УДК 81.161.3’23’33 (045)

Бекреева Юлия Викторовна
кандидат филологических наук
доцент кафедры лексикологии
и стилистики английского языка
Минский государственный
лингвистический университет
г. Минск, Беларусь

Yuliya Bekreyeva
PhD in Philology
Assistant professor
of the Department of English Lexicology
Minsk State Linguistic University
Minsk, Belarus
bekreyeva@mail.ru

КОРПУС ТЕКСТОВ КАК МАТЕРИАЛ ДЛЯ МОДЕЛИРОВАНИЯ ОБРАЗА ИСТОРИЧЕСКОЙ ПЕРСОНАЛИИ

В статье представлена процедура отбора и оценки репрезентативности материала исследования из корпусов белорусскоязычных текстов для моделирования образа исторической персоналии. На примере генерации конкорданса по поисковому запросу «Скарына» описаны результаты автоматической выборки эмпирического материала в Белорусском N-корпусе, подкорпусе белорусского языка Национального корпуса русского языка и веб-корпусе beTenTen 2016. Установлены источники и жанры текстов конкордансной выборки; выявлены погрешности поиска, возможности и ограничения автоматической обработки полученного материала инструментами корпуса. Полученные выводы находят практическое применение в проекте создания специализированного корпуса текстов для алгоритмизации моделирования образов исторических персоналий.

К л ю ч е в ы е с л о в а: *корпус текстов; конкорданс; репрезентативность выборки; образ исторической персоналии; Скарына.*

TEXT CORPUS AS A MATERIAL FOR MODELING THE IMAGE OF A HISTORICAL PERSONAE

The article presents a procedure for selecting and assessing the representativeness of research material from corpora of Belarusian-language texts to model the image of a historical personae. The results of automatic sampling of empirical material in the Belarusian N-corpus, the subcorpus of the Belarusian language of the National Corpus of the Russian Language and the beTenTen 2016 web corpus are described on the example of concordance generation for the search query “Скарына”. The sources and genres of texts of the concordance selection are established; search errors, possibilities and limitations of automatic processing of the material using corpus tools are identified.The results and conclusions have practical application in the project of creating a specialized corpus of texts for algorithmizing the modeling of images of historical personae.

K e y w o r d s: *text corpus; concordance; representativeness of the sample; image of a historical personae; Skaryna.*

По мере того, как большие корпуса оцифрованных текстов становятся все более доступными, исследователи заново открывают для себя потенциальную плодотворность текстовых данных для изучения лингвистических, социальных и культурных явлений. Несмотря на преимущества компьютерных технологий для сбора и хранения данных, которые реализованы в современных корпусах текстов [1, с. 277], актуальной остается проблема отбора и качественной оценки языкового материала, избираемого для конкретной цели исследования.

В настоящей статье описывается процедура анализа данных об употреблении имени исторической персоналии в текстах белорусскоязычных корпусов, которая представляет этап выбора эмпирического материала для лингвокультурологического исследования образов исторических персоналий Беларуси. Объем статьи не позволяет представить теоретическое обоснование исследования образа деятеля на материале контекстов употребления его имени, поэтому ограничимся перечнем задач исследования, реализация которых предполагает обращение к корпусным методам и технологиям:

1) установить категоризацию исторической персоналии в профессиональной, личностной и межличностной сферах;
2) определить ситуативные роли, модели поведения и типичные сценарии действий и событий с участием исторической персоналии;
3) выделить черты внешности и/или свойства характера, ассоциируемые с именем исторической персоналии;
4) выявить артефакты, локации, события, ассоциируемые с именем исторической персоналии;
5) определить тональность высказываний об исторической персоналии.

Подчеркнем, что при моделировании образа исторической персоналии в фокусе внимания находится то, как эта личность представляется авторами речевых произведений, как изображается в текстах и как воспринимается и запоминается читателями. «Важнейшим свойством образа выступает его метафоричность, эмоциональность и узнаваемость, что обеспечивает реализацию коммуникативных функций, а также способность отражать некоторые универсальные ценности» [2, с. 249].

