Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 558–564
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30307

Book review

Review of A.Ya. Shajkevich, V.M. Andryushchenko, N.A. Rebeckaya. 2021. *Distributive-statistical analysis of the language of Russian prose of the 1850–1870s,* vol. 3. Publishing House YaSK, Moscow. ISBN 978-5-907290-61-7

Venera R. BAYRASHEVA

Kazan (Volga Region) Federal University, Kazan, Russia
✉vbayrasheva@gmail.com

For citation:
Bayrasheva, Venera R. 2022. Review of A.Ya. Shajkevich, V.M. Andryushchenko, N.A. Rebeckaya. 2021. *Distributive-statistical analysis of the language of Russian prose of the 1850–1870s,* vol. 3. Publishing House YaSK, Moscow. *Russian Journal of Linguistics* 26 (2). 558–564. https://doi.org/10.22363/2687-0088-30307

Рецензия

Рецензия на книгу: А.Я. Шайкевич, В.М. Андрющенко, Н.А. Ребецкая. 2021. *Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.* Т. 3. М.: Издательский Дом ЯСК. ISBN 978-5-907290-61-7

В.Р. БАЙРАШЕВА

Казанский (Приволжский) федеральный университет, Казань, Россия
✉vbayrasheva@gmail.com

Для цитирования:
Байрашева В.Р. Рецензия на книгу А.Я. Шайкевич, В.М. Андрющенко, Н.А. Ребецкая. 2021. *Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.* Т. 3. М.: Издательский Дом ЯСК. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 558–564. https://doi.org/10.22363/2687-0088-30307

Третий том завершает публикацию фундаментального исследования «Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.», выполненного в Институте русского языка РАН коллективом под руководством авторов издания. Статистические методы систематически применяются в

лингвистических исследованиях с середины прошлого века, и их роль продолжает возрастать. Это объясняется, прежде всего, возможностью быстрой автоматической обработки огромных массивов текстов, появившейся благодаря необычайному прогрессу вычислительной техники, в том числе созданию Интернета. Лингвисты практически сразу после появления компьютеров осознали их возможности и приступили к созданию электронных корпусов текстов. В настоящее время корпусы размером 1 млрд слов и более уже не редкость. Созданный недавно Генеральный интернет-корпус русского языка (ENA, May 22, 2022)[1] содержит 20 млрд слов.

Для анализа корпуса текстов основным инструментом является дистрибутивно-статистический анализ (ДСА), которому авторы дают следующее широкое определение: «Дистрибутивно-статистический анализ есть сумма формальных алгоритмических процедур, направленных на описание языка и опирающихся только на распределение (дистрибуцию) заданных элементов в тексте». Цель всего исследования двоякая – описать и развить ДСА, а также изложить полученные с его помощью конкретные научные результаты, представляющие интерес для русистики. Исследование проводится на материале специально собранного корпуса русской прозы объемом 15 млн слов.

Основным объектом представленной рецензии является вышедший в 2021 г. последний том трехтомного труда, однако методология исследований, важные технические аспекты описаны в первых двух томах, поэтому по необходимости будет затронуто и содержание предыдущих томов. Авторы начинают исследование с противопоставления дескриптивной и теоретической лингвистики. В середине прошлого века проявилось разочарование в дескриптивной методике, явно выраженное в высказывании Н. Хомского: «было бы абсурдным пытаться построить грамматику, которая непосредственно описывала бы наблюдаемое лингвистическое поведение» (Хомский 1962). Однако, несмотря на всю продуктивность развитых Н. Хомским и его последователями абстрактных моделей грамматики, дескриптивный подход нельзя считать отвергнутым. Более того, в последнее время дескриптивный подход получил мощный импульс благодаря развитию компьютерных технологий. Одним из инструментов дескриптивного метода является ДСА, представленный в этой книге.

