ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

УДК 811.113.6 '33

Богданова Наталия Альбертовна
кандидат филологических наук,
доцент кафедры фонетики
и грамматики немецкого языка
Минский государственный
лингвистический университет
г. Минск, Беларусь

Natallia Bahdanava
PhD in Philology, Associate Professor
of the Department of Phonetics
and Grammar of the German Language
Minsk State Linguistic University
Minsk, Belarus
albertowna@mail.ru

Мельник Виталина Андреевна
преподаватель кафедры теории
и практики перевода
Минский государственный
лингвистический университет
г. Минск, Беларусь

Vitalina Melnik
Lecturer at the Department of Theory
and Practice of Translation
Minsk State Linguistic University
Minsk, Belarus
vitalinasmolskaya@gmail.com

ЭТАПЫ РАЗРАБОТКИ КОРПУСНОГО МЕНЕДЖЕРА ДЛЯ АНАЛИЗА ПАРАЛЛЕЛЬНЫХ ТЕГИРОВАННЫХ КОРПУСОВ ТЕКСТОВ

STEPS OF DEVELOPING A CORPUS MANAGER FOR ANALYSIS OF PARALLEL TAGGED TEXT CORPORA

В статье рассматриваются определения и типология текстовых корпусов, описываются процесс создания параллельного тегированного корпуса текстов на базе отрывка трилогии А. Линдгрен о Малыше и Карлсоне в оригинале и в переводе на белорусский язык, а также этапы создания и принцип работы корпусного менеджера для обработки параллельного тегированного корпуса текстов. Приводятся предварительные результаты, полученные в ходе автоматической обработки созданного корпуса.

К л ю ч е в ы е с л о в а: *корпус текстов*; *система тегов*; *семантическая разметка*; *дескриптор*; *тегированная текстовая модель*.

The article deals with the definition and typology of text corpora, describes the process of creating a parallel tagged text corpus based on an excerpt of A. Lindgren's trilogy about the Little Brother and Karlsson-on-the-Roof in the original and translated into Belarusian language, as well as the stages of creation and the principle of operation of a corpus manager for processing a parallel tagged corpus of texts. Preliminary results obtained in the course of automatic processing of the created corpus are described.

K e y w o r d s: *text corpus*; *tagging system*; *semantic markup*; *descriptor*; *tagged text model*.

Лингвистическое исследование неизбежно опирается на анализ языковых данных. Сегодня этап сбора лингвистических данных для последующего анализа значительно упростился в связи с развитием современных информационных технологий и появлением корпусов текстов. *Корпусная лингвистика* представляет собой «раздел компьютерной лингвистики, который занимается разработкой общих принципов построения и использования лингвистических корпусов с применением компьютерных технологий» [1, с. 5]. Это направление основывается на том, что достоверная информация о языке и речи может быть получена только из довольно большого объема текстов. Одним из важных преимуществ корпусного исследования является значительная экономия времени, благодаря чему текстовые корпусы активно применяются во многих областях лингвистики: лексикографии, переводоведении, психолингвистике, социолингвистике, сравнительном языкознании, литературоведении и других. Необходимость исследователей располагать как можно большими объемами лингвистического материала и послужила причиной такого широкого применения корпусов текстов [2].

Существует множество определений корпуса текстов как феномена. А. Н. Баранов понимает под *корпусом текстов* «вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области» [3, с. 115]. В. П. Захаров определяет корпус текстов как «большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [4, с. 3]. Т. МакЭнери и Э. Вилсон рассматривают *корпус* как «собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка» [5]. Таким образом, можно понимать корпус как «уменьшенную модель языка или подъязыка» [4, с. 5].

Лингвист Э. Финеган под корпусом текстов понимает «репрезентативное собрание текстов» [6]. Репрезентативность по отношению к проблемной области является важнейшим свойством любого корпуса текстов. По мнению А. Н. Баранова, репрезентативность – это «способность корпуса текстов отражать все свойства проблемной области, релевантные для данного типа лингвистического исследования, в определенной пропорции, определяемой частотой явления в проблемной области» [3, с. 118].

Исходя из приведенных выше определений понятия корпуса текстов можно сделать вывод, что к основным характеристикам современных

лингвистических корпусов относятся следующие: машиночитаемый вид, цель, репрезентативность. Кроме того, рассмотренные определения корпуса свидетельствуют о двойственной природе корпусной лингвистики. С филологической точки зрения корпус рассматривается как совокупность текстов, т. е. содержит языковой материал. С технической же точки зрения корпусы могут полноценно функционировать только с использованием современных компьютерных технологий.

