# Лингвистические ресурсы: европейский опыт и уроки для России

А.Б. Антопольский

ИНИОН РАН

ale5695@yandex.ru

## Аннотация

Обсуждается проблема создания инфраструктуры для поддержки лингвистических информационных ресурсов (ЛИР). Описывается опыт CLARIN - Общеевропейской исследовательской инфраструктуры для языковых ресурсов и технологий. Описываются существующие сервисы CLARIN и стратегия их развития на ближайшие 3 года. Формулируются предложения по программе развития и поддержки ЛИР в России, включая информационную поддержку, финансирование, стандартизацию, архивацию и другие необходимые сервисы. Лучшим решением выглядит представление ЛИР в виде связанных открытых данных, но проблемой является необходимость мотивации разработчиков для участия в облаке LLOD.

**Ключевые слова**: Лингвистические информационные ресурсы, инфраструктура, CLARIN. программа развития, сервисы, LLOD

**Библиографическая ссылка:** Антопольский А.Б. Лингвистические ресурсы: европейский опыт и уроки для России // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24 – 26 июня 2021 г. Сборник научных статей). — СПб.: Университет ИТМО, 2021. С. 09-16. DOI: 10.17586/2541-9781-2021-5-09-16

## Введение

Лингвистические технологии в последние десятилетия стали одной из самых влиятельных технологических инноваций эпохи науки о данных. Будь то исследователи, использующие ресурсы и инструменты для решения новых научных проблем, правительства и промышленность, применяющие интеллектуальный анализа текста для поиска ценных паттернов в больших данных, СМИ, отличающие достоверную информацию от дезинформации, преподаватели и студенты, изучающие языки, или граждане, использующие автоматическое распознавание речи и машинный перевод, - языковые технологии вездесущи. Центральным элементом лингвистических технологий являются лингвистические информационные ресурсы (ЛИР). Количество ЛИР в мире быстро растет. Крупнейший на сегодня в мире Языковой архив Института Макса Планка [1] содержит около 150 тыс. ЛИР, а суммарное количество ЛИР, отраженных в архивах, вошедших в состав Открытого консорциума лингвистических архивов (OLAC) [2], достигает 400 тыс.

Неудивительно, что потребовалось создание специальной инфраструктуры, обеспечивающее эффективное использование ЛИР, межстрановое и междисциплинарное сотрудничество при их создании и распространении.

В этой связи обратимся к европейскому опыту. В Евросоюзе создана организация CLARIN [3] - общеевропейская исследовательская инфраструктура для языковых ресурсов и технологий.

## 1. Общая характеристика CLARIN

CLARIN – это распределенная инфраструктура данных, которая предлагает доступ к ЛИР, технологиям и знаниям, обеспечивает сотрудничество между академическими кругами, промышленностью, политиками, культурными и образовательными учреждениями и широкой общественностью. CLARIN основана на видении, что все цифровые ЛИР со всей Европы и за ее пределами должны быть доступны через единый онлайн-вход. CLARIN обеспечивает быстрый и устойчивый доступ к ЛИР (в письменной, устной или мультимодальной форме).

CLARIN также предлагает инструменты для поиска, исследования, использования, аннотирования, анализа или интеграции ЛИР. Эти функции реализуются через сетевую федерацию хранилищ ЛИР, сервисных центров и центров знаний с единым входом для всех членов сообщества. Инструменты и данные из разных центров являются функционально совместимыми, поэтому сбор данных можно объединять, а инструменты из разных источников можно интегрировать для выполнения сложных операций. Инфраструктура CLARIN полностью функционирует во многих странах и продолжает создаваться в странах, которые присоединились совсем недавно.

## 2. Сервисы CLARIN

*Депозитные услуги*. Одним из основных сервисов CLARIN является обеспечение архивирования ЛИР и предоставление их сообществу надежным способом. Чтобы помочь исследователям устойчиво хранить свои ЛИР (например, корпуса, лексиконы, аудио- и видеозаписи, грамматики и т. д.), многие центры CLARIN предлагают услугу депозита. Это имеет ряд преимуществ:

- Долгосрочное архивирование: гарантия хранения может быть предоставлена на длительный период (в некоторых случаях до 50 лет);
- ЛИР удобно цитировать через идентификатор;
- ЛИР и их метаданные будут интегрированы в инфраструктуру, что позволит эффективно осуществлять их поиск.
- Защищенные ЛИР будут доступны для легальных пользователей.
- Интеграция ЛИР в инфраструктуре CLARIN упрощает их обработку с помощью различных лингвистических инструментов.

