DOI: 10.15514/ISPRAS-2026-38(2)-15

Применение тезауруса RuThes и векторных представлений Word2vec в задаче лексической типологии

И.К. Полозов, ORCID: 0000-0003-2679-5465 <ilya-polozov@mail.ru>
И.А. Волкова, ORCID: 0009-0007-3211-6517 <irina.a.volkova@gmail.com>
Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

Аннотация. В статье описывается использование тезауруса RuThes и векторных представлений Word2vec для задачи определения лексической типологии языков. Актуальность работы обусловлена необходимостью проводить типологические исследования языков и слабо развитыми средствами автоматизации данного направления. Сделан обзор существующих методов определения лексической типологии, описаны достоинства и недостатки каждого метода, предложен подход автоматизированного выделения типологий. Также описываются различные виды отношений RuThes. Дано описание используемых корпусов текстов. В качестве примера исследуются семантические зоны “тянуть-толкать” и “чинить-портить”. Получены фреймы для данных семантических зон. Проанализированы извлеченные слова, реализующие семантические зоны, и произведено сравнение с ручным методом. Сравниваются три способа выделения лексической типологии: только с помощью тезауруса, с помощью тезауруса и фильтрации по Word2vec и с помощью тезауруса и добавления ближайших слов по Word2vec. Произведена оценка и сравнение с существующими методами. Для каждого способа посчитана полнота, точность и F-мера. Выявлено, что наилучшие результаты для семантической зоны “тянуть-толкать” дает комбинация использования тезауруса и фильтрации по Word2vec. Добавление же дополнительных ближайших слов по Word2vec ухудшает все метрики кроме F-меры для семантической зоны “толкать”. При этом использование только тезауруса уже дает хорошие результаты, которые могут помочь исследователям языков. Для семантической зоны “чинить-портить” самые лучшие результаты показывает подход с тезаурусом, фильтрацией и добавлением ближайших по Word2vec. Предложено объяснение полученных результатов. Программная реализация выполнена с помощью языка программирования Python3, библиотек Gensim для получения векторов Word2vec, Scikit-learn для сравнения векторов, Numpy для работы с массивами, Pymorphy2 для приведения в начальную форму, NLTK для фильтрации стоп-слов и xml.etree для работы с тезаурусом. Практическая значимость заключается в разработке автоматизированного метода помощи лингвистам и оценке его работы.

Ключевые слова: лексическая типология; тезаурус RuThes; модель Word2vec; классификация текстов; компьютерная лингвистика.

Для цитирования: Полозов И.К., Волкова И.А. Применение тезауруса RuThes и векторных представлений Word2vec в задаче лексической типологии. Труды ИСП РАН, том 38, вып. 2, 2026 г., стр. 227–240. DOI: 10.15514/ISPRAS–2026–38(2)–15.

Application of the RuThes Thesaurus and Word2vec Vector Representations in the Lexical Typology Problem

I.K. Polozov, ORCID: 0000-0003-2679-5465 <ilya-polozov@mail.ru>
I.A. Volkova, ORCID: 0009-0007-3211-6517 <irina.a.volkova@gmail.com >

Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. The article describes the use of the RuThes thesaurus and Word2vec embeddings for determining the lexical typology of languages. The relevance of this work stems from the need to conduct typological studies of languages and the underdeveloped automation tools in this area. The article provides an overview of existing methods for determining lexical typology, describing the advantages and disadvantages of each method, and proposing an approach for automated typology extraction. Various types of RuThes relations are also described. The text corpora used are described. The semantic zones "pull-push" and "fix-spoil" are researched. Frames for these semantic zones are obtained. The extracted words implementing the semantic zones are analyzed and compared with a manual method. Three methods for extracting lexical typology are compared: using thesaurus only, using thesaurus and filtering by Word2vec, and using thesaurus and adding the closest words by Word2vec. An evaluation and comparison with existing methods are performed. For each method, recall, precision, and F-score were calculated. It was found that the best results for the "push-pull" semantic zone are achieved by combining the thesaurus and Word2vec filtering. Adding additional Word2vec closest words degrades all metrics except the F-score for the "push" semantic zone. Using the thesaurus alone, however, yields good results that could be helpful to language researchers. For the "fix-spoil" semantic zone, the best results are achieved by using the thesaurus, filtering, and adding Word2vec closest words. An explanation for the obtained results is offered. The software implementation was implemented using Python3, the Gensim library for generating Word2vec embeddings, Scikit-learn for vector comparison, Numpy for array manipulation, Pymorphy2 for priming, NLTK for stopword filtering, and xml.etree for thesaurus manipulation. The practical significance lies in the development of an automated method for assisting linguists and evaluating its performance.

Keywords: lexical typology; RuThes; Word2vec; text classification; computational linguistics.

For citation: Polozov I.K., Volkova I.A. Application of the RuThes thesaurus and Word2vec vector representations in the lexical typology problem. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 2, 2026, pp. 227-240 (in Russian). DOI: 10.15514/ISPRAS-2026-38(2)-15.

