

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ/MATHEMATICAL SOFTWARE FOR COMPUTERS, COMPLEXES AND COMPUTER NETWORKS

DOI: https://doi.org/10.60797/IRJ.2026.167.64 EDN: WRHHEM

АВТОМАТИЗИРОВАННОЕ ПОСТРОЕНИЕ ПРЕДСТАВЛЕНИЙ ФРЕЙМОВ В ОБЛАСТИ ЛЕКСИЧЕСКОЙ ТИПОЛОГИИ НА ОСНОВЕ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ И ПОСЛЕДОВАТЕЛЬНЫХ ПЕРЕВОДОВ

Научная статья

Полозов И.К.[1, *], Волкова И.А.[2]
[1] ORCID : 0000-0003-2679-5465;
[1, 2] Московский государственный университет, Москва, Российская Федерация

* Корреспондирующий автор (ilya-polozov[at]mail.ru)

Предложена: 29.03.2026; Принята: 24.04.2026; Опубликована: 18.05.2026

Аннотация

Работа посвящена использованию мультиязычного корпуса НКРЯ, последовательных переводов и векторных представлений BERT в задаче поиска репрезентаций фреймов для конкретной семантической зоны. Проблема заключается в определении, как семантическая зона представлена в исследуемом языке и в каких ситуациях она может проявляться в виде фреймов. Проведен обзор существующих методов, описаны их достоинства и недостатки. Сравниваются подходы, основанные на кластеризации векторов BERT параллельных переводов, фильтрации по косинусной мере, выделению фреймов с помощью тезауруса WordNet, последовательных переводов через промежуточный язык и последующей кластеризации векторов BERT. Самые лучшие результаты показывает метод с последовательными переводами и последующей кластеризацией. Определение сходства работает лучше, чем кластеризация, для параллельных корпусов. Проведена оценка работы и предложены объяснения полученных результатов. Даны рекомендации по подбору параметров работы алгоритмов для семантической зоны «тянуть — толкать».

Ключевые слова: лексическая типология, параллельные корпуса, последовательные переводы, BERT, классификация текстов, компьютерная лингвистика, семантические фреймы.

AUTOMATED CONSTRUCTION OF FRAME REPRESENTATIONS IN THE FIELD OF LEXICAL TYPOLOGY BASED ON PARALLEL CORPORA AND CONSECUTIVE TRANSLATIONS

Research article

Polozov I.K.[1, *], Volkova I.A.[2]
[1] ORCID : 0000-0003-2679-5465;
[1, 2] Lomonosov Moscow State University, Moscow, Russian Federation

* Corresponding author (ilya-polozov[at]mail.ru)

Suggested: 29.03.2026; Accepted: 24.04.2026; Published: 18.05.2026

Abstract

The paper is devoted to the use of the multilingual NRL corpus, consecutive translations and BERT vector representations in the task of identifying frame representations for a specific semantic domain. The problem lies in determining how a semantic domain is represented in the studied language and in which situations it may manifest itself in the form of frames. A review of existing methods is conducted, and their pros and cons are described. Approaches based on the clustering of BERT vectors of parallel translations, filtering by cosine similarity, frame extraction using the WordNet thesaurus, and consecutive translations via an intermediate language followed by BERT vector clustering are compared. The method involving consecutive translations followed by clustering yields the best results. Similarity detection performs better than clustering for parallel corpora. The work is evaluated and explanations for the results are suggested. Recommendations are provided for selecting algorithm parameters for the «pull—push« semantic domain.

Keywords: lexical typology, parallel corpora, consecutive translations, BERT, text classification, computational linguistics, semantic frames.

Введение

На сегодняшний день количество методов, предназначенных для автоматизации поиска фреймов в пределах заданной семантической зоны, остаётся ограниченным. При этом данная задача является важной частью лексической типологии, сохраняет актуальность и в большинстве исследований по-прежнему решается ручными методами. В настоящей работе для её решения предлагается автоматизированный подход, основанный на использовании параллельного корпуса Национального корпуса русского языка [1] и контекстных векторных представлений BERT. В качестве экспериментального материала рассматривается семантическая зона «тянуть — толкать», что позволяет сопоставить результаты автоматизированного подхода с существующим ручным исследованием, поскольку данная зона была подробно изучена в работе [2]. Источник является надежным: результаты исследования [2] легли в основу

выпускной квалификационной работы, успешно защищённой в Национальном исследовательском университете «Высшая школа экономики» на оценку 10/10.

Лексическая типология занимается изучением способов, с помощью которых язык выражает конкретные явления, а также сопоставляет соответствующие лексические средства в различных языках. Например, в русском языке один и тот же термин может обозначать как пальцы руки, так и пальцы ноги, тогда как в английском для этих понятий используются разные слова — *finger* и *toe*. Помимо межъязыкового сравнения, исследованию подлежат семантические поля внутри одного языка и способы их выражения. Так, в семантической области «чинить — портить» можно выделить различные варианты значений: «делать вновь пригодным», «настраивать инструмент», «изменять деятельность», «ухудшать». Такие варианты употребления образуют отдельные фреймы [3].

На основе выявленных фреймов можно составить таблицу, где строки будут соответствовать самим фреймам, столбцы — языкам, а в ячейках будут приведены лексические единицы, реализующие эти фреймы в разных языках. Цель данной статьи — автоматизированное создание таких фреймов.

Таким образом, задача данного исследования состоит в автоматизированном поиске представлений фреймов, которые реализуют исследуемую семантическую зону «тянуть — толкать».

Работа является актуальной, поскольку до сих пор большинство современных исследований семантических полей используют полностью ручные методы: [3], [6], [8], [9].

