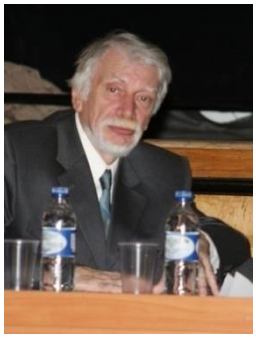
ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ И ТЕХНОЛОГИИ В РОССИИ: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ (Обзор)



Александр Борисович Антопольский

Доктор технических наук, профессор, главный научный сотрудник Институт научной информации по общественным наукам РАН (ИНИОН РАН), Москва, Россия

Аннотация. *Определяется понятие лингвистических информационных ресурсов, приводится обзор их классификаций. Описываются наиболее значимые российские каталоги лингвистических информационных ресурсов и ведущих организаций страны в области компьютерной лингвистики. Обсуждаются первоочередные задачи создания российской инфраструктуры лингвистических информационных ресурсов.*

Ключевые слова: *компьютерная лингвистика; искусственный интеллект; лингвистические ресурсы; информационная инфраструктура.*

Для цитирования: Антопольский А.Б. Лингвистические ресурсы и технологии в России: состояние и перспективы. (Обзор) // Социальные новации и социальные науки. – Москва : ИНИОН РАН, 2021. – № 2. – С. 114–131.

URL: https://sns-journal.ru/ru/archive/ DOI: 10.31249/snsn/2021.02.08

Введение

Лингвистические технологии (т.е. технологии, которые позволяют находить и обрабатывать информацию) в последние десятилетия стали одними из значимых технологических инноваций цифровой эпохи. Среди современных направлений лингвистики можно выделить следующие [Прикладная лингвистика, 2021]:

- машинный перевод;
- автоматическое распознавание речи;
- лингвистическое обеспечение информационного поиска;
- автоматическое извлечение данных из текста;
- автоматическое реферирование текстов;
- создание электронных лексикографических ресурсов (словарей, онтологий);
- создание и использование электронных корпусов текстов (корпусная лингвистика);
- разработка вопросно-ответных систем.

Перечисленные направления часто определяют как отдельную область – автоматическую обработку естественного языка (natural language processing, NLP), – которая является общим полем для технологий искусственного интеллекта и компьютерной лингвистики.

Центральный элемент компьютерной лингвистики представляют лингвистические информационные ресурсы (ЛИР). Количество ЛИР в мире быстро растет. Крупнейший на сегодня Языковой архив (The Language Archive – TLA) Института М. Планка (Германия) содержит около 150 тыс. ЛИР. Суммарное количество ЛИР, отраженных в архивах, которые вошли в состав Открытого консорциума лингвистических архивов (Open Language Archive – OLAC), достигает 400 тыс. [Home, 2011; Main page, 2021].

Классификация и каталогизация лингвистических информационных ресурсов

Главной проблемой при описании ЛИР является их идентификация и классификация, т.е. отнесение разных видов информационных массивов и продуктов к определенным категориям. Составители разнообразных порталов, каталогов, справочных систем, репозиториев и иных собраний ЛИР, либо сведений о них придерживаются существенно различных взглядов на эти вопросы.

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

При первом из них ЛИР – это лингвистические данные и инструменты, которые непосредственно используются в языковых технологиях[1]. Прежде всего, это (лингвистические) корпусы[2], лексиконы (словари), «банки деревьев»[3], лингвистические процессоры[4], описания языков и др. Назовем такой подход узким, а класс ресурсов, который относят к ЛИР сторонники данного подхода, – *специальными ЛИР*.

Второй подход определяет ЛИР более широко и включает любые ресурсы, создаваемые или используемые лингвистами в профессиональной деятельности. Назовем такой подход широким, а такие ЛИР – *тематическими*, поскольку они, как правило, выделяются из универсальных структур по тематическому принципу. К ним относятся, например, электронные библиотеки, библиографии, труды конференций, периодика, энциклопедии, персоналии и тому подобные ресурсы.

Примером узкого подхода к ЛИР может служить типология, предлагаемая англоязычным сайтом Википедии [Language resource, 2021]:

- данные, в том числе:

– лексические ресурсы, например машиночитаемые тексты;

– лингвистические корпусы;

– лингвистические базы данных (БД), такие как коллекции кросс-лингвистических связанных данных;

- инструменты, в том числе:

– аннотации и инструменты для создания таких аннотаций в ручном или полуавтоматическом режиме (синтактический анализ, семантический анализ и т.д.);

– приложения для поиска и извлечения таких данных;

- метаданные и словари, репозитории лингвистической терминологии и языковых метаданных.

Аналогичный подход используется в упоминавшихся выше TLA и OLAC.

Напротив, самый посещаемый в мире справочный портал по лингвистике LINGUIST List придерживается широкого подхода. В данном случае основные разделы выглядят следующим образом [Recent Postings, 2021]:

- люди и организации;
- вакансии;

[1] Языковые технологии, иначе технологии человеческого языка (human language technology, HLT), изучают методы того, как компьютерные программы или электронные устройства могут анализировать, воспроизводить, изменять или реагировать на человеческие тексты и речь. Они включают обработку естественного языка (natural language processing, NLP), компьютерную лингвистику и речевые технологии (по материалам Википедии).

[2] Цифровые коллекции данных на естественных языках.

[3] Проанализированные разными лингвистическими способами текстовые корпуса, которые аннотируют синтаксическую или семантическую структуру предложения.

[4] Компьютерные программы, обеспечивающие анализ, синтез и преобразование текстов на естественном языке.

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

Подобный подход к типологии ЛИР присутствует и в руководстве по лучшим лингвистическим ресурсам в Интернете («Метаиндекс лингвистики, естественного языка и компьютерной лингвистики»), созданном в Стэнфордском университете [A guide to the best linguistic resources ..., 2014].

