Jagellonian University (Польша) Russian Graduate Studies https://ces.uj.edu.pl/central-and-east-european-russian-and-eurasian-studies-ceeres-old

20. Keele University (Великобритания) Russian Section https://www.keele.ac.uk/study/languagecentre/modernlanguages/russian/

21. Klagenfurt University (Австрия) Institut für Slawistik https://www.aau.at/slawistik/

22. Lausanne University (Швейцария) Section de langues et civilisations slaves et de l'Asie du Sud https://www.unil.ch/slas/fr/home.html

23. Leipzig University (Германия) Institut für Slavistik http://www.uni-leipzig.de/~slav/

24. Liceo Canova Treviso Liceo Linguistico (Италия) http://www.liceocanova.it/

25. Ludwig-Maximilians-Universitat Muenchen (Германия) Institut für Slavische Philologie http://www.slavistik.uni-muenchen.de/

26. Lund University (Швеция) Russian Studies https://www.lunduniversity.lu.se/lucat/group/v1000135

27. Padua University (Италия) Sezione di Slavistica http://bibliobeatopellegrino.cab.unipd.it/login_form https://www.disll.unipd.it/

28. Potsdam University (Германия) Institut für Slavistik https://www.uni-potsdam.de/de/slavistik/

29. Pushkin House (Великобритания) https://www.pushkinhouse.org/

30. Queen Mary and Westfield College (Великобритания) School of Modern Languages http://www.modern-languages.qmul.ac.uk/

31. Regensburg University (Германия) Institut für Slavistik https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/institut/index.html

32. RNCC Vriendschap (Нидерланды) https://vk.com/rncc.vriendschap

33. Russian school RUBRIC UK (Великобритания) http://www.birminghamrussianschool.org.uk/ru/index.php

34. Shkola.At Русская школа в Линце (Австрия) https://shkola.at/

35. St. Kliment Ohridski University of Sofia (Болгария) Faculty of Slavic Studies http://www.slav.uni-sofia.bg/

36. SOAS (School of Oriental and African Studies) https://www.soas.ac.uk/library/

37. Stockholm University (Швеция) Slaviska institutionen http://slav.su.se/

38. Szeged University (Венгрия Russian Studies (in Russian) https://u-szeged.hu/english/literature-and/russian-studies-in

39. The University of Edinburgh (Великобритания) Department of Russian https://www.ed.ac.uk/literatures-languages-cultures/delc/russian

40. The University of Oxford (Великобритания) Russian https://www.mod-langs.ox.ac.uk/russian

41. Tuebingen University (Германия) Russisch https://uni-tuebingen.de/international/sprachen-lernen/fremdsprachen-lernen/sprachen-kurskonzepte/russisch/

42. Universitat zu Koln Slavisches Institut (Германия) http://www.uni-koeln.de/phil-fak/slavistik/

43. Universite de Mons-Hainaut (Бельгия) Langue russe à l'Ecole d'Interprètes Internationaux de Mons https://web.umons.ac.be/fti-eii/fr/

44. University College London (Великобритания) The School of Slavonic and East European Studies https://www.ucl.ac.uk/ssees/file/16015

45. University of Basel (ШВейцария) Slavisches Seminar http://www.slavistik.unibas.ch/

46. University of Bath (Великобритания) Russian Studies https://www.bath.ac.uk/departments/department-of-politics-languages-international-studies/

47. University of Birmingham (Великобритания) Centre for Russian and East European Studies http://www.crees.bham.ac.uk/

48. University of Bochum (Германия). Das Seminar für Slavistik / Lotman-Institut Für Russische Kultur http://www.slavistik.rub.de/

49. University of Bradford (Великобритания) Department of Modern Languages. http://www.brad.ac.uk/acad/mod-lang/

50. University of Bristol (Великобритания) Department of Russian Studies http://www.bris.ac.uk/russian/

51. University of Business Administration at Vienna (Австрия) Institut für Slawische Sprachen http://www.wu-wien.ac.at/slawisch

52. University of Cambridge (Великобритания) Department of Slavonic Studies http://www.mml.cam.ac.uk/slavonic/

53. University of Copenhagen Department of Cross-Cultural and Regional Studies (Дания) https://ccrs.ku.dk/research/eastern-europe/

54. University of Erlangen-Nuremberg (Германия) Institut fur Slavistik http://www.phil.uni-erlangen.de/~p2slaw/home.html

55. University of Essex (Великобритания) Department of Language and Linguistics. Russian https://www.essex.ac.uk/departments/language-and-linguistics

56. University of Exeter (Великобритания) Department of Russian https://humanities.exeter.ac.uk/modernlanguages/languages/russian/

