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

Москва 2013

Содержание

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,K, 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 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» [Sokal, Sneath, 1963, 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.

---

[1] Заслуга первого опыта такого рода принадлежит Е. Н. Смирнову [Smirnoff,1925].

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.

«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 исходящею 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.[2] 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

---

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

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

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.[1]

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 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, 1965].

---

[1] М. Рьюз называет этот принцип принципом Максвелла [Рьюз, 1977, 188].

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 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, 1964], 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,

[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.

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, 1960].

Особо следует сказать о работах Дж. Майлз [Miles,1948,1950,1951], 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, et al., 1949, Rolsti, 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,1969].

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, 1969].

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, созданная 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 1953 г. появилась важная статья одного из зачинателей автоматической информатики Г. Луна «Автоматическое создание рефератов научной литературы» [Luhn, 1953], 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» [Sti1es, 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, 1961].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. ниже п. 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, также Boyle, 1973]. 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, 1973b], 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.

---

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

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.

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,1973,12]. 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 100000 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 формальными методами, без опоры на его прямые (семантика) и опосредственные (грамматика) связи с действительностью, не представляется возможным и не ставится в качестве исследовательской задачи.»

Предлагаемые алгоритмы строятся на одном и том же фундаменте — на понятии коррелятивной функции (КФ), определяемой как «отношение условной вероятности языкового элемента при заданном условии к независимой вероятности того же элемента» (с. 22).

На с. 26 перечисляются четыре принципа, основополагающие для любых алгоритмов в рамках СКМ:

«Первый принцип. Система речи устанавливается с необходимой достоверностью лишь на основании достаточной статистики; ошибка наблюдения должна быть существенно меньше измеряемой величины...

Второй принцип. Выявление группировки языковых элементов всегда начинается с узловой точки, обнаруживаемой по максимуму коррелятивной функции, т. е. по наибольшему отклонению условной вероятности от независимой; языковой элемент, которому соответствует этот максимум, считается информантом рассматриваемого подмножества...

Третий принцип. Формирование группировок языковых элементов производится по достаточному сходству вероятностного распределения их комбинаторики...

Четвертый принцип. Распознавание означаемых, стоящих за выделенными группировками означающих, производится на основе критериев, имеющих эксплицитно формальный характер, но привязанных имплицитно к определенному содержанию... Для этой позиции кроме КФ алгоритм учитывает максимальную вероятность буквы (безотносительно к независимой вероятности). Если бы не это требование, максимальную КФ показала бы буква Ы с КФ > 6.

Алгоритмы, построенные на этих четырех принципах, действуют с непрерывным накоплением информации о системе речи и с постоянным переводом последней в информацию о языке. При этом считается вполне допустимой последующая коррекция ранее полученной информации...»

Работы группы математической лингвистики велись широким фронтом: было обследовано 22 языка с алфавитной письменностью. Для русского и чешского языка объем отобранных текстов был по тем временам довольно большим (соответственно 1000 и 770 тыс. словоупотреблений), выборки по 7 языкам (сербохорватский, болгарский, армянский, английский, немецкий, французский, эстонский) колебались от 50 до 100 тыс. Выборки по остальным 13 языкам (украинскому, латышскому, итальянскому, албанскому, хинди, хауса, суахили, венгерскому, вьетнамскому, китайскому, киргизскому) были явно недостаточны для каких бы то ни было суждений об эффективности алгоритма. Языками-антирекордсменами были индонезийский и бирманский (примерно по 3 тысячи словоупотреблений).

В конце 1960-х гг. группа математической лингвистики распалась, сам Н. Д. Андреев вскоре обратился к другим темам.

Поскольку морфология является логическим началом всех алгоритмов Андреева и именно здесь были получены самые важные результаты, рассмотрим в качестве иллюстрации подхода эту стадию СКМ.

Все начинается с поиска информанта в той или иной позиции в слове (позиция +1 - начальная буква, позиция -1 - последняя, позиция +2 — вторая буква слева, позиция -2 — вторая буква справа и т. д.), буква-информант должна обладать максимальной КФ в данной позиции. В позиции +1 максимальной

КФ (5,7) обладает буква П, в позиции -1 максимальную КФ обнаруживаем у Й (8,0), затем идут Я (6,4) и Х (5,4).

Выбрав Й в качестве первого информанта, обнаружим в позиции -2, предшествующей этому Й, букву О с КФ=5,7[1], первым информантным аффиксом становится -ОЙ. В текстах по радиоэлектронике общим объемом 1 миллион словоупотреблений обнаружено около 1900 разных буквенный цепочек, объявляемых теперь «базами» (основами): *котор-*, *соб-*, *больш-*, *одн-*, *высок-*, *так-*, *обратн-*, *баз-* и т. д. У выделенных баз могут быть и другие продолжения («остатки») помимо -ОЙ: -ого, -ых, -ы, -о, -а, -ые, -ый, -ое, -ая и т. д. После многих шагов алгоритма сформируется полная парадигма первого морфологического типа (прилагательные с основой на твердый согласный) у 27 основ: *котор-*, *полупроводников-*, *постоянн-*, *обратн-*, *электронн-*, *мал-*, *некотор-*, *перв-* и т. п.

Как первый морфологический тип прилагательные выделились и в чешской публицистике, в немецких и французских текстах по радиоэлектронике.

В сербских политических текстах получена полная субстантивная парадигма с пятью основами — *закон-*, *колектив-*, *однос-*, *проблем-*, *результат-*. В болгарских технических текстах получен аналогичный класс существительных с 6 основами — *генератор-*, *електрод-*, *елемент-*, *капацитет-*, *сигнал-*, *трансформатор-*.

В английских текстах по радиоэлектронике первый морфологический тип совпадает с традиционным типом правильных глаголов (39 основ — approximat-, cause-, chang- и т. п.). Тот же тип (но только с двумя основами liv- и lov-) обнаружен на материале романа Диккенса «Тяжелые времена».

Общий стратегический план алгоритма Андреева предполагает, что после получения первого морфологического типа для данного языка, тот же алгоритм применяется к его соседям в тексте (слева или справа). Например, после того как во французском языке был выделен первый морфологический тип с флексиями -0, -e, -es, -s и с 11 основами, у его левых соседей алгоритм выделил второй морфологический тип с минимальной парадигмой -0, -s с 127 основами — преимущественно существительными (atome-, bande-, champ-, gain-, partie- etc.), но и с теми адъективами, которые не имеют различий для мужского и женского рода (вроде possible-, difficile-, classique-, inverse-). Если в английском языке первый морфологический тип соответствовал правильным глаголам, то у левых соседей обнаружился второй морфологический тип с парадигмой -0, -ly, включающий 29 основ (accurate-, fair-, great-, normal- etc.).

У правых соседей первого морфологического типа в русском языке (прилагательные) получен второй морфологический тип с парадигмой *-0, -а, -ам, -ами, -ах, -е, -ов, -ом, -у, -ы*, в полном виде представленный лишь у двух основ — *канал-* и *снаряд-*. После объединения («типизации») первого и второго типа у их правых соседей обнаружен третий морфологический тип с полной парадигмой *-е, -ем, -и, -й, -ю, -я, -ям, -ями, -ях* у двух основ: *значени-*, *явлени-*. Восемь флексий *-ей, -и, -й, -ю, -я, -ям, -ями, -ях* обнаружено и в четвертом морфологическом типе с двумя основами — *станци-* и *функци-*. Наконец, выделен и пятый морфологический тип с парадигмой *-ется, -ются, -я* (явно неполной с обычной грамматической точки зрения) с 5 основами *вычисля-*, *использу-*, *применя-*, *рассматрива-*, *счита-*.

Несколько лет работы группы математической лингвистики принесли довольно скромные результаты. С позиций XXI в. (с его прекрасной компьютерной оснасткой) основной причиной неуспеха кажется отсутствие подходящих технических средств обработки больших текстовых массивов. Ограниченность выборок приводила лишь к небольшому числу набора основ с полной парадигмой, требуемых алгоритмом.

Но были и внутренние причины, предопределившие печальную судьбу этого, говоря современным языком, проекта. СКМ был обнародован Н. Д. Андреевым как вполне готовый алгоритм, почти для каждой операции будущих машинных

[1] Для этой позиции кроме КФ алгоритм учитывает максимальную вероятность буквы (безотносительно к независимой вероятности). Если бы не это требование, максимальную КФ показала бы буква Ы с КФ > 6.

экспериментов был заготовлен соответствующий термин[1]. Лингвисту оставалось вручную следовать предписаниям алгоритма.

В таких условиях исследователь начинал нарушать правила алгоритма. Характерно замечание Е. Е. Корди [Корди, 1965a, с. 270]:

«Для того же, чтобы найти слова первого типа в тексте, при ручном тестировании приходится прибегать к экстраполяции. Мы пользуемся знанием языка и считаем, что все слова имеющие парадигму из 4 окончаний: -es, -e, - нулевой аффикс, -s — входят в первый тип».

Другая возможность — локальное дополнение алгоритма. Ср. замечание того же автора [Корди, 1965b, с. 306]:

«Рабочее множество для получения третьего типа было получено снова другим способом. Работа велась с правым маргинальным подмножеством, но не со всем, а только членимыми словоформами, входящими в него. Кроме того, если справа от типизованного слова находилась нечленимая словоформа, производился поиск дальше вправо от него и в рабочее множество включалась первая членимая словоформа, находящаяся справа от типизованного слова в том же предложении». Напомним, для получения третьего типа во французском языке используется обобщенный тип существительных и прилагательных, ясно, что справа от третьего типа снова окажется прилагательное, по какой-то причине не попавшее во второй тип. Действительно, полная парадигма -ux, -l, -lement при трех основах: éga-, généra-, expérimenta-.

Недовольство общим жестким универсальным алгоритмом возрастало, когда исследователи при работе с конкретным языком получали странные или мизерные результаты (например, для немецкого или армянского).

Поучительны общие соображения М. Р. Мелкумяна [Мелкумян, 1965, с. 135):

«Предписания алгоритма статистико-комбинаторного моделирования принадлежат разным уровням. Эту иерархию полезно выявить с самого начала, в описании подпрограммы выделения морфологического типа.

Введем предварительное условное разбиение предписаний подпрограммы на две группы: 1) информантный поиск вдоль схемы подпрограммы (а далее — вдоль всей схемы алгоритма); 2) иные предписания: “ступенчатый спуск”, “проверка на унитарную членимость”...

Первой группе предпосылается фундаментальное утверждение: “... на любом ярусе языка отклонение от случайного распределения — достаточно крупное в абсолютной мере, по медиальному разбиению, и экстремальное в относительной мере, по коррелятивной функции — принимается за некоторую информацию об узловой точке структуры языка” [Андреев, 1963, с. 11]. В данном утверждении сформулировано наиболее общее глубинное свойство некоторого рода структур, а в частности языковых; вообще говоря, предписания первой группы (включенные в незакрепленную многоярусную систему дополнительных — в смысле Бора — классификаций) являются универсальными и достаточны для нашей цели.

Предписания второй группы не следуют естественным образом из приведенного утверждения. Но именно их включение делает наиболее эффективным приложение алгоритма статистико-комбинаторного моделирования к исследованию конкретных текстов.

Становясь с включением предписаний второй группы более эффективным в ряде конкретных приложений, алгоритм статистико-комбинаторного моделирования соответственно утрачивает свою изначальную универсальность...

Очевидный путь повышения эффективности алгоритма статистико-комбинаторного моделирования состоит в установлении ряда новых предписаний второй группы и их дальнейшей экспериментальной проверке. При выборе более или менее окончательного рабочего варианта модели имеет смысл ввести в качестве критерия универсальности систему полевых характеристик».

В той же книге М. Р. Мелкумян [Мелкумян, 1965, с. 256] и создатель метода обмениваются язвительными любезностями:

---

[1] В СКМ-67 предметный указатель (с. 384–398) содержит несколько сот новых терминов. Примерами могут служить: децентрация аффикса... закон ломаного убывания... обратная нормализация... остатки первого вида... рабочий маргинал, радикал (исходный, первообразный, разложимый), разбиение медиальное... разрядный диполь, разрядный шлейф... редукция списка основ... фратрия, фратрийный анализ. Термин «информантная лестница» (очень частый в СКМ-65) почему-то отсутствует в этом перечне.

«Новаторская сущность статистико-комбинаторного моделирования должна состоять, по-видимому, в постоянном учете “обратных связей”. Следовательно, внутренняя структура алгоритма совершенствуется таким образом, чтобы в нем вернее учитывалась специфика каждого моделируемого языка. Это предполагает ветвление алгоритмических предписаний, выработку системы правил условных переходов. С другой стороны, совершенствование алгоритма становится явлением закономерным, непрерывным.

В короткой, но бурной истории математической лингвистики чрезвычайно поучительна судьба жестко фиксированных универсальных моделй (алгоритмов). Сталкиваясь с конкретными языками, модель обнаруживает свою недостаточность, а затем становится инструментом “структурно-типологических исследований”. Между прочим, чтобы сделать возможным такое применение универсальной модели, достаточно только объявить об этом».

На это Н. Д. Андреев в качестве редактора книги дает следующее примечание.

«Если алгоритм не будет “жестко фиксирован”, то он утратит право именоваться алгоритмом — по общеизвестному определению последнего. Следует отличать имманентную фиксированность **алгоритма** в течение эксперимента от свободы **исследователя** варьировать алгоритм в промежутках между экспериментами. Следует также не смешивать **алгоритм** обработки данных речи с получаемой при помощи этого алгоритма **моделью** языковых подсистем: между ними тождества не больше, чем между жерновами и пирогом».

Реальный алгоритм СКМ остановился перед синтаксисом. В область синтаксиса обсуждается как нечто, покоящееся на традиционных частях речи и членах предложения. Характерен пример статьи А. И. Варшавской [Варшавская, 1965] «Некоторые данные статистико-комбинаторного анализа английского предложения на материале подъязыка радиоэлектроники». Автор пишет:

«В статье приводятся данные по частотности отдельных функциональных классов (ФК) и связей их между собой в предложении, полученные статистико-комбинаторным (аппроксимационным) анализом...[1]

«Исходным принципом выделения ФК является соотношение морфологических классов слов (МК) и синтаксических функций (СФ), выполняемых словами в предложении. МК выделяются в достаточной степени произвольно, на основе различных имеющихся классификаций слов по частям речи. Выделенный набор МК не совпадает ни с одной классификацией, так как в некоторых случаях слова, обычно относимые к одной части речи, распределяются нами по нескольким МК. Так, например, отдельными морфологическими классами являются инфинитив, причастие, герундий. Для анализа было выделено 25 морфологических классов [СКМ-65, с.387]. При определении СФ мы исходим из деления предложения на члены предложения в обычном понимании этого термина [СКМ-65, с.389].»

Смена ориентиров никак не обсуждается в главе III СКМ-67 «Синтаксис в статистико-комбинаторном моделировании». На с. 69 говорится: «Опираясь на понятие ВДП [вероятностных дифференциальных признаков], можно создать принципиально новые методы исследования языка, важных как в теоретическом, так и в прикладном аспекте. Одним из таких методов является аппроксимационный анализ.

Сущность этого исследовательского метода состоит в том, что строится некоторая гипотеза о разбиении заданной совокупности языковых объектов на непересекающиеся классы, затем статистически изучается комбинаторика этих классов, на основании полученных данных устанавливаются ВДП и с их помощью проверяется справедливость исходной гипотезы».

По ходу изучения комбинаторики ФК и СФ накоплено множество количественных данных, отраженных в приложениях 13–19 (с. 317–355). Сопоставимость языков здесь затруднена из-за недостаточного объема выборок и разнобоя в классификации групп лингвистических объектов. Интерпретация этих приложений затруднительна для читателя, не погрузившегося в детали подхода. В качестве иллюстрации приведем (с изменением формата) фрагмент таблицы «Независимые вероятности частей речи», ограничимся пятью языками и только сферой публицистики.

---

[1] Характерная смена термина. В области синтаксиса вместо СКМ часто используется сочетание аппроксимационный анализ, причем молчаливо изменяются задачи самого метода.

	Русск.	Англ.	Франц.	Итал.	Исп.
Артикль	-	0,081	0,157	0,093	0,148
Существительные	0,335	0,266	0,245	0,264	0,334
Местоимение	0,083		0,020	0,018	0,026
Прилагательные	0,137	0,058	0,090	0,123	0,065
Глагол	0,109	0,147	0,132	0,140	0,083

В главе IV СКМ-67 «Лексика и семантика в статистико-комбинаторном моделировании» больше всего внимания уделено распределительному словарю, т. е. таблицам частот отдельных слов или групп слов (или числа выборок, в которых представлены слова) в их распределении по подъязыкам. В Таблице 198 в качестве подъязыка первого ранга (публицистика) фигурируют выборки из словаря Э. А. Штейнфельдт [Штейнфельдт, 1963] (художественная литература, радиопередачи для молодежи, публицистические статьи), второй подъязык первого ранга (научно-технические тексты) включает четыре подъязыка второго ранга: 1) гуманитарные науки, 2) химия, физика, математика, электроника, военное дело, 3) биология, медицина, сельское хозяйство и 4) геология, металлургия, транспорт, строительство. Сквозь призму подъязыков термин семантическое поле понимается по-новому и крайне необычно:

«Определим семантическое поле как подмножество слов, имеющих положительную специфичность для одного и того же подъязыка. Из этого определения следует, что семантические поля могут быть стольких рангов, сколько их существует в иерархии подъязыков, все специфичные группы слов, приведенные в 4.2.1 принадлежат к семантическим полям соответственно первого ранга, общенаучно-технического (*процесс*, *исследовать*, *нейтральный*, *приблизительно*), второго ранга, точных наук (*кривизна*, *интегрировать*, *радиальный*), третьего ранга, физического (*квант*, *теплоемкость*, *турбулентный*); сюда можно добавить четвертый ранг, принадлежащий атомной физике (*протон*, *изотоп*, *антипараллельный*), и пятый, относящийся к теории элементарных частиц (*кварк*, *странность*, *барионный*)» [СКМ-67, с. 92]. И здесь ощущается контраст между планами на будущее и весьма небольшим запасом реальных достижений.

Интригует раздел 5 той же главы, озаглавленный «О возможности формализованного распознавания означаемых в лексике». Ср. особенно путь преодоления семантического барьера, обозначенный на с. 106–107. Упорядочив все существительные по числу связанных с ними глаголов, Андреев надеется увидеть во главе списка слово со значением «человек». Прилагательное с максимальным разнообразием ассоциированных существительных должно иметь значение «большой». Открыв слово со значением «человек», надеемся по наибольшим значениям КФ обнаружить антропоклассы среди глаголов (*думать*, *нагибаться*, *разговаривать*, *смеяться*, *телеграфировать*, *судиться*), среди прилагательных (*веснущатый*, *талантливый*, *рыжебородый*, *завистливый*), среди субстантивов (*душа*, *пятки*, *взгляд*, *ненависть*). Высокая КФ субстантивов-подлежащих при словах антропокласса глаголов и тех же субстантивов при словах антропокласса прилагательных даст нам семантический класс имен деятеля [лучше было бы сказать класс обозначений человека] (*юноша*, *няня*, *директор*, *поэт*). Напротив, низкая КФ при именах деятеля выделит хрематоклассы (*плавиться*, *прорастать*, *электромагнитный*, *итерационный*, *теплоемкость*, *ирригация*), которые, в свою очередь, станут базой для выделения все новых и новых частично семантизованных классов слов.

Ну что ж! Все это выглядит очень соблазнительным. Дело за фактами. The proof of the pudding is in the eating!

1.2.2. Изучение языка дешифровочными методами

Независимо от Н. Д. Андреева подход «статистический анализ речи для открытия системы языка» начал разрабатываться в секторе структурной лингвистики Института русского языка. На симпозиуме по структурному изучению

знаковых систем в 1962 г. [Симпозиум 1962] прозвучали доклады двух сотрудников этого сектора — Виталия Викторовича Шеворошкина «О начальном этапе дешифровки буквенных письменностей» и Бориса Викторовича Сухотина «Общая задача дешифровки. Алгоритм установления связи слов в предложении».

В. В. Шеворошкин занимался реальной дешифровкой карийского языка, представленного небольшим числом надписей [См. Шеворошкин, 1965]. Первым шагом на пути к поставленной цели было разделение букв на два класса — гласных и согласных (и далее — соответствующих подклассов). Первая фраза тезисов его доклада звучит так: «При изучении структуры буквенного текста вскрываются некоторые очень простые закономерности, позволяющие чисто комбинаторным путем отделить гласные от согласных, выделить их подклассы, определить структуру слога, выделить лексические и грамматические морфемы и т. д.» [Симпозиум 1962, с. 57][1]. При этом Шеворошкин опирался на закономерности звуковых цепей в языках мира, позднее его исследования были обобщены в книге «Звуковые цепи в языках мира» [Шеворошкин, 1969].

Б. В. Сухотин подошел к задачам дешифровки с самых общих позиций. «Нормализуемый алгоритм является нормальным лингвистическим алгоритмом, если он использует только следующую информацию в тексте: дано, где расположена любая черная точка и где белая (предполагается, что текст написан черным по белому). Размеры точки определяются предельной разрешающей способностью устройства (например, глаза)» [Симпозиум 1962, с. 62].

«Последовательность все более сложных нормальных лингвистических алгоритмов примерно такая: алгоритм выделения букв (этот алгоритм можно построить на основе того, что буквы суть в некотором смысле устойчивые сочетания черных точек); алгоритмы классификации букв (например, на гласные и согласные и т. д.); алгоритм выделения морфем; алгоритм классификации морфем; алгоритм выделения слов; алгоритм классификации слов; алгоритмы выделения предложений; алгоритмы установления связей слов внутри предложения. В этом перечне опущены очень многие возможные алгоритмы.

Однако в практической деятельности по созданию алгоритмов нет необходимости придерживаться некоторой строгой последовательности. Нужно только принимать в качестве известного минимальное количество сведений, причем таких, для получения которых не слишком трудно составить нормальные лингвистические алгоритмы» [там же].

Последнее замечание очень важно. Алгоритм классификации морфем использует перечень уже установленных морфем, но алгоритм выделения морфем еще не продемонстрировал своей эффективности. Для алгоритма выделения предложений и алгоритма установления связей слов в предложении «нужно знать классы слов, т. е. отличать, что данное слово принадлежит к классу X, а не классу Y, но не нужно знать, что X есть, скажем, существительные» [там же, с. 63].

Подробнее вопрос о соотношении исходных данных и результатов работы того или иного алгоритма обсуждается в большой статье Б. В. Сухотина «Алгоритмы лингвистической дешифровки», опубликованной в сборнике «Проблемы структурной лингвистики 1963».

[Одно замечание по поводу термина. Реальная работа по дешифровке в двух отношениях отличается от всех исследований Б. В. Сухотина, сводимых под общую шапку «изучение языка дешифровочными методами».

Дешифровщик стремится как можно скорее нащупать реальное содержание в той или иной точке текста. Успех Г. Гротефенда в дешифровке древнеперсидской письменности в большой мере связан с идентификацией некоторых цепочек знаков как имен персидских царей. В построении общих алгоритмов (тем более — универсальных алгоритмов) рассчитывать на такие случайные прорывы нельзя.

Как правило, корпус реальных текстов, подлежащих дешифровке, невелик. Статистический переход от текстов к системе языка предполагает обращение к большим и очень большим собраниям текстов, обеспечивающим надежность результатов.]

«При работе по составлению дешифровочных алгоритмов не обязательно, конечно, педантично выдерживать подобную последовательность, но нельзя и слишком грубо нарушать ее, поскольку тогда не будет исключена возможность

[1] В. В. Шеворошкин не обсуждает проблем дешифровки консонантных систем письма, характерных для семитских систем письма. Разделение на гласные и согласные натолкнулось бы здесь на серьезные трудности.

принятия в качестве данных сведений, которые не могут быть получены нормальным лингвистическим алгоритмом. Нужно помнить, что нарушая подобную последовательность, т. е. забегая вперед, мы выдаем вексель на восполнение недостающей цепочки алгоритмов...

Алгоритм установления связей станет нормальным лингвистическим после значительно большего количества алгоритмов выделения букв,[1] имеющегося алгоритма выделения морфем, по-видимому, алгоритма классификации морфем на лексические и грамматические, алгоритма выделения основ, алгоритма выделения предложений и, наконец, алгоритма классификаций слов. Однако мы верим, что подобные алгоритмы могут быть созданы» [Сухотин, 1963, с. 78].

С самого начала Б. В. Сухотин подчеркивал важность экспериментальной работы на вычислительных машинах. Три группы задач прошли такие испытания: 1) Разбиение букв на два класса — гласные и согласные; 2) Выделение морфем; 3) Установление связи слов в предложении.

Для решения первой задачи сначала строится симметричная матрица попарных соседств букв в изучаемом тексте (т. е. дается сумма сочетаний русских ПС и СП, РК и КР, английских AE и EA, QU и UQ, французских UX и XU, CH и HC). Затем довольно сложный алгоритм начинает перетасовывать строки матрицы так, чтобы в исходной матрице образовалась два класса, буквы первого класса при этом должны редко сочетаться друг с другом, но часто сочетаться с буквами второго класса. Для русского языка был достигнут искомый результат — в один класс попали О, А, Е, И, У, Ь, Я, Ы, Э, Ю (Ь дистрибутивно ведет себя как гласный). Хороший результат был получен для французского и испанского языков. В английском эксперименте в первом классе оказались E, O, A, I, U, Y, но также и T, что плохо. [Проблемы структурной лингвистики. М., 1962, с. 203–204]. Еще хуже обстояло дело в случае немецкого языка, где с гласными объединились S, H, K. Второй (корректирующий) алгоритм устранил ошибки немецкого и английского языка, но во французском к гласным добавилась буква H. «Таким образом, корректирующий алгоритм выдает ответы или вообще без ошибок, или с ошибками, которые, как правило, менее грубы, чем те, что были ранее» [Сухотин, 1963, с. 98].

С лингвистической точки зрения наиболее поучительной была серия экспериментов по выделению морфем в текстах без пробелов между словами. Эти эксперименты нашли окончательное отражение в книге, специально посвященной этой задаче [Сухотин, 1984]. «Первый вариант алгоритма выделения морфем в текстах без пробелов между словами был опубликован автором настоящей книги в 1963 г. [ПСЛ-63]. Эксперименты по его проверке удалось закончить лишь к 1975 г. Их результаты были опубликованы в монографии [Сухотин, 1976]. Эти результаты можно было считать удовлетворительными только на начальном этапе исследований, но ни в коем случае не окончательными. Поэтому дальнейшее совершенствование алгоритма представлялось необходимым.

Некоторые соображения были самоочевидны. Поскольку программирование и отладка, производившиеся силами самого автора, оказались исключительно трудоемкими и длительными, следовало по возможности сохранить схему алгоритма и программ... Центральным элементом алгоритма является блок вычисления качества цепочек “устойчивости”... Во втором варианте алгоритма внешняя устойчивость, как и внутренняя, представлялась числом, имеющим смысл вероятности... Однако полный эксперимент по проверке нового варианта не оправдал надежд... Поэтому в третьем эксперименте был применен видоизмененный способ модификации устойчивостей, способ искусственный и компромиссный, поскольку правильный учет видоизменения устойчивостей требует радикальной перестройки программы и значительного увеличения расхода машинного времени и машинной памяти... третий вариант обеспечил значительное улучшение результатов по сравнению со вторым и тем более с первым вариантом алгоритма. Ввиду искусственности третьего варианта были предприняты дальнейшие поиски, которые привели к построению четвертого варианта алгоритма. Его особенность заключается в том, что внешняя и внутренняя устойчивости имеют характер абсолютных частом (математических ожиданий), а не вероятностей.[2]

---

[1] Проще принять заданность исходного алфавита букв. Автоматическое *выделение* букв сталкивается с неимоверными трудностями.

[2] Наконец-то! Как и Андреев, Сухотин оперировал лишь вероятностями, теперь осуществляется переход к математическим ожиданиям — первый шаг на пути к

... Четвертый вариант дал наилучшие экспериментальные результаты при сохранении простоты и естественности формулировки» [Сухотин, 1984, с. 12–14].

