ВОЗМОЖНОСТИ ТЕКСТОВОГО ПОИСКА В ШВЕДСКОМ ПОДКОРПУСЕ НРКЯ

Ковалевская К. В.

Минский государственный лингвистический университет

Аннотация. Статья содержит сведения об основных возможностях текстового поиска в шведском корпусе НРКЯ. Раскрывается содержание понятия «текстовый поиск» и «метаразметка». Автор дает обобщенную характеристику классификации метатекстовой разметки шведского подкорпуса НРКЯ и ее структуры. В заключение раскрывается значимость текстовой разметки в корпусе и возможности использования корпуса текста в исследовательских целях.

Ключевые слова: корпус, разметка, корпусная лингвистика, поиск.

Значительной частью поискового процесса является текстовый поиск в корпусе, который предоставляет позволяет пользователям отбирать тексты из корпуса по заданным внешним параметрам: например, тексты художественной литературы, написанные автором женского пола, родившимся между 1920–1930 гг. Из списка выбранных текстов пользователь может создать свой собственный подкорпус или, подмассив, текстов, в котором в дальнейшем может осуществлять как и поиск точных форм, так и лексико-грамматический поиск [3, с.194 –195]. Функция выборки текстов и создания пользовательского подкорпуса становится возможно благодаря так называемой метаразметке или, метаописанию текстов. Данная функция значительно облегчает работу с огромными массивами текстов и предоставляет дополнительную лингвистическую (морфологическую, синтаксическую, семантическую, стилистическую) информацию о текстах.

Подкорпус может быть создан пользователем в следующих целях:

1. Осуществление поиска в текстах за конкретный период времени;
2. Осуществление поиска в текстах определенного автора или определенном тексте;
3. Осуществление поиска в текстах определенных жанров или тематики;
4. Осуществление поиска в текстах, которые появились в НКРЯ недавно или, наоборот, давно.

Пользователь может создать свой подкорпус текстов, выбрав функцию «Задать подкорпус», в любом корпусе, кроме синтаксического и исторического. Однако, набор параметров текста варьируется в зависимости от корпуса. Так, в мультимедийном параллельном корпусе (Мультипарк) в блоке «Жанр и тип текста» можно указывать место и время описываемых событий, в то время как в диалектном корпусе могут быть выбраны район и год записи текста.

Текстовый поиск становится возможным благодаря метаразметке. Под метаразметкой понимают приписывание тексту определенных атрибутов, которые характеризуют обстоятельства его создания, автора, тематику, особенности жанра и т.д. Совокупность этих признаков, характеризующих текст, называют также паспортом текста. Метаразметка служит основой для формирования архитектуры корпуса, а также позволяет контролировать процесс информационного наполнения корпуса, оценивать его представительность и сбалансированность [4, с. 3–5]. Метаразметка является необходимой для лингвистов и языковедов, пользующихся корпусов в исследовательских целях, т. к. функция

метаразметки позволяет делать выборку текстов по определенным заданным параметрам: например, искать тексты, написанные только мужчинами, или тексты только мемуарного характера.

Существует мнение, что тщательно продуманная и качественно реализованная метатекстовая разметка может помочь справиться с проблемой балансировки текстов при составлении репрезентативного корпуса. Для этого разработчикам необходимо снабдить метаразметкой как можно большее количество текстов корпуса и предоставить пользователю возможность самому отбирать необходимые подкорпуса по предложенным признакам. Следовательно, чем больше набор параметров характеристики каждого текста, тем шире возможности для решения различных лингвистических задач [7].

Национальный корпус русского языка включает в себя внушительный объем текстов различных жанров и стилей, территориальных и социальных вариантов и т.п., поэтому метаразметка является незаменимой при работе с внушительным объемом разнородных текстов [6, с. 8–9]. Более того, одной из интересных задач, которую способна решить метаразметка является установление статистически достоверных корреляций между теми или иными метатекстовыми параметрами (например, полом или возрастом автора) и языковыми особенностями текста.

