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 K 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].

> Поскольку слово никогда не существует изолированно, очень важно поместить его в то семантическое множество и те множества, к которым оно принадлежит... Еще более важно, чтобы эти множества выделялись не на основе классификации, навязанной языку извне и исходящей из философских, идеологических, научных и технических соображении; необходимо, чтобы они основывались на «естественном» использовании языка до всякого вмешательства философии, науки и техники [Imbs 1970: 471].

Наконец, большинство семасиологов в настоящее время занято изучением синхронной семантики живых языков, реже синхронный анализ семантики распространяется на языки прошлого, еще реже встречаются работы по диахронической семантике. Заметим при этом, что и в семасиологии еще слабо разработаны принципы сопоставления семантических систем. Если сопоставление отдельных семантических областей в разных языках встречается во многих работах, то сопоставление семантических систем разных этапов истории одного и того же языка наблюдается очень редко[1]. Совсем нет исследований, посвященных семантическим системам, сосуществующим в рамках одного языка.

Указанные три тенденции (1. преобладание процедуры представления, 2. внимание к микросемантике и к локальной семантике, 3. преимущественное изучение живых языков) по-видимому, еще долго (а может быть, и всегда) будут оставаться господствующими в семасиологии. Вместе с тем, лингвистика нуждается в исследованиях другого типа, где господствовали бы противоположные тенденции.

Но есть ли общие черты у противоположных тенденций? Можно ли как-то объединить столь разноплановые стремления, как 1. тенденцию разработки процедуры открытия, 2. большее внимание к макросемантике, 3. обращение к языкам прошлого?

Да, такие общие черты можно заметить у всех трех тенденций. Все они предъявляют повышенные требования к объективности исследования.

Действительно, требование получать новую семантическую информацию приводит лингвиста к дилемме. Либо в начале исследования заданы некоторые четко определенные семантические сведения, а затем с помощью столь же четко определенного метода (лучше — алгоритма) исходные сведения преобразуются в новую семантическую информацию. Либо в начале работы исследователь не обладает никакими семантическими данными, не знает даже языка, на котором написаны изучаемые тексты. Тогда вся семантическая информация будет получена в ходе исследования. Такое исследование с неизбежностью будет формальным, но не в том смысле, о котором говорилось выше, а в другом.

В настоящей работе под словами «формальный анализ» будут пониматься исследования, не использующие смысл как нечто заданное до самого исследования. Такой подход будет противопоставляться «семантическому анализу», «семантической лингвистике».

Как в случае использования ограниченного набора исходных семантических данных, так и в случае строго формального анализа необходимо, чтобы до начала работы уже существовал метод. Именно тогда метод отчужден от самого исследователя, и между субъективизмом исследователя и семантическим материалом воздвигнут непреодолимый барьер.

Существование метода уже есть гарантия объективности. Правда, эта объективность — лишь антитеза субъективности. Ее еще нельзя понимать как синоним реальности, естественности. Для того чтобы понимать объективность в этом смысле, по-видимому, необходимы дополнительные условия. Очевидна желательность, по крайней мере, двух условий. Необходимо, во-первых, чтобы один и тот же метод давал сходные результаты на выборках из одной генеральной совокупности, т. е. работал бы устойчиво: а во-вторых, чтобы сходные результаты получались при работе внешне очень различных методов. Только в этом последнем случае у нас появляется уверенность, что результаты исследования зависят не от прихотей метода, но соответствуют истинному положению дел[2].

Мысль о важности метода и вытекающей отсюда объективности может быть распространена и на две другие тенденции. Если наши субъективные, интуитивно ощущаемые знания семантики слов превратились для нас в психологическую реальность благодаря беспрерывным подкреплениям со стороны четко выделимых кусков действительности, то этого нельзя сказать о семантической системе в целом, о больших семантических классах слов, о семантических категориях, о постоянно воссоздаваемых семантических шаблонах, т. е. о том, что можно назвать макросемантикой. Соответствующие семантические сущности обычно не осознаются, относятся к «криптотипам» [Whorf 1956: 69], а потому исследователь снова оказывается перед необходимостью семантического открытия. Очевидно, что то же верно и в отношении языков прошлого, где предварительные знания исследователя сильно ограничены или даже равны нулю. В значительной мере это справедливо относительно стилистических систем в пределах

---

[1] Работа Й. Трира [Trier 1931] так и осталась блестящим неповторенным образцом.

[2] М. Рьюз называет этот принцип принципом Максвелла [Рьюз 1977: 188].