В качестве примера поисковой единицы избрано имя белорусского просветителя и первопечатника Франциска Скорины. Объектом исследования послужили корпусы белорусского языка в открытом доступе: Белорусский N-корпус [3], белорусский подкорпус Национального корпуса русского языка (НКРЯ) [4], веб-корпус beTenTen 2016 [5].

Белорусский N-корпус – текущий проект Института языкознания НАН Беларуси. Объем корпуса составляет около 409 000 текстов и 124 млн слов. Основной корпус включает тексты официально-делового, художественного, научного, публицистического и религиозного стилей. Подкорпус веб-ресурсов включает тексты массмедийных информационных сайтов belta.by, tvr.by, zviazda.by, а также официального государственного сайта president.gov.by. Подкорпус белорусского языка в НКРЯ – небольшой по объему: 312 текстов и около 10 млн слов. Он включает тексты художественного, публицистического и научного жанров с параллельным переводом на русском языке.

Корпус beTenTen 2016 представляет собой компиляцию текстов на белорусском языке, собранную из ресурсов открытого доступа в сети Интернет в 2016 году по технологии автоматического сбора и обработки лингвистически значимого веб-контента SpiderLing, разработанной исследовательской группой А. Килгариффа [6]. Источники текстов – онлайн-библиотеки, официальные сайты СМИ, блоги, паблики, сайты государственных учреждений и общественных объединений. Объем корпуса составляет 63 млн слов.

Процедура анализа корпусных данных включает:

1) оценку возможностей и ограничений поискового запроса имени исторической персоналии;

2) определение объема полученного конкорданса и погрешностей автоматической выборки конкордансных единиц для выполнения исследовательского задания;

3) оценку репрезентативности выборки на основе метаданных о жанровой спецификации и/или источниках текстов;

4) определение возможностей сортировки и извлечения языковых данных в полученной выборке эмпирического материала.

Как правило, в корпусе текстов представлены несколько вариантов поиска: по слову во всех грамматических формах (лемма), по конкретной словоформе или сочетанию слов, по грамматической форме или морфеме. В Белорусском N-корпусе и подкорпусе белорусского языка НКРЯ имя собственное *Скарына* лемматизировано (отметим, что для некоторых других имен персоналий, например, *Ефрасіння*, в Белорусском N-корпусе отсутствует лемматизация). В корпусе beTenTen лемматизации нет, необходимо осуществлять простой поиск по точной словоформе (*Скарына, Скарыне* и т. д.), объединяя полученные конкордансы, или формулировать сложный поиск SQL со всеми вариантами словоформ.

Избранная поисковая единица – имя собственное, имеющее определенную специфику. С одной стороны, имя собственное исторической персоналии представляет собой сочетание слов: имени и фамилии. Воспроизводство полного имени собственного может сделать поиск более точным, особенно при наличии или омонимичных имен собственных (например, однофамильцы), или имен нарицательных. Так, в Белорусском N-корпусе был обнаружен омоним категории «имя нарицательное»: *Ня бачылі касарыкі Хлеба ні* ***скарыны*** (З. Бядуля). Имя исторической персоналии часто представлено вариантами, в нашем случае: *Францыск Скарына*, *Францішак Скарына*, сокращение *Ф. Скарына* и даже *Георгій Скарына*. Простой поиск по словосочетанию ограничивает возможность обнаружения всех словоформ (например, *Францыску, Францыска* и т. д.), требуется оформление поискового запроса в пределах диапазона совместной встречаемости двух лемм.

С другой стороны, известность персоналии отражается в регулярном употреблении только фамилии или только имени, которое становится прецедентным феноменом и, как следствие, расширяет свою категориальную семантику. Данное явление представляет особый интерес для моделирования образа деятеля, который через тексты закрепляется в коллективной памяти носителей языка и воспроизводится в последующей речевой деятельности (новых текстах). Таким образом, с учетом поставленных задач исследования было принято решение сформировать конкорданс по слову *Скарына*, что позволяет охватить контексты употребления исследуемого прецедентного феномена. Погрешности поиска (омонимия) решаются ручной сортировкой.