По нашему мнению, наиболее перспективным собранием ЛИР в мире является Облако лингвистически связанных открытых данных (Linguistic Linked Open Data, LLOD). В него включены только ЛИР, которые при узком подходе относятся к данным [Home, 2018], в том числе:

- лингвистические корпусы;
- лексиконы / словари;
- терминологические ЛИР, тезаурусы;
- метаданные ЛИР;
- категории лингвистических данных;
- типологические базы данных.

Российские каталоги лингвистических информационных ресурсов

Российских каталогов ЛИР на сегодняшний день существует свыше 50, включая каталоги образовательных ресурсов по русскому языку, которые в этой статье не затрагиваются. Среди их создателей присутствуют сторонники обоих подходов. Примерами каталогов, созданных на основе широкого подхода, служат следующие.

Навигатор информационных ресурсов по языкознанию (НИРЯЗ), разработанный в 2019–2020 гг. группой сотрудников в ИНИОН РАН по гранту РФФИ, включает не только цифровые, но и бумажные ЛИР, в частности библиотечные фонды, архивные и музейные документы. В НИРЯЗ входит около 1,2 тыс. ЛИР, созданных в учреждениях РАН [О проекте, 2021]. Специальные ЛИР выделены в отдельный класс. Сокращенная типология ЛИР этого каталога выглядит следующим образом:

- библиотеки;
- архивы;
- музеи;
- каталоги;

- электронные коллекции и библиотеки;
- информационные системы;
- справочники, энциклопедии;
- персональные ресурсы;
- лингвистические ресурсы, в том числе:
  - корпусы текстов,
  - словарные базы данных и электронные картотеки,
  - лингвистические процессоры,
  - грамматические ресурсы,
  - описания и реестры языков,
  - лингвистические атласы,
  - этно- и социолингвистические БД,
  - комплексные лингвистические ресурсы (сайты),
  - информационные языки;
- периодика;
- библиографии;
- мероприятия;
- неопубликованные материалы;
- медиаресурсы;
- прочие интернет-ресурсы.

Каталог ресурсов для обработки естественного языка (NLPub) появился в 2012 г. (автор проекта Д.А. Усталов) и создается на принципах краудсорсинга. В этом случае используется следующая классификация ЛИР [Заглавная страница, 2020]:

- методы и инструменты, в том числе:

обработка текста, обработка речи, утилиты, методы, алгоритмы;

- ресурсы, в том числе:

словари, тезаурусы, корпус текстов, коллекции n-грамм (последовательности из n слов и их частоты в больших массивах текстов), банки данных;

- эксперты и мероприятия, в том числе:

организации, персоналии, конференции;

- образование, в том числе:

образование (учреждения высшего образования в России и зарубежные аспирантуры, онлайн-курсы и т.д.), литература, темы дипломов;

- проекты.

Портал знаний по к*омпьютерной лингвистике* разрабатывается с 2007 г. сотрудниками Института систем информатики имени А.П. Ершова СО РАН совместно с учеными из других организаций СО РАН, Москвы и Казани [Новости, 2021]. В этом проекте классификация ЛИР представлена наиболее подробно и фундаментально разработана. Фактически построена онтология понятий, относящихся к данной области знаний. В частности, верхние уровни этой классификации и раздел «лингвистические БД» включают следующие составляющие: интернет-ресурсы, методы и средства исследования, научные результаты и продукты, лингвистические ресурсы и корпусы, 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, невербальную коммуникацию, речевые произведения, структурные языковые единицы, научные мероприятия.

Обзор некоторых категорий российских лингвистических информационных ресурсов

Сколько-нибудь полный анализ имеющихся в России ЛИР выходит далеко за рамки журнальной статьи. В связи с этим ограничимся кратким обзором наиболее заметных отечественных ЛИР, а также разработок в области языковых технологий.

Корпусы текстов. Лингвистический, или языковой, корпус текстов – это большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. Основными чертами современного корпуса являются машиночитаемый формат, репрезентативность, наличие металингвистической информации. В настоящее время корпусы стали важнейшим типом ЛИР, так как с их помощью лингвисты решают многие научные и прикладные задачи. Описанию проблем корпусной лингвистики посвящена, например, работа [Захаров, Богданова, 2011].

Создание лингвистических корпусов в России началось в 1980-е годы с работ по формированию Машинного фонда русского языка. В 2012–2014 гг. действовала программа Президиума РАН «Корпусная лингвистика», в рамках которой было реализовано много проектов по созданию и развитию корпусов русского языка и языков народов России.

На сегодняшний день центральное место среди отечественных корпусов принадлежит Национальному корпусу русского языка (НКРЯ[1]), который включает следующие разделы (подкорпусы) [Национальный корпус русского языка …, 2009]:

- основной корпус (современные и ранние письменные тексты);

[1] Работы по его созданию были начаты в 2001 г. специалистами из Москвы, Санкт-Петербурга и Воронежа, к которым в дальнейшем присоединились лингвисты из других научных центров России. Открыт в 2004 г. Современный владелец сайта – Институт русского языка им. В.В. Виноградова РАН (Москва).

- газетный корпус СМИ 2000-х годов;
- газетный региональный корпус;
- диалектный корпус;
- обучающий корпус;
- корпус параллельных текстов;
- поэтический корпус;
- устный корпус;
- акцентологический корпус (истории русского ударения);
- мультимедийный корпус;
- древнерусский корпус;
- берестяные грамоты;
- старорусский корпус;
- церковнославянский корпус;
- другие корпусы.

Суммарный объем НКРЯ в марте 2021 г. составлял: число текстов – 2 419 215 единиц; число предложений – 78 694 781 единиц; число словоупотреблений – 961 081 047 единиц. Кроме того, в разделе «Другие корпуса» приведены ссылки на другие общедоступные онлайн-ресурсы. Функциональные возможности НКРЯ, а также принципы разметки и представления отдельных подкорпусов, описаны в работе [Национальный корпус русского языка …, 2009].

