УДК 81’322=512.31

АВТОМАТИЗАЦИЯ В ЛЕКСИКОГРАФИИ КАК ОСНОВА ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ: ЛИНГВОСТАТИСТИЧЕСКИЙ КОММЕНТАРИЙ НА МАТЕРИАЛЕ БУРЯТСКОГО ХУДОЖЕСТВЕННОГО ТЕКСТА

© ***Дырхеева Галина Александровна***
доктор филологических наук, профессор,
главный научный сотрудник Института монголоведения,
буддологии и тибетологии СО РАН
Россия, 670047, г. Улан-Удэ, ул. Сахъяновой, 6
E-mail: an5dag1@mail.ru

В статье представлен краткий обзор возможностей использования компьютерной лексикографии для лингвистических исследований. Показано, какие типы словарей можно получить, используя современные компьютерные технологии: авторские словари, словари к литературным произведениям, частотные словари, указатели, индексы, конкордансы и др. На примере материалов по бурятскому языку (статистико-комбинаторных характеристик сочетающихся согласных, алфавитно-частотного словаря по прозаическим произведениям бурятского писателя Х. Намсараева, обратного алфавитно-частотного словаря, выявления параметров «богатства» художественного текста, анализа лингвотипологических особенностей языка и др.) показано, какого рода исследования можно провести с помощью автоматической обработки языковых данных. Отмечено, что, к сожалению, эти возможности, несмотря на активный рост информационной поддержки бурятского языка в компьютерной и виртуальной средах в виде корпуса бурятского языка, практически, в настоящее время не используются.
Ключевые слова: автоматизация; лексикография; алфавитно-частотный словарь; бурятский язык; статистика.

Сегодня лингвистические исследования уже невозможно представить без использования различного рода компьютерных технологий. Они представлены во многих языках и охватывают, практически, все многообразие лингвистического анализа. Очевидно, что наиболее широкое применение они нашли в лексикографической области языкознания.

Особую категорию среди многочисленных словарей составляют словари к литературным произведениям, авторские словари, различные словоуказатели, индексы, конкордансы, частотные словари [1-6 и др.]. Такого рода лексикографические описания существенно облегчают работу исследователям, занимающимся изучением творчества того или иного писателя, литературных стилей и жанров, они являются источником разнообразной информации не только литературоведческого, но часто и общеязыковедческого плана. Несомненно, что ценность подобных изданий намного повышается в случае количественной или статистической информации, сопровождающей обследуемый объем текста, что позволяет проводить более глубокий анализ множества лингвистических признаков, характерных и стабильных параметров текста, в которых отражаются стилевые особенности языка автора. Ранее подобного типа словари обычно были результатом кропотливой, трудоемкой многолетней работы. Сегодня современные компьютерные технологии намного облегчили труд составителей подобных словарей.

Необходимо также отметить, что их особенностью является то, что они основываются не на выборке из текстов, а являются перечнем всех слов или всех лингвистических явлений, которые содержатся в данном произведении или слов, которые употребил данный автор в своих работах, а также тем, что указываются «адреса»

этих слов, то есть номера страниц, строк, предложений и другие адресные сведения. Если же к слову приведены все контексты — предложения или более длинные отрывки текста, — то такой список слов называется конкордансом.

В подобных исследованиях одной из наиболее важных задач является определение четких критериев выделения параметров лингвистического и лингвостатистического анализа текстов. Обычно они выделяются интуитивно на основе лингвистических особенностей, лежащих на поверхности текста.

Отправными пунктами при выделении параметров, в первую очередь, являются цели и задачи, которые ставит филолог-исследователь, используя компьютерную обработку данных. Спектр этих задач охватывает, практически, все языковые уровни, начиная от графики и фонетики и заканчивая текстовыми или корпусными массивами. То есть, очевидно, что при составлении подобных словарей наиболее трудный и важный этап работы подготовительный, когда необходимо продумать все детали как организационно-технического плана, так и лингвистического: характер выделяемых единиц, лингвистическая дополнительная информация, которую, возможно, пожелает извлечь потребитель, способ представления материала и т. д.

