И.И. Саженин
Новосибирский государственный педагогический университет

Разработка параметров поисковой системы для Словаря русских народных говоров

Аннотация: В статье изложены результаты исследования, посвященного созданию параметров автоматизированной поисковой системы применительно к Словарю русских народных говоров.

The paper presents the results of research devoted to the development of automated search settings as applied to The Vocabulary of Russian National Dialects.

Ключевые слова: корпус, корпусная лингвистика, поисковая система, поисковые параметры, лексикография, словарь, диалект, диалектизм, помета, тэг.

Corpus, corpus linguistics, search system, search settings, lexicography, vocabulary, dialect, dialectism, mark, tag.

УДК: 81

Контактная информация: Новосибирск, ул. Вилюйская, 28. НГПУ, кафедра современного русского языка. E-mail: sajana84@mail.ru.

Активное развитие в последние десятилетия компьютерных технологий создало ряд новых возможностей для оптимизации исследовательского и образовательного процессов. Компьютерные технологии и, в первую очередь, веб-технологии позволяют разрабатывать инструменты для решения задач исследовательского и прикладного характера в области лингвистики. Возможность оперативного доступа к различным видам информации, являющейся востребованной для проведения разнообразных исследовательских работ и быстрая ее обработка открывают новые возможности для ученых-филологов. И, если на сегодняшний день имеет место наличие разного рода корпусов, снабженных специализированными поисковыми системами, основанных на собрании текстов, позволяющих вести работу с речевыми произведениями, то представленность в Сети лексикографических источников, во-первых, не высокая и разобщенная, а, во-вторых, не создано до сих пор такого ресурса, который бы позволял извлекать информацию из большого количества словарей, а его поисковая система могла бы предоставлять исследователю именно ту информацию, которая ему нужна, оперативно и в должном объеме.

О том, что из себя должен представлять такой ресурс и в соответствии с какими критериями должна быть разработана поисковая система, уже было нами описано, см.: [Саженин, 2011, с. 447]. На начальном этапе в качестве массива данных мы определили материалы следующих лексикографических источников: [Крысин, 2000; Ожегов, 1997; Фасмер, 1986–1987, т. 1–4; Словарь русского языка, 1999].

Для выработки профессионально ориентированных параметров поисковой системы мы попытались определить, какие факторы должны обеспечить выбор таких параметров. В качестве основных факторов мы предложили содержательный и целевой: то есть, во-первых, необходимо учесть, какую информацию может предоставить исследователю тот или иной лексикографический источник в зави-

симости от своего типа, структуры, содержания и целевой направленности, во-вторых, поиск информации зависит от цели, которую исследователь ставит перед собой в своей исследовательской работе. Проанализировав каждый из включенных в массив данных лексикографический источник, мы пришли к выводу, что некоторая информация представлена в словаре «открыто» – например, информацию о принадлежности языковой единицы к некому пласту лексики мы можем считать за счет системы помет, однако, существует иной тип информации, например, о способах токования слова, о типе лексического значения, который формально не выражен, и тем не менее, исследователь, работая с определенной словарной статьей, такую информацию считывает в силу своих знаний, умений и навыков.

Приняв во внимание такое положение дел, словарные статьи мы разметили в соответствии с типом информации, содержащейся в словарях открыто: система помет, язык-источник, сфера функционирования; а также в соответствии с типом информации, которая содержится в словарной статье, но не эксплицирована каким-либо специальным способом: статус языка-донора (источник, посредник), вид лексики с точки зрения ее происхождения, принадлежность языковой единицы к тому или иному языковому объединению в соответствии с генеалогической классификацией языков, способ толкования лексического значения и др. Таким образом, система параметров поискового менеджера, например, для Толкового словаря иноязычных слов Л.П. Крысина выглядит следующим образом:

1) поиск по лемме (универсальный, применим ко всему массиву данных);
2) поиск по языку-донору;
3) поиск по статусу языка-донора: (источник, посредник);
4) поиск по сфере функционирования.

Так, для создания параметров поиска по языку-донору, мы выбираем языки, являющиеся, по мнению авторов словаря, источниками, например: немецкий, английский, французский, итальянский.

Таким образом, параметр № 2 (поиск по языку-донору) будет условно содержать четыре языка-донора для выбора, а так же будет возможность ограничить поиск по языку-донору с учетом параметра № 3 (поиск по статусу языка-донора), если в этом будет исследовательская необходимость. Иначе, если исследователю требуется сделать сплошную выборку слов, например, итальянского происхождения, и при этом учесть статус данного языка (источник или посредник), то он сможет это сделать, поставив метку в соответствующие поля. Параметр № 4 (поиск по сфере функционирования) будет содержать пятьдесят три сферы функционирования, к которым принадлежат лексические единицы.

В зависимости от целей исследования можно варьировать запросы, например, выбирая два и более языка, определяя их статус (источник или посредник), указывая сферу функционирования, к которой должны принадлежать нужные исследователю лексические единицы.

Следующей задачей стало расширение массива данных за счет включения в него ряда других лексикографических источников и разработка параметров поисковой системы для новой части массива данных в соответствии с теми же факторами. В данной статье мы планируем представить результаты анализа Словаря русских народных говоров [1965–1994, вып. 1–28] с точки зрения информации о языковой единице, которую словарь может предоставить исследователю в явном или неявном виде, а также выстроить на основе полученных данных систему параметров поисковой системы.

