Комплексные словарно-текстовые интернет-корпусы (на материале русского и татарского языков)[20] (Complex Internet Dictionary-Textual Corpora (on Basis of Russian and Tatar Languages))

К.Галиуллин, Е.Горобец, Г.Каримуллина, Р.Каримуллина, А.Гизатуллина, Д.Мартьянов

Казанский (Приволжский) федеральный университет

Abstract

The article describes the features of development, structure and filling of dictionary-textual and dictionary corpora (funds), which are created in Kazan University. These funds contain materials of Russian and Tatar languages. Also we touch the problem of their role in dataware of linguistics in general and particularly in linguography.

Современная ситуация информационного обеспечения лингвистики наглядно свидетельствует о том, что наиболее перспективной формой существования языкового справочника является интернет-версия, среди «плюсов» которой [1]:

1) глобальная обобщественность материалов языковых справочников, когда благодаря интернет-технологиям обеспечивается доступ к ним широкому кругу пользователей;

2) обширные возможности сопряжения, установления связи с сетевыми справочниками, которые содержат сходные, сопоставимые или дополняющие материалы, и формирования лингвографических интернет-комплексов на основе ресурсов, размещенных как на одном, так и на разных порталах (сайтах).

Наличие этих (и других) «плюсов» в немалой степени обусловило разработку интернет-компонентов для многих справочных корпусов (фондов), созданных и создаваемых в Казанском университете на материале русского и татарского языков.

Среди них значительную часть составляют словарно-текстовые комплексы, состоящие из двух основных информационных модулей - текстового и словарного компонентов.

Наиболее обширным из этих комплексов является корпус русскоязычных памятников Казанского края XVI-XVII веков, который содержит материалы 883 документов указанного периода общим объемом 1 068 000 словоупотреблений (без учета цифровых обозначений; данные на 15 апреля 2011 г.). В настоящее время подготовлены и размещены в сети в качестве информационного ресурса пять интернет-словарей: а) XVI века // http://www.klf.ksu.ru/kazan/16 (материалы 100 источников общим объемом более 79540 словоупотреблений); б) первой четверти XVII века //

20 Работы по формированию компьютерных фондов, подготовке словарей, изданию и размещению языковых справочников в Интернете поддержаны, в частности, Федеральной целевой научно-технической программой «Исследования и разработки по приоритетным направлениям развития науки и техники» на 2002-2006 гг.; Российским гуманитарным научным фондом (ряд проектов, в их числе «Машинный фонд татарского языка: словарный подфонд», «Компьютерная поддержка русской лексикографии XVIII века»; «Большой корпус русского языка XVIII века»; «Комплексный фонд русскоязычных памятников Казанского края XVI-XVII веков: текстовый и словарный подфонды»; «Комплексный справочный фонд словарей русского языка XVIII – первой половины XIX века», проект 11-04-12076в); Российским фондом фундаментальных исследований; Культурным центром имени Дж. Неру при Посольстве Индии в Российской Федерации; Федеральной целевой программой «Русский язык» (проект «Компьютерный лингвографический фонд русского языка»); Аналитической ведомственной целевой программой «Развитие научного потенциала высшей школы (2009-2010 гг.)»; Республиканской целевой программой «Русский язык в Татарстане», а также Комитетом по реализации Закона «О языках народов Республики Татарстан» при Кабинете Министров Республики Татарстан и др.

http://www.klf.ksu.ru/kazan/1_4_17 (165 источников, более 180560 словоупотреблений); в) второй четверти XVII века // http://www.klf.ksu.ru/kazan/2_4_17 (124 источника, более 255590 словоупотреблений); г) третьей четверти XVII века // http://www.klf.ksu.ru/kazan/3_4_17 (253 источника, более 145180 словоупотреблений); д) четвертой четверти XVII века // http://www.klf.ksu.ru/kazan/4_4_17 (139 источников, более 185090 словоупотреблений).

Кроме того, разрабатываются словарно-текстовые комплексы, описывающие: а) язык памятников, связанных с русско-восточными отношениями XVI-XVII веков; б) язык русских пословиц и поговорок конца XVII – первой половины XVIII века; в) язык писем М.В.Ломоносова (http://www.klf.ksu.ru/lomonosov); г) язык русской поэзии XIX века; д) язык Г.Р.Державина и др.

Материалы письменных источников татарского языка представлены в словарно-текстовых комплексах, описывающих: а) язык поэзии Габдуллы Тукая (http://www.klf.ksu.ru/tukay); комплекс включает материалы 411 произведений поэта общим объемом 45 899 словоупотреблений; б) язык цикла «Моабитские тетради» Мусы Джалиля (http://www.klf.ksu.ru/jalil); комплекс включает материалы 93 произведений общим объемом 12369 словоупотреблений; в) язык татарских пословиц и поговорок; комплекс на данном этапе включает 17089 паремий общим объемом 95405 словоупотреблений; г) язык произведений современной художественной литературы и текстов СМИ и др.

Описание архитектуры, макро- и микроструктуры, а также информационного наполнения указанных комплексов см. [2].

Ряд создаваемых корпусов (фондов) направлен на решение задачи комплексной информатизации исследований и разработок в области лингвографии21, на создание полифункциональных металингвографических систем, аккумулирующих, особым образом организующих словарные данные, обеспечивающих возможность манипуляции этими данными, их сравнения, отбора и перегруппировки в соответствии со стоящими перед исследователем задачами (об информационном потенциале справочников см. [4]).

