Информационные процессы / Information processes

Обзорная статья / Review article
УДК 004.9:025.3
https://doi.org/ 10.31432/1994-2443-2024-19-1-5-20

# Языковые ресурсы в современном информационном пространстве

**А.Б. Антопольский** ✉
*Институт научной информации по общественным наукам Российской академии наук (ИНИОН РАН), г. Москва, Российская Федерация*
✉ *ale5695@yandex.ru*

**Аннотация.** Подчеркивается важное место языковых ресурсов в современной информационной индустрии. Предлагается обзор международной инфраструктуры языковых ресурсов, включая собрания, репозитарии и архивы языковых ресурсов. Перечисляются международные организации, действующие в этой области, особенно в области стандартизации. Излагается история разработки метаданных языковых ресурсов. Кратко характеризуются основные категории языковых ресурсов. В качестве перспективы и основного направления развития языковых ресурсов описывается деятельность сообщества открытых лингвистических открытых данных.

**Ключевые слова:** языковые ресурсы, лингвистические данные, информационное пространство, стандартизация, международная инфраструктура, международные организации

**Финансирование.** Финансирование отсутствовало.

**Для цитирования:** Антопольский А.Б. Языковые ресурсы в современном информационном пространстве. *Информация и инновации*. 2024;19(1):5-20. https://doi.org/10.31432/1994-2443-2024-19-1-5-20

© Антопольский А.Б., 2024
(cc) BY

2024;19(1):5-20 Информация и инновации / Information and Innovations

# Language resources in the modern information space

**A.B. Antopolsky** ✉
*Institute of Scientific Information for Social Sciences of the Russian Academy of Sciences (INION RAS), Moscow, Russian Federation*
✉ *ale5695@yandex.ru*

**Abstract.** The important place of language resources in the modern information industry is emphasized. Provides an overview of the international language resource infrastructure, including collections, repositories, and archives of language resources. International organizations active in this area, especially in the field of standardization, are listed. The history of the development of metadata for language resources is outlined. The main categories of language resources are briefly characterized. The activities of the open linguistic open data community are described as a perspective and main direction for the development of language resources.

**Keywords**: language resources, linguistic data, information space, standardization, international infrastructure, international organizations

**Funding.** No funding.

**For citation:** Antopolsky A.B. Language resources in the modern information space. *Information and Innovations.* 2024;19(1):5-20. (In Russ.). https://doi.org/10.31432/1994-2443-2024-19-1-5-20

## Введение

Организованные языковые данные в цифровой форме, или языковые ресурсы, (далее — ЯР), в последние десятилетия стали важнейшим инструментом в самых различных компьютерных и человеко-машинных технологиях и процессах современной индустрии обработки данных и интеллектуальных систем. Это прежде всего машинный перевод, речевые технологии (в частности, автоматический анализ и синтез устной речи), голосовое общение, информационный поиск, автоматическое реферирование и др. ЯР легли в основу новых дисциплин, таких как корпусная лингвистика или электронная лексикография.

Цифровые ЯР активно используются также для поддержки традиционных задач языковой политики и практики, таких как терминологическая, переводческая, редакторская деятельность; изучение и преподавание родного и иностранных языков, сохранение исчезающих и находящихся в опасности языков и др.

Широкий фронт применения ЯР вызвал к жизни и массовое производство различных ЯР. Действительно, в крупнейших языковых архивах счет идет на десятки и даже сотни тысяч ЯР. 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 профессиональное сообщество. Речь, прежде всего, идет о необходимости перехода к промышленным технологиям создания, обработки и применения ЯР. Это требует внедрения методов управления, стандартизации, применения типовых программных продуктов, архивации ЯР, повторное использование и т.д.

В этих условиях актуальным становится создание информационной инфраструктуры, которая обеспечивала бы сохранность и повторное использование лингвистических информационных ресурсов (ЛИР), коллаборацию и взаимодействие между разработчиками ЛИР, в основном учеными-лингвистами и IT-компаниями, распространение передовой практики, внедрение промышленных методов и прочее.