Авторы развивают интервальный подход в рамках ДСА, состоящий в том, что текст определенным образом разбивается на сегменты, а распределение языковых элементов анализируется в пределах сегментов. В первом томе такими сегментами являются слова (рассматривается распределение букв), во втором томе – биграммы, в третьем томе – отрезки текста длиной 40 слов. На каждом уровне анализа авторы обнаружили интересные закономерности.

[1] http://www.webcorpora.ru/

Первый том посвящен трем вопросам: описанию истории ДСА, результатам ДСА на микроинтервалах (словах) и частотному словарю языка русской прозы 1850–1870-х гг.

История применения статистических методов в лингвистике возводится к работам Т. Менденхолл конца XIX в. Приведенный в первом томе подробный анализ эволюции ДСА представляет несомненный интерес для истории лингвистических учений. Обсуждаются как внутренние мотивы эволюции ДСА, так и внешнее влияние. Основное внимание уделено подходам, развитым в отечественной лингвистике. Подробно изложен статистико-комбинаторный метод Н.Д. Андреева, работы В.В. Шеворошкина и Б.В. Сухотина.

Уже в первом томе авторы вводят математический аппарат, используемый в последующих томах. Для оценки статистической значимости отклонений от математического ожидания события используется следующая формула: $S = (f - m - 1)/ \sqrt{m}$, где f – наблюдаемая частота данного события, m — математическое ожидание этого события, подсчитанное на основе какой-то нулевой гипотезы (пуассоновского распределения). В ДСА событием является совместная встречаемость двух языковых элементов в пределах рассматриваемого интервала. Чем величина S больше, тем совместная встречаемость больше отклоняется от случайной, т.е. можно говорить об их сущностной связи. Экспериментально авторы установили, что при $S > 4$ можно говорить о неслучайности совместной встречаемости пары языковых элементов. Применяя ДСА на уровне слов, авторы выделили 43 агрегированных парадигмы с соответствующими наборами суффиксов. Также в первом томе полностью приведен частотный словарь русской прозы 1850–1870-х гг., содержащий 51 тыс. слов.

Во втором томе завершается описание сочетаемости элементов на уровне слов и приводится полный анализ на уровне биграмм. При анализе на уровне биграмм и бóльших интервалов величина m для пары слов a, b из вышеприведенной формулы принимается равной $m_{ab} = (F_a \times F_b) / N$, где F – частота слова в корпусе, N – объем корпуса (в словоупотреблениях). В используемом для исследования корпусе встретилось почти 6 млн биграмм, изучаются 235 тыс. наиболее частотных.

Авторы показывают, как ДСА биграмм приводит к выявлению элементов грамматики, таких как число, падеж, род, некоторых дистрибутивных классов – предлогов, адвербов, компаративов и др., а также указания на возможность некоторой коррекции классов. Например, формы на -о (*хорошо*) целиком вошли в парадигму адъективов (*хороший*).

Отметим, что авторы не ограничиваются только русским языком и только одним применением ДСА – интервальным анализом корпусов. Рассматривая хорошо известный Брауновский корпус английского языка, они рассчитывают по формулам взаимную близость между 15 его подкорпусами и обнаруживают два кластера подкорпусов – деловой (включающий также научные тексты) и подкорпус художественной литературы.

По степени детальности и полноты описания морфологии русского языка данная монография вполне может быть поставлена в один ряд с Грамматическим словарем русского языка А.А. Зализняка. Причем важно, что описание вытекает из формального квантитативного анализа корпуса текстов.

Третий том посвящен ДСА на средних интервалах и, пожалуй, является наиболее интересным. Прежде чем излагать полученные результаты, авторы предпринимают серьезный методологический анализ. Подробно описываются проблемы, возникшие при лемматизации, и выбранные способы их решения. Как и во многих задачах обработки текстов, проблему представляют омонимы. Хотя потенциально разрешение омонимии возможно на основе дистрибуции, оно представляет большие трудности, и в данной работе использован экспертный семантический анализ.

