**Баранов Виктор Аркадьевич**
Ижевский государственный технический университет
имени М.Т. Калашникова
**Гнутиков Роман Михайлович**
Удмуртский государственный университет
УДК 519.2:801.82(045)

# Дистрибутивный словарь славянских текстов X–XV веков: параметры запросов и визуализация результатов

*корпус славянских рукописей X–XV веков, лексическая дистрибуция, корпусный менеджер*

Известное положение традиционной лингвистики о взаимозаменяемости как критерии синонимии (см., напр.: [Белов 2018]) получил развитие в качестве дистрибутивной гипотезы [Sahlgren 2008], согласно которой слова, использующиеся в окружении одних и тех же слов, близки и по смыслу. Эти положения, а также сведения о количественных и/или статистических характеристиках лексического окружения анализируемых слов в настоящее время легли в основу разработанных прикладных методов и инструментальных средств извлечения из текстов близких по семантике лингвистических единиц (см., напр.: [Белов и др. 2020: 11, 12; Максименко 2018]). Сопоставление сведений о лексическом составе контекстов, в которых встречаются две анализируемые лингвистические единицы, с аналогичными данными для всех других пар единиц в некотором объеме текстов позволяет выявить степень близости лексического окружения слов, а соответственно, найти в текстах близкие по парадигматическим свойствам слова [Шайкевич 1976]. Вычисление отклонения совместной встречаемости двух лингвистических единиц от статистически ожидаемой составляет основу статистических методик и позволяет получить синтагматические характеристики лингвистических единиц – степень их статистической тесноты (близости) [Максименко 2019: 94]. Как было неоднократно показано, между количественными и статическими подходами, а соответственно, парадигматическими и синтагматическими свойствами текстовых единиц существует корреляция: чем чаще слова встречаются в одних и тех же контекстах и чем сильнее их статистическая близость при этом, тем выше их парадигматическая связь [Шайкевич 1976: 370; Захаров 2018: 4; Максименко 2019: 94].

Одним из популярных приемов вычисления семантической близости слов являются операции с векторами слов, полученными в результате извлечения из текстовых корпусов числовых значений их (слов) сочетаемости (см., напр.: [Золотарев и др. 2018: 43-45]). За единицу совместной встречаемости могут быть приняты различные величины – абсолютная или относительная совместная встречаемость двух лингвистических единиц или одна из статистических мер их тесноты, например PPMI [Белов 2020: 13, 16-17], или иная (MI, PMI, Log-Likelihood, T-score и др.).

Особым направлением работы в области автоматического семантического анализа является создание специализированного компьютерного инструментария, предназначенного для обработки больших массивов лингвистических данных. Чаще всего компьютерные общедоступные сервисы, программы, процедуры ориентированы на решение широкого круга задач, объектом которых являются современные тексты (см., напр.: [Золотарев и др. 2018: 45–46; Белов 2020: 18–20]). Хорошо известны, например, веб-инструменты, демонстрирующие возможности автоматического нахождения близких по семантике слов, – проекты «Serelex: поиск семантически связных слов», «WebVectors: word embeddings online», «RusVectōrēs: семантические модели для русского языка».

В публикации [Баранов 2022] дано описание прототипа исторического дистрибутивного словаря – оснований и методов, текстового материала, процедур менеджера, параметров запроса для демонстрации семантически близких слов. Построенный на лингвистических данных исторического корпуса «Манускрипт» (manuscripts.ru), содержащего транскрипции славянских рукописей X–XV вв., модуль позволяет выявить в корпусе семантически близкие леммы или текстовые прецеденты и вывести их на экран в виде списка с указанием статистической величины их близости и графа. В основу алгоритмов вычисления близости слов положена модель дистрибутивной семантики word2vec, единицами которой являются k-skip-n-граммы некоторого подкорпуса, представленные в виде векторов лингвистических единиц и косинусного расстояния между ними.

Вторая версия словаря существенно доработана. Основные изменения касаются способов подготовки данных для выборки, увеличения количества параметров запроса, а также интерфейса.

В настоящее время пользователю предлагается простой и расширенный вариант интерфейса словаря:

http://manuscripts.ru/mns/mns_evp.vec.simple;
http://manuscripts.ru/mns/mns_evp.vec.main.

Простая запросная форма предоставляет возможность ввода с помощью старославянского или современного алфавита в поле маски анализируемого слова в его начальной форме (по мере ввода подгружаются соответствующие маске леммы) с последующим выводом перечня семантически близких ему слов той же части речи и графа, в котором с помощью ребер показаны связи не только между анализируемым словом и его семантическими аналогами, но и между самими аналогами.

