Чумарина Гузель Раисовна

РЕСУРСЫ И АРСЕНАЛ ЭЛЕКТРОННЫХ КОРПУСОВ В СОВРЕМЕННОЙ ЛЕКСИКОГРАФИИ

В статье рассматриваются современные разработки и использование корпусных словарей на базе информационных технологий. В этой связи изучаются общие принципы построения корпусов, методы их исследования и потенциал корпусной лингвистики. Выделяются и описываются характерные особенности некоторых существующих корпусных словарей татарского языка.

Адрес статьи: www.gramota.net/materials/2/2017/3-1/52.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2017. № 3(69): в 3-х ч. Ч. 1. С. 173-175. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2017/3-1/

УДК 81’32:811.512.145

В статье рассматриваются современные разработки и использование корпусных словарей на базе информационных технологий. В этой связи изучаются общие принципы построения корпусов, методы их исследования и потенциал корпусной лингвистики. Выделяются и описываются характерные особенности некоторых существующих корпусных словарей татарского языка.

Ключевые слова и фразы: корпусный словарь; компьютерные технологии; автоматизированный сбор данных; массив языковых данных; корпусная лингвистика.

Чумарина Гузель Раисовна, к. филол. н.
Казанский инновационный университет имени В. Г. Тимирясова (ИЭУП)
chumarina1@yandex.ru

РЕСУРСЫ И АРСЕНАЛ ЭЛЕКТРОННЫХ КОРПУСОВ В СОВРЕМЕННОЙ ЛЕКСИКОГРАФИИ

Введение электронных корпусных словарей в лингвистическую науку является неизбежным и очень значительным вкладом в развитие языка. Исследования на основе корпусов стали основной методологией, используемой во многих отраслях лингвистики. С 1990-х годов корпусные словари стали применяться в переводе. Рассмотрим потенциал и методы исследований на основе корпусов.

В настоящее время приоритетной целью является создание широкого поля лингвистических знаний в форме лингвистических описаний, как можно более полных и используемых многократно, структурированных в большую базу лингвистических знаний или в разные типы связанных между собой лингвистических баз (базы грамматических знаний, лексических, текстуальных).

После упоминания актуальных тенденций в разных областях, а также определения возможностей информационной лингвистической системы, перейдем к факту, что лексикография (являясь профессией языковой индустрии) имеет очень долгую традицию, и создание языковой базы данных с подобающим содержанием и размерами является очень затратным и требует много времени.

Благодаря многочисленным источникам машинных словарей мы получаем лексическую информацию в многоязыковом контексте, для того чтобы создать единую многоязыковую базу лексических знаний, одним из направлений использования которых являются исследования в области перевода. Однако реализация автоматического перевода сталкивается с определенными препятствиями, которые еще предстоит преодолеть. «Системы машинного перевода текстов с одних естественных языков на другие моделируют работу человека-переводчика. Их эффективность зависит прежде всего от того, в какой степени в них учитываются объективные законы функционирования языка и мышления» [1]. Электронные словари значительно облегчают процесс перевода, но требуют от пользователя определенного знания языка и затрат времени на его осуществление. Следует отметить, что «электронный словарь – это особый лексикографический объект, в котором могут быть реализованы и введены в обращение многие продуктивные идеи, не востребованные по разным причинам в бумажных словарях» [3]. Дальнейшее развитие и возможные перспективы информационных технологий неоспоримо оказывают влияние на лексикографию в целом и на программное обеспечение электронных словарей в частности. «Если при теоретических исследованиях лингвисты действительно редко учитывают реальные возможности вычислительной техники, то выбор оптимальных решений конкретных лингвистических задач в рамках… автоматических словарей в значительной степени зависит именно от уровня развития вычислительной техники» [4, с. 10].

Одним из ведущих направлений современной прикладной и математической лингвистики является корпусная лингвистика, занимающаяся разработкой общих принципов построения и применения лингвистических корпусов с использованием информационных технологий. Исследование национальных языков на достоверном материале с использованием современных компьютерных технологий автоматической обработки текстов позволяет выработать новые подходы к решению актуальных проблем изучения и исследования национальных языков.

Термин «корпусная лингвистика» появился в 1980-х годах. Корпусная лингвистика занимается изучением закономерностей языка на материале больших объемов текстов (корпусов), которые систематизированы, размечены и обработаны в электронной форме. Хотя эта методология была давно известна в бумажном виде. Одним из самых значительных примеров корпусного бумажного словаря является Оксфордский словарь английского языка, опубликованный в бумажном варианте в 1928 г., представляющий из себя собрание 5 млн словарных статей, собранных во второй половине XIX века и первой половине XX века. В настоящее время мы можем увидеть Оксфордский словарь английского языка в электронном виде онлайн [10]. С развитием компьютерных технологий корпусные исследования языка стали появляться в конце 1980-х годов, охватывая все большее количество областей лингвистики и связанных с ней дисциплин. Популярность этой области исследования может быть подтверждена целым рядом книг и статей, опубликованных по данной теме за рубежом. Со временем корпусные словари значительно увеличились в размере. Например, Оксфордский корпус [9] английского языка имеет 2 млрд слов, Корпус Современного американского английского языка – более 400 млн слов [7], Американский Национальный корпус – 22 млн слов [5], а Британский Национальный корпус – более 100 млн слов [6].