*Виртуальная языковая обсерватория*. Цель - предоставить простой в использовании интерфейс, обеспечивающий единый процесс поиска разнообразных ЛИР. Развитый синтаксис запроса позволяет выполнять целенаправленный поиск. Виртуальная языковая обсерватория позволяет на основе результатов поиска создавать виртуальные коллекции.

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

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

*Виртуальные коллекции*. Представляют собой последовательные наборы ссылок на цифровые объекты (например, размеченный текст, видео). Ссылки могут происходить из разных архивов, отсюда и термин виртуальный. Виртуальная коллекция удобна для ручного доступа (с помощью веб-браузера), а также для автоматической обработки (например, веб-сервисом).

*Реестр CLARIN* предоставляет сервис, где ученые могут создавать и публиковать свои виртуальные коллекции. Он тесно интегрирован с инфраструктурой и обеспечивает постоянные идентификаторы и федеративный вход в систему.

*Инвентаризация ЛИР.* Предоставляется инструмент, удобный для инвентаризации (каталогизации) ЛИР. Данный каталог отличается от других депозитных услуг тем, что нет необходимости загружать данные или метаданные (достаточно ссылки на веб-сайт и описание) и его можно использовать немедленно, без предварительного обращения в хостинг-центр. Каталог полностью интегрирован в инфраструктуру CLARIN. Его метаданные (собранные с 2008 года) тщательно курируются и общедоступны. Долгосрочное архивирование гарантировано.

*Федеративный поиск*. Чтобы обеспечить исследователям возможность расширенного поиска с использованием конкретных моделей коллекции данных, CLARIN предлагает поисковую систему (пока прототип), данных, которые доступны в центрах хранения. Сами данные остаются у владельца, поэтому поиск называется федеративным. Поисковая система суммирует и отображает то, что доступно. Чтобы выполнить более сложный запрос, нужно перейти к специализированному интерфейсу поиска в центре-владельце ЛИР. Федеративный поиск отличается от поиска метаданных, например, в Виртуальной обсерватории, где все метаданные сначала собираются (копируются на один сервер), а затем индексируются централизованно. Он применяется по нескольким причинам:

- Правовые проблемы делают невозможным копирование некоторых ЛИР;
- Размер многих ЛИР делает децентрализованную индексацию наиболее приемлемым вариантом.
- Большинство ЛИР индексируются в зависимости от коллекции, что затрудняет использование или разработку единой поисковой системы.

Федеративный поиск контента является масштабируемым, он обходится дешевле, чем локальный поиск, хотя некоторые функции отсутствуют, например, ранжирование. Поэтому федеративный поиск будет особенно полезен в качестве первого шага для поиска интересных ЛИР.

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

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

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

*Обмен знаниями.* В CLARIN создается инфраструктура обмена знаниями, включающая как технические и инструментальные средства, так и обученный персонал. Внедряется набор общих согласованных организационных 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. Организационную и содержательную составляющую инфраструктуры обмена знаниями составляют Центры знаний CLARIN, функции и специализация которых представлена на странице [4].

*Мероприятия*. CLARIN ежегодно проводит 12 конференций, семинаров и других профессиональных встреч.

## 3. Стратегия CLARIN

На 2021–2023 гг. CLARIN предлагает план развития, включающий 4 приоритетные области, причем по каждой области предлагаются конкретные направления активности [5].

*Инфраструктура знаний.* Углубление мониторинга ресурсов, технологий и опыта структур, входящих в CLARIN; увеличение способов финансирования обмена и распространения лучших практик.

*Техническая инфраструктура*. Поддержка и развитие системы идентификаторов; контроль качества метаданных, как автоматический, так и интеллектуальный.

*Организационное развитие*. Поддержка сотрудничества на уровне координации, технологии и финансовой, развитие инструментов обмена информацией

*Устойчивость* Расширение состава, в том числе за счет неевропейских стран, финансовая диверсификация, новые формы сотрудничества с индустрией.

Можно констатировать, что европейский опыт демонстрирует разные формы поддержки языковой индустрии, созданию и распространению ЛИР. Важно отметить, что CLARIN последовательно поддерживает принципы открытой науки, что весьма способствует эффективности созданных ЛИР и технологий.

