RUDN Journal of Language Studies, Semiotics and Semantics 2019 Vol. 10 No 1 92—100
Вестник РУДН. Серия: ТЕОРИЯ ЯЗЫКА. СЕМИОТИКА. СЕМАНТИКА http://journals.rudn.ru/semiotics-semantics

ВНУТРЕННЯЯ И ВНЕШНЯЯ ДЕТЕРМИНАНТЫ ТЕКСТА

УДК: 81:005:004
DOI: 10.22363/2313-2299-2019-10-1-92-100

АВТОМАТИЗИРОВАННЫЙ ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ КАК СИСТЕМНАЯ ОБРАБОТКА ТЕКСТА

О.И. Максименко

Московский государственный областной университет
ул. Радио, 10а, Москва, Россия, 105005

В статье рассматривается автоматизированный дистрибутивно-статистический анализ с позиций системной лингвистики, акцент делается на различии в дистрибутивном и статистическом подходах для анализа терминологической лексики, детализируются понятия единицы анализа и интервала анализа, возможности применения метода для автоматической обработки текста с целью формирования как общих, так и информационно-поисковых тезаурусов, атрибуции жанров текстов в тематических корпусах.

Ключевые слова: дистрибутивно-статистический анализ, гиперлексема, корпус, текст, система

> Язык относится к классу таких объектов, в которых системность, внутренняя организованность, взаимосогласованность всех статистических и динамических характеристик представлена в своих высших проявлениях.
>
> Г.П. Мельников, 1978 г.

ВВЕДЕНИЕ

Автоматическая обработка текста из комплекса экспериментальных задач, которую она представляла собой полвека назад, превратилась в повседневную реальность. За прошедшие годы было предложено немало решений, определяемых тем, для чего, собственно, нужна была такая обработка. Это автоматическое индексирование текстов для обеспечения систем информационного поиска, автоматическое аннотирование и реферирование текстов для хранения информации в сжатом виде, автоматическое создание частотных словарей и конкордансов для последующего контент-анализа, создание машинных словарей для систем машинного перевода и информационно-поисковых систем, проведение автоматического выявления тональности текста, работа с Big Data и мн. др.

К числу методов, используемых как для теоретических целей (исследование сочетаемости слов, частотности, распределения), так и сугубо прагматических (распознавание жанров текстов в информационном потоке), относится дистрибутивно-статистический анализ.

Исследования, объектом которых служит естественный язык, всегда были чрезвычайно непросты. В связи с этим за последние сто лет лингвистической науки исследовательские парадигмы менялись не раз. Пытались исследовать язык, так сказать, «по частям», разделяя его на уровни — фонетический, морфологический, синтаксический, семантический — воспринимая естественный язык как множество, т.е. «многое, мыслимое как целое». Однако сколь бы ни был привлекательным такой взгляд на язык, он упрощает то, чем на самом деле естественный язык является. Нам намного ближе системный взгляд на язык, одним из приверженцев которого был Г.П. Мельников.

Казалось бы, что квантитативной лингвистике, в рамках которой существует немало числовых методов обработки естественного языка, проще работать с отдельными множествами, чем с плохо формализуемыми семантическими понятиями, тем не менее, существуют методы, которые вполне можно отнести к системным, работающими с языком как системой, т.е. «целым, мыслимым как многое».

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ

Одним из методов, рассматривающих естественный язык как систему, считается метод дистрибутивно-статистического анализа (ДСА). Этот метод относился в первую очередь к лингвостатистике и существует в рамках этой дисциплины с последней трети XX в., разделяясь на два варианта: дистрибутивный и статистический. Он предполагает использование математического аппарата теории вероятностей и логико-лингвистического аппарата, основанного на понятии дистрибуции.

Многое, сделанное в квантитативной лингвистике этим методом, суммировал А.Я. Шайкевич, который предложил свою точку зрения, рассматривая ДСА как сумму формальных алгоритмических процедур, направленных на описание языка и опирающихся только на распределения (дистрибуции) заданных элементов в тексте. Под заданными элементами могут пониматься буквы (и другие графические символы), цепочки букв между пробелами (слова), цепочки слов между более крупными пробелами (предложения), т.е. любые объекты в тексте, непосредственно доступные нашему восприятию. Сам анализ при этом постоянно использует количественную информацию о встречаемости элементов в тексте [1. С. 355].

