ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ ... 8

ЧАСТЬ 1. ОРГАНИЗАЦИЯ ДЕЯТЕЛЬНОСТИ В СФЕРЕ ЛИР ... 11
ГЛАВА 1. ЛИР – ОПРЕДЕЛЕНИЕ И ТИПОЛОГИЯ ... 11
Вводные замечания ... 11
Типологии специальных ЛИР ... 11
Типологии в рамках широкого подхода к ЛИР ... 15
Литература к главе 1 ... 24
ГЛАВА 2. СОБРАНИЯ ЛИР ... 25
Введение ... 25
Мировые собрания ЛИР ... 26
Европейские собрания ЛИР ... 32
Карта LRE ... 42
Литература к главе 2 ... 45
ГЛАВА 3. МЕЖДУНАРОДНАЯ ДЕЯТЕЛЬНОСТЬ В ОБЛАСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ЦИФРОВОЙ ГУМАНИТАРИСТИКИ ... 46
Общие замечания ... 46
Профессиональные ассоциации ... 46
Консорциумы ... 50
Защита и сохранение исчезающих языков ... 56
Терминологическая и переводческая деятельность ... 62
Цифровая гуманитаристика ... 66
Литература к главе 3 ... 71
ГЛАВА 4. ИНФРАСТРУКТУРА ЯЗЫКОВЫХ РЕСУРСОВ И ТЕХНОЛОГИЙ: ЕВРОПЕЙСКИЙ ОПЫТ ... 72
Введение ... 72
Инициативы и стратегии по научной инфраструктуре ... 73
Информационные ресурсы и проекты открытой науки ЕС ... 74
Инфраструктурные консорциумы ERIC ... 76
Европейские объединения и проекты ... 81

ЧАСТЬ 2. ТЕХНОЛОГИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ 89
ГЛАВА 5. МЕЖДУНАРОДНАЯ СТАНДАРТИЗАЦИЯ ЯЗЫКОВЫХ РЕСУРСОВ И ТЕХНОЛОГИЙ 89
Вводные замечания 89
Международные органы по стандартизации 90
Тематика международных стандартов и спецификаций 96
Информационные системы по стандартизации 107
Российская стандартизация в области ЛИР и языковых технологий 108
Литература к главе 5 109
ГЛАВА 6. МЕТАДАННЫЕ ЛИР 110
Краткая история 110
Проект метаданных IMDI [1] 111
Метаданные OLAC 112
Метамодель META-SHARE 114
Международный стандартный номер ЛИР (ISLRN) 120
Карта LRE 121
Стандартизация метаданных 130
Остинские принципы цитирования лингвистических данных 133
Реестр категорий данных для ЛИР 134
Исследование лексики метаданных российских ЛИР 140
Литература к главе 6 142
ГЛАВА 7. ЛИНГВИСТИЧЕСКАЯ АННОТАЦИЯ 145
Общие сведения 145
Справочник по лингвистической аннотации 146
Семинар по лингвистическим аннотациям (LAW) 148
Стандартизация лингвистического аннотирования 152
Литература к главе 7 155
ГЛАВА 8. ЯЗЫКОВАЯ ДОКУМЕНТАЦИЯ 157
Введение 157
Международные проекты языковой документации 158
Документирование языков, находящихся под угрозой исчезновения DOBES 161
Языковая документация и ресурсы для ревитализации языков Living Languages 171
Российские ЛИР языковой документации 172
Литература к главе 8 175
ГЛАВА 9. КАТАЛОГИ И БИБЛИОТЕКИ ЛИНГВИСТИЧЕСКОГО ИНСТРУМЕНТАРИЯ 176
Каталоги лингвистических программ 176
Европейские каталоги ПО 182
Российские каталоги лингвистического ПО 185
Библиотеки лингвистических программ 186
Российские разработчики лингвистического ПО 191
Литература к главе 9 194

ЧАСТЬ 3. КАТЕГОРИИ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ 195
ГЛАВА 10. ТЕКСТОВЫЕ КОРПУСА 195
Общие замечания 195
Статистика корпусов 196
Классификации корпусов 197
Банки деревьев (treebanks) 201
Инструментальные средства корпусной лингвистики 202
Корпусная лингвистика в России 204
Литература к главе 10 214
ГЛАВА 11. ЛЕКСИЧЕСКИЕ РЕСУРСЫ И КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ 215
Определение, классификация, статистика 215
Международное сотрудничество по электронной лексикографии 217
Функциональность электронных словарей 219
Концептуальные лексико-семантические ЛИР 222
Представление электронных словарей 224
Средства программной поддержки электронных словарей 228
Электронная лексикография в России 231
Литература к главе 11 238
ГЛАВА 12. ТЕРМИНОЛОГИЧЕСКИЕ БАЗЫ ДАННЫХ 240
Общие сведения 240
Европейский опыт 241
Терминологические структуры Еврокомиссии 243
Мировые терминологические структуры 246
Российские ТБД 248
Память перевода 249
Номенклатуры, классификации, таксономии 250
Литература к главе 12 257
ГЛАВА 13. ТИПОЛОГИЧЕСКИЕ ЛИР 259
Общие сведения 259
Зарубежные типологические ЛИР 261
Российские типологические ЛИР 270
Литература к главе 13 275
ГЛАВА 14. РЕСУРСЫ ЗВУЧАЩЕЙ РЕЧИ 277
Общие сведения 277
Классификация речевых корпусов 278
Статистика ЛИР звучащей речи 279
Краткая история создания ЛИР звучащей речи 279
Обзоры ЛИР звучащей речи 280
Проектирование речевых ЛИР 282
Разработки ресурсов устной речи в России 284
Литература к главе 14 293
ГЛАВА 15. ЛИНГВИСТИЧЕСКИЕ КАРТЫ И АТЛАСЫ 295
Общие сведения 295