1. Введение

Задача определения лексической типологии языков является актуальной и малоизученной темой. Большинство подходов ее определения являются ручными и требующими ресурсов ученых-лингвистов и носителей языка. В работе предлагается способ автоматизированной помощи исследователю. Изучаются семантические поля глаголов “тянуть-толкать” и “чинить-портить”, так как данные семантические поля уже исследованы в работах [1] и [2], и есть возможность сравнить автоматизированный метод с ручным.

Лексическая типология изучает, какими лексическими средствами описываются конкретные явления. Также производит сравнение этих лексических средств в разных языках. Например, палец руки и ноги в русском языке обозначается одним словом, а в английском языке палец на руке обозначается лексемой “finger”, а на ноге – “toe”. Также и в рамках одного языка может изучаться определенное семантическое поле и лексические средства, с помощью которых оно выражается. Например, для семантической зоны “тянуть-толкать” могут быть найдены варианты употребления, такие как “перемещать объект вперед”, “нажимать на кнопку”, “подталкивать к решению”, “привлекать внимание”. Найденные описания употреблений являются фреймами [3]. Затем может составляться таблица, в описании строк которой находятся фреймы, в описании столбцов - языки, а в ячейках – лексические единицы, которыми найденные фреймы описываются в разных языках. Данная статья помогает составлять такие фреймы.

2. Обзор литературы

2.1 Классические методы

Сегодня существует четыре основных метода исследования. Первый подход основан на использовании фреймов [3], также он называется подходом Московской лексико-типологической школы. В нем каждая ситуация, принадлежащая семантическому полю, обозначается фреймом - набором описательных слов. Например, фрейм “нажимать предмет вперед” семантического поля “тянуть-толкать”. Фрейм в языке представлен определенными лексемами. Таких фреймов может быть много. Исследователь отбирает их на основе словарей, переводчиков, собственных представлений. Могут анализироваться синхронные переводы текстов. Затем исследователь создает таблицу. В строки заносятся описания фреймов, в столбцах размещаются лексемы, а в ячейках признак того, описывает ли данная лексема фрейм. Вторым вариантом является таблица, в строках которой записывается фрейм, а в столбцах разные языки. В ячейках таблицы указываются лексемы, с помощью которых в конкретном языке описывается конкретный фрейм. Недостатком является ручной метод составления фреймов. Также в случае использования словарей исследователю необходимо выбрать границу, после которой следует прекратить поиск фреймов, так как при переводе будут образовываться новые фреймы, достаточно отдаленно связанные с исследуемым семантическим полем, что связано с существованием многозначных слов. Этот подход также приводит к необходимому привлечению носителей языков.

Второй подход основан на физических чувствах людей [4]. Исследователь собирает набор универсальных стимулов, например, что-то с определенным вкусом, запахом, цветом, формой и предоставляет это носителю языка. Носитель языка своими словами должен как можно более точно описать выбранный предмет. Так, предоставляя элементы разным носителям языка, можно определить, какими лексическими средствами они описываются в разных языках. Недостатком такого подхода является невозможность физически отобразить все многообразие лексических единиц, а также сложность и большие временные затраты. Здесь также необходимы носителя языка.

Третий подход использует универсальные примитивы, из которых можно описать любую ситуацию [5]. В таком подходе используется система из 64 базовых понятий, из которых выводятся все остальные. Недостатком является неоднозначность вывода и сложность подхода.

Четвертый подход основан на использовании параллельных корпусов [6]. Исследователь находит переводные эквиваленты для реализаций семантической зоны. Недостатком является отсутствие параллельных корпусов для редких языков.

Сегодня чаще всего исследователи используют фреймовый подход. Например, в работе [7] исследуется семантическое поле “мешать“ с помощью фреймового подхода. В работе [8] исследуются семантическое поле “домашний скот” для германских и славянских языков. В качестве материала используется лексический фонд. Автор работы [9] исследует семантическое поле “шахматная игра” для русского языка с помощью подхода “центр-периферия”. В центре находятся основные узкоспециализированные семантические признаки, а на периферии менее специализированные. В работе [10] исследуется семантическая зона “острый” в китайском языке. Используются словари, корпуса текстов и носители языка. В главе “Methodology at work Semantic fields sharp and blunt” книги [11] описывается семантическая зона “острый-тупой”. Обнаружено, что основные виды оппозиции – это тип острого объекта и чувство, по которому определяется степень остроты объекта. В работе [12] изучается семантическая зона слова “город”. Исследование проводится с помощью литературных источников.