Дистрибуция, о которой начали говорить еще в структурной лингвистике, предполагает учет всех контекстов, в которых встречается та или иная языковая единица. Таким образом, дистрибутивный вариант ДСА опирается на сходстве распределений слов, т.е. всех возможных контекстов, в которых эти слова могут встречаться. Статистический вариант ДСА, в свою очередь, основан на числовых характеристиках совместной встречаемости элементов текста в контекстах определенной длины. Естественно предположить, что слова, связанные по смыслу, должны часто встречаться в тексте недалеко друг от друга, и наоборот, слова, часто

встречающиеся вместе в осмысленном тексте, связаны друг с другом по смыслу. Реальная совместная встречаемость двух лексических единиц (двух ключевых слов) и ожидаемая (теоретическая) встречаемость этих же единиц дает представление о силе связи этих элементов текста. Оценка расхождений между значениями, полученными эмпирически и путем статистических расчетов, и является основой метода ДСА.

Для любого анализа принципиально важен выбор единицы, это же касается и метода ДСА. Интервал текста также имеет значение. Необходимо отметить, что выбор единицы анализа определяется рядом прагматических факторов, в частности, задачами проводимого анализа. В разное время исследователи выбирали единицы, в большей мере соответствующие целям их работы. Это могли быть фонемы, словоформы, квазиосновы, гиперлексемы, ключевые слова, устойчивые словосочетания и пр.

Вопрос об определении интервала анализа оставался открытым в течение долгого времени. В работах А.Я. Шайкевича было конкретизировано понятие интервала текста и приведены в общих чертах задачи, которые можно решить на том или ином интервале. Под интервалом текста понимался один из множества равных отрезков, на которые разбивается текст. При этом предполагается, что для каждого конкретного исследования точно фиксируется длина интервала, которая служит базой для всех математических расчетов. Переход от одного интервала к другому означает получение семантической информации нового качества. Таким образом, понятие интервала текста оказывается некоторым аналогом понятия «уровень» в содержательной лингвистике [2]. Понятие уровня организации чего-либо характерно именно для системного подхода.

Поскольку текст на естественном языке является сложной системой, имеющей семиотическую природу, в которой обмен информацией проходит на семантическом уровне (по Ю.А. Шрейдеру) [3], то оптимальным интервалом для фиксации «контекстуальных связей» слов, т.е. связей, обеспечивающих семантическую связность текста, считается интервал в одно-два предложения справа и слева от анализируемого слова.

Анализ связи между ключевыми словами основывается, напротив, на предположении о независимости появления двух слов в тексте (т.е. об отсутствии связи между ними), что позволяет вычислить математическое ожидание их совместной встречаемости в интервале текста, исходя из теоремы об умножении вероятностей. Отклонение реальной совместной встречаемости двух слов от их теоретической встречаемости противоречит гипотезе о независимости появления этих слов в тексте, т.е. между ними существует связь (синтагматическая). Помимо синтагматических связей существуют и парадигматические связи, играющие существенную роль при построении информационно-поисковых тезаурусов. Сила парадигматической связи возрастает с увеличением числа синтагматических связей и их силы. С увеличением интервала текста качественное различие между синтагматическими и парадигматическими связями постепенно стирается.

Важно отметить, что ДСА, являясь формальной методикой, дает в отношении выявляемых смысловых связей вероятностные результаты, что вызывает ряд

вопросов и возражений по поводу его применения в информационной практике. В то же время простота метода, которая является его основным достоинством, искупает все недостатки, ему приписываемые. Простота делает ДСА универсально применимым к любой предметной области, в частности, при обработке текстов в режиме Big Data.

Автоматизированный вариант ДСА используется для организации лексики тезаурусов (установление синтагматических и особенно парадигматических отношений, которые наиболее важны для поиска); при индексировании запросов (расширение поискового предписания за счет терминов, ассоциированных с лексическими единицами запроса, что повышает эффективность поиска); при классификации поискового массива (объединение документов, содержащих ассоциированные по смыслу термины) [2].

Наиболее очевидным и актуальным является использование ДСА при построении автоматических тезаурусов. При создании тезаурусов необходимо учитывать различие между тезаурусом как инструментом поиска и тезаурусом как способом описания какой-либо понятийной области, включая терминологию разных дисциплин науки и техники, хотя и тот, и другой в самом общем виде можно определить как множество смыслоразличительных элементов (слов, словосочетаний и т.д.) некоторого языка с заданными смысловыми отношениями.

В информационно-поисковых системах (ИПС) с развитой грамматикой или с элементами грамматики синтагматические отношения слов используются для получения списков устойчивых сочетаний ключевых слов, которые необходимы при решении вопросов о границах словосочетаний, типах используемых текстуальных связей и др.