57. Universita di Genova. Dipartimento di lingue e culture moderne https://lingue.unige.it/node/51

58. University of Greifswald (Германия) Institut fur Slawistik https://slawistik.uni-greifswald.de/

59. University of Groningen (Нидерланды) Кафедра славянских языков и литературр http://www.rug.nl/let/onderwijs/talenenculturen/slavischeTalenCulturen/index

60. University of Halle (Германия) Institut fur Slavistik https://www.slavistik.uni-halle.de/

61. University of Hamburg (Германия) Slavistik http://www.sign-lang.uni-hamburg.de/07/Splan/Slavistik.html

62. University of Hannover (Германия) Leibniz Language Centre https://www.llc.uni-hannover.de/de/leibniz-language-centre/unicert/russisch/

63. University of Heidelberg (Германия) Slavischen Institut http://www.uni-heidelberg.de/institute/fak9/slav/

64. University of Helsinki (Финляндия) Slavic philology https://www.helsinki.fi/en/faculty-arts/research/disciplines/languages/slavic-philology

65. University of Joensuu (Финляндия) Russian Department https://www.uef.fi/en/unit/vera-centre-for-russian-and-border-studies

66. University of Kiel (Германия) Institut für Slawistik http://www.slavistik.uni-kiel.de/

67. University of Konstanz (Германия) Slavistik Fachgruppe Sprachwissenschaft http://www.unikonstanz.de/FuF/Philo/Sprachwiss/slavistik/slavistru.htm

68. University of Leeds (Великобритания) School of Languages, Cultures and Societies https://ahc.leeds.ac.uk/russian-slavonic-studies

69. University of Ljubljana (Словения) Oddelek za slovenistiko http://www.ijs.si/lit/oddel.html

70. University of Manchester (Великобритания) Russian and East European Studies https://www.alc.manchester.ac.uk/modern-languages/study/languages/russian-studies/

71. University of Mannheim (Германия) Slavisches Seminar https://www.phil.uni-mannheim.de/slavistik/

72. University of Muenster (Германия) Slavisch-Baltisches Seminar http://www.uni-muenster.de/SlavBaltSeminar/

73. University of Nottingham (Великобритания) Department of Slavonic Studies http://www.nottingham.ac.uk/slavonic/

74. University of Oldenburg (Германия) Slavische Philologie https://uol.de/slavistik

75. University of Saarland (Saarbrücken) (Германия) Slavistik http://www.uni-saarland.de/fak4/fr44/

76. University of Sheffield (Великобритания) Department of Russian and Slavonic Studies https://www.sheffield.ac.uk/slc/modules/russian-slavonic-studies

77. University of St. Andrews (Великобритания) Department of Russian https://www.st-andrews.ac.uk/modern-languages/subjects/russian/

78. University of Strathclyde (Шотландия) Russian Division https://www.strath.ac.uk/humanities/schoolofhumanities/modernlanguages/

79. University of Surrey (Великобритания) Russian Studies https://www.surrey.ac.uk/russia

80. University of Tampere (Финляндия) Отделение славянской филологии http://www.uta.fi/laitokset/kielet/slaf/

81. University of Tartu (Эстония) Отделение русской и славянской филологии http://www.ut.ee/FLVE/

82. University of Trier (Германия) Slavistik http://slavistik.uni-trier.de/index.php?id=1367

83. University of Turku (Финляндия) The Department of Russian Language and Culture https://www.utu.fi/fi/yliopisto/humanistinen-tiedekunta

84. University of Wales Swansea (Великобритания) Department of Russian https://www.swansea.ac.uk/international-students/my-country/russia/

85. University of Wales, Bangor (Великобритания) https://www.bangor.ac.uk/international/countries/russia

86. University of Wuerzburg (Германия) Institut fur Slavistik und Vergleichende Sprachwissenschaft https://www.phil.uni-wuerzburg.de/maas/startseite/

87. University of Zurich (Швейцария) Slavisches Seminar http://www.unizh.ch/slav/

88. Univerzita Palackého v Olomouci (Чехия) Katedra slavistiky http://www.slavistika.upol.cz/

89. Uppsala Universitet (Швеция) Slaviska institutionen https://www.moderna.uu.se/slaviska/

90. Vitautas Magnus University (Kaunas, Литва) The Centre for Slavistic Studies https://www.vdu.lt/en/research/

91. Warsaw University (Польша) Institute of Russian Studies http://www.ir.uw.edu.pl/

92. Základná škola (Словакия). http://www.skolalamac.stranka.info /

93. Австрийская ассоциация преподавателей русского языка и литературы (Австрия) www.russischlehrer.at