В настоящее время для разметки лингвистических корпусов могут в том числе применяться инструменты обработки естественного языка (англ. natural language processing – NLP), которые позволяют максимально автоматизировать процесс сборки необходимых языковых данных; эта область корпусной лингвистики пересекается с исследованиями искусственного интеллекта и активно развивается [7; 8].

В зависимости от материала, входящего в корпус, способа его организации, а также целей и задач применения корпуса выделяются различные типы лингвистических корпусов. Так, например, существуют многоцелевые и специализированные, исследовательские и иллюстративные, динамические и статические, размеченные и неразмеченные, одноязычные, двуязычные и многоязычные, национальные и сбалансированные корпусы [4; 9].

Особым типом корпуса текстов является параллельный корпус. Корпусы параллельных текстов содержат как исходный текст, так и его перевод на другой язык. Для проведения исследования нами был создан двуязычный параллельный корпус текстов на материале первых трех глав трилогии шведской писательницы А. Линдгрен о Малыше и Карлсоне на шведском языке[1] и в переводе на белорусский язык[2].

Целью исследования являлось создание программно-лингвистической платформы для построения и сопоставительного исследования моделей литературного персонажа в оригинале и переводе с применением разработанного параллельного корпуса текстов. Объем нашего корпуса представляется достаточным для иллюстрации возможностей применения корпусной методологии в рамках подобного сопоставительного исследования.

В процессе создания параллельного корпуса тексты прошли несколько основных этапов обработки. Сначала отобранный текстовый материал был приведен в машиночитаемую форму, чтобы стать распозна-

[1] *Lindgren, A.* Lillebror och Karlsson på taket / A. Lindgren. – Uddevalla : Bohuslänningens AB, 1968. – Kap. 1–3. – S. 5–46.

[2] *Ліндгрэн, А.* Малы і Карлсан, які жыве на даху / А. Ліндгрэн. – Мінск : Папуры, 2019. – Гл. 1–3. – С. 7–54.

ваемым для компьютера. Путем сканирования бумажные версии текстов были переведены в формат «.pdf», после чего последовала конвертация текстов из формата «.pdf» в машиночитаемый формат «.docx».

На следующем этапе при помощи программы memoQ параллельные тексты были автоматически выровнены на уровне предложений с целью установления соответствий между текстом оригинала и переводным текстом. Далее автоматическое выравнивание было отредактировано вручную, так как не для всех шведских предложений были автоматически найдены белорусские соответствия. Полученным 116 пустым целевым вариантам предложений соответствия были установлены вручную. В результате был получен корпус текстов в виде таблицы, состоящей из 618 строк, число которых соответствует количеству предложений произведения оригинала.

Далее отобранный и прошедший все необходимые этапы обработки текстовый материал подлежит тегированию, т. е. разметке. В нашем исследовании для сравнительного описания литературного персонажа мы разработали собственную систему семантических тегов, включающую два вида дескрипторов: характеристики персонажа и способы выражения этих характеристик. Было выделено 16 основных характеристик персонажа: национальность; гендерная принадлежность; внешний вид; возраст; условия жизни; социальный статус; образ жизни; физические способности; интеллектуальные способности; эмоциональное состояние; психологическая характеристика в момент речи; убеждения; личностные качества; пристрастия и привычки; отношения с другими героями; поступки. К способам выражения характеристик были отнесены 5 видов речи, в которых содержатся данные характеристики: авторская речь; прямая речь персонажа; внутренняя речь персонажа; прямая речь других героев; внутренняя речь других героев.

Система семантических тегов включает, соответственно, коды каждого из двух видов дескрипторов. Названия дескрипторов-характеристик персонажа представляют собой начальные буквы каждой характеристики на английском языке. Например, действия и поступки – <*act*>, интеллектуальные способности – <*int*> и т. д. Теги для способов выражения характеристик представляют собой цифры от 1 до 5. Все теги заключаются в угловые скобки в соответствии с правилами языка разметки HTML. Для проведения разметки необходимы открывающие и закрывающие теги. Закрывающие теги отличаются наличием в угловых скобках слэша перед названием тега. Например, *</ap>* – для внешности героя; *</hab>* – для привычек; *</1>* – для авторской речи; *</5>* – для внутренней речи других героев произведения.

Нами была проведена семантическая разметка параллельных текстов с применением разработанной системы тегов. Каждый фрагмент, содержащий выделенные нами характеристики персонажа, в обоих текстах маркируется двумя открывающими и двумя закрывающими тегами: характеристика персонажа и способ выражения характеристики. Тегированию подлежали как отдельные слова (*маленькі, самаўпэўнены, прылятаў*), так и словосочетания (*незвычайная асоба, забыўся на дамок, спрытна схапіў*) и предложения (*Спакойна, толькі спакойна!*). Например, тегированный фрагмент, который мы отнесли к убеждениям Карлсона, содержащийся в его прямой речи, выглядит следующим образом: *<bel><2>Det är en världslig sak</2></bel>* ‘<bel><2>Гэта з’ява банальная</2></bel>’.

