

Раздел
Информационные процессы
Section
Information Processes

Международная деятельность
в области языковых технологий и ресурсов

Антопольский А.Б.

д.т.н., профессор, главный научный сотрудник, Институт научной информации по общественным наукам Российской академии наук (ИНИОН РАН), Россия, Москва,
e-mail: ale5695@yandex.ru

Аннотация. Описываются международные организации, действующие в области языковых технологий и ресурсов, а также реализуемые ими проекты. Выделяются следующие области: компьютерная лингвистика в целом, поддержка исчезающих языков, терминология и переводческая деятельность, стандартизация. Выделены организации и проекты, работающие в технологии Семантической сети и связанных открытых данных. Отдельно рассматриваются организации и проекты в области языковых технологий и ресурсов в Европе.

Ключевые слова: международные организации, международные проекты, языковые технологии, языковые ресурсы, компьютерная лингвистика, связанные открытые данные.

International Activities
in the Field of Language Technologies and Resources

Antopolsky A.B.

Doctor of Technical Sciences, Professor, Chief Researcher, Institute for Scientific Information on Social Sciences of the Russian Academy of Sciences (INION RAS), Russia, Moscow,
e-mail: ale5695@yandex.ru

Abstract. Describes the international organizations active in the field of language technologies and resources, as well as the projects they are implementing. The following areas are highlighted: computational linguistics in general, support for endangered languages, terminology and translation activities, standardization. Highlighted organizations and projects working in the technology of the Semantic Web and related open data. Organizations and projects in the field of language technologies and resources in Europe are considered separately.



Key words: international organizations, international projects, language technologies, language resources, computational linguistics, related open data.

DOI: 10.31432/1994-2443-2021-16-2-5-22

Цитирование публикации: Антопольский А.Б. Международная деятельность в области языковых технологий и ресурсов // Информация и инновации. 2021, Т. 16, № 2. с. 5-22. DOI: 10.31432/1994-2443-2021-16-2-5-22

Citation: Antopolsky A. International Activities in the Field of Language Technologies and Resources // Information and Innovations 2021, T. 16, № 2. p. 5-22. DOI: 10.31432/1994-2443-2021-16-2-5-22

Введение

В быстро развивающемся цифровом мире важное место занимают языковые технологии, прежде всего вычислительная лингвистика, анализ текста, синтез и анализ речи, машинный перевод, и другие задачи обработки естественного языка (NLP), а также лингвистические информационные ресурсы (ЛИР), создаваемые для этих технологий. Развитие этой деятельности, как и других направлений высоких технологий, происходит в условиях все более открытой науки и развивающейся международной коллаборации ученых. В данной статье будет предложен обзор международных организаций и реализуемых ими проектов в области NLP и ЛИР в целом, а также в некоторых специальных сферах, таких как защита исчезающих языков, терминология и поддержка переводческой деятельности, а также стандартизация в сфере языковых технологий. Отдельно выделено наиболее перспективное направление международной интеграции и коллаборации в области ЛИР — связанные лингвистические открытые данные. В силу значительно большего объема этой деятельности в Европе, развитой организации и управления европейский опыт рассматривается отдельно.

Профессиональные ассоциации

Общее взаимодействие между специалистами разных стран осуществляется через профессиональные ассоциации. Они проводят конференции, издают журналы, создают порталы и сайты, организуют учебу и семинары, находят другие формы взаимодействия.

Если брать область компьютерной лингвистики в целом, то наиболее заметными международными профессиональными ассоциациями являются следующие:

- Ассоциация компьютерной лингвистики (ACL) [1];
- Международная ассоциация прикладной лингвистики (AILA) [2];
- Международный комитет по координации и стандартизации речевых баз данных и методов оценки (COCOSDA) [3];
- Международный комитет компьютерной лингвистики (ICCL — COLING) [4];
- Международная ассоциация количественной лингвистики (IQLA) [5];
- Международная ассоциация речевых коммуникаций (ISCA) [6].

Сферу профессиональных интересов профессиональных ассоциаций легко оценить через перечень рабочих групп, которые они создают. Так, ACL имеет следующие рабочие группы (SIG):

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

Приведем также перечень рабочих групп ISCA:

Тематические SIG:

- Синтез Речи — SynSig;
- Аудиовизуальная речь — AVISA;
- Недостаточно обеспеченные ресурсами языки — СИГУЛ;

- Характеристика говорящего и языка — SpLC;
- Речевая Просодия — Спросиг;
- Дискурс и диалог — Сигдиал;
- Речевые и языковые технологии в образовании — SlaTE;
- Машинное обучение — SIGML;
- Речь и язык в Мультимедиа — SLIM;
- Обработка речи и языка для вспомогательных технологий-SLPAT;
- История наук о речевой коммуникации — SIG-HIST;
- Взаимодействие с компьютером ребенка — SIG-CHILD;
- Надежная обработка речи — SIG-RoSP;
- Безопасность и конфиденциальность в речевой коммуникации — SIG-SPSC;
- Устный перевод — SIG-SLT.

Языковые SIG:

- Обработка китайского разговорного языка — SIG-CSLP;
 - Французская ассоциация устной речи — AFCP;
 - Итальянская ассоциация устной речи — AISV;
 - Иберийские языки — SIG-IL;
 - Обработка речи на индийском языке — SIG-ILSP;
 - Анализ русской речи — SIGRU;
 - Недостаточно обеспеченные ресурсами языки — SIGUL.
- Тематические области COCODA:
- Распознавание речи;
 - Синтез речи;
 - Устные корпуса;
 - Инструменты аннотации корпусов;
 - Локальные языки.

