

RU

Группирование токенов в сбалансированном лингвистическом корпусе по семантическому признаку

Горожанов А. И.

Аннотация. Цель исследования – разработать процедуру группирования токенов сбалансированного лингвистического корпуса, исходя из их лексического значения. Научная новизна заключается в том, что впервые на платформе программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» в полностью автоматическом режиме тестируется технология выявления значений лексических единиц, формируя тем самым принципы построения семантических рядов с опорой на инструменты управляемого искусственного интеллекта и в русле корпусного подхода. В результате был написан и интегрирован в корпусный менеджер программный код, позволивший автоматически получить более 1700 релевантных групп имен существительных размерностью от двух до 174 единиц, более 650 групп имен прилагательных размерностью от двух до 79 единиц и более 460 групп глаголов размерностью так же от двух и до 12 единиц на материале немецкоязычного корпуса новостных текстов объемом около 4,18 млн токенов.

EN

Grouping tokens in a balanced linguistic corpus based on semantic criteria

A. I. Gorozhanov

Abstract. The aim of the study is to develop a procedure for grouping tokens from a balanced linguistic corpus based on their lexical meaning. The scientific novelty lies in the first-time fully automated testing – using the “Balanced Linguistic Corpus Generator and Corpus Manager” software package – of a technology for identifying the meanings of lexical units; this establishes principles for constructing semantic series by leveraging managed artificial intelligence tools within a corpus-based framework. As a result, software code was written and integrated into the corpus manager, enabling the automatic generation of over 1,700 relevant noun groups (ranging from 2 to 174 units), over 650 adjective groups (from 2 to 79 units), and over 460 verb groups (from 2 to 12 units) from a German-language news corpus comprising about 4.18 million tokens.

Введение

Корпусные исследования традиционно строятся на процедурах извлечения статистических данных с их последующей мысленной интерпретацией человеком. При этом семантический компонент реализуется эксплицитно, т. е. ученый, как правило, заранее формирует списки лексем, которые подлежат поиску в корпусах (Каменский, Бредихин, 2025; Степанова, 2024а; Бондарчук, Грачев, 2024), или даже работает с одной лексемой и производными от нее (Ермолова, 2026; Степанова, 2026b).

Обилие лингвистических исследований, опирающихся на методы корпусного подхода, а именно – работы, касающиеся семантики, свидетельствует о востребованности нашего исследования.

Актуальность выбранного нами тематического направления обусловлена объективной необходимостью внедрения систем искусственного интеллекта в ключевые области развития нашей страны, в первую очередь, в связи с программным выступлением Президента Российской Федерации В. В. Путина в рамках пленарного заседания Евразийского экономического форума. По словам президента, «к 2030 году подобные технологии и продукты на их основе должны использоваться во всех областях, включая производство, логистику, энергетику, управление и образование» (Доброва, 2026). Также, по нашему мнению, в настоящее время необходима разработка систем управляемого искусственного интеллекта, которые смогут сместить баланс пользовательского интереса в сфере науки с генеративных моделей, оперирующих вопросо-ответной формой, на специализированные ИИ-решения, выполняющие строго регламентированные задачи.

Теоретическую базу предлагаемого исследования формируют современные научные и научно-прикладные работы из области компьютерной лингвистики.

Среди таких работ можно выделить несколько направлений. Самую многочисленную группу образуют публикации, посвященные описанию текущих значений слов, а также изменениям этих значений. Так, Ю. В. Шуйская, работая с Национальным корпусом русского языка (НКРЯ), описывает «изменение значений таких слов, как “благотворительный” и “благотворительность”, в связи с фактами институционализации благотворительности как общественной деятельности в истории России» (2026, с. 110). С. С. Нимаева (2026) с помощью того же корпуса рассматривает динамику семантической деривации лексемы «вдруг» в русском языке. Корпус позволяет получить около 10 тысяч контекстов за 1766–2011 гг., что говорит о высокой степени достоверности и верифицируемости полученных результатов (Нимаева, 2026, с. 26). М. Чжао (2025) находит с помощью НКРЯ доказательства изменений в структуре и содержании семантического поля «чистота» на рубеже XX–XXI веков. При этом большое внимание уделяется переносным значениям и частотности коллокаций (Чжао, 2025).