Рассмотрим вопрос: что такое корпус? Корпусный словарь определяется как репрезентативная информационно-справочная система, основанная на базе данных информатизированных текстов, собранных с целью их лингвистического анализа. Корпус включает в себя различные типы письменных и устных текстов, представленных в данном языке, различные типы словарей, а также разметку – информацию о свойствах текстов. Анализ корпуса лежит в основе корпусной лингвистики. Корпусная лингвистика не является однородной методологией, она используется с разной степенью детализации и разной опорой на количественные и качественные методы, со следующими особенностями: распознаваемый машиной естественный язык, сбалансированный и репрезентативный проект корпуса, систематический и исчерпывающий анализ. Следующие особенности характерны для работы с корпусами:

- анализ основан на корпусе или корпусах естественного языка, который распознается машиной, следовательно, поиск образцов для исследования осуществляется с помощью информационных технологий;
- корпус должен быть сбалансированным, а также репрезентативным в отношении модальности/записи/разноплановости, на которые нацелено исследование;
- анализ является (или стремится к тому, чтобы быть) систематическим и исчерпывающим. Это означает, что корпус не просто служит базой данных примеров, из которых можно выбрать нужные, а другими можно пренебречь, а весь корпус (или образец корпуса) принимается во внимание. Итак, под лингвистическим, или языковым, корпусом текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, филологически компетентный массив языковых данных, предназначенный для решения лингвистических задач.

Корпус показывает, что является центральным и типичным, нормальным и ожидаемым. Он подчеркивает, что использование языка является очень точным и выверенным, и такое разнообразие языка является не случайным, а когнитивно мотивированным. Все находится в сочетании друг с другом. Другое важное влияние корпусов состоит в том, что они предполагают лингвистический анализ за пределами отдельно взятого слова как основы семантической единицы. Отметим, что лингвистическое описание слов должно включать и конструкции, в которых они появляются.

В целом можно обозначить следующие главные преимущества исследований языка на основе корпусов: уменьшенное количество предположений, гипотез и субъективных выводов; аутентичность базы данных; потенциал для систематической проверки исследовательских гипотез, основанных на более обширном лингвистическом материале. Из недостатков можно отметить проблемы с репрезентативностью и балансом: любые заявления и обобщения, которые мы делаем о выборке языка, которую мы исследуем, не является показателем всего языка.

В качестве методов исследования корпусная лингвистика отдает приоритет наблюдению. Оно классифицируется как составная часть научной деятельности, входящая как необходимый элемент в процедуру всех лингвистических методов и приемов, набор правил выделения из текста или потока речи языковых фактов и включения их в изучаемую категорию или систему. Правила наблюдения формулируют закономерности отбора фактов, установления их признаков, уточнения предмета наблюдения и описания наблюдаемых явлений. Наблюдение является индуктивным приемом исследования, который позволяет установить закономерности при помощи индукции, выведения общего правила из наблюдений над ограниченным количеством фактов, подчиняющихся общему правилу. В то же время это количественный метод, который также интегрирует качественные характеристики для выведения гипотезы о базе данных, предоставляемой корпусом, и для формирования генерализаций об использовании языка. С привлечением корпуса текстов определенного языка возможны более обширные и объективные исследования тех или иных аспектов языка и культуры малочисленных народов. Необходимо уточнить, что корпусы текстов могут состоять из текстов как устной, так и письменной речи. Письменные корпусы включают в себя тексты различных жанров (проза, официальные документы и др.), что позволяет им максимально соответствовать критерию репрезентативности. Для языков, обладающих небольшим объемом литературных источников и письменных памятников, а также для языков малочисленных народов с ограниченным ареалом их употребления составить письменный корпус является довольно сложной задачей. Для таких языков, в частности, «корпус устных текстов… представляется именно тем собранием языковых данных, которое обеспечивает наиболее полное отображение реалий исследуемого языка» [2].

Существует полемика: является ли корпусная лингвистика методологией или теорией? Превалирующий взгляд состоит в том, что это не теория или независимая область лингвистики. Она не определена объектом изучения. Объектом изучения является не исследование корпусов, а скорее исследование языка через корпуса. В основном корпусная лингвистика рассматривается как методология, которая разработала свои собственные систематические методы и принципы применения корпусов для исследований использования языка; следовательно, это методология с «теоретическим статусом», используемая во многих областях и теориях лингвистики [8]. Например, она применяется для описания разных областей языка (дескриптивная лингвистика): в семантике (словосочетания, синонимы), синтаксисе (грамматика на основе корпуса), прагматике (запись вариантов, анализ жанров, стилистики). Открытия в исследованиях на основе корпусов применяются в разных областях лингвистики и теоретических основах: лексикографии (корпусные словари), социолингвистике, прикладной лингвистике (изучение языка), диахронических исследованиях, дискурсном анализе, когнитивной лингвистике, а также в контрастивной и компаративной лингвистике и исследованиях перевода.