В настоящей статье предлагается краткий обзор международной инфраструктуры цифровой языковой индустрии, а также основных категорий ЯР.

## Собрания и типология ЯР

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

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

Значительная часть этих собраний включена в регистр лингвистических архивов OLAC (http://www.language-archives.org/archives), перечень лингвистических мета-сайтов Linguist list (https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=25) или каталог репозиториев научных данных RE3 (https://www.re3data.org/).

В обзоре представлены описания наиболее известных собраний собственно языковых ресурсов (Language Resources). Существует и более широкий подход, ко-

торый можно назвать лингвистическими ресурсами (Linguistic Resources), который включает широкий класс научной и практической литературы или справочной информации по языкознанию. Обзор этих ресурсов выходит за сферу данной статьи.

***Облако связанных лингвистических открытых данных LLOD***

Центральным способом сбора, архивации и интеграции ЯР и организации эффективных коллабораций является проект связанных лингвистических открытых данных (LLOD), основанный на платформе Семантического веба. В этом проекте предлагается следующая типология ЯР:

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

Наиболее полное описание современного состояния LLOD имеется в работе [1]. Реализацию проекта LLOD осуществляет международная рабочая группа OWLG, 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. Практические результаты этой коллаборации представлены в Облаке LLOD, включающем десятки крупнейших мировых ЯР (https://linguistic-lod.org/llod-cloud).

***Сообщество открытых языковых архивов OLAC***

Наиболее полным собранием ЯР является собрание Открытого консорциума языковых архивов (OLAC http://www.language-archives.org/). 63 архива, входящие в OLAC, в совокупности содержат свыше 400 тыс. ЯР.

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

- по языкам и семействам языков;
- по наличию ЯР в онлайне;
- по странам и регионам;
- по наименованию архива;
- по типу ЯР;
- по типу дискурса;
- по области лингвистики;
- по типу Дублинского ядра метаданных;
- по формату;
- по предметным рубрикам Библиотеки Конгресса и др.

***Справочная система для поиска информации об языковых ресурсах Linghub (http://linghub.org)***

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

Стандартное описание ЛИР в *Linghub* включает следующие реквизиты:

- создатель;
- описание;
- права доступа;
- источник;
- предмет;
- наименование.

***Каталог Консорциума лингвистических данных (LDC https://catalog.ldc.upenn.edu/)***

Каталог LDC включает около 900 ЯР, по большей части корпусов и лексиконов, созданных в научных или исследователь-

ских целях в университетах, входящих в Консорциум.

***Архив языков и культуры SIL (https://www.sil.org/resources/language-culture-archives )***

SIL — это глобальная некоммерческая организация, которая работает с местными сообществами по всему миру над разработкой языковых решений для улучшения жизни. Среди проектов SIL важное место занимает архив языков и культур. 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 (https://www.elararchive.org)***

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 испанский, мандаринский, английский или русский.

***Проект архивирования лингвистических данных (Langues et Civilisations à Tradition Orale (LACITO).– URL: http://xml.coverpages.org/ lacitoAR-desc-english.html#Resume)***

Целью проекта архивирования языковых 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).

***Российский архив уральских и алтайских языков на платформе ЛингвоДок (http://lingvodoc.tsu.ru)***

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.

***Европейские собрания ЛИР***

В рамках Евросоюза реализовано несколько крупных проектов, направленных на сбор, хранение, поиск и обмен ЯР. Достаточно подробное их описание имеется в [2], поэтому мы ограничимся их перечислением с соответствующими ссылками.

*Европейская координация языковых ресурсов (ELRC: https://lr-coordination.eu/).*

*Европейская исследовательская инфраструктура языковых ресурсов и технологий CLARIN https://www.clarin.eu).*

*Языковой архив Института психолингвистики Общества Макса Планка (TLA https://archive.mpi.nl/tla/).*

*Каталог языковых ресурсов Европейской ассоциации языковых ресурсов (ELRA http://portal.elda.org)*.