Обзор литературы

2.1. Ручные методы

Одной из задач лексической типологии является изучение семантических полей. Например, поле «тянуть — толкать». В задаче необходимо найти все варианты реализации поля в конкретном языке. Данное поле может быть реализовано в следующий вариантах: «открывать от себя», «увеличивать в размере», «привлекать внимание», «перемещать на себя» и т.п. Также дополнительно реализации могут быть сравнены среди разных языков.

Для таких исследований выделяют четыре основных подхода. Первый из них, известный как метод Московской лексико-типологической школы [10], основан на использовании фреймов. В рамках этого подхода каждая ситуация, относящаяся к определённому семантическому полю, описывается с помощью фрейма — набора характеристик, выраженных словами. Например, фрейм «нажимать предмет вперёд» относится к семантической области «тянуть — толкать». В языке этот фрейм реализуется через конкретные лексемы, причём возможное количество фреймов может быть довольно большим. Их выбор исследователь осуществляет на основе словарей, переводных материалов и собственных интуитивных суждений, а также может опираться на синхронные переводы текстов.

На следующем этапе формируется таблица: строки содержат описания фреймов, столбцы — лексемы, а в ячейках фиксируется, соответствует ли конкретная лексема данному фрейму. Существует альтернативная структура таблицы, где строки обозначают фреймы, столбцы — языки, а в ячейках указываются лексемы, через которые этот фрейм реализуется в каждом языке. Основным недостатком такого подхода является ручной характер выделения фреймов. Кроме того, при работе со словарями исследователь вынужден ограничивать область поиска, так как в процессе перевода появляются новые фреймы, зачастую лишь косвенно связанные с изучаемым семантическим полем из-за многозначности слов. Поэтому нередко требуется привлечение носителей соответствующих языков.

Второй подход использует физическое восприятие человека [11]. Исследователь подготавливает набор универсальных стимулов — объектов с определённым вкусом, запахом, цветом или формой — и показывает их носителям языка. Задача информанта заключается в том, чтобы как можно точнее описать представленный объект словами. Сравнивая ответы носителей разных языков, можно определить, какими лексическими средствами выражаются одни и те же стимулы в разных языках. К недостаткам этого метода относятся невозможность отразить всё многообразие лексических единиц с помощью физических стимулов, высокая трудоёмкость и значительные затраты времени. Кроме того, этот подход требует обязательного участия носителей языка.

Третий подход опирается на использование универсальных семантических примитивов, с помощью которых, как предполагается, можно описать любую ситуацию [12]. Этот метод применяет систему из 64 базовых понятий, комбинации которых позволяют выводить все остальные значения. Главными недостатками подхода являются неоднозначность интерпретации получаемых значений и высокая методологическая сложность.

Четвёртый подход основан на анализе параллельных текстовых корпусов. Исследователь определяет переводные соответствия для различных способов реализации конкретной семантической зоны. Основным ограничением этого метода является отсутствие или недостаточная полнота параллельных корпусов для редких и малоизученных языков.

Фреймовый подход является наиболее широко применяемым в лингвистических исследованиях. Так, в работе [3] с его помощью анализируется семантическое поле «мешать». В исследовании [4] рассматривается семантическая область «домашний скот» на материале германских и славянских языков, при этом основным источником данных выступает лексический фонд. В работе [5] при изучении семантических зон «попасть, упасть» и «задеть, попасть» в казымском диалекте хантыйского языка используются корпусные данные, словари и сведения, полученные от носителей языка.

Авторы работы [6] исследуют семантическую зону «острый» в китайском языке с опорой на данные словарей, текстовых корпусов и сведения, полученные от информантов. В работе [7] анализируется семантическое поле «шахматная игра» в русском языке с использованием модели «центр — периферия», где в центре сосредоточены наиболее узкоспециализированные семантические признаки, а на периферии — менее специализированные.

В исследовании [8] для анализа семантической зоны «мягкий», «твёрдый», «жёсткий» применялся метод анкетирования носителей языка. В работе [9] семантическая зона слова «город» исследуется на материале литературных источников.

2.2. Автоматизированные методы

Подходы, основанные на автоматизации, пока остаются слабо развитыми. Так, в работе [13] применяются заранее подготовленные анкеты, которые затем автоматически переводятся на другие языки с использованием словарей и параллельных корпусов. Исследование посвящено семантическим зонам «острый — гладкий» и «толстый — тонкий». Основным ограничением данного подхода является необходимость предварительной подготовки таких анкет.

Авторы исследования [14] используют биграммы Национального корпуса русского языка [1], дополненные различными леммами. Для кластеризации создаются векторные представления: выбираются 10 000 наиболее частотных лексем, после чего для каждого исследуемого слова подсчитывается количество совместных употреблений с каждой из этих лексем в окне шириной пять слов. Для анализа применяются алгоритмы иерархической кластеризации, поскольку методы, не требующие заранее заданного числа кластеров, показали низкую эффективность. Основными недостатками данного подхода являются отсутствие в векторных представлениях семантической информации и данных о контексте употребления.

Современные мультиязычные модели, такие как BERT, показывают способность формировать общее семантическое пространство для разных языков [15] и применяются для изучения лексико-семантических свойств [16]. Это дает возможность использовать их для поиска представлений фреймов в данном исследовании.

Алгоритм работы

3.1. Кластеризация переводов

Исследование проведено на материале текстов Национального корпуса русского языка, а именно на его мультиязычном корпусе [1]. Составляются векторы BERT [17] предложений русского языка и их переводы на английский язык. Затем они кластеризуются методом K-means [18]. Если перевод предложения оказался в другом кластере, возможно, этот перевод дает новое значение семантической зоны. Предложения, чьи переводы оказываются в другом кластере, считаются репрезентациями фреймов, поскольку содержат контекст и позволяют выразить ситуацию, в которой фрейм применяется (как и в работе [2] предложения тоже приводятся для описания фреймов). Также вместо алгоритма K-means исследуется алгоритм DBSCAN [19].

