FOLIA 260

Annales Universitatis Paedagogicae Cracoviensis

ISSN 1689-9911
DOI 10.24917/16899911.11.2
Studia Russologica 11 (2018)

Izabela Kozera

Метод корпусного исследования – преимущества и недостатки
(на примере использования Национального корпуса русского языка)

Введение

Под *корпусом текстов* обычно подразумевается компьютерная коллекция естественных текстов, представляющих письменную или устную речь разных вариантов, функциональных стилей и типов[1]. Указанный массив языковых данных во многом отличается от сборника текстов, собранного случайно («[…] a corpus is not a random collection of text»[2]). Главным признаком корпуса является его представительность, относящаяся к разнообразию языка, отраженного в соответствующих, естественных пропорциях. Коллекция текстов должна быть сбалансированной и иметь достаточную выборку по числу текстов и авторов, чтобы служить основой для статистически достоверных исследований лингвистических феноменов[3]. Сегодня корпусные исследования широко применяются в различных научных областях, таких как лексикография, преподавание иностранных языков, транслатология (так наз. параллельные корпуса), психолингвистика, социолингвистика и др. С помощью корпусов представители *диванной лингвистики* (англ. *armchair linguistics*) получают иллюстративный материал для проверки своих гипотез. Недаром в последнее время у противников корпусной лингвистики появилось определение: *computer-aided armchair linguistics*[4], которое явно несет отрицательный оттенок[5]. В настоящей статье предпринимается попытка осветить вопрос

[1] *Podstawy językoznawstwa korpusowego*, red. B. Lewandowska-Tomaszczyk, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2005, c. 11.

[2] G. Aston, L. Burnard, *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh 1998, c. 21.

[3] С.А. Шаров, *Представительный корпус русского языка в контексте мирового опыта*, «Научно-техническая информация. Серия 2. Информационные процессы и системы» 2003, № 6, с. 10.

[4] C.J. Fillmore, *Corpus linguistics vs. computer-aided armchair linguistics*, [in:] *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, Stockholm 1992, http://is.muni.cz/el/1421/jaro2008/FJ0B738/um/Corpus_linguistics_verze1.pdf [доступ: 2.06.2017].

[5] Там же.

о целесообразности использования метода корпусного исследования. Для указанной цели представлены несомненные преимущества и недостатки этого подхода, проиллюстрированные на материале Национального корпуса русского языка.

1. Извлечение информации с помощью корпуса – преимущества и недостатки

1.1. Большой массив данных

Начиная с 80-тых годов «стали активно появляться корпусные проекты различных масштабов на разных языках и для разнообразных целей»[6]. По словам В.В. Мамонтовой: «В среде учёных появилось понимание того, что ряд корректных лингвистических исследований возможно провести только на большом речевом материале»[7]. Доступ к огромному количеству языковых данных несомненно стал толчком в сторону все большего использования корпуса в качестве достоверного источника языковой информации. Существенным является факт, что сторонники разных теорий и направлений имеют право проверять свои гипотезы на корпусном материале. Однако следует помнить что, пользуясь корпусами, сторонник когнитивизма, структурализма, исследователь антропологии или истории языка представляет в первую очередь свою теорию, так как лишь она, в отличие от эмпирических данных, является для него источником гипотез[8]. Итак, большая коллекция текстов в электронном виде, собранная в одном месте и упорядоченная по избранным параметрам, может послужить для получения иллюстративного материала. Корпусы позволяют проверять интуицию, показывая частотность разных языковых конструкций, а также выявляя образцы сочетаемости слов[9]. Следует заметить, что представленный метод указывает на корпус как на способ извлечения эмпирического материала, своего рода метод поиска. Несомненно, достоинство корпуса заключается в значительном упрощении и ускорении процедуры лингвистической обработки больших массивов текстов. Сторонники лингвистических корпусов обращают внимание не только на аспект поиска в большом речевом материале. Они утверждают, что благодаря корпусу

> […] исследователь получает представление об относительной значимости различных явлений. Такая значимость оценивается при помощи простого количественного коррелята – частоты. Корпусная идеология состоит в том, что языковое явление тем важнее, чем чаще оно встречается в естественном употреблении[10].

---

[6] О.В. Нагель, *Корпусная лингвистика и ее использование в компьютеризированном языковом обучении*, «Язык и культура» 2008, т. 1, № 4, с. 53–54.

[7] В.В. Мамонтова, *Корпусная лингвистика в современной языковедческой парадигме*, «Актуальные вопросы современной науки» 2010, № 12, с. 232.

[8] A. Pawłowski, *Lingwistyka korpusowa – perspektywy i zagrożenia*, «Polonica» 2003, t. XXII–XXIII, с. 22.

[9] *Podstawy językoznawstwa korpusowego*…, с. 9–12.

[10] *Рассказы о сновидениях: Корпусное исследование устного русского дискурс*а, ред. А.А. Кибрик, В.И. Подлесская, Изд. Языки славянских культур, Москва 2009, с. 27.

На указанную пользу скептики обычно приводят аргумент о сомненной представительности корпусных данных. Лингвист, занимающийся корпусным исследованием, должен учесть факт, что его анализы соответствуют лишь тем явлениям, которые зафиксированы в корпусе. «Явления, не встретившиеся в корпусе, не получают отражения в описании»[11]. Поэтому очень важным является критерий репрезентативности корпусных данных, которая «достигается за счет наиболее полного и сбалансированного охвата всех релевантных для конкретного исследования типов продуктов речевой деятельности»[12]. Ведь «корпус часто рассматривается как некая уменьшенная модель языка или подъязыка»[13].

> Электронные корпусы позволяют не только быстрее и эффективнее решать уже давно стоящие перед наукой задачи, но и выдвигать новые, которые раньше были фактически невыполнимы из-за своей трудоемкости[14].

Существует однако угроза слишком радикального эмпиризма. Корпусному исследованию подлежит только язык, зафиксированный в текстах, составляющих корпус. Тексты могут содержать различного рода ошибки. Поэтому некую опасность представляет собой изучение языка без каких-нибудь предварительных предположений о его структуре. Лингвист может заблудиться в сложном массиве данных. Поэтому некой угрозой является радикальный эмпиризм, лишенный научной рефлексии, а лишь сосредоточенный на поверхностном анализе[15].