*Оценочная карта языковых ресурсов (карта LRE http://lremap.elra.info).*

*Универсальный каталог* (*http://www.elra.info/en/catalogues/universal-catalogue).*

*Виртуальная языковая обсерватория* (*VLO https://vlo.clarin.eu).*

*Поиск и обмен языковыми ресурсами (META-SHARE http://www.meta-share.org).*

В работе [2] автором предложена следующая типология. Всего выделено 8 категорий ЯР, обладающих структурным и функциональным сходством:

- Текстовые корпуса;
- Лексические ресурсы;
- Терминологические базы данных;
- Типологические ресурсы;
- Ресурсы звучащей речи;
- Лингвистические карты и атласы;
- Ресурсы жестовых языков;
- Образовательные ресурсы.

Ниже будет представлен краткий обзор категорий ЯР в соответствии с этой типологией.

## Международное сотрудничество в области создания и использование ЯР

В настоящем обзоре кратко перечисляются различные международные коллаборации в сфере создания и использования ЯР. Более подробные сведения об их деятельности можно найти в монографии [2].

### *Ассоциации*

Основной формой взаимодействия между специалистами разных стран в сфере создания и использования ЯР являются профессиональные ассоциации по прикладной и компьютерной лингвистике. Они проводят конференции, издают журналы, создают порталы и сайты, организуют учебу и семинары, находят другие формы взаимодействия. Вот некоторые из них:

- 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);
- Азиатская федерация по обработке естественного языка (AFNLP).

### *Консорциумы*

В отличие от ассоциаций, обычно включающих специалистов как физических лиц, консорциумы объединяют организации, в том числе университетские лаборатории. Консорциумы, как правило, поддерживают постоянные проекты, в том числе проекты собраний ЯР, которые перечислены ниже:

- *Консорциум лингвистических данных LDC;*
- *OLAC — Сообщество открытых лингвистических архивов;*
- *Международное сообщество лингвистов онлайн LINGUIST List;*
- *SIL International — международная некоммерческая организация (бывший Летний институт*

*лингвистики — Summer Institute of Linguistics).*

***Охрана исчезающих языков***

Еще одним типом коллабораций в области ЯР следует считать организации, нацеленные на сохранение исчезающих языков и языков, находящихся в опасности. Этим занимается ЮНЕСКО, а также специализированные организации:

- *Программа ЮНЕСКО по поддержке языкового разнообразия и многоязычия в Интернете;*
- *Атлас ЮНЕСКО по языкам мира, находящимся под угрозой исчезновения;*
- *Фонд исчезающих языков FEL;*
- *Программа документирования исчезающих языков ELDP;*
- *Программа документирование исчезающих языков DOBES;*
- *Электронная метаструктура данных об исчезающих языках E-MELD;*
- *Проект по архивированию лингвистических данных LACITO;*
- *Фонд библиотеки долговременного хранения естественных языков Rosetta.*

***Международная стандартизация***

Особая роль в коллаборации по языковым ресурсам и технологиям принадлежит стандартизации. Прежде всего это Международная организация по стандартизации (ISO), где ведущей структурой по языковой индустрии является Технический Комитет 37 «Язык и терминология» (ISO/TK 37). В состав ТК 37 входят пять подкомитетов:

- 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.

Важнейшим для сферы ЯР является ПК4, в составе которого действуют следующие рабочие группы:

- 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.

Полный аннотированный перечень 70 стандартов, разработанных ISO/TK 37, а также проектов стандартов, доступен на сайте ТК 37 (https://www.iso.org/committee/48104.html).