Так, например, при анализе статистико-комбинаторных характеристик парадигм сочетающихся согласных в бурятском языке (первом опыте использования автоматической обработки данных в бурятоведении) был определен количественный состав сочетающихся согласных в словаре [7]. Ставилась задача выявления функциональной нагрузки фонем сначала в словаре, а, в дальнейшем и в тексте, поскольку система букв и фонем (звуков) в бурятском языке находятся в относительно хорошей корреляции. Проведенный позже графемный анализ текста выявил, что для текста и словаря во всех позициях характерно преимущество гласных твердого ряда. Что касается согласных, то в тексте незначителен перевес сонантов и слабых согласных, а в словаре преимущество за сильно-слабым началом и сонантным концом. Данные результаты, в частности, можно использовать в выявлении типологических особенностей языка, при рационализации бурятской орфографии, при составлении словарей (особенно учитывая распределение букв в начальной позиции) и др.

Очевидно, что, в первую очередь, в подобных исследованиях наибольший интерес представляют лингвотипологические результаты, то есть целью исследования может быть изучение текста как частного случая реализации общей языковой системы. Наиболее эффективными в данном случае являются различного типа словари, полученные в результате компьютерной обработки и сопровождающиеся определенными статистическими показателями. Например, на базе частотного словаря слов и словоформ по прозаическим произведениям классика бурятской литературы Х. Намсараева [8] были выделены такие лингвостатистические типологические критерии как средняя повторяемость слов, средняя длина словоформ, статистическая покрываемость по зонам частотного словаря (в агглютинативном бурятском тексте в отличие флективно-аналитических языков 100 первых словоформ покрывают 23–28% текста (во флективно-аналитических — 43–54%), слова высокочастотной зоны (можно, в частности, отметить, что имена существительные здесь четко образуют тематическую группу слов, в значениях которых проявляется доминация компонентов человеческого микромира и его окружения) и др.

Что касается изучения конкретно текста как особой системы, то интерес может представлять, например, информация о лексико-грамматических классах слов, устойчивых сочетаниях и фразеологизмах, в бурятском языке — о парных словах и т. д.

Так, общеизвестно, что особую языковую систему представляет, практически, язык любого человека. Однако наибольший научный интерес обычно составляет авторский стиль того или иного писателя. А поскольку степень художественности или

образности речи, специфику языка в первую очередь связывают с лексическим составом, словарем, то, очевидно, что чаще эту информацию можно найти в частотных списках или словарях, составленных по произведениям писателя. В частности, данные этих словарей можно использовать для выявления так называемого «богатства» языка писателя. Например, выявлено, что к «языковым константам» «богатства» языка Х. Намсараева относятся: процент покрываемости текста словами высокочастотной зоны (значения К (концентрация словаря)) относительно невысокий, соответственно, высокие значения коэффициента лексического разнообразия, интенсивность использования знаменательных частей речи, высокий параметр «однократные и низкочастотные слова» (или значения индекса исключительности) на уровне словаря. Кроме того, анализ параметров «парные слова», «фразеологические выражения», а также «пословицы и поговорки» подтвердил активность использования данных средств в произведениях писателя.

Полученный материал в виде алфавитно-частотного словаря словоформ был использован в дальнейшем для анализа морфологической структуры бурятского слова. Для этого он был преобразован в обратный алфавитно-частотный список, все словоформы были разделены на морфемы и сопровождены количественными характеристиками. Всего было выделено 5028 корней, 267 суффиксов. Относительно слово– и формообразовательной структуры слова было установлено, что в словаре примерно 10% — корневые слова, 32% состоят из двух морфем, 36% — трех, 16% — четырех, более 5% — из пяти и более морфем. Данная информация позволяет судить о продуктивности или употребительности морфемных элементов, зависимостях распределения определенных классов слов, формообразовательных элементов, и далее о типологии слово– и формообразования, и, соответственно, получения словообразовательного словаря, типологического сравнительного анализа. То есть в некоторой степени способствовать решению грамматической проблематики бурятского языка.