живого языка, где степень эксплицитной осознанности невелика, особенно у народов без сильной традиции кодификации стилистических норм.

Высказанные соображения — важный довод в пользу разработки формальных объективных методов изучения языка.

Дистрибутивно-статистический анализ есть сумма формальных алгоритмических процедур, направленных на описание языка и опирающихся только на распределение (дистрибуцию) 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 (высказывания), короче — любые объекты в тексте, непосредственно доступные нашему восприятию. Самый анализ при этом носит не жестко детерминистский, а статистический характер, постоянно использует количественную информацию о встречаемости элементов в тексте.

Что касается математического обоснования избранных методов, автор (как и многие другие исследователи в этой области) считает их обсуждение преждевременным.

> В моей практике обычно выяснялось, что выбор той или иной техники анализа по чисто математическим основаниям оправдан только тогда, когда существует полная ясность в задачах и возможностях этой техники; в противном случае наиболее разумно выбирать технику после оценки практичности и качества получаемых результатов. Если нет основополагающей теории, поясняющей, что означает, что слово встретилось в тексте один раз, дважды, трижды или n раз, было бы наивным применять сложные, теоретически обоснованные статистические формулы. С другой стороны, интуиция может привести к выбору статистической формулы совершенно ad hoc без всякой опоры на математическую теорию [Doyle 1963].

Главным критерием оценки алгоритма является результат его работы. Можно построить очень остроумный, очень простой (или, наоборот, очень сложный) алгоритм, но обсуждать его бесполезно, пока мы не знаем, какие результаты он дает. Это общая черта алгоритмов, предназначенных для индуктивной методики. Вот почему проверка ДСА должна была вестись на достаточно обширных текстах. Для первого опробования метода в действии нельзя было обращаться к совершенно неизвестному языку, ведь в этом случае нельзя определить, насколько осмысленны полученные результаты. Поэтому в настоящем введении и в Части 3 читатель обнаружит примеры анализа текстов на хорошо известных европейских языках.

1.1.2. Внешние влияния

Помимо дескриптивной лингвистики существенное влияние на ДСА оказали и некоторые другие течения.

Одним из таких источников является традиционная стилистическая статистика, где первой по времени возникла задача статистической характеристики индивидуального стиля автора. Еще в 1887 г. Т. Менденхолл [Mendenhall 1887] обнаружил сходство распределения длин слов у Марло и Шекспира и их отличие от Бен-Джонсона, Бэкона, Бомонта и Флетчера. Длина слов и длина предложений использовались для различения авторских стилей и в дальнейшем [Yule 1939; Elderton 1949; Fucks 1955], но среди лингвистов все более крепло убеждение, что эти параметры обладают слабой различительной силой.

Постепенно в круг анализа все более втягивались новые и новые лингвистические явления, при этом уже не только авторские, но и жанровые различия становятся объектом изучения.

Традиционная задача стилостатистики — определение авторства текста. Можно полагать, что эта задача выполнима только в случае полного жанрового (и частичного тематического) совпадения. Попытки решить эту задачу с помощью какого-нибудь одного явления обычно терпят неудачу. Подобные неудачи заставили Дж. Юла [Yule 1944] искать более сложные математические отношения, скрытые в тексте, как надежный источник решения спорного авторства (в конкретном случае «De Imitatione Christi»).

Так родилась проблема соотношения числа слов в тексте и в словаре (token-type ratio), столь популярная в лингвостатистике (см. [Herdan 1956; Фрумкина 1964]). Однако это направление быстро превратилось в совершенно замкнутую область, где господствуют чисто математические проблемы, слабо связанные с проблемой идентификации автора.

В 1960-х гг. попытки установления авторства соединяются с анализом сразу многих наблюдаемых явлений. На лексические единицы опирались авторы наиболее известных исследований: Ф. Мостеллер и Д. Уоллес, исследовавшие авторство спорных статей («Федералист») с двумя возможными кандидатами — Гамильтоном и Медисоном [Mosteller, Wallace 1963], и А. Эллегорд, изучавший авторство «Писем Юниуса», где было 40 кандидатов, из которых самым вероятным кажется Филипп Френсис [Ellegard 1962].

На русской почве самые интересные результаты были получены М. А. Марусенко и его соавторами [Марусенко 1990; 2001] на первоначальной основе списка из 56 параметров (частот грамматических явлений[1]). Алгоритм Марусенко был реализован в машинных экспериментах по атрибуции статей в «Кине-журнале» (1913—1915), романов «Три страны света» и «Мертвое озеро», романа «Тихий Дон» и пьесы «Ты и Вы» (1827 г.). Было убедительно показано, что в каждой отдельной ситуации атрибуции диагностирующими оказывается лишь часть параметров.