Таким образом, ресурсы и анализ корпусов текстов, а также методы исследования корпусной лексикографии являются перспективным направлением лингвистики. Материалы корпусов позволяют оценить весь

спектр языковых явлений представленных текстов, выделить и исследовать особенности языка. Электронные корпуса являются принципиально новым источником, обеспечивающим автоматизированное изучение отдельных черт языка, обеспечивающим перекрестные исследования различных текстов и облегчающим поиск и выборку необходимых данных.

Список литературы

1. **Белоногов Г. Г.** Системы фразеологического машинного перевода политематических текстов [Электронный ресурс]. URL: http://www.a-z.ru/person/belonogov/ (дата обращения: 06.01.2017).
2. **Лемская В. М.** Потенциал применения методов корпусной лингвистики в рамках дескриптивного подхода в исследовании чулымско-тюркского языка [Электронный ресурс]. URL: http://psibook.com/linguistics/potentsial-primeneniya-metodov-korpusnoy-lingvistiki-v-ramkah-deskriptivnogo-podhoda-v-issledovanii-chulymsko-tyurkskogo-yazyka.html (дата обращения: 06.01.2017).
3. **Селегей В.** Электронные словари и компьютерная лексикография [Электронный ресурс]. URL: http://www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 06.01.2017).
4. **Семенов А. Л.** Современные информационные технологии и перевод. М.: Академия, 2008. 224 с.
5. **American National Corpus** [Электронный ресурс]. URL: http://www.americannationalcorpus.org/ (дата обращения: 06.01.2017).
6. **British National Corpus (BYU-BNC)** [Электронный ресурс]. URL: http://corpus.byu.edu/bnc/ (дата обращения: 06.01.2017).
7. **Corpus of Contemporary American English** [Электронный ресурс]. URL: http://corpus.byu.edu/coca/ (дата обращения: 06.01.2017).
8. **McEnery T., Wilson A.** Corpus Linguistics: An Introduction 2nd edition. Edinburg: Edinburg University press, 2001. 235 p.
9. **Oxford English Corpus** [Электронный ресурс]. URL: https://www.sketchengine.co.uk/oxford-english-corpus/ (дата обращения: 06.01.2017).
10. **Oxford English Dictionary** [Электронный ресурс]. URL: http://www.oed.com/ (дата обращения: 06.01.2017).

RESOURCES AND ARSENAL OF ELECTRONIC CORPUSES IN CONTEMPORARY LEXICOGRAPHY

Chumarina Guzel' Raisovna, Ph. D. in Philology
Kazan Innovative University named after V. G. Timiryasov (IEML)
chumarina1@yandex.ru

The article deals with the modern developments and the use of corpus dictionaries on the basis of information technologies. In this context the paper studies the general principles of corpuses construction, methods of their research and the potential of corpus linguistics. The author singles out and describes the characteristics of some existing corpus dictionaries of the Tatar language.

Key words and phrases: corpus dictionary; computer technologies; automated data collection; language data corpus; corpus linguistics.

УДК 808.2-311

В статье впервые рассматриваются общеизвестные и узколокальные названия сельских поселений и микротерриторий Иркутской области, образованные от нарицательных имен. Исследуются процессы семантической, грамматической онимизации слов, словосочетаний и онимизация апеллятивной лексики с одновременной трансонимизацией антропонимов. Приводятся данные относительно продуктивности указанных способов в образовании топонимии Иркутской области, устанавливаются общеславянские и оригинальные черты, свойственные анализируемым названиям.

Ключевые слова и фразы: онимизация; семантическая онимизация; грамматическая онимизация; топоним; микротопоним; топонимия; апеллятив; апеллятивное сочетание.

Чупановская Мария Николаевна, к. филол. н., доцент
Иркутский государственный университет
mariya-chupanovskaya@yandex.ru

АПЕЛЛЯТИВНАЯ ЛЕКСИКА В РУССКИХ ТОПОНИМИЧЕСКИХ НАЗВАНИЯХ ИРКУТСКОЙ ОБЛАСТИ

Топонимическая система Восточной Сибири, на наш взгляд, заслуживает пристального внимания со стороны филологов, поскольку географические названия Иркутской области не только отражают расселение народов на данной территории, культурно-исторический фон эпох, мировосприятие сибиряков-старожилов, межъязыковые контакты (контакты между автохтонными народами и русским населением), но и лингвистические особенности, демонстрирующие черты местного пользования языком.