Разработкой и внедрением международных стандартов в области языковых ресурсов и технологий занимается еще ряд международных организаций, которые здесь кратко перечисляются:

- *Консорциум Всемирной сети W3C;*
- *Инициатива по кодированию текстов TEI;*
- *Международные стандарты для языковых технологий ISLE;*
- *Международная федерация библиотечных ассоциаций IFLA;*
- *Ассоциация отраслевых стандартов локализации LISA;*
- *Организация по развитию стандартов структурированной информации OASIS;*
- *Ассоциация глобализации и локализации GALA;*
- *Инициатива по языку разметки правил RuleML.*

Тематическая классификация стандартов и спецификаций на ЯР и языковые

технологии представлена в информационной системе стандартизации CLARIN-D (https://clarin.ids-mannheim.de/standards/views/list-topics.xq):

- 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.

## Метаданные языковых ресурсов

Одной из центральных задач разработки языковых ресурсов и технологий является создание и развитие системы метаданных, пригодных для всего разнородного сообщества, работающего с ЯР.

Постановка задачи разработки специальной системы метаданных для ЯР принадлежит, по-видимому, рабочей группе EAGLES/ISLE, которая в 2001 году предложила план разработки соответствующего стандарта. 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 ЯР, известная как Международный стандартный номер ЯР (ISNLR), которая опирается на правила OLAC, и, независимо, — карта оценки ЯР (LRE map), использующая другую модель описания ЯР.

Подробное описание истории разработки систем метаданных для ЯР представлено в работе [4]. Более компактное описание можно найти в [5].

Наконец, в 2015 и в 2019 гг. появились две части стандарта ISO на метаданные для ЛИР, основанные на CMDI[1,2].

[1] 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 (date of application: 01.12.2021).

[2] 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 (date of application: 01.04.2022).

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 должны быть упомянуты.

Реестр категорий лингвистических данных (DCR) под эгидой ISO TC37 был создан в 2008 году в Институте психолингвистики Общества Макса Планка (MPI) в Неймегене, Нидерланды, под названием ISOcat. Однако по мере развития проекта отпала необходимость в регистрирующем органе, и хранилище ISOcat перестало быть проектом ISO. Был создан DatCatInfo (https://datcatinfo.net) — это репозиторий категорий данных (DCR), который заменяет ISOcat.

Следующим этапом развития регистра стало создание нового реестра CLARIN Concept Registry (CCR https://www.clarin.eu/ccr).

Наконец, в качестве модуля метаданных пакета словарей Ontolex OntologyLexicon разработан, а затем размещен на облаке LLOD *Словарь лингвистических метаданных (LIM E https://old.datahub.io/dataset/lime*).

## Категории ЯР

Подробное описание языковых ресурсов, конечно, выходит за рамки статьи. Здесь мы кратко охарактеризуем основные категории ЯР, предложенные выше.

### *Текстовые корпуса*

Лингвистический, или языковой, корпус текстов — это большой, цифровой унифицированный, структурированный, размеченный массив языковых данных, предназначенный для решения конкретных, чаще филологических задач. Основными чертами современного корпуса являются машиночитаемый формат, репрезентативность, наличие метаданных. Корпуса стали базой для новой дисциплины компьютерной лингвистики, а именно — корпусной лингвистики [6].

Корпуса стали самым распространенным видом ЯР. Так, в крупнейшем в мире языковом архиве Института психолингвистики Общества Макса Планка из 150 тыс. ЯР корпуса составляют 91,2 тыс.

Во многих каталогах в качестве отдельной категории выделяют корпуса с синтаксической и/или семантической разметкой. Их принято называть банками деревьев, поскольку разметка обычно имеет древовидную структуру. Они достаточно популярны, так, в англоязычной википедии: (https://en.wikipedia.org/wiki/Treebank) приводится перечень около 300 таких банков.

Корпуса звучащей речи (речевые корпуса) в нашей типологии выделены в отдельную категорию, что определялось спецификой представления ЯР и инструментальных средств для их обработки.

### *Лексикографические ресурсы*

Электронная, или е-лексикография, также стала самостоятельной дисциплиной [7]. Лексикографических ЯР также создается много, в упомянутом архиве TLA представлено 25 тыс. такого типа ЯР. Лексикографические ЯР образуют самостоятельный объект международного сотрудничества: в Евросоюзе действует проект *Европейская сеть электронной лексикографии* (ENEL), а также *Европейская ассоциация лексикографии (EURALEX*). Аналогичные ассоциации созданы и в других регионах мира.