Грандиозность поставленной цели должна быть учтена при оценке полученных результатов. Рассмотрим для двух вариантов по 50 первых цепочек букв, выделившихся как морфемы.

Второй вариант*: человек сказалон головой варнавы прове глаза фук еще что капитан лсякапитан варнавапр братдюк блестел хватил только толсты овали значит делать вокруг сигби рукой подня омрач образ много истал живот ением вшись ается черт хлеб снул сить ойно оего нуты мура коза итут инно дана бенц ляю дюк это лдо письмо.*

Четвертый вариант: *капитан дюк про брат сигби человек глаза марианн нул енно варнава варнав котор сказа лопаткой головой говор пада еще что сказалон прове если сказал айте ами пере лсякапитан торжеств картофел ственно пропада братдюк блестел челове хватил толсты собака сказалдюк значит делать вокруг понял омрач олько образ живот дерев вшись больш.*

Прогресс налицо: число правильно опознанных корней увеличилось с 7 до 13, появились префиксы *пере*, *про-*, флексия *-ами;* но в большинстве случаев выделяются словоформы (*глаза*, *лопаткой*, *головой*, *сказал*...), последовательности морфем внутри слова (*нул*, *енно*, *айте*, *ственно*, *вшись*),пары словорм (*сказалон*, *сказалдюк*, *братдюк*) и даже конец одного слова со следующей словоформой (*лсякапитан*).[1]

Если положить предел, обрабатываем, скажем, 10 букв, в список морфем не попали бы *тебядорогой брат, появилсявдверях, кричалкапитано, капитансказал, инесметьтебе, укоризненно, третьегодня, тебядорогой, пропадайвсе, дляочищения*. Этот способ борьбы с длинными псевдоморфемами кажется предпочтительным в сравнении с альтернативным повышением порога реальной частоты цепочки с 2 до 4 (или 5). В больших текстах нам все равно встречались бы длинные цепочки, преодолевшие этот барьер. Так, в «Войне и мире» найдем цепочку букв *австрийскийгенерал* (f=10),*вглубинедуши* (f=10), *вследствиекоторого* (f=11), *поцеловалаего* (f=11), *видиможелая* (f=13), *военногоминистра* (f=13), *времясражен* (f=15), *всегдабывает* (f=18), *говорилпьер* (f=18), *чистоеделомарш* (f=18), *бородинскогосражения* (f=19), *всякуюминуту* (f=20), *русскиевойска* (=20), *глазамисмотре* (f=25), *князяандреяи* (f=30), *говорилаона* (f=62), *сказаланаташа* (=85), *аннамихайловна* (f=98), *князьвасилий* (f=149), *главнокомандующ* (f=155), *княжнамарья* (f=366), *князьандрей* (f=780).

Совсем неудовлетворительно членение на НС самого текста, о чем можно судить, например, по первому предложению: ((капитан)(дюк)) рано (утром) в (маленьк) (омгороде) (при) ле (гавш) емкодномуиздомиков ((общин)ы) ((голубых) (брат)ьев)) средизацветающ (его) (картофел) я (расс) аж (енно) го (прави) (льным) и (кус) т (ами) (появился) ((челове)к) летсорокаввязанойбез ((рук)ав) ке (морск) ихсуконныхштанаххитробо (образ) н (ойчер) ной ((шляп)е) [Сухотин, 1984, с. 77].

Оценка результатов поиска морфем не составляет труда для русиста. Громадное большинство словоформ текста легко разбивается на морфемы, и результаты машинных экспериментов наглядным образом демонстрируют все отклонения от «истинного» членения.[2]

Намного труднее оценить третью ветвь алгоритмов — эксперименты, направленные на установление синтаксических связей. Во-первых, в традиционной («семантической») лингвистике здесь нет единства мнений. В качестве догмы предполагаются неочевидные решения, например, представление структуры предложения в виде дерева (т. е. запрет двум стрелкам подчинять один узел),

---

статистической устойчивости (robustness) процедур и, следовательно, к работе с большими текстами.

[1] В грамматическом словаре русского языка А. А. Зализняка я обнаружил 41 основу длиной 9 букв, неразложимых на русской почве (из более распространенных упомяну *башибузук*, *инкунабул*, *канделябр*, *карбункул*, *катамаран*, *лапсердак*, *натюрморт*, *пастернак*, *пропаганд*, *репертуар*, *синедрион*, *сомнамбул*, *факультет*, *фейерверк*), 10 основ длиной 10 букв (*вундеркинд*, *гиппопотам*, *ландскнехт*, *спиричуэлс*...), 3 основы длиной 11 букв (*андерграунд*, *падепатинер*, *фольксваген*) и 1 основу длиной 12 букв — *страдивариус*.

[2] Морфемные алгоритмы Сухотина не предусматривали существования алломорфов [Сухотин,1984, с. 18], поэтому на две морфемы разбита словоформа челове-к.

выбор глагола как обязательной вершины предложения[1], необходимость разбивать сложные предложения на комбинацию простых.

Во-вторых, долгой истории разработки синтаксических алгоритмов подводится итог в последней книге Б. В. Сухотина. На одном и том же тексте («Пиковая дама» Пушкина — 1500 предложений) проведено 10 экспериментов. На первых двухстах предложений с 969 связями правильными оказались в третьем эксперименте 78% всех стрелок, выданных машиной, в девятом эксперименте соответственно 81%, а в десятом - 83%. Внимательному синтаксисту было бы полезно погрузиться в изучение конкретного материала [особенно в Сухотин, 1973, с. 464–488].

Однако остается главное замечание. Как и Андреев (с его функциональными классами), Сухотин со своими синтаксическими классами резко нарушает естественную логику движения от недешифрованного текста к системе языка. Вопреки тому, что провозглашалось в [Сухотин, 1963], в той же самой статье намечалась цель алгоритма установления лингвистических [точнее сказать — синтаксических] связей:

«Кодировка текстов для алгоритма установления лингвистических связей складывается из двух этапов. Первый этап не может быть пока формализован ввиду того, что не существует алгоритма определения классов слов. Этот этап должен осуществляться кодировщиком. В дальнейшем, по-видимому, классы слов могут быть выделены алгоритмом, аналогичным алгоритму определения классов букв» [Сухотин, 1963, с. 90].

Последующие десятилетия работы не привели к созданию алгоритма выделения синтаксических классов слов. И понятно - почему. Среди 19 синтаксических классов Сухотина 12 классов (падежи существительных и прилагательных) могут быть сформулированы только после того, как определены синтаксические связи в предложении. Налицо логический круг: информация, используемая в начале работы алгоритма, может быть получена только после завершения работы того же алгоритма.

С течением времени Б. В. Сухотин все более погружался в область логического построения алгоритмов, все более становился математиком как по стилю изложения, так и по сокращению (вплоть до полного исчезновения) лингвистического материала. [Ср. Сухотин, 1983 и вся вторая часть «Описание грамматических категорий средствами тензорной алгебры» в Сухотин, 1990]. Такой математический подход к созданию алгоритмов начался со стремления разбить матрицу соседств букв ровно на два класса, из чего следовал поиск оптимального решения для такого разбиения. Лишь в конце 1970-х гг. Сухотин обращается к ситуациям со многими классами, но и здесь конечный результат достигается через определение экстремума на каждом шаге работы алгоритма.

«... в наилучшем начальном делении последовательно осуществляются наилучшие транспозиции, пока множество разрешенных транспозиций не окажется пустым. Полученное деление на n классов считается наилучшим для данного n» [Сухотин, 1979].

1.2.3. Опыт изучения совместной встречаемости слов

К лингвостатистическим исследованиям я обратился сразу после завершения кандидатской диссертации (1959 г. — «Источники лексической омонимии в германских языках», защищена в 1963 г. в Московском государственном педагогическом институте им. Ленина). В то время я еще ничего не знал о работе Н. Д. Андреева. Результаты исследования были отражены в докладе на Межвузовской конференции по применению структурных и статистических методов исследования словарного состава языка (ноябрь 1961 г.), происходившей в стенах 1-го Московского государственного педагогического института иностранных языков. Суть нового подхода отражает первый абзац опубликованных тезисов:

«До сих пор выделение семантических полей языка основывалось, как правило, на экстралингвистических соображениях, хотя дальнейший анализ мог носить исключительно языковедческий характер. Не решая принципиального

[1] Прямо запрещается иметь в одном предложении два глагола в личной форме [ПСЛ-72, с. 456].

вопроса, что такое семантическое поле — явление собственно лингвистическое или явление смешанного характера, можно лишь предположить следующее: семантические связи слов должны, в частности, проявляться в их относительно частом появлении вместе в текстах. Отсюда ясно, что, изучая статистические корреляции слов в тексте, можно делать и выводы относительно семантических связях.

Для проверки этого предположения проведено исследование на материале английских прилагательных».

Ниже дается соответствующая публикация 1963 г.

А. Я. ШАЙКЕВИЧ

РАСПРЕДЕЛЕНИЕ СЛОВ В ТЕКСТЕ И ВЫДЕЛЕНИЕ СЕМАНТИЧЕСКИХ ПОЛЕЙ

Одна из очередных задач семасиологии заключается в том, чтобы выделить семантические поля отдельных языков. При этом приходится сталкиваться с серьезными теоретическими и практическими трудностями.

Остается до сих пор невыясненным вопрос о сущности понятия «семантическое поле», о месте этого понятия в кругу других лингвистических понятий. Более или менее выявились две противоположные точки зрения: 1) семантическое поле — факт языка и только языка, его выделение и анализ должны основываться исключительно на лингвистических данных; 2) семантическое поле — явление экстралингвистическое (психологическое или логическое), но находящее отражение в фактах языка и поэтому интересующее лингвистов. Нетрудно заметить, что и в том, и в другом случае делается известное допущение: признается наличие некоторых сущностей (в первом случае — групп слов или значений, во втором — групп идей и эмоций), имеющих известную самостоятельность, замкнутость. Считается, что компоненты этих сущностей системно организованы, представляя собой «значимости» в понимании Ф. де Соссюра. Само такое допущение вполне в духе современной лингвистики, рассматривающей объект своего исследования как систему. Тем не менее это допущение до сих пор остается допущением. Быть может, оно имеет силу не для семантики в целом, а лишь для ее отдельных участков (например, для семантики грамматических единиц, для некоторых групп слов). В таком случае факты действительности в разной степени систематизируются языком. Значения, связанные с абстрактными понятиями, легче укладываются в систему, чем значения, связанные с конкретными предметами. Вопрос не может быть решен только в результате теоретических рассуждений. Существование «семантических полей» (в любом их понимании) тогда будет признано окончательно, когда у нас будут для этого достаточные объективные доказательства (лингвистические или экстралингвистические). Возможно, в процессе накопления этих доказательств будет меняться и само понятие «семантическое поле».

Отказываясь (на данном этапе) от определения понятия семантического поля, предполагаем лишь следующее: слова в языке вступают друг с другом в разнообразные отношения. На грамматическом уровне эти отношения сравнительно легко выявляются при формальном анализе (без обращения к анализу значения). На уровне лексики отношения между словами определяются не только системой языка, но в конечном итоге и внешней действительностью, тем не менее эти отношения выражены в языке и, следовательно, тоже допускают формальный анализ (например, изучение дистрибуции лексических единиц).

Иными словами, предполагается, что формальный анализ позволяет выявить отношения языковых единиц независимо от того, чем обусловлены сами эти отношения. Если в результате подобного анализа будет показано, что некоторые группы единиц языка, слабо связанные друг с другом, обнаруживают разнообразные и тесные связи элементов внутри самих групп, мы действительно сможем убедиться в существовании относительно замкнутых подсистем, которые и назовем условно «семантическими полями». При этом нельзя забывать, что выделенные «семантические поля» будут целиком зависеть от метода анализа, сколько методов - столько типов классификации. Это вызовет значительные терминологические затруднения, в конце концов сам термин «семантические поля» закрепится за какой-нибудь одной системой классификации (наиболее близкой к традиционному пониманию этого термина), а для остальных будут созданы новые термины. Несколько упрощая задачу, можно сформулировать ее так: **не обращаясь**

к анализу значения, выявить связи, слов и соответствующие группы слов; в дальнейшем переходить к интерпретации этих отношений уже с учетом соответствующих реалий.

В настоящей статье делается попытка проверить реальность такой задачи на ограниченном материале и с помощью только одного метода. В основе исследования лежала следующая гипотеза: 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.

Материалом исследования служат английские прилагательные в произведениях Чосера, Спенсера, Марло, Шекспира, Мильтона, Уордсворта, Шелли, Арнольда. Выбор именно этих авторов обусловлен чисто техническими соображениями (наличием соответствующих словарей) и поэтому не может быть признан вполне удачным: не соблюден принцип синхронного изучения языка, рассматриваются произведения разных жанров: чистая поэзия и драматургия. Все это необходимо будет учесть при оценке результатов исследования. Делается единственное, но очень важное допущение: предполагается, что мы умеем на основании формальных критериев отличать прилагательные от других частей речи. Это допущение особенно важно для английского языка, в котором прилагательные как класс выделяются не очень четко.[1]

Прилагательные отбирались из словарей-конкордансов[2], лишь для Марло прилагательные отбирались непосредственно из текста[3].

Из общего числа были отобраны для дальнейшего анализа все слова, встретившиеся у данных авторов не реже 15 раз, при условии, что по крайней мере два автора используют данное слово. Таких прилагательных оказалось 1078. Можно полагать, что общая длина текстов, из которых отобран материал исследования, приближается к 2 млн слов (более 250 тыс. строк). Общее число прилагательных у отдельных авторов приводится в следующей таблице:

Автор	Всего прилагательных в тексте (в тысячах)	Из них подверглось дальнейшему анализу	
Чосер	21	17,7	84%
Спенсер	31	29,4	95%
Марло	7	6,1	87%
Шекспир	42	36,9	88%
Мильтон	11	9,1	83%
Уордсворт	38	31,5	83%
Шелли	18	15,8	88%
Арнольд	8	6,4	80%
Всего	176	152,9	87%

[1] Фактически прилагательные выделялись не на основе формальных критериев, а на основе интуиции автора. Отбирались словоформы, которые обычно используются как прилагательные в английском языке. Отсев словоформ, относящихся к другим частям речи, произведен только для слов: just — справедливый, kind — добрый, left — левый, light — легкий, светлый, right — правый, справедливый, wrong —- неправый. Лексические омонимы не различались (например, rare — редкий и rare — сырой считались одним словом). Исключены из анализа слова content, fast, full, last, open, sound. Следует заметить, однако, что метод, примененный в данном исследовании, можно было бы проверить и на других частях речи или даже на всех словах языка без предварительной разбивки на классы. Прилагательные были выбраны потому, что в этом разряде слов, как предполагалось, легче выделить замкнутые группы, зависящие преимущественно от языковой системы.

[2] J. S. Tat1ock, A. G. Kennedy. A Concordance to the Complete Works of Geoffrey Chaucer, Wash., 1927; Ch. G. Osgood, A Concordance to the Poems of Edmund Spenser, Wash., 1915; J. Bartlett, A New ... Concordance ... to the Works of Shakespeare, L., 1894; J. Bradshaw, A Concordance to the Poetical Works of John Milton, L., 1894; L. Cooper, A Concordance to the Poems of William Wordsworth, L., 1912; F. S. Ellis, A Lexical Concordance to the Poetical Words of P. B. Shelley, L., 1892; S. M. Parrish, A Concordance to the Poems of Matthew Arnold, Ithaca, 1959.

[3] The Works of Chr. Marlowe, Oxford, 1941.

Для каждого прилагательного фиксировались все случаи появления его вместе с другими прилагательными в одной строке. Зная общее число прилагательных в тексте, мы можем вычислить вероятность появления двух данных прилагательных в одной строке (для простоты вычисления общий объем текстов приравнивается к 250 000 строк). Рассмотрим механику вычисления на примере пяти прилагательных: bitter, blue, red, sweet, white (предполагается, что мы ничего не знаем о значении этих слов).

	Всего в тексте	Вероятность появления в строке
bitter	341	0,0014
blue	308	0,0012
red	473	0,0019
sweet	2547	0,0102
white	750	0,0030

Для перехода к оценке вероятности совместного появления двух прилагательных в одной строке мы можем воспользоваться формулой: X = np, где для нашего случая: n — число случаев появления первого прилагательного, p — вероятность появления второго прилагательного в строке. Таким образом, для нашего примера при чисто случайном распределении (т. е. при полной независимости этих пяти прилагательных) мы ожидали бы следующее число случаев совместного появления двух данных прилагательных в одной и той же строке (предполагается, что вероятность появления каждого прилагательного остается постоянной):

	blue	red	sweet	white
bitter	0,4	0,6	3,5	1,0
blue	-	0,6	3,1	0,9
red	-	-	4,8	1,4
sweet	-	-	-	7,6

Однако фактическое распределение наших прилагательных в тексте отличается от ожидаемого. В текстах мы наблюдаем следующее число случаев совместного появления данных слов:

	blue	red	sweet	white
bitter	-	-	13	-
blue	-	9	2	16
red	-	-	1	50
sweet	-	-	-	4

В статистике существуют разные способы определения существенности отклонения фактических данных от ожидаемых. Для нашего исследования, где вероятность совместного появления двух событий очень мала, а число наблюдений велико, можно предположить, что при полной независимости отдельных событий вероятности их совместного появления распределялись бы по закону Пуассона. Резкие отклонения от этого распределения указывали бы на наличие корреляции между рассматриваемыми событиями.

По закону Пуассона вероятность появления фактического числа событий (у нас — наблюдаемое число случаев совместного появления двух прилагательных в одной строке) вычисляется по формуле

$$Pm = npme^{-np} / m!,$$

где для нашего случая m — фактическое число случаев совместного появления, а e — известная константа математики (n и p см. выше). Таким образом, вероятность случайного совместного появления двух прилагательных ровно столько раз, сколько они действительно появились, равна:

	blue	red	sweet	white
bitter	0,670	0,549	5,7 10~5	0,368
blue	-	2,6 10-8	0,216	3,6 10-15
red	-	-	0,040	1,6 10-58
sweet	-	-	-	0,070

Для небольших значений np построены соответствующие таблицы. Еще большую ценность для нас представляют таблицы суммарных значений функции Пуассона, в которых определяется вероятность появления данного события m число раз и больше[1].

Итак, для наших пар прилагательных вероятность случайного характера наблюдаемых (и больших) отклонений от ожидаемых составляет:

bitter - sweet	7,6 10-5	blue - red	5,2 10-8
blue - white	7,2 10-15	red - white	3,2 10-58

(величины меньше ожидаемой величины нас уже не интересуют, поскольку они не могут свидетельствовать о положительной корреляции событий).

Из этих данных можно сделать лишь один вывод: наблюдаемые случаи совместного появления приведенных четырех пар прилагательных настолько расходятся с ожидаемыми, что ни в коем случае не могут объясняться случайностью. Между bitter и sweet, между blue, red и white несомненно существует сильная зависимость. Обращение к значениям соответствующих слов (bitter — горький, sweet — сладкий, blue — синий, red — красный, white — белый) совершенно ясно указывает характер этой зависимости. Корреляция в распределении данных слов в тексте, очевидно, обусловлена смысловой связью этих слов.

Для более наглядного изображения степени связи исследуемых слов удобнее пользоваться не самими полученными вероятностями, а логарифмами чисел, обратных этим вероятностям, по формуле

Коэффициент[2] смысловой связи = К.С.С. = -log 20P.

Коэффициент 20, вводимый в эту формулу, позволяет оценивать нулем отклонения, объясняемые случайностью в 5% всех случаев и больше. Отклонения, оцениваемые от 0 до 1, объясняются случайностью в 0,5-5% всех случаев. Отражаемые подобной величиной смысловые связи будем считать слабыми смысловыми связями. При величине от 1 до 3 соответствующие смысловые связи будем считать средними, при величине от 3 до 6 – сильными и при величине К.С.С. больше 6 — очень сильными. В этом случае сила связи наших прилагательных будет оцениваться следующим образом:

bitter - sweet	2,8	blue - red	6,0
blue - white	12,8	red - white	56,2

В нашем исследовании был установлен нижний предел для учитываемых случаев фактического совместного появления (m=2), однократное совместное появление не учитывалось, как бы мала ни была вероятность такого события. Кроме того, число случаев совместного появления уменьшается на единицу, если данная пара прилагательных появляется совместно только у одного автора.

Из 1078 рассмотренных прилагательных 442 обнаружили какие-то смысловые связи с другими прилагательными. Всего выявлено 763 смысловые связи. Прежде чем переходить к анализу этих результатов, необходимо устранить влияние посторонних факторов, связанных с самим характером нашего материала. Речь

---

[1] При исследовании материала использовались таблицы из справочников: Г. З. Купарадзе. Справочник экономиста. Тбилиси, 1960; Я. Янко. Математико-статистические таблицы. М., 1961. Для значений m, отсутствующих в таблицах, суммарное значение принимается равным двум Pm.

[2] Конечно, название этого коэффициента условно. Фактически фиксируется неслучайный характер отношений между двумя единицами, но в нашем случае такими единицами являются слова, и мы можем предполагать, что неслучайный характер их связи объясняется в большинстве случаев смыслом этих слов.

идет о влиянии аллитерации. Если бы этот фактор не оказывал никакого влияния на совместное появление прилагательных в одной строке, мы ожидали бы следующее число связей у прилагательных, начинающихся на один и тот же звук: слабых связей — 17, средних — 16, сильных — 6, очень сильных — 1. На самом деле наблюдаем соответственно: 41 — 34 — 19 — 13, отсюда можно сделать вывод о несомненном искажении интересующих нас данных под влиянием аллитерации. Необходимо ввести какую-то поправку. К сожалению, не удалось выявить теоретических оснований для такого коэффициента, пришлось довольствоваться чисто эмпирической поправкой. Число совместных появлений пары прилагательных, начинающихся на одну и ту же букву, уменьшалось на 1/4 (с дальнейшим уменьшением до целого числа). В результате этой поправки число связей уменьшилось следующим образом: слабые связи — 20, средние — 26, сильные — 7, очень сильные — 6, т. е. стало намного ближе к ожидаемому. У отдельных авторов (у Шекспира, Уордсворта, Шелли) применение такой поправки приводит к снижению числа связей между созвучными прилагательными даже ниже ожидаемой, только у Спенсера аллитерация — настолько обычное явление, что число наблюдаемых связей даже с поправкой в три раза превышает число ожидаемых связей. Следует заметить вместе с тем, что многие аллитерирующие прилагательные интуитивно кажутся связанными по смыслу, что, вероятно, является одной из характерных особенностей английского языка, особенно поэтического.

Сравним такие пары:

baleful	- гибельный	bitter	- горький
black	- черный	blue	- синий
courteous	- учтивый	kind	- добрый
deadly	- смертельный	direful	- ужасный
deaf	- глухой	dumb	- немой
doleful	- скорбный	dreary	- мрачный
faint	- слабый	feeble	- слабый
fair	- красивый	foul	- некрасивый
lief	- любимый	loath	- (в ср. англ.) ненавистный
slow	- медленный	swift	- быстрый
thick	- толстый	thin	- тонкий

После введения поправки на аллитерацию число прилагательных с выявленными смысловыми связями уменьшилось до 426 (715 смысловых связей). Анализ слов с выявленными смысловыми связями позволяет сделать два вывода.

1. Чем больше размер текста, чем больше частотность анализируемых прилагательных, тем больше можно обнаружить смысловых связей. Если разбить все рассмотренные прилагательные на три группы: а) встретившиеся более 100 раз; б) встретившиеся от 40 до 100 раз и в) встретившиеся от 15 до 40 раз, то число прилагательных, обнаруживших смысловые связи, составит в первой группе 248 из 324 (или 77%), во второй — 120 из 297 (или 40%), в третьей — 60 из 411 (или 15%). Это же проявляется и в том, что число подобных слов в произведениях отдельных авторов (Уордсворт и Шелли — 197, Спенсер — 213, Шекспир — 272) ниже, чем в объединенном материале всех восьми авторов (426). Дальнейшее расширение материала, несомненно, увеличило бы число обнаруженных смысловых связей.

2. В нашем материале на выявление смысловых связей влияла также длина слова. Односложные слова гораздо чаще обнаруживали смысловые связи. Процент слов, обнаруживших смысловые связи, среди различных групп слов отражен в следующей таблице:

	Слова, встретившиеся в тексте		
	более 100 раз	40 — 100 раз	15 — 40 раз
Односложные слова	89%	66%	30%
Двусложные слова	70%	35%	12%
Многосложные слова	57%	40%	14%

Отрицательное влияние многосложности на выявление смысловых связей, по-видимому, будет уменьшаться в текстах других жанров, где не будет сказываться влияние стихотворного размера.

Обнаруженные смысловые связи представлены на схеме.[1]

Пунктирной линией показаны слабые смысловые связи, пунктирной линией с точкой — средние, сплошной линией — сильные смысловые связи и двойной сплошной линией - очень сильные связи. Площадь кружков пропорциональна частотности соответствующих прилагательных.

Из слов, не обнаруживших смысловых связей, отметим наиболее важные (в скобках — число слов в тексте):

alone	(831) — один(окий)	piteous	(202) - жалкий;
angry	(222) — сердитый;	present	(391) - настоящий;
certain	(487) — некоторый;	quiet	(310) - спокойный;
common	(491) — обычный;	ready	(453) - готовый;
eternal	(370) — вечный;	royal	(434) - королевский;
former	(244) — прежний;	sovereign	(296) - превосходный;
glorious	(390) — славный;	sure	(596) - уверенный;
native	(284) — родной;	warlike	(220) - воинственный;
natural	(254) — природный;	wicked	(426) - нехороший.

Первая часть поставленной задачи — выявление смысловых связей, оппозиций — выполняется сравнительно легко. Но, хотя установление смысловых связей и их интенсивность само по себе представляет интерес, в настоящей работе оно является лишь подготовительной работой по отношению к более трудной и сложной задаче — выявлению смысловых групп слов — «семантических полей».

Среди всех рассмотренных слов очень небольшая часть организуется в «абсолютно замкнутые» группы, т. е. пары или группы слов, совсем не связанные с другими парами или группами. Сюда относятся 17 парных сочетаний прилагательных, например, cheerful — радостный и cheerless — безрадостный; Christian — христианский и pagan — языческий; female — женский и male — мужской; fleshly — плотский и ghostly — духовный; inward — внутренний и outward — внешний; Greek — греческий и Latin — латинский[2].

К «абсолютно замкнутым» относятся также две группы слов: 1) цепь слов от public — общий до particular — особенный и 2) слово dear — дорогой и его спутники (см. схему). Остальные 382 слова так или иначе связаны друг с другом. Для того чтобы без учета значений распределить это огромное количество слов по группам, необходимо выработать некоторую жесткую программу. Можно принять, например, следующую процедуру: отправляясь от какого-то слова, добавлять к нему по одному слову таким образом, чтобы интенсивность внешних связей получающейся группы не увеличивалась ни в один из моментов присоединения нового слова. Мерилом интенсивности связей будем считать сумму коэффициентов К.С.С. Градация линий, обозначающих смысловые связи, позволяет приближенным образом воспроизводить всю процедуру прямо по схеме. В качестве образца проделаем эту операцию со словом good — хороший: + ill + evil + excellent + honest (совершенно очевидно, что при всех этих присоединениях сумма внешних связей группы не растет, а уменьшается), затем можно присоединить сюда же слово bad (К.С.С. good — bad = 30, т. е. больше, чем сумма К.С.С. bad — worse — 5,0 и К.С.С. bad — bold — 1,9). На этом присоединение новых слов заканчивается: всякое дальнейшее расширение группы приводило бы к росту внешних связей. Расширение группы слова great — великий, большой закончилось бы на присоединении слов huge и brazen (поскольку К.С.С. great — huge не превышает сумму К.С.С. huge с другими словами).

Последовательно применяя этот метод ко всем смысловым связям, представленным на схеме, мы получим 89 групп слов. Некоторые из этих групп, обладая высокой интенсивностью смысловых связей внутри группы, с другими группами связаны сравнительно слабо. Такие группы слов можно назвать «относительно замкнутыми» семантическими полями. К ним можно отнести[3]:

---

[1] Вместо схемы, приложенной к оригинальной классификации, здесь даются рис. 1–4.