От текстуальных связей можно перейти к системным (парадигматическим) или связям второго порядка (например, отношениям иерархии «род—вид», «часть—целое», «причина—следствие», отношениям синонимии и квазисинонимии, деривации). Такая структура тезауруса основана на использовании сложной системы естественного языка, и получающийся информационно-поисковый тезаурус (ИПТ) является более «гибким», что сказывается на результатах поиска.

Тезаурусы входят в лингвистическое обеспечение при индексировании запросов, в частности, при автоматическом индексировании. В настоящее время системы автоматического индексирования способны распознавать слова и словосочетания, включая словоизменительные и словообразовательные варианты слов, определять характеристики информативности (веса) лексических единиц в аспекте выражения предметного содержания текстов, устанавливать связи между терминами, «переводить» ключевые слова на другие естественные или информационно-поисковые языки.

Сведение лексем в гиперлексемы как единицы анализа — это то допустимое при формальном анализе материала обращение к семантике, о котором говорили А.Я. Шайкевич и Э.И. Королёв. Без этого при существующих вариантах дистрибутивно-статистического не обойтись. Неверно созданный список гиперлексем может дать значительный информационный шум при анализе текстов с помощью ДСА.

Автоматизированный вариант дистрибутивного анализа основывается на изучении сочетаемостных характеристик таких единиц текста, как гиперлексема. Для этого проводится компьютерная сортировка выбранных гиперлексем — маркировка тех словоупотреблений текста, которые соответствуют какой-либо из гиперлексем. За этим следует получение программным способом списка терминосочетаний из двух, трех или четырех гиперлексем, выбор из них тех, которые удовлетворяют требованиям по частоте встречаемости, сортировка по алфавиту и удаление повторений. Параллельно с созданием списка терминосочетаний предусмотрено составление матрицы частоты совместной встречаемости гиперлексем, которая используется при дальнейшей обработке. Следующий этап алгоритма заключается в анализе попарной дистрибутивной связанности терминологических однословных единиц и в вычислении меры их смыслового сходства, «расстояния» в каком-либо избранном интервале текста. Эта степень смыслового сходства контекстуальных окружений становится мерой смысловой связанности, ассоциативности двух слов.

На основе этой информации ручным (или компьютерным) способом строится граф связности гиперлексем для дальнейшего исследования текстового массива. Граф — это один из наиболее удачных вариантов получения визуальной информации, т.к. в нем указывается не только прямой или опосредованный характер связей, но и уровни и сила связей. В результате выявляются градации силы и последовательности связей терминологических гиперлексем, формируется смысловая, семантическая карта данного текста, которая может выступать в качестве иерархически организованного поискового образа документа. При обработке достаточно представительного массива можно получить той или иной степени полноты словарь ассоциативных отношений гипертерминов данной предметной области, который может использоваться, например, для построения тезауруса этой предметной области.

Способ построения графа связности заключается в последовательном переводе матричной информации в графическую форму. Прежде всего, задается порог силы связи между членами пар, ниже которого связи в совокупный граф не включаются. Построение иерархического кластера начинается с обозначения пары гиперлексем (или гиперлексемных терминосочетаний), находящихся на самом близком «расстоянии» в тексте, т.е. имеющих самую сильную связь. Далее обозначаются с указанием силы связи все гиперлексемы, связанные с первым членом центральной пары, затем — со вторым. Так организуется второй уровень иерархии зависимостей. Выделяются межуровневые и внутриуровневые связи, формируются кластеры, определяется степень изолированности узлов. Полученная таким образом наглядная информация служит основой для анализа как внутренней структуры текста, так и взаимных связей между текстами в случае смешанных выборок.

Для выявления статистической связи между единицами текста используется алгоритм, основанный на оценке частоты одновременного появления двух единиц (гиперлексем) в одном интервале текста.

Результаты, также как при дистрибутивном анализе, представляются сначала в матричном виде, а затем автоматически переводятся в графическую форму. Для

построения графа статистической связности при установленном уровне значимости выбирается пара гиперлексем или гиперлексемных терминосочетаний с максимальной силой связи, и от каждого члена пары проводятся связи (ребра графа) к коррелирующей с ним гиперлексемой второго уровня, далее — третьего и т.д. Так же как при построении дистрибутивного графа формируются терминологические гиперлексемные кластеры, каждый из которых может быть интерпретирован.