Д. В. Степанова исследует актуальные образцы употребления лексемы “security” в сбалансированном лингвистическом синхроническом корпусе «актуальных англоязычных текстов новостной сети США CNN за период с января 2025 года по февраль 2026 года, общим объемом 120 942 предложения или более 3 млн токенов» (2026а, с. 218), что говорит в пользу работы не только с национальными корпусами, но и специальными авторскими корпусными решениями.

Сходным по технологическим параметрам является исследование О. И. Писарик (2024), в котором проводится тщательный корпусный и контекстуальный анализ контекстов немецкоязычных СМИ, содержащих лемму “Kunst”.

Отметим важность работы Е. В. Рахилиной и В. А. Плунгяна, в которой постулируется тренд на развитие цифровой лексикографии в условиях, когда происходит «интеграция словаря и корпуса» (2025, с. 363).

Также мы встречаем публикации, обращенные к вопросам стилистики, которые, однако, основываются на корпусных данных и затрагивают семантический компонент. А. Ж. Сейсебаев и Е. Д. Сулейменова привлекают гигакорпусы английского языка и предлагают классификацию оксюморонов, которая «основана на семантическом противостоянии компонентов, степени контекстуальной зависимости и корпусной частотности как показателе лексикализации» (Seisebayev, Suleimenova, 2026, p. 1110). А. С. Кожухметова и А. К. Казкенова (2025) описывают корпусную базу данных «Эмоции в метафорическом представлении», охватывающую четыре языка: английский, казахский, немецкий и русский. Среди параметров аннотирования выделяются «семантические роли» лексем (Кожухметова, Казкенова, 2025, с. 71).

Непосредственно проблеме семантической разметки корпусов посвящено относительно немного работ. М. А. Кузина предпринимает попытку ручной разметки новостного текстового массива тегами, касающимися проявления идеологических концептов: «успех», «скорость», «реванш», «превосходство» и пр. Разметке подлежат не только отдельные лексемы, но и словосочетания. Процедуры автоматизации проявляются в виде регулярных выражений Python (Кузина, 2026). А. А. Михеев и Е. С. Ярыгина (2024) предлагают создать новый тип семантической разметки текстов на основе онтологического подхода, например для поиска схожих событий в различных текстах. Однако работа имеет обзорно-аналитический характер и пока еще не дает готового технического решения (Михеев, Ярыгина, 2024).

Таким образом, мы можем заключить, что при высоком интересе к проблемам семантической разметки и семантического анализа лингвистических корпусов наблюдается нехватка программных авторских решений. В попытке частично закрыть эту лауну мы ставим перед собой следующие задачи:

- 1) собрать лингвистический материал для апробации созданного технического решения;
- 2) написать программные функции автоматического группирования токенов;
- 3) провести тестирование функций;
- 4) встроить созданные функции в интерфейс программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер».

При решении задач, определенных в работе, использовались следующие методы исследования: методы профессионально ориентированного программирования, обработки естественного языка и дистрибутивной семантики (на основе векторных представлений) используются для сбора лингвистического материала и написания алгоритмов; методы тестирования и эксперимента привлекаются в связи с апробацией написанных функций; методы системной интеграции программных продуктов и моделирования архитектуры графического интерфейса пользователя послужили решению задачи встраивания полученных технологических решений в корпусный менеджер.

Лингвистическим материалом исследования выступают немецкоязычные новостные тексты издания “Focus” за период с 14.02.2026 по 04.10.2026, собранные в сбалансированный лингвистический корпус объемом 4 180 446 токенов или 266 880 контекстов (предложений). Технологическим материалом явились программные коды корпусного менеджера размером около 230 строк.