1.2. Разметка

Корпус не является лишь коллекцией текстов с системой автоматического поиска. Роль корпуса значительно растет с момента добавления разметки, так как он больше не является просто совокупностью текстов. Итак, так наз. *размеченный корпус* является источником лингвистической информации, которая внесена в результате обработки исходного текста[16] и явно выражена рядом с ним. Под термином *разметка* (англ. *tagging*, *annotation*) подразумевается приписывание текстам и их компонентам специальных помет, иначе тегов (содержащих информацию о лексемной принадлежности), причем употребляются они в соответствии с известной для пользователя схемой. Подход к аннотированию языковых единиц является самым существенным

[11] Там же.

[12] А.М. Лаврентьев, *Влияние корпусных технологий на развитие диахронической лингвистики: пример Франции*, https://halshs.archives-ouvertes.fr/halshs-01071863/document [доступ: 11.06.2017].

[13] М.Ю. Колокольникова, *Дискурс-анализ и корпусный анализ в исследованиях в области исторической лексикологии*, «Известия Саратовского университета» 2010, Журналистика, т. 10, вып. 2, http://www.philology.ru/linguistics1/kolokolnikova-10.htm [доступ: 11.06.2017].

[14] Там же.

[15] B. Habert, A. Nazarenko, A. Salem, *Les linguistiques de corpus*, Armand Colin, Paris 1997, с. 216, цит. за: A. Pawłowski, *Lingwistyka*..., с. 21.

[16] С.А. Шаров, *Представительный*..., с. 12.

вопросом, в котором выделяются две крайние точки зрения. Первый, «формально-морфологический» способ, позволяет приписать каждой словоформе некий ярлык независимо от особенностей ее употребления. Указанное может оказаться проблематическим в случае омонимичных словоформ и ввести малоопытного пользователя корпуса в заблуждение. Формально-морфологический способ применяется в случае систем автоматического аннотирования, что «позволяет разметить огромные массивы текстов без участия человека (программа-парсер приписывает информацию, руководствуясь электронными морфологическими словарями-указателями словоформ)»[17]. Второй подход намного сложнее, так как дает более подробную семантическую информацию о словоформе, учитывая ее употребление. Указанное невозможно без постоянного участия опытного лингвиста, поэтому не является подходящим для корпусов больших объемов.

Выделяется несколько типов разметок в зависимости от содержащейся в ней информации. Экстралингвистическая разметка (так наз. метаразметка) заключается в прибавлении сведений об авторе и тексте (заглавие, год и место издания, жанр, тематика). Структурная разметка выделяет такие метатекстовые элементы, как: главы, абзацы, предложения, словоформы. Однако наиболее ценной является так наз. лингвистическая разметка, так как она делает корпус источником информации о языковой системе с помощью лингвистической терминологии.

Лингвистическая разметка подразделяется на несколько типов. Первый из них представляет собой морфологическая разметка (иначе: *частеречная разметка*, англ. *part-of-speech tagging*, *POS-tagging*), которая обычно состоит из следующих элементов: лемма, признак части речи, признаки грамматических категорий. Каждый корпус имеет свою частеречную разметку в зависимости от использованного языка и способа формализации (см. дальше). Более сложной задачей является составление семантической разметки, так как среди лингвистов не существует общее мнение на тему того, какие семантические характеристики должны быть аннотированы.

Анализируемый признак, отличающий корпус от обычной совокупности текстов, также обладает некими недостатками. Во-первых, подготовка грамматического описания единиц, составляющих корпус, является очень сложной задачей, обусловленной типом языковой системы. Во-вторых, несмотря на факт, что «пока не существует компьютерных программ, которые были бы способны заменить человека»[18], каждая разметка искажена субъективностью эксперта и требует дальнейшей проверки другим лингвистом.

1.3. Метод корпусного исследования

Метод корпусного исследования состоит из нескольких шагов. На первом этапе в корпусе делается запрос. Поиск осуществляется в зависимости от выбранных параметров. Следующим шагом является создание базы данных:

---

[17] О.Н. Ляшевская, В.А. Плунгян, Д.В. Сичинава, *О морфологическом стандарте Национального корпуса русского языка*, [в:] *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*, Изд. Индрик, Москва 2005, с. 112.

[18] Там же.

полученные выдачи (словоформы/леммы вместе с информацией о частоте) помещаются в таблицу (обычно для этого используется формат Excel), где подлежат дальнейшей ручной обработке. Необходимым является корректирующий этап работы по формированию базы данных: проверка их достоверности, а также устранение некоторых технических особенностей выдачи материала в данном корпусе. С помощью данных относительно частоты языковых единиц возможен количественный анализ. В свою очередь, качественный анализ является следствием самостоятельного умственного труда исследователя. Достоверность метода корпусного исследования во многом зависит от использованного корпуса и его технических особенностей поиска. В свою очередь, степень представительности корпусных данных влияет на возможность приходить к общим выводам, не только имеющим отношение к корпусу, но и к языку в целом. Поэтому в дальнейшей части статьи приводятся примеры на материале Национального корпуса русского языка.

2. Национальный корпус русского языка (НКРЯ)

Корпусная лингвистика в России начинает свое развитие лишь в 80-тые годы. Старейшим русскоязычным корпусом стал Упсальский корпус русских текстов, который в дальнейшем был включен в Тюбингенский корпус, появившийся в Интернете в открытом доступе[19]. Следующие русскоязычные корпусы были созданы с учетом конкретных научных целей. В течение 2000–2002 годов был подготовлен Корпус газетных текстов (КГТ), затем Хельсинкский аннотированный корпус (ХАНКО), предназначенный прежде всего для учебных целей. Самым большим достижением корпусной лингвистики в России следует считать появление двух представительных корпусов: Национального корпуса русского языка (см. дальше) и Национального корпуса русского литературного языка, которые представляют настолько разнообразную, морфологически аннотированную коллекцию текстов, что могут стать основой многих лингвистических исследований.

НКРЯ, доступный на сайте www.ruscorpora.ru, был создан в Институте русского языка им. В.В. Виноградова РАН. Корпус включает подлинные тексты с половины XVIII в. к началу XXI в. В данный момент НКРЯ состоит из 12 подкорпусов: основной, синтаксический, газетный (подкорпус СМИ 2000-х годов, региональный подкорпус), параллельные корпуса, обучающий, диалектный, поэтический, устный, акцентологический, мультимедийный, исторический, а также новейший подкорпус мультиПАРК (мультимедийный параллельный корпус)[20].