Эксперимент для оценки алгоритма ДСА [2] заключался в анализе путем дистрибутивной и статистической методик (двух вариантов ДСА) текстов на естественном языке и получении матриц дистрибутивных расстояний между односоставными терминами и терминосочетаниями (двух-, трех- и четырехсоставными терминами), встречающимися в текстах, а также матриц статистической зависимости между ними. Путем анализа полученных данных, а именно путем построения графов дистрибутивных связей с различными порогами и графов статистической связности с разными уровнями значимости, выяснялись степени покрытия текстов терминологической лексикой при варьировании дистрибутивных порогов и уровней значимости, что играет существенную роль при построении поискового образа документа и других прикладных целей.

При дальнейшем лингвистическом анализе семантических сетей (графов), полученных при разных дистрибутивных порогах и статистических уровнях значимости, выяснялись специфические отношения между отдельными членами семантической сети и целыми кластерами и проводилась лингвистическая интерпретация полученных корреляций.

Сравнивать и оценивать варианты ДСА можно лишь в том случае, когда существует основание для сравнения, т.е. тождество основных существенных параметров при варьировании значения одного и не более сопоставительного параметра. Но лингвистический материал редко предоставляет такую возможность. Зачастую выделить один параметр для сопоставления, чтобы получить «чистый» результат, трудно, так же трудно, как уравнять все остальные параметры системы, чтобы фиксировать их изменение в ответ на изменение одного избранного параметра.

Как уже говорилось, возможные параметры сравнения — полнота и точность лексических единиц графов. Чтобы провести сравнение на этом основании, можно избрать для анализа два графа (дистрибутивный и статистический) при определенном значении дистрибутивного порога и статистического уровня значимости (доверительной вероятности) связей. Сопоставляемым параметром будет разное для двух методик соотношение полноты и точности выявленных результатов. Фиксация, уравнивание в этом случае проводится по экстенсивному параметру — общему объему элементов, связываемых при подбираемом значении дистрибутивного иди статистического критерия. Варьирование величин дистрибутивных порогов и статистических уровней значимости предоставляет разнообразную гамму возможностей как в выделении ключевых для данного текста терминов, так и для получения общей информации о покрытии текста терминологической лексикой и специфических отношениях между терминами. Методика автомати-

зированного ДСА позволяет отражать даже слабые связи между заданными словами (терминами), что дает возможность учесть все многообразие связей отдельного термина с другими элементами системы и всей системой в целом.

Важным является то, насколько дистрибутивный и статистический методы чувствительны к типу смысловых связей, т.е. их способность по-разному выделять синтагматические и парадигматические связи из обрабатываемых текстовых массивов. Как при одном, так и при другом методе в совокупных графах представлены оба типа связи. Иначе говоря, применение каждой из двух методик, по всей видимости, не может служить средством преимущественного извлечения того или иного типа связи. Следует оговориться, что этот вывод справедлив для того интервала, на котором проводилось исследование — одно предложение. При изменении интервала, например, при переходе на контактный, двухсловный интервал, результаты могут быть иными. Возможна более сильная реакция статистической методики на синтагматические связи. Расширение интервала за пределы предложения должно, по-видимому, привести к преимущественной чувствительности на парадигматические связи слов со стороны дистрибутивной методики.

Большая часть связей, полученных в эксперименте при оптимальных и жестких дистрибутивных порогах и статистических уровнях значимости, была интерпретирована и могла бы включаться в понятийные классификации. Вместе с этим необходимо отметить, что ДСА имеет принципиальные ограничения, связанные с вероятностным характером той информации, которая в нем используется. Собственно семантические характеристики слова в полном объеме и в точном соответствии не могут быть реализованы в сочетаемостных характеристиках.

Качественные различия между дистрибутивным и статистическим методами заключаются в особенностях их функционирования на текстах с разными характеристиками. Применение дистрибутивной и статистической методик на разных в жанровом отношении текстах позволяет автоматически их различать. Жанровая атрибуция текстов проводится по различиям в типах графов. Соотношение терминологических гиперлексем, не связанных по дистрибутивному критерию и связанных по статистическому и наоборот для одного и того же текста, характеризует его структуру: смешанный или однородный, написанный одним автором или группой авторов, что можно активно использовать при атрибуции текста или при процедуре лингвистической экспертизы текста.

ЗАКЛЮЧЕНИЕ

Подводя итоги эксперимента, стоит отметить, что, прежде всего, исследования подобного рода важны для оценки существующих формальных методов исследования языка с целью их последующего использования в автоматизированных словарно-терминологических службах информационно-поисковых систем. Обработка корпусов научно-технических текстов методом дистрибутивно-статистического анализа дает возможность поддерживать в актуальном состоянии отраслевые тезаурусы, классификаторы, исследовать системные отношения и изменения ядерной и периферийной лексики, проводить автоматическое индексирование документов определенной предметной области, создавать отраслевые словари, различать

жанры в лингвистических корпусах, т.е. описанный комплексный метод входит в состав современных методов автоматизированной обработки текста и является вполне продуктивным аппаратом анализа текстовой информации.