В явном виде, как правило, любой лексикографический источник представляет информацию о принадлежности языковой единицы к определенной группе (частеречная принадлежность, категориальная принадлежность языковой единицы, сфера ее употребления, степень употребления и др.). Такого рода информация

содержится в системе помет. Анализируемый словарь обладает следующей системой помет: грамматические, семантические, исторические и описательные. Грамматические пометы указывают на принадлежность слова к части речи, на его грамматические значения и формы. Система этих помет и основание их применения такие же, как и в семнадцатитомном «Словаре современного русского литературного языка» 1948 – 1965 гг. [Филин, 1965, с. 11].

К семантическим пометам составители относят *перен.* – переносное значение (или оттенок значения), которое стоит в явной связи с прямым значением; образное употребление слова – в таких случаях авторы используют знак *.

Еще одна функция помет в данном лексикографическом источнике – дать представление пользователю словаря о месте, занимаемом словом в диалектной лексике, о том, насколько широко употребляется слово в системе одного говора, о сфере употребления слова. Например, *устар.* – устарелое, *фольклор.* – фольклорное, «так говорили раньше», «это слово знали старики», *редко* – редкое, *детское* – в детском языке и т. д. [Там же, с. 11].

Таким образом, основанием для определения параметров поисковой системы может стать система помет словаря. Иными словами, каждой словарной статье, каждой лемме будет присвоен соответствующий код или иначе – тэг, характеризующий статью и лемму, как содержащие или не содержащие ту или иную помету.

В неявном виде материалы словаря могут так же нести достаточно широкий массив сведений о языковой единице. Например, в данный лексикографический источник включена диалектная лексика всех русских народных говоров XIX – XX вв. Диалектные слова по своему значению чрезвычайно разнообразны. Среди них имеются названия явлений живой и неживой природы, термины сельского хозяйства, скотоводства, охоты, рыболовства, всевозможных ремесел и занятий, бытовая лексика, слова с отвлеченными значениями и т. п. [Филин, 1965, с. 5]. Таким образом, следующим типом параметров поисковой системы может стать принадлежность языковой единицы к определенной сфере функционирования. Присвоение соответствующих кодов той или иной словарной статье и лемме будет уже обусловлено не наличием каких-либо маркеров в тексте словарной статьи (грамматических или семантических помет), а осознаваемым специалистом фактом принадлежности языковой единицы к определенной сфере ее бытования. Автоматизировать данный процесс полностью пока не представляется возможным. Иными словами, перед нами стоит два типа задач: аналитическая и практическая. Первая – на основании материалов словаря составить список сфер функционирования лексики, представленной в словаре, вторая – присвоить леммам и словарным статьям необходимые коды-соответствия.

Есть еще один тип информации, заключенный в самом диалектном слове, но никаким специальным образом не эксплицированный. Дело в том, что, формируя словник, авторы учитывали разные типы лексических диалектных отличий. Так, они включили собственно словарные диалектизмы, то есть, слова, корни которых отсутствуют в литературном языке (*кво́лый* – слабый, больной; слова, образованные от общенародных корней, но имеющие иные аффиксы и особые диалектные значения (*сузе́м* – большой массив леса на севере; слова с теми же корнями и значениями, что и в литературном языке, но в ином аффиксальном оформлении (*насдогнать* – догнать; *взгреметься* – начать усиленно греметь); слова с такими особенностями произношения, которые не являются элементами фонетических закономерностей и имеют «индивидуальный», лексикализованный характер (*анбар* – амбар); семантические диалектизмы, то есть слова, в своем оформлении ничем не отличающиеся от слов литературного языка, но имеющие в говорах особые значения (*руда* – кровь, губы – грибы, виски – волосы); фразеологические диалектизмы (*андроны едут, поехали* – о чем-либо вздорном, о небылице и т. п.). [Там же, с. 6]. Таким образом, еще одной категорией параметров поисковой системы

может стать тип диалектизмов: фонетические, грамматические, словообразовательные, собственно лексические, семантические, этнографические, а так же фразеологические.

Подводя итог, можно сказать следующее: основанием для построения поисковой системы применительно к данному лексикографическому источнику могут быть: система помет (как грамматических, так и семантических); сфера функционирования лексики, типы диалектизмов. В случае необходимости у исследователя будет возможность отбирать в кратчайшие сроки массив словарной информации, соответствующий его запросу.

Литература

Крысин Л.П.Толковый словарь иноязычных слов. М., 2000.

Ожегов С.И.Толковый словарь русского языка. М., 1997.

Саженин И.И. Словарный корпус как элемент оптимизации исследовательского и учебного процессов // Информатизация образования – 2011: Материалы Международной научно-практической конференции. Елец, 2011. Т. 1. С. 447–453.

Словарь русского языка: В 4-х т. / Гл. ред. А.П. Евгеньева. М., 1999.

Словарь русских народных говоров / Гл. ред. Ф.П. Филин; ред. Ф.П. Сороколетов; Ин-т русского языка, Словарный сектор АН СССР. Л., 1965–1994. Вып. 1–28.

Филин – Словарь русских народных говоров / Гл. ред. Ф.П. Филин; ред. Ф.П. Сороко-летов; Ин-т русского языка, Словарный сектор АН СССР. Л., 1965. Вып. 1.

Фасмер М. Этимологический словарь русского языка. М., 1986–1987. Т. 1–4.