[2] Среди этих пар слов встречаются такие, интерпретация которых вызывает известные трудности, например, almighty — всемогущий и vengeful — мстительный, careless — беспечный и little — маленький.

[3] Метод исследования позволяет нам дать нумерацию групп как единственное средство их различения. Для наглядности, однако, английские слова

1) «семантическое поле жизни» с центральными словами dead — мертвый и alive — живой (20 : 4);
2) «семантическое поле высоты» с центральными словами high — высокий и low — низкий (45 : 4);
3) «семантическое поле общего размера» с центральными словами great — большой, small — маленький, huge — огромный (16 : 2);
4) «семантическое поле тонкости» с центральными словами thin — тонкий, thick — толстый, subtle — тонкий (7 : 0,2);
5) группа слов с центральным словом right — правый (50 : 4);
6) группа слов с центральным словом fine — чистый, нежный, хороший (7 : 0,4);
7) «семантическое поле упрямства и раздражительности» с центральным словом sullen — угрюмый (2,2 : 0,2);
8) «семантическое поле бессмертия» с центральными словами mortal — смертный и immortal — бессмертный (6 : 0,2);
9) «семантическое поле гордости» с центральными словами proud — гордый и humble — смиренный (17 : 2).

Также «относительно замкнутыми» являются две пары слов: 1) keen — острый, резкий и frosty — морозный; 2) Elysian — райский и fortunate — счастливый. С известными сомнениями к «относительно замкнутым» можно отнести: «семантическое поле правды» с центральными словами true — верный, правильный и false — лживый (36 : 6) и «семантическое поле святости» с центральным словом holy — святой (7 : 1,2). В последних случаях выделяемая группа слов обнаруживает по нескольку слабых и средних внешних связей, но направлены они к разным группам и не дают оснований для предположения о их большей близости к какой-то определенной группе. До сих пор мы выделили 32 «семантических поля» (12 групп и 20 пар), на схеме остается еще 76 групп и пар и 30 одиночных прилагательных, не вошедших в группы при описанной выше операции. Полученные группы слов могут оказаться в дальнейшем «семантическими полями» или подгруппами в составе «семантических полей», в последнем случае мы будем называть их «элементарными группами».

Прежде всего рассмотрим некоторые группы и пары слов, имеющие лишь одну внешнюю связь, мало уступающую по интенсивности внутренним связям в группе. В качестве вспомогательного критерия при решении вопроса о том, считать эту группу самостоятельным «семантическим полем» или нет, можно привлечь данные о частотности слов. Условимся считать данную группу самостоятельным «семантическим полем» в том случае, если совокупная частотность всех слов данной группы превышает частотность того слова, с которым эта группа связана; в противном случае подключаем данную группу к указанному слову. В этом случае получаем три новых «семантических поля»:
1) «семантическое поле немоты» со словами dumb, mute — немой, deaf — глухой (всего 386 слов в тексте, т. е. больше, чем частотность слова dull, с которым эта группа связана);
2) «семантическое поле узости со словами narrow и steep — узкий;
3) пара слов Grecian — греческий, Trojan — троянский.

С другой стороны, «элементарная группа» perfect — совершенный — подключается к семантическому полю с центральным словом good — хороший; группа barren — голый, бесплодный — присоединяется к группе bare-naked — голый; группа modest — скромный присоединяется к группе old-young-new — старый-молодой-новый; группа fearless — бесстрашный — присоединяется к слову free — свободный, которое и становится центром «семантического поля свободы»; группа ancient — старинный — присоединяется к слову new —— новый. Благодаря последней операции число выявленных «семантических полей» увеличилось до 36 (14 групп и 22 пары), не уточнено еще 68 групп и пар и 30 одиночных прилагательных.

На этом новом этапе мы можем повторить прием, уже применявшийся к отдельным словам; начиная с какой-то элементарной группы, присоединять по одной элементарной группе или одиночному слову таким образом, чтобы интенсивность внешних связей ни в один из моментов присоединения новой группы или слова не увеличивалась. Например, элементарная группа с центральным словом fair — красивый (э. г. fair) присоединяет к себе э. г. foul — отвратительный, грязный, затем э. г. vile-base — подлый, затем э. г.

сопровождаются переводом и дается условное название семантического поля, если интуитивно обнаруживается общность смысла слов, входящих в группу.

abominable — отвратительный; в результате получаем большое «семантическое поле морально-эстетической оценки». Применяя эту методику, получаем следующие «семантические поля» (с. п.):
1) с. п. «общей оценки» с тремя э. г.: good — хороший, perfect — совершенный, worse — хуже;
2) с. п. «земли и неба» с двумя э. г.: heavenly — небесный, earthly — земной и э. г. divine — божественный — human — человеческий;
3) с. п. «объема» с четырьмя э. г.: а) deep — глубокий; б) hollow — пустой; в) wide — широкий; г) broad, large — широкий;
4) с. п. «стремительности» с центр. словом rash;
5) с. п. «необычности» с центр, словом strange;
6) с. п. «сладости» с э. г. sweet — сладкий и э. г. bitter — горький;
7) с. п. «тяжести» с двумя э. г.: dull, heavy — тяжелый, light — легкий и slow — медленный, swift — быстрый;
8) с. п. «температуры и влажности» с двумя э. г.: cold — холодный, hot — горячий и dry — сухой, wet, moist — влажный.
9) с. п. «бледности» с двумя э. г.: pale — бледный и lean — тощий, fat — жирный;
10) с. п. «слабости» с четырьмя э. г.: faint, feeble — слабый; weary — усталый; sick — больной, whole — здоровый, целый и tedious — утомительный;
11) с. п. «смертельности» с двумя э. г.: deadly — смертельный и woeful — скорбный;
12) с. п. «мягкости» с двумя э. г.: tender — нежный, sharp — острый и soft — мягкий.

В нашем списке уже 49 с. п. (27 групп и 22 пары), остались неуточненными 40 групп и пар и 17 одиночных прилагательных.

До сих пор уменьшение внешних связей происходило при попарном соединении слов или элементарных групп. На новом этапе анализа производятся пробные соединения элементарных групп и отдельных слов по три, если при этом интенсивность внешних связей уменьшается, например, если сумма связей после объединения трех групп окажется меньше половины суммы внешних связей всех трех групп до объединения, рассматриваемые группы признаются элементарными группами одного и того же семантического поля. Таким образом, мы выделяем еще два семантических поля: 1) с. п. «умственной неполноценности» с тремя э. г.: mad, fond — безумный, foolish — глупый и idle — ленивый, vain — тщетный, тщеславный; 2) с. п. «покоя и нежности» (слова gentle, mild, calm). Две элементарные группы: 1) blue — синий и 2) white — белый, black — черный, red — красный, yellow — желтый при этом могут быть объединены в четыре возможные группы: сочетаясь с э. г. green — зеленый, э. г. grey — серый, brown — коричневый, pale — бледный и со словом purple — пурпурный. Мы выбираем последнюю возможность, поскольку в этом случае сумма внешних связей оказывается наименьшей. Получившееся семантическое поле можно назвать с. п. цвета. После этого на схеме остается еще 14 одиночных, 11 пар и 23 группы; группы и пары слов мы и признаем особыми семантическими полями, поскольку они не вошли в другие семантические поля. Назовем важнейшие из оставшихся семантических полей:

с. п. «бедности» (с центральными словами poor — бедный и rich - богатый);

с. п. «веселья и грусти» (с центральными словами sad — грустный и merry — веселый);

с. п. «радости» (с центральным словом glad — радостный);

с. п. «дикости» (со словами wild, savage — дикий, tame — смирный и rude — грубый);

с. п. «времени» (с центральными словами old — старый, new — новый и young — молодой);

с. п. «свежести» (с центральным словом fresh — свежий);

с. п. «силы» (с центральными словами strong — сильный и weak — слабый);

с. п. «яркости и ясности» (с центральными словами bright — яркий и clear — ясный);

с. п. «счастья» (с центральным словом happy — счастливый);

с. п. «чистоты» с двумя э. г.: э. г. pure — чистый и э. г. innocent — невинный.

Некоторые из полученных «семантических полей» можно было бы объединить в более крупные единицы. Схема показывает наличие связей между с. п. wise —

мудрый, с. п. valiant — отважный, с. п. worthy — достойный, с. п. noble — благородный или у с. п. «цвета» (выделенного выше) с с. п. green — зеленый и с с. п. grey — серый. Всего благодаря указанным выше методам 412 слов оказались сгруппированными в 53 группы и 33 пары, которые мы назвали «семантическими полями», вне «семантических полей» осталось 14 прилагательных. Некоторые из них имеют по две слабые внешние связи (например, lofty — высокий, imperial — царственный), другие имеют по несколько связей с разными «семантическими полями» (например, serene — безмятежный, azure — лазурный, smooth — гладкий и т. д.).

Анализ результатов исследования показывает общую правильность выдвинутой нами гипотезы. Выделенные «семантические поля» лишь в редких случаях противоречат нашему интуитивному представлению о смысловой классификации английских прилагательных. Вместе с тем примененный нами метод имеет целый ряд недостатков:

1. Выявлены смысловые связи и их интенсивность, но не выяснен характер смысловых связей, не показано качественное отличие синонимических, антонимических и других видов связи.

2. Метод не позволяет различать значения слов. Между тем разные значения одного и того же слова и омонимы (которые тоже не различались) могут входить в разные семантические поля. Это обстоятельство создает трудности при попытке четкого разделения семантических полей. Можно полагать, что детализация дистрибуции слов позволит преодолеть эту трудность.

3. Как и всякое другое применение статистики, данный метод сопряжен с большой затратой труда и времени (около 400 часов). Собирание материала непосредственно из текстов увеличило бы затраты времени еще в несколько раз.

Наряду с указанными недостатками отметим и важнейшие преимущества примененного нами метода:

1. Метод исключает всякую субъективность и зависит от минимального числа предварительных условий.

2. Сама техника работы очень проста и может быть полностью автоматизирована. Это обстоятельство особенно облегчает использование в данном случае электронно-счетной машины, благодаря которой можно было бы в несколько раз увеличить скорость исследования, работать непосредственно с текстом, брать не только строку, но и другие интервалы текста.[1]

Комментарий:

Это было первое исследование совместной встречаемости слов, направленное на обнаружение смысловых связей. Сопоставление реальной частоты события с математическим ожиданием этого события оказалось шагом вперед по сравнению с простым сравнением условной вероятности с независимой вероятностью (КФ Андреева). Оно позволило ввести соображения статистической значимости в ДСА.

В ретроспективе очень важным кажется заключительное примечание к статье. Впервые вводится словосочетание «интервал текста» (правда, без эксплицитного объяснения) и намек на мысль о семантической специфичности разных интервалов. [См. ниже, п. 5]

Увлеченность тогдашней семасиологией идеей семантического поля, восходящей к замечательной работе Йоста Трира (J. Trier, Der deutsche Wortschatz im Sinnbezirk des Verstandes, Hdlb, 1931), сказалась и в этой статье. Сейчас я бы не стал так настойчиво стремиться к выделению дискретных полей и групп. Предлагаемые конкретные меры, способствующие разделению групп, не должны абсолютизироваться. Идеи сети, пространства, поля выступают как некоторые антитезы глобальной дискретности современной лингвистики. Более сдержанный и взвешенный подход отражен в последующих публикациях.

[1] Применение другого интервала (скажем, 5 строк или страница) существенно изменило бы и сами смысловые связи, и получающиеся при этом семантические поля.

1.3. Принципы работы ДСА

Основная идея дистрибутивного анализа заключается в следующем.

Располагая пространным текстом (или большим собранием текстов) и не зная языка, на котором написан текст, исследователь пытается открыть систему этого языка. Заданными считаются буквы соответствующего алфавита, причем считаются уже отождествленными заглавные и строчные буквы (русские А и а, Е и е, латинские B и b, D и d, G и g, H и h, L и l, N и n, R и r, T и t). Сохранены пробелы между словами; в алфавит включены цифры, дефис и апостроф. Все остальные знаки при работе ДСА игнорируются.[1]

Дистрибутивно-статистический анализ мыслится как медленный процесс последовательного применения некоторых алгоритмов, объединенных общей идеей. На каждом новом этапе исследования строится какая-то нулевая гипотеза, основанная на предположении о независимости событий (с их вероятностями), зафиксированных на предыдущем этапе. На основе такой нулевой гипотезы подсчитывается математическое ожидание осуществления тех или иных событий. Статистически значимое расхождение между математическим ожиданием и реально наблюдаемой частотой события и есть то новое, что мы получаем на каждом новом этапе. Новое знание дает возможность пересчитать вероятности каких-то новых событий и построить новые нулевые гипотезы для следующего этапа.

Как правило, вероятности событий, наблюдаемых в ДСА, достаточно малы. Вероятность двух букв оказаться рядом в русском тексте едва ли превысит 0,01; вероятность двух слов оказаться рядом заведомо меньше 0,002. Так, в «Войне и мире» вероятность слова И равна 0,047, вероятность слова В равна 0,024, в предположении независимости для вероятности появления И + В имеем $P = 0{,}047 \times 0{,}024 = 0{,}0011$. Математическое ожидание последовательности И + В в этом романе (около 450 тысяч словоупотреблений) составляет 493. Реальная частота И + В (482) оказалась удивительно близкой к математическому ожиданию, из чего следует несомненный вывод: И никак не прогнозирует В. Реальная частота В + И равна 1. Без какого бы то ни было статистического критерия всякому ясно, что последовательность В + И практически запрещена в русском языке.[2]

Чрезвычайно малые вероятности, характерные для языковых элементов и тем более для комбинации языковых элементов, заставляют подозревать, что соответствующие события распределялись бы в тексте согласно закону Пуассона (в предположении взаимной независимости этих элементов и при взгляде на порождение текста как на случайный процесс).[3]

В дальнейшем в качестве приближения к пуассоновскому распределению в оценке статистической значимости отклонений от математического ожидания 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.[4]

В тех случаях, когда изучаемый корпус текстов состоит из очень разнородных подкорпусов, удобнее пользоваться следующей модификацией нашей основной формулы:

---

[1] Если на каком-то этапе потребуется учесть некоторые из опущенных знаков, такое отступление от исходного алфавита должно быть оговорено.

[2] Этот численный пример хорошо иллюстрирует разницу в подходе современного лингвиста и адепта ДСА. Современный лингвист скажет — такая последовательность допустима, и тут же придумает контекст: *в и вне.* Реально эта последовательность встретилась в романе в таком контексте: «Посетили меня братья Г. В. и О.».

[3] Во избежание недоразумений подчеркнем, что эти предположения есть часть нулевой гипотезы, которая и должна быть опровергнута в ходе работы ДСА.

[4] Основное внимание в ДСА уделяется случаям превышения реальной частоты над математическим ожиданием. В тех редких случаях, когда нас будет интересовать сильное превышение математического ожидания над реальной частотой, наша формула примет вид: $S = (f-m+1)/\sqrt{m}$.

$$S = (f-m-1)/ \sqrt{m},$$

где m — сумма математических ожиданий в отдельных подкорпусах.

Если S > 2, исследователь должен насторожиться, при S > 3 возникнут подозрения в неслучайности события, при S > 4 подозрения превратятся в уверенность. При очень больших S (например, при выделении идиоматических выражений или двусловных терминов) этот показатель может логарифмироваться. В вышеприведенном численном примере для И в S = -0,45: отклонения статистически не значимы, в случае же В и S = -21,8 отклонение значимо в высшей степени и должно быть интерпретировано как запрет для И появляться после В.

Ясно, что при таком подходе мы заведомо отказываемся от полного описания языка. В русском языке, например, не будет учтено ударение. Правда, косвенным свидетельством существования ударения окажутся стихи. Особенности русской орфографии не позволят обнаружить многие случаи ассимиляции. Не вводя в исходный алфавит русскую букву ё, мы не только теряем информацию о месте ударения, но и многие детали грамматики и лексики. Одним словом, исследователь работает с текстом как с текстом на мертвом языке, даже если это язык живой, и к нему можно обратиться при оценке результатов, полученных формальным способом.

Слово **формальный** употребляется здесь в смысле ‘не использующий семантическую информацию’[1]. В работе алгоритма не используются прямо многие понятия и термины традиционной (‘семантической’) лингвистики.

Все результаты основываются, в конечном счете, на комбинаторике букв и графических слов, т. е. цепочек букв между пробелами. Так, в ходе ДСА может быть открыта сопряженность графических слов *видел, видела, видело, видели* и множества подобных четверок с другими «основами», будет выявлен неслучайный характер сочетаний

он (андрей, пьер, ростов), ты, я - видел
она (графиня, марья, наташа, сама), я - видела
он (балашев, государь, граф, наполеон), я - сказал
она (наташа, соня) - сказала;
сказал - он (адъютант, анатоль, багратион, балашев, берг, билибин, болконский, борис, виконт, генерал, граф, денисов, доктор, долохов, дрон, жерков, ильин, казак, капитан, каратаев, князь, кто-то, кутузов, масон, наполеон, несвицкий, николай, офицер, петя, пьер, растопчин, ритор, ростов, старик, сын, тихон, тушин, француз, шиншин, эсаул)
сказала - она (m-lle, анна, вера, гостья, графиня, губернаторша, девушка, дуняша, жюли, княжна, марья, мать, наташа, соня, элен).

Но тот факт, что формы на *-л* и на *-ла* относятся к прошлому, что их различие связано с различием пола, останется недоступным для исследователя. Однако исследователь заметит, что слова, соседствующие с *видел* и *сказал*, не кончаются на -А или -Я (за единственным исключением *петя*), а слова, соседствующие с *видела* и *сказала*, часто кончаются на *-а* или *-я* (за исключением *-lle*, *жюли*, *элен*, *мать*).

Вернувшись к внутреннему строению графических слов, мы откроем среди «существительных»[2], не кончающихся на *-а* или *-я* и соседствующих с формами на *-л*, следующие «парадигмы», т. е. возможность присоединять к «основе» такие-то «аффиксы».

1. -0, -а, -е, -ом, -у : борис, генерал, граф, наполеон, офицер, пьер, ритор, старик, сын, француз (средняя частота всех членов «парадигмы» = 497);

---

[1] Единственное исключение сделано для европейских (а теперь — международных) цифр.

[2] Употребляемые в этом параграфе лингвистические термины даются в кавычках, ведь у нас на этом этапе еще нет строгих доказательств, что это существительные.

2. -0, -а, -ом, -у : адъютант, багратион, берг, доктор, дрон, казак, капитан, тихон, эсаул (с.ч.= 101);
3. -0, -а, -у : виконт, масон (с.ч.= 35);
4. -0, -а, -е, -у, -ым : денисов, долохов, каратаев, кутузов, растопчин, ростов (с.ч. = 389);
5. -0, -а, -у, -ым : балашев, билибин, жерков, ильин, тушин, шиншин (с.ч. = 64);
6. -ий, -им, -ого, -ом, -ому : болконский (ч. = 167);
7. -ий, -им, -ого, -ому : несвицкий (ч. = 73);
8. -е, -ем, -ь, -ю, -я : князь (ч. = 2023);
9. -ь, -ю, -я : анатоль (ч. = 216);
10. -ем, -й, -ю, -я : николай (ч. = 548).

Различия в частоте позволяют отождествить «парадигмы» 2 и 3 с «парадигмой» 1, «парадигму» 5 с «парадигмой» 4, «парадигму» 7 с «парадигмой» 6 и «парадигму» 9 с «парадигмой» 8 — из-за меньшей частоты основ какие-то «аффиксы» случайно не появились.

Свои «парадигмы» появились и у слов с исходом на -А или -Я:
1. -а, -е, -ой, -ою, -у, -ы : анна, вера, княжна;
2. -а, -е, -и, -ой, -у девушка;
3. -а, -е, -ей, и, -у : дуняша, наташа;
3а. -а, -е, -ей, -и : губернаторша;
4. -е, -ей, -ею, -и, -ю, -я : марья;
4а. -е, -ей, -и, -ю, -я : графиня, петя, соня;
4б. -е, -и, -ю, -я гостья.

Доказать, что *-ой* и *-ою*, *-ей* и *-ею* не разные «морфемы», а лишь «алломорфы одной морфемы», очень не просто в рамках ДСА. Решающим аргументом будет сходство дистрибуции обоих аффиксов относительно соседних слов (в скобках дана частота), ср. *анной михайловной* (6) и *анною михайловной* (5); *за анной* (2) и *за анною* (2), *перед княжной* (2) и *перед княжною* (1); *с анной* (4) и *с анною* (2); *с княжной* (18) и *с княжною* (1); *княжной марьей* (27), *княжною марьей* (1) и *княжною марьею* (1); *марьей дмитриевной* (3) и *марьею дмитриевной* (1).

Может возникнуть вопрос: зачем строить систему алгоритмов (ДСА), если конечные результаты окажутся лишь структурными сущностями (единицами и классами единиц плана выражения), а не полнокровными двусторонними знаками?

Ответ первый (чисто психологический) — все дело в методическом азарте, поддерживающем мой интерес на протяжении сорока лет.

Ответ второй (лингвистический и филологический) — по ходу применения ДСА будут уточнены (или даже пересмотрены) некоторые традиционные таксономические понятия грамматики (идеи частей речи и членов предложения, идеи лемм и парадигм), особое поле исследований открывается в области лексики и стилистики. Даже если мы не откроем пути в Индии, мы по дороге узнаем много нового.

Ответ третий (совершенно фантастический) — давайте подготовимся к появлению громадного текста на неизвестном языке, представляющего собой последовательность элементарных единиц, укладывающихся в обозримый алфавит.[1]

1.4. Черты поведения лингвистических элементов, используемые в ДСА

1.4.1. Синтагматическая сочетаемость

Последовательность лингвистических элементов — важнейший ключ в поисках осмысленных единиц языка, его общей структуры. В самом трудном случае — при отсутствии каких-либо границ в тексте — Б. В. Сухотин обратился именно к синтагматической сочетаемости букв в поисках морфем.

[1] Вспоминаю, как Борис Викторович Сухотин пригласил меня поехать в Харьков на конференцию по радиоастрономии, где он делал доклад в секции «Общение с внеземными цивилизациями». Сейчас я забыл детали, но помню ощущение чего-то грандиозного и захватывающего.

При нашем допущении пробелов между словами, т. е. признании существования графических слов (цепочек букв между пробелами) появятся и другие критерии, но синтагматическая сочетаемость останется важным источником информации. Рассмотрим конкретный пример.

Зная частоту и вероятности русских букв в романе «Война и мир», мы подсчитаем математическое ожидание цепочки *леж* в графических словах этого текста. Оно составит 105, реальная же частота этой цепочки равна 304, статистическая значимость этого отклонения заставляет нас обратить внимание на *леж* как на потенциальную единицу языка. В 213 случаях за *леж* следует *а* (m=24,7, S=37), в 46 случаях — *ит* (m=1,0, S=44), в четырех случаях *н* (m=19,8), в трех случаях *ск* (m=1,6) (в составе буквенной цепочки *коллежск*), в пяти случаях *у* (m=7,9). Обратимся к самой значимой цепочке *лежа*. В девяти случаях за ней следует пробел, в 158 — *л* (m=9,2, S=48), в 35 — *вш* (m=0,58, S=43), в 26 — *щ* (m=0,58, S=31), в 17 — т (m=12,8). Первые три цепочки выглядят кандидатами в основы. У этих трех цепочек находим такие продолжения:

лежал 0 = 64 (S=5), а = 37 (S=7), и = 31 (S=7) , о = 26;

лежавш ее 4 (S=11), ие 3 (S=5), ий 5 (S=19), им 3 (S=6), их 3 (S=8), ую 9 (S=44);

лежащ его 11 (S=31), ее 4 (S=13), ий 2 (S=4).

Если обратится к левому окружению цепочки *леж*, то окажется, что в 217 случаях она следует за пробелом, а в 79 случаях за *д* (S=24). Дальнейшее движение налево покажет, что цепочке *длеж* предшествует *на* в 46 случаях (S=12), *по* в 30 случаях (S=8), *пре* в трех случаях. Цепочке *надлеж* предшествует *при* в 44 случаях (S=88).

С позиций семантической лингвистики эти результаты кажутся неплохими. Правые буквенные продолжения намечают путь к полной глагольной парадигме глагола *лежать* (хотя опора на буквенные цепочки не опознала значимость *лежу*, *лежат*, *лежа*, *лежало* как особых форм парадигмы). Изучение левых предшественников в конечном счете правильно опознало *под*, *над* и *принад* как вероятных префиксов, но придало слишком большую значимость букве *д*, подталкивая исследователя к ложной интерпретации — *по-д-лед, на-д-леж, прина-д-леж*.

Подход, описываемый в 1.4.4 и в Части 3, окажется более эффективным.

Предположим, однако, что алгоритмы ДСА сработали в полную силу, и у нас есть три глагола ЛЕЖАТЬ, ПОДЛЕЖАТЬ и ПРИНАДЛЕЖАТЬ; пусть одновременно правильно лемматизировались слова субстантивного и адъективного склонения. Тогда у глагола ЛЕЖАТЬ (f=216) обнаружатся левые статистически значимые соседи: КОТОРЫЙ (f=16, S=9), *он* (f=19, S=7), *она* (f=7, S=4), *неподвижно* (f=3, S=15), *долго* (f=3, S=6), ПРЕДМЕТ (f=2, S=4).

Справа от глагола обнаружим: *на* (f=43, S=21), *в* (f=35, S=11), *у* (f=8, S=8), ОБЛОКОТИТЬСЯ (f=3,S=16), *высоко* (f=3, S=13), *там* (f=3, S=5), *ничком* (f=2, S=24), *пред* (f=2, S=6). Семантика пространства проглядывает здесь очень ярко.

У глагола ПОДЛЕЖАТЬ (f=30) слева находим две значимые связи *не* (f=8, S=8) и *более* (f=4, S=17). Справа представлены исключительно формы дательного падежа: ЗАКОН (f=7, S=59), *законам* (f=4, S=70), *закону* (f=3, S=73), *необходимости* (f=3, S=71) *разуму* (f=2, S=49), *одному* (f=2, S=47), *тем* (f=2, S=4).

Глагол ПРИНАДЛЕЖАТЬ (f=41) показывает всего одну значимую связь слева с КОТОРЫЙ (f=4, S=4) и с двумя словами справа К (f=9, S=13), *ему* (f=3, S=4).

Подведем итог этому маленькому экскурсу. Рассмотренные три глагола демонстрируют идентичную морфологию, но существенно различаются в своих правых окружениях. Этот результат интерпретируется содержательно следующим образом: при этимологически одном и том же корне наши три глагола настолько расходятся семантически, что уже не относятся к одному словообразовательному гнезду.

Статистически значимые сочетания слов могут стать яркими маркерами, различающими тексты, авторов, жанры. Рассмотрим в качестве примеры бинарные сочетания с предлогом *до* у двух авторов (числа указывают на значения S).

Толстой «Война и мир»		Достоевский
30	до свидания	130
148	до сих (пор)	425
21	до вечера	30
17	до утра	30
11	до обеда	9

до завтрашнего 97, до поздней 18, до приезда 11, до смерти 11, до весны 9

до ночи 43, до гроба 12, до могилы 11, до седых (волос) 22, до свадьбы 17, до барабана 17

до Вильны 19, до Дриссы 19, до Смоленска 18, до Вязьмы 16, до Москвы 16, до барьеров 18

до поворота 35, до печки 21, до дому 14, до ворот 11, до дверей 8, до дивана 8, до квартиры 6, до кладбища 6, до железной (дороги) 6

53	до конца	107
24	до последней	139
13	до самой	63
15	до земли	26

до бесконечности 27, до высшей 26, до такой 30, до малейших 60

до крайности 99, до невероятности 84, до глубины 45, до ушей 35, до головы 32, до невозможности 31, до нитки, до потолка, до nec plus ultra, до копейки

26	до слез	52

до жестокости 9

до безумия 38, до безобразия 27, до бесчувствия, до бешенства, до болезни, до восторга, до гадливости, до грубости, до дикости, до забвения, до идиотства, до изнеможения, до исступления, до истерики, до комизма, до крику, до мучения, до неприличия, до обморока, до обожания, до отчаяния, до самозабвения, до столбняка, до сладострастия, до страдания, до страсти, до странности, до суеверия, до сумасшествия, до тошноты, до ярости

(не) до игрушек 9, (не) до смеху 7, (не) до тону, (не) до вас, (не) до тебя.

