

ПЛЕНАРНЫЕ ЗАСЕДАНИЯ

УДК 81'33

Алексей Иванович Горожанов, д. филол. н.
Московский государственный лингвистический университет, Москва, Россия
э-почта: a.gorozhanov@linguanet.ru

Alexey Ivanovich Gorozhanov, Dr. of Sc.
(Philology)
Moscow State Linguistic University,
Moscow, Russia
e-mail: a.gorozhanov@linguanet.ru

Иннара Алиевна Гусейнова, д. филол. н.
Московский государственный лингвистический университет, Москва, Россия
э-почта: ginnap@mail.ru

Innara Alievna Guseynova, Dr. of Sc.
(Philology)
Moscow State Linguistic University,
Moscow, Russia
e-mail: ginnap@mail.ru

О ТЕХНОЛОГИИ ОПРЕДЕЛЕНИЯ КОННОТАТИВНЫХ ХАРАКТЕРИСТИК ПУБЛИЦИСТИЧЕСКОГО ТЕКСТА

Описываются этапы технологии для составления специального словаря оценки тональности текста публицистического жанра. Пропагандируется тезис о необходимости такого рода словарей для эффективной работы с отдельными жанрами и издательствами.

Ключевые слова: коннотативные характеристики; сбалансированный лингвистический корпус; тональный словарь; публицистический жанр; новостная статья.

ON THE TECHNOLOGY OF DETERMINING CONNOTATIVE CHARACTERISTICS OF A PUBLICISTIC TEXT

The stages of a technology for compiling a special dictionary for evaluating the tonality of a publicistic text are described. The thesis about the necessity of such dictionaries for effective work at individual genres and publishing houses is promoted.

Key words: connotative characteristics; balanced linguistic corpus; dictionary of tonality; journalistic genre; news article.

Исследования в области анализа тональности текста не теряют своей актуальности в настоящее время. Более того, современные программные методы позволяют постепенно уменьшать долю ручного труда при определении коннотативных характеристик текстов различных жанров.

В настоящей работе предпринимается попытка описать в общих чертах технологию определения коннотативных характеристик текстов публицистического жанра (новостных статей) в рамках корпусного подхода.

В целом определение тональности текста происходит сегодня двумя способами: а) при помощи библиотек обработки естественного языка (технология искусственного интеллекта), б) при помощи тональных словарей.

По нашему мнению, первый способ все еще имеет высокую погрешность и может дать более-менее сносный результат при определении семантического сходства лексических единиц. Говоря о втором способе, мы исходим из того, что универсальные тональные словари (например, семейства SentiLex) плохо подходят для точного анализа текстов определенных жанров, не говоря уже об учете специфики идиостиля издания и конкретного автора.

Далее в круг задач нашего исследования входит развитие уже существующего программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер», разработанного в лаборатории фундаментальных и

прикладных проблем виртуального образования Московского государственного лингвистического университета (Свидетельство Роспатент о регистрации программы для ЭВМ RU 2023683209 от 03.11.2023). Это означает, что технические параметры, связанные с выбором языка программирования (Python), дополнительных библиотек (PyQt, spaCy) и типа базы данных (SQLite), жестко детерминированы, и речь может идти только об интеграции разрабатываемого решения в уже имеющийся рамочный продукт.

Исходя из самого общего определения технологии как совокупности методов и инструментов, служащих для достижения некоторой цели, заключим, что мы имеем только относительно свободный выбор в отношении метода. Представляя последний как путь к достижению цели, выделим следующие этапы этого пути, относящиеся к анализу тональности публицистического (новостного) текста: 1) создание сбалансированного лингвистического корпуса текстов изучаемого новостного издания объемом не менее 1 млн токенов; 2) получение частотного списка прилагательных, существительных (исключая имена собственные) и глаголов из корпуса (именно эти части речи несут основную оценочную нагрузку); 3) ручная разметка полученных единиц коннотативными значениями (например, «-1», «0», «1»), исходя из специфики их употребления в текстах данного жанра и издательства. При возможности одной и той же лексической единицы выражать полярные оценочные значения, такую единицу предпочтительно оставить без разметки, возможно, маркируя ее особо для возможности работы с ней позже; 4) внесение значений в базу данных корпуса; 5) расчет коннотативных параметров с помощью блока функций «Коннотации» корпусного менеджера [1]. 6) анализ полученных данных. Представляется, что этапы 3–6 могут быть повторены несколько раз для корректировки коннотативных значений и разметки тех единиц, которые не были размечены ранее ввиду их многозначности.

Таким образом, целью проводимых итераций будет создание специального тонального словаря, отражающего характер оценки, транслируемой данным изданием.

Перспективой исследования является программная реализация и апробация этапов обозначенной технологии.

ЛИТЕРАТУРА

1. Горожанов А. И., Степанова Д. В. Расширение базы данных сбалансированного лингвистического корпуса значениями тонального словаря (корпусный эксперимент) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 7(888). С. 29–35.