Консорциумы

В отличие от ассоциаций, обычно объединяющих специалистов как физических

лиц, консорциумы объединяют организации, в том числе университетские лаборатории. В области компьютерной лингвистики наиболее известны такие консорциумы:

- LDC — Консорциум лингвистических данных [7];
- OLAC — Сообщество открытых лингвистических архивов [8];
- LINGUIST List — Международное сообщество лингвистов онлайн [9];
- SIL International — международная некоммерческая организация, (бывший Летний Институт Лингвистики — Summer Institute of Linguistics) [10];
- GOLD. Общая онтология лингвистического описания [11].

Приведем краткие описания этих консорциумов и реализуемых ими проектов.

LDC — это открытый консорциум университетов, библиотек, корпораций и государственных исследовательских лабораторий. Первоначально основная роль LDC заключалась в репозитории и точке распространения языковых ресурсов. С того времени и с помощью своих членов LDC превратилась в организацию, которая создает и распространяет широкий спектр языковых ресурсов. LDC также поддерживает спонсируемые исследовательские программы и оценки языковых технологий, предоставляя ресурсы и внося свой организационный вклад. LDC находится в Университете Пенсильвании.

LDC поддерживает образование, исследование и развитие технологий путем создания и обмена ЛИР, такими как данные, инструменты и стандарты.

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

- выработки консенсуса в отношении наилучшей современной практики циф-

рового архивирования языковых ресурсов;

- создания сети взаимодействующих хранилищ и служб для обеспечения доступа к таким ресурсам.

OLAC объединяет 63 лингвистических архивов из всех регионов мира, в которых содержатся до 400 тыс. описаний ЛИР, включающих данные на св. 8 тыс. языков. OLAC поддерживает фундаментальный каталог ЛИР самых разных типов. Имеется сводный анализ метаданных для ЛИР, включая инструменты поддержки и использования ЛИР. Также имеется развитая поисковая машина для каталога, позволяющая получать различную статистику.

OLAC уделяет большое внимание технологиям работы с ЛИР, протоколам и процедурам обмена ЛИР, а также разработке и поддержке системы метаданных. Основу метаданных составляет известный международный стандарт Dublin Core.

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

Mailing Lists [12]. *Архив лингвистических сайтов* — проект по созданию единого постоянного и доступного для поиска архивного сайта для сотен языковых форумов и дискуссий прошлого и настоящего, чтобы информация, которую они содержат, могла

быть свободно доступна любому специалисту в этой дисциплине

LL-MAP [13]. Лингвистические геоданные. Проект, предназначенный для интеграции языковой информации с данными физических и социальных наук с помощью географической информационной системы (ГИС). Онлайн-механизм сбора геоданных позволит ученым создавать карты на основе собственных наблюдений и дополнять информацию о распределении языков, полученную из открытых источников.

MultiTree [14]. Генеалогические деревья. В рамках проекта создается цифровая библиотека научных гипотез о языковых отношениях и подгруппах. Эта информация систематизируется в базе данных. Имеется возможность поиска с помощью веб-интерфейса. Каждая гипотеза представляется графически в виде интерактивного гиперболического отображения генеалогического древа, сопровождаемого информацией обо всех задействованных языках, а также об авторах и библиографических источниках гипотезы.

SIL International — международная религиозная некоммерческая организация, (бывший Летний Институт Лингвистики — Summer Institute of Linguistics). Миссия SIL — служение языковым сообществам в качестве международного защитника. SIL обслуживает языковые сообщества по всему миру, наращивая их потенциал для устойчивого развития языка посредством исследований, перевода, обучения и разработки материалов. SIL участвует в 1660 языковых проектах в 162 странах.

SIL является лидером в области идентификации и документирования мировых языков. Он поддерживает авторитетную базу данных *Ethnologue* (Этнолог; языки мира) [15]. *Ethnologue* объединяет больше данных, чем любой другой ресурс подобного рода — от населения до карт, диалектов,

сведений об использовании языков, тенденции языков и многое другое. Имеется статистика, публикации, справочные данные. В ресурсе возможен поиск по странам, названиям и кодам языков, языковым семьям. Имеются коммерческие сервисы.

GOLD. Проект, начавшийся в 2004 г. вырос в сообщество, цель которого — объединить ученых, заинтересованных в наилучшей практике кодирования лингвистических данных. Проект продвигает передовую практику, развивает интероперабельность данных за счет использования онтологии GOLD, облегчает поиск по разрозненным наборам данных и предоставляет платформу для обмена существующими данными и инструментами. GOLD не просто таксономия лингвистических терминов, она основана на принципах онтологической инженерии: например, обеспечивается богатая аксиоматизация классов и отношений.

На базе сообщества GOLD реализуется ряд проектов, создаются программные инструменты и методики.

Электронная лингвистика [16]. Функция — содействие использованию электронных данных и связанных с ними инструментов в области лингвистики. Пакет e-Linguistics содержит код Python для:

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

Метасхема языка SIL [17]. Это язык семантической интерпретации, который используется для определения значения элементов и атрибутов в схеме разметки XML в терминах понятий, определенных в формальной семантической схеме (такой как схема RDF или онтология OWL).

FIELD. Среда ввода лингвистических дан- ны [18]. FIELD-это веб-инструмент для по- левых лингвистов, позволяющий вставлять свои языковые данные в полностью доступ- ную для поиска онлайн-базу данных.

