А. И. Горожанов, Д. В. Степанова

КОРПУСНЫЙ ПОДХОД К МОДЕЛИРОВАНИЮ ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ И КАЧЕСТВЕННОЙ СТРУКТУР ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ

Корпусная лингвистика как относительно новое направление компьютерной лингвистики занимается изучением построения, методологии создания и использования лингвистических корпусов, а также способов обработки данных в корпусах. Суть этого направления сводится к тому, что достоверные данные о фонетической, морфологической, синтаксической и семантической структуре языка и речи могут быть получены только из достаточно большого массива текстов.

Современные компьютерные технологии позволяют накапливать и оперативно обрабатывать сверхбольшие объемы данных. Работа с электронными корпусами текстов стала одним из основных методов лингвистических исследований, с помощью которого решаются самые различные задачи.

В настоящее время для проведения корпусных исследований используются «гигакорпусы», или лингвистические корпусы третьего поколения, которые позволяют решить проблему преодоления несовместимости с прог-

раммным обеспечением пользователя и достижения высокой скорости исполнения поисковых запросов (Солнышкина, Гатиятуллина 2020). Но даже такие вычислительные мощности не позволяют в полной мере решать актуальные задачи корпусной лингвистики, в частности, связанные с интерпретацией художественных произведений точными методами.

Исследование ставит перед собой цель определить в рамках корпусного подхода оптимальный с точки зрения достоверности и универсальный метод интерпретации художественных произведений, который позволит извлекать данные для моделирования пространственно-временной и качественной структур произведения, установить черты идиостиля автора.

Для достижения поставленной цели по применению корпусного подхода к интерпретации художественного произведения точными методами необходимо выполнить следующие задачи.

1. Определить параметры интерпретации художественных произведений. Разрабатываемый метод направлен на получение результатов, которые бы позволили, во-первых, описать параметры художественной реальности произведения, во-вторых, составить представление о персональных характеристиках героев, в-третьих, установить черты идиостиля автора. Описание художественной реальности в рамках нашей задачи сводится прежде всего к исследованию таких категорий, как пространство, время и качество, которые могут быть представлены в различных комбинациях, например, пространство и время, «хронотоп» (Бахтин 1975; Ноздрина 1997), пространство и качество (Горожанов, Гусейнова 2021). Составление языковых портретов персонажей требует тщательной работы над текстом художественного произведения с высокой долей «ручного» труда, касающейся маркирования реплик и внутренней речи персонажей, причем работа осложняется еще и тем, что отделить языковой портрет героя от языкового портрета автора чрезвычайно трудно. Планируется охарактеризовать языковые портреты героев, составив предварительно подкорпусы их речи как подмножества корпуса всего художественного произведения (Potapova, Komalova 2018). Что касается исследования идиостиля автора, то в современной предметно-специальной литературе мы встречаем как работы, направленные на установление идиостиля в целом (Баранов, Добровольский, Фатеева 2021), так и попытки описать его отдельные компоненты, в рамках чего могут быть проанализированы целые текстообразующие категории или отдельные языковые явления (Тарасевич 2014; Соколова, Степанова 2019; Бойчук, Джонсон 2020; Горожанов 2021). Мы планируем применить здесь принцип «от простого к сложному», двигаясь от отдельных языковых явлений к текстообразующим категориям (Горожанов, Степанова 2022).

2. Определить тип создаваемого лингвистического корпуса, необходимого для исследования всех заданных параметров. Это может быть либо национальный («гигакорпус»), либо сбалансированный (специальный) лингвистический корпус. Для интерпретации произведений художественной литературы целесообразным представляется использовать сбалансированные корпусы,

которые будут включать, например, тексты всех или определенных работ того или иного писателя, а возможно, даже тексты произведений писателей одного литературного направления. Такого рода корпусы применяются для решения специализированных задач и не имеют универсальных жестких требований к объему. В рамках исследования ставится задача создания собственного сбалансированного (специального) корпуса объемом не менее одного полного текста художественного произведения (с опцией выделения подкорпусов речи персонажей), а именно письменный одноязычный (русский, немецкий, английский языки) литературный художественный исследовательский статический неразмеченный полнотекстовый синхронический корпус.

3. Отобрать подходящие программные решения, которые определяют не только скорость обработки данных, что немаловажно при объемах современных лингвистических корпусов, но и достоверность получаемых результатов. Для решения поставленной задачи был выбран язык программирования Python, который на текущий момент является мировым лидером по популярности благодаря своей универсальности и наличию большого набора библиотек для решения различных прикладных задач, например, библиотека spaCy, представляющая собой набор разнообразных инструментов для расширенной обработки естественного языка (Добровольский, Кротова, Цветаева и др. 2021).

4. Предварительная апробация разрабатываемого метода. Экспериментальная проверка использования корпусного подхода для интерпретации художественного произведения проводилась на материале романа Ф. Кафки «Замок», текст которого составил сбалансированный неразмеченный лингвистический корпус, программно разбитый на предложения, каждое из которых начинается с новой строки. Токенизатор spaCy выделил в корпусе 13 6410 элементов (токенов), к которым были отнесены не только словоформы, но также знаки препинания и знаки новой строки. В автоматическом режиме были выделены так называемые «сущности», т. е. имена собственные и локации в порядке упоминания от самого частого к самому редкому (своего рода ономастическая модель или художественный ономастикон романа (Косиченко 2017). Результат работы программы нетривиален, поскольку spaCy проводит не простую частотную выборку и сравнение токенов с имеющимися в ее базе данных образцами, но и анализирует их окружение. После ручного объединения различных падежных форм одной и той же леммы был получен следующий список персоналий (более 10 раз упоминаются в полученном результате): K. – 724 раза, Frieda – 296 раз, Barnabas – 177 раз, Klamm – 120 раз, Amalia – 88 раз, Hans Brunswick – 83 раза, Olga – 70 раз, Pepi – 53 раза, Bürgel – 33 раза, Jeremias – 30 раз, Sordini – 24 раза, Sortini – 21 раз, Gerstäcker – 15 раз, Erlanger – 15 раз, Lasemann – 11 раз. Зная содержание романа, можно сказать, что в пропорции количественных показателей достаточно точно отражается важность тех или иных героев. При рассмотрении языкового портрета персонажей очевидным становится тот факт, что частотные герои романа обязательны для такого рода анализа. Среди клю-

чевых локаций романа программа зафиксировала следующие (в алфавитном порядке): *Brükkenhof* ‘постоялый двор «У моста»’, *Dorf* ‘деревня’, *Erde* ‘Земля’, *Fenster* ‘окно’, *Haus* ‘дом’, *Herrenhof* ‘господский постоялый двор’, *Kirche* ‘церковь’, *Schloss* ‘замок’. Интересны данные о временной структуре романа, отраженной в глагольных формах. Расчет показывает, что глаголы, имеющие в spaCy метку «настоящее», употребляются в романе 6 928 раз, а глаголы с меткой «прошедшее» – 8 028 раз. Интерпретировать этот результат можно различным образом. Например, такое большое количество форм настоящего времени объясняется тем, что в романе значительное место занимают диалоги персонажей, тогда как в прошедшем времени приводятся в основном различного рода описания. Особенно важным представляется тот факт, что этот результат был получен оперативно и без предварительной разметки корпуса художественного произведения.

Таким образом, полученные результаты показали, что использование неаннотированных корпусов текстов может быть эффективным для решения поставленной задачи в совокупности с современными программными инструментами обработки естественного языка.