К заметным отечественным информационным ресурсам относятся также следующие:

- проекты по корпусной лингвистике Школы лингвистики НИУ ВШЭ [Проекты, 2021];
- корпусы Института этнологии и антропологии РАН [Корпусы ИЭА РАН, 2019].

Отечественными специалистами созданы учебные, диалектные и диахронические корпусы русского языка, корпусы языков народов России, а также корпусы для ряда иностранных языков. Перечень последних приведен на портале [Корпуса языков России, 2021].

Корпусы звучащей речи. Речевой корпус – база данных аудиофайлов и транскрипций текстов, – разновидность корпусов текстов, которую, однако, часто выделяют в отдельный тип ЛИР в силу очевидной специфики. Речевые корпусы используются в фонетических исследованиях, диалектологии, при создании акустических моделей и в других областях. Существует два типа речевых корпусов: базы начитанных текстов и базы аудиозаписей спонтанной речи.

Одним из первых российских ЛИР для звучащей речи была БД ISABASE, разработанная в 1990-х годах в Институте системного анализа РАН [База речевых фрагментов …, 1998]. В настоящее время в России создано значительное количество ЛИР звучащей речи.

Например, в 2009 г. на базе фонограммархива Института русской литературы РАН (Пушкинский Дом, Санкт-Петербург) при финансовой поддержке Правительства Ямало-Ненецкого АО был создан Национальный электронный звуковой депозитарий с уникальными фольклорными записями народов Крайнего Севера, Сибири и Дальнего Востока России. Хотя этот депозитарий не являлся сугубо лингвистическим – в нем, кроме текстов, хранилось большое количество музыкальных записей. В электронной базе было представлено почти 2000 различных записей на восьми языках, носителей некоторых из которых осталось буквально единицы [Институт русской литературы …, 2009; Коллекция Национального электронного звукового депозитария …, 2019]. Однако на момент написания статьи доступ к этому ресурсу заблокирован.

ЛИР звучащей речи также присутствуют в «устном» и «мультимедийном» подкорпусах НКРЯ, а также в специальном проекте Школы лингвистики НИУ ВШЭ [Проекты Школы лингвистики, 2021]. Можно еще отметить устный подкорпус Национального корпуса калмыцкого языка[1] [Главная, 2012], Корпус русской устной речи СПбГУ[2] [Корпус русской устной речи, 2021] и корпусы, созданные в Санкт-Петербургском институте информатики и автоматизации РАН (СПИИ РАН) [Инновационная продукция, 2021].

Лексические ресурсы и компьютерная лексикография

Создание и применение электронных (машиночитаемых, цифровых) словарей было исторически первым направлением компьютерной лингвистики, и до сих пор остается самым популярным. Уже давно сформировалась специальная дисциплина, – компьютерная лексикография, – которая занимается этими проблемами, а лексические ресурсы – самый распространенный тип ЛИР.

Лексические ресурсы. Согласно наиболее распространенному определению, электронный словарь – это любой упорядоченный, относительно конечный массив лингвистической информации, представленный в виде списка, таблицы или перечня, удобного для размещения в памяти ЭВМ и снабженного программами автоматической обработки и пополнения [Компьютерная лексикография, 2021].

Существует множество классификаций электронных словарей, поскольку традиционная лексикография породила великое множество их разновидностей. Например, в работе [Попова, 2012] выявлены 155 классификационных признаков различных словарей. Представляется, что слишком детальная классификация нерелевантна для компьютерной лексикографии, поскольку все эти раз-

[1] Создан в 2012–2014 гг. в Калмыцком национальном центре (КИГИ) РАН при финансовой поддержке РГНФ.
[2] Создается с 2009 г. за счет гранта РФФИ.

новидности структурно можно объединить в ограниченное количество классов ЛИР. Например, в обзоре ЛИР, сделанном в CLARIN[1], выделяется всего пять классов [Resource families, 2021]:

- лексиконы – в основном используется в NLP-приложениях. Они обычно содержат обширный лексический запас с конкретной лингвистической информацией (например, морфосинтаксис);
- словари – создаются в основном для использования человеком (например, для изучения языка, перевода, лексикологии) и, как правило, являются семасиологическими, т.е. организованы вокруг слов и содержат информацию об их значениях, определениях, произношении и т.д.;
- концептуальные ресурсы – включают такие лексические ресурсы, как словарные и фреймовые[2] сети, таксономии, тезаурусы и онтологии; обычно связаны семантическими отношениями (например, гипернимии и гипонимии[3]);
- глоссарии – специализированные словари, содержащие специфичную для данной предметной области терминологию и / или выражения. К ним относятся и терминологические БД;
- списки слов – алфавитные или частотные лексические перечни (конкордансы).

В настоящее время пользователям Интернета доступны сотни как оцифрованных традиционных словарей на различных языках, так и специализированных лексикографических БД различного назначения. Например, каталог «Топ 100» Рамблера по запросу «словари» дает ссылки почти на 300 сайтов (https://top100.rambler.ru/search?query=словари).

Среди агрегаторов лексикографических ресурсов наиболее популярны порталы: Грамота.ру (http://new.gramota.ru), Словари.онлайн (https://slovaronline.com), Мультитран (https://www.multitran.com), Академик (https://translate.academic.ru/), Slovar.cc (https://slovar.cc).

Сведения о российских лексикографических ЛИР для научных исследований собраны в Навигаторе информационных ресурсов по языкознанию, созданном в 2019–2020 гг. [О проекте, 2021]. Всего там описано около 100 лексикографических ЛИР, разработанных в учреждениях РАН, в том числе различные словари на основе НКРЯ. Разнообразные лексикографические ЛИР используются во всех современных лингвистических процессорах: переводчиках, образовательных ресурсах, системах анализа и синтеза речи, поисковых машинах.