Преимуществом НКРЯ является очень простой интерфейс пользователя. Запрос осуществляется им автоматически путем выбора соответствующих признаков, определенных разметкой (см. дальше). Указанный способ поиска

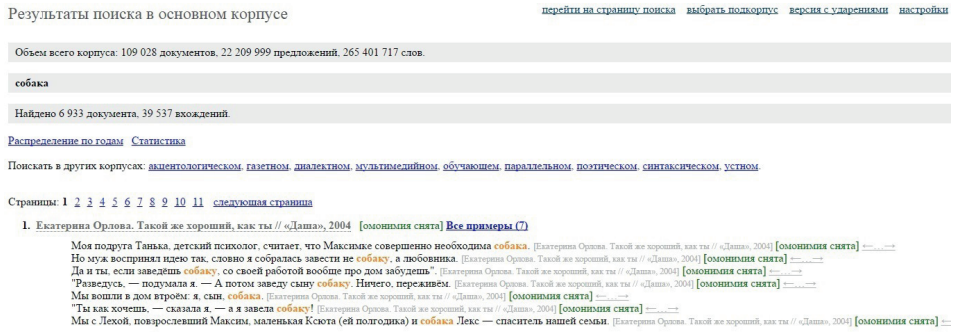
---

[19] М.В. Копотев, А. Мустайоки, *Современная корпусная русистика*, [в:] *Инструментарий русистики: корпусные подходы*, ред. А. Мустайоки, Л.А. Бирюлин, Е.Ю. Протасова, Helsinki University Press, Helsinki 2008, с. 13.

[20] Пилотный вариант мультимедийного корпуса (МультиПАРК) был открыт лишь 8 июля 2015 г. и включает в себя две театральные постановки и одну экранизацию пьесы Н.В. Гоголя «Ревизор».

не требует от пользователя знаний на тему синтаксиса запроса или продвинутых информационных умений. Поиск возможен относительно точных форм, а также слов, соответствующих выбранным грамматическим параметрам. Результаты запроса, содержащие интересующие пользователя единицы, демонстрируются в виде конкорданса (в «обычном формате» и в формате KWIC с указанным словом-ключом посередине – см. Таб. 1А и Таб. 1Б).

Таблица 1А. Результаты запроса в НКРЯ – «обычный формат»



Результаты поиска в основном корпусе

перейти на страницу поиска | выбрать подкорпус | версия с ударениями | настройки

Объем всего корпуса: 109 028 документов, 22 209 999 предложений, 265 401 717 слов.

собака

Найдено 6 933 документа, 39 537 вхождений.

Распределение по годам | Статистика

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, синтаксическом, устном.

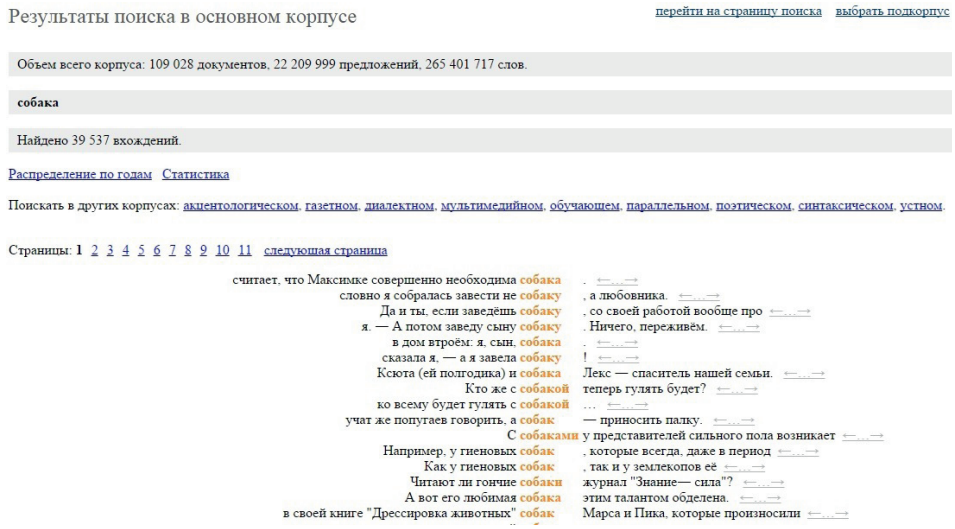
Страницы: **1** 2 3 4 5 6 7 8 9 10 11 следующая страница

1. Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004 [омонимия снята] **Все примеры (7)**

Моя подруга Танька, детский психолог, считает, что Максимке совершенно необходима **собака**. [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
Но муж воспринял идею так, словно я собралась завести не **собаку**, а любовника. [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
Да и ты, если заведёшь **собаку**, со своей работой вообще про дом забудешь". [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
"Разведусь, — подумала я. — А потом заведу сыну **собаку**. Ничего, переживём. [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
Мы вошли в дом втроём: я, сын, **собака**. [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
"Ты как хочешь, — сказала я, — а я завела **собаку**! [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]
Мы с Лехой, повзрослевший Максим, маленькая Ксюта (ей полгодика) и **собака** Лекс — спаситель нашей семьи. [Екатерина Орлова. Такой же хороший, как ты // «Даша», 2004] [омонимия снята]

Источник: www.ruscorpora.ru [доступ: 17.12.2017].

Таблица 1Б. Результаты запроса в НКРЯ – формат KWIC



Результаты поиска в основном корпусе

перейти на страницу поиска | выбрать подкорпус

Объем всего корпуса: 109 028 документов, 22 209 999 предложений, 265 401 717 слов.

собака

Найдено 39 537 вхождений.

Распределение по годам | Статистика

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, синтаксическом, устном.

Страницы: **1** 2 3 4 5 6 7 8 9 10 11 следующая страница

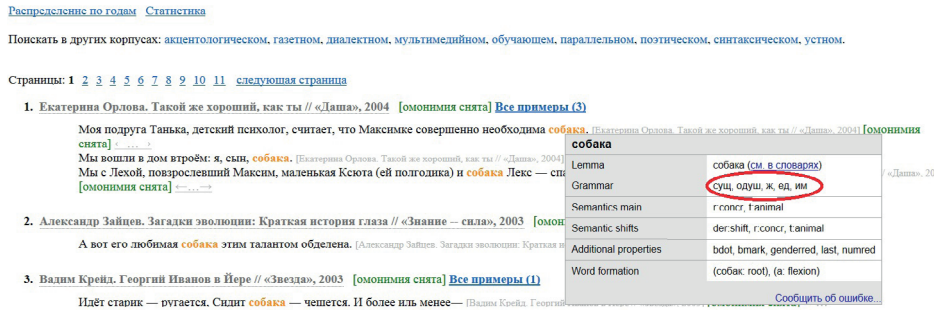
считает, что Максимке совершенно необходима **собака**	.
словно я собралась завести не **собаку**	, а любовника.
Да и ты, если заведёшь **собаку**	, со своей работой вообще про
я. — А потом заведу сыну **собаку**	. Ничего, переживём.
в дом втроём: я, сын, **собака**	.
сказала я, — а я завела **собаку**	!
Ксюта (ей полгодика) и **собака**	Лекс — спаситель нашей семьи.
Кто же с **собакой**	теперь гулять будет?
ко всему будет гулять с **собакой**	…
учат же попугаев говорить, а **собак**	— приносить палку.
С **собаками**	у представителей сильного пола возникает
Например, у гиеновых **собак**	, которые всегда, даже в период
Как у гиеновых **собак**	, так и у землекопов её
Читают ли гончие **собаки**	журнал "Знание— сила"?
А вот его любимая **собака**	этим талантом обделена.
в своей книге "Дрессировка животных" **собак**	Марса и Пика, которые произносили

Источник: www.ruscorpora.ru [доступ: 17.12.2017].