Псевдокод алгоритма:

```
Вход: мультиязычный датасет НКРЯ
Выход: фреймы

sents = corpus[тянуть|толкать]
embeddings = BERT(sents)
KmeansModel = KMeans.fit(embeddings)
DBSCANModel = DBSCAN.fit(embeddings)

func FindSentsInDiffCluster(model, embeddings):
 frames = []
 for index in embeddings, step 2:
   ruCluster model.getCluster(index)
   # Английский перевод хранится за русским в списке
   enCluster = model.getCluster(index + 1)
   IF ruCluster != enCluster:
     frames += corpus[index]
 return frames

framesKMeans = FindSentsInDiffCluster(KmeansModel, embeddings)
framesDBSCAN =  FindSentsInDiffCluster(DBSCANModel, embeddings)

return (framesKMeans, framesDBSCAN)
```

3.2. Кластеризация по сходству

Данный подход аналогичен методу, описанному выше, однако для определения того, попадают ли оригиналы предложений и их переводы в разные кластеры, используется сходство их векторов по косинусной мере. Если разница векторов больше определенного значения, то они попадают в разные кластеры, а иначе — в один. Это значение подбирается экспериментальным методом для достижения наилучшего качества. Предложения, векторы переводов которых оказались в другом кластере, считаются репрезентацией фреймов.

Псевдокод алгоритма:

```
Вход: мультиязычный датасен НКРЯ
Выход: фреймы

sents = corpus[тянуть|толкать]
embeddings = BERT(sents)
func FindSentsInDiffCluster(threshold, embeddings):
 frames = []
 for index in embeddings, step 2:
   #Английский перевод хранится за русским в списке
   similarity = cos(embeddings[index], embeddings[index + 1])
```



```
  IF similarity > threshold:
    frames += corpus[index]
 return frames

frames = FindSentsInDiffCluster(threshold, embeddings)

return (framesKMeans, framesDBSCAN)
```

3.3. Последовательные переводы через промежуточный язык с кластеризацией K-means

В следующем предлагаемом данной статьей подходе используется тезаурус WordNet [20]. Слова из исследуемой семантической зоны переводятся на английский язык. Для них находятся синонимы из тезауруса. Затем каждый синоним переводится в промежуточный язык. Используется переводчик Google через API с помощью библиотеки Googletrans (pypi.org/project/googletrans/). В качестве промежуточного языка выбран польский, поскольку это один из языков в мультиязычном корпусе НКРЯ. Затем слова из промежуточного языка переводятся обратно на русский. Далее итерация повторяется не больше 3 раз, чтобы снизить вероятность перехода в другую семантическую зону. Итерация также прерывается, если после перевода не получено новых слов.

Затем в основном корпусе НКРЯ находятся предложения, содержащие найденные слова. Берутся их векторы BERT, далее выполняется кластеризация методом K-means. Для каждого кластера находится центроид и предложение, вектор которого наиболее близок к центроиду. Найденные предложения считаются репрезентацией фреймов.

Псевдокод алгоритма:

```
Вход: мультиязычный датасет НКРЯ
Выход: фреймы

func FindTranslations(semanticZoneWord, maxIterationCount):
 translations = set(Google.translate('ru - en', semanticZoneWord))
 iteration = 0
 while (iteration <= maxIterationCount AND !isSubset(newTranslations, translation))
   englishWords = set()
   polishWords = set()
   russianWords = set()
   for word in translations:
     englishWords.add(WordNet.synonyms(word))
   for word in englishWords:
     polishWords.add(Google.translate('en-pl', word))
   for word in polishWords:
     russianWords.add(Google.translate('pl-ru', word))
   for word in russianWords:
     englishWords.add(Google.translate('ru-en', word))
   translations = translations.union(englishWords)

 return Google.translate('en-ru', translations)

translations = FindTranslations(semanticZoneWord, 3)
sents = corpus[тянуть|толкать]
embeddings =  BERT(sents)
centroids =  KMeans(embeddingsBERT)

func FindNearest(centroids, embeddings):
 frames []
 for centroid in centroids:
   MaxSim = 0
   MaxSimFrame = 0
   for emd in embeddings:
     simCetroidSent = cos(centroid, emd)
     frame = corpus[number(emd)]
     IF (simCetroidSent > MaxSim and NOT(frame in frames)):
       MaxSim = simCetroidSent
       MaxSimFrame = frame
   frames + = MaxSimFrame
 return frames

frames = FindNearest(centroids, embeddings)

return (frames)
```

Эксперименты

4.1. Кластеризация переводов

Материалом для исследования является мультиязычный датасет Национального корпуса русского языка [1]. Он состоит из 17 597 параллельных переводов (версия от 02.03.2022). Единицами анализа для поиска всех предложений являются лексемы «тянуть» и «толкать».

Находятся все предложения, содержащие данные лексемы. Всего таких предложений для зоны «тянуть» найдено 494, а для зоны «толкать» — 22. Далее все найденные предложения и их переводы кластеризуются. Кластеризация необходима для того, чтобы объединить одинаковые по смыслу предложения в один кластер. Предполагается, что все они являются различными репрезентациями одного и того же фрейма, а предложения в разных кластерах представляют разные фреймы. Таким образом, кластеры состоят из предложений. Если предложение и его перевод попали в разные кластеры, то, возможно, в английском переводе появляется новый фрейм. Находятся все такие предложения, переводы которых попали в другой кластер. Они считаются репрезентациями фреймов для исследуемой семантической зоны. Сам фрейм является смыслом исследуемого слова (представляющим исследуемую семантическую зону) в данном предложении.