LEGO. Расширение лексикона с помо- щью онтологии GOLD [19]. Это проект по созданию инструментов и стандартов для облегчения обмена и взаимодействия лек- сических данных. Создается сеть взаимо- действия данных путем сопоставления лек- сических элементов из нескольких лекси- конов и списков слов с понятиями в GOLD . Пользователи могут осуществлять поиск по морфосинтаксической информации, а так- же по определению, языку, орфографии и т.д. LEGO также предлагает набор требо- ваний к данным, которые создатели лекси- конов могут реализовать для того, чтобы присоединиться к сети интероперабельно- сти.

ODIN. Онлайн-овая база данных межли- нейных примечаний [20], которая форми- руется на основе сбора лингвистических данных из Интернета. ODIN содержит дан- ные по различным языкам мира подвер- гнувшиеся лингвистическому анализу по методу IGT [21]. IGT содержит фонетиче- скую транскрипцию, морфосинтаксический анализ, (морфемный глоссарий, грамма- тическую информацию различного рода), а также свободный перевод. В настоящее время ODIN дает ссылки на научные статьи в Интернете, которые содержат примеры IGT. Имеются разные виды поиска. ODIN в настоящее время интегрируется с лингви- стической онтологией, чтобы пользователи могли искать IGT, используя терминологию GOLD.

Защита исчезающих языков

Одним из важнейших направлений в сфере языковых технологий является де-

ятельность, направленная на сохранение, защиту и возрождение исчезающих языков и языков, находящихся в опасности. Этим занимается ЮНЕСКО, а также ряд специали- зированных организаций, которые описа- ны ниже вместе с реализуемыми ими про- ектами.

ELF [22]. Фонд исчезающих языков — это некоммерческая организация, базиру- ющаяся в Нью-Хейвене, штат Коннектикут. ELF поддерживает проекты поддержки языков и документации, находящихся под угрозой исчезновения, которые направле- ны на сохранение языков мира, одновре- менно предоставляя редкие лингвистиче- ские данные научному сообществу. С 1997 года Фонд спонсировал более 100 языко- вых проектов в 30 странах, а недавно начал разработку большого цифрового архива языковых данных, находящихся под угро- зой исчезновения. Основным механизмом поддержки ELF является финансирование отдельных лиц, племен и музеев.

ELD [23]. Программа документирова- ния исчезающих языков была основана в 2002 году. Целью Программы ELDP явля- ется сохранение находящихся под угрозой исчезновения языков во всем мире. С этой целью ELDP предоставляет гранты по всему миру отдельным лицам для документиро- вания языков, находящихся под угрозой ис- чезновения. Каждый год предоставляется от 30 до 40 грантов на проекты документа- ции по всему миру. На эти средства прово- дятся полевые работы по записи носителей языков, находящихся под угрозой исчезно- вения, на аудио- и видеозаписи, составле- нию документального описания по языку или жанру, находящемуся под угрозой ис- чезновения. Эти документальные коллек- ции затем архивируются и сохраняются и становятся свободно доступными через цифровой онлайн-архив исчезающих язы-

ков (ELAR), который является частью библиотеки Лондонского университета SOAS [24].

DOBES [25]. Документирование исчезающих языков. В 2000 году Фольксвагенфонд начал программу DOBES (Dokumentation bedrohter Sprachen) с целью документирования языков, которые потенциально находятся под угрозой исчезновения в течение нескольких лет. Было профинансировано 67 проектов документирования. Языковое документирование — это реакция языкового сообщества на имманентное исчезновение большинства мировых языков. Оно преследует три основные цели:

- Техническое обслуживание и ревитализация;
- Сохранение информации об языковом разнообразии и культурных ценностях человечества для будущих поколений носителей языка и исследователей;
- Введение подотчетности в лингвистические исследования.

E-MELD [26]. Электронная метаструктура для данных об исчезающих языках. Цель проекта — создание архитектуры эффективного сотрудничества между лингвистами, работающими над исчезающими языками. Одной из основных целей проекта является выработка консенсуса в отношении стандартов для метаданных, лингвистических аннотаций и языковой идентификации, что позволит обеспечить широкий доступ к данным в максимально полезной форме.

LACITO. Проект по архивированию лингвистических данных [27]. Целями проекта являются сохранение и распределение речевых данных. Основным источником данных для проекта является множество документов, зафиксированных в полевых условиях членами LACITO за последние тридцать лет. Это уникальные записи, в основном спонтанной речи на неписьменных

языках. Транскрипции и переводы некоторые из них были опубликованы, но оригиналы записей никогда не публиковались и должным образом не архивировались. Данные LACITO включают в себя как звук, так и текст — как минимум фонологическую транскрипцию и свободный перевод, а также, где это возможно, слово в слово глоссы, заметки и т. д. Текст индексируется по звуку на уровне «предложения» или интонационной группы.

Rosetta. Фонд библиотеки долговременного хранения естественных языков [28] реализуется на основе сотрудничества языковых специалистов и носителей языка, работающих над созданием общедоступной цифровой библиотеки естественных языков. Прототип долгосрочного архива — Rosetta Disk — никелевый диск диаметром 3 дюйма с почти 14 тыс. страницами информации, выгравированной на его поверхности. Поскольку каждая страница представляет собой изображение, а не цифровое кодирование 1 и 0, она может быть прочитана человеческим глазом с помощью оптического увеличения мощностью 500. Диск покрыт сферой из нержавеющей стали и стекла, что защищает диск от случайных ударов и истирания. При минимальном уходе он мог бы легко сохраняться и читаться в течение тысяч лет.

В основе коллекции дисков Rosetta лежит набор «параллельной» информации для более чем 1000 человеческих языков. Идея собрать параллельные тексты была вдохновлена оригинальным Розеттским камнем, который имел один и тот же основной текст (декрет), написанный тремя различными письменами. С тех пор коллекция проекта «Розетта» выросла до более чем 100 тыс. страниц языковых записей на более чем 2500 языках. В настоящее время коллекция размещена в интернет-архиве и продолжает расширяться.

