С. И. Монахов, В. В. Турчаненко, Е. А. Федюкова, Д. Н. Чердаков

Санкт-Петербург

# АНАЛИЗ ТЕРМИНОЛОГИИ В СОВРЕМЕННЫХ ШКОЛЬНЫХ УЧЕБНИКАХ МЕТОДАМИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

При финансовой поддержке РФФИ, проект № 19-29-14032

# СПОСОБЫ ВЫЧЛЕНЕНИЯ ТЕРМИНОВ ИЗ ТЕКСТОВ

## Традиционные «ручные» способы

отличаются надежностью результатов, но

плохо приложимы к большим массивам данных;

не выявляют частотности употребления терминов;

не выявляют специфику их синтагматических связей.

## Традиционные автоматизированные способы

используют статистический подход и опираются на сравнение частоты встречаемости лексических элементов в двух корпусах:

— целевом (откуда необходимо вычленить термины)
и
— референционном (представляющим систему языка в целом).

Единицы целевого корпуса, отличающиеся неожиданно высокой частотностью по сравнению с частотностью в референционном корпусе, получают в системе статистическую меру ключевого слова (keyness score) и при превышении данной мерой определенного порога определяются в качестве кандидатов на присвоение им терминологического статуса.

## ЦЕЛЕВОЙ КОРПУС
### 212 учебников с 5-го по 11-й класс по 21 дисциплине

1) алгебра, 2) астрономия, 3) биология
4) всеобщая история и история России,
5) география, 6) геометрия, 7) естествознание
8) изобразительное искусство,
9) информатика, 10) литература, 11) математика,
12) математический анализ,
13) мировая художественная культура, 14) музыка,
15) обществознание, 16) право, 17) русский язык,
18) технология, 19) физика, 20) физическая культура
21) химия

Учебники входят в федеральный перечень изданий, рекомендованных Министерством просвещения.

Было получено согласие издательства «Просвещение» на использование текстов этих учебников в исследовательских целях.

Общий объем целевого корпуса — около 14 370 000 слов.

По требованиям правообладателя корпус доступен только исследовательской группе.

Разделен на подкорпусы согласно учебным дисциплинам и годам обучения и загружен на платформу Sketch Engine (https://www.sketchengine.eu).

## РЕФЕРЕНЦИОННЫЙ КОРПУС
### Russian Web 2011 Sample (ruTenTen11)

Доступен на платформе Sketch Engine, содержит более 900 миллионов слов из русскоязычных интернет-текстов.

## ПРИМЕР ВЫДЕЛЕНИЯ КАНДИДАТА В ТЕРМИНЫ

(для однословных единиц)

Для однословных единиц высчитывалась метрика keyness score по формуле:

((Lt * 1,000,000 / Ct) +1) / ((Lr * 1,000,000 / Cr) +1),

где Lt — частота употребления лексемы в целевом корпусе, Ct — общее количество токенов в целевом корпусе, Lr — частота употребления лексемы в референционном корпусе, Cr — общее количество токенов в референционном корпусе.

Если значение метрики keyness score превышало 1, данная лексема включалась в список терминологических кандидатов.

Так, для термина «многочлен», встречающегося в целевом подкорпусе по учебной дисциплине «Алгебра» (7–9 классы), было получено следующее значение метрики keyness score: 1003.785 + 1 / 0.352 + 1 = 743.18, что обусловливает зачисление слова в кандидаты в термины.

**КАНДИДАТЫ В ТЕРМИНЫ**
**выделены на основании автоматического сравнения относительной частоты употребления слов и сочетаний в целевом и референционном корпусе**

Все дисциплины, все классы с 5-го по 11-й — 48 911 единиц.

Данные по некоторым учебным дисциплинам:

алгебра — 1655
биология — 3863
всеобщая история и история России — 5198
география — 3352
геометрия — 806
информатика — 1294
русский язык — 3780
физика — 3254
химия — 2095

# ВЫСОКОЧАСТОТНАЯ ЛЕКСИКА В УЧЕБНИКЕ ПО РУССКОМУ ЯЗЫКУ ДЛЯ 5 КЛАССА

**собственно термины**

| | | | | | |
|---|---|---|---|---|---|
| **19** | местоимение | 0 | 4 | 5 | ruslang |
| **20** | морфологический_признак | 0 | 4 | 5 | ruslang |
| **22** | глагольный_форма | 0 | 4 | 5 | ruslang |
| **23** | наречие | 0 | 4 | 5 | ruslang |
| **37** | глагол | 0 | 4 | 5 | ruslang |
| **75** | существительное | 0 | 4 | 5 | ruslang |
| **78** | форма | 0 | 4 | 5 | ruslang |
| **80** | настоящий_время | 0 | 4 | 5 | ruslang |
| **83** | множественный_число | 0 | 4 | 5 | ruslang |
| **106** | форма_глагол | 0 | 4 | 5 | ruslang |
| **110** | лицо | 0 | 4 | 5 | ruslang |
| **132** | прилагательное | 0 | 4 | 5 | ruslang |
| **133** | краткий_форма | 0 | 4 | 5 | ruslang |
| **138** | предлог | 0 | 4 | 5 | ruslang |