2.2 Автоматизированные методы

Автоматизация типологических исследований развита недостаточно. В работе [13] используются биграммы НКРЯ, дополненные различными леммами слов. Для кластеризации используются вектора, полученные следующим образом: выбираются 10 000 наиболее частотных лексем, для выбранного слова считается, сколько раз встретилась каждая из 10 000 частотных лексем на расстоянии 5-и слов. Используются алгоритмы иерархической кластеризации, так как алгоритмы с неизвестным количеством кластеров показали плохие результаты. Недостатком подхода является ручное задание количества кластеров, которое надо подбирать, при этом оно будет разное для разных семантических зон. В работе [14] используются готовые анкеты, которые затем автоматически переводятся на другие языки с помощью словарей и параллельных корпусов. Исследование проводится для семантической зоны “острый-гладкий” и “толстый-тонкий”. Недостатком является необходимость создания готовых анкет.

3. Тезаурус RuThes

RuThes [15] является тезаурусом русского языка. Он включает 31 тысячу понятий и 111 тысяч отношений. Состоит из 4 файлов:

1) concepts – понятия;
2) relations – отношения между понятиями;
3) text_entry – текстовые входы;
4) synonyms – отношения между текстовыми входами и понятиями;

В тезаурусе представлены 3 вида отношений: “выше-ниже”, “часть-целое”, “асц1-асц2”. Понятия “ниже” являются экземпляром понятия “выше”. Например, “сфера связи” ниже “отрасли, оказывающей услуги”. Связь “часть-целое” означает, что одно понятие является частью другого и не может существовать вне него. Например, “телекоммуникационная связь” является частью “телекоммуникационной компании”. Однако “дерево” не является частью “леса”, так как может расти вне него. Понятия, связанные отношением “асц1-асц2”, являются зависимыми. Например, “сфера связи” зависит от понятия “Министерство транспорта и связи”.

4. Эксперименты без Word2Vec

В работе исследуются семантические зоны “тянуть-толкать” и “чинить-портить” на примере русского языка (также подход может быть применен и к другим языкам и семантическим зонам). Сначала была исследована возможность работы без поиска ближайших слов по модели Word2vec [16]. Для семантического поля были найдены все связанные концепты из тезауруса. Концепты искались на первом уровне со всеми отношениями вида: “выше-ниже”, “часть-целое”, “асц1-асц2”.

4.1 Алгоритм работы программы

Шаги алгоритма:

1) Для каждого из слов “тянуть”, “толкать”, “чинить” и “портить” находится соответствующий концепт в тезаурусе. Все концепты в тезаурусе переводятся в начальную форму для максимального совпадения. Концепты могут состоять из нескольких слов, например, “острый на вкус”, поэтому сравнение происходит с главным словом концепта, которое указано в файле text_entry тезауруса RuThes.
2) В RuThes находятся все концепты, связанные отношением с исследуемыми словами;

3) Составляется список объектов, в каждом из которых хранятся лексемы “тянуть”, “толкать”, “чинить”, “портить” и все связанные отношением с ними слова на первом уровне из тезауруса RuThes;
4) Удаляются одинаковые слова;
5) Остаются группы слов, в которых первое слово означает найденный концепт для исследуемой лексемы, а последующие слова означают связанные слова из RuThes;
6) Найденные группы слов считаются фреймами исследуемой семантической зоны;

Использованные ресурсы:

1) Библиотека Xml.etree (https://docs.python.org/3/library/xml.etree.elementtree.html;
2) Язык реализации – Python 3;

4.2 Результаты

Для семантической зоны “тянуть” найдено 42 фрейма. Ниже приведены несколько примеров найденных фреймов:

1) “притянуть (придвинуть), тянуть (тащить направляя куда-либо)”;
2) “дергать (тянуть, тащить резким движением), одернуть вниз, вздернуть вверх”;
3) “вытянуть в длину”;
4) “тянуться, потягиваться”;
5) “тянуться, чтобы достать”;

Аналогично рассматривается связанная с этим полем лексема “толкать”. Найдено 6 фреймов. Примеры фреймов:

1) “толкать от себя”;
2) “толкаться, толкать друг друга, давка в толпе”;
3) “вытолкнуть, выпихнуть, толкать от себя”;
4) “растолкать спящего”;
5) “протолкаться сквозь толпу, растолкать в разные стороны”;

Для сравнения была взята работа [1], в которой ручным способом найдены следующие фреймы для русского языка:

1. Для группы “толкать-нажимать”:

   Найдены 4 семантические лексемы: “пихать”, “давить”, “тыкать”, “нажимать”, “толкать”. Они реализуются в следующих ситуациях:

   1.1) Неодушевленные объекты:

   1.1.1) “перемещать ногами”:

   1.1.2) “нажимать на кнопку”;

   1.1.3) “открывать от себя”;

   1.1.4) “перемещать перед собой”;

   1.1.5) “помещать внутрь”;

   1.2) Одушевленные объекты:

   1.2.1) “будить”:

   1.2.2) “привлекать внимание”;

   1.2.3) “перемещать человека от себя”;

   1.2.4) “сталкивать с высоких объектов”;

1.2.5) “проявлять агрессию”;

1.2.6) “толкать из-за тесноты”;

2. Для группы “тянуть-выдергивать”:

   Найдены 4 семантические лексемы: “волочить”, “дергать”, “тащить”, “тянуть”, “выдергивать”. Они реализуются в следующих ситуациях:

   2.1) Неодушевленные объекты:

   2.1.1) “открывать на себя”;

   2.1.2) “извлекать”;

   2.1.3) “двигать на себя”;

   2.1.4) “двигать за собой”;

   2.1.5) “выдергивать объект”;

   2.1.6) “тянуть шею, чтобы увидеть”;

   2.1.7) “достать/дотянуться”;

   2.1.8) “увеличивать в размере”;

   2.2) Одушевленные объекты:

   2.2.1) “перемещать на себя”;

   2.2.1) “привлекать внимание”;

Аналогичным образом была проведена работа для семантической зоны “чинить-портить” и сделана оценка с работой [2]. Для зоны “чинить” было найдено 3 фрейма, для зоны “портить” 17. Это такие фреймы, как “чинить”, “застояться (испортиться)”, “износиться от употребления”, “ржаветь”, “коверкать”, “застояться (испортиться)”, “изъесть (испортить)”, “портить отношения”, “портить репутацию”.

4.3 Оценка подхода на основе тезауруса RuThes

При подсчете точности, полноты и F-меры для семантической зоны “тянуть-толкать” за золотой стандарт взята работа [1], в которой были найдены 22 семантические зоны для русского языка.

Табл. 1. Оценка работы подхода на основе тезауруса RuThes для семантической зоны “тянуть-толкать”.
Table 1. Evaluation of the performance of the RuThes thesaurus-based approach for the “push-pull” semantic zone.

	Тянуть	Толкать	Тянуть-толкать
Точность	12%	83.3%	21%
Полнота	50%	46%	48%
F-мера	19.2%	59%	29%

В табл. 1 представлена оценка работы подхода на основе тезауруса RuThes. Найдены не все ситуации, которые есть в работе [1], но при этом найдены и ситуации, которых нет в работе [1], хотя они связаны с семантической зоной “тянуть-толкать”, например, “вытянуть в длину”, “перетянуть на себя”, “длиться по времени”. При этом полнота больше точности, и есть похожие найденные ситуации. Точность можно повысить фильтрацией лишних найденных ситуаций. Для этого необходимо найти векторы найденных слов и убрать векторы, которые слишком похожи по косинусной мере. Для увеличения полноты нужно

добавить похожие по Word2vec слова к словам “тянуть-толкать” и повторить алгоритм уже для совокупности найденных слов.

Для семантической зоны “чинить-портить” взята работа [2]. В табл. 2 представлена работа алгоритма на ней.

Табл. 2. Оценка работы подхода на основе тезауруса RuThes для семантической зоны “чинить-портить”.
Table 2. Evaluation of the RuThes-based approach for the semantic zone “fix-spoil”.

	Чинить	Портить	Чинить-портить
Точность	67%	77%	75%
Полнота	20%	88%	50%
F-мера	31%	82%	64%

5. Использование Word2vec для фильтрации найденных слов

Векторы Word2vec [18] обладают таким свойством, что ближайшие по косинусной мере вектора также близки и по смыслу. Эту особенность можно использовать для фильтрации найденных в предыдущих главах 42-х словосочетаний для зоны “тянуть-толкать”. Для модели Word2vec используется корпус НКРЯ (https://ruscorpora.ru/) с почти 250 миллионами словоупотреблений, объем словаря 195 071, размер вектора 300, модель обучения Continuous Skipgram. Корпус сбалансированный. Из словосочетаний берется первое. Для представления словосочетаний векторов находится средний вектор из входящих слов в словосочетание. Убираются стоп-слова с помощью библиотеки NLTK (https://www.nltk.org/), и все слова преобразуются в начальную форму. Также удаляются все знаки препинания. Сравнивается каждое слово с каждым по косинусной мере, затем удаляются слишком близкие друг к другу слова. Не все слова есть в словаре модели Word2vec, поэтому такие слова не обрабатываются.

На рис. 1 изображено попарное сходство слов. Эмпирически был выбран порог схожести векторов по косинусной мере 0.4. При этом результаты не меняются при пороге от 0.3 до 0.4. При пороге 0.29 количество найденных ситуаций для зоны “тянуть” падает с 6 до 5. При пороге 0.41 количество найденных ситуаций возрастает с 6 до 8, при этом совпадающих с работой [1] не становится больше. Косинусная мера рассчитывалась как

$$\text{similarity(A, B)} = \frac{\sum_{i=1}^{n} Ai \times Bi}{\sqrt{\sum_{i=1}^{n} (Ai)\text{^}2} \times \sqrt{\sum_{i=1}^{n} (Bi)\text{^}2}},$$

где A и B векторы признаков, Ai и Bi измерения признаков, n - размерность векторов признаков.

Для слова “тянуть” получено 6 ситуаций, совпадает с работой [1]:

1) “тянуть (тащить направляя куда-либо)”;
2) “растянуться (упасть всем телом)”.

Для слова “толкать” найдено 5 фреймов. Они все совпадают с фреймами в работе [1].

