Научная статья
УДК 81’42:811.112.2

Расширение базы данных сбалансированного лингвистического корпуса значениями тонального словаря (корпусный эксперимент)

А. И. Горожанов[1], **Д. В. Степанова**[2]

[1]*Московский государственный лингвистический университет, Москва, Россия, a.gorozhanov@linguanet.ru*
[2]*Минский государственный лингвистический университет, Минск, Республика Беларусь, daryastepanova79@gmail.com*

Аннотация. В предлагаемом исследовании ставится цель разработать и апробировать алгоритм расширения сбалансированного динамического лингвистического корпуса объемом более 3 млн токенов коннотативными характеристиками. При этом авторы опираются на оригинальные программные решения, созданные в лаборатории фундаментальных и прикладных проблем виртуального образования ФГБОУ ВО МГЛУ. В результате получен штатно функционирующий корпус с возможностью дополнения отдельных его фрагментов данными о коннотациях токенов и предложений.

Ключевые слова: корпусная лингвистика, корпусный менеджер, сбалансированный лингвистический корпус, коннотативные характеристики, тональный словарь, немецкий язык, Франкфуртер альгемайне цайтунг

Для цитирования: 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. Вып. 7 (888). С. 29–35.

Original article

Expanding the Database of a Balanced Linguistic Corpus with Values from a Dictionary of Tonality (corpus experiment)

Alexey I. Gorozhanov[1], **Darya V. Stepanova**[2]

[1]*Moscow State Linguistic University, Moscow, Russia, a.gorozhanov@linguanet.ru*
[2]*Minsk State Linguistic University, Minsk, Republic of Belarus, daryastepanova79@gmail.com*

Abstract. The proposed research aims to develop and test an algorithm for expanding a balanced dynamic linguistic corpus of more than 3 million tokens with connotative characteristics. To achieve this, the authors rely on original software solutions created at the laboratory for fundamental and applied issues of virtual education at Moscow State Linguistic University. As a result, a properly functioning corpus was obtained with the ability to supplement its individual fragments with data on the connotations of tokens and sentences.

Keywords: corpus linguistics, corpus manager, balanced linguistic corpus, connotative characteristics, dictionary of tonality, German language, Frankfurter Allgemeine Zeitung

For citation: Gorozhanov, A. I., Stepanova, D. V (2024). Expanding the database of a balanced linguistic corpus with values from a dictionary of tonality (corpus experiment). Vestnik of Moscow State Linguistic University. Humanities, 7(888), 29–35. (In Russ.)

ВВЕДЕНИЕ

Настоящее исследование является продолжением в цепочке прикладных работ, объединенных проблемой автоматической генерации сбалансированного лингвистического корпуса с помощью инструментов обработки естественного языка.

Ранее нам удалось найти технические решения, позволяющие сборку и оперирование лингвистическими корпусами с морфологической разметкой [Горожанов, Гусейнова, Степанова, 2024], в том числе и в динамическом режиме [Степанова, 2023], а также добавить числовые данные о коннотациях токенов и предложений в таблицу базы данных лингвистического корпуса малого объема (до 150 тыс. токенов) [Горожанов, 2023]. Тем не менее практика показывает, что работа с более крупными корпусами (свыше 1 млн токенов) в рамках наших исследований требует отдельного подхода, поскольку стандартная база данных *SQL* еще до расширения дополнительными сведениями занимает ок. 70 Мб дискового пространства на 1 млн размеченных единиц, что делает процесс ее обработки на условно «обычной» ЭВМ весьма времязатратным. Здесь мы исходим из характеристик разрабатываемого в ходе исследования программного обеспечения, одним из требований к которому является возможность функционирования на персональных компьютерах средней мощности.

Одним из требований к программному обеспечению, расширяющему базу данных корпуса дополнительными (коннотативными) характеристиками, является также возможность работы с прерыванием процесса и возобновлением его с указанной оператором точки, например, идентификационного номера предложения.

Отметим, что проблема определения тональности текста (в другой популярной формулировке – сентимент-анализа) является весьма актуальной, о чем свидетельствует значительное количество публикаций в высокорейтинговых журналах (например, [Черничкин, Кривенко, 2023; Комарова, 2023; Глушак, 2023]). Также имеют место исследования, практическим результатом которых является разработка прототипов промышленных систем для проведения автоматического сентимент-анализа [Чернышевич, 2018].