Для алгоритма K-means необходимо выбрать количество кластеров, на которые разбивать множество предложений. По методу локтя [21] для зоны «тянуть» было выбрано разбиение на 10 кластеров, а для зоны «толкать» — на 11. Затем после кластеризации для каждого предложения на русском языке было найдено, находится ли его перевод на английский язык в одном с ним кластере или в другом.

Результаты экспериментов показали, что описанный выше алгоритм дает приемлемые по сравнению с работой [2] результаты только для зоны «толкать». Для нее было найдено 11 предложений на русском языке, переводы которых (тоже 11) попали в отличные от оригинала кластер. Переводов, которые попали в один кластер с оригиналом, не найдено.

Ниже приведены некоторые из найденных предложений и их переводов (Здесь и далее примеры взяты из Национального корпуса русского языка [1]). Они представляют собой репрезентацию фреймов.

Но вот на улице появился запыхавшийся и вспотевший человек, который с большим трудом один толкал две тачки с углем.

Just then a man passed by, worn out and wet with perspiration, pulling, with difficulty, two heavy carts filled with coal.

- Да при чем здесь «толкнул»? - сердясь на общую бестолковость, воскликнул Иван, — такому и толкать не надо! Он такие штуки может выделывать, что только держись! Он заранее знал, что Берлиоз попадет под трамвай!

«What are you talking about?» exclaimed Ivan, irritated by his listener's failure to grasp the situation. «He didn't have to push him! He can do thing».

Иван впал в беспокойство, растолкал окружающих, начал размахивать свечой, заливая себя воском, и заглядывать под столы. Тут послышалось слово, «доктора!» — и чье-то ласковое мясистое лицо, бритое и упитанное, в роговых очках, появилось перед Иваном.

Ivan was by now in a state of some excitement. Pushing the bystanders aside he began waving his candle about, pouring wax on himself, and started to look under the tables. Then somebody said «Doctor!» and a fat, kindly face, clean-shaven, smelling of drink and with horn-rimmed spectacles, appeared in front of Ivan.

Для зоны «тянуть» было найдено 490 предложений, переводы которых попали в кластеры, отличные от кластеров оригиналов (при этом только для 4-х предложений соответствующие им переводы оказались в том же кластере). При варьировании параметра значения количества кластеров для алгоритма K-means от 2 до 20 в разные кластеры попадало от 490 до 492 переводов на английский язык. Таким образом, параметр количества кластеров очень слабо влиял на результат. В связи с этим для зоны «тянуть» описанный в данной главе алгоритм не работает, поскольку значение 490 намного больше найденных в работе [2] предложений (в ней для зоны «тянуть» было найдено 10).

Поэтому оценка была произведена только для зоны «толкать»: найденные представления фреймов в виде предложений на русском языке были сравнены с представлением фреймов из работы [2]. Была подсчитана точность, полнота и F-мера. Результаты приведены в таблице 1.

Таблица 1 - Оценка кластеризации переводов K-means

DOI: https://doi.org/10.60797/IRJ.2026.167.64.1

	Толкать
Точность	80
Полнота	36
F-мера	50

Для дальнейших исследований был выбран вариант алгоритма с кластеризацией DBSCAN. Его параметры позволяют задать работу более тонко. Для зоны «тянуть» оптимальными были выбраны следующие параметры:

eps: 0.08

min_samples: 5

При этих параметрах находится 10 предложений. Возможные значения min_samles лежат в диапазоне от 1 до 15. При значениях min_samples больше 15 в разных кластерах становится больше предложений - уменьшается точность. Если уменьшать eps, падает точность, если увеличивать — падает полнота.

Для зоны «толкать» такие же параметры дают только 2 предложения. Оптимальными были выбраны следующие параметры:

eps: 0.06

min_samples: 5

Данные параметры позволяют найти 12 предложений. Отличные значения eps дают меньше предложений и уменьшают полноту. Значение min_samples могут быть в диапазоне от 5 до 9. Значения min_samples меньше 5 дают больше предложений и уменьшают точность, значения min_samples больше 9 дают меньше предложений и уменьшают полноту.

Оценка приведена в таблице 2

Таблица 2 - Оценка кластеризации переводов DBSCAN

DOI: https://doi.org/10.60797/IRJ.2026.167.64.2

	Тянуть	Толкать	Тянуть - толкать
Точность	70	80	75
Полнота	60	46	53
F-мера	65	58	62

4.2. Кластеризация по сходству

Для зоны «тянуть» оптимальный порог близости был найден 0.82, поскольку на нем достигаются наилучшие значения точности и полноты. Для зоны «толкать» оптимальный порог сходства был найден 0.905. Было найдено 11 предложений в разных кластерах. Оценка приведена в таблице 3.

Таблица 3 - Кластеризации по сходству BERT

DOI: https://doi.org/10.60797/IRJ.2026.167.64.3

	Тянуть	Толкать	Тянуть - толкать
Точность	57	78	70
Полнота	60	73	67
F-мера	59	76	69

Таким образом, фильтрация по близости векторов позволяет задавать более тонкую настройку и более гибко оптимизировать работу.

4.3. Последовательные переводы через промежуточный язык с кластеризацией K-means

Эксперименты показали, что со второй итерации алгоритма переводы стали достаточно отдаленные от первоначальной зоны «тянуть», поэтому была сделана одна итерация алгоритма. Найдено 23 слова. Затем в корпусе НКРЯ были найдены все предложения с этими словами, найдены их векторы BERT и кластеризованны в 17 кластеров по методу локтя [21].

Некоторые из найденных предложений:

Артист вытянул вперед руку, на пальцах которой сверкали камни, как бы заграждая уста буфетчику, и заговорил с большим жаром,

Белая волокнистая пелена, затянувшая почти все болото, с каждой минутой приближалась к дому.