Терминологические БД и глоссарии. Данный тип ЛИР большинство специалистов выделяет как самостоятельный, поскольку терминологические БД создаются не столько для решения лин-

[1] Common Language Resources and Technology Infrastructure («Общая инфраструктура для языковых ресурсов и технологий») создана в 2012 г. в ЕС в рамках Европейского консорциума исследовательской инфраструктуры (European Research Infrastructure Consortium – ERIC) для поддержки исследователей в области гуманитарных и социальных наук.

[2] Специфический способ представления знаний в технологии искусственного интеллекта, представляющий собой схему действий в реальной ситуации.

[3] Иерархические отношения, отражающие взаимосвязь между более общим (гиперонимом) и конкретным (гипонимом) понятиями.

гвистических задач, сколько в помощь отраслевым специалистам, переводчикам и редакторам. В России имеется несколько терминологических БД, среди которых наиболее известна БД «Российская терминология» (БД РОСТЕРМ) [Российская терминология …, 2021], созданная в 1980-х годах на основе массива стандартизованных терминов и поддерживаемая ФГУП «СТАНДАРТИНФОРМ» (ранее ВНИИКИ).

Концептуальные ресурсы. В России накоплен богатый опыт создания информационно-поисковых тезаурусов – в 1970–1980-х годах сотни их были разработаны по различным отраслям науки. Однако позже тезаурусы вышли из употребления из-за распространения более простых и дешевых методов полнотекстового поиска. Впрочем, традиционные информационно-поисковые тезаурусы продолжают функционировать в ряде организаций: ИНИОН РАН, ЦНСХБ, ЦНМБ, ООО «Интегрум Медиа». Обзоры применения тезаурусов содержатся в работах [Лукашевич, 2011; Антопольский, 2003]. Появление и распространение онтологий, как универсального средства представления понятийной структуры предметной области, снова привлекли внимание к тезаурусам, которые являются промежуточным этапом создания онтологий.

В сфере онтологий в России лидирует ООО «Лаборатория информационных исследований» [Лаборатории информационных исследований, 2021]. Разработанный здесь тезаурус RU TEZ служит базой для многих онтологических проектов и используется в университетской информационной системе (УИС Россия). Одним из заметных российских ЛИР является также созданный в СПбГУ компьютерный тезаурус RussNet [RussNet, 2005].

Ценным источником онтологической и лексико-семантической информации выступает русский Викисловарь [Русский Викисловарь, 2021]. Из проектов последних лет нужно отметить YARN – открытый электронный тезаурус русского языка, создаваемый на принципах краудсорсинга [Yet Another RussNet, 2018]. Разработчик YARN Д.А. Усталов также реализовал проект интеграции нескольких тезаурусов русского языка в семантическую сеть LLOD [Усталов, 2017].

Компьютерная лексикография в России. В России компьютерная лексикография развивалась очень активно, начиная с 1970-х годов, когда создавались разнообразные машинные словари. В 1990-х годах появились коммерческие лексикографические продукты на переносимых носителях, среди которых наибольшую популярность получила система Lingvo компании ABBYY[1]. Сейчас продукты этой компании широко доступны онлайн [Lingvo Live, 2021].

Одним из актуальных направлений развития компьютерной лексикографии является организация обмена и повторного использования лексических ЛИР, поскольку их создание представляет собой достаточно трудоемкий и затратный процесс. Еще до эпохи Интернета для обмена лексиче-

[1] Создана в 1989 г. в России. В настоящее время – успешная международная компания – разработчик решений в области интеллектуальной обработки информации и анализа бизнес-процессов, распознавания текстов и лингвистики.

скими ЛИР появились специальные коммуникативные форматы, последний из них получил название MARTIF [Computer applications in terminology …, 1999]. Сейчас этот стандарт отменен, и ему на смену пришли современные форматы обмена на основе языков разметки, прежде всего XML.

Наиболее популярным методом формализованного представления лингвистической информации в цифровом виде стала инициатива TEI (Text Encoding Initiative). Применение этого метода для кодирования словарей подробно описано в работе [Захаров, 2013]**.** На основе указанной инициативы разработаны международные стандарты для представления лексикографических данных. В настоящее время существует целая серия таких стандартов, которые представлены на портале Международной организации по стандартизации (ISO) [Management of terminology resources …, 1985]. Некоторые из этих стандартов переведены на русский язык и утверждены в качестве национальных. В России этим занимается Технический комитет по стандартизации «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле» (ТК 055) [О Техническом комитете по стандартизации …, 2016].

Для концептуальных лексикографических ресурсов, отражающих семантические отношения, основным инструментом представления стал SKOS (Simple Knowledge Organization System – «простая система организации знаний») – разработанная Консорциумом W3[1] модель организации знаний, призванная облегчить взаимодействие различных информационных систем за счет стандартизации тезаурусов, систем классификации, таксономий, фолксономий[2] и др. [Home Page, 2012].

Ведущие организации России в области компьютерной лингвистики

В России действует несколько исследовательских центров и коммерческих компаний, имеющих заметные достижения в компьютерной лингвистике. К их числу относятся следующие.

ООО «Яндекс». Компания начала свою деятельность именно с языковых технологий. Даже название «Языковой иНДЕКС» она получила от морфологического анализатора (разработанного одним из основателей компании, И. Сегаловичем). Этот анализатор Mystem и сейчас является одним из лучших для русского языка. В компании разработана и действует система словарей и машинного перевода, а также голосовой помощник Алиса со встроенной системой анализа и синтеза речи. Технологии «Яндекса», включая языковые, описаны на отдельной странице портала компании [Технологии, 2021].

ООО «ABBYY». Компания предлагает словари и продукты для переводчиков (например, ABBYY Lingvo x6) на различных платформах, а также популярный оптический распознаватель

[1] World Wide Web Consortium (W3 C) – независимое международное сообщество, созданное для разработки и внедрения единых стандартов работы сети Интернет на основе унификации, общедоступности и безопасности. Создан в 1994 г. и в настоящее время объединяет более 400 членов.