У наших писателей представлены четыре семантические группы сочетаний: сочетания времени, сочетания пространства, сочетания логического предела, сочетания эмоционального предела. Последняя группа непомерно раздута у Достоевского. Только у Достоевского мы находим группу «*не до*».

1.4.2. Статистические распределения лингвистических элементов

Первоначально заданные элементы (буквы и графические слова) каким-то образом распределены в тексте. При формальном методе вроде ДСА изучение статистических распределений может дать какую-то информацию о классах соответствующих единиц, а часто — и о структуре текста.

В конце 1960-х гг. была сделана попытка проверить оправданность такой надежды. Вот (в сокращении) первая публикация на эту тему.

А. Я. Шайкевич

Интервал текста и характер статистических распределений языковых единиц

В 1962 году появилась важная статья Р. М. Фрумкиной «О законах распределения слов и классов слов». С тех пор в советской лингвостатистике неоднократно обращались к изучению статистических распределений лингвистических элементов. Цель настоящей статьи — уточнить постановку задач, связанных с этой проблемой, в свете понятия «интервал текста».

Под интервалом текста здесь понимается один из множества равных отрезков, на который разбивается текст; при этом предполагается, что для каждого конкретного исследования точно фиксируется длина такого интервала. Можно, например, один и тот же текст расчленить на интервалы по 100 слов в каждом, по 1000, по 5000 и т. п., на интервалы длиной в строку, в страницу, 10 страниц и т. п. В дальнейшем постараемся показать, что выбор того или иного интервала существенно влияет на характер статистического распределения тех или иных лингвистических элементов.

Сначала рассмотрим статистический материал, касающийся букв поэтического текста в русском и английском языках. Для русского языка был взят текст «Евгения Онегина» Пушкина, для английского — текст «Дона Жуана» Байрона и «Королевы фей» Спенсера.

В данной статье мы будем последовательно сопоставлять эмпирические распределения с теоретическим распределением Пуассона, используя для этого аппарат критерия согласия χ^2 (так же, как и в исследовании Р. М. Фрумкиной).

Первый интервал, который будет рассмотрен для букв, — строка поэтического текста (1311 строк для Пушкина, 1912 строк для Байрона, 900 строк для Спенсера). Распределения представлены в табл. 1.1 и 1.2.

Таблица 1.1

	А	Т	Д	К	Й	Ь
0	264	435	577	749	876	884
1	468	523	525	394	356	347
2	361	254	174	141	75	74
3	150	83	34	23	4	4
4	53	11	1	4	-	2
5	11	3	-	-	-	-
6	3	2	-	-	-	-
x	1,47	1,03	0,75	0,58	0,40	0,39
s	1,28	0,93	0,62	0,60	0,37	0,39
P	2,2%	1,8%	0,06%	13%	22%	51%

Как видим, гипотеза о Пуассоновом распределении должна быть принята для К, Й, Ь, она ставится под сомнение в случае А и Т и решительно отвергается для Д.

Таблица 1.2

	Byron			Spenser		
	B	D	F	B	D	F
0	1155	470	1011	534	115	403
1	569	702	658	292	306	284
2	157	466	185	73	236	418
3	26	193	42	7	148	49
4	5	64	14	4	69	12
5	-	15	2	-	20	4
6	-	2	-	-	6	-

x	0,54	1,34	0,64	0,53	1,82	0,87
s	0,53	1,25	0,66	0,51	1,60	1,04
P	76%	37%	4,4%	60%	1,9%	2,4%

Гипотеза распределения Пуассона здесь ставится под сомнение для F у обоих авторов и для D у Спенсера.

Переход к интервалу в 1 строфу резко меняет всю картину (Пушкин — 93 строфы, Байрон - 239, Спенсер - 100). Суммарные данные представлены в табл. 1.3.

Таблица 1.3

	x	s	P		x	s	P
	Пушкин				Byron		
А	20,6	27,3	31%	B	4,1	4,6	79%
Т	14,4	22,0	14%	D	10,7	14,9	1,0%
Д	10,4	11,6	88%	F	5,0	5,0	18%
К	8,1	8,6	67%		Spenser		
Й	5,6	9,1	0,03%	B	4,8	3,4	57%
Ь	5,4	12,9	0,02%	D	17,0	10,2	72%
				F	8,0	10,2	94%

Итак, интервал «строфа» заставляет нас сомневаться в распределении Пуассона для D у Байрона, и отказаться от этой гипотезы для Й и Ь у Пушкина.

Как раз значительные отклонения от теоретического распределения и представляют наибольший интерес для лингвостатистики, поэтому рассмотрим их более детально. Обратимся к графикам, отражающим эмпирические распределения (сплошная линия) в сравнении с теоретическими (пунктирная линия).

На рис. 1 сплошная линия эмпирического распределения ниже пунктирной линии теоретического распределения для крайних значений (0 и 3, 4), хвост эмпирического распределения короче, чем у теоретического. Это ясно указывает на то, что возмущающие факторы, вызывающие отклонения от теоретического распределения, носят «ограничительный» характер. Они лимитируют появление экстремальных значений, т. е. для данного случая мешают слишком большому скоплению Д в одной строке: ожидая теоретически восемь строк, где Д встречается четыре раза или больше, фактически сталкиваемся только с одной такой строкой. Вмешательство данного фактора лимитирует распределение «сверху». В какой-то степени этот же фактор вмешивается и в распределении буквы А, хотя здесь и встретилась строка с восемью А (*Рвалась и плакала сначала*). Для буквы D у Спенсера «ограничительный» фактор сильнее проявляет себя снизу (115 строк, где не встретилась эта буква, против ожидаемых 146; для строк с пятью и более D соответствующие цифры 26 и 34)...

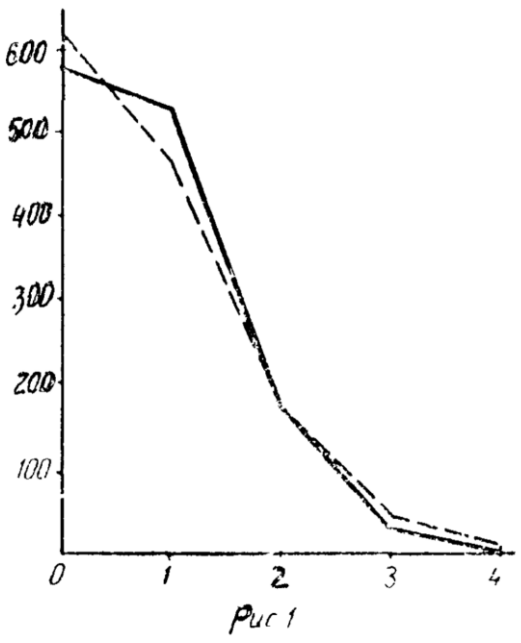

Рис 1

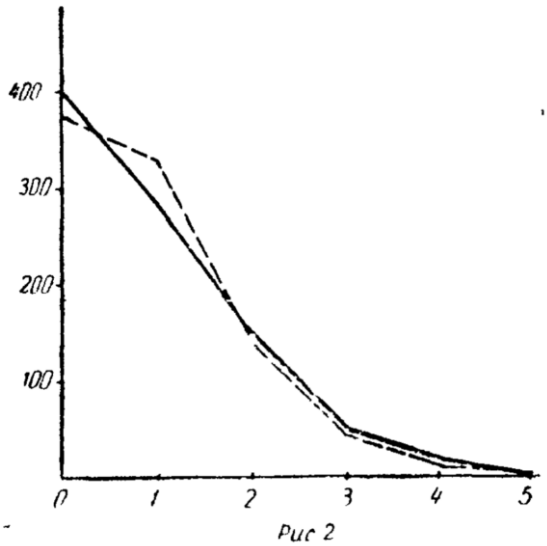

Рис 2

Рис. 3 и 4, наоборот, иллюстрируют действие «поощрительного» фактора. Экстремальные значения эмпирических распределений здесь выше, хвосты длиннее. Для интервала «строка» поощрительный фактор проявлял себя только сверху. У Байрона 16 строк с четырьмя и более F (например, where different talents find their different marts) против ожидаемых восьми; у Спенсера 16 против 11 (например, But fly, ah, fly far hence away, for fear). Основной поощрительный фактор — любовь к аллитерации у Спенсера, частое обращение к аллитерации и производным словам латинского происхождения у Байрона (рис. 3).

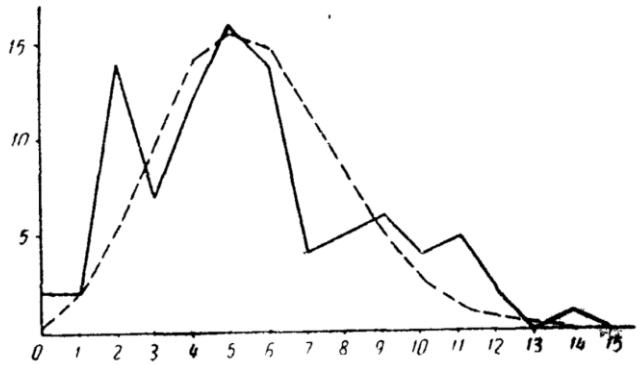

Рис 3

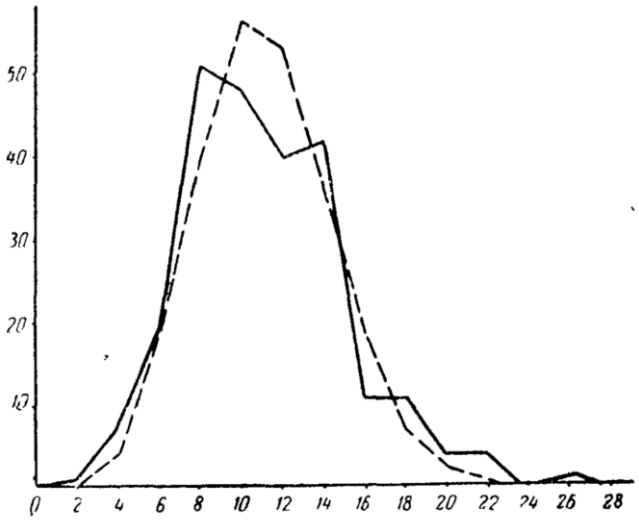

Рис 4

Содержательно интерпретируя данный «ограничительный» фактор, можно указать на фонетические и эвфонические причины для интервала «строка» у Пушкина и на грамматические причины в случае D у Спенсера (см. далее о частях речи).

При интервале «строфа» поощрительные факторы действуют как сверху, так и снизу, т. е. способствуют появлению сгустков и пустот данного элемента. Так для Й (рис. 4) и Ь в «Евгении Онегине»:

		Теоретически	Фактически
Й	2 и менее	7,6	18
	9 и более	10,6	18
Ь	2 и менее	8,8	17
	9 и более	9,0	17

То же характерно и для D у Байрона (см. рис. 2).

Для интервала «строфа» действуют уже не фонетические причины, а причины грамматические и семантические, например, сгустки прилагательных и существительных женского рода: вызывающие появление Й, ср.:

Ночей Италии златой
Я негой наслажусь на воле,
С венециянкою младой,
То говорливой, то немой,
Плывя в таинственной гондоле;

синтаксические повторы инфинитивов и отглагольных существительных (ср. 11 строфу первой главы: *Как он умел казаться новым, Шутя невинность изумлять…* и т. д.); в этой строфе 22 мягких знака, т. е. в 4 раза выше средней. У Байрона это — чередование сюжетной линии с глаголами в прошедшем времени: а следовательно, и буквой d, и лирических отступлений с глаголами в настоящем времени. Напротив, у Спенсера повествование ведется последовательно в прошедшем времени, и значит появление d в тексте (в интервале «строфа») хорошо описывается распределением Пуассона.

Следует особо обратить внимание на то обстоятельство, что во всех случаях, где обнаружился «поощрительный» фактор, дисперсия была не меньше арифметической средней, наоборот, при наличии «ограничительного» фактора дисперсия всегда была ниже средней. Это дает возможность использовать дисперсию как первое указание на существование возмущающих факторов. Этот прием становится особенно важным при недостаточном числе наблюдений, когда критерий χ^2 нельзя применять. <…>

Переходим к анализу статистического распределения трех частей речи (существительных, прилагательных и глаголов) на материале четырех глав «Евгения Онегина» при разных интервалах.

В качестве интервала выбираем двустишие. Сопоставление эмпирических и теоретических распределений для этого интервала приводятся в табл. 1.4.

Таблица 1.4

	Глагол		Прилагательное		Существительное	
x	эмпир.	теорет.	эмпир.	теорет.	эмпир.	теорет.
0	284	346	394	423	72	106
1	522	456	495	474	215	272
2	372	300	368	262	410	334
3	166	131	89	97	350	274
4	20	43	10	27	198	168
5	4	14	2	7	38	83
6					15	36
7						12
8						5
x	1.31		1,11		2,46	
s	0,92		0,98		1,57	

Данные распределения резко отличаются от теоретических: для прилагательных PX = 0,05%, для глагола и существительного — в несколько раз меньше.

Анализ распределений показывает, что это отклонение вызвано «ограничительным» фактором: для прилагательных он действует сверху, для глаголов и существительных – и сверху, и снизу. Конечно, встречаются двустишия с большим скоплением слов, принадлежащих к одной и той же части речи, но количество их значительно меньше ожидаемого, ср.

Здоровье, жизни цвет и сладость,
Улыбка, девственный покой.

Блестящей, ветреной, живой,
И своенравной, и пустой.

Еще не перестали топать,
Сморкаться, кашлять, шикать, хлопать.

Вероятно, такой же ограничительный фактор действует и в английских поэтических текстах, что и сказалось на распределении буквы D у Спенсера.

Переход к интервалу «строфа» (176 строф) снова вводит распределение частей речи в рамки распределения Пуассона. Суммарные данные таковы:

	x	s	PX
Глагол	9,07	9,28	33%
Прилагательное	7,83	8,61	35%
Существительное	17,42	13,47	20%

Итак, было показано, что на одном и том же тексте, для одних и тех же лингвистических единиц мы получали совершенно разные статистические распределения при изменении интервала текста.

Отсюда следует важные для лингвостатистики выводы:

— нельзя просто задавать вопрос: «Как распределены данные языковые единицы?», необходимо всегда добавлять «при таком-то интервале текста»;

— нельзя полагать, что лингвистические единицы одного уровня (при данном интервале текста) подчинены одному и тому же закону распределения;

— следует всегда четко разделять распределения: полученные на сплошном тексте, и распределения, основанные на выборке (это скорее соображение априори).

[Центральный научно-исследовательский институт патентной информации и технико-экономических исследований, Вопросы лингвостатистики и автоматизации лингвистических работ, выпуск III, М., 1970, с. 15–22]

В том же выпуске ЦНИИПИ опубликована статья «Средняя и дисперсия как критерий классификации слов», где указанные критерии прилагались к текстам совершенно иного типа — американским патентным формулам (из «Official Gazette» за 1966 г.). Было рассмотрено 1500 документов 35 классов, относящихся к автотранспорту; подсчитана средняя и дисперсия у 64 частых слов.

В группу слов с высокой дисперсией вошли слова, называющие детали машин arm, axle, bearing, brake, cam, channel, обобщенные названия элементов конструкции: body, element, наконец, несколько слов, входящих в состав сочетаний — названий деталей: air, discharge, drive. Короче говоря, это — группа предметных слов.

В группу слов с низкой дисперсией входят:

1) названия машин или конструкции как целого: apparatus, assembly, device, engine;

2) слова, указывающие на пространственные соотношения частей: adjacent, angle, area, axis, back, bottom, center, define, direction, elongate, etc.;

3) слова, обозначающие функции, причины и следствия: carry, engage, cause, effect;

4) слова — указания изменения расположения или назначения деталей: adapt, arrange, bias, dispose и т. д.

Получилось разделения на «лексические» (предметные) слова и слова «грамматические» (функциональные и топографические).

Различия статистических распределений дают возможность обнаружения контрастов между единицами и в художественной прозе. Вот несколько примеров из «Войны и мира». Разбив текст романа на фрагменты по 50 слов (около 9000 фрагментов), получим такие распределения для следующих слов:

```
Николай   1=312 2=36 3=1 4=1          x=0.0436 s=0.0522
Наполеон  1=206 2=45 3=3 4=1          x=0.0345 s=0.0467
сам       1=374 2=9  3=1              x=0.0441 s=0.0448

ни     1=418 2=150 3=29 4=7 5=5 6=1 9=1   x=0.0974 s=0.1771
ты     1=395 2=148 3=26 4=6 5=1 6=3       x=0.0910 s=0.1572
Наташа 1=519 2=121 3=20 4=1               x=0.0920 s=0.1259
теперь 1=734 2=80  3=4                    x=0.1010 s=0.1123
```

Слово *сам* «ограничено» сверху, здесь находим всего 9 фрагментов, где *сам* встречается по два раза, и 1 фрагмент с тремя *сам*. Имена собственные (*Николай* и *Наполеон*) «поощряются» сверху, дисперсия превышают среднюю. Слово *теперь* идеально соответствуют распределению Пуассона; *ты* и *Наташа* демонстрируют высокую дисперсию, еще выше дисперсия у графического слова *ни* (намек на существование союза *ни…... ни*).

При увеличении размеров изучаемого интервала дисперсия растет быстрее средней, но все-таки не очень сильно отличается от нее.

	Интервал 500 слов		Интервал 2000 слов	
	x	s	x	s
бледный	0.1048	0.127	0.42	0.59
грудь	0.1093	0.150	0.44	0.56
второй	0.1093	0.144	0.44	0.61
дурной	0.111	0.157	0.45	0.67

Слова, не связанные с внутренней структурой текста, хорошо подчиняются закону Пуассона даже при увеличении размеров интервала:

	Интервал 500 слов		Интервал 2000 слов	
	x	s	x	s
жест	0.0836	0.105	0.33	0.42
выражать	0.0836	0.090	0.33	0.37
волнение	0.0825	0.082	0.33	0.44
замечать	0.0814	0.094	0.33	0.38
дожидаться	0.0491	0.053	0.16	0.22
договорить	0.0479	0.052	0.19	0.21
доложить	0.0669	0.075	0.27	0.28

Особенно характерно такое поведений для наречий и других слов на -О:

	Интервал 500 слов		Интервал 2000 слов	
	x	s	x	s
громко	0.0658	0.070	0.26	0.31
дурно	0.0647	0.085	0.26	0.30
беспрестанно	0.0792	0.099	0.32	0.33
действительно	0.1026	0.107	0.41	0.44
вероятно	0.0825	0.093	0.33	0.40
вообще	0.0513	0.053	0.20	0.20
верхом	0.0491	0.055	0.20	0.25
высоко	0.0479	0.050	0.19	0.19
гораздо	0.0457	0.045	0.18	0.16

Экстремальные значения дисперсии обычно связаны с именами или другими обозначениями персонажей романа:

	Интервал 500 слов		Интервал 2000 слов	
	x	s	x	s
Анна	0.351	1.969	1.40	22.04
Василий	0.272	1.324	1.09	11.75
Анатоль	0.249	1.703	1.00	28.98
Багратион	0.177	1.100	0.71	9.54
доктор	0.180	0.854	0.72	5.66
гусар	0.160	0.576	0.64	5.08
главнокомандующий	0.172	0.535	0.69	4.02
Александр	0.148	0.489	0.59	3.50
гость	0.165	0.412	0.66	2.77

Впрочем, очень высокую дисперсию могут показать любые «сюжетные» слова и шире — слова, так или иначе связанные со структурой текста:

	Интервал 500 слов		Интервал 2000 слов	
	x	s	x	s
бал	0.1048	0.345	0.42	2.68
зала	0.120	0.286	0.48	1.78
ворота	0.0792	0.235	0.32	1.21
действие	0.177	0.636	0.71	5.46
закон	0.170	1.008	0.68	10.56
деятельность	0.146	0.445	0.58	2.98

Последние три слова концентрируются в последней части романа, где движение сюжета уступает место изложению взглядов Толстого на историю. Таким образом, сравнение средней и дисперсии может помочь в установлении внутренней структуры текста.

1.4.3. Позиционный анализ

В тех случаях, когда в тексте имеются явно обозначенные границы, эти границы могут стать важным источником информации для работы ДСА. В качестве первого примера рассмотрим пробел между словами как элементарнейшую границу в тексте. Тогда в цепочках букв между пробелами (графических словах) можно выделить отдельные позиции относительно пробела — левую букву (позицию +1 в терминологии Андреева), правую букву (позицию -1). Именно с такого позиционного анализа начинал работать алгоритм Андреева, открывающий морфологические типы.

Возможность позиционного анализа проиллюстрируем материалом художественных произведений Достоевского. В отличие от процедур Андреева общая статистика букв здесь дифференцирована — для всех слов, для слов с относительной частотой > 0,001, для остальных слов. Соответствующие данные приводятся в табл. 1.5, где F — абсолютная частота буквы (в тысячах), P — относительная частота буквы (в промилле).

Таблица 1.5

	Все слова		Слова с P >0,001		Слова с P <0,001	
Буквы	F	P	F	P	F	P
А	719	80	131	79	588	80
Б	160	18	47	28	113	15
В	419	47	81	49	338	46
Г	165	18	20	12	145	20
Д	283	31	47	28	236	32
Е	797	89	180	108	617	83
Ж	101	11	32	19	69	10
З	144	16	25	15	119	16
И	577	64	116	70	461	63
Й	92	10	2	1	90	12
К	294	33	76	46	218	30
Л	415	46	48	29	367	50
М	296	33	30	18	266	36
Н	585	65	135	81	450	61
О	1021	113	219	131	802	110
П	237	26	17	10	220	30
Р	360	40	16	10	344	47
С	476	53	54	32	422	58
Т	577	64	148	89	429	59
У	253	28	31	19	222	30
Ф	13	1,5			13	2
Х	70	8	3	2	67	9
Ц	27	3			27	4
Ч	165	18	48	29	117	16
Ш	79	9			79	11

Щ	25	3	7	4	18	2
Ъ	2	0,2			2	0,2
Ы	154	17	52	31	102	14
Ь	193	21	29	17	164	22
Э	30	3	18	11	12	2
Ю	57	6			57	8
Я	206	23	59	35	147	20
Всего	8992		1671		7331	
Средняя длина слова	4,9		2,4		6,5	

Применяя нашу формулу оценки статистической значимости (см. стр. 28), получим такие значения S у правых хвостов (отдельных букв и диграмм):

```
-а        41   -ал 135  -ам 61  -ах 79  -аю  116  -ая 232
                            -бя      66
                            -ву      54
                            -го     236
                            -гу      65
                            -да      62
                            -ду      55
                            -дь      92
                            -дя      51
-е       152    -ее 153  -ей 193  -ем 171  -ет 118  -ех 77  -ец 133  -ею 77
                            -же 63  -жу 88
-и        46    -ие 203  -ии 151  -ий 198  -ил 107  -им 164
                -их 157  -ич  50  -ию 117  -ия 170
-й       462
                            -ке 107  -ки 98  -ку 100

-л         42     -ла   68  -ли 157  -ло 78
-м        189     -ми  149  -му 210  -мя 78
                            -но 237  -нь 54  -ню 70
-о         40     -ое  178  -ом 220  -ою  138
                            -се 57  -сь 344  -сю 65  -ся   370
                            -те 77  -ту 58  -ть 434  -тя    71
-у        122     -ую  162
-х        175
                            -ца 76  -цо 52  -цу  66  -цы  59
                            -чу 42  -чь 70
                            -шо 47  -шу  53  шь  156
-ы         51     -ые 185  -ый 254  -ым 156  -ых 147
-ь        540     -ью 116  -ья 64
-ю        298     -ют  86
-я        384     -ям  51  -ят 111  -ях 90  -яю 78
```

Одиночные буквы с большими значениями S либо сами по себе являются морфемами, либо (*-м, -х*) — хорошими диагностирующими сигналами морфем. 38 диграмм со значениями S>100 представляют собой либо одиночный аффикс, либо комбинацию аффиксов. Много «хороших» аффиксов обнаруживаются и у правых диграмм с S<100.

В левой позиции одиночные буквы демонстрируют более пеструю картину: лишь две из них интерпретируются как префиксы — *в-* S=252, *с-* =262; шесть букв морфемами не являются: *д-* =253, *з-* =168, *к-* =124, *п-* =645, *ф-* =126, *э-* =209. Потенциальные морфемы (префиксы и корни) составляют большинство у левых диграмм с S>100: *вз-* =154, *вс-* =316, *вы-* =160, *дв-* =137, *до-* =126, *за-* =197, *зн-* =112, *мн-* =182, *мо-* =218, *на-* =182, *не-* =217, *об-* =118, *по-* =491, *св-* =250, *со-* =150, *су-* =100, *эт-* =217. Левые триграммы легко интерпретируются как морфемы: *без-* =63, *бес-* =107, *воз-* =73, *вос-* =80, *люб-* =127, *пре-* =217, *при-* =282, *про-* =310, *раз-* =104.

Как видим, левые и (особенно) правые позиции в графическом слове дают нам много информации относительно потенциальных кандидатов в аффиксы. Объединение этих кандидатов в парадигмы подробно разбирается в Части 3.

Пробелы между графическими словами 0151 не единственный пример четких границ в тексте. Поэтические тексты в европейской традиции самым наглядным образом подаются в «расчлененном» виде (с разделением на строки, а иногда — и на строфы). Опыт использования этого обстоятельства в рамках ДСА представлен в [Шайкевич, 1978]

ПОЗИЦИОННЫЙ АНАЛИЗ ПОЭТИЧЕСКОЙ СТРОФЫ

Позиционным анализом мы будем называть разновидность дистрибутивно-статистического анализа текстов, при которой изучается распределение лингвистических элементов по определенным позициям в тексте. Если в качестве целого выступает предложение (или высказывание), то его позициями могут быть, например, первое слово предложения, второе слово, третье и т. д., или последнее слово, предпоследнее и т. д. Если в процессе анализа выяснится, что те или иные слова статистически часто появляются в той или иной позиции, это будет свидетельствовать о семантической структурированности выбранного целого (в данном случае — предложения или высказывания). Чем больше позиционно связанных элементов мы обнаружим, тем более структурированным следует признать участок текста, выбранный в качестве целого. Позиционный анализ обнаруживает грамматическую структуру предложения в языках с твердым порядком слов [см. Шайкевич, 1976, 360-362], однако, его применимость к большим отрезкам текста еще подлежит проверке.

Можно ожидать, что позиционный анализ даст какие-то результаты лишь в тех случаях, когда изучаемый текст (или отрезок текста) имеет устойчивую структуру. Интересно поэтому применить позиционный анализ к поэтическим текстам с четкой строфической структурой.

Удобным предметом подобного исследования является елизаветинский сонет. Материал исследования содержится в конкордансе пяти елизаветинских сонетистов: Даниеля, Дрейтона, Сидни, Спенсера, Шекспира [Donow, 1969]. Общая длина текста достаточно велика (60294 слова).

Как известно, четырнадцать строк классического елизаветинского сонета организованы рифмой в четыре строфы: три катрена (обычно с перекрестной рифмой) и рифмованное двустишие. Рифмы могут быть замкнуты в катрены (например, у Шекспира представлена схема abab-cdcd-efef-gg), а могут и скреплять катрены друг с другом (как у Спенсера, где схема рифм — abab-bcbc-cdcd-ee). Задача позиционного анализа состоит в том, чтобы обнаружить закономерности появления слов в строках сонета и, тем самым, обнаружить внутреннюю семантическую структуру.