Принципиально важной проблемой стал выбор длины интервала, сделать который можно было только эмпирически. Авторы рассматривают три длины интервала: в 40 слов, в 200 слов и в 1000 слов, а затем проводят содержательный анализ результатов каждого из экспериментов. Лучшие результаты были получены при выборе интервала длиной 40, при котором совместная встречаемость слов в пределах интервалов отражает семантические (авторы используют термин ‘текстуальные’) связи между словами. Для больших интервалов (1000 слов) обнаруживаются сюжетные связи анализируемых произведений. Например, на материале произведений И.С. Тургенева выявляются связи *резать – лягушка – нигилист*, не характерные для русского языка в целом. Таким образом, ДСА на больших интервалах может быть полезным для литературоведческого анализа. Это, однако, выходит за рамки рассматриваемой работы.

В электронной версии монографии приведены все пары слов, для которых $S > 3$. Разумеется, на выбор величины S влияет размер обрабатываемого корпуса. Выбор $S > 3$ релевантен для корпуса размера около 15 млн слов. При этом для конкретных групп слов при выявлении связей в монографии рассматриваются и другие пороговые значения. Интересны следующие приводимые авторами иллюстрации. В множестве слов {*год весна лето осень зима*} наибольшая сила связи ($S = 76$) обнаруживается у пары *лето – зима*, что, вероятно, отражает антонимический характер семантических отношений между ними. А далее идут пары соседних времен года со значительно меньшей силой связи. Например, для пары *осень – зима* $S = 51$. Что же касается пары *весна – осень,* то для нее $S = 15$, т.е. связь менее сильная. Этот простой пример демонстрирует возможности ДСА; результат, который не может быть получен иными методами. Авторы также рассматривают ряд других примеров: месяцы, дни недели, числительные и проч.

Однако более интересны результаты применения ДСА ко всем словам используемого корпуса. В монографии приводится перечень из 3000 пар слов с аномально высоким значением S. Это перечень трактуется как шаг на пути к созданию частотного словаря фразеологических единиц. Первыми

в алфавитном порядке указаны следующие пары: *ааронов жезл, авторское самолюбие, ад вымощенный*. Существующие фразеологические словари русского языка создаются вручную, и представляется интересным сопоставить их со словарем, созданным компьютерным способом.

Для многозначных слов (омонимия и полисемия) указывается, что полностью проанализировать и компактно представить все их устойчивые связи представляется затруднительным, и приводится лишь некоторое число примеров. Например, выявлены следующие устойчивые текстуальные связи со словом *медведица*: 1) *медвежонок* (45[2]), *медведь* (18), *объятие* (9); 2) *Орион* (101), *Южный крест* (79), *звезда* (19), *небо* (6). Формальный ДСА без учета семантики не всегда дает хорошие результаты. Так у слова *мир* выявлено лишь одно значение – ‘peace’. Другие значения не показывают значимых ($S > 3$) сочетаний в рассматриваемом корпусе текстов. В других корпусах ситуация может быть иной, но этот пример демонстрирует определенные ограничения представленного в монографии варианта ДСА. Разумеется, чем больше корпус текстов, к которому применяется ДСА, тем полнее и точнее будут результаты.

Множество слов и их текстуальные связи представляют собой сеть, содержащую около 26 тыс. слов и 500 000 связей между ними. Между двумя словами (лемматизированными) из корпуса устанавливается связь, если для них $S > 3$. Причем с этой связью ассоциируется число S. Сеть напоминает популярные в последнее время карты слов (см., например, ENA, May 22, 2022)[3]. Однако карты слов обычно строятся на основе ассоциаций, описанная же в монографии сеть текстуальных связей построена по корпусу ДС методом. Было бы интересно провести специальное исследование по их сопоставлению.