[2] Народная классификация, практика совместной категоризации информации (текстов, ссылок, фото, видеоклипов и т.п.) посредством произвольно выбираемых меток, называемых тегами (по материалам Википедии).

символов FineReader. Продукты и решения ABBYY описаны на портале компании [Lingvo Live, 2021].

ООО «Лаборатория информационных исследований». Компанией разработаны следующие технологии [Лаборатория информационных исследований, 2021]:

- анализа текстов (классификация, аннотирование, многоязычный поиск) на основе больших лингвистических онтологий;
- оценки тональности, извлечения фактографической информации из текста;
- кластеризации, классификации и обзорного реферирования новостного потока.

Группа компаний «Центр речевых технологий» – разработчик инновационных систем в сфере технологий синтеза и распознавания речи [Группа компаний ЦРТ, 2021].

ООО «PROMT» – разработчик систем машинного перевода; занимается также другими исследованиями и разработками в области ИИ. На сайте компании представлен комплекс переводчиков для различных приложений и платформ для 40 языков мира [Продукты, 2021].

Школа лингвистики НИУ ВШЭ. Научные группы школы проводят исследования в области типологии, социолингвистики и ареальной лингвистики, корпусной лингвистики и лексикографии, древних языков и истории языка. Кроме того, разрабатываются лингвистические технологии и ресурсы: корпусы, обучающие тренажеры, словари и тезаурусы, технологии для электронного представления текстов культурного наследия. Проекты Школы перечислены на ее сайте [Проекты Школы лингвистики, 2021].

Кафедра математической лингвистики СПбГУ. Специалисты кафедры ведут исследования в области автоматической обработки текстов на разных языках, лингвистической семантики, синтаксиса, теории моделирования, терминоведения, автоматической лексикографии, стилеметрии, автоматической атрибуции текстов, квантитативной лингвистики[1]. Результаты их деятельности отражены, например, в коллективной монографии [Николаев, Митренина, Ландо, 2017], а проекты перечислены на сайте кафедры [Научно-исследовательские проекты кафедры …, 2021].

Лаборатория компьютерной лингвистики Института проблем передачи информации им. А.А. Харкевича РАН (ИППИ РАН). Основные направления научной деятельности включают [Лаборатория № 15 …, 2021]:

- развитие модели языка «смысл <=> текст»;
- совершенствование «многоцелевого лингвистического процессора ЭТАП-3» – компьютерной системы анализа и синтеза текстов на основе модели «смысл <=> текст»;
- поддержание одной из составляющих НКРЯ – глубоко аннотированного лингвистического корпуса.

[1] Раздел математической лингвистики, исследует язык при помощи статистических методов анализа.

Лаборатория «Цифровая документация русского языка» ИППИ РАН ведет работу по следующим направлениям [Лаборатория № 20 …, 2021]:

- разработка лингвистических корпусов;
- корпусные и экспериментальные исследования русского языка;
- русская корпусная грамматика.

Отдел корпусной лингвистики и лингвистической поэтики Института русского языка им. В.В. Виноградова РАН (ИРЯ РАН). Основное направление деятельности – развитие НКЯР [Отдел корпусной лингвистики …, 2021].

Лаборатория общей и компьютерной лексикологии и лексикографии филологического факультета МГУ им. М.В. Ломоносова осуществляет разнообразные проекты в рассматриваемой сфере. Их перечень представлен на сайте Лаборатории [Главная страница, 2021].

Институт прикладной семиотики АН РТ выполняет комплексные разработки ЛИР и процессоров для автоматической обработки татарского языка, перечень которых приведен на сайте института [Фундаментальные и прикладные разработки, 2021].

Европейский опыт создания инфраструктуры лингвистических информационных ресурсов

Российские достижения в области компьютерной лингвистики очевидны. Однако широкий фронт этих исследований требует координации и повышения уровня повторного использования ЛИР. В этой связи можно обратиться к европейскому опыту, в частности деятельности *CLARIN* [The research infrastructure …, 2021].

CLARIN –это распределенная инфраструктура, которая предоставляет доступ к ЛИР, технологиям и знаниям, а также обеспечивает сотрудничество между академическими кругами, промышленностью, политиками, культурными и образовательными учреждениями и широкой общественностью. Цель CLARIN состоит в том, чтобы обеспечить единый онлайн-вход ко всем цифровым ЛИР Европы и за ее пределами в письменной, устной или мультимодальной форме. За период, прошедший с момента начала ее функционирования, созданы следующие сервисы [Resource families, 2021].

– *депозитные услуги и архивирование:* для обеспечения устойчивого хранения ЛИР (например, корпусы, лексиконы, аудио- и видеозаписи, грамматики и т.д.);

– *виртуальная языковая обсерватория:* предоставляет простой в использовании интерфейс с развитым синтаксисом, который поддерживает единый процесс поиска разнообразных ЛИР и создания виртуальных коллекций;

– *федеративный поиск* (проект): система поиска данных в центрах хранения. Сами данные остаются у владельца, поэтому поиск называется федеративным. Поисковая система суммирует и

отображает то, что доступно, а для более сложного запроса нужно перейти к специализированному интерфейсу поиска в центре – владельце ЛИР;

– *легкий доступ к защищенным ресурсам:* благодаря федеративному входу в систему защищенные приложения и наборы данных доступны всем, у кого есть учетная запись, запросить которую могут любые пользователи;

– *коммутатор языковых ресурсов:* помогает найти доступные инструменты (список) для обработки веб-приложения или анализа данных пользователя;

– *виртуальные коллекции* (из разных архивов): последовательные наборы ссылок на цифровые объекты (например, размеченный текст, видео);

– *реестр:* создание и публикация отдельных виртуальных коллекций, а также обеспечение постоянных идентификаторов и федеративного входа в систему;

– *ресурсные семьи:* обзоры доступных ЛИР;

– *инвентаризация ЛИР:* удобный инструмент для каталогизации, причем гарантировано долгосрочное архивирование, а метаданные общедоступны;

– *для исследователей и студентов в области цифровой гуманитаристики:* онлайновая коллекция учебных материалов, тематических исследований и контактов с экспертами из всей сети;

– *реестр курсов по цифровой гуманитаристике:* перечень курсов, предлагаемых европейскими университетами, и сведения о дисциплинах, местоположении, кредитах ECTS[1] или присуждаемых академических степенях

– *обмен знаниями:* набор общих согласованных правил, мер и соглашений, которые должны обеспечивать бесперебойное взаимодействие между пользователями инфраструктуры, операторами и компонентами, включая стандарты, условия доступа, лицензии, контроль качества и т.д.