## 4. Проблемы инфраструктуры ЛИР В России

Обратимся к России. В стране создано множество ЛИР по всем направлениям как прикладной, так и теоретической лингвистики и языковой индустрии, а также образовательных ЛИР. Только в учреждениях РАН выявлено свыше тысячи ЛИР различных типов и назначения, которые внесены в Навигатор информационных ресурсов по языкознанию [6].

Напомним, что в этой системе под ЛИР понимались как обычные информационные ресурсы, относящиеся к лингвистике лишь по тематике (библиотечные и архивные фонды, периодика, материалы конференций, энциклопедические данные, диссертации, отчеты, каталоги, сайты и др.), так и специальные ЛИР, к которым отнесены:

1. Корпуса текстов и звучащей речи.
2. Словарные БД и электронные лексиконы.
3. Памятники письменности (кодикологические БД).
4. Лингвистические процессоры.
5. Грамматические ресурсы, описания языков.
6. Типологические БД, реестры языков.
7. Лингвистические атласы, ГИС.
8. Этно- и социолингвистические БД.
9. Комплексные лингвистические АИС.
10. Информационные языки.

Нужно отметить, что в этой классификации могут отсутствовать ЛИР, которых нет в учреждениях РАН, например, учебные ЛИР. Если же оценивать общее количество российских ЛИР, создаваемых в вузах, школах, промышленности, на любительских сайтах, то их число, вероятно, превысит 10 тыс. Несмотря на широкий фронт работ по созданию ЛИР, координация в этой области развита совершенно недостаточно.

Приведем два примера. Действовавшая в 2012-2014 гг. Программа Президиума РАН «Корпусная лингвистика» весьма эффективно способствовала реализации многих проектов: активно развивается НКРЯ, создано множество других корпусов, как для русского, так и для других языков народов России, разработан сайт «Лингвистические корпуса и сервисы» [7], объединяющий корпусных лингвистов. Таким образом, очевиден позитивный эффект от этой программы.

Теперь противоположный пример. В российском Техническом комитете по стандартизации ТК 55 "Терминология, элементы данных и документация в бизнес-процессах и электронной торговле" разработаны, т.е. переведены на русский язык.

несколько стандартов, по управлению ЛИР из числа разработанных в ИСО и утверждены в качестве национальных стандартов. Выбор стандартов для перевода, производит впечатление случайного, а качество переводов при этом чрезвычайно низкое: кажется, что результаты автоматического перевода вообще не редактировались.

Главный же недостаток деятельности российского Технического комитета 55 заключается в том, что разработанные им стандарты вообще не применяются при разработке российских ЛИР. Это неудивительно, ведь в составе этого ТК практически нет разработчиков ЛИР. Исключение только одно – возглавляет ТК 55 институт Стандартинформ, который поддерживает известный российский банк терминологических данных Ростерм. Однако этот банк данных не замечен в активном сотрудничестве с другими разработчиками отечественных ЛИР, а также в разработке открытого доступа и интеграции терминологических данных.

Очевидно, что деятельность этого ТК полностью оторвана от современной российской индустрии ЛИР.

Как мне представляется, деятельность этого комитета не оказывает вообще никакого влияния на российскую индустрию в области ЛИР, хотя в мире и, особенно в Европе, соответствующие стандарты активно применяются и оказывают значительное влияние на совместимость и повторное использование ЛИР.

Анализ показывает, что в числе ЛИР множество дублирующих, неиспользуемых, вообще некачественных.

## Заключение

Какие шаги следует предпринять, чтобы направить вектор развития теории и практики создания ЛИР в нужную сторону?

Мои предложения носят заведомо идеалистический характер, но это сделано сознательно; мне кажется, что нужно представлять себе идеальную картину будущего, чтобы появился ориентир, с которым можно сверять реальные практические шаги. Некоторые из излагаемых ниже предложений были представлены в работе [8].

Прежде всего, нужна программа развития ЛИР, но только не финансовая, а именно ориентирующая, рассчитанная на руководителей учреждений, формирующих различные научные и образовательные программы.

В этой программе желательно определить, какие ЛИР и сервисы имеет смысл централизовать, а какие должны формироваться и поддерживаться на местах. При этом очевидно, что централизация может быть реализована на различных уровнях, например, только на уровне метаданных. Централизованные сервисы также желательно распределить по разным учреждениям и городам, как это сделано в CLARIN.

Для тех ЛИР и сервисов, централизация которых кажется предпочтительной, необходимо определить, имеет ли смысл делать это на национальном уровне, или разумней присоединиться к мировому или европейскому сервису.