Представляется, что результаты исследования применимы в практическом плане и могут быть использованы для оперативного анализа новостного потока, а также для изучения черт идиостиля авторов произведений художественной литературы.

Обсуждение и результаты

Технически для работы со значениями лексем в корпусе мы прибегнем к *векторным представлениям слов*, или *эмбедингам*.

В отечественном научном поле распространено применение технологии Word2Vec (Колосов, Майсурадзе, 2026; Арабов, Седых, 2025; Вершинин, 2022). Тем не менее мы продолжаем использовать spaCy с библиотекой fastText по ряду причин. Во-первых, spaCy является основным инструментом обработки естественного языка программного комплекса. Во-вторых, векторные модели spaCy уже предварительно обучены и готовы к применению. В-третьих, fastText работает с n-граммами, что позволяет учитывать составные слова.

Поскольку в рамках настоящей работы мы оперируем немецкоязычными текстами, нами была выбрана модель spaCy de_core_news_lg, содержащая 300-мерные векторные представления n-грамм. Согласно принципу безопасного функционирования, заложенного в основу программного комплекса, модель установлена локально, поэтому ее использование не требует подключения к сети Интернет. Сбор лингвистического материала производился с веб-сайта журнала «Фокус» (Focus-online...) с помощью специализированной авторской программы на Python. Процедура сбора такова, что программа выходит на главную веб-страницу журнала, ищет на ней новостные заголовки, сверяется со своей базой (файлом TXT) на предмет совпадений и переходит далее на веб-страницы новых статей. Там производится чтение текста новости, он добавляется в новый файл TXT и сохраняется с временной меткой, например, "articlesFO_2026-09-27_21-59-13.txt". Сбор новостей осуществляется сплошной выборкой, без разделения на рубрики («Политика», «Финансы», «Здоровье», «Спорт» и пр.), поскольку ставится задача набрать как можно большее количество текстовых данных. Таким образом, при регулярном старте программы возможно накопить значительные объемы текста. По нашим наблюдениям, каждый день прирост новых текстов составляет 200-250 КБ.

Сбалансированный лингвистический корпус был создан в автоматическом режиме из «сырых» файлов TXT новостей с помощью программы-генератора (Степанова, 2024b), которая после перехода к версии 2.0 увеличила производительность сборки, в зависимости от этапа, в 100-120 раз. Накопленные и отобранные в отдельную папку файлы TXT анализируются по очереди генератором и добавляются в базу данных SQLite сплошным текстовым потоком. Таким образом, исследователь может выбирать временной диапазон новостей и примерный размер будущего корпуса.

Загрузка базы данных в корпусный менеджер показала отсутствие ошибок при сборке.

Построение программных функций автоматического группирования предварил подготовительный этап экспериментов по получению коэффициентов семантического сходства между текстами различной длины, получаемыми с помощью spaCy.

В первую очередь были проверены одинаковые слова и словосочетания. Полученный результат очевиден:

Baum: Baum (дерево) → 100%;

gehen: gehen (идти) → 100%;

rot: rot (красный) → 100%;

die schöne Natur: die schöne Natur (прекрасная природа) → 100%.

В самом деле, каждое слово или словосочетание заменяемы на самих себя в любом возможном контексте. Подчеркнем, что spaCy позволяет сравнивать между собой не только слова, но и словосочетания и целые контексты с помощью одного и того же метода similarity().

Поскольку fastText работает с n-граммами, мы рассмотрели односоставные слова и композиты, например:

Meistersinger: Meister → 72%;

Linde: Lindenbaum → 66%;

Naturprodukt: Ersatzprodukt → 63%;

Baum: Baumwolle → 13%;

Eiche: Eichhörnchen → 38%.