CLARIN также ежегодно проводит конференции, семинары и другие профессиональные встречи.

Предложения по инфраструктуре лингвистических информационных ресурсов для России

При значительных масштабах работ по созданию ЛИР, координация в этой сфере деятельности в России развита совершенно недостаточно.

В частности, одним из эффективных инструментов по обеспечению совместимости и повторного использования ЛИР является стандартизация. В профильном ТК 055 разработаны и утверждены в качестве национальных несколько стандартов ISO по управлению ЛИР. Однако выбор этих стандартов производит впечатление случайного, а качество переводов остается низким. Глав-

[1] Англ. European Credit Transfer and Accumulation System (Европейская система перевода и накопления баллов) – общеевропейская система учета учебной работы студентов при освоении образовательной программы или курса.

ный же недостаток деятельности ТК 055 заключается в том, что разработанные им стандарты вообще не применяются при создании ЛИР. В составе этого ТК практически нет разработчиков ЛИР, он полностью оторван от практической деятельности и не оказывает на нее влияния.

При этом среди российских ЛИР множество дублирующих, а их повторное использование минимально.

Для того чтобы направить вектор развития теории и практики создания ЛИР в нужную сторону, представляются необходимыми следующие шаги.

Прежде всего, нужна стратегия развития ЛИР, рассчитанная на руководителей учреждений, которые формируют различные научные и образовательные программы. В ней следует определить, какие ЛИР и сервисы имеет смысл централизовать, а какие должны формироваться и поддерживаться на местах. Очевидно, что централизация может быть реализована на различных уровнях, например только на уровне метаданных. Централизованные сервисы также желательно распределить по разным учреждениям и городам (как это сделано, например, в CLARIN).

Для тех ЛИР и сервисов, централизация которых представляется предпочтительной, необходимо определить, имеет ли смысл делать это на национальном уровне, или разумней присоединиться к мировой или европейской структуре. Например, созданное облако LLOD является вполне удовлетворительным инструментом, и создавать ему альтернативу не имеет смысла.

Кроме того, многие европейские лингвистические структуры формируют специальные зоны в Википедии, где размещаются сведения, которые данное сообщество считает правильными. Думаю, что в русской Википедии можно то же самое сделать для русскоязычных лингвистических терминов.

Вообще Википедия – прекрасный пример коллаборации. Представляется убедительным, что быстрое и качественное развитие ЛИР может быть организовано только при помощи коллабораций. Однако коллаборация эффективна, когда она основана на общепринятой системе мотивации участия в ней отдельных ученых и научных коллективов / учреждений, а также их соответствующего вознаграждения (причем далеко не всегда финансового).

Конечно, в настоящих условиях, когда фактически единственным инструментом оценки качества и эффективности научной деятельности в России стал «комплексный балл публикационной активности», такой подход выглядит утопией. Напомню, однако, что все современные декларации по развитию науки и ее инфосферы, начиная от декларации DORA [Declaration on …, 2021] и вплоть до последнего проекта рекомендаций ЮНЕСКО по открытой науке [Preliminary report …, 2020], единодушно призывают изменить систему оценки научной деятельности. При этом особое внимание обращается на учет научных результатов в форме открытых научных данных, ориентированных на обмен и повторное использование. Очевидно, что к области ЛИР это относится в

полной мере. Вероятно, лучшей современной формой для реализации ЛИР как открытых научных данных было бы размещение их в облаке LLOD.

Конечно, в России нужен централизованный архив ЛИР. Зарубежный опыт создания таких архивов весьма значителен, достаточно примера архивов, входящих в OLAC [Home, 2011].

Следует также пересмотреть отношение к стандартизации ЛИР. С одной стороны, стандарты должны соответствовать реальным потребностям отрасли (сейчас это совершенно не так). С другой стороны, необходимо потребовать от разработчиков ЛИР реального соблюдения этих стандартов, что должно быть зафиксировано в проектах, заявках на грант, экспертных заключениях, в общем, во всей документации, связанной с разработкой ЛИР.

В концепциях по цифровизации науки, которые размещены на сайте Минобрнауки [Совет по цифровому развитию и ИТ, 2021], в качестве отдельного направления выделено совершенствование цифровой инфраструктуры**.** Очевидно, что в этом контексте развитию отечественных ЛИР и сервисов должно уделяться повышенное внимание.