В табл. 3 представлена оценка работы подхода с использованием Word2vec.

Точность для слова "тянуть" увеличилась с 12% до 33.3%, полнота для слова "тянуть" уменьшилась с 50% до 20%, точность для слова толкать увеличилась с 83.3% до 100%, полнота осталась на 46%, F-мера для слова "тянуть" увеличилась с 19.2% до 25%, F-мера для слова "толкать" увеличилась с 59% до 62.5%, общая точность увеличилась с 21% до 64%, общая полнота уменьшилась с 48% до 33.3%, общая F-мера увеличилась с 29% до 44%.

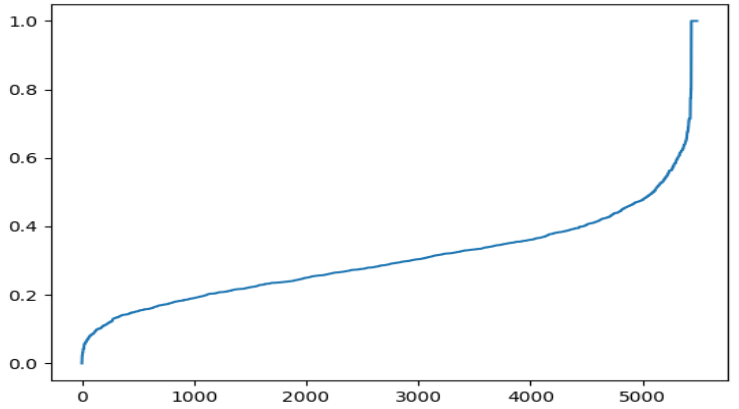

Рис. 1. Попарное сходство найденных слов.
Fig. 1. Pairwise similarity of found words.

Табл. 3. Оценка работы подхода с использованием Word2vec для семантической зоны "тянуть-толкать".
Table 3. Evaluation of the performance of the Word2vec approach for the "push-pull" semantic zone.

	Тянуть	Толкать	Тянуть-толкать
Точность	33.3%	100%	64%
Полнота	20%	46%	33.3%
F-мера	25%	62.5%	44%

Таким образом, увеличилась точность и F-мера, уменьшилась полнота. Значит соответствий с золотым стандартом стало больше. Алгоритм в целом стал работать лучше.

Оценки для семантической зоны "чинить-портить" приведены в таблице 4. Для зоны "чинить" было найдено 3 фрейма, для "портить" тоже 3 фрейма. Для данной зоны фильтрация напротив ухудшила результаты.

Табл. 4. Оценка работы подхода с использованием Word2vec для семантической зоны "чинить-портить".
Table 4. Evaluation of the performance of the Word2vec approach for the "fix-spoil" semantic zone.

	Чинить	Портить	Чинить-портить
Точность	67%	67%	67%
Полнота	18%	37.5%	26%
F-мера	29%	48%	38%

6. Использование Word2vec для увеличения количества найденных слов

Для слов "тянуть-толкать" были найдены 9 синонимов по Word2vec. Берутся 9 наиболее похожих слов. Переводятся все слова в начальную форму и удаляются повторяющиеся.

Для слова "толкать" был получен следующий расширенный список слов:

"подталкивать", "толкать", "подтолкнуть", "пихать", "отпихивать", "подталкивать", "расталкивать", "толкнуть".

Для слова "тянуть" получен следующий расширенный список слов:

"тянуться", "увлекаться", "тащиться", "рваться", "тащить", "потянуть", "волочить", "тянуть".

Дальше алгоритм работает аналогично уже описанному, только коэффициент фильтрации по похожести был уменьшен до 0.3. Порог похожести может быть выбран любой в диапазоне от 0.24 до 0.32. При пороге похожести 0.23 количество найденных фреймов падает с 15 до 14, и количество совпадающих с работой [1] фреймов уменьшается на 1. При 0.33 количество найденных увеличивается с 15 до 18, при этом не появляется новых, совпадающих с работой.

Для слова “тянуть” получено 15 фреймов, 4 совпадают с работой [1]:

1) “тянуть (тащить направляя куда-либо)”;
2) “растянуться (упасть всем телом), утащить (унести), волочить, тащить волоком”;
3) “прорваться (продырявиться), разорваться на части, разорваться (взорваться изнутри), лопнуть (треснуть), надорваться (разорваться), расползтись от ветхости, рваться по швам”;
4) “дергать (тянуть, тащить резким движением), тащить тяжелое”.

Для слова “толкать” получено 7 фреймов, 4 из которых совпадают с работой [1].

В табл. 5 представлена оценка работы подхода с добавлением ближайших Word2vec.

Табл. 5. Оценка работы подхода с добавлением ближайших по Word2vec для семантической зоны “тянуть-толкать”.
Table 5. Evaluation of the approach with adding the closest Word2vec terms for the “pull-push” semantic zone.