Следует сказать еще об одном направлении в статистической стилистике, имеющем прямое отношение к ДСА. Речь идет о выделении ключевых слов авторов, школ, направлений в литературе и вытекающем отсюда семантическом анализе. Предшественником современных исследователей подобного толка можно считать К. Сперджен [Spurgeon 1935], чья книга произвела большое впечатление на литературоведов. Правда, К. Сперджен использовала статистику в рамках содержательной стилистики, но искомые результаты можно было бы получить и при помощи ДСА. Более формальный характер носило раннее исследование Э. Рикерт [Rickert 1927]. В обоих случаях авторы обходились самыми простыми статистическими показателями, необходимость введения специальных статистических мер была осознана лишь к 1950 гг. [Guiraud 1954; 1959].

Особо следует сказать о работах Дж. Майлз [Miles 1965], в которых предпринята первая попытка статистического изучения эволюции поэтического языка на протяжении нескольких столетий. Ее методика очень несовершенна статистически. Форма представления результатов не позволяет проводить сопоставление и оценку результатов, тем не менее, богатое собрание материала заставляет предполагать, что аналогичное исследование, проведенное по правилам статистики, дало бы очень богатый материал для истории стиля.

Дж. Майлз подсчитывает частые слова по выборкам (по 20 поэтов) для десятилетий, разделенных веком (1540-е, 1640-е, 1740-е, 1840-е, 1940-е). Судя по частым словам, XVIII в. сильнее всего отличается от всех остальных периодов. Максимальные сдвиги приходятся на время после XVII в. и после XVIII в. (в значительной мере возврат к лексике XVII в.). Менее значительна разница между XIX в. и XX в. Нормальный ход поэтической эволюции был как бы прерван XVIII в. (наименее поэтическим, по общему мнению). Наиболее подвижными оказываются прилагательные, значительно устойчивее существительные и особенно глаголы.

В той мере, в какой стилистическая статистика сознательно становится на путь «от формальных показателей к семантике», она сближается с контент-анализом, вторым внешним методическим источником ДСА.

Потребность в этом направлении ощущалась очень давно. Само название родилось в 1948 г. «Контент-анализ — исследовательская техника для объективного, систематического и количественного описания содержания сообщения» [Berelson, Lazarsfeld 1948; Berelson 1952]. С конца 1940-х гг. появляются работы, где контент-анализ применялся для целей социологии и современной истории (образцами могут служить [Lasswell, Leites 1949; Holsti 1972]).

Важная особенность контент-анализа заключается в том, что он «редко интересуется явным содержанием сообщения. Скрытые отношения, идентификация авторов, статистические тенденции в использовании символических форм и т. п. вопросы захватили воображение многих ученых, в то время как сознательные намерения выступают как частный случай исследования» [Krippendorf 1959].

Часто подчеркивалась необходимость сближения контент-анализа с лингвистикой. Интересные мысли по этому поводу высказал А. Рапопорт:

> Структурная лингвистика «жестка», но она далека от того, что составляет «суть» человеческого общения. Литературная критика «мягка» и часто безответственна, но она пытается ухватить некоторые очень важные вещи, о которых пытается говорить люди. По-моему, науки о поведении не всегда будут мучиться дилеммой — чем жертвовать: человеческой важностью ради надежности знания или наоборот. То, что эта дилемма часто возникает, происходит, как я думаю, от того, что все труднее «делать науку» по мере того, как материал все более ускользает от ученого, а также от того, что людей разного темперамента и разных интересов влекут к себе либо та, либо другая линия исследования. Будущее науки может быть обеспечено, если «жесткие» методы будут применяться к «мягким» областям, очень постепенно консолидируя каждый кусок «захваченной территории», а также если «твердоголовые» и «мягкоголовые» будут больше прислушиваться друг к другу. Мне кажется, что контент-анализ очень подходит на роль поля исследования, где такая консолидация может быть достигнута [Rapoport 1959].

Чем более в контент-анализ проникают компьютеры, тем сильнее тяга исследователей к максимальной десемантизации, т. е. к превращению в ДСА. Логическим завершением этой тенденции явилась система программ, создан-

[1] В качестве примера назовем несколько параметров: 01 — число слов в цельном предложении, 02 — число графем в цельном предложении, 17 — число знаменательных слов, 18 — число служебных слов, 24 — число предлогов, 30 — число слов в аккузативе, 38 — число причастных оборотов и т. п.