Linguistic diversity. Программа ЮНЕСКО по поддержке языкового разнообразия и многоязычия в Интернете [29].

ЮНЕСКО призывает государства-члены выделять ресурсы и использовать соответствующие инструменты для поощрения и облегчения языкового разнообразия и многоязычия, включая интернет и средства массовой информации. В этой связи организация поддерживает включение новых языков в цифровой мир, создание и распространение контента на местных языках в интернете и каналах массовой коммуникации, поощряет многоязычный доступ к цифровым ресурсам в киберпространстве. Программа поддерживает следующие ресурсы по проблеме языкового разнообразия:

- Нормативные документы
- Языковая политика
- Продвижение местного контента в Интернете
- Измерение языкового разнообразия в Интернете
- Атлас языков, находящихся в опасности
- Интернационализованные доменные имена
- Специальные инициативы и мероприятия
- Глоссарий по управлению Интернетом

Атлас ЮНЕСКО по языкам мира, находящимся под угрозой исчезновения [30], насчитывает около 2500 языков (из которых 230 языки вымершие с 1950 года), т.е. приближается к общепринятым оценкам, что около 3000 языков находится под угрозой. Для каждого языка атлас содержит его название, степень опасности исчезновения и страны, в которых на нем говорят. Онлайн-издание предоставляет дополнительную информацию о количестве докладчи-

ков, соответствующих политиках и проектах, источниках, кодах ISO и географических координатах

Терминология и переводческая деятельность

Терминологическая и переводческая деятельность является предметом многих международных организаций, прежде всего ООН и ЮНЕСКО, а также многих отраслевых организаций. Здесь мы приведем сведения лишь о некоторых профессиональных структурах.

LTAC [31]. Консорциум терминологии и переводов создан для решения языковых проблем в глобальном масштабе. Как некоммерческий консорциум, LTAC использует обширный опыт, методы и технологии для терминологии, овладения языком и локализации/перевода. LTAC работает над продвижением управления терминологией, овладением языком и управлением переводческими проектами. LTAC предоставляет бесплатный доступ к языковым ресурсам для поддержки этой деятельности. LTAC реализует следующие проекты:

- *Linport* — формат представления переводческих материалов
- *TBX* — формат обмена терминологическими данными)
- *Tranquality* — “качество перевода” и модель его оценки с помощью системы многомерных показателей качества (MQM).
- *GEvTerm* — Глобальные терминологические вызовы
- *TerminOrg* Терминология для крупных организаций

IAMT [32]. Международная ассоциация машинного перевода.

IAMT включает три региональных ассоциации:

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

Архив IAMT [33] представляет собой электронное хранилище и библиографию статей, книг и документов по темам машинного перевода и компьютерных средств перевода. Среди его списков вы найдете полные материалы для всех конференций EAMT и IAMT.

Среди проектов EAMT — справочник коммерческих систем машинного перевода и средств поддержки автоматизированного перевода.

Международная стандартизация

Основной организацией по стандартизации в области языковых технологий и ЛИР является, конечно, ИСО, где имеется специализированный **Технический комитет ИСО 37 Язык и терминология ТК ISO37 [34]** Сфера ответственности комитета — стандартизация описаний, ресурсов, технологий и услуг, связанных с терминологией, письменным, устным переводом и другой языковой деятельностью в многоязычном информационном обществе. Комитет разработал 70 стандартов и 31 проектов находятся в стадии разработки. В состав ТК входят подкомитеты:

ISO/TC 37/SC 1 Принципы и методы

ISO/TC 37/SC 2 Терминология: рабочий процесс и языковое кодирование

ISO/TC 37/SC 3 Управление терминологическими ресурсами

ISO/TC 37/SC 4 Управление языковыми ресурсами

ISO/TC 37/SC 5 Письменный, устный перевод и связанные с ним технологии

Однако стандарты в данной области разрабатываются и в других организациях. Например, ведущей международной организацией по стандартизации для World Wide Web является Консорциум W3 **W3C [35]**.

Консорциум состоит из организаций-членов, которые поддерживают штатных сотрудников, совместно работающих над разработкой стандартов для Всемирной паутины. Среди стандартов и других нормативных документов W3C есть ряд документов, регламентирующих представление ЛИР, в частности, в формате связанных данных, например, Рекомендации по созданию лингвистических связанных данных: двуязычные словари [36].

Многие решения по стандартизации в области ЛИР были разработаны в результате деятельности профессиональных групп, не связанных с ИСО или W3C, но потом легли в основу многих стандартов. Таков, например, был проект **ISLE [37]. Международные стандарты для языковых технологий** Целью этого проекта было улучшение доступности языковых ресурсов при помощи стандарта описания метаданных ЛИР. С помощью такого стандарта стало возможным создание просматриваемого и доступного для поиска пространство ЛИР в Интернете.

Другой инициативой, положившей начало многим стандартам в сфере ЛИР, были разработки консорциума **TEI Инициатива по кодированию текстов [38]**. TEI — это международная организация, миссия которой заключается в разработке и поддержке руководящих принципов цифрового кодирования литературных, лингвистических и исторических текстов. Консорциум публикует руководящие принципы для электронного кодирования текста и обмена. Фактически руководство TEI превратилось в международный и междисциплинарный стандарт, который широко используется библиотекарями, музеями, издателями и отдельными уче-

ными для цифрового представления всех видов текстовых материалов.