	Тянуть	Толкать	Тянуть-толкать
Точность	27%	57.2%	36.36%
Полнота	40%	36.36%	38%
F-мера	32%	44.4%	32.2%

Точность упала и для слова “тянуть”, и для слова “толкать”, но полнота для слова “тянуть” увеличилась, а для слова “толкать” уменьшилась, общая полнота увеличилась. F-мера для слова “тянуть” увеличилась, F-мера для слова “толкать” уменьшилась. Общая F-мера уменьшилась. Добавление ближайших слов увеличило полноту для слова “толкать”, но для слова “тянуть” точность и полнота уменьшилась. Это объясняется тем, что ближайшие слова по Word2vec часто встречаются в похожем контексте, но близость по Word2vec не всегда означает, что слова являются разным выражением семантического поля.

Аналогично для семантической зоны “чинить-портить” оценка работы в табл. 6. Для зоны “чинить” было найдено 11 фреймов, для “портить” тоже 9 фреймов. Удалось значительно увеличить полноту для зоны чинить с 20% в методе только с тезаурусом до 60%. Полнота для зоны “портить” осталось на значении 88%, при этом упала точность для зоны “чинить”.

Табл. 6. Оценка работы подхода с добавлением ближайших по Word2vec для семантической зоны “чинить-портить”.
Table 6. Evaluation of the approach with adding the closest Word2vec terms for the “fix-spoil” semantic zone.

	Чинить	Портить	Чинить-портить
Точность	27%	78%	50%
Полнота	60%	88%	72%
F-мера	37%	83%	59%

7. Доверительные интервалы подходов

Для оценки доверительных интервалов использовалась формула

$$\rho \pm Z_\alpha \sqrt{\frac{\rho \times (1-\rho)}{n}},$$

где p – доля признака, n – размер выборки, Zα – константа для определения выбранной вероятности.

Подсчеты сделаны для общей точности, полноты и F-меры с константной Zα равной 1.96 для 95% интервала для всех трех подходов: только на основе тезауруса RuThes, с тезаурусом и фильтрацией Word2vec, с тезаурусом, фильтрацией и поиском ближайших слов по Word2vec. Результаты расчетов приведены в табл. 7.

Табл. 7. Доверительные интервалы.
Table 7. Confidence intervals.

	RuThes	RuThes + фильтр. Word2vec	RuThes + фильтр. Word2vec + ближ. Word2vec
Точность	[26%, 49%]	[42%, 88%]	[27%, 56%]
Полнота	[47%, 51%]	[16%, 44%]	[38%, 70%]
F-мера	[27%, 58%]	[26%, 56%]	[31%, 63%]

Наибольший разброс получен для подхода RuThes + фильтрация по Word2vec. Это связано с тем, что подход помогает значительно улучшить результаты в случае, когда из RuThes получено много лишних зон, однако слабо влияет на результат, если RuThes дал изначально мало зон. При этом полнота для подхода только с тезаурусом RuThes остается самой стабильной.

8. Сравнение результатов

В табл. 8 и 9 представлены оценки всех трех способов для семантической зоны “тянуть-толкать” и “чинить-портить” соответственно. Жирным выделен самый лучший результат независимо от слов, серым фоном выделен самый лучший результат для каждой из зон “тянуть”, “толкать”, объединенной зоны “тянуть-толкать", а также зон “чинить”, “портить” и объединенной зоны “чинить-портить”. Самые лучшие результаты для зоны “тянуть-толкать” показал подход с использованием тезауруса RuThes и последующей фильтрацией Word2vec. При этом по полноте лучшие результаты показал подход с использованием одного тезауруса RuThes, аналогичные результаты получились только для зоны “толкать” в подходе комбинации RuThes с фильтрацией Word2vec. Самые плохие результаты для зоны “тянуть-толкать” показал подход комбинации RuThes с добавлением ближайших слов по Word2vec и фильтрацией Word2vec. Он показал лучшие результаты только в F-мере для зоны “тянуть”. Для зоны “чинить-портить” самые лучшие результаты показал подход с добавлением ближайших слов по Word2vec и фильтрацией Word2vec. Особенно увеличение видно по полноте подзоны “чинить”. Это связано с тем, что она плохо представлена в тезаурусе RuThes, поэтому добавление ближайших по Word2vec помогает улучшить результаты. Однако упала точность для зоны “чинить”. Подход только с тезаурусом RuThes показал лучшие результаты по точности для зоны “чинить”, объединенной “чинить-портить”, и по F-мере для объединенной зоны “чинить-портить”. Подход с использованием тезауруса и фильтрацией Word2vec показал лучшие результаты по точности для зоны “тянуть”. Таким образом, использование тезауруса помогает повысить полноту, т.к. дает много новых вариантов употреблений. Но также дает и лишние словоупотребления. Их помогает убрать последующая фильтрация Word2vec, которая показывает лучшие на данном наборе экспериментов результаты для семантической зоны “тянуть-толкать”. Добавление же