Г.П. Мельников в свое время писал, что в такой системе, как язык, непрерывно происходит «стихийный „естественный отбор“ языковых средств, т.е. процесс самоорганизации, на каждом ярусе вырабатывается оптимальные динамическое решение между субстантными возможностями и структурой, наиболее подходящей для этой субстанции, причем структура и субстанция каждого яруса влияет на все остальные ярусы» [4 С. 97; 5]. Это наблюдение поддерживает понятие системной лингвистики, а метод дистрибутивно-статистического анализа, являющийся с одной стороны сугубо квантитативным, лингвостатистическим, с другой стороны, благодаря обращению к дистрибуции, выявляющей смысловую близость языковых единиц, можно считать системным и вполне применимым в рамках заявленного Г.П. Мельниковым направления — системной лингвистики.

Дата поступления: 08.10.2018
Дата приема в печать: 16.12.2018

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Шайкевич А.Я.* Дистрибутивно-статистический анализ текстов: дисс. ... докт. филол. наук. Москва, 1979.
2. *Максименко О.И.* Формализованная лингвистика. Москва: изд-во МГОУ, 2013.
3. *Шрейдер Ю.А.*, Шаров А.А. Системы и модели. Москва: Радио и связь, 1982.
4. *Мельников Г.П.* Азбука математической логики. Москва: Знание, 1967.
5. *Мельников Г.П.* Бодуэновское понимание системности языка // Рождественский Ю.В. (ред.) Языковая практика и теория языка, вып. 2. Москва: Изд-во Московского университета, 1978. С. 32—51.

УДК: 81:005:004
DOI: 10.22363/2313-2299-2019-10-1-92-100

AUTOMATIC DISTRIBUTIVE-STATISTIC ANALYSIS AS SYSTEM TEXT PROCESSING

Olga I. Maksimenko

Moscow Regional State University,
10A, Radio str., Moscow, Russia, 105005,

Abstract. The article is devoted to the to the description of the automatic distributive-statistic analysis, distinction in distributive and statistical approaches for the analysis of terminological lexicon, an opportunity of application of the method for automatic text processing, formation of thesauruses and possibility to use it for the identification text genre in the corpse of technical texts.

Key words: distributive-statistical analysis, hyper lexeme, corpse, text, system

REFERENCES

1. Shaikevich, A.J. (1979). Distributive and statistical text analysis [dissertation]. Moscow. (In Russ.).
2. Maksimenko, O.I. (2013). Formalized linguistics. Moscow: MGOU. (In Russ.).
3. Shrejder, J.A. & Sharov, A.A. (1982). Systems and Models Moscow: Radio and communication. (In Russ.).
4. Mel'nikov, G.P. (1967). The ABC of the mathematical logic. Moscow: Znaniye. (In Russ.).
5. Mel'nikov, G.P. (1978). Boduenovskoye's understanding of language systemacy In J.V. Rozhdestvenskij (Ed.) *Language practice and the theory of language*, *2*. Moscow: Publishing house of the Moscow University. pp. 32—51. (In Russ.).

Для цитирования:
Максименко О.И. Автоматизированный дистрибутивно-статистический анализ как системная обработка текста // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика, 2019. Т. 10. no 1. С. 92—100. doi: 10.22363/2313-2299-2019-10-1-92-100.

For citation:
Maksimenko O.I. (2019). Automatic distributive-statistic analysis as system text processing. *RUDN Journal of Language Studies, Semiotics and Semantics*, 10 (1), 92—100. doi: 10.22363/2313-2299-2019-10-1-92-100.

Сведения об авторе:
Максименко Ольга Ивановна, доктор филологических наук, профессор, профессор кафедры теоретической и прикладной лингвистики Института лингвистики и межкультурной коммуникации Московского государственного областного университета; *научные интересы:* квантитативная лингвистика, лингвосемиотика, теория интерпретации, теория номинации; *e-mail:* maxbel7@yandex.ru

Information about the author:
Olga I. Maksimenko, Doctor of Philology, Professor, Professor, Department of theoretical and applied linguistics, Institute of linguistics and intercultural communication, Moscow state regional University; research interests: quantitative linguistics, linguosemiotics, theory of interpretation, theory of nomination; *e-mail:* maxbel7@yandex.ru