Очевидным недостатком отбора материала исследования является невозможность охватить инструментом поиска дейктические единицы и иные номинации исторической персоналии в текстах. Например, конкордансная единица в Белорусском N-корпусе включает контекст *Менавіта на людзей простых і паспалітых зарыентаваныя прадмовы* ***Скарыны*** *перад кожнай з кнігаў*… но не охватывает последующее предложение: *У гэтых прадмовах* ***доктар Франьцішак*** *даваў разнастайныя зьвесткі па гісторыі, геаграфіі, культуры…* Обращение к полному фрагменту текста (в ручном режиме) позволяет дополнить материал выборки, но увеличивает время сбора материала.

В целом конкордансный список во всех трех корпусах формируется автоматически из неполных текстовых отрезков с маркированной единицей поиска (KWIC). В НКРЯ приводится связный абзац или полное предложение. В Белорусском N-корпусе единой формы вывода KWIC нет: встречается полное предложение, словосочетание, фрагмент предложения с маркированной единицей поиска, фрагменты предложения с пропусками (отмечены многоточиями). В beTenTen есть возможность классического формата KWIC (левый и правый фрагменты текста от центральной единицы поиска) и формат полного предложения. Расширенный контекст доступен в отдельном окне по запросу для каждой конкордансной единицы.

Рассмотрим объем конкорданса, полученный по запросу *Скарына* в трех корпусах. В НКРЯ ожидаемо обнаружено наименьшее количество конкордансных единиц: 62 примера употребления из 5 текстов. В 2 примерах поисковое имя не обозначает деятеля Франциска Скорину. В Белорусском N-корпусе не указывается количество текстовых отрезков, сгенерированных по запросу, поэтому потребовались дополнительные инструменты для определения объема выборки: сохранение конкорданса в таблице excel и подсчет позиций. Объем конкорданса (основной корпус и веб-корпус) составил 4905 единиц. В 3 примерах поисковое имя не обозначает деятеля, в 116 примерах из пьесы Д. Язэпа поисковое имя обозначает роль и открывает прямую речь. В beTenTen 2016 конкорданс по словоформе *Скарына* составил 1121 пример. В выборках Белорусского N-корпуса и beTenTen обнаружено дублирование конкордансных единиц. Если в корпусе текстов СМИ или веб-корпусе факты дублирования в какой-то мере объясняются спецификой дискурса (тиражируется одна и та же информация разными источниками), то включение в конкорданс одинаковых фрагментов из одного художественного произведения – погрешность конкордансера. Таким образом, выборка материала из Белорусского N-корпуса более представительна в количественном отношении.

Рассмотрим репрезентативность выборки из корпусов по жанрам и источникам. Примеры употребления имени *Скарына* из белорусского подкорпуса НКРЯ взяты из двух публицистических текстов и трех художественных произведений. Лишь в одном источнике, «Память о легендах: белорусские старины голоса и лица» К. Тарасова, представлен связный текст о деятель-

ности Франциска Скорины, примеры из остальных источников – это упоминания имени. В основной части Белорусского N-корпуса конкорданс также составлен из текстов художественного и публицистического стилей (см. Таблицу 1). При анализе источников материала из конкорданса были удалены дублированные конкордансные единицы и примеры омонимии поискового слова.

Таблица 1

Жанровое распределение конкорданса по запросу «Скарына» в Белорусском N-корпусе (основная часть)

Жанр текста	Количество источников	Количество конкордансных единиц
Художественный стиль	17	142
рассказ	2	4
роман	7	11
пьеса	2	118
стихотворение	4	5
повесть	2	4
Публицистический стиль	24	87
статья	8	52
выступление	1	1
предисловие	2	3
очерк	4	10
эссе	5	15
не определен	4	6

В подкорпусе веб-ресурсов несколько источников, но достаточно репрезентативный объем конкорданса (см. Таблицу 2). Интернет-публикации представляют особую сложность для жанровой спецификации [7, с. 23], поэтому здесь мы ограничимся общим описанием: большинство текстов – это новостные и обзорные статьи.