Автоматизация лексикографических работ способствует также решению лингвистических нормативных задач, исследованиям по культуре речи, истории изменения и формирования лексики, методике преподавания языка и т.д. То есть, в целом, можно отметить, что спектр лингвистических проблем и задач, в решении которых возможно использование компьютерных технологий достаточно много. Но, к сожалению, наблюдаемый сегодня рост интереса к информационной поддержке бурятского языка в компьютерной и виртуальной средах, сопровождаемый увеличением различных сайтов, виртуальных сервисов, созданием корпуса бурятского языка, пока не способствовал увеличению числа научных исследований, в которых привлекались бы эти современные средства. Хотелось бы также пожелать расширить ряды частотных списков, охватив тексты других писателей, и тексты наиболее развитых бурятских литературных жанров: публицистику, газетный жанр, а также разговорную речь, что имело бы большое значение для нормализации литературного языка, для исследований по культуре речи и т. д.

Литература

1. Словарь поэтического языка М. Цветаевой: в 4 т. / сост. И. Ю. Беляева, И. П. Оловянникова, О. Г. Ревзина, Рук. О. Г. Ревзина. М.: Дом-музей М Цветаевой, 1996–1999.

2. Словарь языка Пушкина: в 4 т. / гл. ред. В. В. Виноградов; [Предисл. В. В Виноградова]. М., ГИС, 1956–1961. Т. 1–4.

3. Частотный словарь автобиографической трилогии М. Горького / авт.-сост. П. М. Алексеев. СПб.: Изд-во СПбГУ, 1996. 208 с.

4. Частотный словарь рассказов А. П. Чехова / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб.: Изд-во СПбГУ, 1999. 172 с.

5. Частотный словарь романа Л. Н. Толстого "Война и мир" / сост. З. Н. Великодворская, Г. С. Галкина, Г. В. Куперман, В. М. Цаппикова. Тула: Изд-во Тул пед ин-та, 1978. 381 с.

6. Шоу Д. Т. Конкорданс к стихам А. С. Пушкина: в 2 т. М.: Языки рус. культуры, 2000. 1300 с.

7. Дырхеева Г. А. Статистический анализ сочетаемости согласных в бурятском языке // Сегментные и просодические единицы языков байкальского региона. Улан-Удэ: Изд-во БНЦ СОРАН, 1991. С. 77–88.

8. Дырхеева Г. А. Бурятский художественный текст: лингвостатистическое описание (на материале прозы Х. Намсараева). Улан-Удэ: Изд-во БНЦ СО РАН, 2007. 205 с.

AUTOMATIZATION IN LEXICOGRAPHY AS THE LINGWARE BASIS: THE LINGVOSTATISTIC COMMENT ON MATERIAL OF THE BURYAT ART TEXT

Galina A. Dyrkheeva
Doctor of Philology, professor, main research worker of the Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the Russian Academy of Sciences
Russia, 670047, Ulan-Ude, Sakhjyanova St., 6
E-mail: an5dag1@mail.ru

In the article the short review of opportunities of a computer lexicography use for linguistic researches is presented. It is shown, what types of dictionaries can be received, using modern computer technologies: author's dictionaries, dictionaries of literary works, frequency word books, indexes, concordances, etc. On the example of materials on the Buryat language (statistic-combinatory characteristics of the combined concordance, the alphabetic frequency word book on prosaic works of the Buryat writer Kh. Namsaraev, the return alphabetic frequency word book, identification of the wealth parameters of the art text, analysis the typological features of language, etc.) it is shown, what sort of research it is possible to carry out by means of automatic processing of language material. It is noted that, unfortunately, these opportunities, despite the active growth of information support of the Buryat language in computer and virtual environments in the form of the case of the Buryat language, practically, now aren't used.
Keywords: automatization, lexicography, alphabetic frequency word book, Buryat language, statistics.