STUDIA PHILOLOGICA

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ РУССКОГО ЯЗЫКА им. В. В. ВИНОГРАДОВА

А. Я. ШАЙКЕВИЧ, В. М. АНДРЮЩЕНКО, Н. А. РЕБЕЦКАЯ

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ ЯЗЫКА РУССКОЙ ПРОЗЫ 1850—1870-х гг.

Том 1

ЯЗЫКИ СЛАВЯНСКОЙ КУЛЬТУРЫ
МОСКВА 2013

УДК 811.161.1
ББК 81.2 Рус
Ш 17

Издание подготовлено при финансовой поддержке
Российского гуманитарного научного фонда (РГНФ)
проект № 13-04-16009

Р е ц е н з е н т ы :
д. ф. н. *А. Ф. Журавлев*, д. ф. н. *С. А. Крылов*

Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.

Ш 17 Дистрибутивно-статистический анализ языка русской прозы 1850—1870-х гг. Т. 1. М.: Языки славянской культуры, 2013. — 504 с. — (Studia philologica.)

ISBN 978-5-9551-0668-7

Цель дистрибутивно-статистического анализа состоит в открытии структуры языка на основе большого корпуса текстов. В настоящей трехтомной монографии этот формальный метод в полной мере прилагается к текстам русской прозы 1850—1870 гг. (около 15 млн словоупотреблений); а частично (в виде иллюстраций) к текстам на других языках.

Первый том включает три части:

Очерк развития метода;

Открытие регулярной морфологии в рамках графического слова;

Частотный словарь языка русской прозы 1850—1870 гг.

Первые две части адресованы лингвистам, особенно тем, кто интересуется лингвостатистикой. Частотный словарь будет интересен филологам-русистам. В существенно расширенном виде он представлен на компакт-диске.

ББК 81.2 Рус

ISBN 978-5-9551-0668-7

Предисловие

Лингвостатистика стала кристаллизоваться в автономную лингвистическую дисциплину в середине XX века. Первый частотный словарь (*Kaeding F. W.* Häufigkeitswörterbuch der deutschen Sprache, Berlin, 1898) базировался на текстах общим объемом 11 миллионов словоупотреблений. В дальнейшем частотные словари создавались как подспорье в педагогической практике преподавания языков. Важными вехами оказались издания: *Thorndike E. L., Lorge I.* The Teacher's Word Book of 30,000 Words. NY, 1944 (созданный на текстах общим объемом более 20 миллионов словоупотреблений); *West M.* A General Service List of English with Semantic Frequencies. NY, 1953 (до сих пор остается единственным словарем, где количественно представлена полисемия) и вершина достижений в этом педагогическом направлении — *Carroll J. B. e. a.* Word Frequency Book, Boston, 1971 (более 5 миллионов словоупотреблений, с максимальной дифференциацией по школьным предметам).

В рамках частотных словарей алфавитный словарь легко преобразуется в ранговый словарь, где слова расположены в порядке убывания частоты. Дж. Ципф (*Zipf G. K.* The Psycho-biology of Language. Boston, 1935) был первым, кто стал всерьез изучать количественные соотношения единиц в ранговом словаре. Вскоре к этим проблемам присоединились математики Г. Хердан, Дж. Кэррол и многие другие, искавшие в ранговых словарях новые интересные виды статистических распределений. В глазах лингвистов и математиков лингвостатистика стала ассоциироваться исключительно с законом Ципфа и сопутствующими проблемами. Эта ассоциация оказалась фатальной для нарождавшейся дисциплины, где лингвистическое содержание и методическая эффективность стремились к нулю[1].

Между тем, вдалеке от лингвостатистики в 1940-х гг. в США сформировалась дескриптивная лингвистика. Под сильным влиянием бихевиоризма ученые этого направления (Б. Блок, З. Хэррис, Ч. Хоккет, Ю. Нида и др.) следующим образом представляли себе задачу лингвиста, изучающего какой-либо язык: дан большой корпус текстов (зафиксированных акустически или графически), изучая распределение (дистрибуцию) элементов по отношению друг к другу, постараться описать структуру языка, как можно реже задавая информанту вопрос: Что означает такая-то цепочка звуков (цепочка букв)? Колоссальная трудоемкость метода, психологически допустимая лишь для энтузиаста, изучающего экзотический язык, должна была привести в отчаяние и обречь на безработицу лингвистов, изучающих родной язык, где все, казалось бы, ясно. Понятно поэтому, что лингвисты вздохнули с облегчением, когда молодой иконоборец Н. Хомский в 1961 г. категорически заявил «было бы абсурдным пытаться построить грамматику, которая непосредственно описывала бы наблюдаемое лингвистическое поведение» [Хомский 1962].

Но отвергнуть задачу не значит решить ее.