Крупнейшие международные проекты лингвистических карт и атласов ... 297
Собрания цифровых лингвистических карт ... 304
Российские проекты ... 307
Литература к главе 15 ... 311
ГЛАВА 16. РЕСУРСЫ ЖЕСТОВЫХ ЯЗЫКОВ ... 312
Общие сведения ... 312
Список жестовых языков ... 313
Обследование ЛИР жестовых языков ... 314
Мировые жестовые (знаковые) языки ... 317
Литература к главе 16 ... 319
ГЛАВА 17. ОБРАЗОВАТЕЛЬНЫЕ ЛИР ... 320
Общие сведения ... 320
Каталоги лингвистических ЭОР ... 321
Рекомендательные сервисы ... 324
Литература к главе 17 ... 326
ГЛАВА 18. РЕСУРСЫ ПО РУССКОМУ ЯЗЫКУ В ЗАРУБЕЖНЫХ СОБРАНИЯХ ... 327

ЧАСТЬ 4. ПЕРСПЕКТИВЫ РАЗВИТИЯ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ ... 335
ГЛАВА 19. ЛИНГВИСТИЧЕСКИЕ СВЯЗАННЫЕ ОТКРЫТЫЕ ДАННЫЕ (LLOD) ... 335
Общие сведения ... 335
Облако LLOD ... 336
Семинар по LLOD (LDL) ... 340
Проекты по развитию LLOD ... 341
Литература к главе 19 ... 351
ГЛАВА 20. ЛИР В КОНТЕКСТЕ ЦИФРОВОЙ ГУМАНИТАРИСТИКИ ... 353
Общие сведения ... 353
Консорциум DARIAH ... 354
Huma-Num. Программа цифровой гуманитаристики ... 357
Российские гуманитарные ресурсы с лингвистическим компонентом ... 364
Литература к главе 20 ... 368
ГЛАВА 21. ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ РОССИЙСКОЙ ИНФРАСТРУКТУРЫ ЛИР ... 369
Вводные замечания ... 369
Российская ситуация ... 370
Справочно-информационная система по языкознанию ... 372
Стратегия информационной инфраструктуры языковых технологий и ресурсов ... 376
Литература к главе 21 ... 377
УКАЗАТЕЛЬ АКРОНИМОВ ... 378
РУССКИЕ (КИРИЛЛИЧЕСКИЕ) СОКРАЩЕНИЯ ... 384

ПРИЛОЖЕНИЯ 385
ПРИЛОЖЕНИЕ 1. КАТАЛОГИ, АРХИВЫ И РЕПОЗИТАРИИ ЛИР 385
ПРИЛОЖЕНИЕ 2. РОССИЙСКИЕ КАТАЛОГИ ЛИР 391
ПРИЛОЖЕНИЕ 3. МЕЖДУНАРОДНЫЕ ОРГАНИЗАЦИИ В ОБЛАСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ЦИФРОВОЙ ГУМАНИТАРИСТИКИ. 393
ПРИЛОЖЕНИЕ 4. ПРОЕКТЫ, СТАНДАРТЫ, ФОРМАТЫ, РЕСУРСЫ 397
ПРИЛОЖЕНИЕ 5. ЦЕНТРЫ ЗНАНИЙ CLARIN 412
ПРИЛОЖЕНИЕ 6. РОССИЙСКИЕ СТАНДАРТЫ НА ЛИР И СМЕЖНЫЕ ВОПРОСЫ 417
ПРИЛОЖЕНИЕ 7. ИНСТРУМЕНТЫ ЛИНГВИСТИЧЕСКИХ ТЕХНОЛОГИЙ . 424
ПРИЛОЖЕНИЕ 8. МИРОВЫЕ ТЕРМИНОЛОГИЧЕСКИЕ БАНКИ ДАННЫХ . 446
ПРИЛОЖЕНИЕ 9. ЗАРУБЕЖНЫЕ ЦЕНТРЫ И РЕСУРСЫ ПО РУСИСТИКЕ .. 450
ПРИЛОЖЕНИЕ 10. СЛОВАРИ В СОСТАВЕ БД ОПТЕЛ 463

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).