Он быстро поел, а в столовую еще тянулись сгорбленные старцы и старухи.

Мероприятие, предполагавшее живой обмен мнениями, затянулось на два часа и порой напоминало лекцию на юрфаке.

— Да — как будто удивился он, потом протянул в раздумье: — Оч-чень хо-ро-шо!

Для подзоны «толкать», как и для подзоны «тянуть», со второй итерации идет сильный переход в другие зоны, поэтому была сделана только одна итерация переводов.

Был найден фрейм «нажимать на кнопку», который не был найден ни в одном их предыдущих алгоритмов:

«Человек нажал кнопку, включавшую аппарат, и тонкий луч света, пронзив пространство, прямой наводкой попал в глазок камеры.»

Количество кластеров по методу локтя для зоны «тянуть» было выбрано 11, а для «толкать» 10.

Кластеры представлены векторами BERT предложений, поэтому для их визуализации необходимо уменьшить размерность до 2-х главных компонент. Это сделано с помощью метода PCA [22]. Визуализация представлена на рисунках 1 и 2 для зоны «тянуть» и «толкать» соответственно с помощью библиотеки matplotlib.

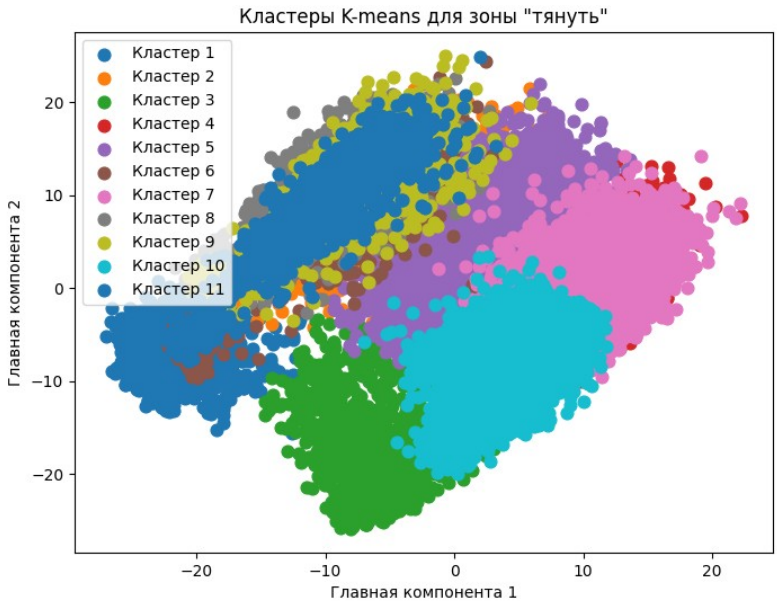

Рисунок 1 - Кластеры K-means для зоны «тянуть»
DOI: https://doi.org/10.60797/IRJ.2026.167.64.4

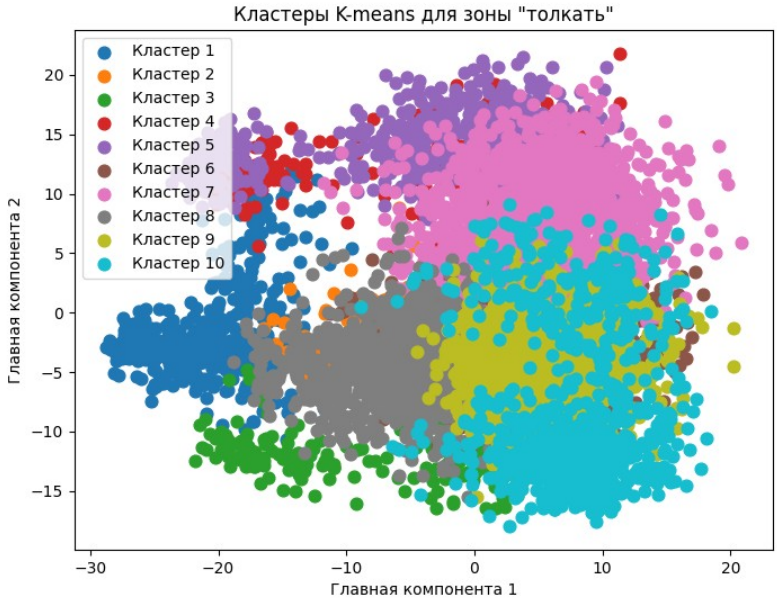

Рисунок 2 - Кластеры K-means для зоны «толкать»
DOI: https://doi.org/10.60797/IRJ.2026.167.64.5

Оценка приведена в таблице 4.

Таблица 4 - Последовательные переводы и кластеризация

DOI: https://doi.org/10.60797/IRJ.2026.167.64.6

	Тянуть	Толкать	Тянуть - толкать
Точность	71	90	78
Полнота	80	60	70
F-мера	75	72	74

Результаты

В Таблице 5 приведены сравнительные оценки работы алгоритмов. Жирным выделен самый лучший результат среди всех алгоритмов по соответствующей оценке. Подчеркиванием выделен самый лучший результат для каждой из зон «тянуть», «толкать», «тянуть — толкать».