Примеры показывают, что сходства высчитываются не только на основании сравнения простых последовательностей символов в словах, но также и с учетом значений, точнее говоря – с учетом возможности взаимозаменяемости сравниваемых элементов в равных контекстах (ср. парадигматические отношения по Ф. де Соссюру). Слова скорее взаимозаменяемы при коэффициенте более 50%: Meistersinger (мейстерзингер) и Meister (мастер); Linde (липа) и Lindenbaum (липа); Naturprodukt (натуральный продукт) и Ersatzprodukt (продукт-заменитель) (здесь и далее перевод наш. – А. Г.). И наоборот, внешне похожи, но не взаимозаменяемы: Baum (дерево) и Baumwolle (хлопок); Eiche (дуб); Eichhörnchen (белка).

Пары из одного семантического ряда также показывают высокое сходство (более 50%):

строительные материалы – Beton (бетон): Zement (цемент) → 72%;

овощи – Gurke (огурец): Tomate (помидор) → 76%;

цветы – Tulpe (тюльпан): Rose (роза) → 65%;

средства передвижения – Flugzeug (самолет): Auto (автомобиль) → 66%;

эмоциональные состояния – Zorn (гнев): Wut (ярость) → 73%

и т. д.

Информативно также сравнение целых предложений:

Der Himmel war blau: Der Himmel ist blau → 91%.

Здесь при различии только во временной форме глагола «быть» (Небо было/*есть голубое) указывается на почти полное совпадение.

Или:

Ich kann nicht umhin die Natur zu lieben: Ich liebe die Natur sehr → 68%.

Несмотря на употребление другой конструкции (Я не могу не любить природу: Я люблю природу), коэффициент имеет довольно высокое значение.

Неожиданно высокий коэффициент оказывается и при употреблении фразеологии:

Er rannte Hals über Kopf: Er rannte blitzschnell → 85% (Он бежал сломя голову: Он бежал быстро как молния);

Er wurde rot wie eine rote Bete: Er schämte sich → 71% (Он покраснел как свекла: Он застыдился);

Wir sprachen unter vier Augen: Wir redeten ohne Augenzeugen → 71% (Мы разговаривали с глазу на глаз: Мы говорили без свидетелей).

При этом отдельное сравнение обстоятельств дает очень низкий коэффициент сходства:

unter vier Augen: ohne Augenzeugen → 28%.

Итак, встроенные эмбединги spaCy для немецкого языка позволяют работать как со словами, так и со словосочетаниями и даже в какой-то мере учитывают фразеологию, показывая если не точные синонимические отношения, то, во всяком случае, вероятность встретиться в том же окружении, что по результату может расцениваться как возможность наличия общей семы.

Затем был осуществлен переход к этапу моделирования алгоритма. Идея состояла в том, чтобы выделить из корпуса знаменательные части речи (или одну из них), привести полученные токены к леммам и составить их список без дублирования (планировалась функция `get_unique_lemmas()` → `List[str]`).

Далее из этого списка уникальных лемм каждая пара слов должна быть оценена на предмет семантического сходства при помощи метода `similarity()`. Пары не должны дублироваться, т. е. если уже проводилось сравнение пары «токен1-токен2», то сравнение «токен2-токен1» не должно иметь место (планировалась функция `get_vectors()` → `Tuple[np.ndarrays, List[str]]`).

Непосредственно для группирования лексем предусматривалось наличие порогового значения, которое должно выступать аргументом функции `group()` → `List[List[str]]`. Функция возвращает список групп лемм, объединенных высоким семантическим сходством между собой, т. е. в одной группе каждая из возможных пар лемм имеет значение коэффициента выше порогового.

После моделирования началось написание и тестирование программного кода запланированных функций. В итоге был создан и отлажен класс `wordGrouper`, сохраненный в отдельном файле `vector.py`. Таким образом, задачи 2 и 3 были решены.

Встраивание новых функций в интерфейс корпус нового менеджера проходило в два этапа. Сперва было модифицировано главное меню, в группу «Семантика» которого был добавлен пункт «Группирование ЧР» (Иллюстрация 1).