Принятый в НКРЯ подход к аннотированию представляет собой «золотую середину» среди представленных в пункте 1.2. трактовок. По мнению создателей корпуса, своего рода баланс представляет собой разметка словоформ на основе традиционной словарной грамматики русского языка, а омонимичным формам присваивается только одна характе-

ристика[21]. Однако учитывая сложность омонимии, разработчики НКРЯ решили включить в корпус два вида размеченных текстов – со снятой и неснятой омонимией, которые представляют собой два подкорпуса. Разметка подкорпуса с неснятой омонимией осуществляется автоматически, поэтому поиск в нем дает намного больше результатов, в отличие от поиска в подкорпусе со снятой омонимией[22]. Поскольку разметка подкорпуса со снятой омонимией требует частичного участия человека, она не лишена ошибок в такой же степени, как и автоматическая. Автоматическая морфологическая разметка осуществляется с помощью парсеров, т.е. программ со встроенными словарями (в случае НКРЯ образцовым является *Грамматический словарь русского языка* А.А. Зализняка[23]). Разметка корпуса с неснятой омонимией использует программу *mystem* компании Яндекс, которая производит морфологический анализ текста на русском языке, умеет также приводить слова к начальной форме и строить их гипотетические разборы[24]. Каждой словоформе приписывается набор грамматических признаков, инвентарь которых был разработан в 2001–2004 годов О.Н. Ляшевской, В.А. Плунгяном, Д.В. Сичинавой, Г.И. Кустовой и А.Е. Поляковой. В НКРЯ ожидаемая информация в виде отдельной таблички появляется лишь в момент нажатия на конкретную словоформу (см. Таб. 2). Такой способ полезен для пользователя корпуса, так как полученные контексты речи не перегружены избыточными сведениями. Пользователь сам решает, когда ему нужна более точная информация о грамматических признаках лексемы. Таким образом возможно быстрое извлечение лингвистической информации.

Таблица 2. Морфологическая разметка в Национальном корпусе русского языка[25]



Источник: www.ruscorpora.ru [доступ: 17.12.2017].

---

[21] О.Н. Ляшевская, В.А. Плунгян, Д.В. Сичинава, *О морфологическом...*, с. 113.

[22] Там же, с. 114.

[23] А.А. Зализняк, *Грамматический словарь русского языка*, Русский язык, Москва 1977 и его новое издание 2003 года.

[24] По содержанию официального сайта компании Яндекс: https://tech.yandex.ru/mystem [доступ: 2.01.2017].

[25] Грамматические признаки зачеркнуты автором.

Для семантической разметки НКРЯ характерно небольшое количество таксономических классов для каждой из частей речи. Указанный подход объясняется, во-первых, стремлением к быстрой выдаче результатов, во-вторых, упрощением интерфейса пользователя, которому легче разобраться в принятой классификации, если она представлена в одном окне компьютера[26]. Семантическая разметка НКРЯ позволяет искать словоформы, отвечающие созданной семантической классификации[27]. В ней выделяется: таксономия (тематический класс лексемы), мереология, топология, каузация, служебный статус, оценка. Следует заметить, что вышеуказанные классы лексем не имеют ярких границ. Редко лексема принадлежит лишь к одному классу, что часто приводит к разным ошибкам в полученном материале (см. дальше).

Кроме морфологической и семантической разметок, НКРЯ обладает еще метатекстовой разметкой. Метатекстовая разметка обеспечивает поиск по названию текста, его автору, полу и году рождения автора, литературному жанру (нехудожественные тексты и тексты художественной литературы), по типу текста, по месту и времени описываемых событий, сфере функционирования и по тематике.

> В настоящее время перед разработчиками корпуса стоит задача повышения точности разметки и снижения уровня «шума» в результатах поиска. […] При этом разрабатываемые методы должны быть относительно просты, не требовать ручной разметки большого массива данных[28].

3. Преимущества и недостатки работы с корпусом текстов на примере исследования глагола по данным НКРЯ

В настоящей подглаве представлен пример анализа материала, полученного из Основного подкорпуса НКРЯ. Указанная часть корпуса является наиболее представительной, сбалансированной с точки зрения выборки текстов и их жанрово-стилистической разнообразности. Несмотря на преимущества, вытекающие из теоретических положений создателей НКРЯ, указанный корпус обладает также некими недостатками, которые возникают лишь в момент его практического использования. Для иллюстрации некоторых трудностей работы с этим корпусом послужило исследование глагола. Среди всех частей речи глагол является самым сложным классом лексем не только с точки зрения набора грамматических категорий, но и с учетом степени полисемии. Существенным оказался вопрос, насколько корпус сможет справиться с описанием глагола, учитывая формальный и семантический уровни этой части речи, а также чисто технические особенности выдачи результатов.

---

[26] Б.П. Кобрицов, О.Н. Ляшевская, С.Ю. Толдова, *Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей*, http://download.yandex.ru/IMAT2007/kobricov.pdf [доступ: 11.06.2017].

[27] См. сайт: http://www.ruscorpora.ru/corpora-sem.html [доступ: 11.06.2017].

[28] Б.П. Кобрицов, О.Н. Ляшевская, С.Ю. Толдова, *Снятие…*

3.1. Особенности поиска глагольных форм по грамматическим, семантическим и морфемным параметрам

На первом шагу вследствие автоматического запроса по глаголу пользователь получает результаты: контексты речи, леммы[29] и словоформы. НКРЯ демонстрирует результаты поиска в виде последующих подстраниц с указанными частотными данными найденных словоформ и лемм. Табличка с подсчетом результатов приводится отдельно для каждой подстраницы. Большим недостатком НКРЯ является возможность скачать лишь ограниченное количество результатов запроса, что сводится к формулировке «скачать несколько первых результатов выдачи в формате Excel, OpenOffice Calc, XML»[30]. На практике придуманное ограничение позволяет пользователю собрать максимально до 5 тысяч результатов. Указанное количество не является представительной эмпирической базой для научного исследования, так как в случае поиска глаголов полная выдача составляет свыше двух миллионов результатов. Неким выходом является поиск глаголов по более узким параметрам (см. дальше) или использование специализированного скрипта (необходима тогда помощь программиста).