Еще один проект, направленный на разработку и/или гармонизацию стандартов в области ЛИР, — это проект **RELISH [39]. Обеспечение функциональной совместимости лексиконов языков, находящихся под угрозой исчезновения, посредством гармонизации стандартов.** Данный проект направлен на установление единого способа представления структуры лексики в ЛИР, а также разработкой процедуры миграции гетерогенных лексиконов в совместимый со стандартами формат XML.

Лингвистические связанные открытые данные

Наиболее перспективным направлением международной коллаборации в области ЛИР является, по мнению автора, технология *Лингвистических связанных открытых данных* (LLOD) [40]. Это технология Семантической сети, разрабатываемой многими исследовательскими группами под общим руководством Фонда открытого знания [41], председателем которого является создатель Всемирной паутины Тим Бернерс-Ли.

Краткое описание проекта представлено по адресу [42]. Наиболее полное описание современного состояния LLOD имеется в работе [43].

Реализацию проекта LLOD осуществляет международная рабочая группа OWLG [44]. Проект LLOD представляет собой международную коллаборацию, занимающееся созданием, обеспечением совместимости, повторным использованием ЛИР в соответствии с принципами связанных данных.

Стандарты LLOD

Помимо сбора метаданных и создания облачной диаграммы LLOD, сообщество

OWLG стимулирует разработку стандартов сообщества в отношении словарей, метаданных и рекомендаций передового опыта. Согласно цит. выше обзору Cimiano et al. (2020), к ним относятся:

- для моделирования лексических ресурсов — OntoLex-Lemon, — стандарт сообщества для лексических ресурсов
- для моделирования лингвистических аннотаций
 - Web Annotation, стандарт W3C для аннотации веб-ресурсов (текстовых или иных)
 - NIF, стандарт сообщества для грамматической аннотации текста
 - CoNLL-RDF, словарь на основе NIF для представления RDF корпусов в традиционных форматах TSV («CoNLL»)
 - POWLA, словарь для общих лингвистических структур данных, которые можно использовать для дополнения NIF.
- для категорий лингвистических данных
 - Онтологии лингвистической аннотации (OLiA) для лингвистической аннотации
 - LEXINFO для грамматических и других функций в лексических ресурсах
- для идентификация языка
 - в виде строк с языковыми тегами с использованием языкового тега IETF BCP 47
 - с ISO 639-3
 - Glottolog URI для языков, не охваченных ISO 639
- для метаданных
 - Dublin Core , стандарт, который может использоваться для описания веб-ресурсов

- Словарь каталогов данных (DCAT), стандарт W3C для каталогов данных, опубликованных в Интернете
- METASHARE-OWL, словарь для метаданных

Кроме работ по созданию и развитию облака LLOD, технология связанных данных применяется еще во многих проектах, как правило, реализуемых международными сообществами лингвистов. Перечислим их.

CLLD — кросс-лингвистические связанные данные [45] осуществляет перевод в формат LLOD более десятка лингвистических баз данных, охватывающих языки мира. Это привело к созданию программной среды, которую можно использовать для разработки отредактированных коллекций баз данных, представленных лингвистами со всего мира.

Описание методологии CLLD можно найти в работе [46]. Список баз данных, реализованных как приложения CLLD и опубликованных на платформе CLLD, доступен по ссылке Datasets [47]. *Dictionaria* [48] — журнал словарей менее широко изучаемых языков, который работает на платформе CLLD, уже опубликовал 10 словарей.

LIDER: FP [49]. Связанные данные для аналитики. Проект предполагает определение эталонной архитектуры, которая включает в себя общие и частые задачи, требующие услуг NLP и бесплатных, открытых и совместимых ресурсов для анализа многоязычного и мультимедийного контента. Эталонная архитектура LIDER основана на открытых стандартах и обеспечивает:

- Эталонную модель, которая идентифицирует те задачи, в которых лингвистические связанные данные могут поддерживать контент-аналитику, и обеспечивает стандартную декомпозицию этих задач на элементы, которые совместно

могли бы решать эти задачи вместе с потоками данных между элементами

- Каталог архитектурных шаблонов, описывающий типы элементов, которые могут быть использованы в вышеуказанных задачах, типы связей между этими элементами, и ограничения на то, как они могут использоваться.

Модель OntoLex-Lemon, возникшая в результате работы Группы сообщества W3C, первоначально была разработана с целью обеспечить полное лингвистическое обоснование онтологий. Это означает, что выражения естественного языка, используемые в метках, определениях или комментариях элементов онтологии, снабжены подробным лингвистическим описанием.

Онтологии являются важным компонентом семантической сети, но современные языки онтологий, такие как OWL и RDF(S), не поддерживают обогащение онтологий лингвистической информацией, в частности информацией о том, как объекты онтологии, то есть свойства, классы, индивиды и т. д., могут быть реализованы на естественном языке. Модель OntoLex-Lemon направлена на то, чтобы закрыть этот пробел, предоставив словарь, который позволяет онтологиям обогащаться информацией о том, как описанные в них элементы словаря реализуются лингвистически, в частности в естественных языках.

Проект PRET-a-LLOD. Основная цель проекта заключается в обеспечении многоязычного междисциплинарного доступа к ЛИР, используемых в многоязычных трансграничных ситуациях. Это достигается за счет предоставления инструментов обнаружения данных, основанных на метаданных, агрегированных из нескольких источников, методологий описания свойств данных и услуг, а также инструментов для

вывода возможных значений ресурса, полученного после сложного конвейера.

В проекте реализуются методы, для обнаружения, преобразования и связывания лингвистических данных таким образом, чтобы они могли быть опубликованы как LLOD. С этим связана разработка трансформационной платформы, которая отображает наборы данных в форматы и схемы, которые могут быть использованы LLOD. Наконец, проект развивает экосистему для поддержки разработки языковых технологий, основанных на LLOD.