Список литературы

1. Ананьева М.И., Кобозева М.В. Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // DIALOGUE-2016. 22 nd INTERNATIONAL CONFERENCE on Computational Linguistics and Intellectual Technologies. – 2016. – 06. – URL: https://www.researchgate.net/publication/311618194_Razrabotka_korpusa_tekstov_na_russkom_azyke_s_razmetkoj_na_osnove_teorii_ritoriceskih_struktur/link/5851582908ae0c0f321a6265/download (дата обращения 01.03.2021).
2. Антопольский А.Б. Лингвистическое обеспечение электронных библиотек. – Москва : Информрегистр, 2003. – 302 с.
3. База речевых фрагментов русского языка ISABASE / Богданов Д.С., Кривова О.Ф., Подрабинович А.Я., Фарсобина В.В. // Интеллектуальные технологии ввода и обработки информации. – Москва : Эдиториал УРСС, 1998. – С. 74–85.
4. Главная // Национальный корпус калмыцкого языка. – 2012. – URL: http://kalmcorpora.ru (дата обращения 03.03.2021).
5. Главная страница // Лаборатория общей и компьютерной лексикологии и лексикографии. – 2021. – URL http://www.philol.msu.ru/~lex/main.htm (дата обращения 05.03.2021).
6. Группа компаний ЦРТ // ЦРТ. – 2021. – URL https://www.speechpro.ru/about/ (дата обращения 03.03.2021).
7. Другие корпуса // Национальный корпус русского языка. – 2021. – URL : https://ruscorpora.ru/new/corpora-other.html (дата обращения 02.03.2021).
8. Заглавная страница // NLPub. – 2020. – 18.10. – URL: https://nlpub.ru/ (дата обращения 03.03.2021).
9. Захаров В.П. Электронный обменный формат для словарей проекта Text Encoding Initiative : учебное пособие. – Санкт-Петербург : СПбГУ. РИО. Филологический факультет, 2013. – 80 с.
10. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск : ИГЛУ, 2011. – 161 с.
11. Инновационная продукция // СПИИ РАН. – 2021. – URL: http://www.spiiras.nw.ru/ru/scientific-activity/unique-equipment.html (дата обращения 01.03.2021).
12. Коллекция национального электронного звукового депозитария пополняется результатами этнографической экспедиции в Ямальский и Приуральский районы // Салехард. BezFormata. – 2019. – 11.01. – URL: https://salehard.bezformata.com/listnews/etnograficheskoj-ekspeditcii-v-yamalskij/12575916/ (дата обращения 09.06.2021).
13. Компьютерная лексикография // Википедия. – 2021. – URL: https://ru.wikipedia.org/wiki/Компьютерная_лексикография (дата обращения 01.03.2021).
14. Корпус русской устной речи // Проект «Корпус русской устной речи». – 2021. – URL: http://russpeech.spbu.ru/project.htm (дата обращения 06.03.2021).
15. Корпусы ИЭА РАН // ИЭА РАН. – 2019. – URL: https://corpora.iea.ras.ru/corpora/ (дата обращения 10.03.2021).
16. Корпуса языков России // Лингвистические корпуса и сервисы. – 2021. – URL: http://web-corpora.net/ (дата обращения 03.04.2021).
17. Лаборатории информационных исследований // Сайт лаборатории информационных исследований. – 2021. – URL: http://www.labinform.ru/ (дата обращения 09.03.2021).