94. El Instituto de lengua rusa A. Pushkin Институт русского языка им. А.С. Пушкина в Барселоне (Испания) http://www.centroruso.es/

95. Sussex Centre for Language Studies (Великобритания) http://www.sussex.ac.uk/languages/resources/russian

96. Фонд славянской письменности и культуры Республики Молдова (Молдова) https://russkie.md/category/slaveanskaia-pisimennosti/

***Другие регионы***

1. Hokkaido University (Япония) Slavic Research Center https://src-h.slav.hokudai.ac.jp/index-e.html

2. Lingvistika Multilingual Ventures, LLP (Индия) https://connect2india.com/LINGVISTIKA-MULTILINGUAL-VENTURES-LLP/5457347

3. Macquarie University (Австралия) Russian Studies https://coursehandbook.mq.edu.au/2020/units/RUSS2010

4. Melbourne University (Австралия) Russian Program https://arts.unimelb.edu.au/school-of-languages-and-linguistics/discipline-areas/russian-studies

5. Nagoya University (Япония) Department of Russian https://www.lang.nagoya-u.ac.jp/bunai/dep/roshiag/index-e.html

6. Russian Language Center in Hong Kong.(Гонконг) http://rlc.edu.hk/

7. Seoul National University (Ю. Корея) Russian Language and Literature https://humanities.snu.ac.kr/en/academics/department?deptidx=6

8. The Hebrew University of Jerusalem (Израиль) The Division of Russian and East European Studies in the Department of Central and East European Cultures https://en.russian.huji.ac.il/

9. The University of Auckland (Новая Зеландия) Germanic Languages & Literature & Slavonic Studies https://www.auckland.ac.nz/en/arts/about-the-faculty/school-of-cultures-languages-and-linguistics.html

10. The University of Canterbury (Новая Зеландия) Department of Russian https://www.canterbury.ac.nz/study/subjects/russian/

11. Victoria University of Wellington (Новая Зеландия) Language Learning Centre http://www.vuw.ac.nz/llc/languages/other-langs/russian.aspx

12. Visva – Bharati University, Santiniketan. Rabindranath Tagore's University (Индия). https://www.studyfrenchspanish.com/visva-bharati-university/

13. Казахский государственный национальный университет имени аль-Фараби (Казахстан). Филологический факультет https://www.kaznu.kz/ru/356/page/

14. Ру́сский культу́рно-образова́тельный центр «Пу́шкинский дом» г. Сеул (Ю. Корея) http://www.pushkinhouse.co.kr/

15. Даляньский университет иностранных языков (Китай). Факультет русского языка https://www.dlufl.edu.cn/ru/info/1006/1652.htm

16. Каталог Русских центров в Китае https://polusharie.com/index.php?PHPSESSID=d69620c24b1a168d6a55fe7d267b989f&topic=133370.0

17. Международный университет Макао (Китай). Магистратура по русскому языку https://www.masterstudies.ru/%D0%A0%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9/

18. Университет Альзахра (Иран). Пушкинский центр обучения и тестирования русского языка https://en.alzahra.ac.ir/pushkin-russian-language-center

19. Центры Россотрудничества https://rs.gov.ru/ru/contacts

20. Центры фонда «Русский Мир» https://russkiymir.ru/rucenter/catalogue.php

***Зарубежные каталоги образовательных ресурсов по русскому языку***

1. Eurolinguiste http://eurolinguiste.com/

2. FluentU Russian Language and Culture Blog https://www.fluentu.com/blog/russian/learn-russian-online/

3. Lingvist – это самый быстрый способ выучить русский язык https://lingvist.com/course/learn-russian-online/

4. Live Fluent, учить русский язык https://livefluent.com/russian-learning-resource/

5. Online Resources for the Study of the Russian language / https://linguaholic.com/topic/1943-online-resources-for-the-study-of-the-russian-language-russian/

6. Online resources in Russian Studies
https://www.mmll.cam.ac.uk/slavonic/resources/online-resources/russian

7. Russian Language Resources For Self Study
https://expressrussian.com/russian-language-resources-for-self-study/

8. Russian Online Resourses https://www.ed.ac.uk/literatures-languages-cultures/delc/russian/online-resources/russian-online-resource

9. Some very useful resources for Russian.
https://forum.duolingo.com/comment/27304619/Some-very-useful-resources-for-Russian

10. The Ultimate Resource Guide for Learning Russian
https://thelanguagesherpa.com/russian-guide/

11. What are good online resources for learning Russian? – Quora
https://www.quora.com/What-are-good-online-resources-for-learning-Russian