Для частых лингвистических элементов (слов) составляется таблица распределения по строкам сонета. Примером табличной записи могут служить два слова:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Всего
ever	9	6	6	5	6	5	5	-	7	6	5	9	6	6	81
eye	39	22	19	13	35	26	22	19	35	22	25	16	18	22	333

Как видим, эти слова распределены по строкам сонета неравномерно. Если принять условный порог 4/3x, то мы обнаружим превышение порога в первой и тринадцатой строках у слова ever и в строках 1, 5, 9 у слова eye. Теперь можно попарно сопоставить строки сонета и определить число общих превышений (n). Результаты приведены в табл. 1.6. Левая нижняя часть матрицы соответствует списку 144 грамматических слов, правая верхняя - списку остальных слов (381 слов).

Таблица 1.6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		34	25	20	33	27	27	16	26	17	15	19	24	30
2	8		24	29	27	32	21	22	24	22	23	27	17	18
3	13	10		26	28	28	19	27	20	16	20	27	22	26
4	3	9	10		24	28	30	35	22	16	27	24	30	22
5	20	11	13	5		30	21	24	27	23	23	21	25	25
6	4	3	5	2	6		26	42	23	25	25	26	26	22
7	8	7	9	8	8	1		30	15	21	27	19	22	23
8	3	8	9	5	6	3	6		31	33	28	34	22	28
9	20	5	11	3	12	4	7	9		26	24	25	28	29
10	5	7	6	3	7	5	6	7	6		15	21	26	16
11	8	7	9	4	6	7	7	11	14	6		25	21	23
12	7	6	7	4	4	4	7	10	10	7	11		24	31
13	19	9	8	10	19	6	6	7	23	7	13	11		37
14	12	8	10	13	9	7	10	8	13	9	13	13	21	

Долю суммы общих превышений (m) к общему числу слов списка (N) будем считать показателем сходства двух строк. Результаты показаны на рис. 5 (грамматические слова) и на рис. 6 (остальные слова). Список имен и глаголов не обнаружил сколько-нибудь заметной структуры; напротив, список грамматических слов доказывает глубокую структурированность сонета: одни и те же слова часто появляются в строках 1, 5, 9, 13 (в меньшей степени — в строках 3 и 11). Сходны строки 13 и 14. Есть сходство у строк 4, 12 и 14. Такое распределение строк ясно указывают на то, что внутри сонета существуют некие единства по четыре строки, внутри которых проглядывают меньшие двустрочные единства. Каждое такое единство обнаруживает внутреннюю структуру: в катренах первая строка противопоставлена остальным, в двустишиях противопоставлены нечетные и четные строки. Степень структурирования может варьировать от автора к автору.<…>

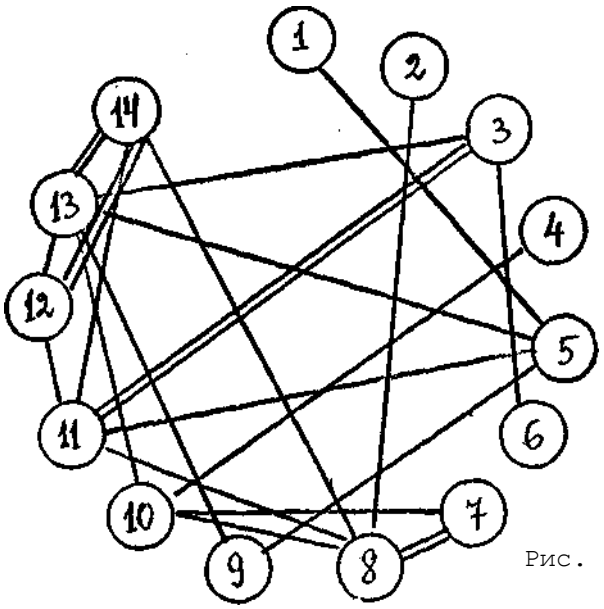

Рис. 5

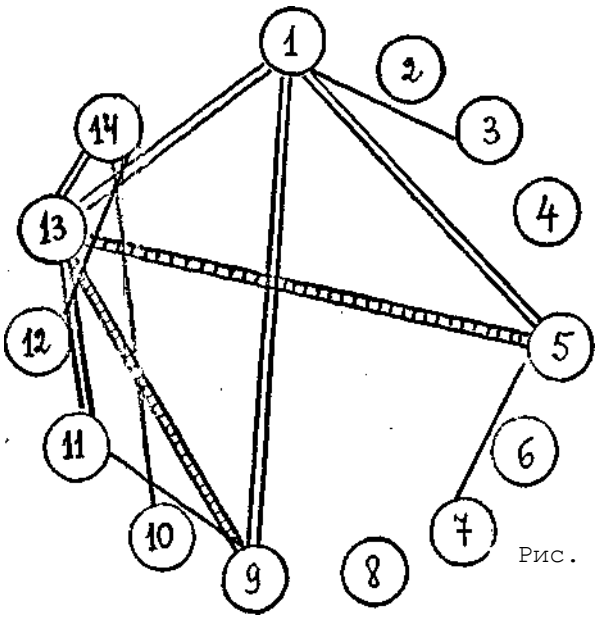

Рис. 6

Структура двустишия носит грамматический характер: нечетные строки совпадают с началом предложения, четные – с его концом:

	Нечетные	Четные
Субъектные местоимения:		
I, thou, he, she, we, they	654	375
Вопросительные слова:		
how, what, when, who	233	118
Вспомогательные глаголы:		
be, do, have	679	449

В основе структуры катрена лежит не грамматика, а сюжетный фактор. К началу катрена тяготеют следующие группы слов:

Строки катрена	1	2	3	4
I, am, me, mine, my	437	293	281	214
thou, thee, thy	123	93	72	76
Delia, Stella	30	16	10	13
ah, o, alas	82	13	19	23
cruel, fair, happy, true	125	72	70	79
love, Cupid, power, admire	208	136	101	121
see, eye, gaze, behold, hear	194	123	106	82

Итак, четверостишие обычно начинается с «расстановки сил»: поэт и его любовь — вот отправное начало. Отсюда осуществляется переход к детализации внутреннего мира и к миру внешнему:

	Строки катрена			
	1	2	3	4
aspire, bear, bend, bless, bring, brought, burn, consume, decay, die, leave, live, made, make, please, praise, prey, prove, remove, show, suffer	154	212	243	308
all, any, both, every, another, other	96	118	170	153
they, their, them	22	39	58	62
and, more, new, again	128	221	243	283
in, on, from, to	233	335	305	376

Семантическая структура катрена частично повторяется в структуре целого сонета. Об этом красноречиво свидетельствуют группы слов, представленные в табл. 1.7.

Экспозиция сонета включает поэта и его возлюбленную, а также страсть поэта (гр. 1, 2), однако, в отличие от катрена та же тема звучит и в конце сонета. Главная тема елизаветинского сонета обычно образует рамку. Как и в структуре двустишия, в начале сонета часто появляются вопросительные слова (гр. 6). Красота возлюбленной (гр. 12, 15) требует всей изобретательности поэта, чтобы найти себе достойное описание (гр. 4, 13). Альтернативная тема, начинающая сонет, — жестокость неприступной возлюбленной (гр. 5), страдания поэта (гр. 3) в настоящем и прошлом (гр. 7).

От прямого называния предмета любви идет путь к орнаментальным иносказаниям, через визуальные сравнения (гр. 16), ко всему миру (гр. 17). Описание это статично (гр. 18), иногда появляется побочная тема ложности поверхностного впечатления (гр. 19).

Третий катрен вносит новый элемент динамизма в развитие сонета (гр. 20, 22, 25). Конец его предваряет заключительное двустишие, где сила страсти достигает наивысшего накала (гр. 28), преодолевая страдания прошлого (гр. 32, 33). Появляется твердая уверенность в том или ином исходе любовных страданий (гр. 27, 31, 29, 30). Снова утверждается неизменность любви (гр. 1, 2, 23) уже как некая неоспоримая истина (гр. 31).

Таблица 1.7

	Строки						
	1 2	3 4	5 6	7 8	9 10	11 12	13 14
1. I, me, mine, my, thou, thy, thee, thine, let, o	477	293	371	278	454	322	532
2. admire, adore, heart, kiss, love, lover, passion	182	106	133	112	137	108	215
3. alas, burden, captive, care, cruel, cruelty, cry, despair, fear, grief, languish, poor, prey, sad, sigh, sorrow, starve, suffer, tear, toil, torment, tyrant, unkind, weak, weary, weep, weigh, woe, work, wound	160	135	106	151	123	132	105
4. line, muse, paper, pen, verse	48	20	28	19	29	21	27
5. cruel, cruelty, disdain, frown, pride, proud, unkind	45	25	26	36	30	35	30
6. how, what, when, where, which, who, who, whose, why	129	115	79	84	79	93	79
7. after, before, ere, last, late, long, oft, past, since, when	118	79	75	60	65	72	72
8. age, day, hour, new, old, time, young, youth	82	74	68	46	74	63	61
9. begin, begun, brought, did, found, gone, had, known, pass, saw, sought, was, were	58	50	45	42	45	39	38
10. came, come, fly, gone, run, way	39	26	25	26	32	19	23
11. brain, breast, heart, mind, soul	88	57	66	75	72	61	77
12. beauty, brow, cheek, eye, face, fair, fame, hair, hand, head, shape	193	117	174	122	157	139	122
13. as, compare, like	83	62	78	69	64	58	55
14. dwell, lay, lie, place, shore, rest, sit, stand, stay	51	47	44	50	37	43	29
15. behold, blind, eye, gaze, glass, hear, look, see, sense, sight, view	149	88	93	114	172	103	116
16. beam, black, bright, clear, colour, crystal, dark, fade, hue, light, lightning, moon, ray, red, rose, shade, shadow, sun, white	58	84	83	69	59	59	50
17. air, bud, cloud, earth, field, moon, nature, rose, shade, sky, sun, tempest, tree, wing	55	52	78	58	50	50	28
18. above, at, from, in, on, out, unto, up, upon, within	138	157	183	193	162	159	135

19. conceit, deceive, false, hide, idle	7	11	14	14	21	5	6
20. great, increase, high, more, much, so, than	80	84	100	84	121	102	137
21. because, cause, effect, therefore	7	10	11	12	19	17	20
22. become, cease, change, grow, made, make, new, turn	50	61	51	44	59	80	89
23. again, back, hold, keep, remain, rest, return, same, stay, still	52	72	56	79	80	76	99
24. away, forth, from, to, up, unto	128	146	137	147	130	182	138
25. add, and, both, either, or, other, sum	98	176	126	173	142	214	177
26. he, him, his, her, she, they, their, them	83	100	90	134	115	142	143
27. must, need, shall, will	35	38	41	38	44	62	81
28. dead, death, die, eternal, grave, immortal, kill, life, live, mortal, murder	37	42	36	53	42	66	94
29. bad, curse, ill, worse	4	11	6	8	6	16	27
30. best, better, good, perfect, praise, virtue, well, worthy	57	57	46	68	52	71	91
31. fool, idea, know, known, learn, proof, prove, reason, seem, show, teach, think, thought, true	91	93	75	87	87	76	156
32. cure, save	1	2	-	4	4	10	11
33. bless, bliss, delight, glad, happy, triumph	28	21	19	25	22	44	45

Конечно, такая структура сонета отнюдь не обязательна. Очень немногие из 33 групп, представленных в табл. 1.7, преимущественно сконцентрированы в какой-то определенной части сонета. Таковы тема жизни и смерти (гр. 28), тема счастья (гр. 33), тема избавления от страдания (гр. 32) в конце сонета; тема лжи (гр. 19) в его середине и тема поэзии (гр. 4) — в начале. В большинстве случаев превышения над средней на тех или иных участках не слишком велики (хотя и значимы статистически).

Большие колебания обнаруживаются у отдельных авторов. Так, группа «Время» (гр. 8) крайне характерна для Шекспира; у него слова этой группы распределены по частям сонета равномерно. <…>

Результаты позиционного анализа английского сонета интересно сравнить с аналогичными результатами для онегинской строфы. Онегинская строфа включает 14 строк (как и сонет) с жесткой схемой рифмовки: abeb-ccdd-effe-gg. Таким образом, и здесь мы сталкиваемся с тремя четверостишиями, за которыми следует заключительное двустишие. Однако в отличие от сонета, весьма автономного в семантическом отношении, строфа «Евгения Онегина» вплетена в ткань всего романа, а потому можно ожидать, что она не обнаружит такого грамматического и лексического единства, как сонет.

По списку грамматических слов (82 слова) связи между строками оказались очень слабыми. Это может свидетельствовать либо о том, что части строфы не связаны строго с предложением, либо о том, что само предложение не имеет твердого порядка слов. Список существительных, прилагательных и глаголов (176 слов) дал много слабых связей между строками, которые, однако, не укладываются в четкую структуру.

Четверостишие онегинской строфы обнаруживает некоторую семантическою структурированность, к его началу тяготеют такие группы слов:[1]

	Строки четверостишия			
	1	2	3	4
ГЕРОЙ (Евгений, Ленский, няня, Онегин, Таня, Татьяна)	126	62	56	56
НО (а, же)	120	57	74	63
КАКОЙ (где, как, когда, кто, ли, что)	170	124	133	121
ЭТОТ (сей, тот)	74	61	51	41

[1] Одно из слов выбирается в качестве представителя всей группы и выносится вперед, остальные слова группы приводятся в скобках.

Противоположную тенденцию показали отдельные слова и группы слов:

	Строки четверостишия			
	1	2	3	4
И	155	187	215	235
СЕБЯ (свой)	19	35	37	39
ОЧИ (глаза, голова, грудь, лицо, ножка, рука, сердце, уста)	26	44	45	51

Позиционные предпочтения выявлены и для строфы в целом. В начале строфы встречаются некоторые из тех же групп, которые предпочитают начало четверостишия (ГЕРОЙ, НО, ЭТОТ). Кроме них, в начале строфы преимущественно появляются группы КОГДА (*тогда*, *уже*, *еще*, *вновь*, *бывало*, *новый*, *старый*, *пора*), ВОТ (*теперь*, *тут*, *здесь*); ЛЮБОВЬ (*любить*, *сердечный*, *чувство*); ДУМАТЬ (*дума*, *знать*, *мысль*); БЫ (*казаться*, *ли*, *мочь*}. По краям строфы появляются преимущественно Я (*мой*) и небольшая группа МОДА (*модный*, *слава*). Напротив, в середину строфы упакована группа МУЗА (*перо*, *писать*, *стихи*). Группы И (*да*, *ни*) и КАЖДЫЙ (*везде*, *всегда*, *люди*, *мир*, *много*, *никто*, *ничто*) встречаются в центре строфы и на ее конце. В конце строфы повышает свою частоту группа слов - индикаторов чувства. СЕРДЦЕ (*глаза*, *грудь*, *душа*, *краса*, *нежный*, *ножка*, *огонь*, *очи*, *слеза*, *улыбка*, *уста*). <…>

Основной эффект позиционного анализа — выявление внутренней структуры строфы. Эта структура оказалась весьма жесткой и определенной в случае елизаветинского сонета, напротив, она еле заметна в «Евгении Онегине». Побочный эффект позиционного анализа - выделение некоторых семантических групп слов (особенно, тематических групп в сонете).

[Ученые записки Тартуского государственного университета. Linguistica X. Тарту, 1978, с. 96–113.]

1.4.4. Сходство и различие в окружениях лингвистических единиц

Дистрибутивный анализ дескриптивной лингвистики находит себе широкое применение в ДСАТ. Правда, два центральных понятия дистрибутивного анализа (контрастная дистрибуция и свободное варьирование) едва ли здесь уместны, поскольку они предполагают обращение к смыслу, на что наложен запрет в нашем формальном методе.

Хорошим примером количественного анализа дистрибутивных сходств и различий может служить комбинаторика русских букв в пределах графических слов. В отличие от первых опытов Б. В. Сухотина, работавшего над текстом без межсловных пробелов, не будем заранее ставить перед собой задачу разделения ровно на два класса (гласные и согласные). Как и в п. 3.3, устраним из текста «Войны и мира» самые частые слова (P>0,001). Тогда, зная вероятности букв и пробела между словами (#), можно подсчитать математическое ожидание сочетания букв и определить статистическую значимость (S) отклонений реальных частот от математического ожидания. Положим крайний предел для значений S (S=5) и представим результаты в следующей матрице (табл. 1.8).[1]

Сочетание *#ь* (совершенно невозможное в русском языке) получает оценку S=-5, а *ь#* — S=5, для сочетания *йо* имеем S=-2, а для *ой* S=5.

Как инструмент количественного выражения дистрибутивного сходства букв применим хорошо известную формулу коэффициента корреляции

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

Область значений r лежит между -1 и +1. Если рассматривать значения S в вышеприведенной матрице как отклонения от средней $(x - \overline{x})$, данная формула примет вид:

[1] Сюда не включены буквы #, ф, ъ, э, обладающие наименьшей вероятностью и крайне ограниченной сочетаемостью.

$$R_{xy} = \frac{\sum S_{ix} S_{iy}}{\sqrt{\sum S^2_{ix} \; S^2_{iy}}}$$

Попарные коэффициенты дистрибутивного сходства представлены в следующей матрице (табл. 1.9). В правой верхней части матрицы показано сходство букв по их правому окружению; в левой нижней части сходство определено по левому окружению. Буквы в матрице сгруппированы теперь не по алфавиту, а по дистрибутивному сходству. Цифра в клетках читается как 10R.

Таблица 1.8

Матрица сочетаний букв

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	х	ц	ч	ш	щ	ы	ь	ю	я	#
#	-5	5	5	5	5	-5	1	4	-4	-3	5	-3	2	3	-3	5		5		1	1		2		-1	-3	-5	-2	-3	
а	-5		5		5	-4	2	6	-5		1	5	3	2	-5	-2		1	3	-3	2	-1		1		-2	-3	1	3	3
б		-1		-1	4		-1		-1		1	-1	-1	2	-1	2	-2	-2	2						3	3	-1			-2
в	5	-1	-3		-2	4	-1		2	-1	-2	-1		-1	5	-2	-2		-3		-1	-1	-1	5		5	-2	-1	-1	-3
г		-1	-2	-1	1		-1	-1		-1	-1		3	-2	-1	5	-2	1	-2	-2	1	-1		-1	-1		-1	-1			-2
д	4			-5	-2	5	-2	-1	2	-1	-1	-2	-2	1	2	-2	1	-2	-2	4			-1			1				-3
е	-5		-5	1	2	-4	1	1	-5	4	-1	5	2	3	-1	-2	1	1		-3	1		1			-2	-3		-2	3
ж	3		-1	-1	3	5		-1	4			-1	-1	3	-2	-1	-1	-1	-1								-1			-2
з	5		1	1	1	-2		-1	-1	-1		-1		4	-1	-2	-1	-1	-2		-1		-1	-1			2	-1		1	-2
и	-5		1					2	-3	3	2	4	3	1	-5	-3	-3	1	2	-3	5	5	2			-2	-3		3	3
й	-2		-1	-1		-2		-1	-2		-1	-1	-1		-2	-1	-1	1		-1										5
к	5	-1	-1	-1	-2	-2	-1	-1	3	-1	-2	-1	-2		5	-2	2	-1	-1	4	-1		-1	-1		-1	-2		-1	-3
л	4	-1	-3	-2	-2	2		-2	5	-1	-1	-4	-3	-3	2	-3	-4	1	-1		-1	-1	-1	-1		1	5	4	2	2
и	2		-3	-1	-2	2	-1	-1	5	-1	-2	-2	-2	-1	1	-2	-2	-3	-3	4	-1		-1	-1		1	-2	-1	-1	5
н	5	-1	-4	-2	-1	4	-1	-2	5	-2	-2	-4	-3	1	5	-3	-4	-3	-3	2	-1	3	-1	-1		5			3	-4
о	-5	5	5	5	4	-4	3		-5	5		4	5	-2	-5	-3	2	4	2	-4		-1		1		-3		-2		
п	-1	-1	-3	-2	-2		-1	-5	-2	-1	-2	-1	-2	-3	5	-3	5	-3	-3		-1		-1	-1			-3		-1	-5
р	5	-1	-2	-1	-2	5		-1	5	-1	-2	-3	-2	-2	5	-3	-4	-3	-2	5	-1		-1			2	-1		1	-5
с	-2	-2		-2	-2	-1	-1	-2	-1	-2	5	1		-2	-1		-3	-2	5			-1		-1	-1		5		5	-5
т	3	-2	3	-2	-2	2	-1	-2	2	-2	-1	-3	-3	-2	5	-3	2	-2	-3	1	-1		-1	-1		2	5		-2	
у	-3	1		4	2	-3	3	2	-3		-1		1	-2	-4				1	-2	1		3	3	3	-1	-2	5	-1	4
х	1			-1	-1	-2			-1			1		-1	5	-1	-1	-1	-1								-1			5
ц	2				5														4					1						
ч	5	-1	-1	-1	-1	5		-1	3	-1		-1	-1		-1	-1	-2	-2		2		-1		-1			-2			
ш	5	-1	-1	-1	-1	5		-1	5		1		-1		-1		-1	-1	-1					-1			-2			
щ	1				5			5							-1		-1										-1			
ы	-2	1	3		-1			-2	5		-1	5	-1	-3	-5				-1	5			3		-1	-1		-1	2	
ь	-3	-1	-2		-2	-2	-1	-1	-3			-3	-1		-4	-2	-3		-3	-2	-1		-1	1		-1	-2	2		5
ю	-1	4	-1			-1			-1			-1	-1	-1		-1	-1	-1	1	-1					5		-1		-1	5
я	-3	1	-1			-2	1	1	-2		-1			-3	-2	-2			1	-2					2	-1	-2			5

Таблица 1.9

Коэффициенты дистрибутивного сходства букв

	а	о	у	ы	е	и	я	ю	ь	й	б	в	г	д	м	н	к	п	т	ж	ч	ш	щ	ц	л	р	з	с	х
а		+7	+5	+4	+5	+6	+5		+3	-4	-5	-3	-6	-5	-6	-5	-4	-5	-5	-5	-4	-3	-4	-5	-6					
о	+7		+6	+6	+3	+5	+5			-4	-6	-3	-7	-6	-8	-5	-3	-8		-5	-6	-3	-4	-5	-7					
у	+7	+5		+3	+5	+4	+6	+4	+3	+4		-4	-3	-8		-5	-4		-5		-5	-5	-3	-3	-5	-5			-4	
ы	+4	+6			+3	+6	+4	+3			-3		-3	-3										-3						
е	+7	+3	+6	+4		+5	+5		+4	+4		-3		-6	-4	-5	-3		-7	-3	-6	-5	-5	-4	-5	-6				
и	+9	+5	+6	+5	+7		+5		+3	+3		-4		-4	-3	-3	-6	-3	-7		-3			-4	-5					
я	+5	+5	+6	+4	+5	+5		+6	+7	+8	-3	-4	-3	-3			-5		-5	-3	-3	-4	-3	-4		-5				
ю			+4	+3			+3	+6																					+4	
ь					+3				+7							-3	-3		-3			-3				-3					
й	-5	-5	-6	-5		-5				-4	-4	-3	-4			-3			-5	-4	-5	-4	-3		-3		+3	+3		
б	-3	-4	-3	-4	-4	-4	-3				+5	+4	+4		+4	+3	+5	+5	+3	+3	+3	+3	+5		+5					
в	-4	-5		-4	-3	-4	-4					+5		+3	+6	+4	+7	+5	+4	+4	+4	+5	+4	+3	+4	+4	+6			
г	-5	-5	-4	-4	-5	-5	-3		-3	-5	+6	+4		+3	+3	+4	+6	+6	+4							+4		+3		
д	-5	-5	-4	-6	-4	-5			-4		+6	+5	+6		+4	+7	+6	+4	+8	+6	+6	+7	+6	+4	+5	+7					
м	-6	-6	-5	-5	-4	-6				+4	+5	+6	+6	+6		+5	+5		+3	+5	+4	+3	+3	+5	+6		-3			
н		-3		-3		-2		-4				+3			+3	+6	+3	+8	+7	+4	+3	+6	+8							
к	-5	-5	-3	-4	-3					+3	+5	+5	+3	-4	+5		+5	+7	-4	+4	+3			+3	+6	+3		+5		
п	-3	-4			-3						+4	+5				+6		+4	-4					+3						
т	-4	-5	-4	-5		-4					+4	+3	+3	+5		+3	+3		+4	+3			+6	+5	+3	+4					
ж	-5	-6	-4	-4	-4	-5					+6	+5	+5	+7	+6	+3		+3	+4		+5	+9	+8				+5	-3		
ч	-6	-6	-5		-3	-4	-3				+3	+4		+3	+5		+5	+4	+5	+6		+9	+7	+6		+5	+3		-3		
ш		-3			-4								+4				+3	+3		+8	+5		+5			-4				
щ					-3																		+5	+3						
ц					-4																+6									
л	-4	-5	-6	-4	-4	-6			+4	+3			+3	+6		+5		+4								+5				
р	-3		-4	-4																	+6						+4			
з	-4	-5		-4	-4	-4		-3		+4	+5	+4	+7	+6	+5	+3	+3	+6	+6	+4	+3						+3			
с	-5	-4		-4	-5	-5			-3	+6	+5	+7		+7	+4		+6	+4	+5	+3	+3									
х	-5	-5	-4	-4		-4					+5			+6		+4		+4	+3	+5	+3		+3	+3	+3		+3			

В левом верхнем углу матрицы расположилось ядро группы гласных (от А до Ю), по правому окружению к ним присоединяются Ь и Й. Буква Й оказалась среди гласных по исключительно высокой вероятности ее появления перед пробелом (ср. п. 1.3.3). Она явно не относится к гласным по своему левому окружению, где невозможны согласные, т. е. наблюдается сочетаемость, прямо противоположная сочетаемости гласных. Появление этой буквы в некоторых аффиксах (ОЙ, ЫЙ) объясняет ее некоторое сходство с М и Л (ср. ОМ, ЫЛ).

Среди согласных обратим внимание на особое поведение некоторых букв по их правым окружениям. Буква С легко сочетается с последующими согласными, а буква Х столь характерна для положения перед пробелом.

Примером классификации слов по их ближайшему окружению может служить случай русских предлогов (из корпуса литературных текстов Достоевского).[1]

Нарушим естественный ход анализа, предполагающий первоначальное обращение к микроинтервалу (ср. 1.5.1 и Часть 3) и открытие морфологии языка. Пусть нам известен только алфавит и кое-что из репертуара графических слов (например, некоторый список частых коротких слов). Сопоставим список 18 таких слов с концами (буквами и диграммами) правых соседей и снова определим статистическую значимость превышений реального числа подобных словосочетаний над математическим ожиданием. Результаты (двоичные логарифмы S) даны в табл. 1.10.

[1] В поисках грамматической системы полезно исключать из анализа крайне маловероятные события (например — с $P < 0{,}000001$). В данном случае были исключены слова с частотой 1 и 2.

Таблица 1.10

	без	в	для	до	за	из	к	на	над	о	об	от	по	под	при	про	с	у
	Финальные буквы (включая диграммы) у слов справа																	
а	3		5	3		5		3				4						4
е		7					3	5		7	4		4		4			
и	3		6	5		5		5				4					6	4
й	3	7		5	3	5	4	6	4	5	4	4	6	3	4		7	5
м		8			6			4	5	6	7		7	4	6		6	8
о		5			6	6		6		5			6		6	5	6	6
у		6			6	3		8		6				7		4		4
х	4	7	4	7	3	7		5		6	4	5			6		4	7
ы	3					3		5									3	3
ь		4						3										
ю		4			5		5	6		5			5	3		3		6
я	3		6		4	3		6				6				6		7
	Финальные буквы (без диграмм) у слов справа																	
а	4		5	3	6	4						6				3	4	5
е		7					5	6		5	3		6		5			
и	4	5	6	7	5	4		5		5		3						5
о					6			5								5		
у	3	6			6	3	7	7					7	4		4		
х								5								3		
ы	3		3	5		5		3									3	
ь					3			4										
ю							5	6	4									
я	4		5	4	3	6						6				7		7
	Финальные диграммы у слов справа																	
ам							7						9					
ах		8						9		4								3
го	6	4	7	7	5	6		6				7				5	7	7
ее	3	5	5	3	3	3		6			3	5				3	3	4
ей		6		5	5	5	7	5	5	5	4	4	5		4		7	6
ем		7			3			5	4	8	5			6			7	
ех			3	5	6	4		7		5		4			5	6	5	5
ею					5		3	4	4								7	
ие		5			3			4										
ии		7				4							4	5	4			
ий		5						5										
им					6		7		7				5	4	3		6	
их	5	6	5	7	3	7		5		5	5	6			5	3	3	7
ию		5					6						6				4	
ия	4		5	5		5						6						
ми					6												6	
му							8						7					
ов	3					4												
ое		5			4			5								3		
ой	4	6		5	4	4	6	5	4	5	4	5	4	7	7		4	
ом		8			5			6		5	8			6	7		7	
ою		5			6				4					5		3	3	7
ую		6			4			5								4		
ые		4						3										
ый		5						3										
ым					3		4		3				6				8	
ых		5	3	3		3		3		5	3	4			4		3	4
ье		5			3			5		4								
ью					4		4	3						5			7	
ья				3								3						

Одиночных букв у правых соседей оказалось слишком мало для классификации левых слов. Лишь знание русского языка позволяет заметить дистрибутивное сходство в парах В и НА (предложный падеж с финальным Е), К и ПО (дательный падеж с конечным У).