Отметим, что при проведении семантической разметки возникло множество спорных моментов, связанных с невозможностью однозначного толкования тех или иных контекстов: некоторые текстовые фрагменты можно было отнести сразу к нескольким характеристикам персонажа. Однако для достижения точного результата мы придерживались наиболее очевидной трактовки контекстов и относили их лишь к одной из возможных характеристик. Вместе с тем в отобранном нами объеме материала в процессе проведения семантической разметки существенных расхождений между текстом оригинала и текстом перевода не наблюдалось, так как в большинстве случаев теги расставлялись «симметрично».

Необходимой частью практически любого корпуса текстов является *корпусный менеджер* – встроенная в корпус поисковая система, которая позволяет пользователю в простой и удобной форме извлекать статистические данные [3, с. 6]. Именно благодаря ему осуществляется управление текстовыми данными в корпусе и извлекается лингвистическая информация для последующего анализа.

Совместно с международной группой IT-компаний SoftTeco нами был разработан корпусный менеджер, представленный в ограниченном доступе в сети Интернет в виде приложения.

При разработке проекта корпусного менеджера мы ориентировались на решение двух видов задач:

1) получать статистическую информацию по метаданным (абсолютную и относительную частоту тегов);

2) получать информацию по языковым данным (какое вербальное выражение получил тот или иной тег, а также какова абсолютная и относительная частота отмеченных тегами текстовых фрагментов). Первоначально данная информация была оформлена нами в виде таблиц (в том числе с использованием инструментов Microsoft Excel), которые впослед-

ствии и стали основой для создания приложения. Работа приложения апробирована на материале созданного нами параллельного корпуса текстов.

Для работы в приложение загружаются представленные в виде таблицы параллельные тексты на двух языках, выровненные на уровне предложений и размеченные при помощи разработанной нами системы семантических тегов.

Программа имеет функцию представления так называемой «тегированной модели текста», т. е. выводит на экран только тегированные текстовые фрагменты в том порядке, в котором они следуют в тексте. Это позволяет проанализировать, какие лингвистические средства использует автор оригинального произведения и каким образом данные средства трансформируются при переводе на другой язык. Эта функция доступна во вкладке «Main statistics». Текстовые фрагменты, не отмеченные тегами, на экран не выводятся. Так, нами был выявлен ряд различий в вербальном оформлении тегированных текстовых фрагментов в двух текстах. В частности, автор белорусского перевода активно использует уменьшительно-ласкательные суффиксы с целью снижения категоричности описаний (*lilla trinda kropp* ‘маленькая круглявая постаць’; *ett särskilt litet hus* ‘у малюпасенькім дамочку’; *farbror* ‘дзядзечка’), а также перефразирует сложные для детского восприятия слова шведского языка (*världens bästa konstflygare* ‘найлепшы ў свеце лятун’; *världens bästa ångmaskinsskötare* ‘найлепшы ў свеце запускальнік паравых машын’; *världens bästa motorskötare* ‘найлепшы ў свеце змазвальнік мацараў’).

Во вкладках приложения «Tags using swe» и «Tags using bel» доступна информация об абсолютной частоте (т.е. количестве упоминаний) тегированных текстовых фрагментов в оригинале и в переводе, что позволяет проследить, какие вербальные выражения наиболее типичны для той или иной характеристики персонажа. Кроме того, при нажатии на тот или иной фрагмент, относящийся к какой-либо характеристике персонажа, во всплывающем окне пользователю доступен и контекст, в котором данный фрагмент употреблен. Сравнительный анализ показал, что при описании внешности Карлсона в оригинальном тексте наиболее частотными являются следующие фрагменты: *liten* (F=5), *tjock* (F=4), *lagom tjock* (F=3). В белорусском тексте наибольшую частотность имеет словосочетание *пульхнай ручкай* (F=3) и прилагательные *таўсматы* (F=3), *маленькі* (F=2), *узорна тоўсты* (F=2). Это позволяет заключить, что акцент в описании внешнего вида персонажа в обоих текстах делается на его полном телосложении. Вместе с тем в шведском тексте более частотным, чем

в белорусском тексте, является прилагательное *liten*, которое упоминается 5 раз, в то время как его белорусское соответствие *маленькі* встречается всего 2 раза. Таким образом, в оригинальном тексте автор в сочетании с полнотой Карлсона выделяет еще и его миниатюрность.

