Теоретическая и прикладная лингвистика, 2023, 9 (4), 35–48

УДК 81'33 doi: 10.22250/24107190_2023_9_4_35

Груздев Дмитрий Юрьевич, Коджебаш Дмитрий Олегович
Военный университет
г. Москва, Российская Федерация
gru@inbox.ru

Поиск лингвистической информации в аннотированном электронном корпусе текстов

Аннотация
В статье рассматриваются способы повышения эффективности запросов в специальных корпусах текстов. Исходя из сложившейся в практике профессиональных переводчиков тенденции отказа от поиска по ключевому слову в пользу шаблонных запросов, выдвигаются гипотеза, что регулярные выражения и аннотация, поддерживаемые современными корпус-менеджерами, имеют потенциал для более продуктивного извлечения лингвистической информации из корпуса текстов. Регулярные выражения уже полноценно используются программистами для анализа текста, в то время как аннотация легла в основу всех современных технологий обработки естественного языка. Это может быть использовано для повышения эффективности решения переводческих задач. Очевидной проблемой специальных корпусов является узкое окно релевантности, в результате чего скорость составления и подготовки ресурсов становятся главными критериями наряду с репрезентативностью. Это соответствует требованиям переводчиков, постоянно сталкивающихся с потоком текстов из различных областей, но в то же время препятствует дальнейшему совершенствованию подготовленных корпусов текстов. Таким образом, жизненный цикл большинства специализированных корпусов завершается поиском и аккумуляцией текстов в соответствии со стоящей лингвистической задачей. Однако благодаря автоматизации аннотации дальнейшее совершенствование ресурсов становится перспективным. Чтобы оценить потенциал аннотированного специализированного корпуса, мы провели разметку корпуса новостных текстов за 2019–2020 гг. С помощью морфологической разметки был сгенерирован промежуточный язык между корпус-менеджером и корпусом текстов для реализации сбалансированных запросов для извлечения конструкций и фраз на основе частей речи без привязки к конкретным лексическим единицам. Для проверки гипотезы использовались корпус-менеджер AntConc и программа для морфологической разметки TagAnt. В результате подтвердился потенциал морфологической разметки по подготовке специальных корпусов для шаблонного извлечения лингвистических данных при поиске путей преодоления переводческих трудностей. Составленные конкордансы отличаются более ёмкими и релевантными совпадениями, что способствуют сокращению времени анализа.

Ключевые слова: теги, частеречная разметка, регулярные выражения, извлечение лингвистической информации, запрос, аннотированный корпус

Для цитирования: Gruzdev D. Yu., Kodzhebash D. O. POS-powered queries for neat and lean concordances in ad-hoc corpora analysis // Теоретическая и прикладная лингвистика. 2023. Вып. 9, № 4. С. 35–48. https://doi.org/10.22250/24107190_2023_9_4_35