Европейские организации и проекты в области языковых технологий и ресурсов

В Европе создана разветвленная сеть организаций и консорциумов, которые занимаются языковыми технологиями и ресурсами.

Прежде всего, это Объединенный исследовательский центр (JRC), в рамках которого создан **Центр компетенций JRC по интеллектуальному анализу текста [50]**. JRC разработал инструменты языковой технологии для более чем двадцати языков и с 2004 года анализирует до 300 000 онлайн-новостных статей в день, создавая таким образом ценные метаданные. JRC также помогает распространять лингвистические ресурсы, выпускаемые другими организациями Евросоюза.

Вот перечень ресурсов, разработанных JRC:

- DGT-Acquis: многоязычный параллельный корпус официального журнала ЕС
- DCEP: Цифровой корпус Европарламента
- DGT Память перевода Генеральной дирекции переводов

- ECDC Память перевода Европейского центра профилактики и контроля заболеваний

- EAC Память перевода Генеральной дирекции образования и культуры

- JRC-Names: имена и варианты их написания

- JRC Тезаурус Eurovoc

JRC призван обслуживать непосредственно органы управления ЕС и их терминологические службы. Что же касается координации деятельности по науке и технологиям в странах-членах ЕС, то для этого созданы Европейские инфраструктурные исследовательские консорциумы (ERIC).

Специализированным ERIC в области языковых технологий является **CLARIN [51] — Общеввропейская исследовательская инфраструктура для языковых ресурсов и технологий**.

CLARIN основана на видении, что все цифровые языковые ресурсы и инструменты со всей Европы и за ее пределами должны быть доступны через единый онлайн-вход для поддержки исследователей.

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

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

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

Депозитные услуги. Одним из основных сервисов инфраструктуры CLARIN является обеспечение архивирования ЛИР и предоставление их сообществу 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.

Виртуальная языковая обсерватория. Цель — предоставить простой в использовании интерфейс, обеспечивающий единый процесс поиска большого количества ЛИР. Развитый синтаксис запроса позволяет выполнять более целенаправленный поиск.

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

ях идентификации, можно запросить учетную запись CLARIN.

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

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

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

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

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

CLARIN для исследователей. CLARIN для исследователей — это онлайн-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 ЛИП для исследователей в области цифровой гуманитаристики и NLP.

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

Обмен знаниями. В CLARIN создается инфраструктура обмена знаниями, включающая технические средства, которые предоставляют пользователям доступ к данным и инструментам, людей, которые управляют этими средствами, стандарты, условия доступа, лицензии, обеспечение качества и т.д. Основу инфраструктуры обмена знаниями составляют Центры знаний CLARIN [52].

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

Кроме CLARIN, имеется еще один ERIC, в котором реализуются языковые технологии. Это **DARIAH — Цифровая исследовательская инфраструктура для искусств и гуманитарных наук** [53]. DARIAH — это сеть людей, экспертных знаний, информации, знаний, контента, методов, инструментов и технологий из стран-членов организации. Она разрабатывает, поддерживает и эксплуатирует инфраструктуру в научно-исследовательской деятельности, основанной на ИКТ, и предоставляет ее исследователям для создания, анализа и интерпретации цифровых ресурсов. Одна из рабочих групп в DARIAH — **Лексические ресурсы**

[54]. Целями этой РГ являются исследование, оценка и рекомендации по стандартным инструментам и методам для создания, применения и распространения цифровых ЛИП и других видов структурированных данных; поощрение, развитие и публикации цифровых лексикографических исследований.

Среди ресурсов и сервисов DARIAH отметим следующие:

Vocabs [55] предоставляет услуги и инструменты, которые позволяют совместно создавать, поддерживать и публиковать словари и таксономии любого рода. Система основана на программном обеспечении с открытым исходным кодом *Skosmos*, которое использует SKOS в качестве базовой модели данных.

ISIDOR [56] — платформа и поисковая система, предоставляющая доступ к цифровым публикациям и данным гуманитарных и социальных наук.

CodiMD [57.] Онлайн-инструмент для совместного редактирования текста. CodiMD-это веб-редактор с открытым исходным кодом, который позволяет нескольким пользователям работать с одним текстом одновременно из разных мест.

TaDiRAH [58]. Таксономия цифровой исследовательской деятельности в гуманитарных науках

NeMO [59]. Онтология методов *NeMO* — это комплексная онтологическая модель научной практики в области социальных и гуманитарных наук, разработка которой осуществляется через исследовательскую сеть ESF NeDiMAN [60].

DARIAH-DE Topics Explorer [61] — метод анализа распределения семантических кластеров слов, так называемых “тем” в текстовой коллекции.

Кроме Консорциумов ERIC в Европе создано много других профессиональных международных объединений, имеющих

форму ассоциаций, советов, консорциумов других видов, и просто временных проектных сообществ, которые работают в области языковых технологий. Далее будут перечислены основные из них:

- ELRA [62]. Европейская ассоциация лингвистических ресурсов;
- TELRI [63]. Трансъевропейская инфраструктура лингвистических ресурсов;
- META-NET [64]. Сеть передового опыта;
- NexusLinguarum [65]. Европейская сеть лингвистических данных, ориентированная на Интернет;
- ELRC3 [66]. Европейская координация ЛИР;
- ELG [67]. Европейская Языковая Сеть;
- EACL [68]. Европейское отделение Ассоциации компьютерной лингвистики;
- EADH [69]. Европейская ассоциация цифровой гуманитаристики;
- EAGLES [70]. Консультативная группа экспертов по стандартам языковых технологий;
- MULTEXT-East [71]. Многоязычные текстовые инструменты и корпуса для языков Центральной и Восточной Европы;
- MLIA [72]. Многоязычный доступ к информации по Ковид-19;

Жанр журнальной статьи не позволяет представить подробные описания всех этих проектов и объединений. Ограничимся описанием проектов ELRA, тем более что недавно подписано соглашение о сотрудничестве ELRA и CLARIN, которое включает ELRA в состав органов Нврокомиссии.