ная Ф. Стоуном и его сотрудниками и названная им General inquirer (GI) (что-то вроде «всеобщий вопрошатель», см. [Stone et al. 1966]). Эта система позволяет практически автоматизировать все традиционные методы контент-анализа (кроме, может быть, больших матриц совместной встречаемости).

Последний источник ДСА — широкое и мощное направление поисков автоматизации в информатике[1]. В этой области, хорошо вооруженной ЭВМ, обладающей большими массивами текста в машинной записи, достигнуты наибольшие успехи ДСА в изучении больших интервалов текста.

В 1958 г. появилась важная статья одного из зачинателей автоматической информатики Г. Луна «Автоматическое создание рефератов научной литературы» [Luhn 1958], с ней родилась задача автоматического реферирования и индексирования. Подход Г. Луна был еще очень несовершенным, предполагалось в качестве ключевых терминов отбирать слова, часто появляющиеся в данном документе. Неудивительно, что очень скоро стали слышны критические замечания и предложения использовать не абсолютную частоту термина как показатель его значимости, а некоторое превышение его относительной частоты в документе над относительной частотой в некоторой нормативной выборке (см., например, [Edmundson, Wyllys 1961]). Как бы то ни было, интерес к данной проблеме возник, и после недолгого чисто теоретического обсуждения в США и Англии приступили к экспериментам.

Для развития автоматической информатики и ДСА ключевым был 1961 г. В январе 1961 г. редакция журнала Journal of the Association for Computing Machinery получила статью Мейрона «Автоматическое индексирование: экспериментальное исследование» [Maron 1961], в апрельском номере появилась статья Г. Стайлза «Ассоциативный фактор в информационном поиске» [Stiles 1961], в марте в редакцию пришел переработанный вариант статьи Л. Дойла «Карты семантических путей для поиска литературы», в которой приводились графические отражения связей нескольких слов [Doyle 1962]. А уже в ноябре на Межвузовской конференции по применению структурных и статистических методов исследования словарного состава языка демонстрировалась большая семантическая карта, отражающая связи нескольких сот слов. (См. [Шайкевич 1961] и ниже п. 1.2.3.) Тем самым было положено начало нескольким направлениям в автоматической информатике и ДСА.

В статье М. Мейрона ясно сформулирован принцип статистического индексирования документов:

> Настоящий подход к проблеме автоматического индексирования является статистическим. Он основан на довольно прямолинейном представлении, что индивидуальные слова в документе функционируют как ключи (clues), на основе которых можно сделать предсказание о той содержательной рубрике, к которой возможно относится документ. По существу, основной тезис состоит в том, что статистика характера, частоты, положения, порядка и т. п. избранных слов достаточна для удовлетворительных предсказаний о содержании документа, включающего эти слова [Maron 1961].

Первые опыты проводились на собрании рефератов по электронике. Как и в последующих экспериментах подобного рода часть собрания рефератов использовалась как исходная база для отбора ключевых слов (в данном случае — 260 рефератов), а другая часть (145 документов) для контроля результатов. Из общего числа 3263 слов М. Мейрон на глазок отобрал 90 слов (как потом стало ясно — очень мало). В основной группе 80 % документов было опознано правильно, 15 % неправильно, а 5 % не опознано, поскольку они не содержали ключевых слов. Если в документе содержалось хотя бы два ключевых слова, доля правильных решений повышалась (до 91 %), но в контрольной группе в таких условиях она составляла лишь 52 %.

Автоматическая индексация стала важным элементом автоматической информатики (см. обзор [Stevens 1965—70]). Конечно, в 1960—1970-х гг. усилия специалистов были направлены, в основном, на поиски путей автоматического отбора ключевых слов статистическими методами. Особенно важна статья К. Спарк-Джоунз «Взвешивание терминов индексирования» [Sparck Jones, Kay 1973], в которой на основании многочисленных экспериментов, тестирующих альтернативные процедуры индексирования, делается вывод о преимуществах взвешенных ключевых слов, а также важный вывод о решающем значении большого числа ключевых слов, индексирующих документ.

Статья Г. Стайлза открывает целое направление в информатике — направление статистических ассоциативных методов.

Эксперименты охватили поисковые образы 100 тысяч документов Министерства обороны США. Для каждого слова фиксировались его текстуальные связи в порядке уменьшения ассоциативного фактора, такие списки связей были названы «ассоциативными профилями термина».