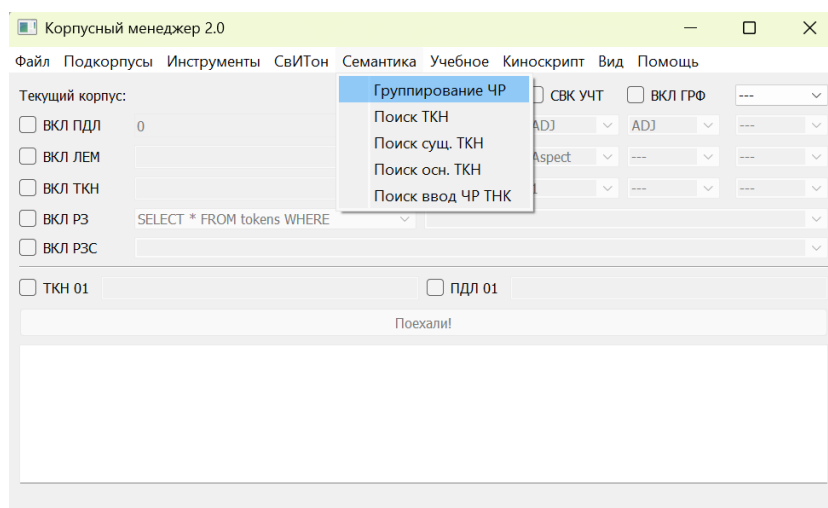


Иллюстрация 1. Пункт «Группирование ЧР» группы главного меню «Семантика» корпусного менеджера программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер»

Затем в главном файле программного комплекса был создан новый сигнал, привязанный к функции `semanticaAutoGroups()`, которая импортирует класс `wordGrouper`, запрашивает значения ниспадающих меню частей речи с флажками ВКЛ ЧР и ВКЛ ЧРМ и производит вывод результата в текстовую область корпусного менеджера. В зависимости от выставленных флажков и выбранных пунктов меню группирование производится по одному, двум или трем частям речи. Программа сама определяет язык корпуса с помощью модуля `detect` и в зависимости от этого выбирает модель spaCy с нужными предобученными эмбедингами. Пороговое значение по умолчанию было выставлено на 75%.

Приведем результаты тестирования функций класса `wordGrouper` на сбалансированном лингвистическом корпусе немецких СМИ: для имен существительных, имен прилагательных и глаголов.

В результате группирования по именам существительным было составлено 1723 группы. Первая группа оказалась самой многочисленной и получила рабочее название *обесцвеченной группы*, поскольку включила 29 218 токенов, которые не имеют между собой связи выше 75%. Например:

Bundesapothekerkammer (Федеральная палата аптекарей): Dieselmotor (дизельный двигатель) → 11%
 Dividendenerhöhung (повышение дивидендов): Lederlenkrad (кожаный руль) → 7%
 и т. п.

Обесцвеченная группа выводится только по той причине, что в дальнейшем планируется разработать алгоритмы ее анализа на установление каких-либо семантических сходств, которые, по предварительной оценке, могут доходить в отдельных случаях до 30%. Приведем примеры некоторых полученных групп, следующих за первой группой.

Группу № 2 (174 слова) составляют существительные типа “11-Jährige” (11-летний), “12-Jährige” (12-летний), “13-Jährige” (13-летний) и пр.

Группу № 4 (28 слов) образуют географические наименования, национальности и названия языков: “Griechisch” (греческий), “Hessen” (Гессен), “Lateinisch” (латынь), “Mitteldeutschland” (средняя Германия) и др.

Группа № 5 (20 слов) состоит из существительных с компонентом “Name” (имя): “Doppelname” (двойное имя / двойная фамилия), “Familiennname” (фамилия), “Frauennname” (женское имя), “Geburtsname” (имя, данное при рождении) и пр.

Группу № 12 (13 слов) составляют животные: “Ameise” (муравей), “Echse” (ящерица), “Schildkröte” (черепаха), “Spinne” (паук) и др. Интересно, что птицы и бабочки были выделены в отдельную группу.