Напротив, включение диграмм уже дает некоторый намек на синтаксис, полученный в обход морфологии. Прямое использование S здесь может стать тормозом на пути поиска системных грамматических свойств; ведь какое-то частое сочетание зачастую оказывает слишком большое влияние на значение S: сочетание О *НЕМ* (f=253) обусловило связь между о и *-ем*. Свою роль играет и семантический фактор: ассоциация *без* с абстрактными существительными (*без сомнения* f=93, *без движения* f=15, *без намерения* f=9 и т. п.) вызывает связь между *без* и *-ия*. При поиске системы ключевыми становятся уже не уникальные сочетания с очень высокими S, а число самих сочетаний, хотя бы и с низкими S. Удобным инструментом для таких ситуаций может быть коэффициент связи,[1] определяемый согласно

$$Q = \frac{ad - bc}{ad + bc},$$

где a, b, c, d — количественные данные в клетках четырехклеточной таблицы. Рассмотрим в качестве примера предлоги ИЗ и ОТ из табл. 1.10 с ее 41 строкой.

		ИЗ	не-ИЗ	Итого
a	b	ОТ	11	1	12
c	d	Не-ОТ	4	25	29
		Итого	15	26	41

У предлогов *из* и *от* находим 11 общих строк (*-а, -и, -го, -ее, -ей, -ех, -их, -ия, -ой, -ых, -я*), в четырех строках заполнены клетки столбца *из* (*-у, -ы, -ии, -ов*), в одной строке заполнена клетка столбца *от* (*-ья*). При этих данных Q = +0,97. Попарные значения Q даны в табл. 1.11.[2]

Таблица 1.11

	без	для	до	из	у	от	про	с	над	под	при	об	о	на	в	за	к	по
без		6	8	9	8		8											
для	6		9	9	8	10	4	7			4	6		7				
до	8	9		9	9	10	7	8			6	9						
из	9	9	9		10	9	7	7				10		6				
у	8	8	9	10		9		7				7	6	4	4			
от		10	10	9	9			4			4							
про	8	4	7	7										9		8		
с		7	8	7	7	4			10	8		5	4		5	8		
над							10			8	5		8			7	8	5
под							8	8				4				9		
при		4	6			4		5				9	9		8			
об		6	9	10	7			5			9		9	7	9	7		
о					6			4	8		9	9			9			
на		7		6	4		9					7			8	6		
в					4			5			8	9	9	8				
за							8	8	7			7		6				
к									8									7
по									5									

В левом верхнем углу матрицы расположились предлоги, управляющие родительным падежом; в противоположном углу — два предлога с дательным падежом. Причина неопределенности в центре матрицы коренится как в омонимии флексий, так и в омонимии, проистекающей из различия одушевленных и неодушевленных существительных мужского рода.

Более интересные результаты будут получены, если оценке статистической значимости подвергнутся не любые концы правых соседей, а лишь те, которые находят себе хотя бы одного «партнера», обнаруженного при одной и той же основе. Такие партнеры начинают создавать «минипарадигмы» при данном предлоге.

---

[1] Юл Дж., Кендэл М. Теория статистики. М., 1960, с. 59.

[2] Клетки таблицы заполнены значениями 10r>3.

Приведем пример с предлогом В. Здесь обнаружено тринадцать наборов «флексий» у правого соседа, граничивших, по крайней мере, с пятью основами[1]:

-е	-у				38	(в Америке, в больнице, в воде, в глубине, в гостинице, в дороге, в душе, в Европе, в избе, в карете, в каторге, в квартире, в коляске, в конторе, в лавках, в Москве, в отставке, в половине, в стене, в темноте, в толпе, в тысяче, в тюрьме, в уголке*…)
-е	-у	-ах			10	(в виде*, в газете, в голове, в зале, в книге, в науке, в столице, в школе…)
-е	-ах				16	(в письме, в подвале, в семействе, в силе…)
-е	-ю				7	(в бане, в деревне, в земле, в келье, в кухне, в пустыне, в спальне)
-е	-о				7	(в беспамятстве, в бешенстве, в зеркале, в обществе, в Степанчикове, в числе…)
-е	-о	-ах	-а		5	(в деле, в кресле, в лице, в окне, в слове)
-и	-ь				18	(в груди, в грязи, в двери, в должности, в жизни, в задумчивости, в крови, в ночи, в постели, в части, в ярости…)
-ие	-ии				9	(в исступление, в недоумение, в отчаяние, в течение, в удивление…)
-ии	-ию				6	(в гимназии, в губернии, в России…)
-ой	-ую				6	(в левой, в столовой…)
-ой	-ую	-ом			6	(в белой, в глубокой, в маленькой, в русской, в страшной…)
-ом	-ых				5	(в главной, в черной…)

К этим тринадцати «минипарадигмам» присоединяются два набора с нулевым аффиксом:

-0	-у	10	(в год, в гроб, в лес, в сад, в ход…)
-0	-е	53	(в буфет, в воздух, в воксал, в восторг, в гнев, в голос, в город, в кабинет, в Лондон, в Мордасов, в народ, в обморок, в острог, в ответ, Павловск, в Париж, в Петербург, в свет, в Спасов, в суд, в театр, в трактир, в ужас…)

Конечно, пятнадцать «минипарадигм» образуют лишь ядра подлинных наборов, пополняемых за счет основ, не преодолевших наш порог в пять основ. К последнему набору добавляются основы: *вагон*, *дом*, *ум* (*-0, -е, -ах*), *карман*, *разговор* (*-0, -е, -ах, -ы*). Особенно характерно такое расширение для адъективных парадигм, ср. основы:

больш- (-ем, -ие, -ое, -ой, -ом, -ую)
ин- (-ой, -ом, -ую, -ые, -ых)
кажд- (-ое, -ой, -ом, -ую, -ые, -ый)
котор- (-ое, -ой, -ом, -ую, -ые, -ый, -ых).

Подобным образом можно продвигаться к морфосинтаксису языка еще до открытия морфологии в микроинтервале.

1.4.5. Совместная встречаемость лингвистических единиц

Термин «совместная встречаемость» (co-occurrence) будет пониматься здесь в более узком смысле по сравнению с тем, как это традиционно принято в лингвистике. Противопоставим этот термин термину «синтагматическая сочетаемость» (п. 1.4.2). При синтагматической сочетаемости чрезвычайно существенно взаимное расположение элементов (а иногда и их синтаксическая

[1] После набора флексий указывается число соответствующих основ. Знаком * отмечены случаи «псевдонаборов», связанные с омонимией основ или флексий.

связь). В тех случаях, когда информация о буквально взаимном расположении элементов (слева направо или наоборот) не используется, мы будем говорить о совместной встречаемости элементов. Рассмотренная в п. 1.2.3 публикация «Распределение слов в тексте…» и есть пример анализа совместной встречаемости слов.

Главная область этого подхода лежит за пределами графического слова и бинарного словосочетания, поэтому он будет обсуждаться в следующем параграфе.

1.5. Интервалы текста в дистрибутивно-статистическом анализе

Выше уже несколько раз говорилось об интервалах текста (с .8, 27, 32–38). Остановимся на этой важной составляющей ДСА. В рамках формального подхода, каким является ДСА, на том или ином этапе исследования текст членится на фрагменты равной длины.[1] Размер такого фрагмента, принятый для данного этапа, называется интервалом текста.

Главная мысль, мотивирующая необходимость понятия «интервал текста», сводится к следующему:

Обращаясь к разным интервалам текста и используя одни и те же формальные подходы (описанные в п. 1.4**), мы получаем качественно различающиеся результаты.**

В ходе проведенных исследований намечены следующие интервалы:
Микроинтервал — анализ ведется внутри графических слов.
Минимальный интервал — анализируются непосредственные сочетания графических слов.
Малый интервал — 5–10 слов текста.
Средний интервал — 40–60 слов текста.
Большой интервал — более 200 слов.
Максимальный интервал — целый текст или группы текстов внутри большого корпуса

Примеры обращения к микроинтервалу были даны в п. 1.4.1 и 1.4.3. Этот интервал дает информацию о морфологии языка, а также о фонетической системе языка в той мере, в какой она проявляется в буквенной записи. Микроинтервал подробно описывается в Части 3.

1.5.1 Минимальный интервал

Минимальный интервал — потенциальный источник информации в трех направлениях. Во-первых, он позволяет достроить морфологию и начать переход к открытию синтаксиса. Во-вторых, на этом интервале можно легко получить фразеологические единицы и тем пополнить репертуар лексем. В-третьих, этот интервал начинает открывать некоторые смысловые ассоциации слов.

В том, что касается грамматики, роль минимального интервала особенно важна. В языках без морфологии с минимального интервала и начинается ДСА. Беда в том, что для реализации поставленных задач требуются большие частотные словари словосочетаний, включающие все слова, на что традиционные конкордансы и не претендовали. В этом отношении знаменательным стало появление конкорданса Спевака к Шекспиру — печатное издание, целиком подготовленное при помощи компьютера.[Spevack, 1968] Неудивительно, что уже первый том (комедии) стал для меня полигоном для проверки ДСА в минимальном интервале.

Первые результаты были опубликованы в статье «Выделение классов слов и парадигм посредством дистрибутивно-статистического метода (на материале комедий Шекспира)» [Шайкевич, 1976], которая приводится здесь с сокращениями.

В данной статье описывается ДСА на стыке субсегментного и минимального интервалов. В качестве сегмента выбирается естественный для европейской письменности — графическое слово. Основная цель — получение парадигм и

[1] Членение на фрагменты может быть задано в самом тексте; если эти фрагменты (например, строки, строфы, главы, страницы печатного издания) не очень сильно варьируют по размеру, они могут быть использованы в ДСА.

грамматических дистрибутивных классов. По целям и методам эта статья перекликается с работой [Прицкер, 1971]. Но в данном случае задача расширена, снимаются некоторые ограничения, несколько изменена техника корреляционного анализа и выделения диагностирующих признаков.

Материал анализа — текст комедий Шекспира (276 тыс. слов текста), в котором предполагаются заданными 1) пробелы между словами и 2) границы между высказываниями (= реплики действующих лиц). Общий план работы таков: сначала формируются пробные парадигмы на основе анализа внутрисловных цепочек букв; затем изучается слов в тексте на минимальном интервале, выделяются дистрибутивные классы слов; наконец, информация о дистрибутивных классах для корректировки парадигм.

Формирование пробных парадигм

При создании пробных парадигм разумно исходить из предположения о преимущественно агглютинирующем способе соединения морфем. В этом случае мы ожидаем, что грамматические парадигмы образуются прибавлением префиксов или суффиксов к основе. Значит, и выявление парадигмы должно начинаться с выделения крайних цепочек букв, присоединяемых к одной и той же центральной цепочке букв. Тогда центральные цепочки букв выступают как аналоги аффиксов (левые цепочки аналогичны префиксам, правые — суффиксы). В дальнейшем для краткости будем использовать термин «аффикс» и «основа», помня, однако, что речь идет о буквенных цепочках, лишь предположительно относимых к комбинациям морфем.

Назовем парадигмой набор аффиксов, сочетающихся о одной я той же основой (основами). Например, аффиксы -ation и -е образуют парадигму, поскольку они сочетаются со следующими основами: accus-, admir-, convers-, examin-, imagin-, observ-, purg-, reput-.

Число аффиксов, входящих в парадигму, будем называть длиной парадигмы. Число основ, с которыми сочетается парадигма, назовем широтой парадигмы. Относительно приведенного примера длина парадигмы — 2, а широта — 8.

Если какая-либо основа, допускающая присоединение аффикса, в то же время может выступать и без всякого аффикса (т. е. совпадает со словом), будем говорить, что в этом последнем, случае основу сопровождает нулевой аффикс (0). Таким образом, возникают парадигмы, включающие нулевой аффикс, например, парадигма /0 - ly - s – ‘s/, сочетающаяся с основами: clerk, cuckold, father, heaven, hour, night, patient, scholar.

В некоторых случаях может возникать затруднение в отнесении цепочки букв к аффиксам или к основам. При решении этого вопроса будем руководствоваться двумя соображениями. 1) Аффиксы обычно короче основ (по числу букв), ср. только что приведенный пример. 2) Сочетаемость аффикса с разными основами обычно значительно шире, чем сочетаемость основы с разными аффиксами. Так, хотя цепочка vex (в слове vexation) короче цепочки -ation, именно последнюю следует признать аффиксом. Суффикс -ation сочетается с 30 разными основами, между тем как vex- сочетается лишь с 5 разными аффиксами (0, -ation, -ations, ‘d, -est). Подобные соображения заставят нас признать суффиксом цепочку -man, которая сочетается с основами: bond-, country-, foot-, French-, gentle-, good-, hang-, mad-, noble-, wo-. Тем не менее, остнутся сомнения в отношении некоторых цепочек с узкой сочетаемостью, ср. -hood (child-, false-, man-, sister-); -dom (duke-, free-, king-).

Для формирования пробных парадигм необходимо сократить объем материала, подлежащего анализу. Такое сокращение, конечно, очень важно при ручном эксперименте, но оно может оказаться полезным и в случае использования ЭВМ. Исключим из рассмотрения все слова, частота которых не попадает в зону от 10^{-5} до 10^{-3}. Для выбора этого критерия важны два мотива. Исключая редкие слова (в нашем тексте — с частотой 1 и 2), мы значительно сокращаем широту парадигм и тем самым облегчаем себе труд. Так, парадигма 0 — un- уменьшает при этом свою широту с 207 до 20. С другой стороны, исключая частые слова, мы заведомо отбрасываем важнейшие служебные слова, либо совсем не изменяемые, либо

принадлежащие к уникальным классам словоизменения.[1] Тем самым, в центре исследования остается регулярная парадигматика. Этот шаг уменьшает вероятность ложных морфемных членений (например, деления слова and на an + d) и уникальных парадигм, например, 0 — s в словах it и its).

Далее исключил из рассмотрения парадигмы с длиной > 4. На оставшемся материале получаем 19 префиксальных парадигм (все по два аффикса) и 110 суффиксальных парадигм, из которых 67 включают два аффикса. 35 — три аффикса, 8 — четыре аффикса. Большая широта суффиксальных парадигм показывает, что именно здесь сосредоточено словоизменение. Приведем парадигмы, отличающиеся наибольшей широтой:
0 — d (30), 0 — 'd — ing (34), 0 — ‘d — s (47), 0 — e (33), 0 — ‘d (125), 0 — ed (66), 0 — ing — s (73), 0 — ing (102), 0 — er (63), 0 — ly (122), 0 — es (30), 0 — s (370), 0 — 's — s (30), 0 — 's (93), 0 — y (31), 0 — a (34), ‘d — e (81), ‘d — es (38), ‘d — ing (30), ing —. s (86).

Полученные суффиксальные парадигмы сильно пересекаются друг с другом, наблюдаются и полные вхождения одной парадигмы в другую (ср. 0 — ment и 0 — ‘d — ment). Преждевременная коррекция парадигм (их объединение или разделение) была бы ошибкой, т. к. грозила бы потерей полезной информации. Так, явной ошибкой было бы простое объединение парадигм 0 — ing — s и 0 — s, поскольку первая соответствует глагольным основам, а вторая, главным образом, — именным. Их смешение лишь тормозило бы разбиение множества слов на дистрибутивные классы.

Предварительный характер пробных парадигм очевиден при качественной их оценке с позиций семантической лингвистики, т. е. обычной лингвистики, в той или иной мере учитывающей смысл. По большей части, списки основ для наших парадигм: содержат очень мало ошибок, а многие вполне безошибочны. Лишь в 28 парадигмах доля ошибок превышает 5%. Впрочем, получена и полностью ошибочная парадигма 0 — e — s с основами bar-, bid-, hat-, pin-, rag-, shin-, star-, war-, win-. Принятие подобной парадигмы в качестве окончательной вело бы к ошибочному морфемному членению и к созданию ложных аффиксов.

Вместе с тем, многие парадигмы остаются неудовлетворительными, хотя формально и не содержат морфемных ошибок. Главный их недостаток — смешение омонимичных парадигм. В качестве примера рассмотрим парадигмы 0 — en и 0 — th. Первая парадигма есть комбинация нескольких случаев: 1) глагольная основа и причастие: beat-, eat-, 2) существительное (название вещества) и прилагательное: gold-, lead-, silk-, 3) имя и отыменной глагол: quick-, 4) варианты основ: list-, maid-, oft-. Вторая парадигма также представляет собой смешение нескольких случаев: 1) количественное и порядковое числительные four-, seven-, 2) глагол и существительное: heal-, steal-, 3) инфинитив и форма 3 л. ед. ч.: choose-, please-. Совершенно ясно, что подобные парадигмы можно расщепить только тогда, когда станет известна неоднородность соответствующих основ, т. е. после того, как будут получены дистрибутивные классы слов. Впрочем, первое приближение к цели можно получить и не выходя за пределы круга слов, на основе которых выделены пробные парадигма.

Будем сравнивать друг с другом группы слов, соответствующие суффиксам парадигм. Мерой сходства или различия будет служить появление в тексте рассматриваемых групп слов рядом с наиболее частыми словами. Самые частые слова выступают здесь (и далее) как диагностирующее средство, ведущее к формированию дистрибутивных классов.

Можно полагать, что для получения хороших результатов желательно иметь, во всяком случае, не менее 50 самых частых слов. В настоящем исследовании самыми частыми считались слова с относительной частотой более 1/2000 (всего 251 слово).

Дистрибутивные свойства парадигмы выявляются при анализе текстовых окружений групп слов, соответствующих данной парадигме. Желательно полностью изучить дистрибуцию таких групп, однако ручное исследование и здесь заставило ввести ограничения. Для каждой парадигмы отбиралось не более 10 самых частых основ. Если широта парадигмы 0 — ‘s равна 93, то для дистрибутивного анализа отбиралось всего 10 основ: brother, daughter, day, fortune, god, heaven, husband, men, other, woman.

---

[1] Отнесение частых слов к регулярным парадигмам осуществляется на более позднем этапе исследования.

Параллельно создавался список основ, принадлежащих только данной парадигме (для этой же парадигмы mother, Claudio, Ford, men, mother, Page, queen, wife, woman, youth). Если такой параллельный список отличался от основного, он выделялся в особый вариант парадигмы (в данном случае — № 63а), всего выделено 17 таких вариантов.

Статистика окружений фиксировалась в таблице с 244 столбцами и 251 строкой. Столбцы соответствовали группам слов (= основа + соотв. аффикс) парадигмы, строка — упомянутым выше самым частым словам.

Каждый столбец включает четыре колонки, для четырех позиций окружения: 1) слово через одно слева, 2) слово непосредственно слева, 3) слово непосредственно оправа, 4) слово через одно справа.

В табл. 1.12 дается фрагмент рабочей таблицы 244×251 для парадигмы № 72а: 0 — s (соответствующие основы — desire, eye, hand, leg, letter, lip, spirit, thought, way, word).

Каким же образом зафиксировать сходство (или различие) дистрибуции двух групп слов (или вообще двух языковых объектов)? Классический дистрибутивный анализ мыслил категориями «все или ничего», элемент встречается в данном окружении или не встречается. Рассуждая подобным образом, можно сказать: слова с суффиксом -0 встречаются (а слова с суффиксом -s не встречаются) непосредственно перед doth, hath и непосредственно после a, an, another, one. Соответственно слова с суффиксом -s встречаются, а слова с суффиксом -0 не встречаются непосредственно после these. Такая формулировка следует из таблицы и соответствует нашему предварительному знанию английской грамматики. Но беда в том, что такая формулировка прямо обусловлена интуитивным знанием грамматики, а дистрибутивные критерии лишь подбираются для подтверждения этого знания. Ведь обычный дистрибутивист оставляет нераскрытым смысл слов «не встречается». Что означает, что слова с сусфиксом -s не встречаются после I: никогда не встречаются или не встретились в данном корпусе текстов? Следует ли придавать какое-либо значение отсутствию all перед словами с суффиксом -0? И напротив, следует ли как-то учитывать тот факт, что слово is часто встречается после слов с суффиксом -0, а слова our и their часто встречаются перед словами с суффиксом -s?

Классический дистрибутивный анализ решает такие затруднения, лишь забывая об аменталистском подходе, лишь включив в понятие позиции синтаксические соображения, которые заведомо не могут быть известны на этой первоначальной стадии исследования.

Возможность преодолеть эти трудности открывается при статистическом подходе. Надо не задавать вопрос: «встречается или не встречается?», а спрашивать: «как часто встречается?». Естественно, что эпитеты «частый» и «редкий» осмысленны только при сравнении абсолютных фактических частот с какими-то теоретическими показателями, например, со средней, с математическим ожиданием и т. п. Важны не сами частоты, а их отклонения от теоретических ожиданий. Очевидно, что частые элементы при этом играют очень важную роль, ведь их отклонения тоже очень велики. Такой подход можно назвать сильным, статистическим подходом.

Но можно предложить еще один вариант применения статистики. При слабом статистическом подходе до определенного момента мы рассуждаем, как и при сильном статистическом подходе: чем отклонения больше, тем более они значимы. Однако после определенного порога происходит как бы возврат к критерию «все или ничего». Дальнейший рост отклонений уже не увеличивает их значимости. Критерием оценки отклонения выступает наша основная формула.

При этом S округляется в сторону уменьшения до целого числа, и для него устанавливается крайний предел — ±5. Таким образом, отклонения, превышающие 5, ведут себя так же, как и отклонения при S=5.

Основное отличие слабого подхода — усиление роли более редких слов: слова another, every, one, например, имеют здесь такой же вес, как и слово a, хотя абсолютное отклонение этого последнего в несколько раз больше. Лингвистический смысл известного уравнивания частых и редких слов заключается в том, что оно направлено против одиночных синтагматических связей лексического или фразеологического характера и, напротив, поощряет регулярные синтагматические связи. Понятно, что речь вдет о вариантах одного и того же метода, отнюдь не противоречащих друг другу, но в то же время, возможно, полезных для разных этапов дистрибутивно-статистического анализа. Знак отклонения остается одним и тем же при двух подходах, а это должно приводить

к совпадению в главном. И в том, и в другом случае мы наблюдаем одни и те же окружения для идентичных позиций. Именно слабый подход отражен в табл. 1.12.

Таблица 1.12

	Слова с суффиксом -0				Слова с суффиксом -s			
	Общая частота группы слов 1221				679			
Слова окружения	Позиции окружения 1	2	3	4	1	2	3	4
1. a	+3	+5	-2	-1	-1	-4	-1	-2
6. all	-1	-2	-1			+3		
8. an	+1	+3						
9. and	-2	-2	+5	-3		-2	+2	-4
10. another		+5						
11. any		+3						
12. are	-2	-1		-1			+5	+1
14. at	+3	+2		-1	+2	-1	+1	-1
27. by	+5	-1		-1				
40. do	-1	+1	-1	-2		-1	+3	
42. doth			+2					
48. every		+5						
73. hath	-1	-1	+1					
79. her		+4		+1	+2	+5	-1	
83. his		+5				+5	-2	-1
90. I		+1		-4		-1	-5	-1
93. in	+5	+5	-3		+1	+2		
95. is	-2	-2	+3	-2	-2	-2	-1	-3
130. of	+3	-2	+5	-2	+4		+4	-3
133. one	+1	+5	-1	-1	-1	-1	-1	
136. our	-1					+5		
164. that	-2	+4	+1	-2	-1	-2	+2	-1
166. the		+5	-3	-1	-1	+2	-3	-2
168. their				-1		+5		
174. these					+3	+5		
175. this	-1	+5	-1		-2	-1		

Для группы слов с суффиксом -0 отметим следующие окружения:1) слева: a, an, one; 2) через одно слово слева: by; 3) непосредственно слева: another, any, every, that, this; 4) непосредственно справа: doth, is. Для группы слов с суффиксом -s характерны; 1) слева: these 2) непосредственно слева: all, our, their; 3) непосредственно справа: are, do. Наконец, для обеих групп характерны: 1) слева: in; 2) через одно слово слева; at, of; 3) непосредственно слева; her, his, the; 4) непосредственно справа: and, of, that.

Дистрибутивный анализ парадигм начинается с проверки сходства групп слов с идентичными суффиксами, входящими в разные парадигмы. Мерой сходства двух групп (обозначаемых, соответственно, через x и y) будем считать коэффициент корреляции:

$$R_{xy} = \frac{\sum S_{ix} S_{iy}}{\sqrt{\sum S^2_{ix} \; S^2_{iy}}}$$

В данном ручном эксперименте этот коэффициент был существенно модифицирован: не учитывались S=1 и все отрицательные S.

Проверка дистрибутивного сходства идентичных суффиксов позволяет объединить многие парадигмы. Например, у парадигм 1) 0 — er — est, 2) 0 — er — ly, 3) 0 — er, 4) 0 — est, 5) 0 — ly — r, 6) er — est, 7) er — ly — ness, 8) er — ly обнаружена высокая корреляция (r=0,5) между группами слов,

относящимися к разным парадигмам, но обладающими идентичными суффиксами. Отсюда следует только один вывод: указанные 8 парадигм необходимо объединить в одну общую парадигму[1]: 0 — er/r — (est — ly). На этом этапе могут объединиться друг с другом и разные суффиксы. Высокая корреляция суффиксов 0 и ly позволяет предположить дистрибутивное сходство суффиксов —er и —r. Проверка подтверждает это предположение. Такие дистрибутивно тождественные суффиксы становятся вариантами одного и того же члена парадигмы. При этом могут обнаружиться какие-то закономерности графических чередований. Например, из приведенного ниже списка парадигм следует, что подобные варианты неизменно включают чередование буквы e с ' и 0.

В результате объединения число суффиксальных парадигм уменьшилось до 43 (звездочкой отмечены парадигмы, где отбирались только те слова, которые не встречаются в других парадигмах):

1. 0 – able
2. 0 – age
3. 0 – ance
4. 0 – d
5. 0 – 'd – ed
6. 0 – e
7. 0 – e – es
8. 0 – ed/d (ing – s)
9. 0 – en
10. 0 — er – 'd/ed
11. 0 – er/r (est – ly)
12*. 0 – es
13. 0 – es – ing
14. 0 – ful –less – s
15. 0 – ing (er – s – 'st)
16. 0 – ion
17. 0 – ity
18*. 0 – ly
19. 0 – ly ('s – s)
20. 0 – man
21. 0 – ment ('d)
22. 0 – n
23. 0 – ness
24. 0 — or
25. 0 — ous
26. 0 — r
27*. 0 — s
28. 0 – 's (s)
29. 0 – s — t
30. 0 — st
31. 0 – t (ing)
32. 0 – th
33. 0 — ty
34. 0 – y (s)
35. 0/e — ation
36. 0/e – 'd (ing – s/es)
37. e — ity
38. e — y
39. ied – ies — y
40. ies — y
41. ily — ly
42. nce — nt
43. ncy — nt

На этом формирование пробных парадигм заканчивается.