Достижение цели работы планируется в ходе последовательного решения следующих задач:

- экстраполировать процедуру разметки авторского сбалансированного лингвистического корпуса на базу данных объемом более 1 млн токенов;
- провести автоматическую разметку корпуса текстов газеты «Франкфуртер альгемайне цайтунг» (FAZ) коннотативными характеристиками;
- проверить целостность полученной базы данных;
- провести серию пробных запросов к полученной базе данных.

В работе применяются методы экстраполяции и объектно ориентированного программирования, а также корпусный эксперимент.

Лингвистическим материалом исследования являются тексты онлайн-версии немецкоязычной газеты *FAZ*, собранные в корпус за период конца 2023 – начала 2024 годов. Объем корпус составил 200856 предложений или 3653195 токенов. В качестве технических инструментов применяются язык программирования *Python* 3.10, библиотека *PyQt5*, базы данных типа *SQL*, а также словарь *SentiWS*, привлекаемый в качестве экспериментального источника тональных данных для немецкого языка.

ХОД ИССЛЕДОВАНИЯ И ЕГО РЕЗУЛЬТАТЫ

При работе с малым корпусом процедура разметки коннотативных характеристик протекала по следующему алгоритму:

1) чтение исходной базы данных;
2) чтение тонального словаря (трансформирован в формат *CSV*);
3) из таблицы токенов выбираются существительные, прилагательные и глаголы, которым присваивается числовое значение от -1 до 1 либо -1, 0 или 1, в зависимости от характера тонального словаря;
4) из таблицы предложений выбираются все предложения;
5) с помощью данных таблицы токенов для каждого предложения производится расчет пяти показателей:
   а) коннотативной плотности (ρ_{con}) – общего количества положительно и отрицательно коннотированных токенов;
   б) суммы положительно коннотированных токенов;
   в) суммы отрицательно коннотированных токенов;
   г) суммы значений всех коннотированных токенов;
   д) коннотативной амплитуды (A_{con}) – суммы значений по модулю всех коннотированных токенов [Горожанов, 2023, с. 3890];
6) полученные числовые значения записываются в таблицу предложений, в дополнительные пять ячеек.

Эксперимент с малой базой данных показал, что модификация таблицы токенов происходит

относительно быстро, в то время как работа с таблицей предложений занимает значительное время. В этой связи нами было предусмотрена возможность заполнения этой таблицы частями, по диапазонам идентификационных номеров предложений (от 1 до N, где N – количество предложений в корпусе).

Далее в несколько фаз было произведено заполнение базы данных сбалансированного лингвистического корпуса текстов газеты *FAZ* коннотативными характеристиками. В первой фазе была заполнена таблица токенов, в последующих (пяти) фазах – таблица предложений, что в общей сложности заняло ок. 60 часов чистого времени работы программного обеспечения. В среднем доразметка одного предложения заняла 1 сек. Объем базы данных корпуса увеличился с 258 Мб до 301 Мб.

После заполнения база данных была проверена на предмет целостности путем загрузки в авторский корпусный менеджер [Горожанов, Гусейнова, Степанова, 2024, с. 202]. Загрузка произошла штатно, без выдачи ошибок (см. рис. 1).

Для решения четвертой задачи исследования необходимо было провести ряд поисковых запросов к базе данных с участием расширенных ячеек. Подчеркнем, что наша цель здесь проверить скорее техническую сторону созданного программного решения, поскольку смысловая компонента тесно связана с качеством исходного тонального словаря, которое в рамках нашего исследования не подвергалось проверке.

В качестве первого запроса узнаем общую тональную характеристику корпуса. Для этого воспользуемся функцией ручного запроса корпусного менеджера:

SELECT COUNT (*) FROM tokens WHERE tokenoption01 > 0.0

Результат:
Токенов: 128711 (3,52 %).

Таким образом, 3,25 % всех токенов базы данных получили положительную коннотативную характеристику.

Для получения количества отрицательно окрашенных токенов применим следующий запрос:

SELECT COUNT (*) FROM tokens WHERE tokenoption01 < 0.0

Результат:
Токенов: 73200 (2,0 %)

Полученные данные позволяют заключить, что (согласно данным разметки) корпус имеет преимущественно положительную коннотацию.