Кроме обычных языковых словарей к лексикографическим ресурсам относят разнообразные концептуальные или лексико-семантические ЯР, такие как тезаурусы, онтологии, семантические сети и др. Проблематика этого типа ЯР подробно описана в [8].

Создается большое количество средств программной поддержки электронных словарей. Активно идет международное сотрудничество по представлению электронных словарей. В последние годы основным стандартом формата стала *Простая система организации знаний* (SKOS, https://www.w3.org/TR/skos-reference), а языком разметки для словарей - LMF [9].

***Терминологические базы данных***

Особым типом лексикографических ЯР являются терминологические базы данных (ТБД)[3]. Их выделяют в отдельный тип, поскольку создание и применение ТБД обычно выходит за пределы лингвистических технологий и предназначено для международных организаций, отраслевых специалистов, переводчиков, редакторов.

Особенно важен европейский опыт управления терминологией, ведь для Евросоюза преодоление языкового барьера при сохранении равенства языков является магистральной политической задачей. Поэтому ТБД Евросоюза обслуживают целую сеть сотрудников и переводчиков. Эту деятельность координирует несколько структур. Среди них *TermCoord* — группа по координации терминологии при Европарламенте. Другая структура - *Центр знаний по интерпретации,* поддерживающая свыше тысячи ТБД. Для переводчиков разработана специальная поисковая система *Lithos*[4]. Она позволяет переводчикам искать термины во всех ТБД.

Важную роль в международной организации терминологической деятельности играют ООН и ЮНЕСКО. Их ТБД доступны через сервисы *Многоязычная терминология (UNTERM, https://unterm.un.org/unterm/portal/welcome)* и *ЮНЕСКОТЕРМ (https://web.archive.org/web/ 20110728143741/ http://termweb.unesco.org/).*

Еще одной известной международной организацией, работающей в сфере ТБД, является Международный консорциум *TerminOrgs,* в котором разработаны *Руководство по терминологической деятельности для корпораций (http://www.terminorgs.net)*, а также спецификация упрощенного варианта формата TBX-Basic (https://www.tbxinfo.net/).

Разновидностью ТБД являются базы данных, называемые «памятью перевода» (translation memory, TM), широко распространенные в системах поддержки переводческой деятельности. Популярные программные системы TM представлены на ресурсе[5].

***Типологические ресурсы***

Самостоятельную категорию ЯР образуют типологические БД (ТИПБД), обеспечивающие структурированное представление данных о свойствах языков, в том числе содержащие классификации языков в различных аспектах: генеалогиче-

[3] 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 (date of application: 01.04.2022).

[4] Lithos. URL: https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/terminology-tools-and-resources/terminology-dg-interpretation_en#lithos

[5] Топ-10 программ памяти переводов. URL: https://translationrating.ru/top-10-cat-tools-2017/

ском, ареальном, эволюционном. ТИПБД создаются для задач контрастивной лингвистики, лингвистической типологии и языковой систематики и других лингвистических дисциплин.

Анализ функциональности ТИПБД можно найти в работе В.Д. Соловьева [10].

Наиболее полным собранием ТИПБД является портал Ассоциации лингвистической типологии, на котором есть специальный раздел *Базы данных* (*https://linguistic-typology.org/databases).* Достаточно полный обзор типологических ЯР, представленных в Интернете, имеется также в работе «Введение в лингвистическую типологию» [11].

***Ресурсы звучащей речи***

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.