Нам по-прежнему кажется интересной цель, поставленная дескриптивистами 60 лет тому назад — по корпусу текстов (в традиционной графической форме) описать структуру языка. Главное методическое приращение — введение статистики в этот процесс открытия структуры языка.

В рамках нижеследующего исследования **дистрибутивно-статистическим анализом (ДСА)** будем называть набор статистических процедур, выявляющих дистрибуцию элементов корпуса текстов (букв, цепочек букв, графических слов, иероглифов) и не использующих смысл как исходное данное. Можно и мягче определить задачу, требуя на каждом этапе исследования четкой фиксации исходного смысла.

ДСА, как он сложился за 50 лет, в качестве центрального понятия использует понятие **интервала текста**. На том или ином этапе исследования текст членится на фрагменты равной длины, что позволяет количественно сравнивать реальные совместные появления элементов (или появления элементов в тех или иных позициях в тексте) с математическим ожиданием тех же событий в предположении независимости элементов (и позиций). Эмпирически показано, что интервалы разной длины дают разную лингвистическую информацию. ДСА не зависит от изучаемого языка, хотя до сих пор еще нет примеров его использования для языков с иероглифической письменностью.

Для того чтобы получить статистически значимые результаты, необходимы большие собрания текстов. К началу 1970-х гг. создатели «Trésor de la langue française» имели громадный корпус литературных текстов, нашедший статистическое отражение в Dictionnaire des fréquences, P., 1971 — уникальном частотном словаре (более 70 мил-

[1] Вклад математиков был очень важен в теории выборочных методов в языкознании, определение репрезентативности результатов, в таких узких проблемах, как установление авторства анонимных текстов. К сожалению, разнообразие форм существования языка таково, что в лингвистике очень трудно понять, что называть генеральной популяцией, в какой мере в текстах наблюдаются какие-то фундаментальные статистические распределения (вроде нормального распределения), служащие основой для статистических методов в естественных науках.

лионов словоупотреблений, более 70 тысяч разных лемм, различение прозы и поэзии, четыре периода от 1789 г. до 1950 г.).

Для дальнейшего развития ДСА существенную роль сыграл подготовленный электронным образом многотомный труд М. Спевака (*Spevack M*. A Complete and Systematic Concordance to the Works of Shakespeare, Hildesheim, 1968—1970). Этот конкорданс явился грандиозной статистической базой для решения многих задач ДСА, но объем корпуса текстов Шекспира (около 900 тысяч словоупотреблений) все еще был недостаточным.

Прогресс в компьютерной технике привел к появлению электронных корпусов текстов. От первого корпуса Университета Брауна (миллион словоупотреблений), ср.: *Kučera H., Francis W. N.* Computational Analysis of Present-day American English, Providence, 1967, до стомиллионных корпусов чешского и русского языков — таков прогресс в корпусной лингвистике за сорок лет, ср.: *Čermak F.* M. Křen, Frekvenčni slovník češtiny. Praha, 2004; *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка. М., 2009.

Появление корпусной лингвистики открывает перед исследователями множество новых направлений исследования. От построения общей теории языка лингвисты все чаще будут возвращаться к эмпирическим исследованиям. Об этом, в частности, говорит важная книга Дж. Сэмпсона (*Sampson G.* Empirical Linguistics. L.; NY, 2001). Существование больших общедоступных корпусов текстов чрезвычайно облегчило решение многих частных задач, ДСА может способствовать распространению более широких подходов.

Но для дальнейшей разработки ДСА весьма важно иметь *свой* корпус текстов. Такой корпус текстов русской прозы в середине 1990-х гг. начал разрабатываться в бывшем Отделе машинного фонда русского языка Института русского языка РАН, к настоящему времени он достиг объема в 15 миллионов словоупотреблений. Вся эта работа была бы немыслима без многолетней поддержки Российского гуманитарного научного фонда (гранты 96-04-06264, 01-04-00247а, 04-04-00060а).

Авторы благодарят своих коллег Елену Николаевну Ловлю (сканирование и корректура текстов) и Елену Николаевну Морозову (поддержка сайта отдела, подготовка макета издания). Авторы благодарны также Татьяне Евгеньевне Реутт за сканирование многих текстов, Григорию Самойловичу Цейтину и Егору Аношкину за некоторые эффективные программы сортировки. Труд авторов был распределен следующим образом: Н. А. Ребецкая подготовила программы дистрибутивно-статистического анализа; В. М. Андрющенко подготовил электронную версию словаря на компакт-диске; вся остальная работа проделана А. Я. Шайкевичем.