**лжетермины**

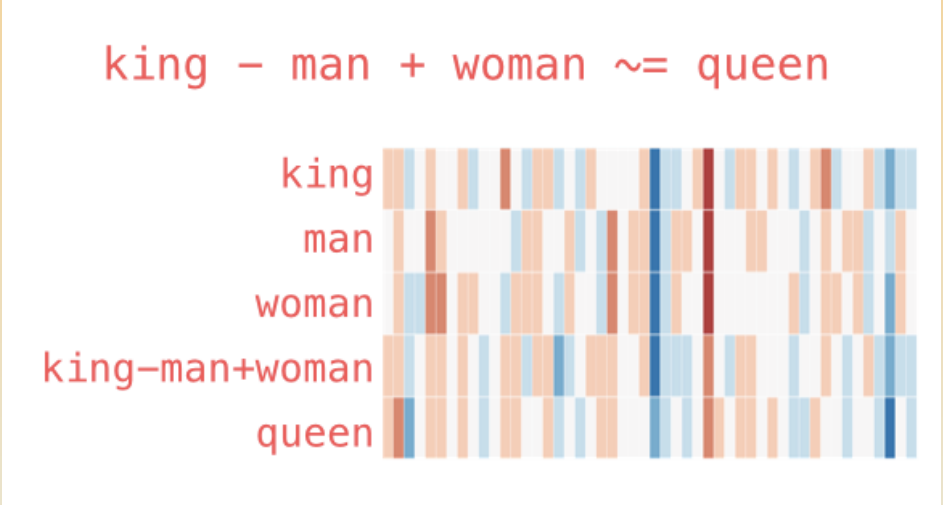
| | | | | | |
|---|---|---|---|---|---|
| **173** | соловей | 1 | 5 | 5 | ruslang |
| **203** | роща | 1 | 5 | 5 | ruslang |
| **215** | вьюга | 1 | 5 | 5 | ruslang |
| **216** | метель | 1 | 5 | 5 | ruslang |
| **220** | весна | 1 | 5 | 5 | ruslang |
| **227** | туча | 1 | 5 | 5 | ruslang |
| **241** | чаща | 1 | 5 | 5 | ruslang |
| **242** | шалаш | 1 | 5 | 5 | ruslang |
| **265** | иней | 1 | 5 | 5 | ruslang |
| **292** | шорох | 1 | 5 | 5 | ruslang |
| **337** | снежный | 1 | 5 | 5 | ruslang |
| **351** | синий_небо | 1 | 5 | 5 | ruslang |
| **362** | лесник | 1 | 5 | 5 | ruslang |
| **372** | жаворонок | 1 | 5 | 5 | ruslang |