Группа № 13 (12 слов) состоит из профессий: “Bauer” (крестьянин), “Bergmann” (горняк), “Förster” (лесник), “Schneider” (портной) и пр.

Группа № 14 (11 слов) – это единицы измерений веса и объема: “Deziliter” (децилитр), “Gramm” (грамм), “Kilogramm” (килограмм), “Liter” (литр), “Mikrogramm” (микрограмм) и т. д.

Интересными для анализа новостного дискурса представляются группы с общим наименованием «виды мусора», «обувь» или «валюты».

Программа сгруппировала и пары, которые представляют собой разное написание одного и того же слова: “Plug-In” и “Plug-in” или “Nuss” и “Nuß”.

Для имен прилагательных было получено 659 групп. Аналогичным образом первая группа вместила 3061 слово, проявляющее слабые семантические связи, например: angewandt (прикладной): demografisch (демографический) → 21%.

Далее наиболее релевантными для нас группами оказались, например:

№ 6 (18 слов) «размерность»: “niedrig” (низкий), “gering” (незначительный), “groß” (большой), “hoch” (высокий).

№ 7 (17 слов) «питательный»: “eiweißreich” (богатый белком), “fettarm” (обезжиренный), “fetthaltig” (жиросодержащий), “fettreich” (жирный), “kalorienarm” (низкокалорийный), “kalorienreich” (калорийный).

№ 12 (10 слов) «профессиональный»: “fachkundig” (компетентный), “inkompetent” (некомпетентный), “kompetent” (компетентный), “kundig” (сведущий), “professionell” (профессиональный). Заметим, что в одном ряду оказались антонимы “kompetent”/“inkompetent”.

№ 54 (5 слов) «спортивный»: “athletisch” (атлетический), “muskulös” (мускулистый), “sportiv” (спортивный), “sportlich” (спортивный), “unsportlich” (неспортивный).

Группа № 108 (4 слова) показала погрешность в работе n-грамм, поскольку в ней были объединены прилагательные “astronomisch” (астрономический) и “gastronomisch” (гастрономический).

Наконец, часть речи «глагол» дала 466 групп. Обесцвеченная группа содержала 4062 слова, например: stattfinden (находиться): fahren (ехать) → 38% или beweisen (доказывать): säubern (чистить) → 16%.

Группа № 2 «копать» включила 12 слов, например: “buddeln” (копать), “graben” (копать), “ausgraben” (выкапывать), “umgraben” (перекапывать).

Группу № 6 (7 слов) «зарабатывать» составили такие глаголы, как “verdienen” (зарабатывать), “dazuverdienen” (дополнительно зарабатывать), “mitverdienen” (иметь долю в заработке).

Группа № 30 (5 слов) «ощущать» включила, например, “schmecken” (иметь вкус), “riechen” (нюхать), а также их словоформы.

В принципе, работа с глаголами оказалась самой неинформативной, с высоким уровнем «шума» из-за попадания в одну группу словоформ одних и тех же глаголов (“pushen/pusht”, “steh/stehn”, “bezuschussen/bezuschusst” и др.), что говорит о более высокой доле погрешности в работе лемматизатора именно с этой частью речи.

Штатная работа алгоритмов группирования позволяет говорить о правильности выбора технологических решений и о высокой степени достоверности полученных результатов.

Заключение

Итак, все задачи нашего исследования были выполнены, пусть и с различной степенью предсказуемости, что в целом характерно для современных систем, базирующихся на технологиях искусственного интеллекта.

Прежде всего был собран объемный текстовый материал исследования, причем его организация в сбалансированный лингвистический корпус заняла около 40 минут работы программы-генератора при объеме около 4,18 млн токенов (ЭВМ Lenovo Y7000P IRX10, Intel Core i9-14900HX 2.20 GHz, RAM 16 GB).