12. Британский Совет. Образовательный пакет по русскому языку и культуре
https://www.britishcouncil.org/school-resources/find/classroom/russian-language-culture

13. Изучай русский язык онлайн
http://study-languages-online.com/useful-resources.html

14. Изучение русского языка в сети http://www.sussex.ac.uk/languages/ruslang/

15. Каталог ресурсов для изучающих русский язык
http://web.colby.edu/russian/resources/

16. Каталог ресурсов для студентов, изучающих русский язык
http://folkways.today/resources-students-russian/

17. Кембриджский словарь
https://dictionary.cambridge.org/dictionary/english-russian/resource

18. Лучшие ресурсы для изучения русского языка
https://www.101languages.net/resources/russian/

19. Ресурсы для изучения языков https://mangolanguages.com/

20. Ресурсы для русских лингвистов http://www.royfc.com/lingo.html

21. Ресурсы по изучению русского языка
https://russianreport.wordpress.com/russian-language/russian-language-resources /

22. Русистика на вебе https://www.ruthenia.ru/web/europe/

23. Форум для изучающих русский язык
https://forum.language-learners.org/viewtopic.php?t=5376

## ПРИЛОЖЕНИЕ 10. СЛОВАРИ В СОСТАВЕ БД ОПТЕЛ

| № словаря | Имя словаря | Имя БД | Тип БД | Объем словаря | Словарная статья |
|---|---|---|---|---|---|
| 1. | Тезаурус Языкознание | БД «Словари» для проекта по интеграции ИПЯ | Лексикограф. | 521 | есть |
| 2. | Список ключевых слов ENG | БД Лингвистика АИСОН | Библиогр. | 9920 | нет |
| 3. | Список ключевых слов RUS | БД Лингвистика АИСОН | Библиогр. | 39 224 | нет |
| 4. | Список слов аннотаций ENG | БД Лингвистика АИСОН | Библиогр. | 147 468 | нет |
| 5. | Список слов аннотаций RUS | БД Лингвистика АИСОН | Библиогр. | 226 485 | нет |
| 6. | Словарь Типы ИР | НИРЯЗ | Справ. | 69 | есть |
| 7. | Морфологические признаки | Синтагрус | Корпус | 131 | есть |
| 8. | Синтаксические отношения | Синтагрус | Корпус | 154 | есть |
| 9. | Лексические функции | Синтагрус | Корпус | 90 | есть |
| 10. | Метатекстовые признаки | БД метатекстовой разметки | Корпус | 331 | есть |
| 11. | Морфологические признаки | Лексико-морфологическая БД | Корпус | 165 | есть |
| 12. | Лексика и структура семантических классов | Лексико-семантическая БД | Корпус | 355 | есть |
| 13. | Инвентарь помет | БД метрической разметки | Корпус | 187 | есть |
| 14. | Метатекстовые признаки | Акцентологический корпус | Корпус | 162 | есть |
| 15. | Электронные БД по русским народным говорам | Русские народные говоры | Лексикограф. | 39 | есть |
| 16. | Русский толково-этимологический словарь | Русский толково-этимологический словарь | Лексикограф. | 66 | есть |
| 17. | Статистический словарь языка Ф.М. Достоевского | БД статистического словаря Достоевского | Лексикограф. | 19 | есть |
| 18. | Словарь БД компаративных тропов | БД компаративных тропов | Лексикограф. | 5 | нет |
| 19. | Семантические классы предметов сравнения | БД компаративных тропов | Лексикограф. | 77 | есть |
| 20. | Семантические классы образов сравнения | БД компаративных тропов | Лексикограф. | 12 | есть |