Например, если речь идет о лингвистических связанных открытых данных, очевидно, что уже созданное облако LLOD (Linguistic Linked Open Data) [9] является необходимым и достаточным инструментом и создавать ему альтернативу нет никакой необходимости.

Или другой пример. В Европе многие лингвистические структуры, (например, Ontolex), которые создают концептуальные и/или энциклопедические данные, формируют специальные зоны в Википедии, где размещаются сведения, которые данное сообщество считает правильными. Думаю, что для русскоязычных лингвистических терминов можно то же самое сделать с русской Википедией.

Вообще Википедия - прекрасный пример коллаборации и кажется очень убедительным, что быстрое и качественное развитие ЛИР должно быть организовано только при помощи коллабораций. Однако коллаборация при создании и поддержки централизованных ресурсов и сервисов в российских условиях эффективна, когда ресурс делается не кучкой

любителей-энтузиастов, а с подключением ведущих академических учреждений и университетов, как это делалось, например, при создании НКРЯ. А это, в свою очередь, требует разработки и реализации системы мотивации и вознаграждения (причем, далеко не всегда финансового) участия в коллаборации отдельных ученых и научных учреждений.

Конечно, в наших условиях, когда фактически единственным инструментом оценки качества и эффективности научной деятельности стал пресловутый Комплексный балл публикационной активности, такой подход выглядит утопией. Напомню, однако, что все современные декларации по развитию науки и ее инфосферы, начиная от декларации DORA [10] и вплоть до последнего проекта рекомендаций ЮНЕСКО по открытой науке [11], единодушно призывают изменить систему оценки научной деятельности. При этом особое внимание нужно обратить на учет научных результатов в форме открытых научных данных, ориентированных на обмен и повторное использование. Очевидно, что к области ЛИР это относится в полной мере. Вероятно, лучшей современной формой для реализации ЛИР как открытых научных данных, было бы размещение их в облаке LLOD.

Отдельно нужно сказать о грантовой поддержке деятельности по созданию ЛИР и сопутствующих сервисов. Это особенно актуально в связи с реорганизацией наших научных фондов.

Считаю необходимым поставить вопрос о специальной форме грантов, направленных на поддержку открытых ЛИР, которые могут использоваться как наукой, так и промышленностью. Эти ЛИР требуют финансовой поддержки не только на этапе создания, но и для постоянного пополнения и развития. К сожалению, большое число ЛИР, созданных за счет грантов, далее не поддерживаются, что во многих случаях означает их гибель. Соответственно нужно менять и правила экспертизы заявок на получение грантов для создания или модернизацию ЛИР.

Конечно, нужен централизованный архив ЛИР, который тоже должен получать отдельное и постоянное финансирование, обеспечивая доступ к ранее созданным ЛИР и прием вновь создаваемых. Зарубежный опыт таких архивов весьма велик, достаточно посмотреть на архивы, входящие в Консорциум OLAC [12].

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

И, конечно, хотелось бы, что бы в России появилось учреждение, которое бы профессионально поддерживало инфраструктуру российских ЛИР, как это в Европе делает CLARIN.

Напомним, что в материалах по цифровизации науки, которые размещены на сайте Минобрнауки [13], есть Концепция цифровой автоматизированной системы предоставления сервисов научной инфраструктуры коллективного пользования (АС УСНИКП) и Концепция создания Единой цифровой платформы научного и научно-технического взаимодействия, организации и проведения совместных исследований в удаленном доступе, в том числе с участием зарубежных ученых (ЦПСИ).

Очевидно, что при реализации этих концепций российские ЛИР и соответствующие сервисы должны занять достойное место. При этом, как представляется автору, хорошим инструментом для решения многих проблем является конвертирование существующих российских ЛИР в формат связанных открытых данных и дальнейшее их развитие в этом облаке в тесной коллаборации с международным сообществом.

## Литература

[1] The Language Archive URL: https://archive.mpi.nl/tla/ (дата обращения: 22.03.2021).