Поиск в НКРЯ обеспечивает выбор грамматических параметров среди всех глагольных категорий: числа, наклонения, времени, лица, залога, вида. Возможен выбор формы глагола. Учитывается категория переходности / непереходности и экспоненты именных категорий, характерные для некоторых глагольных форм (напр. выбор граммем категории рода). Способ выбора грамматических признаков является интуитивным и не представляет собой особых трудностей для стандартного пользователя (см. Таб. 3).

Проблемы могут возникнуть лишь в момент получения результатов, несовсем подходящих для нашего поиска. Итак, наиболее затруднительным для глагола является поиск видовых форм. Разработчиками корпуса принят подход, в котором категория вида глагола имеет переходный характер: от словоизменительного к словоклассифицирующему. Вопреки общепринятой грамматической трактовке, формы глагола, имеющие своего видового партнера, оформлены как слова, образованные от двух инфинитивных форм. Итак, форма *снял* воспринимается корпусом как член парадигмы от инфинитивов *снять* и *снимать*. По этой причине в случае парных глаголов предлагается задать два отдельных запроса – на глагол несовершенного вида и на глагол совершенного вида. Из каждого поиска следует вручную отобрать примеры нужного вида с целью получения наиболее представительной глагольной базы. Однако учитывая технические особенности поиска в НКРЯ, заранее известно, что больше примеров для обоих видов получается при запросе инфинитива в несовершенном виде[31]. Аналогично к интересным выводам

---

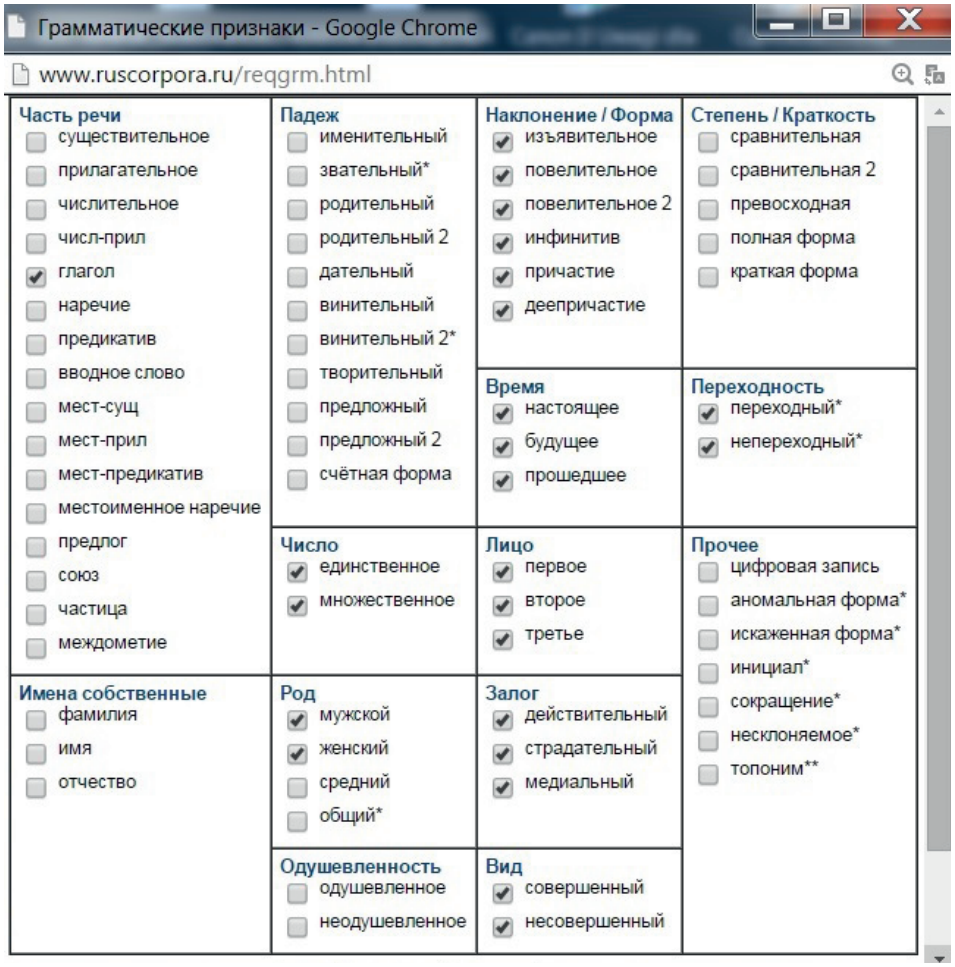
[29] «Лемма – «аналог» лексемы, результат автоматического сведения текстоформ к начальной форме», см. М.В. Копотев, Л. Янда, *Национальный корпус русского языка (www.ruscorpora.ru)*, «Вопросы языкознания» 2006, № 5, с. 150.

[30] По содержанию сайта НКРЯ: http://www.ruscorpora.ru [доступ: 12.06.2017].

[31] По содержанию http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=73&Itemid=84 [доступ: 11.06.2017].

можно прийти на основании поиска причастий в НКРЯ. Согласно исследованию, проведенному С.А. Ковалем на корпусном материале, причастные формы большинства частотных русских глаголов практически не употребляются[32].

Таблица 3. Грамматические признаки для глагола в НКРЯ



Источник: www.ruscorpora.ru [доступ: 17.12.2017].

Семантическа разметка НКРЯ для глагола включает характеристику в областях: таксономии, каузации, служебности этой части речи (фазовые и служебные каузативные глаголы), словообразования. Среди глагольных семантических признаков выделяются следующие: движение (в том числе изменение положения тела, части тела), помещение объекта, физическое

[32] С.А. Коваль, *Роль корпуса в создании реалистичных моделей словоизменительной морфологии*, [в:] *Труды международной конференции «Корпусная лингвистика – 2006»*, Санкт-Петербург 2006, цит. за М.В. Копотев, А. Мустайоки, *Современная...*, с. 17.