В настоящее время в Казанском университете разрабатываются два сводных фонда словарей (СФС) – один по русским лингвографическим источникам, другой – по татарским.

Фонд академических словарей русского языка XVIII – первой половины XIX века объединяет материалы основных словарей русского языка указанного периода (в первую очередь Словаря Академии Российской (1-е и 2-е издания – САР-1, САР-2), Словаря церковнославянского и русского языка 1847 года (СЦРЯ), а также других крупных словарей). Свидетельством интереса исследователей к указанным академическим словарям являются, в частности, фототипическое и наборное переиздания САР-1, САР-2, СЦРЯ и в России и за рубежом. Однако эти издания недоступны большинству специалистов. Размещение словарей в pdf-формате в Интернете расширяет круг пользователей, но не позволяет осуществлять необходимые исследователю операции по поиску, отбору, сравнению словарных материалов.

На основе материалов 317 публикаций автономных татарских языковых справочников, вышедших из печати в Российской Федерации во второй половине XX - начале XXI века (1951-2008 гг.), строится справочный фонд татарских словарей соответствующего периода.

СФС состоит из трех основных информационных модулей – составных частей системы, обладающих определенной самостоятельной ценностью: модуля "Источники", модуля "Лингвографические характеристики (признаки, параметры)", модуля "Словник".

Комплексный характер работы предопределил использование гипертекстовой технологии, которая в современных исследованиях называется в числе наиболее перспективных форм организации информационных фондов. Как показывает опыт, гипертекстовая организация позволяет представить материал в наиболее оптимальной форме, использовать сильные стороны компьютерной техники (по сравнению с традиционными разработками), обеспечивает условия для организации многоаспектных информационных запросов.

---

21 Лингвография – междисциплинарная область языкознания, теория и практика создания языковых справочников, словарей (о лингвографии см.: [3]); подразделами лингвографии являются лексикография, фразеография, морфемография, паремиография и др.

Модули СФС объединены общим объектом описания, которым являются лингвографические источники соответствующего периода.

Из модуля "Источники" пользователь может получить разнообразную информацию о макроструктуре и микроструктуре словаря, его адресатах, объеме словника и т.п.

Основная задача модуля "Лингвографические характеристики (признаки, параметры)" – снабдить пользователя сведениями о той информации, которая содержится в словарях, показать ее объем, характер и способы представления.

Для СФС разрабатывается система характеристик (параметров), позволяющая описать различные типы информации, содержащейся в словарях базы.

В рамках модуля "Словник" объединяются и описываются материалы словников словарей-источников СФС. Основу данного модуля составляет сводный словник.

Необходимость полной инвентаризации материалов словарей в ряде случаев обусловлена также наличием внутристатейных единиц, нередко остающихся вне поля зрения пользователя.

Отличительной особенностью СФС является также фиксация различного рода сверхсловных единиц (фразеологизмов, пословиц и поговорок и др.), описанных в словарях-источниках. Подфонд сверхсловных единиц снабжен указателем слов, связывающим его со сводным словником.

Программный комплекс модуля "Словник" позволяет получать разнообразные сведения о лексическом составе источников в виде списка слов, дополненного количественными характеристиками:

а) по одному словарю - все слова, зафиксированные в словаре, в том числе алфавитный список единиц источников, где слова расположены не по алфавиту; заголовочные единицы основных и отсылочных словарных статей; внутристатейные слова (все или их определенные типы, в соответствии с характеристиками, отраженными в СФС); омонимы; слова, стоящие в словаре не в начальной форме; имеющие какие-нибудь формальные признаки, например, определенные буквы, сочетания букв, морфемы и т.д.

б) по двум и более словарям (дополнительно к указанным в предыдущем пункте сведениям) - объединенный словник; слова, общие для рассматриваемых источников; лексические единицы, зафиксированные в одном источнике (в одних источниках) и отсутствующие в другом (других) и т.д.

Интегрирование СФС в автоматизированное рабочее место лингвографа (словарника), как показывает опыт эксплуатации, способствует оптимизации лингвографической деятельности, дальнейшему повышению ее качества.

Использованная литература

1. Галиуллин К.Р. Интернет-лингвография: русские текстоописывающие словари // Проблемы истории, филологии, культуры.- Вып. 2(24).- Магнитогорск; Новосибирск: Аналит, 2009.- С.635-639.
2. Галиуллин К.Р. Лингвография и тексты: инвентаризующие языковые справочники / К.Р.Галиуллин, Р.Н.Каримуллина // Ученые записки Казан. гос. ун-та.- Т.151.- Серия Гуманитарные науки.- Кн.3.- Казань, 2009.- С.222-229.
3. Компьютерная лингвография / науч.ред. Н.К.Замов, К.Р.Галиуллин.- Казань: Изд-во Казан. ун-та, 1995.- 119 с. (Интернет-версия: http://www.ksu.ru/f10/publications/1995/comp_ling.php)
4. Галиуллин К.Р. Металингвографические фонды: информационный потенциал / К.Р.Галиуллин, Р.Н.Каримуллина, М.Р.Загидуллин, Д.В.Ковшарева // IV Международные Бодуэновские чтения (Казань, 25-28 сент. 2009 г.): тр. и матер.: в 2 т.- Казань: Казан. гос. ун-т, 2009.- Т.1.- С.170-172.

SPECOM 2011

The 14th International Conference

"Speech and Computer"

27 – 30 September 2011

Kazan, Russia

PROCEEDINGS

Moscow State Linguistic University

Moscow, Russia

Kazan (Privolzhsky) Federal University

Kazan, Russia