[2] Open Language Archives Community http://olac.ldc.upenn.edu/ (дата обращения: 22.03.2021).
[3] Common European Research Infrastructure for Language Resources and Technology. URL: https://www.clarin.eu (дата обращения: 22.03.2021).
[4] Knowledge Infrastructure URL: https://www.clarin.eu/content/knowledge-infrastructure (дата обращения: 22.03.2021).
[5] CLARIN strategy at a glance URL: https://www.clarin.eu/content/vision-and-strategy (дата обращения: 22.03.2021).
[6] Навигатор информационных ресурсов по языкознанию URL: http://niryaz2.alexo.beget.tech/ (дата обращения: 22.03.2021).
[7] Лингвистические корпуса и сервисы URL: http://web-corpora.net/ (дата обращения: 22.03.2021).
[8] Антопольский А.Б. О создании центра лингвистических ресурсов РАН // Известия Российской академии наук. Серия литературы и языка. 2019. Т. 78, № 4. С. 5-12.
[9] Linguistic Linked Open Data URL: https://linguistic-lod.org/llod-cloud (дата обращения: 22.03.2021).
[10] Declaration on Research Assessment (DORA) URL: https://sfdora.org/ (дата обращения: 22.03.2021).
[11] Preliminary Report on thé flrst draft of thé Recommendation on Open Science URL: https://unesdoc.unesco.org/ark:/48223/pf0000374409 (дата обращения: 22.03.2021).
[12] Open Language Archives Community http://olac.ldc.upenn.edu/ (дата обращения: 22.03.2021).
[13] Совет по цифровому развитию и ИТ URL: https://minobrnauki.gov.ru/colleges_councils/kollegialnye-organy/digitalcouncil/ (дата обращения: 22.03.2021).

## Linguistic Resources: European Experience and Lessons for Russia

A.B. Antopolskii

INION RAS

The problem of creating an infrastructure to support linguistic information resources (LIR) is discussed. The experience of CLARIN - Common European Research Infrastructure for Language Resources and Technology, is described. It describes the existing CLARIN services and their development strategy for the next 3 years. Proposals are formulated for a program for the development and support of LIR in Russia, including information support, financing, standardization, archiving, and other necessary services. The best solution seems to be to present the LIR as linked open data, but the problem is the need to motivate developers to participate in the LLOD cloud.

**Keywords:** Linguistic information resources, infrastructure, CLARIN. development program, services, LLOD

**Reference for citation:** Antopolskii A.B. Linguistic resources: European experience and lessons for Russia // Computer Linguistics and Computing Ontologies. Vol. 5 (Proceedings of the XXIV International Joint Scientific Conference «Internet and Modern Society», IMS-2021, St. Petersburg, June 24-26, 2021). - St. Petersburg: ITMO University, 2021. P. 09 – 16. DOI: 10.17586/2541-9781-2021-5-09-16

## Reference

[1] The Language Archive URL: https://archive.mpi.nl/tla/ (data obrashcheniya: 22.03.2021).

[2] Open Language Archives Community http://olac.ldc.upenn.edu/ (data obrashcheniya: 22.03.2021).
[3] Common European Research Infrastructure for Language Resources and Technology. URL: https://www.clarin.eu (data obrashcheniya: 22.03.2021).
[4] Knowledge Infrastructure URL: https://www.clarin.eu/content/knowledge-infrastructure (data obrashcheniya: 22.03.2021).
[5] CLARIN strategy at a glance URL: https://www.clarin.eu/content/vision-and-strategy (data obrashcheniya: 22.03.2021).
[6] Navigator informacionnyh resursov po yazykoznaniyu URL: http://niryaz2.alexo.beget.tech/ (data obrashcheniya: 22.03.2021).
[7] Lingvisticheskie korpusa i servisy URL: http://web-corpora.net/ (data obrashcheniya: 22.03.2021).
[8] Antopol'skij A.B. O sozdanii centra lingvisticheskih resursov RAN // Izvestiya Rossijskoj akademii nauk. Seriya literatury i yazyka. 2019. 78 (4). S. 5-12.
[9] Linguistic Linked Open Data URL: https://linguistic-lod.org/llod-cloud (data obrashcheniya: 22.03.2021).
[10] Declaration on Research Assessment (DORA) URL: https://sfdora.org/ (data obrashcheniya: 22.03.2021).
[11] Preliminary Report on thé flrst draft of thé Recommendation on Open Science URL: https://unesdoc.unesco.org/ark:/48223/pf0000374409 (data obrashcheniya: 22.03.2021).
[12] Open Language Archives Community http://olac.ldc.upenn.edu/ (data obrashcheniya: 22.03.2021).
[13] Sovet po cifrovomu razvitiyu i IT URL: https://minobrnauki.gov.ru/colleges_councils/kollegialnye-organy/digitalcouncil/ (data obrashcheniya: 22.03.2021).