Таблица 5 - Сравнение работы алгоритмов

DOI: https://doi.org/10.60797/IRJ.2026.167.64.7

	К-П-К-Т1	К-П-К-Т2	К-П-К-ТО	К-П-В-Т1	К-П-В-Т2	К-П-В-ТО	К-П-П-Т1	К-П-П-Т2	К-П-П-О
P	70	80	75	57	78	70	71	**90**	78
R	60	46	53	60	73	67	**80**	60	70
F	65	58	62	59	**76**	69	75	72	74

Примечание: К-П-К-Т1 - Кластеризация переводов K-means для зоны «тянуть», К-П-К-Т2 - Кластеризация переводов K-means для зоны «толкать», К-П-К-ТО - Кластеризация переводов K-means для зоны «тянуть - толкать»; К-П-В-Т1 - Кластеризация переводов BERT для зоны «тянуть», К-П-В-Т2 - Кластеризация переводов BERT для зоны «толкать», К-П-В-ТО - Кластеризация переводов K-means для зоны «тянуть - толкать»; К-П-П-Т1 - Кластеризация последовательны переводов для зоны «тянуть», К-П-П-Т2 - Кластеризация последовательны переводов для зоны «толкать», К-П-П-ТО - Кластеризация последовательны переводов для зоны «тянуть - толкать»

Самые лучшие результаты показал подход с параллельными переводами и последующей кластеризацией K-means. Он достиг лучших результатов по точности (зона «толкать») и по полноте (зона «тянуть»). Также достиг самых лучших показателей по всем зонам кроме полноты для зоны «толкать» и F-меры для зоны «толкать».

Немного хуже показал результаты подход с фильтрацией BERT. Он достиг самые лучшие результаты по F-мере (для зоны «толкать»), также показал самый лучший результат по полноте для зоны «толкать».

Подход с кластеризацией параллельных переводов занял 3 место. Однако отставание незначительное. По точности для зоны «тянуть» он показал почти такое же значение, как подход с последовательными переводами, и выше, чем подход с фильтрацией по близости BERT. По полноте для зоны «тянуть» показал такой же результат, как второй подход. Низкие результаты он дал только по полноте для зоны «толкать».

Выводы

Была поставлена задача содействия решению лексико-типологической проблемы определения фреймов семантической зоны «тянуть — толкать» с использованием автоматизированных подходов. Традиционно задача решается ручными методами, которые требуют привлечения лингвистов, носителей языка, лингвистических ресурсов, времени и трудозатрат.

Для автоматизации ее решения были предложены и реализованы методы, использующие современные контекстуализированные векторные представления BERT, алгоритмы кластеризации DBSCAN и K-means, мультиязычный корпус НКРЯ и словарь переводов.

Был исследован алгоритм кластеризации векторов BERT переводов с помощью K-means и DBSCAN с последующим выявлением предложений, чьи переводы попали в другой кластер в отличие от оригинала. Такие предложения считались репрезентацией фреймов для исследуемой семантической зоны. Его вариант с K-means показал неудовлетворительные результаты для зоны «тянуть». Возможно, это связано с тем, что векторы переводов достаточно близки, и их разница недостаточна для нахождения новых фреймов. При этом при замене K-means на DBSCAN алгоритм начинает работать. Это может быть связано с тем, что DBSCAN позволяет более гибко регулировать работу алгоритма.

Также был исследован алгоритм кластеризации векторов BERT переводов с помощью сходству по косинусной мере. В нем для определения того, что предложения и их переводы попали в разный кластер, вместо DBSCAN и K-means используется сходство векторов по косинусной мере. Если сходство больше опредлённого порога, найденного экспериментально, то вектор перевода попадает в отличный от оригинального предложения кластер. Все предложения, где вектор перевода попал в другой кластер, считаются репрезентацией фреймов исследуемой зоны. Подход показал более стабильные результаты и занял второе место среди исследуемых алгоритмов. Это говорит о том, что порог сходства по косинусной мере BERT позволяет достаточно точно находить репрезентации фреймов.

Самых лучших результатов удалось достичь с помощью подхода с последовательными переводами с русского на английский через промежуточный язык, расширением слов из тезауруса и последующей кластеризацией K-means. В нем слова исследуемой семантической зоны последовательно переводятся с русского на английский через промежуточный язык, расширяются синонимами из тезауруса, затем в корпусе НКРЯ находятся все предложения, содержащие данные слова, и эти предложения кластеризуются алгоритмом K-means. Для каждого кластера находится центроид и наиболее приближенный к нему вектор предложения из кластера. Предложения, которые соответствуют данным векторам, считаются репрезентациями фреймов. Подход показал наилучшие результаты, поскольку генерация слов через последовательные переводы и добавление синонимов через тезаурус дает много слов, которые принадлежат исследуемой семантической зоне, при этом не сильно выходят за ее пределы. Кластеризация их векторов BERT

помогает объединить по смыслу одинаковые предложения с данными словами и получить предложения, наиболее точно представляющие каждый из фреймов семантической зоны.

Заключение

Было реализовано три подхода к автоматизации поиска репрезентаций фреймов в области лексической типологии на параллельных корпусах и с помощью последовательных переводов через промежуточный язык. Метод с последовательными переводами демонстрирует результаты лучше, чем методы на параллельных корпусах. При этом все три подхода дают близкие результаты. Поэтому для содействия решению данной задачи можно применять любой из трех алгоритмов. Они показали свою работоспособность и гибкость. Таким образом, были разработаны методы, позволяющие автоматизировать часть работы лингвистов и сделать ее более эффективной.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Савчук С.О. Национальный корпус русского языка 2.0: новые возможности и перспективы развития / С.О. Савчук, А.А. Архангельский, А.А. Бонч-Осмоловская и др. // Вопросы языкознания. — 2024. — 2. — С. 7–34.

2. Савельева А.Ю. Глаголы семантических зон «ТЯНУТЬ» и «ТОЛКАТЬ» в типологической перспективе / А.Ю. Савельева // Проблемы компьютерной лингвистики и типологии: сб. Всерос. конф. — Воронеж: Издательский дом ВГУ, 2017. — С. 142–152.

3. Дунаева К.О. Семантическое поле «мешать» в типологической перспективе / К.О. Дунаева, В.В. Маринина // XXVI Открытая конференция студентов-филологов в СПбГУ. — Санкт-Петербург: СПбГУ, 2023. — С. 34–37.

