

О СОЗДАНИИ ЦЕНТРА ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ РАН*

© 2019 г. А.Б. Антопольский

доктор технических наук,
главный научный сотрудник Института научной информации
по общественным наукам РАН (ИНИОН РАН),
Россия, 117997, г. Москва, Нахимовский проспект, д. 51/21
ale5695@yandex.ru

Дата поступления материала в редакцию: 06 июля 2019 г.

Дата публикации: 31 августа 2019 г.

CONCERNING DEVELOPMENT OF A CENTER OF LINGUISTIC RESOURCES OF THE RAS

© 2019 Alexander B. Antopolskii

Doct. Sci. (Tech.),
Head Researcher at the Institute of Scientific Information
for Social Sciences of the RAS (INION RAN),
Nakhimovskiy Prospect 51–21, Moscow, 117997, Russia
ale5695@yandex.ru

Received: July 06, 2019

Date of publication: August 31, 2019

Резюме. Предлагается концепция создания Центра лингвистических ресурсов РАН как организационно-технологической структуры, обеспечивающей агрегацию и интеграцию российских цифровых (электронных) лингвистических информационных ресурсов (ЛИР), реализующих различные задачи теоретической и прикладной лингвистики. Центр должен стать частью Цифровой системы управления сервисами научной инфраструктуры коллективного пользования, создание которой предусмотрено в Национальном проекте “Наука”. Описывается состояние ЛИР в РАН на основе проведенной инвентаризации и предлагается их типология. Формулируются задачи Центра ЛИР. Предлагаются способы действий Центра по отношению к различным типам ЛИР, как документным, так и структурированным. Особое внимание предлагается уделить корпусным и лексикографическим ресурсам. Кратко формулируются организационные, экономические и правовые аспекты функционирования Центра.

Ключевые слова: лингвистические информационные ресурсы, инвентаризация, типология, агрегация, навигация, интеграция, Центр лингвистических ресурсов.

Abstract. The paper discusses a conception for a Center of linguistic resources of the RAS as organizational and technological structure providing aggregation and integration of the Russian digital (electronic) linguistic information resources (LIR) providing implementation of various tasks in theoretical and computational linguistics. The center is thought as a part of the Digital system of management of services of scientific infrastructure of collective use, provided in the National project “Science”. The state of LIR in RAS is described on the basis of a previous investigation, and their typology is discussed. The tasks of the LIR Center are formulated. Models of actions of the Center in relation to various types of LIR, both documentary and structured, are offered. It is proposed to pay special attention to corpus and lexicographic resources. Organizational, economic and legal aspects of the Center are summarized.

Keyword: linguistic information resources, inventory, typology, aggregation, navigation, integration, Center of linguistic resources.

* Статья отражает результаты работы по исследовательскому проекту № 18-00-002\18 “Интеграция научно-информационных ресурсов учреждений РАН (на примере языкознания) как части единого цифрового пространства РАН”, реализуемому при поддержке РФФИ.

Для цитирования: Антопольский А.Б. О создании центра лингвистических ресурсов РАН // Известия Российской академии наук. Серия литературы и языка. 2019. Т. 78. № 4. С. 5–12.
DOI: 10.31857/S241377150006107-0

For citation: Antopolskii, A.B. *O sozdanii tsentra lingvisticheskikh resursov RAN* [On Creation of a Center of Linguistic Resources of the RAS]. *Izvestiya Rossijskoj akademii nauk. Seriya literatury i yazyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2019, Vol. 78, No 4, pp. 5–12. (In Russ.)
DOI: 10.31857/S241377150006107-0

DOI: 10.31857/S241377150006107-0

Постановка задачи

Быстрое развитие информационных технологий в лингвистике как науке, а также необходимость решения различных прикладных задач, использующих методы и средства компьютерной лингвистики (таких как обработка текста, анализ и синтез речи, автоматический перевод, обучение языкам и др.), привели к созданию большого числа электронных лингвистических информационных ресурсов (ЛИР) различного назначения. Деятельность по формированию ЛИР, обеспечению к ним доступа, сохранности, открытости, возможности многократного использования, требует оптимизации, координации и системного подхода, т.е., обобщенно говоря, — требует управления деятельностью в сфере ЛИР. Поскольку наиболее продвинутыми в этой области являются учреждения РАН, предлагается начать организацию этой деятельности в РАН.

Управление деятельностью в сфере ЛИР должно стать частью Цифровой системы управления сервисами научной инфраструктуры коллективного пользования, создание которой предусмотрено в Национальном проекте “Наука”.

Создание системы управления ЛИР обеспечит значительный экономический эффект — при тех же затратах можно получить больший результат.

Определения

Лингвистические информационные ресурсы (ЛИР) — организованные массивы речевых и языковых данных, представленных на машинных носителях и предназначенных для использования в науке и различных сферах практической деятельности, а также массивы документов, отражающих результаты научных исследований в сфере языкознания.

Интеграция информационных ресурсов — создание централизованных ресурсов, полностью включающих данные и функции интегрируемых ресурсов.