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

Для достижения этой цели ELRA проводит такие действия, связанные с ЛИР, как идентификация ЛИР, распространение, производство, проверку, оценку технологий, информирование о ЛИР. При ELRA создано Агентство по оценке и распространению ЛИР [73] (ELDA) которое является оперативным органом ассоциации и реализует выполнение задач ELRA, определенных Советом ассоциации, в том числе коммерческих проектов.

ELRA и ELDA реализуют следующие сервисы на основе ЛИР:

- Идентификация и каталогизация ЛИР;
- Распространение ЛИР;
- Производство ЛИР;
- Проверка и оценка ЛИР;
- Поддержка правовых вопросов, связанных с ЛИР.

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

Для задач идентификации разработан Международный стандартный номер языковых ресурсов (ISLRN) — уникальная и универсальная модель идентификации для ЛИР, использующая стандартизированные метаданные.

Каталогизация. В каталоге ELRA ЛИР распределены по 4 категориям: «Речь и смежные ресурсы», «Письменные ресурсы», «Терминологические ресурсы» и «Мультимодальные / мультимедийные ресурсы».

META-SHARE. Под эгидой ELRA разработано META-SHARE, средство обмена и распространения открытых ЛИР, которое направлено на расширение доступа к таким ресурсам в глобальном масштабе. META-SHARE — это открытое, интегрированное,

безопасное и совместимое средство совместного использования и обмена для ЛИП (наборов данных и инструментов) для автоматической обработки естественного языка (NLP). META-SHARE спроектирована как сеть распределенных репозиториях ЛИП, включая языковые данные и базовые инструменты NLP.

Карта LRE. Для каталогизации ЛИП разработана модель их описания (карта LRE). В соответствии с этой схемой описано св. 6 тыс. ЛИП.

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

Специальная группа по интересам: «Языки с ограниченными ресурсами (SIGUL)». Созданная в 2017 г., SIGUL является совместной Группой по интересам Европейской ассоциации языковых ресурсов (ELRA) и Международной ассоциации речевого общения (ISCA). Эта группа сосредоточена на конкретных потребностях и требованиях для языков с ограниченными ресурсами.

Это далеко не все европейские проекты по языковым технологиям. Значительный список законченных проектов, в которых принимала участие ELRA, приводится по адресу [74].

Заключение

Из представленного обзора очевидно, что в мире и особенно в Европе созданная разветвленная сеть объединений и временных форм сотрудничества, образующих инфраструктуру языковых технологий. Наличие инфраструктуры значи-

тельно повышает эффективность исследований, сокращает дублирование, создает условия для коллабораций и, в конечном счете, повышает эффективность данной отрасли высоких технологий.

Ученым, работающим в этой сфере в разных странах, следует пристально изучить этот опыт. В каких-то случаях имеет смысл повторить аналогичные разработки на национальном уровне или в масштабах международного сотрудничества, в других — разумно присоединиться к европейским ресурсам и сервисам, особенно учитывая приверженность европейских организаций открытой науке. Во всяком случае, стратегия развития языковых технологий должна опираться на соответствующую инфраструктуру.

ЛИТЕРАТУРА / REFERENCES

1. Associatio for Computational Linguistics. URL: <https://www.aclweb.org/portal>.
2. Association Internationale de Linguistique Appliquée. URL: <https://aila.info/>.
3. International Committee for the Coordination & Standardisation of Speech Databases and Assessment Techniques. URL: <http://www.cocosda.org/>.
4. International Committee on Computational Linguistics (ICCL). URL: https://wiki2.org/en/International_Committee_on_Computational_Linguistics.
5. The International Quantitative Linguistics Association. URL: <http://www.iqla.org>.
6. International Speech Communication Association. URL: <https://www.isca-speech.org/iscaweb/index.php/about-isca>.
7. Linguistic Data Consortium (LDC). URL: <https://www ldc.upenn.edu>.
8. OLAC, the Open Language Archives Community. URL: <http://olac ldc.upenn.edu/>.
9. The Linguist List International Linguistics Community Online. URL: <https://linguistlist.org>.