Далее в несколько этапов была протестирована система векторных представлений слов библиотеки spaCy, написан и апробирован программный код, что позволило выявить как достоинства, так и недостатки оперирования эмбедингами fastText. К достоинствам мы отнесем:

- способность работать с целыми предложениями, а не только со словами;
- возможность выявлять составные слова за счет n-грамм;
- наличие векторного представления слов для нескольких языков;
- относительно высокую скорость работы (менее одной минуты на всю процедуру группирования);
- совместимость с библиотеками Python.

Недостатками, на наш взгляд, являются:

- наличие погрешностей;
- работа не собственно с семантикой, а со статистической вероятностью нахождения в одном и том же контексте;
- галлюцинации n-грамм («астрономический»/«гастрономический»);

Впрочем, глубокое понимание недостатков в совокупности с фундаментальной лингвистической подготовкой и высокой технологической компетентностью может обратить их в достоинства в руках умелого исследователя.

Наконец, созданный класс Python был встроен в интерфейс программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер», что делает работу по группированию удобной для оператора.

Важно, что проведенное и описанное выше исследование протекало в русле корпусного подхода, в рамках которого корпус выступает стимулом к получению качественно нового результата (*англ.* corpus-driven approach) – в отличие от ситуации, когда корпус является всего лишь техническим инструментом, позволяющим получать в целом заранее предсказуемую информацию о жестко заданных лексических единицах (*англ.* corpus-based approach).

В первом случае исследователь получает сведения о языке, которые он заранее не может предположить. Например, наличие тех или иных употребляемых в конкретных текстах синонимов или компонентов семантического ряда для какого-либо базового слова (см. “Verdiener”: “Alleinverdiener”, “Besserverdienende”, “Doppelpverdiener”, “Geldverdienen”, “Geringverdiener”, “Gutverdiener”, “Hauptverdiener”, “Normalverdiener”, “Spitzenverdiener”, “Topverdiener”).

Безусловно, эту информацию могут в какой-то степени предоставить и современные словари, но в нашем случае источником примеров является не «черный ящик», а известные, самостоятельно загруженные в базу данных корпуса тексты.

Исследование открывает широкую перспективу в нескольких направлениях. Во-первых, необходимо совершенствовать автоматические процедуры фиксации семантического сходства и группирования. Во-вторых, целесообразно провести серию экспериментов по выставлению других значений порогового коэффициента. В-третьих, возможно привлечение исходных текстов различных типов дискурса и жанров, например, научных или художественной литературы.

Материалы исследования | Research materials

1. Доброва В. Путин поручил создать национальный план внедрения ИИ // РБК. 2026. 11 апреля. https://www.rbc.ru/technology_and_media/10/04/2026/69d950709a794768374a7d51?ysclid=mv41fgtghx47482475
2. НКРЯ – Национальный корпус русского языка. <https://ruscorpora.ru/>
3. Focus-online. Aktuelle Nachrichten. <https://www.focus.de/>

Источники | References

1. Арабов М. К., Седых В. В. Сравнительный анализ методов моделирования семантических представлений слов в условиях ограниченных языковых ресурсов: случай таджикского языка // Научно-технический вестник Поволжья. 2025. № 6.
2. Бондарчук Г. Г., Грачев Г. В. Функционирование английских наименований питания в современном общественно-политическом дискурсе // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 13 (894).
3. Вершинин В. Е. Разработка рекуррентно-сверточной нейросетевой модели для оценки эмоциональной окраски текста // Известия Института инженерной физики. 2022. № 3 (65).
4. Ермолова К. А. Глагольная (предикативная) сочетаемость существительного «театр» как вершины лексико-семантического поля «театр» // Филологический вестник. 2026. Т. 5. № 4. <https://doi.org/10.66621/2949-4656-2026-5-4-114-121>
5. Каменский М. В., Бредихин С. Н. Алгоритмические процедуры идентификации рекламных текстов в дискурсивном пространстве средств массовой информации // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2025. Т. 24. № 1. <https://doi.org/10.15688/jvolsu2.2025.1.6>
6. Кожахметова А. С., Казкенова А. К. Мультиязычная база данных «Эмоции в метафорическом представлении»: принципы создания и возможности использования // Вопросы лексикографии. 2025. № 38. <https://doi.org/10.17223/22274200/38/4>