Таблица 2

Распределение конкорданса по запросу «Скарына» по источникам в подкорпусе веб-ресурсов Белорусского N-корпуса

Источник	Количество конкордансных единиц
статьи газеты «Звезда»	3 526
статьи информационного сайта belta	225
статьи информационного сайта Белтелерадиокомпании tvr.by	696
публикации сайта president.gov.by	229

Веб-корпус beTenTen 2016 – более разнообразен в аспекте выбора источников и, соответственно, типов текста, чем подкорпус веб-ресурсов Белорусского N-корпуса. Конкорданс в корпусе сгенерирован из 126 интернет-ресурсов. Метаданные корпуса содержат только указание на источник и гиперссылку. Многие гиперссылки не активны, поэтому количественный анализ жанрового распределения конкордансной выборки не представляется возможным. Отметим лишь наличие примеров из текстов художественных произведений (романы, повести, рассказы из онлайн-библиотек), новостных и обзорных статей, научно-популярных статей и очерков, рефератов, учебных сочинений, энциклопедических справок и биографических нарративов, интервью, сообщений из блогов и бесед. Можно сказать, что конкорданс beTenTen 2016 репрезентирует все типы текстов об исторической персоналии, доступные для массового читателя, а, значит, на таком материале целесообразно проводить лингвистическое моделирование образа Франциска Скорины. Однако проблема верификации источников текста, отсутствие жанровой сортировки инструментами корпус-менеджера снижает оценку репрезентативности материала.

В корпусах есть инструмент построения списков сочетаемости по позиции: слова слева и справа от поисковой единицы в заданном диапазоне окна поиска (например, -1 в препозиции и +1 в постпозиции от имени). Возможность выстраивания списков сочетаемости с учетом синтаксических отношений или семантики слов не предусмотрена. Инструмент Word Sketch для корпуса beTenTen недоступен из-за отсутствия морфологической разметки. Статистический анализ, например, вычисление мер ассоциации, требует дополнительных программ обработки вне исследуемых корпусов. Нет возможности автоматически отсортировать конкордансные примеры единичного упоминания имени исторической персоналии и примеры текстов «об исторической персоналии», т. е. описаний и рассказов о жизни и деятельности.

Оценка материала из доступных корпусов белорусскоязычных текстов выявила ряд проблемных вопросов, связанных с репрезентативностью выборки и применением корпусных методов в исследовании исторических образов национальных деятелей Беларуси. Решение видится в создании специализированного корпуса текстов для алгоритмизации моделирования образов исторических персоналий, проект которого начат в рамках гранта Ивановского государственного университета по программе «Поддержка партнерств». В планах проекта – формирование репрезентативной коллекции текстов об исторических персоналиях разных жанров, морфологическая и синтаксическая разметка текстов, разработка базы данных семантической разметки с учетом вариантов номинаций исторической персоналии, разработка системы управления базами данных, обеспечивающими возможность статистического анализа, сентимент-анализа, генерации списков сочетаемости на основе семантико-синтаксической роли слов в предложении-высказывании.

ЛИТЕРАТУРА

1. *Мамонтова В. В.* Корпусная лингвистика и лингвистические корпусы // Язык. Текст. Дискурс. 2007. № 5. С. 275–283.

2. Мазур Л. М. Образ прошлого: формирование исторической памяти // Известия Уральского федерального университета. Сер. 2, Гуманитарные науки. 2013. № 3 (117). С. 243–256.

3. Беларускі N-корпус. URL: https://bnkorpus.info/korpus.be.html (дата обращения: 11.04.2023).

4. Национальный корпус русского языка. 2003–2024. URL: https://rus-corpora.ru (дата обращения: 21.04.2023).

5. BeTenTen – Belarusian corpus from the web. 2016. URL: www.sketchengine.eu/betenten-belarusian-corpus/?highlight=Betenten (accessed: 02.04.2024).

6. Kilgarriff A. The TenTen corpus family // 7th International Corpus Linguistics Conference CL, July 2013 [Electronic resource]. URL: https://www.sketchengine.eu/documentation/tenten-corpora/#toggle-id-4 (accessed: 02.04.2024).

7. *Сантини М.* Веб-страницы, типы текстов и лингвистические характеристики: некоторые вопросы // Жанры речи. 2019. № 1 (21). С. 22–33. DOI: 10.18500/2311-0740-2019-1-21-22-33.