10. SIL International, URL: <https://www.sil.org/>.
11. General Ontology for Linguistic Description. URL: <http://linguistics-ontology.org/info/about>.
12. Mailing Lists. URL: <https://old.linguistlist.org/lists/>.
13. Language and Location — A Map Annotation Project. URL: <http://llmap.org/>.
14. MultiTree. URL: <https://old.linguistlist.org/projects/multi-tree.cfm>.
15. Ethnologue Languages of the World. URL: <https://www.ethnologue.com/about>.
16. URL: <http://purl.org/linguistics/eltk>.
17. URL: <http://www.sil.org/~simonsg/metaschema/>.
18. Field Input Environment for Linguistic Data. URL: <http://emeld.org/tools/fieldinput.cfm>.
19. Lexicon Enhancement via the GOLD Ontology. URL: <http://lego.linguistlist.org/>.
20. The Online Database of Interlinear Text. URL: <http://odin.linguistlist.org/>.
21. IGT (interlinear glossed text) — interlinear text notes.
22. The Endangered Language Fund. URL: <https://ogmios.org/>.
23. The Endangered Languages Documentation Programme. URL: <https://www.eldp.net/en/about>.
24. URL: <https://www.soas.ac.uk/library/>.
25. Documentation of endangered languages. URL: <https://dobes.mpi.nl/research/>.
26. Endangered Metadata for Endangered Languages Data (E-MELD). URL: <http://emeld.org/>.
27. Linguistic Data Archiving Project (LACITO). URL: <http://xml.coverpages.org/lacitoAR-desc-english.html>.
28. Rosetta. URL: <https://rosettaproject.org/>.
29. Linguistic diversity and multilingualism on Internet. URL: <http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/linguistic-diversity-and-multilingualism-on-internet/>.
30. UNESCO Register of Good Practices in Language Preservation. URL: <http://www.unesco.org/languages-atlas/>.
31. Language Terminology/Translation and Acquisition Consortium. URL: <http://ltacglobal.org/about.htm>.
32. International Association for Machine Translation. URL: <http://eamt.org/international-association-for-machine-translation/>.
33. Machine Translation Archive/ Electronic repository and bibliography of articles, books and papers on topics in machine translation, computer translation systems, and computer-based translation tools. URL: <http://www.mt-archive.info>.
34. ISO / TC ISO 37 Language and terminology. URL: <https://www.iso.org/ru/committee/48104.html>.
35. World Wide Web Consortium. URL: <https://www.w3.org/Consortium/>.
36. URL: <https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/>.
37. ISLE International Standard for Language Engineering : URL: <https://www.mpi.nl/ISLE/>.
38. Text Encoding Initiative Consortium. URL: <https://tei-c.org/>.
39. Rendering Endangered Language Lexicons Interoperable through Standards Harmonization. URL: <https://old.linguistlist.org/projects/relish.cfm>.
40. Linguistic Linked Open Data. URL: <https://linguistic-lod.org/llod-cloud>.
41. The Open Knowledge Foundation. URL: <https://okfn.org/>.
42. URL: https://ru.qaz.wiki/wiki/Linguistic_Linked_Open_Data.
43. Cimiano, Philipp; Chiarcos, Christian; McCrae, John P.; Gracia, Jorge (2020). Linguistic Linked Data: Representation, Generation and Applications. Springer International Publishing.

44. The-Open-Linguistics-Working-Group. URL: <https://ru.scribd.com/document/126882831/The-Open-Linguistics-Working-Group>.
45. URL: <https://clld.org/>.
46. Reconciling Heterogeneous Descriptions of Language Resources /John P. McCrae, Philipp Cimiano CIT-EC, Bielefeld University Bielefeld, Germany,Victor Rodríguez Doncel, Daniel Vila-Suero Jorge Gracia Universidad Politecnica de Madrid, Madrid, Spain @fi.upm. es Luca Matteis, Roberto Navigli University of Rome, La Sapienza, Rome, Italy Andrejs Abele, Gabriela Vulcu Paul Buitelaar Insight Centre, National University of Ireland Galway, Ireland. URL: <https://www.aclweb.org/anthology/W15-4205.pdf>.
47. URL: <https://clld.org/datasets.html>.
48. URL: <https://dictionaria.clld.org/>.
49. <https://docplayer.net/139432478-Lid-er-fp-linked-data-as-an-enabler-of-cross-media-and-multilingual-content-analytics-for-enterprises-across-europe.html>.
50. URL: <https://ec.europa.eu/jrc/en/language-technologies>.
51. Common European Research Infrastructure for Language Resources and Technology. URL: <https://www.clarin.eu/>.
52. URL: <https://www.clarin.eu/content/knowledge-centres>.
53. The Digital Research Infrastructure for the Arts and Humanities (DARIAH). URL: <https://www.dariah.eu>.
54. URL: <https://digilex.hypotheses.org/>, <https://github.com/DARIAH-ERIC/lexicalresources>.
55. URL: <https://vocabs.dariah.eu/en/>.
56. URL: <https://isidore.science/>.
57. URL: <https://pad.gwdg.de/>.
58. URL: <http://tadirah.dariah.eu/vocab/index.php>, <https://vocabs.dariah.eu/tadirah2/en/>.
59. URL: <http://nemo.dcu.gr>.
60. URL: <http://archives.esf.org/coordinating-research/research-networking-programmes/humanities-um/nedimah.html>.
61. URL: <https://de.dariah.eu/en/web/guest/topicexplorer>.
62. European Language Resources Association. URL: <http://www.elra.info/en/>.
63. Trans-European Language Resources Infrastructure. URL: <http://telri.nytud.hu/>.
64. Multilingual Europe Technology Alliance. URL: <http://www.meta-net.eu/>.
65. NexusLinguarum. URL: <https://nexuslinguarum.eu/>.
66. European Language Resources Coordination. URL: <http://www.elra.info/en/projects/current-projects/elrc3/>.
67. European Language Grid. URL: <http://www.elra.info/en/projects/current-projects/european-language-grid/>.
68. Association for Computational Linguistics, European Chapter. URL: <http://eacl.org/>.
69. European Association for Digital Humanities. URL: <https://eadh.org/>.
70. Expert Advisory Group on Language Engineering Standards. URL: <http://www.ilc.cnr.it/EAGLES/home.html>.
71. MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. URL: <http://nl.ijs.si/ME/>.
72. COVID-19 Multilingual Information Access (MLIA) initiative. URL: <http://eval.covid19-mlia.eu/projects/covid-19-mlia-init/>.
73. The Evaluations and Language resources Distribution Agency. URL: <http://www.elra.info/en/about/elda/>.
74. Completed and Archived Projects. URL: <http://www.elra.info/en/projects/archived-projects/>.