4. Шешкина Т.Ф. Германо-славянские параллели семантического поля «Домашний скот» в немецких лексикографических источниках / Т.Ф. Шешкина // Филологические науки. Вопросы теории и практики. — 2020. — 6. — С. 303–307. — DOI: 10.30853/filnauki.2020.6.57

5. Рыжова Д.А. Фрагмент лексической системы казымского диалекта хантыйского языка: глаголы pitti «упасть, попасть» и χɔjti «задеть, попасть» и их аргументная структура / Д.А. Рыжова // Урало-алтайские исследования. — 2022. — 2(45). — С. 123–140. — DOI: 10.37892/2500-2902-2022-45-2-123-140

6. Холкина Л.С. Семантическое поле ОСТРЫЙ в китайском языке: диахроническое развитие и его отражение в современных диалектах / Л.С. Холкина, Л.О. Наний, Ц. Сы // Journal of Language Relationship. — 2023. — 20(3-4). — С. 280–298. — DOI: 10.31826/jlr-2023-203-410

7. Влавацкая М.В. Лексико-семантическое поле «шахматная игра» в современном русском языке / М.В. Влавацкая // Мир науки, культуры, образования. — 2022. — 2(93). — С. 293–297.

8. Кашкин Е.В. Категоризация качественных признаков «мягкий», «твердый», «жесткий» в горномарийском языке / Е.В. Кашкин // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. — 2022. — 1. — С. 140–150. — DOI: 10.17308/lic.2022.1/9009

9. Григорьева О.Н. Лексико-семантическая группа «город» в современных российских масс-медиа / О.Н. Григорьева // Вестник Московского государственного областного университета. Серия: Русская филология. — 2018. — 5. — С. 31–37. — DOI: 10.18384/2310-7278-2018-5-31-38

10. Рахилина Е.В. Фреймовый подход к лексической типологии / Е.В. Рахилина, Т.И. Резникова // Вопросы языкознания. — 2013. — 2. — С. 3–31.

11. Berlin B. Color Terms: Their Universality and Evolution / B. Berlin. — Berkeley: Berkeley: University of California Press, 1969. — 178 p.

12. Wierzbicka A. Semantic and lexical universals: Theory and emperical findings / A. Wierzbicka // Linguistic Investigations. — 1994. — 21. — P. 249–261.

13. Kyuseva M. Automatic data collection in lexical typology / M. Kyuseva, E. Parina, D. Ryzhova // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018». — 2018. — 1. — P. 29–55.

14. Рыжова Д.А. Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики / Д.А. Рыжова // ВЕСТНИК РГГУ. Сер.: История. Филология. Культурология. Востоковедение. — 2016. — 18. — С. 140–150.

15. Karthikeyan K. Cross-Lingual Ability of Multilingual BERT: An Empirical Study / K. Karthikeyan, Z. Wang, S. Mayhew et al. // International Conference on Learning Representations. — 2020. — 1. — DOI: 10.48550/arXiv.1912.07840

16. Ruder S. Survey of Cross-lingual Word Embedding Models / S. Ruder, I. Vulić, A. Søgaard // Journal of Artificial Intelligence Research. — 2019. — 65. — P. 569–631. — DOI: 10.1613/jair.1.11640

17. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2019. — 1. — P. 4171–4186. — DOI: 10.18653/v1/N19-1423

18. Jin X. K-Means Clustering / X. Jin, J. Han // Encyclopedia of Machine Learning. — 2011. — 1. — P. 563–563. — DOI: 10.1007/978-0-387-30164-8_425

19. Martin E. A density-based algorithm for discovering clusters in large spatial databases with noise / E. Martin, K. Hans-Peter, S. Jorg // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. — 1996. — 1. — P. 226–231. — DOI: 10.13140/RG.2.1.4420.1448

20. Miller A.G. Introduction to WordNet: An On-line Lexical Database / A.G. Miller, R. Beckwith, C. Fellbaum et al. // International Journal of Lexicography. — 1991. — 3(4). — P. 235–244. — DOI: 10.1093/oso/9780199292332.003.0022

21. Thorndike L.R. «Who Belongs in the Family?« / L.R. Thorndike // Psychometrika. — 1953. — 18 (4). — P. 267–276.

22. Hotelling H. Analysis of a complex of statistical variables into principal components / H. Hotelling // Journal of Educational Psychology. — 1932. — 24(6). — P. 417–441. — DOI: 10.1037/h0071325

Список литературы на английском языке / References in English

1. Savchuk S.O. Nacional'ny'j korpus russkogo yazy'ka 2.0: novy'e vozmozhnosti i perspektivy' razvitiya [The National Corpus of the Russian Language 2.0: New Opportunities and Prospects for Development] / S.O. Savchuk, A.A. Arxangel'skij, A.A. Bonch-Osmolovskaya et al. // Issues of Linguistics. — 2024. — 2. — P. 7–34. [in Russian]

2. Saveleva A.Yu. Glagoli semanticheskikh zon "TYaNUT" i "TOLKAT" v tipologicheskoi perspektive [Verbs belonging to the semantic domains "PULL" and "PUSH" from a typological perspective] / A.Yu. Saveleva // Issues in Computational Linguistics and Typology: Proceedings of the All-Russian Conference. — Voronezh: VSU Publishing House, 2017. — P. 142–152. [in Russian]

3. Dunaeva K.O. Semanticheskoe pole "meshat" v tipologicheskoi perspektive [The semantic field of "to interfere" from a typological perspective] / K.O. Dunaeva, V.V. Marinina // The XXVI Open Conference for Philology Students at St Petersburg State University. — Saint Petersburg: SPbGU, 2023. — P. 34–37. [in Russian]