похожих по Word2vec приводит только к ухудшению результатов в случае, когда зона по полноте уже хорошо представлена в тезаурусе. Такое может быть связано с тем, что векторы Word2vec не учитывают контекст слова, при этом слово может иметь разные смыслы в разных контекстах. Поэтому в дальнейшем необходимо исследовать применение векторов ELMO (Embeddings from Language Models) или BERT (Bidirectional Encoder Representations from Transformers), обладающие тем свойством, что для одного и того же слова в разных контекстах будут разные векторы. Однако если зона представлена в тезаурусе плохо, то подход с добавлением ближайших позволяет наоборот улучшить результаты, а фильтрация в этом случае работает хуже, так как фильтрует слабо представленные в тезаурусе зоны. Аналогично в табл. 9 представлена оценка работы для семантической зоны “чинить-портить”.

Табл. 8. Сравнение работы подходов для семантической зоны “тянуть-толкать”.
Table 8. Comparison of the approaches for the “pull-push” semantic zone.

	T1-R	*T2-R*	*O-R*	*T1-R-W*	*T2-R-W*	*O-R-W*	*T1-R-2W*	*T2-R-2W*	*O-R-2W*
P	12%	83.3%	21%	33.3%	**100%**	64%	27%	57.2%	36.36%
R	**50%**	46%	48%	20%	46%	33.3%	40%	36.36%	38%
F	19.2%	59%	29%	25%	**62.5%**	44%	32%	44.4%	32.2%

T1-R – тянуть RuThes, T2-R – толкать RuThes, O-R – общая RuThes, T1-R-W - тянуть RuThes +фильтрация Word2vec, T2-R-W - толкать RuThes +фильтрация Word2vec, O-R-W - общая RuThes +фильтрация Word2vec, T1-R-2W - тянуть RuThes +фильтрация Word2vec + добавление похожих Word2vec, T2-R-2W - толкать RuThes + фильтрация Word2vec + добавление похожих Word2vec, O-R-2W – общая RuThes +фильтрация Word2vec + добавление похожих Word2vec.

Табл. 9. Сравнение работы подходов для семантической зоны “чинить-портить”.
Table 9. Comparison of the approaches for the “fix-spoil” semantic zone.

	Ч-R	*П-R*	*O-R*	*Ч-R-W*	*П-R-W*	*O-R-W*	*Ч-R-2W*	*П-R-2W*	*O-R-2W*
P	67%	77%	75%	67%	67%	67%	27%	**78%**	50%
R	20%	**88%**	50%	18%	37.5%	26%	60%	**88%**	72%
F	31%	82%	64%	29%	48%	38%	37%	**83%**	59%

Ч-R – чинить RuThes, П-R – портить RuThes, O-R – общая RuThes, Ч-R-W -чинить RuThes +фильтрация Word2vec, П-R-W - портить RuThes +фильтрация Word2vec, O-R-W общая RuThes +фильтрация Word2vec, Ч-R-2W - чинить RuThes +фильтрация Word2vec + добавление похожих Word2vec, П-R-2W - портить RuThes +фильтрация Word2vec + добавление похожих Word2vec, O-R-2W – общая RuThes +фильтрация Word2vec + добавление похожих Word2vec.

8.1 Анализ ошибок

Не был найден фрейм “сталкивать с высоких объектов”. Это связано с тем, что такой концепт напрямую не представлен в тезаурусе. При этом добавление ближайших по Word2vec тоже не добавляет слов, из которых можно вывести такой концепт в тезаурусе. Например, ближайшие по Word2vec для слова “тянуть” будут “тянуть”, “тащиться”, “тащить”, “волочить”, “рваться”, “тянуться”, “потянуть”. Они не приводят к нужному концепту в тезаурусе для фрейма “сталкивать с высоких объектов”. Также не были найдены такие фреймы, как “тянуть шею, чтобы увидеть”, “штопать”, “вредить”, “разладить”, “омрачать”, “испохабить”. Они тоже не представлены в тезаурусе. При этом были найдены фреймы, которых нет в ручных работах. Фреймы, такие как “утащить (увезти вопреки желанию)”, “вобраться (втянуться внутрь)”, “втянуться вовлечься”, “вытянуть в длину”, “перетянуть на себя”, “длиться по времени” “волочить, тащить волоком”, “тащить тяжелое”, “просидеть (испортить)”, “вымерзнуть, портиться от холода”, “мокнуть, портиться от влаги”.

9. Заключение

Проведенные эксперименты показали работоспособность подхода для автоматизированного выделения лексической типологии и помощи в работе лингвистов. Предложены 3 варианта поиска фреймов семантической зоны. Лучшие результаты показал подход с использованием тезауруса RuThes и фильтрацией Word2vec для хорошо представленных в тезаурусе зон и подход с добавлением ближайших по Word2vec для плохо представленных в тезаурусе зон. Преимуществом предложенного подхода является отсутствие необходимости привлекать носителей языка, вручную изучать словари и переводы, меньшая сложность и большая быстрота. Недостатком является меньшая точность по сравнению с ручными подходами, но она может быть увеличена последующей ручной обработкой, которая будет проще. Таким образом, предложенный в работе подход не позволяет заменить работу профессиональных лингвистов, но может быть использован ими как инструмент для повышения скорости и эффективности своей работы.