Настоящий труд задуман как трехтомное издание. Каждый том решает две задачи — продвижение методов дистрибутивно-статистического анализа и публикацию результатов, интересных для русистики. В соответствии с этими двумя задачами первый том содержит три части:

Часть 1. Эволюция дистрибутивно-статистического анализа текстов;
Часть 2. Дистрибутивно-статистический анализ в микроинтервале (Статистическое открытие регулярной морфологии);
Часть 3. Частотный словарь языка русской прозы 1850—1870-х гг.

Во втором томе предполагается представить

Часть 4. Минимальный интервал в дистрибутивно-статистическом анализе;
Часть 5. Хронологическая, жанровая и авторская дифференциация языка русской прозы 1850–1870-х гг.

Третий том должен быть посвящен текстуальным связям лексических единиц в прозе 1850—1870-х гг.

Каждый том должен быть опубликован в двух версиях — бумажной и электронной.

Часть 1

Эволюция дистрибутивно-статистического анализа текстов

1.1. Исторические предшественники

1.1.1. Таксономические проблемы в филологии и задачи ДСАТ

В середине XX в. в трудах сторонников дескриптивной лингвистики очень ясно была поставлена задача — алгоритмическим образом описать язык по данным текста, не обращаясь к смыслу. При таком подходе естественной основой всей процедуры становятся анализ и обработка данных, вытекающих из комбинаторики элементов текста (дистрибутивный анализ).

Однако эти замыслы не могли быть практически осуществлены в то время. Во-первых, электронная вычислительная техника тогда только зарождалась, а вручную обрабатывать огромные массивы текстов — чрезвычайно трудно. Во-вторых, мало кто из лингвистов осознавал тогда необходимость использования статистики для решения этой задачи. Вероятностный подход лишь нащупывался в общетеоретических рассуждениях и не проник еще в конкретные методы анализа, в царство жесткого детерминизма. Наконец, большинство дескриптивистов ставило перед собой ближайшие задачи описания конкретных языков в короткие сроки, а потому не преодолевало принципиальных трудностей, а обходило их. В результате за четверть века существования дескриптивной лингвистики не появилось на свет ни одного последовательно алгоритмического дистрибутивного описания какого-либо языка. Дистрибутивный анализ не превратился в алгоритм. Во многих случаях показывалась принципиальная возможность формально-дистрибутивного решения отдельных проблем, но исчерпывающая регистрация фактов проводилась не на формальной, а на содержательной основе, т. е. с использованием знания семантики языка.

К середине 1950-х гг. уже ясно проявляется разочарование в дистрибутивной методике, а в настоящее время большинство лингвистов считают эти методы пройденным этапом в истории языкознания. Характерно замечание Н. Хомского: «предпринимаются попытки сформулировать методы анализа, которые исследователь реально может использовать, если у него есть время, чтобы построить грамматику языка, исходя непосредственно из сырых данных. По-моему, весьма сомнительно, чтобы этой цели можно было достигнуть сколько-нибудь интересным путем, и я подозреваю, что всякая попытка достичь ее должна завести в лабиринт все более и более подробных и сложных аналитических процедур, которые, однако, не дают ответа на многие важные вопросы, касающиеся природы лингвистической структуры» [Хомский 1962: 459]. Слова Н. Хомского падали на благодарную почву. «За последние сорок лет и особенно с 1957 г. необычайно усилился интерес лингвистов к абстрактным теориям и математическим моделям, можно спорить о том, в какой степени эти теории и модели помогли понять функционирование языка и отточить методы решения практических проблем. Но многие лингвисты теперь считают себя учеными (scientists) чистой воды и часто отмахиваются от всего, что пахнет техникой или практическими приложениями» [Sparck Jones, Kay. 1973: 5].

Но отмахнуться от проблемы не значит закрыть проблему. Можно полагать, что отход от дистрибутивных методов вызван не их принципиальной бесплодностью, а теми временными обстоятельствами, которые были перечислены выше. Развитие лингвистики не отменяет задачи формального описания языка по тексту, наоборот, эта задача становится все более важной особенно в связи с развитием вычислительной техники, лингвистической статистики и прикладной лингвистики.

Решение таксономических проблем котируется не слишком высоко в современной лингвистике. Между тем, таксономические проблемы существуют в лингвистике всегда, независимо от того, какое именно направление преобладает в данный момент.

С этой точки зрения, лингвистам полезно обратиться к опыту биологии.

И там в настоящее время есть более актуальные проблемы (например, молекулярная биология, генетика, экология). Тем не менее, проблемы таксономии сохраняют свою важность в биологии. Характерно, что в своей «Философии биологии» [Рьюз 1977] М. Рьюз отводит две главы проблемам таксономии. Именно в биологии впервые родилась попытка найти объективные количественные методы для определения таксонов. На рубеже 1950—1960-х гг. в биологии сформировалось направление количественной таксономии, чьи задачи сформулированы в книге Р. Сокала и П. Снита «Принципы количественной таксономии» [Sokal, Sneath 1963][1].