7. Колосов А. М., Майсурадзе А. И. Улучшение качества векторных представлений слов за счёт использования нескольких источников представлений // Интеллектуальные системы. Теория и приложения. 2026. Т. 30. № 1.
8. Кузина М. А. Семантическая разметка как инструмент анализа манипулятивных тактик в немецкоязычном идеологическом дискурсе // Когнитивные исследования языка. 2026. № 1 (66).
9. Михеев А. А., Ярыгина Е. С. Концепция моделирования семантической разметки корпусов текстов на базе событийной онтологии // Русский лингвистический бюллетень. 2024. № 1 (49). <https://doi.org/10.18454/RULB.2024.49.37>
10. Нимаева С. С. Темпоральные и модальные значения в семантике слова «вдруг» // Вестник Бурятского государственного университета. Филология. 2026. № 2. <https://doi.org/10.18101/2686-7095-2026-2-26-35>
11. Писарик О. И. Репрезентация сферы «культура и искусство» в корпусе современных средств массовой информации ФРГ // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 11 (892).
12. Рахилина Е. В., Плунгян В. А. О цифровой лексикографии // Труды института русского языка им. В. В. Виноградова. 2025. № 1. <https://doi.org/10.31912/pvrli-2025.1.32>
13. Степанова Д. В. Лексема “Security” («Безопасность») в новостном дискурсе: от частотности к смыслам (корпусное исследование) // Актуальные проблемы филологии и педагогической лингвистики. 2026а. № 2. <https://doi.org/10.29025/2079-6021-2026-2-218-229>
14. Степанова Д. В. Особенности употребления лексемы «терпимость» в СМИ ФРГ и США (корпусное исследование) // Изоморфные и алломорфные признаки языковых систем: сборник научных статей. Ставрополь: Параграф, 2026b. Вып. 8.
15. Степанова Д. В. Презентация существительных семантического ряда «традиционные энергоносители» в современных СМИ ФРГ (на материале текстов газеты «Франкфуртер Альгемайне Цайтунг») // Филологические науки в XXI веке: актуальность, многополярность, перспективы развития: сборник научных трудов. Краснодар: Кубанский государственный университет, 2024а.
16. Степанова Д. В. Разработка метода генерации лингвистического корпуса инструментами обработки естественного языка ход и результаты международного проекта // Язык-Семиотика-Культура: сборник научных статей по итогам международной научной конференции: в 2 ч. (г. Минск, 20-21 ноября 2023 г.). Мн.: Минский государственный лингвистический университет, 2024b.
17. Чжао М. Динамика семантического поля «чистота» в условиях цифровизации: диахронический анализ на данных НКРЯ // Современный ученый. 2025. № 11.
18. Шуйская Ю. В. Семантический сдвиг в значении слова «благотворительный» и его производного: институциональный фактор // Litera. 2026. № 4. <https://doi.org/10.25136/2409-8698.2026.4.79124>
19. Seisebayev A. Zh., Suleimenova E. D. Corpus-Based Study of English Oxymorons: Typology, Semantics, and Contextual Verification // Journal of Siberian Federal University. Humanities and Social Sciences. 2026. Vol. 19. No. 5.

Информация об авторах | Author information



Горожанов Алексей Иванович¹, д. филол. н., доц.

¹ Московский государственный лингвистический университет



Alexey Ivanovich Gorozhanov¹, Dr

¹ Moscow State Linguistic University

¹ a.gorozhanov@linguanet.ru

Информация о статье | About this article

Дата поступления рукописи (received): 08.09.2026; опубликовано online (published online): 20.10.2026.

Ключевые слова (keywords): корпусная лингвистика; сбалансированный лингвистический корпус; семантические признаки; векторное представление слов; новостной текст; corpus linguistics; balanced linguistic corpus; semantic features; word embedding; news text.