Выделение дистрибутивных классов слов

Переходим к центральному пункту грамматического этапа ДСА — выделению дистрибутивных классов слов. Материалом служат все те же самые частые слова (всего 251 слово), но к ним добавлен в качестве особого элемента раздел между высказываниями (#); кроме того, анализ распространен на 65 групп слов, каждая из которых соответствует суффиксу одной из 43 парадигм.[2] Таким образом, получается 317 единиц (251+65+1), строки их соответствуют левым элементам, а столбцы — правым, На пересечении столбцов и строк записываются S, принимающие значения 2, 3, 4,, 5. Первая матрица соответствуют непосредственному соседству слов в тексте третьим словом.

Далее задача сводится к расчету попарных коэффициентов корреляции для всех 317 элементов, составляющих матрицу. Для любых двух слов подсчитывается сумма (S1 × S2) для четырех позиций (второе слово слева, первое слово слева, первое слово справа и второе слово справа). Иначе говоря, по обеим матрицам сравниваются две строки, а затем два соответствующих столбца. Для каждой пары

[1] В скобках даны факультативные члены парадигмы.

[2] В 43 парадигмах — 111 суффиксов, но у многих из них соответствующие группы слов обладают совокупной частотой, слишком незначительной для надежных статистических выводов. Поэтому эти суффиксы на данном этапе анализа не рассматриваются.

слов необходимо совершить 4 × 317 = 1268 сравнений. Если учесть, что число пар в матрице равно 50086, то общее число сравнений превысит 60 миллионов — цифра значительная даже для ЭВМ. Поскольку нас не будут интересовать близкие к нулю или отрицательные корреляции, появляется возможность сократить число сравнений на порядок. Тем не менее обработка матрицы оказалась очень трудоемкой (более трех месяцев работы).

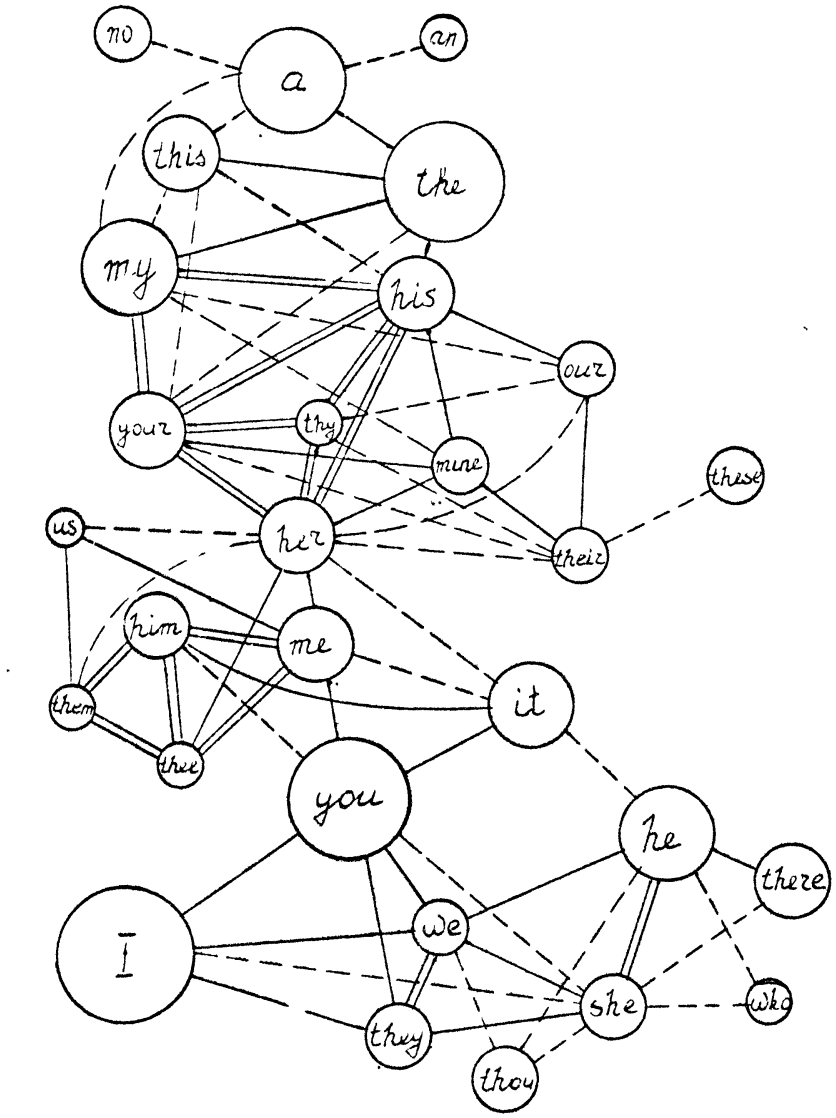

Рис. 7

Результаты представляются на графе, где ребра соответствуют положительным корреляциям между словами (r>0,3). Картина оказывается весьма яркой, легко выделяется несколько сгустков (т. е. дистрибутивных классов), хотя и остается много частных проблем при попытке разбить граф на области. О характере этих проблем можно судить по рис. 7, где приведена часть основного графа. Градация линий соответствует величине r (пунктирная линия — r>0,3, сплошная — r>0,4, двойная — r>0,5).

На рис. 7 хорошо видны две большие группы слов, соединенные одним общим словом her. Сравнительно легко заметить две области (в нижней части графа) с общим словом you, хотя граница здесь не столь очевидна из-за слабой связи между he и it. Конечно, можно было бы предложить какое-либо общее правило разбиения графа на части, но выбор того иного решения труден. При теперешних наших знаниях о подобных графах разумнее визуально анализировать рисунок, стремясь к большей дифференциации. Как мы увидим далее, ошибки чрезмерной дифференциации затем сравнительно легко устраняются, ошибки же неоправданного смешения групп исправить очень трудно.

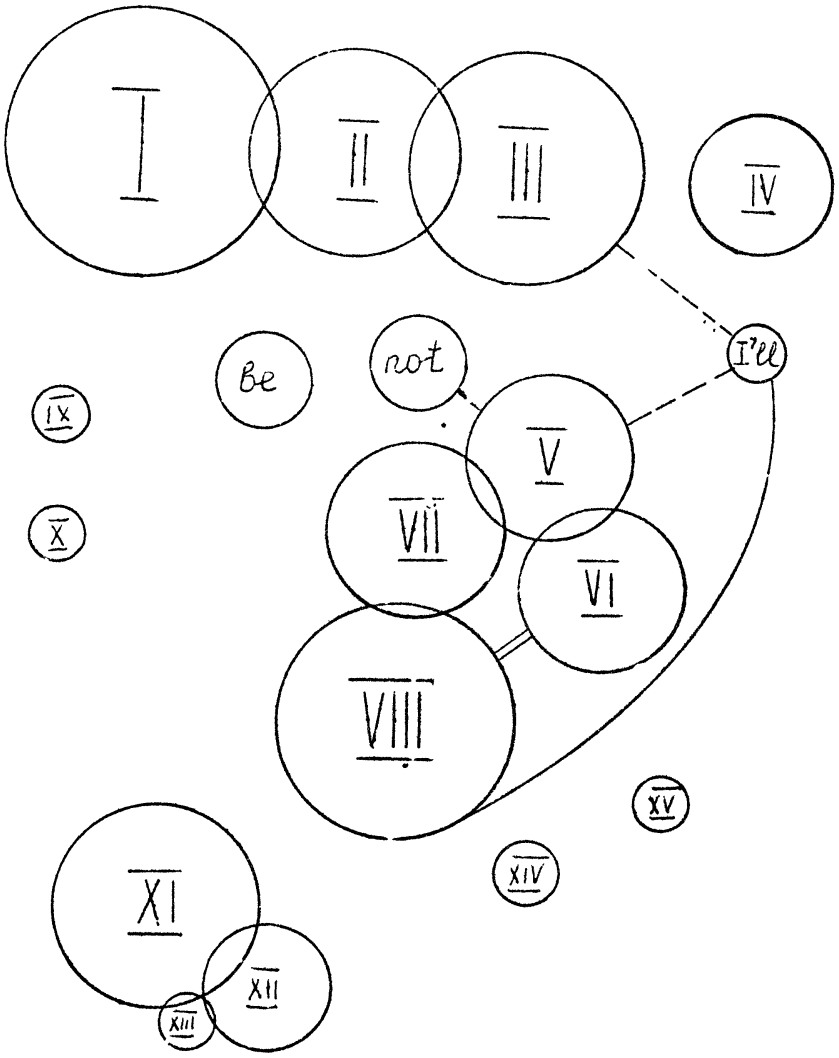

Рис. 8

На рис. 8 приводятся обобщенные результаты разбиения множества слов на дистрибутивные классы.

Класс I (F=31287): a, an, her, his, mine, my, no, our, the, their, these, this, thy, your [приименные детерминативы]

Класс II (F=17396): her, him, it, me, thee, them, us, you [объектные местоимения]

Класс III (F=21785): he, I, she, there, they, thou, we, who, you [субъектные местоимения]

Класс IV (F=11126): as, if, that, though, till, what, when, where, which, whom [союзы и союзные слова, вводящие придаточные предложения]

Класс V (F=9514): can, cannot, could, did, do, doth, may, might, shall, should, were, will, would [модальные глаголы]

Класс VI (F=12207): am, are, art, did, dost, had, has, hast. hath, have, is, was, were; а также группы слов с суффиксами -es (парадигма № 13), -s (№ 15), -'st (№ 15), -st (№ 30), [личные формы глаголов]

Класс VII (F=14705): about, after, bring, call, come, did, do, fear, find, give, go, hear, hold, keep, know, leave, let, live, look, make, marry, please, put, say, see, show, speak, stand, take, tell, thank, think, а также группы слов с суффиксами -0(№ 12),

-0(№ 14), -0 (№ 16), -0 (№ 21), -0 (№ 32)
-0 (№ 35), -0 (; 36), –e (№ 36), –e (№ 7)б
–y (№ 39),[по преимуществу глаголы в основной форме]

Класс VIII (F=25459): about, after, against, at before, by, for, from, in, into, of, on, to, up, upon, with [предлоги]

Класс IX (F=1295): been, done, made, и группы слов с суффиксами: -‘d (№ 36), –ied (№ 39), –n (№ 22)[пассивные причастия]

Класс X (F=630): long, much и группа слов с суффиксом –y (№ 38)

Класс XI (F=17582): best, boy, brother, daughter, day, duke, eye, eyes, face, fair, faith, father, fool, friend, gentleman, grace, hand, head, heart, honour, house, husband, king, lady, life, little, lord, love, man, master, mind, mistress, name, night, own, part, place, son, thing, time, tongue, wife, wit, woman, word, world; группы слов с суффиксами:-0 (№ 27), -0 (№ 28), -0 (№ 31), -0 (№ 34), –ance (№ 3), –ation (№ 35), -d (№ 4), -e (№ 9),–es (№ 9), ies (№ 40), –ing (№ 12),–ing (№ 15), –ion (№ 16), –man (№ 20),–nce (№ 43), -ness (№ 23), –or (№ 24), -r (№ 26),–‘s (№ 28), –s (№ 27), –s (№ 28), –s (№ 34), –t (№ 29)[имена – преимущественно существительные]

Класс XII (F=6164): dear, first, gentle, good, most, old, poor, sweet, true, young; группы слов с суффиксами -0 (№ 11), -0 (№ 23), –able (№ 1), –est (№ 11), –ful (№ 14), –nt (№ 43) [прилагательные]

Класс XIII (F=1613): better, great, more и слова с суффиксом –er (№ 11) [компаративы]

Класс XIV (F=866): any, some

Класс XV (F=404): three, two

Краткая содержательная характеристика классов дана для удобства читателей, чтобы показать осмысленность результатов. Дешифровочный характер процедуры пока не дает еще оснований для полной семантической интерпретации. Попарная корреляция и ее результат — 15 дистрибутивных классов — только начало дистрибутивной классификации слов. Наши классы построены на очень недифференцированном основании — на учете только самых частых слов, среди которых довольно много бесполезных, вносящих лишь элемент случайности. С семантической точки зрения видны и некоторые ошибки (некоторые предлоги среди глаголов, great среди компаративов, best, fair, little, own — среди существительных). Наконец, дистрибутивные классы охватывают далеко не все слова (правда, они составляют 60% всех слов в тексте). Все это заставляет считать полученные группы слов только ядрами дистрибутивных классов, которые должны корректироваться и расширяться при помощи ряда приемов.

Дальнейший шаг процедуры — выделение для каждого класса его собственных диагностирующих признаков, которые позволят количественно определить отнесенность любого слова (не слишком редкого) к тому или иному классу.

Искомые коэффициенты для каждого класса (индексы классов) должны быть построены по общему принципу:

$$G_x = \frac{\sum k_i f_i - \sum k_j f_j}{f_x}$$

иными словами, для каждого класса выделяются положительные (i) и отрицательные (j) диагностирующие признаки (ДП) с их весами (k_i и k_j соответственно), а затем для любого слова (или группы слов) x производится суммирование ДП, появившихся в тексте рядом с x, и полученная сумма делится на частоту x.

Но каким образом отобрать ДП и определить их вес? Как и прежде, основной фактор — расхождение фактических частот слов (и групп слов) и математического

ожидания. Например, общая частота класса XI составляет 17582, а слов класса I — 31287. Общая длина текста 276 тыс. слов и 12 тыс. разделов между высказываниями. При независимости классов I и XI математическое ожидание слов класса I рядом со словами класса XI (17582 × 31287)/(276000+1200) = 1910. Фактические же частоты таковы: второе слово слева = 1730; первое слово слева = 6968; первое слово справа = 824; второе слово справа = 1268. Мы видим, таким образом, концентрацию класса I непосредственно слева от класса XI (d = f — m = 6968—1910 = 5058); наоборот, справа от класса XI слова класса I встречаются редко. Следовательно, слово класса I является положительным ДП для класса XI , если оно встречается непосредственно слева, и отрицательным признаком, если оно встречается справа. Естественно, что при этом должна учитываться значимость наблюдаемых отклонений (S).

Наконец, полезно учесть еще один момент. Вес диагностирующего признака должен быть тем больше, чем больше отношение данного отклонения ко всей частоте признака во всей выборке. Тогда основа веса ДП (K) определяется согласно формуле:

$$K_i = S \times d \times d/F_i = Sd^2/F_i$$

Чтобы ограничить число ДП, отобрав самые эффективные, введем три порога K. K_i должно удовлетворять неравенствам:

1) $K_i > 0{,}001n$ (n -- частота рассматриваемого класса);
2) $K_i > 0{,}01K_{max}$.

Оба эти порога отбрасывают множество слабых признаков. Для индивидуальных слов, которые входят в состав какой-либо группы, полезно ввести еще один порог в зависимости от веса этой группы ($K_{гр}$):

3) $K_i > K_{гр} \times F_i/F_{гр} - 0{,}1\ K_{гр}$

Окончательные значения k_i, k_j пропорциональны K_i, K_j, а количественно подбираются так, чтобы индекс какого-либо класса (G) был равен 100 для всей совокупности слов данного класса и был равен 0 для всей совокупности. Таким образом, на этом этапе исследования мы снова возвращаемся к сильному статистическому подходу.

Предполагается, что процедура взвешивания ДП применяется многократно, и после каждого применения состав класса (или его ядра) меняется: отбрасываются слова, получившие низкие G по данному классу, и добавляются слова с высоким G по данному классу.<…>

Рассмотрим результаты корректировки классов на основе ДП (после второго шага процедуры). Для краткости ограничимся значениями k только для положительных признаков (столбцы соответствуют четырем позициям, как и раньше).

Класс I

ДП	Позиции				ДП	Позиции			
Кл. VIII	—	100	—	—	#	—	—	—	85
Кл. XI	—	—	240	—	all	—	80	—	-
Кл. XII	—	—	160		of	—	150	—	-

ДП этого класса почти полностью совпали с ДП классов XIV и XV. Тем самым, последние могут быть включены в класс I. Ни одно из слов этого класса не получило даже среднего значения по другим классам; исключение представляет собой слово her (как и следовало ожидать!): $G_I = 84$, $G_{II} = 61$ Это свидетельствует о четких границах рассматриваемого класса. К нему примкнули новые слова: another, each, every, ‘s, such, th’, thine, those.

Класс II

ДП	Позиции				ДП	Позиции			
Кл. I	—	—	—	6	give		750	—	-
Кл. II	4	—	—	4	go	130	—	—	-
Кл. V	55	—	—	—	hath	130	—	—	-
Кл. VII	—	90	—	—	have	70	70	—	-

Кл. VIII	—	25	—	—	hear	—	95	—	-
Кл. IX	—	50	—	—	I’ll	350	—	—	-
#	—	—	25	—	let	—	1400	—	-
bring	—	600	—	—	make	—	300	—	-
call	—	550	—	—	see	—	200	—	-
did	110	—	—	—	tell	—	800	—	-
for	—	50	—	—	will	160	—	—	-
from	—	95	—	—	with	—	180	—	-

Из первоначального состава класса три слова вошли одновременно в другие классы, это her, you (G_{II} = 114, G_{III} = 05)и it (G_{II} = 65, G_{III} = 60). К данному классу присоединилось ‘em.

Класс III

ДП	Позиции				ДП	Позиции			
Кл. IV	—	70	—	—	have	—	—	110	-
Кл. V	—	10	170	—	is	20	20	18	-
Кл. VI	—	—	120	—	not	—	—	—	160
Кл. VII	—	14	60	60	pray	—	—	400	-
Кл. IX	—	—	50	50	think	—	85	85	-
#	35	35	—	—	what	50	—	—	-
art	220	70	170	—	will	—	—	260	-
be	—	—	—	12	would	—	—	300	-

Одно слово покинуло этот класс (who), и одно – присоединилось ‘a.

Класс IV

ДП	Позиции				ДП	Позиции			
Кл. III	—	—	85	—	have	—	—	—	70
Кл. V	—	—	90	90	he	—	—	140	-
Кл VI	—	—	120	120	I	—	—	50	-
Кл. X	—	130	130	—	is	—	—	200	200
#	—	25	—	—	it	—	—	130	-
am	—	—	—	130	know	65	65	—	-
as	250	—	—	250	me	—	30	—	-
be	—	—	—	60	thou	—	—	70	-
but	—	55	—	—	was	—	—	80	80
do	—	—	—	150	would	—	—	—	75
ever	—	—	400	-					

Класс несколько расширяется: ere, how, indeed, lest, methinks, since, unless, yet.

Класс V

ДП	Позиции				ДП	Позиции			
Кл. II	—	—	—	14	it	—	40	—	-
Кл. III	—	130	—	—	never	—	85	85	-
Кл. IV	25	25	—	—	not	—	—	40	-
be	—	—	240	35	you	—	65	—	-

В этот класс перешло (из кл. VI) и добавились I’ll, you’ll.

Класс VI

ДП	Позиции				ДП	Позиции			
Кл. I	—	—	5	—	it	—	200	—	-
Кл. III	—	100	—	—	like	—	—	65	65
Кл. IV	85	85	—	—	no	—	—	55	-
Кл. VIII	—	—	—	9	not	—	—	60	-
Кл. IX	—	—	220	—	she	—	180	—	-
#	25	—	—	—	this	—	75	—	-

a	—	—	60	—	thou	—	95	95	-
he	—	160	—	-					

В этом классе сосредоточились, во-первых, глагольные форм 2 и 3 л. ед. ч. и, во-вторых, глаголы, вводящие пассивные причастия. Класс неустойчив, шаги процедуры меняют его состав. Потеряны am, are и have, присоединились – being, comes, didst, does, doth, goes, having, knows, lies, lives, loves, speaks, stands.

Класс VII

ДП	Позиции				ДП	Позиции			
Кл. II	—	—	75	—	not	—	50	50	-
Кл. III	80	80	20	—	shall	—	110	—	-
Кл. V	—	120	—	—	should	—	95	—	-
Кл. VIII	—	—	25	—	sir	—	—	80	-
#	35	35	—	—	to	—	120	—	-
he	18	—	—	—	well	—	—	100	-
I	—	110	—	—	will	—	110	—	-
may	—	80	—	—	would	—	120	—	-
must	—	350	—	—	you	130	130	130	130

В этот класс вошли be, do, have, но основной рост происходит за счет более редких слов: answer, appear, attend, bear, begin, grow, heard, lay, lie, meab, said, spoke, thought и др.

Перечисленные три класса характеризуются сильным пересечением. В орбиту глагольных классов втянуты также классы VIII и IX. Вот важнейшие области пересечения классов:

V–VI: canst, shalt, wilt$ did, had, were;
V–VI–VII: came, went;
V–VI–VII–VIII: knew;
V–VII: dare, lov’d, needs, prithee;
V–VII–VIII: met, saw, took;
VI–VIII: gives, makes, says;
VII–VIII: writ.

Смешение классов, по-видимому, свидетельствует о существовании большого глагольного класса, объединяющего классы V, VI и VII. Другой характер имеет пересечение классов VII и VIII. Уже первый шаг процедуры обнаружил большую группу слов с высокими G_{VII} и G_{VIII}: gave, hold, keep, let, make, please, see, show, take. Второй шаг процедуры присоединил еще: ask, beat, become, brought, get, help, kiss, lead, pardon, put, told. Если не создавать глагольный сверхкласс, эту группу слов (содержательно – переходные глаголы) можно выделить и автономный класс VIIa.

Класс VIII

ДП	Позиции				ДП	Позиции			
Кл. I	—	—	160	—	his	—	—	130	-
Кл. II	—	—	65	—	love	—	120	120	-
Кл. VI	40	—	—	—	me	—	190	190	-
Кл. VII	—	75	—	—	no	8	—	—	-
Кл. X	100	100	—	—	out	—	700	—	-
#	—	—	—	14	own	—	—	—	600
all	—	—	220	—	see	—	—	120	-
be	12	—	—	—	speak	—	200	—	-
come	—	140	—	—	thee	—	200	200	-
go	—	150	—	—	them	—	200	200	-
his	—	—	200	—	us	—	—	100	-

Очень устойчивый класс, постепенно расширяющийся за счет относительно редких слов: a', above, between, here's, near, o'er, over, past, there's, through, till, under, unto, within, without.

Класс IX

ДП	Позиции				ДП	Позиции			
Кл. II	—	—	9	—	has	—	240	—	-
Кл. III	10	—	—	—	hast	150	150	—	-
Кл. V	25	—	—	—	hath	300	300	—	-
Кл. VI	—	45	—	—	have	45	45	—	-
Кл. IX	—	30	30	—	he	30	30	—	-
be	—	100	—	—	me	—	—	18	-
had	400	400	—	—	she	50	—	—	-

Этот класс открыт для присоединения новых слов: call’d, forsworn, found, given, got, known, lost, married, seen, sent, sworn, taken. Отмечено слово, обладающее одновременно высоким G_{VIII} (left).

Класс X

ДП	Позиции				ДП	Позиции			
Кл. IV	—	20	20	—	more	—	—	45	-
Кл. VI	18	—	—	—	so	—	200	—	-
Кл. VIII	—	—	5	6	thus	—	110	—	-
Кл. XIII	—	—	80	80	to	—	—	14	14
as	—	120	120	—	too	—	110	—	-
how	—	70	—	-					

Как видно по диагностирующим признакам, слова этого класса выражают разную степень интенсивности. При корректировке добавились слова far, many и often.

Из следующих трех классов только класс XIII характеризуется вполне четкими границами.

Класс XIII

ДП	Позиции				ДП	Позиции			
Кл.X	35	35	—	—	no	—	75	—	-
such	55	55	—	—	than	—	—	500	500

Корректировка расширила этот класс за счет слов.further, less, longer, worse. Слово great отпало.

Что касается классов XI и XII, то попытка полностью разделить их не удается. Как правило, если слово обладает высоким G_{XI}, то и G_{XII} у него высок. Это указывает на родство двух классов. Чтобы добиться их разделения, можно предложить следующий прием. В качестве генеральной совокупности рассматриваются не все слова текста, а только множество слов, относящихся к классам XI и XII.Вытекающее отсюда изменение математического ожидания и отклонений приводит к совершенно новым ДП. Назовем первоначальные индексы G-0, а индексы, полученные на основе ДП, — G-1. Окончательные результаты как усреднение обоих индексов.

Класс XI

ДП	Позиции G-0				ДП	Позиции G-1			
Кл. I	—	200	—	—	Кл. I	300	65	—	-
Кл. VIII	18	—	12	—	Кл. III	—	—	130	-
Кл. XII	—	55	—	—	Кл. V	—	—	220	-
#	—	—	80	12	Кл. VI	—	—	500	-
and	—	—	12	—	Кл. VIII	—	—	180	-
good	—	220	—	—	Кл. XII	—	25	—	-
in	50	—	—	—	#	—	—	500	-
my	—	300	—	—	good	—	350	—	-
of	—	—	60	—	his	350	350	—	-
your	—	100	—	—	I	—	280	280	-

is	—	—	60	-
it	—	—	350	-
my	200	130	—	-

Двойная корректировка сделала этот класс идеальным с точки зрения семантической лингвистики. Отброшены слова best, fair, little, own (они перешли в класс XII). Среди многочисленных добавлений (более 120) нет ни одной ошибки.

Класс XII

	G			
ДП	Позиции			
Кл.I	—	120	—	—
Кл. XI	—	—	50	—
#	20	20	—	20
a	—	350	—	—
an	—	100	—	—
as	—	50	50	—
friend	—	—	550	—
lady	—	—	300	—
lord	—	—	200	200
man	—	—	190	—
master	—	—	220	—
most	—	260	—	—
o	140	140	—	—
of	25	—	—	25
very	—	350	—	—

	G			
ДП	Позиции			
Кл. II	—	100	—	-
Кл. IV	—	—	—	70
Кл. VI	65	200	—	-
Кл. VII	—	90	—	-
Кл. VIII	—	—	—	35
Кл. XI	—	—	400	-
a	—	45	—	-
am	140	140	—	-
an	—	200	—	-
and	—	—	—	190
are	280	280	—	-
as	—	200	—	-
be	—	240	—	-
it	—	300	—	-
lord	—	—	400	400
man	—	—	80	80
most	—	650	—	-
nay	500	500	—	-
o	85	85	—	-
of	—	—	—	240
sir	—	—	120	-
so	—	260	—	-
‘tis	—	300	—	-
very	—	220	—	-

Этот класс пополнился пятью частыми словами: best, fair, little, own, very и более чем тридцатью относительно редкими: black, dead, free, high, mad, proud, strong, wise и др. Замечена лишь одна ошибка — включение слова ass.

Итак, процесс корректировки исходных дистрибутивных классов слов закончен. Число классов уменьшилось до 13. Все они хорошо интерпретируются семантически. Конечно, и этот список классов нельзя считать окончательным. Вне классов осталась довольно большая группа частых слов (49 из 251), охватывающая, в частности, союзы (and, but, nor, or и т. п.), наречия (again, away, down, ever, here, out, still, well и т. п.), слова-обращения (madam, o, sir) и некоторые другие (all, even, god, myself, not, so и т. д.). Среди менее частых слов довольно велика доля имен собственных, также оставшихся вне классов. По-видимому, к недифференцированной части полезно снова применить первоначальную процедуру, направленную на выявление попарных корреляций. При этом число признаков окружения можно уменьшить, используя уже полученную информацию о дистрибутивных классах.

Суммарные статистические характеристики дистрибутивных классов приведены в табл. 1.13 (P — относительная частота).

Таблица 1.13

	Зона частых слов (P>0,0005)		Зона слов средней частоты (0,00001 < P < 0,0005)		Всего	
	Разных слов	Общая частота	Разных слов	Общая частота	Разных слов	Общая частота
ВСЕГО	251	183.582	473	36.738	724	220.320
I	21	33.632	5	598	26	34.230
II	8	17.396	1	38	9	17.434
III	9	24.371	1	44	10	24.415
IV	18	14.964	4	281	22	15.245
V	16	10.084	14	843	30	10.927
VI	14	8.183	19	1.112	33	9.295
VII	37	16.207	92	6.188	129	22.395
VIII	19	25.932	13	1.039	32	26.971
IX	3	690	15	887	18	1.557
X	3	733	3	221	6	954
XI	41	9.415	137	8.271	178	17.686
XII	16	4.691	39	2.626	55	7.317
XIII	3	1.141	3	82	6	1.223

Как видим, наши классы распадаются на две группы по их поведению в двух частотных зонах. Первая группа объединяет классы I–IV и VIII. Доля частоты второй зоны к общей частоте колеблется здесь от 0,2% (кл. II, III) до 4,4% (кл. VIII). Отметим, что именно эти классы не связаны с пробными парадигмами. Отсюда может быть только один вывод: перед нами закрытые или почти закрытые классы.