Классификация речевых корпусов предложена в другой работе О.Ф. Кривновой [12]. Более подробный обзор и анализ речевых корпусов можно найти в коллективной монографии [13]. Можно указать также на обзор «Устные корпуса» (https://www.clarin.eu/resource-families/spoken-corpora), в котором описывается 90 корпусов устной речи, которые входят в инфраструктуру CLARIN. 79 из них содержат как транскрипции устной или спонтанной речи, так и соответствующие аудиозаписи, а 11 — только транскрипции. Корпуса содержат транскрипции, например, новостей в эфире или повествований и диалогов.

***Лингвистические карты и атласы***

ЯР, опирающиеся на географические информационные системы (ГИС) и предназначенные для изучения пространственных аспектов языков, диалектов и языковых свойств, называются лингвистическими картами.

Обзор и анализ проблем методов лингвогеографии и применения цифровых карт можно найти в работах [14,15]. Наиболее полные собрания лингвистических карт и атласов имеется на ресурсах *Сокровищница лингвистических карт TLM (https://www.degruyter.com/database/tlm/html), Лингвистическая картография Эстремадуры* (http://www.geolectos.com/atlas.htm), *Карты Мэтью Драйера (https://www.acsu.buffalo.edu/~dryer/family_maps.htm)* и некоторых других. Из российских проектов нужно отметить ресурсы *Общеславя́нский лингвисти́ческий а́тлас (ОЛА, http://www.slavatlas.org)* и *Лингвариум. Языковые карты (http://www.lingvarium.org/maps.shtml).*

***Ресурсы жестовых языков***

Жестовые языки являются языками, которые используют визуально-ручную модальность, чтобы передать смысл. Языки жестов — это полноценные естественные языки со своей грамматикой и лексикой.

Основной ресурс, где имеется перечень около 300 жестовых языков с необходимыми ссылками и описаниями, это *Список жестовых языков (https://ru.abcdef.wiki/wiki/List_of_sign_languages).* Следует также отметить комплексное обследование доступных в Интернете словарей жестовых языков, которое выполнил рос-

сийский исследователь А.Е. Харламенков из НИИ «Русского жестового языка»[6].

***Образовательные ЯР***

Применение цифровых ЯР в образовательных целях, т.е. для изучения и/или преподавания иностранных или неродных языков представляет собой огромное поле разнообразных ресурсов. Соответственно существует множество обзоров, классификаций, рейтингов, методических материалов по их использованию. Ограничимся ссылками на российский и иностранный обзоры этой категории ЯР [16, 17].

Также существует множество каталогов образовательных ЯР. Укажем на наиболее известный российский каталог *Единое окно доступа к образовательным ресурсам (http://window.edu.ru)*. Только каталогов интернет–ресурсов в помощь учителю русского языка и литературы известно полтора десятка.

Точно также имеется множество каталогов ресурсов по изучению других иностранных языков, в том числе по онлайн обучению. Среди них выделяются рекомендательные сервисы. Приведем примеры:

- Как выбрать цифровые ресурсы для использования в вашем языковом классе (https://www.cambridge.org/elt/blog/2020/06/29/how-to-select-digital-resources-for-use-in-your-language-classroom);
- 60 сайтов и приложений для изучения иностранных языков (https://blog.mann-ivanov-ferber.ru/2019/02/21/60-sajtov-i-prilozhenij-dlya-izucheniya-inostrannyx-yazykov/).

Нужно отметить, что образовательные ЯР скорее относятся к сфере педагогики, нежели лингвистики и поэтому в мировых каталогах ЯР почти не фиксируются.

## Заключение: перспективы развития ЯР

Распространенное мнение, к которому присоединияется автор, заключается в том, что наиболее перспективной платформой для развития ЯР и коллаборации в этой сфере является платформа Семантического веба, на которой ЯР будут развиваться в форме открытых связанных лингвистических данных (LLOD). Выше мы упоминали проект Облака LLOD, в котором уже интегрированы крупнейшие мировые ЯР (https://linguistic-lod.org/llod-cloud).

Далее изложим основные черты этого направления.

***Принципы LLOD***

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;

[6] 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

- 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.

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.