| № словаря | Имя словаря | Имя БД | Тип БД | Объем словаря | Словарная статья |
|---|---|---|---|---|---|
| 21. | Типы тропеических конструкций | БД компаративных тропов | Лексикограф. | 29 | есть |
| 22. | Словарь русских синонимов | Автоматический вариант Словаря русских синонимов | Лексикограф. | 10 | есть |
| 23. | Словарь лексикостатистической базы данных | Лексикостатистическая БД МФРЯ | Лексикограф. | 7 | нет |
| 24. | Словарь русской идиоматики | Русская идиоматика ИПС | Лексикограф. | 5 | нет |
| 25. | Структура словаря языка русской поэзии XX века | БД словаря языка русской поэзии XX века | Лексикограф. | 10 | есть |
| 26. | Грамматические и стилистические пометы | БД словаря языка русской поэзии XX века | Лексикограф. | 357 | есть |
| 27. | Автор, дата создания произведения | БД словаря языка русской поэзии XX века | Лексикограф. | 10 | нет |
| 28. | Словарь графических слов | Частотный словарь языка русской прозы 1850–1870 | Лексикограф. | 186 712 | нет |
| 29. | Словарь лемм | Частотный словарь языка русской прозы 1850–1870 | Лексикограф. | 62 181 | есть |
| 30. | Обратный словарь лемм | Частотный словарь языка русской прозы 1850–1870 | Лексикограф. | 49 790 | нет |
| 31. | Ранговый словарь лемм | Частотный словарь языка русской прозы 1850–1870 | Лексикограф. | 20 151 | нет |
| 32. | Языки | ИС для описания языков малочисленных народов | ГИС | 248 | есть |
| 33. | Словари | ИС для описания языков малочисленных народов | ГИС | 1 | нет |
| 34. | Аудио | ИС для описания языков малочисленных народов | ГИС | 3 | есть |
| 35. | Источник данных | ИС для описания языков малочисленных народов | ГИС | 3 | есть |
| 36. | Годы | ИС для описания языков малочисленных народов | ГИС | 1 | Нет |
| 37. | Населенный пункт | ИС для описания языков малочисленных народов | ГИС | 1 | нет |
| 38. | Авторы | ИС для описания языков малочисленных народов | ГИС | 1 | нет |
| 39. | Степень угрозы вымирания языка | ИС для описания языков малочисленных народов | ГИС | 6 | есть |
| 40. | Грамматика | ИС для описания языков малочисленных народов | ГИС | 115 | есть |
| 41. | Теги Албанского корпуса | Албанский национальный корпус | Корпус | 137 | есть |
| 42. | Грамматические пометы | Русская академическая неография. Неологизмы русского языка | Лексикограф. | 33 | есть |
| 43. | Стилистические пометы | Русская академическая неография. Неологизмы русского языка | Лексикограф. | 41 | есть |
| 44. | Список языков | Корпус вепсского и карельского языка | Корпус | 16 | есть |
| 45. | Список диалектов | Корпус вепсского и карельского языка | Корпус | 97 | есть |

| № словаря | Имя словаря | Имя БД | Тип БД | Объем словаря | Словарная статья |
|---|---|---|---|---|---|
| 46. | Части речи | Корпус вепсского и карельского языка | Корпус | 43 | есть |
| 47. | Грамматические признаки | Корпус вепсского и карельского языка | Корпус | 55 | есть |
| 48. | Семантические отношения | Корпус вепсского и карельского языка | Корпус | 9 | есть |
| 49. | Семантические формулы топонимов | Тематический сайт по топонимии Европейского Севера России | Лексикограф. | 45 | есть |
| 50. | Ключевые слова журнала «Вопросы языкознания» | Сайт журнала «Вопросы языкознания» | Сайт | 1345 | нет |
| 51. | Генеалогические классы языков | Языковая энциклопедия Лингвисто. орг | Справ. | 446 | есть |
| 52. | Рубрикатор языкового фасета | НИРЯЗ | Справ. | 106 | есть |
| 53. | Тезаурус по педагогике (Языкознание) | Электронный каталог ГНПБ им. Ушинского | Библиогр. | 126 | есть |
| 54. | Рубрикатор «Языкознание» | АИСОН | Библиогр. | 421 | есть |
| 55. | Тезаурус «Лингвистика» | АИСОН | Библиогр. | 2948 | есть |

**А.Б. Антопольский**

**ЛИНГВИСТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ**

***Под научной редакцией Д.В. Ефременко***

**Монография**

**Серия**
«Наука, образование и технологии»

Оформление обложки И.А. Михеев
Техническое редактирование
и компьютерная верстка И.К. Летунова
Корректор Д.Г. Валикова

Гигиеническое заключение
№ 77.99.6.953.П.5008.8.99 от 23.08.1999 г.
Подписано к печати 2 / VIII – 2022 г.
Формат 70×100/16 Бум. офсетная № 1 Печать офсетная
Усл. печ. л. 37,5 Уч.-изд. л. 27,5
Тираж 300 (1–100 экз. – 1-й завод) Заказ № 29

**Институт научной информации**
**по общественным наукам Российской академии наук (ИНИОН РАН)**
Нахимовский проспект, д. 51/21, Москва, 117418
http://inion.ru

**Отдел маркетинга и распространения**
**информационных изданий**
Тел.: +7 (925) 517-36-91, +7 (499) 134-03-96
e-mail: shop@inion.ru

Отпечатано по гранкам ИНИОН РАН
ООО «Амирит»,
410004, Саратовская обл., г. Саратов,
ул. Чернышевского, д. 88, литера У