Агрегация информационных ресурсов — создание частично распределенных ресурсов, в которых

централизованы метаданные и некоторые функции, как правило, поиск и навигация,

Зарубежный опыт

В мире создан ряд организаций, занимающихся разработкой, интеграцией и агрегацией лингвистических ресурсов, а также координацией деятельности в этой области. К их числу относятся:

LDC (Linguistic Data Consortium, USA)¹,

ELRA (European Language Resources Association)²

TELRI (TransEuropean Language Resources Infrastructure)³.

Перед этими организациями стоят следующие задачи:

- разработка единых стандартов создания ресурсов;
- сбор, учет, каталогизация ресурсов, создаваемых партнерами;
- разработка способов защиты от несанкционированного доступа;
- создание единых экспертных требований;
- планирование единой стратегии разработки лингвистических ресурсов;
- создание многофункциональных лингвистических ресурсов большого объема для использования в разных странах.

Зарубежный опыт должен быть использован при проведении аналогичных разработок в России. Следует добавить, что проблемам создания ЛИР ежегодно посвящается большое количество научных конференций во всем мире.

Состояние проблемы в России

В результате проведения мониторинга академических информационных ресурсов был создан

¹ <https://www ldc.upenn.edu>

² <http://www.elra.info/en/>

³ <http://telri.nytud.hu>

Навигатор информационных ресурсов по языкознанию⁴. Количество электронных ЛИР, отраженных в Навигаторе, по каждому типу приведено в табл. 1.

Следует отметить, что эти данные указывают количество не только собственно ресурсов по языкознанию, но также ресурсы универсальные по тематике, включающие лингвистические ресурсы как часть. Особенно это касается документных ресурсов (библиографии, каталоги, электронные библиотеки, периодика).

Таблица 1. Функциональная типология и статистика лингвистических информационных ресурсов в РАН

№	Типы лингвистических ресурсов	Кол-во
1.	Библиографии, библиотечные каталоги, архивные описи, каталоги ссылок	93
2.	Электронные коллекции и библиотеки полных текстов (книги, диссертации, отчеты, труды конференций и др.)	138
3.	Периодические, продолжающиеся издания и архивы периодики	77
4.	Корпуса	15
5.	Лексикографические ресурсы	83
6.	Этнолингвистические и социоллингвистические БД	10
7.	Лингвистические географические системы, атласы	5
8.	Электронные представления памятников письменности	5
9.	Активные лингвистические ресурсы (алгоритмы, процессоры, программы)	28
10.	Граматики	11
11.	Описания языков и комплексные лингвистические сайты	14
12.	Информационные языки	12
13.	Энциклопедии, справочники, реестры языков	75
14.	Сведения об отдельных персонах (сайты и страницы ученых, личные фонды, биобиблиографии)	471
15.	Списки, перечни, указатели персон	31
16.	Медиаресурсы	8
17.	Сайты учреждений – владельцев ЛИР	116
18.	Сайты-сателлиты и ресурсы во внешних АИС	46
	Всего	1238

В таблице 1 также представлена функциональная типология электронных лингвистических ресурсов, разработанная на основе мониторинга российских академических ресурсов. В данном

варианте она несколько упрощена по сравнению с двухуровневой типологией, использованной в Навигаторе.

При классификации ЛИР иногда выделяют активные и пассивные лингвистические ресурсы. К пассивным формам относят словари, письменные текстовые массивы (корпуса текстов), фонетические ресурсы, электронные библиотеки и т.д., к активным формам – алгоритмы, модели, программы, базы знаний.

В ходе мониторинга использовалась также тематическая классификация информационных ресурсов на основе Государственного рубрикатора научно-технической информации (ГРНТИ), однако опыт показал, что эта классификация устарела и требует модернизации. Однако на данном этапе было принято решение сохранить классификацию ЛИР на основе ГРНТИ как основной официальной российской классификации научной информации. Одновременно будет разработана новая тематическая классификация ЛИР, отвечающая современным реалиям в теоретическом и прикладном языкознании. Также признана необходимой разработка специального фасета, отражающего принадлежность ЛИР к конкретному языку, а также к группам языков – как генеалогическим, так и ареальным.

Приведенные данные свидетельствуют, что в академических учреждениях создается значительное число лингвистических информационных ресурсов, поэтому координация деятельности по их созданию даст значительный экономический эффект. Очевидно, что в ближайшей перспективе мониторинг и каталогизацию ЛИР следует распространить на все отечественные ЛИР.

Конечно, нужно иметь в виду, что множество документных ресурсов по языкознанию содержатся в общенациональных информационных системах, таких как Национальная электронная библиотека, Электронная библиотека диссертаций, Киберленинка, Научная электронная библиотека и других. Поэтому в рамках взаимодействия с этими организациями следует выработать решения, минимизирующие дублирование при обработке этих документов.

Значительный объем ресурсов по языкознанию практически всех категорий представлен также в информационных системах вузов, а также в различных общественных и коммерческих проектах, которые на данном этапе не рассматривались.

Данные мониторинга убедительно показывают, что деятельность учреждений РАН по созданию, поддержке и обеспечению сохранности ЛИР всех

⁴ <http://nirya.alexo.beget.tech/web?page=1>