Перейдем к таблице предложений. Построим запросы к сумме значений всех коннотированных токенов, в котором проверим положительные, отрицательные и нулевые значения:

SELECT COUNT (*) FROM sents WHERE sentoption04 > 0.0

Результат:
Предложений: 69422
Предложений всего: 200856

SELECT COUNT (*) FROM sents WHERE sentoption04 < 0.0

Результат:
Предложений: 49898
Предложений всего: 200856

SELECT COUNT (*) FROM sents WHERE sentoption04 = 0.0

Результат:
Предложений: 81536
Предложений всего: 200856

Здесь фиксируем превалирование положительных значений над отрицательными. Сложив результаты, получим 69422 + 49898 + 81536 = 200856,

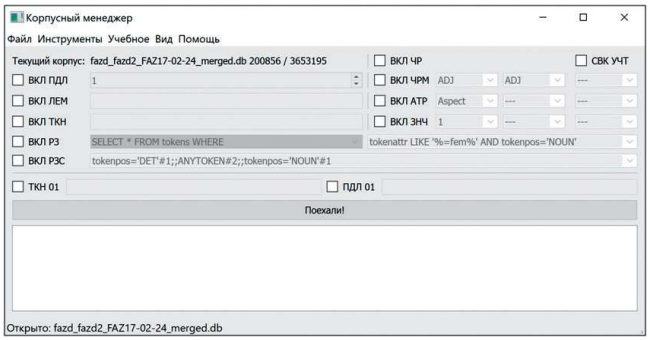

Рис. 1. Интерфейс корпусного менеджера после загрузки расширенной базы данных fazd_fazd2_FAZ17-02-24_merged.db

что дополнительно подтверждает правильность процедуры разметки.

Самым «отрицательным» предложением оказался образец № 88789:

> *Verboten* sein sollen demnach fortan: Videos, die *falsche* Informationen zur Prävention von *Krankheiten* verbreiten, Videos, die *unwirksame* Therapien für *Krankheiten* bewerben sowie Videos, die gar die Existenz von *Krankheiten leugnen*. – В будущем следует запретить: видео, распространяющие ложную информацию о профилактике заболеваний, видео, рекламирующие неэффективные методы лечения болезней, и видео, которые даже отрицают существование болезней[1].

Предложение имеет семь отрицательно коннотированных токенов (выделены полужирным) с суммой значений -3,315.

Самым «положительным» стало предложение № 24556 с шестью положительными токенами (выделены полужирным) и с суммой значений 1,9108:

> Der vorgeschlagene Pakt *mag* nicht *perfekt* sein, aber wenn 27 Staaten über ein so *sensibles* Thema *verhandeln*, kann es kein *perfektes* Ergebnis geben, daher *unterstütze* ich ihn im Großen und Ganzen. – Предлагаемый пакт, возможно, не идеален, но когда 27 государств ведут переговоры по такому деликатному вопросу, идеального результата быть не может, поэтому я, в общем и целом, его поддерживаю.

Обратим внимание на то, что отрицание *nicht* не получило никакой коннотации, а глагол *mögen* употребляется не в прямом значении, а в значении предположения, что может сказываться на валидности интерпретации этого и других образцов. Однако, как было отмечено выше, проблема качества исходного тонального словаря в настоящем исследовании нами не рассматривается и может быть подвергнута изучению в отдельной научной работе.

В ходе эксперимента по вводу запросов к таблице предложений выяснилось, что работа со знаками «больше» и «меньше» имеет свою специфику, что обусловлено особенностями языка SQL. Например, при запросе:

SELECT COUNT (*) FROM sents WHERE sentoption04 < -2.0

– мы получим результат 49847 предложений, что не соответствует действительности, так как таких предложений в корпусе только 51 единица. Для корректного результата необходимо изменить синтаксис запроса на следующий:

SELECT COUNT (*) FROM sents WHERE sentoption04 BETWEEN -2.0 AND -4.0

Предпримем далее попытку осуществить более предметный поиск, в рамках которого возможно было бы оценить контекст определенных лексических единиц. Например, возьмем наименования стран: *Deutschland* (*Германия*), *Frankreich* (*Франция*), *Türkei* (*Турция*).

Простой поиск по лемме показывает, что эти существительные упоминаются в корпусе в 3937, 636 и 242 предложениях соответственно. Оценим коннотации этих предложений, для чего в очередной раз воспользуемся функцией ручного запроса. Найдем все предложения с отрицательной суммарной коннотацией, в которых встречается существительное *Deutschland*. Поскольку поиск производится по таблице предложений, а не по таблице токенов, и при этом поиск осуществляется одной строкой, то наименование страны необходимо будет искать как последовательность символов в ячейке, содержащей текст предложения, исключив первый символ:

SELECT COUNT (*) FROM sents WHERE sentoption04 < 0.0 AND senttext LIKE '%eutschland%'

Результат:
Предложений: 1109
Предложений всего: 200856

Проведем аналогичные запросы для положительных и нулевых значений:

SELECT COUNT (*) FROM sents WHERE sentoption04 > 0.0 AND senttext LIKE '%eutschland%'

Результат:
Предложений: 1705
Предложений всего: 200856

SELECT COUNT (*) FROM sents WHERE sentoption04 = 0.0 AND senttext LIKE '%eutschland%'

Результат:
Предложений: 1390
Предложений всего: 200856

Результат говорит о том, что положительные и нейтральные значения превалируют.