***Стандарты и рекомендации LLOD***

Помимо сбора метаданных и создания облачной диаграммы LLOD, сообщество OWLG стимулирует разработку стандартов сообщества в отношении словарей, метаданных и рекомендаций передового опыта. К ним относятся стандарты:

- 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);
- для категорий лингвистических данных;
- для идентификации языка;
- для метаданных.

Кроме того, на сайте LLOD (https://linguistic-lod.org/) опубликован комплекс рекомендаций для различных видов деятельности по применению LLOD.

***Дискуссии и обучение LLOD***

С момента своего создания серия семинаров по связанным данным в лингвистике (LDL) зарекомендовала себя как основная площадка для обсуждения того, как связанные открытые данные (LOD) и технологии Семантического веба могут использоваться для обработки, анализа, публикации и управления лингвистическими данными. Последний 8-й семинар по этой проблеме состоялся в 2022 г. (https://ldl2022.linguistic-lod.org/).

***Исследовательские проекты LLOD***

Активное развитие данного направления компьютерной лингвистики выражается в том числе во многих международных проектах. Приведем некоторые примеры последних лет:

- LiODi. Связанные открытые словари (2015-2020);
- FREME. Открытая платформа электронных услуг для многоязычного и семантического обогащения цифрового контента (2015-2017);
- POSTDATA. Стандартизация поэзии и связанные открытые данные (2016-2021);

- Linking Latin Связывание латиницы (2018-2023);
- Pret-a-LLOD Подготовка лингвистических данных к использованию (2019-2021);
- NexusLinguarum. Европейская сеть веб-исследований лингвистических данных (2019-2023);
- ELEXIS Создание лексикографической инфраструктуры (2020 -2022).

Все это свидетельствует, что связанные открытые лингвистические данные являются мейнстримом современной языковой индустрии. Исследователи, работающие в данной области, а также органы управления наукой и информационной индустрией должны учитывать это при планировании научной и практической деятельности, особенно связанной с информационной инфраструктурой.

## КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

## CONFLICT OF INTERESTS

The author declare no relevant conflict of interests.

## СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Linguistic Linked Data: Representation, Generation and Applications / Cimiano P., Chiarcos C., McCrae John P., Gracia J. — Springer International Publishing, 2020.
2. Антопольский А.Б. Лингвистические информационные ресурсы: монография/ ИНИОН РАН, Фундам. б-ка; науч. ред Д.В. Ефременко. - Москва: ИНИОН РАН, 2022. - 466 с.
   Antopol`skij A.B. Lingvisticheskie informacionny`e resursy`: monografiya/INION RAN, Fundam. b-ka; nauch. red D.V. Efremenko. - Moskva: INION RAN, 2022. - 466 p.
3. Антопольский А.Б. Международная деятельность в области языковых технологий и ресурсов. *Информация и инновации*. 2021;16(2):5-22. https://doi.org/10.31432/1994-2443-2021-16-2-5-22
   Antopolsky A.B. International Activities in the Field of Language Technologies and Resources. *Information and Innovations*. 2021;16(2):5-22. (In Russ.) https://doi.org/10.31432/1994-2443-2021-16-2-5-22
4. 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 (date of application: 01.12.2021).
5. Антопольский А.Б. Метаданные лингвистических ресурсов: история и современное состояние. *Известия Российской академии наук. Серия литературы и языка.* 2022;81(1):21–36. https://doi.org/10.31857/S160578800018917-4
   Antopolsky A.B. Metadannye lingvisticheskih resursov: istoriya i sovremennoe sostoyanie [Metadata of Linguistic Resources: History and Current State]. *Izvestiâ Rossijskoj akademii nauk. Seriâ literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2022;81(1):21–36. (In Russ.). https://doi.org/10.31857/S160578800018917-4

6. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. — Санкт-Петербург: Издательство СПбГУ, 2019. — ISBN 978-5-288-05997-1.
Zaxarov V.P., Bogdanova S.Yu. Korpusnaya lingvistika. — Sankt-Peterburg: Izdatel`stvo SPbGU, 2019. — ISBN 978-5-288-05997-1.
7. 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; 14:304–324.
Zaxarov V.P. E`lektronnaya leksikografiya XXI veka. *Slovo i slovar`. = Vocabulum et vocabularium: sbornik nauchny`x statej,* red. O.N. Kry`lova, S.A. My`znikov. — Sankt-Peterburg: Nestor-Istoriya. 2016; 14:304–324.
8. 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.
Ontologii i tezaurusy`: modeli, instrumenty`, prilozheniya / Dobrov B.V., Ivanov V.V., Lukashevich N.V., Solov`ev V.D. — Moskva: Izd-vo INTUIT, 2008. — 176 p.
9. LMF Lexical Markup Framework, ISTE / Gil Francopoulo (edited by). — Wiley, 2013. — ISBN 978-1-84821-430-9.
10. Соловьев В.Д. Типологические базы данных: перспективы использования. *Вопросы языкознания*. 2010;1:94–110. EDN MBCGIZ.
Solov'ev V.D. Tipologicheskie bazy dannyh: perspektivy ispol'zovaniya. *Voprosy yazykoznaniya*. 2010;1:94–110.
11. Edith A. Moravcsik. Introducing Language Typology. — Cambridge: Cambridge University Press, 2013. — ISBN:9780511978876. https://doi.org/10.1017/CBO9780511978876.
12. Кривнова О.Ф. Речевые корпуса на новом технологическом витке. *Речевые технологии = Speech technology*. 2008;2:14–23.
Krivnova O.F. Rechevy`e korpusa na novom texnologicheskom vitke. *Rechevy`e texnologii = Speech technology.* 2008;2:14–23.
13. 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.
Rechevaya kommunikaciya v informacionnom prostranstve: kollektivnaya monografiya / Potapova R.K., Potapov V.V., Dolinskij V.A., Xitina M.V., Xarlamov A.A., Bazhenova I.Yu., Komalova L.R., Bobrov N.V., Gordeev D.I., Os`kina K.A.; otv. red. Potapova R.K. — Moskva: LENAND, 2017. — 112 s. — ISBN 978-5-9710-3940-2.
14. Jay Lee, Jiajun Qiao, Dong Han. GIS in Linguistic Research. — 2018. https://doi.org/10.1016/B978–0-12–409548–9.09662–7
15. Candice R. Luebbering, Korine N. Kolivras, Stephen P. Prisley. Visualizing Linguistic Diversity Through Cartography and GIS. *The Professional Geographer.* 2013; 65(4). https://doi.org/10.1080/00330124.2013.825517
16. Есенина Н.Е. Обзор электронных образовательных и информационных ресурсов для обучения иностранному языку. *Информатика и образование.* 2012;12:103–105.

Esenina N.E. Obzor e`lektronny`x obrazovatel`ny`x i informacionny`x resursov dlya obucheniya inostrannomu yazy`ku. *Informatika i obrazovanie*. 2012;12:103–105.
17. 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. https://doi.org/10.2991/aebmr.k. 200324.035

## ИНФОРМАЦИЯ ОБ АВТОРЕ

**Александр Борисович Антопольский,** доктор технических наук, профессор, главный научный сотрудник, Институт научной информации по общественным наукам Российской академии наук (ИНИОН РАН), Нахимовский проспект, д. 51/21, 117418, г. Москва, Российская Федерация; https://orcid.org/0000-0002-1670-8746; e-mail: ale5695@yandex.ru

## INFORMATION ABOUT THE AUTHOR

**Alexander B. Antopolsky,** Dr. Sci. (Eng.), Professor, Chief Researcher, Institute of Scientific Information for Social Sciences of the Russian Academy of Sciences (INION RAS), Nakhimovsky ave., 51/21, 117997, Moscow, Russian Federation; https://orcid.org/0000-0002-1670-8746; e-mail: ale5695@yandex.ru

**Поступила в редакцию / Received** 26.03.2024