4. Sheshkina T.F. Germano-slavyanskie paralleli semanticheskogo polya «Domashnij skot» v nemeczkix leksikograficheskix istochnikax [Germanic-Slavic parallels in the semantic field of "livestock" in German lexicographical sources] / T.F. Sheshkina // Philological Studies: Theory and Practice. — 2020. — 6. — P. 303–307. — DOI: 10.30853/filnauki.2020.6.57 [in Russian]

5. Ry'zhova D.A. Fragment leksicheskoj sistemy' kazy'mskogo dialekta xanty'jskogo yazy'ka: glagoly' pitti «upast', popast'» i χɔjti «zadet', popast'» i ix argumentnaya struktura [A fragment of the lexical system of the Kazym dialect of the Khanty language: the verbs pitti "to fall, to hit" and χɔjti "to touch, to hit" and their argument structure] / D.A. Ry'zhova // Ural-Altaic Studies. — 2022. — 2(45). — P. 123–140. — DOI: 10.37892/2500-2902-2022-45-2-123-140 [in Russian]

6. Xolkina L.S. Semanticheskoe pole OSTRY'J v kitajskom yazy'ke: diaxronicheskoe razvitie i ego otrazhenie v sovremenny'x dialektax [The semantic field of ‘SHARP’ in the Chinese language: diachronic development and its reflection in modern dialects] / L.S. Xolkina, L.O. Nanij, Cz. Sy' // Journal of Language Relationship. — 2023. — 20(3-4). — P. 280–298. — DOI: 10.31826/jlr-2023-203-410 [in Russian]

7. Vlavaczkaya M.V. Leksiko-semanticheskoe pole «shaxmatnaya igra» v sovremennom russkom yazy'ke [The lexical-semantic field of ‘chess’ in modern Russian] / M.V. Vlavaczkaya // The world of science, culture and education. — 2022. — 2(93). — P. 293–297. [in Russian]

8. Kashkin E.V. Kategorizaciya kachestvenny'x priznakov «myagkij», «tverdy'j», «zhestkij» v gornomarijskom yazy'ke [The categorisation of the qualitative features ‘soft’, ‘hard’ and ‘stiff’ in the Hill Mari language] / E.V. Kashkin // VSU Bulletin. Series: Linguistics and Intercultural Communication. — 2022. — 1. — P. 140–150. — DOI: 10.17308/lic.2022.1/9009 [in Russian]

9. Grigor'eva O.N. Leksiko-semanticheskaya gruppa «gorod» v sovremenny'x rossijskix mass-media [The lexical-semantic group ‘city’ in contemporary Russian mass media] / O.N. Grigor'eva // Bulletin of Moscow State Regional University. Series: Russian Philology. — 2018. — 5. — P. 31–37. — DOI: 10.18384/2310-7278-2018-5-31-38 [in Russian]

10. Raxilina E.V. Frejmovy'j podxod k leksicheskoj tipologii [A frame-based approach to lexical typology] / E.V. Raxilina, T.I. Reznikova // Issues of Linguistics. — 2013. — 2. — P. 3–31. [in Russian]

11. Berlin B. Color Terms: Their Universality and Evolution / B. Berlin. — Berkeley: Berkeley: University of California Press, 1969. — 178 p.

12. Wierzbicka A. Semantic and lexical universals: Theory and emperical findings / A. Wierzbicka // Linguistic Investigations. — 1994. — 21. — P. 249–261.

13. Kyuseva M. Automatic data collection in lexical typology / M. Kyuseva, E. Parina, D. Ryzhova // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018». — 2018. — 1. — P. 29–55.

14. Ry'zhova D.A. Opy't avtomaticheskogo postroeniya ankety' dlya leksiko-tipologicheskogo issledovaniya prilagatel'ny'x i odnomestny'x glagolov s pomoshh'yu modelej distributivnoj semantiki [An experiment in automatically constructing a questionnaire for a lexical-typological study of adjectives and monovalent verbs using distributional semantics models] / D.A. Ry'zhova // RSHU Bulletin. Series: History. Philology. Cultural Studies. Oriental Studies. — 2016. — 18. — P. 140–150. [in Russian]

15. Karthikeyan K. Cross-Lingual Ability of Multilingual BERT: An Empirical Study / K. Karthikeyan, Z. Wang, S. Mayhew et al. // International Conference on Learning Representations. — 2020. — 1. — DOI: 10.48550/arXiv.1912.07840

16. Ruder S. Survey of Cross-lingual Word Embedding Models / S. Ruder, I. Vulić, A. Søgaard // Journal of Artificial Intelligence Research. — 2019. — 65. — P. 569–631. — DOI: 10.1613/jair.1.11640

17. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2019. — 1. — P. 4171–4186. — DOI: 10.18653/v1/N19-1423

18. Jin X. K-Means Clustering / X. Jin, J. Han // Encyclopedia of Machine Learning. — 2011. — 1. — P. 563–563. — DOI: 10.1007/978-0-387-30164-8_425

19. Martin E. A density-based algorithm for discovering clusters in large spatial databases with noise / E. Martin, K. Hans-Peter, S. Jorg // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. — 1996. — 1. — P. 226–231. — DOI: 10.13140/RG.2.1.4420.1448

20. Miller A.G. Introduction to WordNet: An On-line Lexical Database / A.G. Miller, R. Beckwith, C. Fellbaum et al. // International Journal of Lexicography. — 1991. — 3(4). — P. 235–244. — DOI: 10.1093/oso/9780199292332.003.0022

21. Thorndike L.R. «Who Belongs in the Family?« / L.R. Thorndike // Psychometrika. — 1953. — 18 (4). — P. 267–276.

22. Hotelling H. Analysis of a complex of statistical variables into principal components / H. Hotelling // Journal of Educational Psychology. — 1932. — 24(6). — P. 417–441. — DOI: 10.1037/h0071325