18. Лаборатория № 15 «Компьютерная лингвистика» // ИППИ РАН. – 2021. – URL http://iitp.ru/ru/researchlabs/245.htm (дата обращения 06.03.2021).
19. Лаборатория № 20 «Цифровая документация русского языка» // ИППИ РАН. – 2021. – URL: http://iitp.ru/ru/researchlabs/digital_documentation (дата обращения 05.04.2021).
20. Лукашевич Н.В. Тезаурус в системах информационного поиска. – Москва : Изд-во МГУ, 2011. – 512 с.
21. Национальный электронный звуковой депозитарий // Институт русской литературы (Пушкинский дом) РАН. Научная деятельность. – 2009. – URL: http://old.pushkinskijdom.ru/Default.aspx?SearchID=20246&tabid=9146 (дата обращения 09.06.2021).
22. Научно-исследовательские проекты кафедры математической лингвистики // Лингвистические информационные технологии : обзорная информация. – 2021. – URL: http://mathling.phil.spbu.ru/node/9 (дата обращения 01.04.2021).
23. Национальный корпус русского языка: 2006–2009. Новые результаты и перспективы. – Санкт-Петербург : Нестор-История, 2009. – 502 с. – URL: https://ruscorpora.ru/new/sbornik2008/00.pdf (дата обращения 03.03.2021).
24. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика. – Москва : ЛЕНАНД, 2017. – 320 с.
25. Новости // Компьютерная лингвистика. Портал знаний. – 2021. – URL: https://uniserv.iis.nsk.su/cl/ (дата обращения 02.03.2021).
26. О нас // Integrum World Wide. – 2008. – URL: http://www.integrumworld.com/rus/about.html (дата обращения 01.02.2021).
27. О проекте // Навигатор информационных ресурсов по языкознанию. – 2021. – URL: http://niryaz2.alexo.beget.tech/ (дата обращения 04.01.2021).
28. О Техническом комитете по стандартизации ТК 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле». Федеральное агентство по техническому регулированию и метрологии. Приказ от 3 декабря 2010 года № 4850 // Электронный фонд правовых и нормативно-технических документов. – 2016. – 14.03. – URL: https://docs.cntd.ru/document/902260132 (дата обращения 03.04.2021).
29. Отдел корпусной лингвистики и лингвистической поэтики // Институт русского языка им. В.В. Виноградова. РАН. – 2021. – URL: http://www.ruslang.ru/node/79 (дата обращения 03.01.2021).
30. Поиск библиотечных материалов // Президентская библиотека им. Б.Н. Ельцина. – 2021. – URL: https://www.prlib.ru/search?f%5B0%5D=field_book_publisher%3A131892 (дата обращения 01.01.2021).
31. Попова Л.В. Типологии и классификации словарей // Вестник ЧелГУ. – 2012. – № 20 (274). – С. 106–113.
32. Прикладная и компьютерная лингвистика. – Санкт-Петербург : ЛЕНАНД, 2017. – 320 с.
33. Прикладная лингвистика / Фонд знаний «Ломоносов». – 2021. – URL: http://lomonosov-fund.ru/enc/ru/encyclopedia:01206:article (дата обращения 03.02.2021).
34. Продукты // PROMT. – 2021. – URL: https://www.promt.ru/. (дата обращения 07.04.2021).
35. Проекты // НИУ «ВШЭ». Лингвистическая лаборатория по корпусным технологиям. – 2021. – URL: https://cfi.hse.ru/corpora/projects (дата обращения 04.03.2021).
36. Проекты Школы лингвистики / НИУ ВШЭ. – 2021. – URL: https://linghub.ru/ (дата обращения 01.04.2021).
37. Российская терминология (терминологические словари) / ФГУП «СТАНДАРТИНФОРМ». – 2021. – URL http://nd.gostinfo.ru/catalog/databank.aspx (дата обращения 02.04.2021).
38. Русский Викисловарь // Викисловарь. – 2021. – URL: https://ru.wiktionary.org/wiki/Викисловарь:Заглавная_страница (дата обращения 02.02.2021).
39. Совет по цифровому развитию и ИТ // Сайт Минобрнауки России. – 2021. – URL: https://www.minobrnauki.gov.ru/colleges_councils/kollegialnye-organy/digitalcouncil/ (дата обращения 02.03.2021).
40. Технический комитет / Федеральное агентство по техническому регулированию и метрологии (Росстандарт). – 2021. – 22.01. – URL: http://webportalsrv.gost.ru/portal/TKSUGGEST/TK2006.nsf/84eb0d5919ea20bac325653100289c4a/35f7dc4669e72340c325718f0040138d?OpenDocument (дата обращения 01.04.2021).
41. Технологии // Яндекс. – 2021. – URL: https://yandex.ru/company/technologies (дата обращения 03.04.2021).
42. Усталов Д.А. Семантические сети и обработка естественного языка // Открытые системы. – 2017. – № 2. – С. 46–47.
43. Устные корпуса / НИУ «ВШЭ». Международная лаборатория языковой конвергенции. – 2021. – URL: https://ilcl.hse.ru/corpora (дата обращения 01.02.2021).
44. Фундаментальные и прикладные разработки / НИИ «Прикладная семиотика» АН РТ. – 2021. – URL: http://www.antat.ru/ru/ips/science/rnd/ (дата обращения 10.02.2021).
45. A guide to the best linguistic resources on the web // Linguistics, natural language, and computational linguistics meta-index. – 2014. – 15.04. – URL: https://nlp.stanford.edu/links/linguistics.html (дата обращения 03.02.2021).
46. Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange // ISO 12200. – 1999. – 10.01. – URL: https://infostore.saiglobal.com/preview/iso/updates2013/wk31/iso_12200-1999.pdf?sku=224662 (дата обращения 13.03.2021).
47. Declaration on research assessment // DORA. – 2021. – URL: https://sfdora.org/ (дата обращения 10.03.2021).
48. Home // Linguistic Linked Open Data. – 2018. – URL: http://linguistic-lod.org/ (дата обращения 03.02.2021).
49. Home // Open language archives community. – 2011. – URL: http://olac.ldc.upenn.edu/ (дата обращения 03.04.2021).
50. Home Page // SKOS Simple Knowledge Organization System. – 2012. – URL: https://www.w3.org/2004/02/skos/ (дата обращения 03.03.2021).
51. Language resource // QW. – 2021. – URL: https://ru.qaz.wiki/wiki/Language_resource (дата обращения 01.02.2021).

52. Lingvo Live // ABBYY. – 2021. – URL: https://www.lingvolive.com/ru-ru (дата обращения 05.04.2021).
53. Main page // The Language archive. – 2021. – URL: https://archive.mpi.nl/tla/ (дата обращения 01.03.2021).
54. Management of terminology resources (ISO/TC 37/SC 3) // ISO. – 1985. – URL: https://www.iso.org/committee/48136.html (дата обращения 11.03.2021).
55. Preliminary report on thé flrst draft of thé recommendation on Open science // UNESKO. Цифровая библиотека. – 2020. – URL: https://unesdoc.unesco.org/ark:/48223/pf0000374409 (дата обращения 09.04.2021).
56. Recent Postings // LINGUIST List. – 2021. – URL: https://linguistlist.org/ (дата обращения 04.02.2021).
57. Resource families // Clarin. – 2021. – URL: https://www.clarin.eu/resource-families (дата обращения 01.04.2021).
58. RussNet // Сайт проекта RussNet. – 2005. – 14.06. – URL: http://project.phil.spbu.ru/RussNet/index_ru.shtml (дата обращения 01.02.2021).
59. The research infrastructure for language as social and cultural data // Clarin. – 2021. – URL: https://www.clarin.eu/ (дата обращения 10.04.2021).
60. W3 C Recommendation // Simple knowledge organization system reference. – 2009. – 18.08. – URL: https://www.w3.org/TR/skos-reference/ (дата обращения 02.03.2021).
61. Yet Another RussNet // Сайт проекта Yet Another RussNet. – 2018. – URL: https://russianword.net/ (дата обращения 01.03.2021).

LANGUAGE RESOURCES AND TECHNOLOGIES IN RUSSIA: STATE AND PROSPECTS (Review)

Antopolsky Alexander

DrS (Tech. Sci.), Chief Researcher, Institute of Scientific Information for Social Sciences, Russian Academy of Sciences (INION RAN), Moscow, Russia

Abstract. *The author determines the concept of linguistic information resources, gives an overview of their classifications. Describes the most significant Russian catalogs of linguistic information resources and the country's leading organizations in the field of computational linguistics. Discusses the priority tasks of creating a Russian infrastructure for linguistic information resources.*

Keywords: *computational linguistics; artificial Intelligence; linguistic resources; information infrastructure.*

For citation: Antopolsky A.B. Language resources and technology in Russia: state and prospects (Review) // Social Novelties and Social Sciences. – Moscow : INION RAN, 2021. – N 2. – Pp. 114–131.

URL: https://sns-journal.ru/ru/archive/ DOI: 10.31249/snsn/2021.02.08