Г. Стайлз предусмотрел возможность создания профилей второго поколения. В ответ на запрос потребителя машина выдавала все термины, связанные средним ассоциативным фактором с терминами запроса.

---

[1] Слово *информатика* употребляется здесь как эквивалент английского information science.

Увлечение ассоциативными методами в первой половине 1960-х гг. было настолько велико, что на Симпозиуме в Вашингтоне в 1964 г. присутствовало более сотни специалистов [Stevens 1965]. После этого наступает известное разочарование. Как пишет К. Спарк-Джоунз, «многие участники Вашингтонского симпозиума 1964 г., который может считаться кульминационным пунктом периода энтузиазма в связи с новыми подходами к основной проблеме описания документов статистическими методами, уже не заняты в исследованиях в данной области» [Sparck Jones, Kay l973: 12]. Основная причина, по ее мнению — перспектива долгого упорного труда с неясным исходом. Более оптимистически смотрят на период 1965—1967 гг. (и это же можно распространить на последующие десятилетие) авторы самого большого из опубликованных трудов по ассоциативным статистическим методам в информатике [Jones et al. 1968: 1]. Они пишут: «По сравнению с теми днями творческих исследований, размах публикаций и явное продвижение вперед заметно сократились. Некоторым даже стало казаться, что тема эта, в основном, завершена и заброшена. Но внешняя сторона дела обманчива. Деятельность в этой области в последнее время была приурочена больше к консолидации — в нашем случае, к проблемам обработки очень больших массивов документов». Действительно, Джоунз, Кэртис, Джулиано и Шерри сумели обработать массив в 100 000 документов с объемом словаря 18 тыс. слов. Их программа рассчитана на обработку массивов до 1 млн документов и словаря до 32 тыс. слов. Одновременно в машине обрабатывались матрицы размером 3000 × 3000.

Автоматизация в информатике поощряет контакты этой науки с лингвистикой. Об этом свидетельствует книга К. Спарк Джоунз и М. Кея «Лингвистика и информатика» [Sparck Jones, Kay 1973]. Это сближение, в частности, проявляется в попытках разнообразить то окно, в рамках которого изучается совместная встречаемость элементов текста («интервал» в терминах настоящей книги).

Таковы те источники, которыми питался дистрибутивно-статистический анализ. Выделение этого метода как особого направления в лингвистике не означает, конечно, разрыва с течениями, его породившими. Напротив, ДСА в свою очередь начнет способствовать развитию дескриптивной лингвистики, стилистики, литературоведения, социологии, информатики.

1.2. Первые шаги на пути к формальному открытию системы языка по фактам речи

1.2.1. Статистико-комбинаторный метод Н. Д. Андреева

Первопроходцем на пути построения системы языка на основе статистического анализа речи был Николай Дмитриевич Андреев (1920—1997). Его первая краткая публикация относится к 1959 г. [Андреев 1959]. А в «Материалах по математической лингвистике и машинному переводу». Сб. II. Л., 1963 появилась развернутая статья «Алгоритмы статистико-комбинаторного моделирования морфологии, синтаксиса, словообразования и семантики» [Андреев 1963]. Статья эта произвела на меня большое впечатление общей масштабностью замысла, проработкой деталей преодоления потенциальных барьеров на пути к открытию системы языка. Оставалось лишь ждать результатов.

Н. Д. Андрееву удалось в Ленинградском отделении Института языкознания АН СССР создать группу математической лингвистики и дополнительно привлечь к работе научных работников из Тарту, Риги, Еревана, Фрунзе, Иркутска, Синельникова. Коллективная монография «Статистико-комбинаторное моделирование языков» вышла в свет в 1965 г. (далее [СКМ-65]), а в 1967 г. опубликована книга Н. Д. Андреева «Статистико-комбинаторные методы в теоретическом и прикладном языковедении».

Цель своего метода Андреев формулирует следующим образом (с. 6):

> При разработке статистико-комбинаторного метода считалось, что принципиально правильным будет только такой порядок исследования:
>
> 1) вскрыть систему языковых форм, исходя из статистики и комбинаторики, но совершенно не используя никаких значений (ни лексических, ни грамматических) и не обращаясь к критерию грамматической правильности;
>
> 2) лишь после этого, опираясь на смысл текстов, установить значения выявленных форм, а через них — и грамматическую правильность.
>
> Иначе говоря, была принята та точка зрения, что в исследовании языка имеется граница, ниже которой можно формализовать исследование вплоть до полной алгоритмизации; выше этой границы исследовать язык чисто фор-