# МЕТОДЫ ДИСТРИБУТИВНОЙ СЕМАНТИКИ, РЕАЛИЗОВАННЫЕ В НАБОРЕ КОМПЬЮТЕРНЫХ АЛГОРИТМОВ WORD2VEC

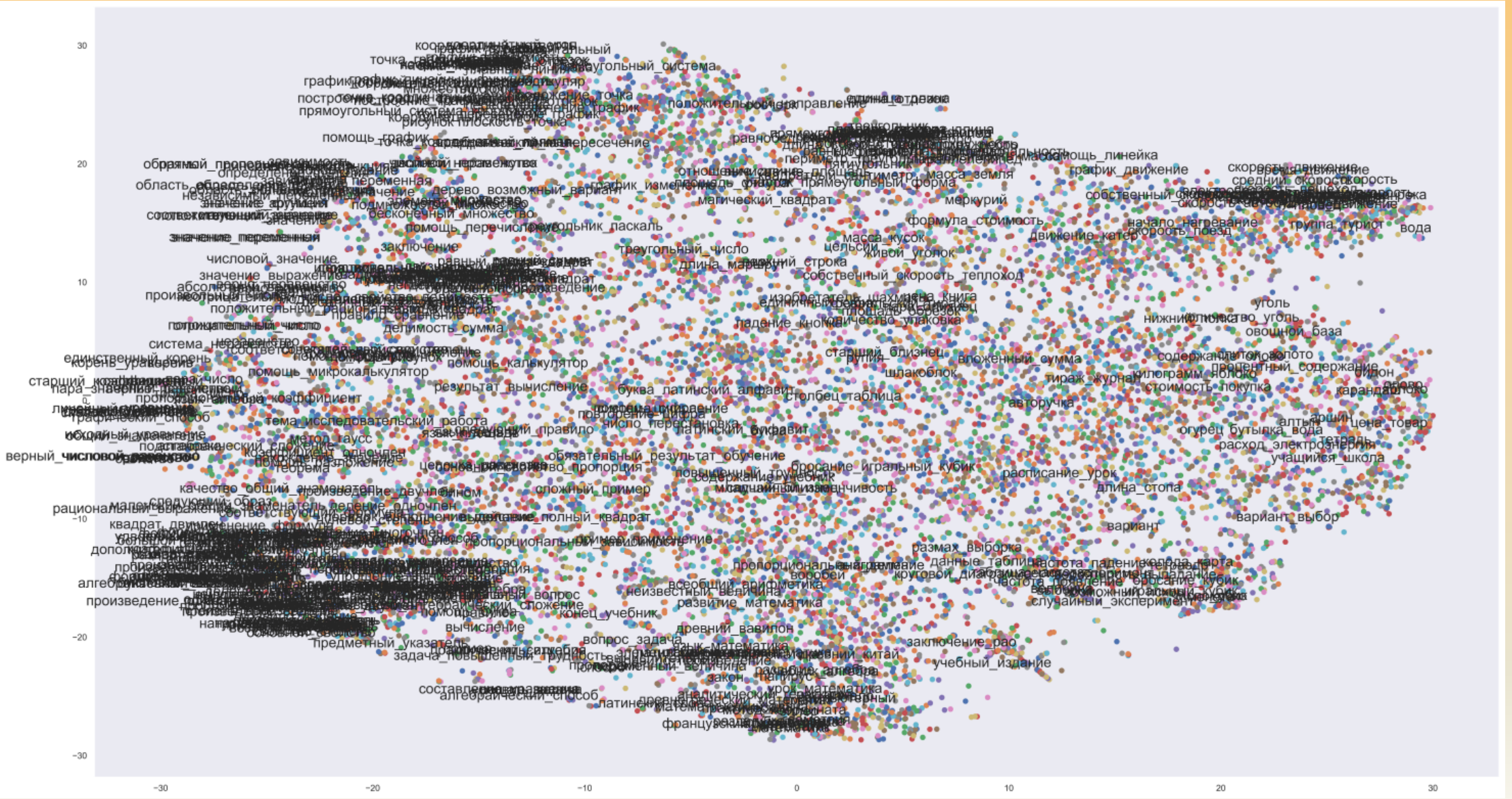
Значение слова может быть выведено из его лексического окружения в текстах.

Значение слова, выведенное из его сочетаемостных связей, может быть представлено в векторном пространстве.

Векторы слов, как их семантическое представление, могут сопоставляться, складываться, вычитаться.




Слова могут быть автоматически сгруппированы по степени семантической близости на основе контекстов, в которых они встречаются.




Векторное пространство терминологических кандидатов
подкорпуса «Алгебра. 7 класс»

**УТОЧНЕННЫЕ ДАННЫЕ**
**о кандидатах в термины после применения алгоритмов Word2Vec**
**и автоматического отделения терминов от лжетерминов**

Все дисциплины, все классы с 5-го по 11-й — 48 911 единиц; терминов — 26 282; лжетерминов — 22 629.

Данные по некоторым учебным дисциплинам:

| | | |
|---|---|---|
| алгебра — всего 1655 | терминов — 1526 | лжетерминов — 129 |
| биология — всего 3863 | терминов — 2324 | лжетерминов — 1539 |
| всеобщая история и история России — всего 5198 | терминов — 2491 | лжетерминов — 2707 |
| география — всего 3352 | терминов — 1635 | лжетерминов — 1717 |
| геометрия — всего 806 | терминов — 570 | лжетерминов — 236 |
| информатика — всего 1294 | терминов — 682 | лжетерминов — 612 |
| русский язык — всего 3780 | терминов — 2633 | лжетерминов — 1147 |
| физика — всего 3254 | терминов — 2836 | лжетерминов — 418 |
| химия — всего 2095 | терминов — 1087 | лжетерминов — 288 |

# ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

**методологическая область**

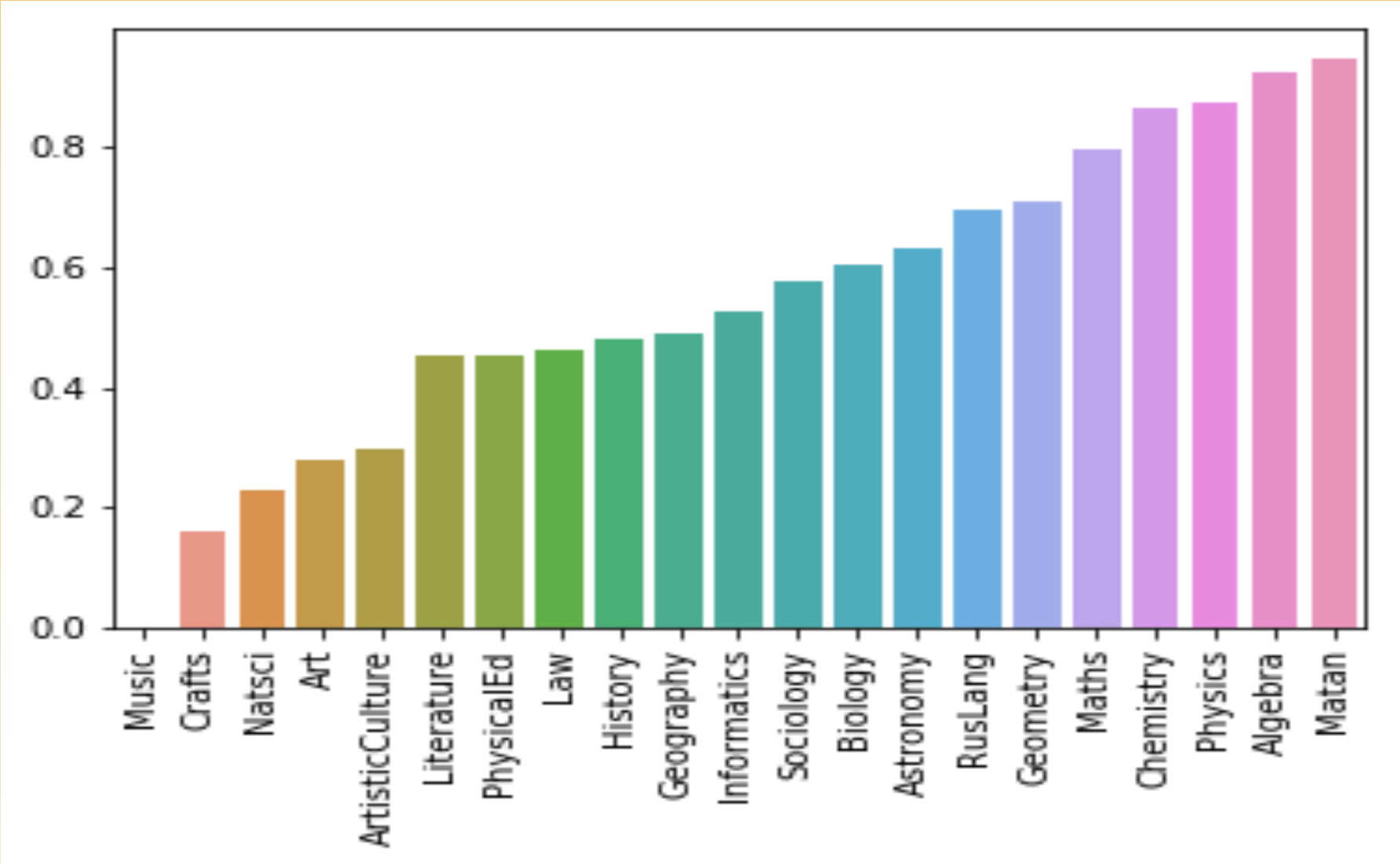
— возможность применения алгоритмов Word2Vec и методов дистрибутивной семантики в целях оптимизации автоматического вычленения терминов из целевого корпуса

**дидактическая область**

— возможность оценить степень терминологической насыщенности учебников в сопоставительном аспекте

— проблема высокочастотной лексики, не являющейся терминологической

— проблема принципиального расхождения в частотности между коррелирующими терминами




Коэффициенты терминологической насыщенности
(доля терминов от общего числа высокочастотной лексики)
для учебников по разным дисциплинам

Все материалы, необходимые для воспроизведения результатов и верификации выводов, за исключением текстов учебников, являющихся собственностью правообладателя, размещены в постоянном научном хранилище Zenodo и доступны по адресу: https://zenodo.org/record/4079198#.X4Mrfy1h29Y.

**СПАСИБО ЗА ВНИМАНИЕ!**