Существует несколько стандартов кодирования корпусной информации, однако наиболее авторитетными их них считаются TEI (Text Encoding Initiative), XCES (XML Corpus Encoding Standard), EAGLES (European Advisory Group on Language Engineering Standards) [1, с. 279–280]. Наиболее детальным стандартом считается кодировка TEI, которая подходит для представления абсолютно любых текстов и элементов текстовой информации, включая:

структуру, заголовки, типы речи, страницы, цитаты, ссылки, сноски, исправления, таблицы, формулы, специальные символы, лингвистические аннотации и др. Стандарт TEI не ориентирован на корпусные приложения, но используется во многих существующих корпусах, включая BNC (Britisch National Corpus), Чешский национальный корпус и другие. Если рассматривать TEI и некоторые другие стандарты кодировки детально, то они являются слишком сложными для массовой разметки текстов. По этой причине первоначально для выделения характеристики текстов в НКРЯ была использована классификация EAGLES (European Advisory Group on Language Engineering Standards), предложенная Джонам Синклером [2, с. 251–263]. Данная классификация была использована для разметки материала в чешском, британском и американском национальных корпусах. Для описания материала русского языка классификация была адаптирована С.А. Шаровым.

Классификация в отечественной лингвистике речи близка с оригинальной классификациуй EAGLES, которая отражает структуру акта коммуникации, и включает в себя использование категорий «сфера функционирования текста» и «речевой жанр» как тип текста. Речевые жанры не привязываются на прямую к коммуникативным целям, как в классификации Синклера.

При создании НКРЯ у разработчиков была задача отразить картину реального употребления языка определенного периода и при этом было разработано решение минимального вмешательства в соотношение реальных текстов, употребляемых в различных сферах речевой практики. Однако, разработчики были ограничены отбором изданий на этапе наполнения корпуса текстами. Учитывались такие факторы, как общественная значимость и популярность произведений, оценка критиков и т.п.

На данный момент для описания текстов национального корпуса русского языка используются 25 признаков, 9 из которых характеризуют текст, 3 параметра характеризуют автора, 3 – возможную аудиторию, 4 параметра содержат библиографическую информацию о тексте, 5 параметров представляют собой служебную информацию [5, с. 244–245]. Однако, в параллельном корпусе при задании подмножества корпусов доступны не все параметры. Так, при задании подкорпуса в параллельном корпусе шведского языка существует возможность указывать информацию об авторе, а именно: его имя и возраст; выбрать тип корпуса (двуязычный или многоязычный); год создания текста; имя переводчика; язык оригинала и перевода; дату перевода и сферу функционирования произведения.

Таким образом, текстовый поиск позволяет сузить поиск информации среди внушительных массивов текстов в НКРЯ и отобрать необходимые тексты по указанным характеристикам. Создав свой подкорпус, пользователь может осуществить фразовый и лексико-грамматический поиск в текстах по выбранным заранее параметрам. Благодаря внушительной метатекстовой разметке НКРЯ, которая включает в себя лингвистическую (морфологическую, синтаксическую, семантическую, стилистическую), социологическую и библиографическую информацию возможности поиска для решения различных лингвистических задач значительно расширяются.

ЛИТЕРАТУРА

1. Аброскин, А. А. Поиск по корпусу: проблемы и методы их решения / А. А. Аброскин. –СПб.: Нестор-История, 2009. – С. 279–280.

2. Амиева А. М., Филимонов В. В., Сергеев А. П. Основные методики исследования структуры текста // Передача, обработка, восприятие текстовой и графической информации. – Екатеринбург, 2015. С. 251-263.

3. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. – М.: Индрик, 2005. – С.194 –195.

4. Захаров, В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов/ В.П. Захаров, С.Ю. Богданова. – Иркутск: ИГЛУ, 2011. – С. 3–5.

5. Кретов А. А. Анализ семантических помет в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. – С. 244–245.

6. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Корпуса современного русского языка // Научная и техническая информация, сер. 2. Информационные процессы и системы, 2005, № 6. С. 8–9.

7. Монгуш, Ч. М. Метатекстовая разметка в Национальном корпусе тувинского языка: структура и функциональные возможности [Электронный ресурс] /Монгуш Ч.М. // Новые исследования Тувы. 2016, № 4. – Режим доступа: http://nit.tuva.asia/nit/article/view/613. – (дата доступа: 07.05. 2020).