Наконец, приложение предоставляет данные об абсолютной и относительной частоте тегов в обоих текстах. Эти данные доступны во вкладках «Tags statistics swe» и «Tags statistics bel» для оригинального и переводного текстов соответственно. В результате проведенного исследования было установлено, что в оригинальном тексте содержатся 946 тегов, а в тексте перевода на белорусский язык – 928 тегов. В оригинальном тексте наибольшее количество характеристик персонажа представлено в авторской речи и в прямой речи самого персонажа. Так, авторская речь в оригинальном тексте составляет 43 % от всех видов речи, а прямая речь персонажа – 33 %. Ситуация в белорусском переводе практически идентична: 44 % от всех видов речи составляет авторская речь, а 33 % – прямая речь персонажа. В обоих текстах описания автора доминируют в характеристиках внешности, национальности персонажа, а также его психологической характеристики в момент речи. Прямая речь Карлсона доминирует в описаниях его возраста, образа жизни, интеллектуальных и физических способностей, убеждений, привычек и поступков. При этом о своих способностях персонаж говорит исключительно с положительной стороны, что свидетельствует о его эгоистичной, самовлюбленной натуре. Вместе с тем в исследуемом объеме материала совершенно не представлена внутренняя речь персонажа, что позволяет судить о Карлсоне как о довольно поверхностном герое, лишенном определенной глубины эмоций и суждений.

Доминирующими характеристиками в нашем корпусе текстов являются отношения между персонажами (26 % в оригинальном и 27 % в переводном тексте), психологическая характеристика в момент речи (19 % и 18 % соответственно для оригинального и переводного текстов) и эмоциональное состояние персонажа (12 % от общего количества характеристик в обоих текстах).

Основываясь на данных, полученных с использованием корпусного менеджера, можно заключить, что в сформированном корпусе параллельных текстов не было выявлено значительных различий в частотном распределении тегов между оригинальным и переводным текстом, что объясняется высокой точностью белорусского перевода по отношению к оригинальному тексту на шведском языке. Однако ряд различий в вербальном наполнении тегов свидетельствует о стремлении переводчика

смягчить возможные отрицательные описания персонажа и упростить текст для детского восприятия, в результате чего Карлсон становится более понятным и привлекательным для детей героем.

Разработанный корпусный менеджер позволяет исследовать параллельные и непараллельные тегированные тексты разных жанров, автоматически получать статистическую информацию по разным видам дескрипторов, предоставляя исследователю данные для изучения оригинала и перевода/переводов текста, с целью интерпретации текста, а также выявления индивидуальных особенностей авторского стиля.

ЛИТЕРАТУРА

1. *Захаров, В. П.* Корпусная лингвистика : учеб. пособие / В. П. Захаров, С. Ю. Богданова. – Иркутск : ИГЛУ, 2011. – 161 с.
2. *Мамонтова, В. В.* Корпусная лингвистика в современной языковедческой парадигме [Электронный ресурс] / В. В. Мамонтова. – Режим доступа: https://cyberleninka.ru/article/n/korpusnaya-lingvistika-v-sovremennoy-yazykovedcheskoy-paradigme. – Дата доступа: 13.12.2022.
3. *Баранов, А. Н.* Введение в прикладную лингвистику : учеб. пособие / А. Н. Баранов. – Изд. 5-е. – М. : URSS, 2017. – 368 с.
4. *Захаров, В. П.* Корпусная лингвистика : учеб.-метод. пособие / В. П. Захаров. – СПб., 2005. – 48 с.
5. *McEnery, T.* Corpus Linguistics / T. McEnery, A. Wilson. – Edinburgh : Edinburgh Univ. Press, 1996. – 209 p.
6. *Finegan, E.* Language: its structure and use / E. Finegan. – N. Y. : Harcourt Brace College Publishers, 2004. – 546 p.
7. *Горожанов, А. И.* Создание лингвистического корпуса на основе инструментов обработки естественного языка: планирование программных решений / А. И. Горожанов // Филологические науки. Вопросы теории и практики. – 2023. – Т. 16, № 5. – С. 1616–1620.
8. *Горожанов, А. И.* Стандартизированная процедура получения статистических параметров текста (на материале цикла рассказов Дж. Лондона «Смок Белью. Смок и Малыш») / А. И. Горожанов, И. А. Гусейнова, Д. В. Степанова // Вестник МГЛУ. Сер. 1, Филология. – 2022. – № 4 (119). – С. 7–13.
9. *Горожанов, А. И.* Интерпретация художественного произведения: корпусный подход / А. И. Горожанов, Д. В. Степанова // Филологические науки. Вопросы теории и практики. – 2022. – Т. 15, № 1. – С. 203–208.

Поступила в редакцию 24.11.2023