Список литературы / References

[1]. Савельева А.Ю. Глаголы семантических зон 'ТЯНУТЬ' и 'ТОЛКАТЬ' в типологической перспективе. Проблемы компьютерной лингвистики и типологии: сб. Всерос. конф. № 6, 2017, стр. 142-152.

[2]. Короткая М.Б. Семантическая структура лексического поля ЧИНИТЬ-ПОРТИТЬ, диссертация на соискание степени бакалавриата, факультет Фундаментальной и компьютерной лингвистики НИУ ВШЭ, 2020, 73 с.

[3]. Рахилина Е. В., Резникова Т. И.. Фреймовый подход к лексической типологии. Вопросы языкознания, № 2, 2013, стр. 3-31.

[4]. Berlin B. Kay P. Basic Color Terms: Their Universality and Evolution. University of California Press, 1969. 178 p.

[5]. Wierzbicka A., Goddard C., Semantic amd lexical universals: Theory and emperical findings, Amsterdam. Lingvisticae Investigationes, vol. 21, no. 1, 1994. pp. 249-261. DOI: 10.1075/LI.21.1.11CHA.

[6]. Viberg, A., Seeing the lexical profile of Swedish through multilingual corpora. The case of Swedish aka and other vehicle verbs, Advances in corpus-based contrastive linguistics, Amsterdam, vol b, 2013, pp. 25–56. DOI: /10.1075/scl.54.04vib.

[7]. Дунаева К.О., Маринина В.В. Семантическое поле 'мешать' в типологической перспективе. XXVI Открытая конференция студентов-филологов в СПбГУ, 2023 стр. 34-37.

[8]. Шешкина Т.Ф. Германо-славянские параллели семантического поля «Домашний скот» в немецких лексикографических источниках. Филологические науки. Вопросы теории и практики. т. 13, № 6, 2020, стр. 303-307. DOI: 10.30853/filnauki.2020.6.57.

[9]. Влавацкая М.В., Журавлева И.Н. Лексико-семантическое поле «Шахматная игра» в современном русском языке. Мир науки, культуры, образования, №2(93), 2022, стр. 293-297.

[10]. Холкина Л.С., Наний Л.О., Сы Ц. Семантическое поле ОСТРЫЙ в китайском языке: диахроническое развитие и его отражение в современных диалектах. Journal of Language Relationship, № 20(3-4), 2023, стр. 280-298. DOI: 10.31826/jlr-2023-203-410.

[11]. Rakhilina E, Reznikova T., Kyuseva M., Parina E., Ryzhova D., Panina A., Kruglyakova V., Kozlov A., Vinogradova O. I., Vyrenkova A. S., Orekhov B. The Typology of Physical Qualities, Amsterdam: John Benjamins Publishing Company, part 2, 2022, pp. 29-55.

[12]. Григорьева О.Н., Цзян Н., Лексико-семантическая группа “город” в современных российских масс-медиа. Вестник Московского государственного областного университета. Серия: Русская филология, № 5, 2018, стр. 31-37. DOI: 10.18384/2310-7278-2018-5-31-38.

[13]. Рыжова Д.А. Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики. Вестник РГГУ. Сер.: История. Филология. Культурология. Востоковедение, том 18, 2016, стр. 140-150.

[14]. Kyuseva M., Parina E., Ryzhova D. Automatic data collection in lexical typology. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. № 2, 2018, стр. 29-55.

[15]. Loukachevitch N., Dobrov B., RuThes Thesaurus for Natural Language Processing. The Palgrave Handbook of Digital Russia Studies, 2021, pp. 319-334. DOI: 10.1007/978-3-030-42855-6_18.
[16]. Mikolov T., Sutskever I., Chen K., Corrado G.S., Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems, 2013, pp. 1-9.

Информация об авторах / Information about authors

Илья Константинович ПОЛОЗОВ – аспирант кафедры Алгоритмических языков факультета Вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова. Научные интересы: компьютерная лингвистика, семантический анализ, вопросно-ответные системы, машинное обучение.

Ilia Konstantinovich POLOZOV – postgraduate student at the Department of Algorithmic Languages, the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Research interests: computational linguistics, semantic analysis, question-answering systems, machine learning.

Ирина Анатольевна ВОЛКОВА – кандидат физико-математических наук, доцент кафедры Алгоритмических языков факультета Вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова. Научные интересы: компьютерная лингвистика, синтаксические анализаторы, лингвистические процессоры, интерфейсы взаимодействия с ЭВМ.

Irina Anatolyevna VOLKOVA – Cand. Sci. (Phys.-Math.), associate Professor at the Department of Algorithmic Languages, the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Research interests: computational linguistics, parsers, linguistic processors, computer interfaces.