В биологической систематике давно известна разница между логической классификацией и естественной классификацией. В системе классификации, восходящей к Аристотелю, главная задача — открытие и определение сущности таксономической группы. Эта сущность проявляется в диагностирующих признаках, каждый из которых обязателен для любого члена группы. Однако натуралисты, имея в руках некоторые естественные критерии разделения групп (вроде репродуктивного барьера), давно обнаружили, что некоторые несомненные естественные группы

[1] Заслуга первого опыта такого рода принадлежит Е. Н. Смирнову [Smirnoff 1924].

не подходят под такое понимание классификации. Так родилось представление о монотетических и политетических группах.

«Основная идея монотетической группы заключается в том, что она формируется на основе жестких последовательных логических делений, так что обладание уникальным набором признаков необходимо и достаточно для членства в группе, определенной таким образом... Политетическая классификация группирует вместе организмы, обладающие наибольшим числом общих признаков, но ни один из признаков не является необходимым и достаточным для включения организма в группу...

Класс обычно определяется по отношению к признакам, одновременно необходимым и достаточным (по определению) для членства в классе. Возможно, однако, определить группу К в терминах набора G признаков f_1, f_2 ... f_n по-другому. Предположим, мы имеем собрание организмов (мы еще не называем их классом), таких, что:

1) каждый обладает большим (но не указанным точно) числом признаков в G;
2) каждый f в G принадлежит большому числу организмов, и
3) ни один f из G не принадлежит всем организмам собрания.

Такие условия задают полностью политетический класс» [Sokal, Sneath 1963: 13—14].

Как правило, различны и пути создания классификации. «Классификация сверху» неизбежно приводит к монотетическим таксонам, «классификация снизу» часто приводит к политетическим таксонам.

> Логики давно уже поняли, что центральная идея, лежащая в основе «естественных» группировок, состоит в практической ценности метода, который группирует объекты таким образом, что члены группы обладают многими общими признаками. Действительно, мы полагает, что неуловимое свойство естественности есть просто степень осуществления этого принципа [Ibid.: 18].

В филологических науках таксономические проблемы имеют не меньшее значение. Более того, можно полагать, что политетические группы встречаются в них чаще, чем в биологии, границы между группами чаще размыты. В этих условиях количественная таксономия в филологии не только имеет право на существование, но и может оказаться чрезвычайно полезной для дальнейшей эволюции лингвистики и литературоведения.

В наибольшей степени этот новый подход может оказаться полезным в тех частях лингвистики, которые отличаются нежесткой структурой, т. е. за пределами фонологии и большей части грамматики. Но и «жесткие» участки структуры языка легко интегрируются в рамках количественной таксономии как частный случай.

Выбор именно формального подхода к анализу языка (и особенно семантики языка) выглядит парадоксальным лишь с первого взгляда. Мотивы такого предпочтения становятся ясными из следующих соображений.

Для современной лингвистики в целом и для семасиологии, в частности, характерно чрезвычайное разнообразие подходов, как в выборе предмета исследования, так и в выборе метода. Тем не менее, можно отметить некоторые явно преобладающие тенденции. Во-первых, усилия лингвистов, в основном, направлены на разработку способов представления семантики.

Обычно предполагается, что семантика языковых единиц задана исследователю, владеющему изучаемым языком, и дело лингвиста — довести свое интуитивное владение смыслом до такой степени расчлененности и эксплицитности, которая соответствовала бы взыскательности самого исследователя и его коллег. В максимальной степени такая требовательность предполагает возможность прямом передачи результатов какому-либо автомату (ЭВМ) для дальнейшего использования в лингвистическом анализе. Именно с точки зрения возможностей автомата обычно говорится о формальном характере семантического описания.

Значительно реже исследование направлено на получение семантических результатов, на открытие чего-то такого в семантике, что было неизвестно лингвисту до начала исследования.

Во-вторых, работа семасиологов, как правило, ограничена либо анализом одного слова, например, анализом полисемии, состава сем (все это можно назвать «микросемантикой»), либо небольшими группами слов — словообразовательными гнездами, синонимическими рядами, семантическими полями («локальная семантика»). Очень редко предпринимаются попытки описания лексико-семантической системы в целом или крупных ее фрагментов («макросемантика»). Правда, в практике лексикографии известны опыты создания тезаурусов или идеологических словарей для отдельных отраслей знания (специальных языков), так и для естественного языка вообще[1]. Однако жесткие логические схемы построения подобных словарей производят впечатление искусственных моделей, весьма далеких от предполагаемых «естественных» систем языка. Необходимость поисков этой естественной системы начинает осознаваться лингвистами.

[1] См. обзоры в [Морковкин 1970; Караулов 1976].