воздействие (в том числе: создание физического объекта, уничтожение), изменение состояния или признака, бытийная сфера (существование/начало существования/прекращение существования), местонахождение (в том числе положение тела в пространстве), контакт и опора, посессивная сфера, ментальная сфера, восприятие, психическая сфера (в том числе эмоция, воля), речь, поведение человека, физиологическая сфера, природное явление, звук, свет, запах. Итак, каждый глагол, у которого можно выделить разные контекстуальные значения, получает в корпусе семантические пометы, свидетельствующие о принадлежности к семантическим типам, напр. *пилить* (*бревно*) – «физическое воздействие», *пилить* (*мужа*) – «речь»; (*вьюга) разбушевалась* – «природное явление», (*сосед) разбушевался* – «поведение человека»[33]. При поиске каждому слову приписываются сразу все пометы, которые были помещены в словаре корпуса. Пытаясь противостоять глагольной многозначности, создатели корпуса все время работают над семантическими фильтрами, с помощью которых корпус сможет продемонстрировать лишь те признаки, которые обнаруживаются в данных контекстуальных условиях (принцип контекстной однозначности)[34]. К сожалению, глагольная многозначность все еще присутствует в корпусе. Причиной является не только система автоматического аннотирования, которого приводит к неправильному разделению глаголов. Недостатком являются выделенные таксономические классы, необладающие строгими границами. Итак, например семантический компонент ДВИЖЕНИЕ входит одновременно как в класс «движение», так и в класс «помещение объекта». Поиск глаголов в НКРЯ показывает, что согласно семантической разметке они могут подходить одновременно для нескольких семантических классов, напр. *вдалбливать* («физическое воздействие», «эмоция»), *взвешивать* («помещение объекта», «эмоция»), *взрывать* («физическое воздействие», «уничтожение», «прекращение существования», «эмоция»), *воссоздавать* («начало существования», «эмоция»), *вслушиваться* («восприятие», «эмоция»), *вымачивать* («изменение состояния или признака», «эмоция», «физическое воздействие») и т.п. Не все глаголы получают в корпусе однозначную семантическую характеристику – в таком случае они обладают пометой «disamb», что является сокращением для английского слова *disambiguation*. Данное определение употребляется в случае семантической неоднозначности слова, которое трудно классифицировать по выделенным признакам. Указанный факт свидетельствует о необходимости усовершенствования семантической разметки создателями НКРЯ.

С точки зрения словообразования, корпус обеспечивает поиск приставочных глаголов, семельфактивов и вторичных имперфективов (с суффиксами -ива-, -ва-, -а-). Автоматический морфемный разбор по формальным показателям нередко ведет к недостоверным результатам. Примером может послужит

---

[33] Г.И. Кустова, С.Ю. Толдова, *Национальный корпус русского языка: Семантические фильтры для разрешения многозначности глаголов*, [в:] *Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы*, Санкт-Петербург 2009, с. 259, http://ruscorpora.ru/sbornik2008/12.pdf [доступ: 11.06.2017].

[34] Там же.

поиск вторичных имперфективов – глаголов с приставочно-суффиксальной морфематикой. Среди результатов по запросу «V d:impf d:impf» пользователь может найти простые немотивированные глаголы (напр. *жить*, *писать*), бесприставочные глаголы (напр. *ловить*, *стрелять, сматривать*), глаголы без обязательного суффикса вторичной имперфективации (напр. *встречаться*), а также ряд глаголов, которые только ошибочно могут быть признаны приставочными (напр. *продавать*).

В зависимости от выбора вышеописанных параметров, результаты поиска скачиваются в виде электронной таблицы Excel и на следующем шагу должны подлежать ручной обработке по формированию базы данных. Корректирующий этап работы является существенным для дальнейшего исследования материала: выдачи (прежде всего словоформы и леммы вместе с информацией о частоте) проверяются с точки зрения достоверности и наличия ошибок. К примеру следует привести три таблицы, иллюстрирующие вышеуказанные этапы анализа. Таб. 4А указывает на контексты в формате KWIC, полученные при поиске вторичных имперфективов (запрос «V d:impf d:impf») на сайте НКРЯ. Таб. 4Б показывает те же результаты поиска, помещенные в таблицу Excel (каждая словоформа помещается в отдельную колонну). Таб. 4В указывает на результаты работы по устранению возможных ошибок эмпирического материала.

Созданная таким образом база данных подлежит дальнейшим анализам – количественному и качественному, которые являются основными компонентами каждого корпусного исследования. В зависимости от выбранных параметров относительно поиска глагольных форм, корпус дает широкий спектр возможностей с привлечением статистически более представительного материала. Об указанном свидетельствует ряд научных работ, посвященных в частности: выделению частотных русских глаголов[35], исследованию глагольной сочетаемости в сравнении со словарной нормой[36], выявлению системных различий между языками на основании параллельных подкорпусов[37], стилометрическому анализу текстов известных писателей[38]. На самом деле, указанными вопросами исследователи занимались задолго до возникновения технологий и программ автоматической обработки текста. Однако «корпусные методы позволяют сделать такие исследования более аккуратными и тонкими»[39].

---

[35] См. С.А. Коваль, *Роль…*

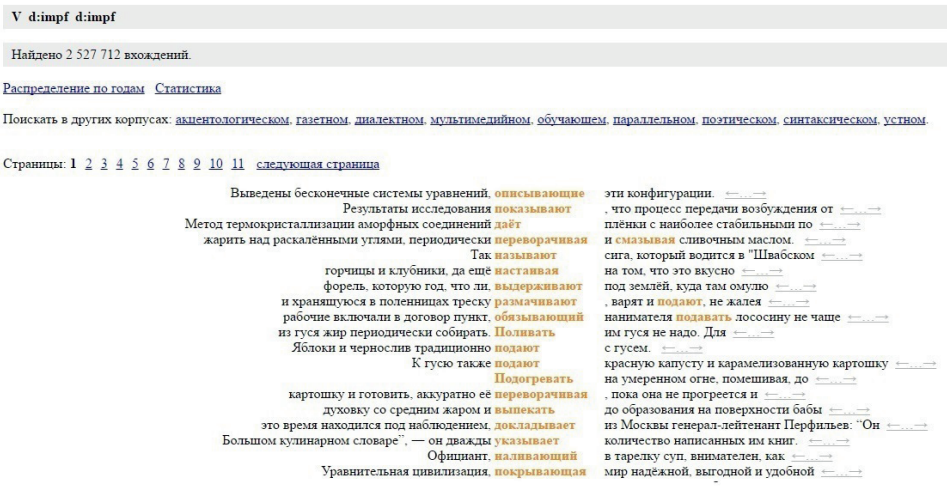
[36] См. X. Guo, *Modal auxiliaries in phraseology: a contrastive study of learner English and native speaker English*, [in:] *Proceedings from the Corpus Linguistics Conference Series*, 2005, www.corpus.bham.ac.uk/PCLC/CL%202005%20xiaotian%20guo.doc [доступ: 11.09.2017].