Объединим эти данные с показателями для Франции и Турции и представим в табличной форме (табл. 1).

[1]Зд. и далее перевод наш. – *А. Г., Д. С.*

Таблица 1

КОННОТАТИВНЫЕ ХАРАКТЕРИСТИКИ ПО СТРАНАМ

Страна	+	–	0	Вывод
Deutschland	1705	1109	1390	+
Frankreich	225	201	226	0/+
Türkei	82	77	89	+

Подчеркнем, что таблица отражает некоторые данные о контекстах только существительных, обозначающих страны, а не о том, как в целом представлена страна в текстах корпуса. Для понимания последнего было бы необходимо работать менее формально и учесть целую группу лемм, образующих семантическое поле, связанное с Германией, Францией или Турцией, и охватывающих несколько частей речи.

Полученные данные говорят о том, что, исходя из маркировок в используемом тональном словаре, топоним *Deutschland* лидирует по положительным контекстам употребления относительно отрицательных. *Türkei* также преимущественно упоминается в положительных контекстах, а *Frankreich* имеет скорее нейтрально-положительное окружение.

Приведем несколько примеров. Образец № 187937 (Deutschland) имеет суммарную коннотацию 1,2644 (4 «положительных» токена, выделены полужирным):

> Deutschland ist ein toller Standort, ein *guter* Standort, aber es ist jetzt *wichtig*, dass wir die *richtigen* Dinge tun, damit er wettbewerbsfähig bleibt. – Германия – крутое место, хорошее место, но сейчас важно, чтобы мы делали правильные вещи, чтобы она оставалась конкурентоспособной.

Образец № 165619 187937 (Deutschland) имеет суммарную коннотацию -1,8059 (4 «отрицательных» токена, выделены полужирным):

> Einerseits trifft die Charakterisierung unzweifelhaft die Welt eines Politikers, der in der Wiedervereinigung, in Europa, in *Krisen* aller Art, aber auch im Alltag seines politischen Wirkens die demokratische und wirtschaftliche Ordnung bewahren wollte, die Deutschland nicht, wie mancher „Konservative" meint, *schwach* und *dekadent* zu machen *droht*, sondern so lebenswert hat werden lassen wie nie zuvor. – С одной стороны, характеристика, несомненно, отражает мир политика, который после падения Берлинской стены, в Европе, в кризисах всех видов и также в повседневной политической жизни хотел сохранить демократический и экономический порядок, который не угрожает сделать Германию, как полагают многие «консерваторы», слабой и упадочной, но комфортной для жизни, как никогда раньше.

Примечательно, что существительное *Wiedervereinigung* (обычно переводится как «падение Берлинской стены» или «объединение Германии», выделено подчеркиванием) имеет в тональном словаре положительную коннотацию равную 0,004.

ЗАКЛЮЧЕНИЕ

Итак, цель нашего исследования достигнута. Во-первых, процедура разметки авторского сбалансированного лингвистического корпуса была экстраполирована на базу данных объемом 3653195 токенов, в ходе выполнения этой задачи при разметке таблицы предложений была предусмотрена работа с заданным диапазоном, чтобы оператор имел возможность прервать и затем возобновить процесс из любой точки; во-вторых, была проведена автоматическая разметка корпуса текстов газеты FAZ коннотативными характеристиками, при этом был использован тональный словарь немецких существительных, прилагательных и глаголов с числовыми значениями от -1 до 1; в-третьих, была проверена целостность полученной базы данных; и в-четвертых, была проведена серия пробных запросов к полученной базе данных, результаты которых показали состоятельность предложенного программного решения.

Кроме того, наличие функции ручного запроса позволило избежать доработки корпусного менеджера для работы с базой данных, расширенной коннотативными характеристиками, что доказывает универсальность корпусного менеджера как системы управления базами данных с вариативными компонентами.

Перспективой исследования могут стать такие направления, как добавление в тональных словарь других частей речи (например, частиц) и в целом – создание собственных тональных словарей, а также совершенствование программного обеспечения для генерации сбалансированных лингвистических корпусов и осуществления поисковых запросов к ним.