В сети выделяются кластеры слов – группы слов, внутригрупповые связи которых преобладают над внешними. Описан метод выделения кластеров. Сначала ручным способом выделяется небольшая группа слов (центр кластера) с большими коэффициентами связи S. Затем она пополняется словами, имеющими хотя бы две связи со словами центра. Затем еще раз пополняется словами, имеющими хотя бы одну связь со словами ранее построенного множества. В монографии применяется подход case study, анализируется состав нескольких построенных таким способом кластеров. Один из примеров – кластер «Дуэль». Оказалось, что в него входят глаголы, преимущественно совершенного вида, что объясняется краткосрочностью дуэлей. Представляется интересным довести формальный подход до конца, применив для выделения кластеров современные алгоритмы кластеризации. Однако с учетом огромного объема сети и ее неоднородностью применение алгоритмов кластеризации нетривиально и требует отдельного исследования.

[2] В скобках указано значение *S*.

[3] https://wordassociations.net/ru/

Авторы обращают особое внимание на проблемы, возникающие при неоднородности текстов. Например, во входящих в исследуемый корпус произведениях А.И. Герцена аномально часто по сравнению с корпусом в целом встречаются слова *революция* и *республика*. Возможность существования подобных неоднородностей следует учитывать как на стадии формирования или выбора корпуса, так на стадии компьютерной обработки данных и на стадии анализа результатов.

Сопоставляя приведенный в монографии вариант ДСА с другими аналогичными методами, следует обратить внимание на следующее. По большому счету, этот метод сближается с латентным семантическим анализом (Landauer 1998), который также направлен на выявление контекстно-зависимых значений слов при помощи статистической обработки больших корпусов. Принципиальная разница состоит в том, что в методе ДСА, используемом в монографии, рассчитывается только близость слов, а в латентном семантическом анализе одновременно рассчитывается еще и близость фрагментов корпуса, при этом близость фрагментов является основной. Дистрибутивная близость слов является и предметом векторной семантики, классическая реализация которой представлена в (Mikolov 2013). С появлением в последние годы нейронных сетей глубокого обучения именно векторная семантика стала основным инструментом дистрибутивного анализа. При этом следует учесть сложность реализации таких задач, требующих серьезно квалификации в компьютерных технологиях. С этой точки зрения ДСА, предложенный в рассматриваемой монографии, выгодно отличается простотой реализации, что делает его доступным большому числу лингвистов.

В заключение отметим, что рецензируемая монография является прекрасным ответом на известное высказывание Н.С. Трубецкого «Язык лежит вне меры и числа» (Трубецкой 1960: 15). Несмотря на, казалось бы, сухой предмет монографии – статистику – книга написана живым языком, не характерным для современных научных публикаций, и, несомненно, привлечет внимание всех заинтересованных исследователей.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgements

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Трубецкой Н.С. Основы фонологии. М., 1960. [Trubetskoi, Nikolai S. 1960. Osnovy fonologii. Moscow. (In Russ.)].

Хомский Н. Синтаксические структуры // Новое в лингвистике. 1962. Вып. 2. М. [Chomsky, Noam. 1962. Syntactic structures. *Novoe v Lingvistike*. 412–528. Moscow. (In Russ.)].

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* 25. 259–284.

Book review history:
Received: 20 December 2021
Accepted: 25 February 2022

Bionote:
Venera R. BAYRASHEVA holds a doctoral degree in physics and mathematics. She is Associate Professor of the Theoretical Cybernetics Department at Kazan (Volga Region) Federal University. Her research interests include theoretical cybernetics and computer linguistics.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: vbayrasheva@gmail.com
ORCID: 0000-0002-1728-034X

Сведения об авторе:
Венера Рустамовна БАЙРАШЕВА – кандидат физико-математических наук, доцент кафедры теоретической кибернетики Казанского федерального университета. Сфера научных интересов: теоретическая кибернетика, компьютерная лингвистика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, 18
e-mail: vbayrasheva@gmail.com
ORCID: 0000-0002-1728-034X