[37] См. М. Михайлов, *Параллельные корпуса художественных текстов*, Tampere 2003.

[38] См. напр. М.В. Копотев, *Из наблюдений над публицистикой Ф.М. Достоевского (человек и его дело)*, «Slavic Almanac: The South African Year Book for Slavic, Central, and East European Studies» 2003, с. 153–164.

[39] М.В. Копотев, А. Мустайоки, *Современная…*, с. 17.

Таблица 4А. Поиск глаголов по запросу «V d:impf d:impf» – выдача в формате KWIC



V d:impf d:impf

Найдено 2 527 712 вхождений.

Распределение по годам Статистика

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, синтаксическом, устном.

Страницы: **1** 2 3 4 5 6 7 8 9 10 11 следующая страница

Выведены бесконечные системы уравнений,	**описывающие**	эти конфигурации. ⟵ ⟶
Результаты исследования	**показывают**	, что процесс передачи возбуждения от ⟵ ⟶
Метод термокристаллизации аморфных соединений	**даёт**	плёнки с наиболее стабильными по ⟵ ⟶
жарить над раскалёнными углями, периодически	**переворачивая**	и **смазывая** сливочным маслом. ⟵ ⟶
Так	**называют**	сига, который водится в "Швабском ⟵ ⟶
горчицы и клубники, да ещё	**настаивая**	на том, что это вкусно ⟵ ⟶
форель, которую год, что ли,	**выдерживают**	под землёй, куда там омулю ⟵ ⟶
и хранящуюся в поленницах треску	**размачивают**	, варят и **подают**, не жалея ⟵ ⟶
рабочие включали в договор пункт,	**обязывающий**	нанимателя **подавать** лососину не чаще ⟵ ⟶
из гуся жир периодически собирать.	**Поливать**	им гуся не надо. Для ⟵ ⟶
Яблоки и чернослив традиционно	**подают**	с гусем. ⟵ ⟶
К гусю также	**подают**	красную капусту и карамелизованную картошку ⟵ ⟶
	Подогревать	на умеренном огне, помешивая, до ⟵ ⟶
картошку и готовить, аккуратно её	**переворачивая**	, пока она не прогреется и ⟵ ⟶
духовку со средним жаром и	**выпекать**	до образования на поверхности бабы ⟵ ⟶
это время находился под наблюдением,	**докладывает**	из Москвы генерал-лейтенант Перфильев: "Он ⟵ ⟶
Большом кулинарном словаре", — он дважды	**указывает**	количество написанных им книг. ⟵ ⟶
Официант,	**наливающий**	в тарелку суп, внимателен, как ⟵ ⟶
Уравнительная цивилизация,	**покрывающая**	мир надёжной, выгодной и удобной ⟵ ⟶

Источник: www.ruscorpora.ru [доступ: 17.12.2017].

Таблица 4Б. Результаты по запросу «V d:impf d:impf» помещенные в таблицу Excel

L5	L4	L3	L2	L1	**Втор. имперф.**	P1	P2	P3	P4	P5
	Выведены	бесконечные	системы	уравнений,	описывающие	эти	конфигурации			
			Результаты	исследования	показывают	что	процесс	передачи	возбуждения	от
	Метод	термокристаллизации	аморфных	соединений	даёт	плёнки	с	наиболее	стабильными	по
жарить	над	раскалёнными углями		периодически	переворачивая	и	смазывая	сливочным	маслом.	
				Так	называют	сига	который	водится	в	Швабском
горчицы	и	клубники	да	ещё	настаивая	на	том	что	это	вкусно
форель	которую	год	что	ли	выдерживают	под	землёй	куда	там	омулю
и	хранящуюся	в	поленницах	треску	размачивают	варят	и	подают,	не	жалея
рабочие	включали	в	договор	пункт,	обязывающий	нанимателя	подавать	лососину	не	чаще
из	гуся	жир	периодически	собирать.	Поливать	им	гуся	не	надо	Для
	Яблоки	и	чернослив	традиционно	подают	с	гусем.			
		К	гусю	также	подают	красную	капусту	и	карамелизованную	картошку
					Подогревать	на	умеренном	огне	помешивая	до
картошку	и	готовить	аккуратно	её	переворачивая	пока	она	не	прогреется	и
духовку	со	средним	жаром	и	выпекать	до	образования	на	поверхности	бабы
это	время	находился	под	наблюдением	докладывает	из	Москвы	генерал-лейтенант	Перфильев:	"Он
Большом	кулинарном	словаре	— он	дважды	указывает	количество	написанных	им	книг.	
				Официант,	наливающий	в	тарелку	суп	внимателен	как
			Уравнительная	цивилизация	покрывающая	мир	надёжной	выгодной	и	удобной

Источник: собственный материал [доступ: 17.12.2017].

Таблица 4В. Обработка материала в виде таблицы Excel

L5	L4	L3	L2	L1	**Втор. имперф.**	P1	P2	P3	P4	P5
форель	которую	год	что	ли	выдерживают	под	землёй	куда	там	омулю
и	хранящуюся	в	поленницах	треску	размачивают	варят	и	подают,	не	жалея
из	гуся	жир	периодически	собирать.	Поливать	им	гуся	не	надо	Для
					Подогревать	на	умеренном	огне	помешивая	до
духовку	со	средним	жаром	и	выпекать	до	образования	на	поверхности	бабы

Источник: собственный материал [доступ: 17.12.2017].

Выводы

Анализ возможностей корпусов, в частности при поиске глагольных форм на примере НКРЯ, показывает, что корпусное исследование наполняется смыслом лишь при совмещении автоматизированной процедуры извлечения данных с качественным анализом опытного лингвиста. Сдвиг от эксцерпции данных к качественному анализу является признаком достоверного корпусного исследования. Корпусная лингвистика предоставляет инструменты для обработки данных и обеспечивает их поиск с целью решения лингвистических задач и получения конкретной лингвистической информации. Несмотря на недостатки работы с корпусом текстов, этот метод является в данный момент единственным путем к быстрому анализу больших массивов текстов и способом прийти к выводам, недостигаемым без использования новейших технологий. Таким образом, корпусная лингвистика позволяет не только заново заняться известными вопросами, но и проблемами, решение которых до сих пор было связано с рядом трудностей. Адекватность использования корпуса для исследования глагола подтверждается словами А.Д. Шмелева: «Есть лингвистические задачи, для решения которых обращение к компетенции носителей языка ничего не дает и которые могут быть решены только посредством обращения к представительному корпусу текстов»[40]. Таким образом корпусная лингвистика открывает дальнейшие научные перспективы.