СПИСОК ИСТОЧНИКОВ

1. Горожанов А. И., Гусейнова И. А., Степанова Д. В. Обработка естественного языка и художественный текст: база для корпусного исследования // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2024. Т. 15. №1. С. 195–210. DOI 10.22363/2313-2299-2024-15-1-195-210.
2. Степанова Д. В. Программный комплекс для генерации динамического корпуса текстов СМИ // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2023. № 6 (127). С. 123–130. EDN FMBTKO.
3. Горожанов А. И. Расширение стандартного сбалансированного лингвистического корпуса, построенного по правилам spaCy, коннотативными характеристиками // Филологические науки. Вопросы теории и практики. 2023. Т. 16. № 11. С. 3888–3893. DOI 10.30853/phil20230594. EDN FVUIUL.
4. Черничкин Д. А., Кривенко А. И. Медиаобраз России в казахстанских телеграм-каналах // Политическая экспертиза: ПОЛИТЭКС. 2023. Т. 19. № 4. С. 565–586. DOI 10.21638/spbu23.2023.404. EDN POURDG.
5. Комарова Е. В. Проблема цифрового этикета в русских и английских медиатекстах: на материале миграционного дискурса // Медиалингвистика. 2023. Т. 10. № 2. С. 253–264. DOI 10.21638/spbu22.2023.207. EDN MFJOQV.
6. Глушак В. М. Отрицание немецких полярных слов и выражений в автоматизированном анализе тональности текста // Филологические науки. Вопросы теории и практики. 2023. Т. 16. № 10. С. 3287–3292. DOI 10.30853/phil20230510. EDN CWDXEU.
7. Чернышевич М. В. Принципиальная схема решения задачи ACAT и его лингвистическое обеспечение // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2018. № 3 (94). С. 72–80. EDN WXUUJR.

REFERENCES

1. Gorozhanov, A. I., Guseynova, I. A., Stepanova, D. V. (2024). Natural Language Processing and Fiction Text: Basis for Corpus Research. RUDN Journal Of Language Studies, Semiotics And Semantics, 15(1), 195–210. DOI 10.22363/2313-2299-2024-15-1-195-210.
2. Stepanova, D. V. (2023). Software package for generating a dynamic media texts corpus. Minsk State Linguistic University Bulletin. Series 1. Philology, 6(127), 123–130. EDN FMBTKO. (In Russ.)
3. Gorozhanov, A. I. (2023). Extension of a standard balanced linguistic corpus built according to spaCy rules by connotative characteristics. Philology. Theory & Practice, 11(16), 3888–3893. DOI 10.30853/phil20230594. EDN FVUIUL. (In Russ.)
4. Chernichkin, D. A., Krivenko, A. I. (2023). Media image of Russia in Kazakh Telegram channels. Political Expertise: Politex, 4(19), 565–586. DOI 10.21638/spbu23.2023.404. EDN POURDG. (In Russ.)
5. Komarova, E. V. (2023). Digital ethics challengers in Russian and English media texts: Migrant Discourse Case Study. Media Linguistics, 2(10), 253–264. DOI 10.21638/spbu22.2023.207. EDN MFJOQV. (In Russ.)
6. Glushak, V. M. (2023). Negation of German polar words and expressions in automated analysis of text tonality. Philology. Theory & Practice, 10(16), 3287–3292. DOI 10.30853/phil20230510. EDN CWDXEU. (In Russ.)
7. Chernyshevich, M. V. (2018). The architecture of sentiment-analysis system and its linguistic resources. Minsk State Linguistic University Bulletin. Series 1. Philology, 3(94), 72–80. EDN WXUUJR. (In Russ.)

ИНФОРМАЦИЯ ОБ АВТОРАХ

Горожанов Алексей Иванович
доктор филологических наук, доцент,
профессор кафедры грамматики и истории немецкого языка факультета немецкого языка
Московского государственного лингвистического университета

Степанова Дарья Валерьевна
кандидат филологических наук, доцент
доцент кафедры теории и практики английской речи факультета английского языка
Минского государственного лингвистического университета

INFORMATION ABOUT THE AUTHORS

Gorozhanov Alexey Ivanovich
Doctor of Philology (Dr. habil), Associate Prof. ,
Professor in the Department of German Language Grammar and History
Faculty for German Language
Moscow State Linguistic University

Stepanova Darya Valeryevna
PhD (Philology), Associate Prof.
Associate Professor in the Department of Theory and Practice of English Speech
Faculty for English Language
Minsk State Linguistic University

Статья поступила в редакцию	01.04.2024	The article was submitted
одобрена после рецензирования	29.04.2024	approved after reviewing
принята к публикации	06.05.2024	accepted for publication