Расширенный вариант запросной формы предлагает пользователю настроить запрос в соответствии с решаемыми задачами. Так, помимо тех параметров, которые были в версии модуля, описанной в [Баранов 2022], (выбор подкорпуса, типа лингвистической единицы, величины косинусного расстояния и нек. др.), в текущем варианте а) реализован алгоритм приведения текстовых прецедентов к одной и только одной лемме, б) предусмотрен анализ гиперлемм и лемм, снабженных частеречной характеристикой, в) предоставлена возможность при вычислении векторов использовать не только абсолютную величину совместной встречаемости слов, но и величину их статистической тесноты (Mutual Information, Point MI, Log-Likelihood, T-score и нек. др.).

Для вывода результатов на экран пользователь может а) задать диапазон расстояний как между векторами анализируемого слова и его семантических эквивалентов, так и между векторами самих эквивалентов, б) использовать частеречный фильтр, в) запустить процедуру извлечения из корпуса биграмм, на основе которых вычислена близость слов, и просмотреть их контексты и др.

Результатом запроса является анализируемое слово (текстовый прецедент) (или список из нескольких слов (текстовых прецедентов), соответствующих маске поиска), его семантические аналоги с указанием статистической близости анализируемому, перечень слов, являющихся лексическим окружением пары *анализируемое – семантический аналог*, и количественные сведения об этом окружении. Вот один из примеров: в подкорпусе славянских Паримейников XII–XIV вв. (объем 195 628 текстовых форм) пятью первыми семантическими эквивалентами слова *господь* являются *аврамъ* 0.739, *заступьникъ* 0.737, *богъ* 0.709, *люди* 0.703, *отьць* 0.678.

**Благодарности**

Исследование выполнено при финансовой поддержке РНФ (проект № 20-18-00206).

**Литература**

Баранов В.А. Дистрибутивный словарь исторического корпуса «Манускрипт»: постановка задачи, материал, методы // Актуальные проблемы филологии и педагогической лингвистики. – 2022. – № 2. – С. 94–106.

Белов В.А. Взаимозаменяемость как критерий синонимии (экспериментальное и корпусной исследование) // Вестник Санкт-Петербургского университета. Серия: Язык и литература. – 2018. – Т. 15. – Вып. 3. – С. 390–410.

Белов С.Д. Обзор методов автоматической обработки текстов на естественном языке / С.Д. Белов, Д.П. Зрелова, П.В. Зрелов, В.В. Кореньков // Системный анализ в науке и образовании. – 2020. – № 3. – С. 8–22. – URL: https://www.elibrary.ru/download/elibrary_44288349_53267953.pdf (дата обращения: 14.10.2022).

Захаров В.М. Дистрибутивно-статистический анализ как инструмент автоматизации формирования семантических полей (на примере поля «империя») // Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018). – Kazan, Russia, November 1, 2018. – URL: http://ceur-ws.org/Vol-2303/paper11.pdf (дата обращения: 14.10.2022).

Золотарев О.В. Современные подходы к обработке многоязычных текстов, основанные на методах дистрибутивной семантики / О.В. Золотарев, М.М. Шарнин, А. Еромасова, Ф.М. Тезадова // Труды Международной научной конференции по физико-технической информатике CPT2018, 2018. С. 43–47. – URL: https://www.elibrary.ru/download/elibrary_35081401_42451913.pdf (дата обращения: 14.10.2022).

Максименко О.И. Квантитативные методы в языковых исследованиях: ретроспективный анализ // Ученые записки национального общества прикладной лингвистики. – 2018. – № 2 (22). – С. 30–39. – URL: https://www.elibrary.ru/download/elibrary_36782916_13097817.pdf (дата обращения: 14.10.2022).

Максименко О.И. Автоматизированный дистрибутивно-статистический анализ как системная обработка текста // Вестник Российского

университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2019. Т. 10. № 1. С. 92–100. – URL: https://www.elibrary.ru/download/elibrary_37217201_33053109.pdf (дата обращения: 14.10.2022).

Шайкевич А.Я. Дистрибутивно-статистический анализ в семантике // Принципы и методы семантических исследований. – М., 1976. – С. 353–378.

Sahlgren M. The Distributional Hypothesis. From context to meaning // Rivista di Linguistica. – 2008. – Vol. 20, № 1. – P. 33–53.

Serelex: поиск семантически связных слов. – URL: http://serelex.org/ru (дата обращения: 14.10.2022).

RusVectōrēs: семантические модели для русского языка. – URL: https://rusvectores.org/ru/about/ (дата обращения: 14.10.2022).

WebVectors: word embeddings online. – URL: http://vectors.nlpl.eu/explore/embeddings/en/ (дата обращения: 14.10.2022).