Литература

Aston G., Burnard L., *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh 1998.

Fillmore C.J., *Corpus linguistics vs. computer-aided armchair linguistics*, [in:] *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, Stockholm 1992, http://is.muni.cz/el/1421/jaro2008/FJ0B738/um/Corpus_linguistics_verze1.pdf [доступ: 2.06.2017].

Guo X., *Modal auxiliaries in phraseology: a contrastive study of learner English and native speaker English*, [in:] *Proceedings from the Corpus Linguistics Conference Series*, 2005, www.corpus.bham.ac.uk/PCLC/CL%202005%20xiaotian%20guo.doc [доступ: 11.09.2017].

Habert B., Nazarenko A., Salem A., *Les linguistiques de corpus*, Armand Colin, Paris 1997.

Pawłowski A., *Lingwistyka korpusowa – perspektywy i zagrożenia*, «Polonica» 2003, t. XXII–XXIII, c. 19–31.

Podstawy językoznawstwa korpusowego, red. B. Lewandowska-Tomaszczyk, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2005.

Зализняк А.А., *Грамматический словарь русского языка*, Русский язык, Москва 1977.

Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю., *Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей*, http://download.yandex.ru/IMAT2007/kobricov.pdf [доступ: 11.06.2017].

[40] А.Д. Шмелев, *Языковые факты и корпусные данные*, «Русский язык в научном освещении» 2010, № 1, с. 236–265.

Коваль С.А., *Роль корпуса в создании реалистичных моделей словоизменительной морфологии*, [в:] *Труды международной конференции «Корпусная лингвистика – 2006»*, Санкт-Петербург 2006, с. 148–158.

Колокольникова М.Ю., *Дискурс-анализ и корпусный анализ в исследованиях в области исторической лексикологии*, «Известия Саратовского университета» 2010, Журналистика, т. 10, вып. 2, http://www.philology.ru/linguistics1/kolokolnikova-10.htm [доступ: 11.06.2017].

Копотев М.В., *Из наблюдений над публицистикой Ф.М. Достоевского (человек и его дело)*, «Slavic Almanac: The South African Year Book for Slavic, Central, and East European Studies» 2003, с. 153–164.

Копотев М.В., Янда Л., *Национальный корпус русского языка (www.ruscorpora.ru)*, «Вопросы языкознания» 2006, №5, с. 149–155.

Копотев М.В., Мустайоки А., *Современная корпусная русистика*, [в:] *Инструментарий русистики: корпусные подходы*, ред. А. Мустайоки, Л.А. Бирюлин, Е.Ю. Протасова, Helsinki University Press, Helsinki 2008, с. 7–24.

Кустова Г.И., Толдова С.Ю., *Национальный корпус русского языка: Семантические фильтры для разрешения многозначности глаголов*, [в:] *Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы*, Санкт-Петербург 2009, http://ruscorpora.ru/sbornik2008/12.pdf [доступ: 11.06.2017].

Лаврентьев А.М., *Влияние корпусных технологий на развитие диахронической лингвистики: пример Франции*, https://halshs.archives-ouvertes.fr/halshs-01071863/document [доступ: 11.06.2017].

Ляшевская О.Н., Плунгян В.А., Сичинава Д.В., *О морфологическом стандарте Национального корпуса русского языка*, [в:] *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*, Изд. Индрик, Москва 2005, с. 111–135.

Мамонтова В.В., *Корпусная лингвистика в современной языковедческой парадигме*, «Актуальные вопросы современной науки» 2010, № 12, с. 230–238.

Михайлов М., *Параллельные корпуса художественных текстов*, Tampere 2003.

Нагель О.В., *Корпусная лингвистика и ее использование в компьютеризированном языковом обучении*, «Язык и культура» 2008, т. 1, № 4, с. 53–59.

Рассказы о сновидениях: Корпусное исследование устного русского дискурса, ред. А.А. Кибрик, В.И. Подлесская, Изд. Языки славянских культур, Москва 2009.

Шаров С.А., *Представительный корпус русского языка в контексте мирового опыта*, «Научно-техническая информация. Серия 2. Информационные процессы и системы» 2003, № 6, с. 9–18.

Шмелев А.Д., *Языковые факты и корпусные данные*, «Русский язык в научном освещении» 2010, № 1, с. 236–265.

http://www.ruscorpora.ru [доступ: 12.06.2017].

http://www.ruscorpora.ru/corpora-morph.html [доступ: 11.06.2017].

http://www.ruscorpora.ru/corpora-sem.html [доступ: 11.06.2017].

http://www.ruscorpora.ru/corpora-stat.html [доступ: 11.06.2017].

https://tech.yandex.ru/mystem [доступ: 2.01.2017].

Метод корпусного исследования – преимущества и недостатки (на примере использования Национального корпуса русского языка)

Резюме

Настоящая статья посвящена методу корпусного исследования. В ней широко представлены несомненные преимущества и недостатки этого подхода, проиллюстрированные на материале Национального корпуса русского языка. Достоинство корпуса заключается в значительном упрощении и ускорении процедуры лингвистической обработки больших массивов текстов, что однако не лишено многих недостатков, связанных с автоматическим аннотированием данных.

Ключевые слова: корпус; корпусное исследование; аннотирование; корпусная лингвистика, НКРЯ, глагол

The Method of Corpus Study – Advantages and Disadvantages (On the Example of Russian National Corpus)

Abstract

This paper is devoted to the method of corpus study. It presents the advantages and disadvantages of this approach illustrated on the material of the Russian National Corpus. The advantage of the corpus is the considerable simplification and acceleration of the procedure of the linguistic analyze of large amount of texts, which, however, does not avoid many of errors associated with automatic annotation of the data.

Key words: corpus, corpus study, annotation, corpus linguistics, RNC, verb

Izabela Kozera
doktor nauk humanistycznych
Uniwersytet Jagielloński
Instytut Filologii Wschodniosłowiańskiej UJ
ul. Ingardena 3, 30-060 Kraków, Polska

Izabela Kozera, PhD
Jagiellonian University
Institute of Eastern Slavonic Studies
e-mail: izabela.kozera@uj.edu.pl
+48 602281515