Начиная с класса V (9,3%), доля частот второй зоны нарастает, достигая максимума в классах XI и XII. Итак, классы V–VII и IX–XIII относятся к открытым.

Все предыдущее построение основано на отклонении фактических частот от математического ожидания. Теперь можно выделить экстремальные отклонения. Если положительное отклонение в какое-то ситуации сохраняет высокую значимость (S=5) при удвоении математического ожидания, будем говорить, что эта ситуация поощряется языком. Если отрицательное отклонение сохраняет S=5 при сокращении m вдвое, будем говорить, что на эту ситуацию наложен запрет. Система запретов и поощрений для позиционной сочетаемости дистрибутивных классов представлена в следующей матрице:[1]

	I	II	III	IV	V	VI	VII	VIII	IX	XI	XII	#	and	not
I	-/-	-/-	-/-	—	—			-/-		+	+	—	—	-/-
II		—				—		/-				+		
III	—	-/-	—	—	+/-	+/-	+	—	+				-/-	/+
IV		-/-	+		/+	+		—				—	-	
V		/+				-/-	+	—					—	+
VI	+	—			—	—			+	—	—	—	—	+
VII		+								-				
VIII	+	/-	-/-		-/-	-/-	/-		—			—	-/-	-/-
IX														
XI			/-							—	—	—	+	
XII	—	—					-/-			+				-
#		—								-				
and		—	—					-/-				—	-	
not													-/-	

Позиционные связи классов отражены на рис. 9. Стрелками здесь показаны превышения фактической частоты еад математическим ожиданием, двойной стрелкой — поощрения, усиленной двойной стрелкой — поощрения при одновременном запрете обратного порядка слов. Стрелки ведут от слова, стоящего в тексте слова, к

[1] Классы IX и XIII не показали запретов и поощрений и из матрицы исключены. Добавлены раздел между высказываниями и два самых частых слова из числа тех, что не вошли в классы. Знак / предшествует запретам и поощрениям на позиционную близость через одно слово.

слову справа. Этот ориентированный граф вместе с матрицей запретов и поощрений предназначен для выявления всевозможных нерегулярностей как непосредственно в тексте, так и в обобщенных таблицах сочетаемости.

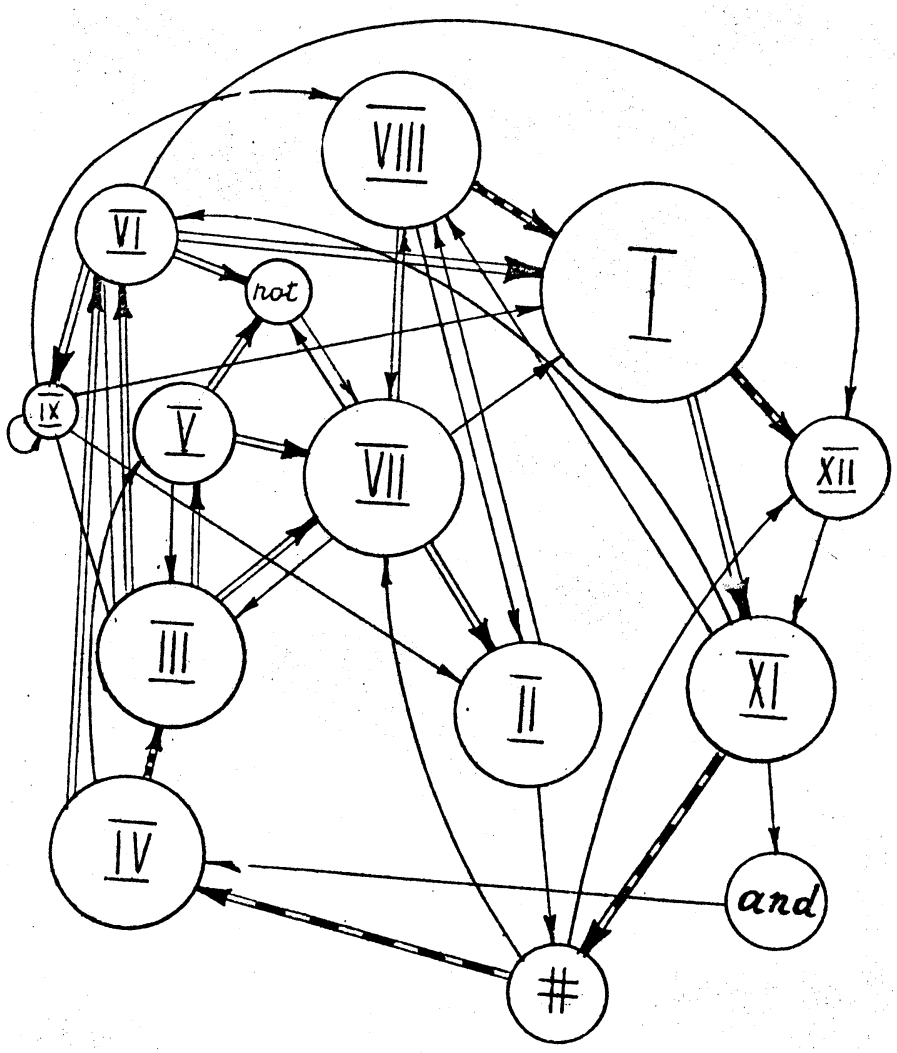

Рис. 9

Предположим, мы анализируем сочетаемость слова art, которое относится к классу VI. В матрице запретов и поощрений для класса VI зафиксировано несколько запретов, в частности, запрет на слово того же класса справа, запрет на слово класса VIII слева (непосредственно слева и через одно слово). Между тем, в таблице сочетаемости этого слова отмечено два раза слово is справа, два раза слово in слева через одно слово. Эти отклонения — сигнал для начала поиска в тексте. Каждый случай появления данного слова в тексте оценивается на принадлежность его к тому или иному классу.[1] Такая проверка приводит к двум возможным результатам: либо в тексте создаются новые границы, либо анализируемое слово расщепляется. В таких контекстах, как

> I have with such provision in mine art
> So safely ordered that there is no soul
> (The Tempest, 1.2.26)

слово art попадает в класс XI (ДП in mine слева). Расщепление art приводит к конструированию нового слова art-2, омонимичного исходному art-1. Проверка

[1] Два несоответствия, давшие повод для проверки, согласно графу связей могут указывать только на слово класса XI; ведь только слово класса XI может одновременно предшествовать слову is (класс VI) и следовать через одно слово за in (класс VIII).

всех контекстов слова art показала, что из 34 случаев употреблений art как существительного процедура дала правильный ответ (т. е. отнесла слово art к классу XI) в 29 случаях. Расчет индексов классов для редких слов позволил бы правильно разрешить все случаи.

В контексте

Thought I ____ thy spirits were stronger than thy shame,
Myself would, ____ on the rearward of reproaches,
Strike at thy life/ Griev'd I ____ I had but one?
(Much Ado about Nothing, 4.1.125)

отмечаются три несоответствия графу связей классов (подчеркнуты). Одно из них приходится на запятую, для двух других приходится создавать новую границу в тексте.[1]

Систематизация парадигм

Мы оставили суффиксальные парадигмы в тот момент, когда их число сократилось до 43. Информация о дистрибутивных классах используется теперь для дальнейшего сокращения числа парадигм и для их коррекции.

Для каждой группы слов с определенным суффиксом, относящимся к данной парадигме, устанавливается принадлежность к какому-то дистрибутивному классу. Вот, например, результат этой операции для парадигм 12, 27 и 40:

№ 12	-0	Кл. XI	-es	Кл. XI
№ 27	-0	Кл. XI	-s	Кл. XI
№ 40	-y	Кл. XI	-ies	Кл. XI

Таким образом, обнаруживается идентичность дистрибутивных классов и сходство формы правых членов парадигм. Мы уже видели, что 0 и e чередуются друг с другом, теперь же мы обнаружили еще одну графическую альтернацию. Дистрибутивное тождество суффиксов позволяет перевести -y и -i в основу, допустив альтернацию y/i (y в конце слова, i — перед e). Все три парадигмы объединяются в одну.

Сильная дистрибутивная неоднородность обоих членов двучленной парадигмы говорит о фиктивности данной парадигмы. Так, в парадигме № 4 0 — d не удалось установить принадлежность к какому-либо дистрибутивному классу 0, ни для группы, ни для группы d. Отдельные слова, входящие в эти группы, также не создают однородной картины. В группе с суффиксом -0 представлены классы: VII: bear, die, hate, woo; XI: age, mote, war; не опознаны слова amen, fee, ten. В группе с суффиксом d представлены: кл. VII: amend, feed, tend; кл. IX: died, hated, noted; кл. XI: beard, wood; кл. XII: aged, не опознано ward. Отсюда следует вывод: парадигма фиктивна.

Таким же образом доказывается фиктивность парадигм № 6 0 — e, № 9 0 — en, № 26 0 — r, № 29 o — s — t, № 30 0 — st, № 31 0 — t, № 32 0 — th, № 38 e — y, № 43 ncy — nt. Все эти парадигмы отбрасываются. Правда, в некоторых случаях (№ 9, 32, 43) можно подозревать, что отрицательный результат объясняется лишь ограниченностью материала. Во всяком случае, правила процедуры должны предусматривать возможность расщепления пробной парадигмы на две. Так произошло с парадигмой № 14.[2] Выяснилось, что за ней скрываются две парадигмы: 0 (кл. VII) — ful (кл. XII) и 0 (кл. XI) — less (кл. XII). Неоднородность одного члена парадигмы еще не дает основания для разрушения всей парадигмы.

Слабая неоднородность групп слов, входящих в парадигму, ведет к удалению некоторых основ. Так в парадигме 0 (кл. XII) — ty (кл. XI) должна быть отброшена основа wit-, поскольку слово wit относится к кл. XI, а слово witty — к кл. XII, т. е. прямо противоположно тому, что мы ожидали бы от соответствующих суффиксов.

[1] Напомним, что знаки препинания не вводились в материал анализа.
[2] Впрочем, больший по объему корпус мог бы и сохранить эту парадигму.

Одновременно группы слов парадигмы могут быть пополнены за счет частых основ. Например, парадигма № 0 (кл. VII) — n (кл. IX) расширяется основами give-, know-, see-, take-, поскольку с нулевым суффиксом эти основы попадают в кл. VII, а с суффиксом -n — в кл. IX, и, следовательно, ничем не отличаются от основ анализируемой парадигмы.

Ниже приводится итоговый список парадигм (в новой нумерации). При суффиксах указаны их дистрибутивные классы (если их удалось выявить). Для каждой парадигмы приводятся характерные основы. Факультативные члены парадигм даны в квадратных скобках.

1. 0 (VII) – s (VI) – ‘st (VI) – ing – [er (XI)]:
add, approach, bear, call, catch, come, fear, follow, give, hang, keep, kiss, know, live, look, make, mean, pass, play, say, seem, show, sleep, speak, stand, stay, teach, tell, think, touch, wish;
2. 0 (VII) – s (VI) — ing – ‘d (IX):
approve, ask, call, carry, cry, cry, curse, decisive, deny, enforce, follow, hang, kill, look, move, promise, remove, show, study, thank, turn, wrong;
3. 0 (VII) – ance (XI):
acquaint, assist, deliver, import, pen, suffer,temper;
4. 0 (VII) – ation (XI):
commend, damn, expect, habit, lament, protest, tempt, vex, visit;
5. 0 (VII) – ed (IX):
add, commend, follow, paint;
6. 0 (VII) – ful (XII):
care, fear, hate, need, thank, wonder;
7. 0 (VII) – ion (XI):
act, confess, corrupt, infect, instruct, pass, possess, profess, quest;
8. 0 (VII) – ment (XI):
agree, amaze, appoint, banish, employ, entertain, govern, pay, punish;
9. 0 (VII) – n (IX):
broke, draw, give, know, ope, see, show, take, throw;
10. 0 – able (XII):
comfort, commend, damn, honour, lament, reason;
11. 0 – age (XI):
bond, cozen, mess, pass, person;
12. 0 – or (XI):
confess, credit, debt, govern, suit, tail;
13. 0 (XI) – ‘s (XII) – s (XI) – [ly (XII)]:
brother, daughter, day, ear, earth, fortune, friend, god, grace, heaven, hour, husband, king, night, part, world;
14. 0 (XI) – s (XI):
ass, blush, business, company, courtesy, day, dish, enemy, eye, fancy, folly, friend, glass, hand, heart, hour, inch, jealousy, lady, leg, letter, lip, master, mistress, oath, pain, part, quality, speech, spirit, thing, time, way, wench, wit, witch, word;
15. 0 (XI) – less (XII)
boot, guilt, hap, harm, peer, sense, shape, speech
16. 0 (XI) – ous (XII):
clamor, courage, danger, humour, slander, villain;
17. 0 (XII) – er (XIII) – est (XII) – [ly (X)]:
dear, deep, easy, eld, fair, farth, fresh, great, happy, heavy, low, merry, plain, rich, strong, swift, true, worthy, young;
18. 0 (XII) – ity (XI):
adverse, antique, chaste, civil, divine, familiar, mortal, native, prosper, pure solemn, virgin;
19. 0 (XII) – man (XI):
country, French, gentle, noble;
20. 0 (XII) –ness (XI):
base, forgive, foul, gentle, idle, kind, noble, strange, white, wicked;
21. 0 (XII) – ty (XI):
certain, cruel, frail, loyal, proper, royal, sovereign, sure;
22. nt (XII) – nce (XI):
differe, excelle, ignore, impatie, innoce, invite, obedie, prese, revere, sile.

Для суффиксов итоговых парадигм могут быть получены свои диагностирующие признаки аналогично тому, как это было сделано для классов XI и XII. Вот, например, ДП суффикса -s (парадигма 14):

	ДП	Позиции				ДП	Позиции			
Кл. IV	—	—	9	—		mine	—	70	—	-
Кл. VI	—	—	20	—		our	—	110	—	-
all	—	400	—	—		their	—	80	—	-
and	20	—	—	—		three	—	1000	—	-
are	85	85	350	60		two	—	600	—	-
do	—	—	100	—		with	40	40	—	-
many	—	700	—	—		your	14	14	—	-

Смысл выделения ДП для членов парадигмы направлен на выявление несоответствий между дистрибутивными свойствами суффиксов и свойствами того дистрибутивного класса, к которому они относятся.

Рассмотрим в качестве примера класс и относящиеся к нему суффиксы. Сильным ДП для него являются слова класса VI; оказывается, однако, что они обладают известной избирательностью по отношению к суффиксам -s и -‘st (парадигма 1). Например, слово thou связано поощрениями с суффиксом -‘st и словом art и запретом со словом is. Группа слов he, it, she, there связана поощрением с hath и is и запретом со словом art. Наконец, группа слов I, they, we, you обнаруживает на сочетаемость со всем классом. Обобщая условия согласования, легко обнаружить соответствующие ограничения и для классов V и VII

Левое окружение (слова класса III)	Соответствующие единицы из классов V - VII
1. Все слова класса III	be, must
2. thou	-‘st, -st (canst, didst, dost, hast), -t (art, shalt, wast, wirt, wilt)
3. Остальные слова класса III	Кл. V (can, cannot, could, may, might, shall, should, will, would) Кл. V – VI (did, had, were) Кл. V – VII (came, knew, met, saw, took, went) часть слов кл. VII (brought, gave, heard, said, spoke, told)
4. he, it, she, there	-s (Кл. VI)
5. he, I, it, she, there	was
6. I	am
7. I, they, we, you	большая часть кл.

Картина не прояснилась окончательно, некоторые выводы все же можно сделать. Парадигма 1 частично связана согласованием с группами кл. III: суффикс -s — с группой 4, суффикс -‘st — со словом thou. Что касается суффикса 0 (Кл. VII), то он не сочетается со словом thou. Слова, концентрирующиеся вокруг класса V (т. е. показавшие высокий G-V), обладают редуцированной парадигмой: 0 — ‘st/st/t. Вне этой схемы остаются am, art, be, must, was.

Доказав согласовательный характер суффиксов парадигмы 1, мы тем самым доказали, что это парадигма словоизменения.

Оценивая грамматические результаты работы ДСА, мы можем признать их вполне удовлетворительными. Английский язык беден аффиксами, и это облегчало их формальное выделение и формирование парадигм. Если бы анализ проводился на русском материале, для достижения аналогичного результата потребовался бы больший текст. С другой стороны, развитая омонимия затрудняла работу. Тем не менее, процедура преодолевала эту трудность как при омонимии основ, так и при омонимии суффиксов.

Общий результат — 13 дистрибутивных классов и 22 парадигмы, довольно хорошо совпадающие с обычными грамматическими представлениями.

Открытие основных дистрибутивных классов не исчерпывает возможностей минимального интервала. Рассмотрим два параллельных примера из романов Диккенса и из «Войны и мира».

	f	S		f	S
more than	958	73	louder than	22	16
better than	351	44	stronger than	23	14
worse than	117	38	higher than	24	13
less than	152	32	shorter than	12	13
older than	57	30	shyer than	5	13
rather	153	23	longer than	38	11
oftener than	18	18	sooner than	33	11
otherwise than	54	18	faster	16	11
younger than	36	16	fairer than	7	11

Очень тесная связь со словом than обнаружили bigger, brighter, cleaner. clearer, darker, dearer, deeper, earlier, easier, fewer, fresher, further, handsomer, happier, harder, heavier, larger, later, lower, paler. plainer, prettier, quicker, redder, richer, sharper, slower, smaller, taller, tighter, truer, weaker, wider, wiser. Синтаксическая природа суффикса -er здесь совершенно очевидна. Что еще важнее – в той же конструкции находятся more, worse и less, чья принадлежность к компаративам никак не может быть доказана в микроинтервале. Семантическую близость к компаративам обнаружили rather и otherwise.

Неразличение Е и Ё создает трудности при анализе аналогичной русской конструкции (обозначены знаком *):

	f	S		f	S
более чем	21	26	о чем *	61	31
больше чем	26	38	понятнее чем	2	12
в чем *	82	18	прежде чем	32	40
важнейшем чем	2	19	при чем *	4	9
лучше чем	7	14	сильнее чем	4	10
веселее чем	3	13	хуже чем	3	7

Слева от компаративов часто появляется слово *гораздо*:

	f	S		f	S
гораздо более	2	4	гораздо менее	2	10
гораздо больше	3	11	гораздо после	3	8
гораздо лучше	9	71	гораздо прежде	2	5

С этим же словом ассоциированы *благороднее, богаче, большую, важнейшем, вероятнее, выше, дальше, меньше, опытнее, покойнее, правее, проще, сильнее, сильнейшие, слабее, умнее, худшем, хуже, чаще.* И здесь обнаружены компаративные суффиксы *-ее* и *-ш-*.

Точно так же обнаруживается связь суффикса множественного числа -s со словом few:

	f	S		f	S
few moments	146	138	few yards	13	18
few minutes	138	86	few pence *	7	16
few seconds	54	82	few grains	4	12
few paces	41	71	few months	15	12
few days	69	32			
few words	95	32	few moments’	12	49
few hours	50	28	few minutes’	6	14
few weeks	24	21	few hours’	4	9

Тесную связь со словом few показывают articles, boxes, chairs, drops, feet *, halfpence *, hundred *, inches, lines, men *, miles, mornings,

necessaries, notes, people *, pounds, questions, remarks, sentences, shillings, steps, turns, whiffs, years. И снова мы находим попадание в этот же дистрибутивный класс либо супплетивных форм плюралиса, либо слов с семантикой множественности (отмечены звездочкой).

Обобщая, можно сказать, что высокие значения S, обнаруженные на минимальном интервале, — самый надежный путь идентификации супплетивных пар и разнообразных морфологических исключений. Как пример рассмотрим максимальные S у трех форм английского глагола go:

	go	gone	went
F	4035	1799	3385
abroad	8	4	4
along	17		11
down	5	3	12
down-stairs	8		38
home	16	3	11
on	10		23
out	4	8	23
straight	9	4	12
through	4	5	4
up-stairs	10		23

Минимальный интервал способствует расщеплению омонимичных аффиксов, полученных в микроинтервале. Вот пример аффикса –АЯ (числа указывают на абсолютную частоту).

Первая группа слов с финальным –ая 1127 (какая 71, такая 53, маленькая 51 (княгиня 41!), другая 47, большая 39, старая 37 (графиня 15!), экая 35, новая 30 самая 29, милая 27… багрово-румяная, миниатюрненькая).

Концы правых соседей:

-ая=130, -яя=8 -ла=18 -лась=8;
-а=314 княжна=22, сила=17, женщина=17, девушка=13, минута=10, толпа=9, война=7, голова=8, дама=11, дорога=6, душа=4, жена=4, квартира=8, Наташа=5, причина=6, рана=4, собака=5;
-я=88 княгиня=51, графиня=18;
-ия=40 армия=16, история=11;
-ь=152 жизнь=22, часть =16, цель=11, прелесть=8, мысль=6, мудрость=5, боль=4, дверь=4, лошадь=4.

Вторая группа слов с финальным –ая 37 (Перонская 13, Болконская 5, Друбецкая 5, горничная 14).

Концы правых соседей:

-ла=6

Слово *которая* 274:

Концы правых соседей:

-ла=85, была =35;-лась=15; -ет 18 бывает =6;
-ит=5, ется=5, о=9 должна=8, не=9, с=8, так=10.

Концы левых соседей:

-а=31, -е=15, -и=34, Ии=7, ой=14, -у=29, -ы=11, -ь=26, -я=6.

Четвертая группа слов с финальным –ая 1387 (указывая 90, зная 88, ожидая 85, желая 75, отвечая 65, слушая 51, понимая 50, спрашивая 33, делая 32,поднимая 31, отыскивая 31, вставая 30…).

Концы правых соседей:

-0=82, -а=65, -ами=10, -его=10, -ее=7, -ей=14, -и=51, -ие=20, -им=15, -их=12, -ия=34, -ние=10, -о=25, -ого=5, -ое=13, -ой=26, -ти=6, -ть=93, -ться=13, -у=53, -ую=20, -ы=43, -ые=14, -ый=14, -ым=7, -ых=5, -ь=8, -ью=6, -я=13, -л=44, лся=6; в=51, -их=12, друг=10, его=42, ее=7, ему=25, за=10, к =32, как=23, на=138, не=13, о=13, от=28, с=37, себя=11, что=82.

Тем самым прокладывается путь к открытию таких лингвистических феноменов, как падеж, согласование и управление.

Понятие «минимального интервала» ДСА, по существу, распространяется на широко распространенную практику поиска коллокаций, столь популярную в информационных приложениях. Вот небольшой список таких словосочетаний из футбольного подкорпуса газеты «The Times» за 1995 год. Здесь мы находим терминологические сочетания grand slam, Achilles tendon, free kick, bookable offence, dressing room, acute angle, grammar school, corner flag, American football…; имена собственные Hong Kong, British Isles, New Zealand, Great Britain, Cape Town, New Guinea… Crystal Palace, Manchester United, Blaclburn Rovers, Bolton Wanderers, Manchester City, Coca-cola cup…; всевозможные идиомы общего языка – goodness knows, breathing space, each other, even though, great deal, for example…[1]

Многие уникальные сочетания с высокими значения S станут полноправными лексемами наряду с лексемами в обычном смысле.

Минимальный интервал открывает широкие возможности открытия семантических классов. Снова вернемся к романам Диккенса.

За графическим словом ' исключительно часто следуют verba dicendi: said (f=8362, S=108), cried (f=1276, S=57), returned (f=1114, S=40), asked, assented, demanded, echoed, ejaculated, exclaimed, faltered, gasped, growled, grumbled, inquired, interposed, murmured, muttered, observed, pursued, pursues, quoth, rejoined, remarked, repeated, replied, resumed, retorted, roared, screamed, shouted, suggested, urged, whispered. Совершенно неожиданно перед исследователем открывается большая группа глаголов с общей семантикой.

К неожиданным семантическим группам приводит нас минимальный интервал и в случае прилагательных цвета.

	white	black	red	blue	green	brown	grey
F	694	999	643	459	431	374	271
			Одежда				
cap	2	6	13	19			
cloak		7		4			31
clothes		6		30			
coat		6	3		14	15	9
frock	7	9					
gown		5			9		24
neckcloth	10		7				
neckerchief	26		16				
ribbon		13		15			
stockings	7	3					
surtout		9		14			
trousers	23	5					
waistcoat	47	81	2	3			
			Ткань				
cloth	19		8				
satin	22	18		20			
silk	8	26		5			
velvet			7		23		
			Голова				
eye		7					10
eyebrows		17					5
eyes		8	5	15		16	2
hair	17	39	2				38

Те же семантические группы существительных мы находим и у отдельных прилагательных цвета:

white: apron, cravat, gloves, great-coat, handkerchief, hat, linen, nightcap, robes; muslin; locks, teeth;
black: bonnet, breeches, dress, gaiters, glove, mantle, shawl;

[1] Подобные сочетания широко представлены во второй части настоящего тома.

	bombazeen, calico;	curls, moustache, whiskers;
red:	worsted	face, nose, whisker;
blue:	cockade;	cotton;
green:	veil;	baize;
grey:	livery;	head[1]

Специфическая семантическая сочетаемость обнаруживается у частых слов даже там, где семантическая лингвистика молчит. Обратимся к двум важнейшим грамматическим словам — артиклям. В романах Диккенса the встретился 182580 раз (P=0,048), неопределенный артикль a — 84853 раза (P=0,022) Традиционные грамматики, конечно, отметят употребление артикля the со словами same (f=2720, S=35), world (f=1026, S=12), sun (f=315, S=7), с прилагательными в превосходной степени (commonest, deepest, easiest, etc.).У Диккенса ясно видна группа слов, обозначающих персонажей романов: Jew (f=301, S=10, locksmith (f=257, S=10), baron, beadle, captain, coachman, coroner, dean, doctor, dwarf, hangman, hostess, landlady, landlord, major, manager, marquis, notary, schoolmaster, sergeant, spinster, turnkey, undertaker, waiter; a также собирательные — crowd, mob, rioters. Вплетены в сюжет и обычные слова — указатели предметов окружающего мира, на фоне которых действуют и говорят наши персонажи (все они тесно связаны с определенным артиклем): door (f=2343, S=17), fire-place (f=15, S=10), fire (f=1071, S=9), table (f=903, S=9), balcony, banisters, ceiling, chimney(-piece), dining-room, drawing-room, floor, hall, hearth, keyhole, kitchen, latch, mantel-shelf, piano, shutters, threshold, wall, window-sill.

Неопределенный артикль тесно связан с именами, указывающими на размер и количество в пространстве и во времени (особенно — малое количество):

	f	S		f	S
a few	1153	43	a pair	227	13
a little	3318	24	a large	434	14
a dozen	229	21	a great	1523	13
a moment	779	16	a low	408	12
a moment's	160	20	a week	229	11
a minute	286	17	a hundred	256	10
a couple	208	17	a word	650	10
a bit	256	16	a mere	180	8

bottle, century, considerable, fortnight, full-sized, gallon, guinea, handful, long, mile, million, month, morsel, mouthful, penny, piece, pint, quantity, quarter, score, shilling, short, slice, small, spoonful, syllable, thousand, trifle, twelvemonth, vast, year. В орбиту неопределенного артикля попадают и некоторые слова, обозначающие людей (все с изрядной долей негативной оценки): beggar, coward, fool, liar, hypocrite, madman, miser.[2]

Заметим, наконец, ассоциацию неопределенного артикля с неожиданными событиями: sudden (f=176, S=10), pang, swoon.

Изучение семантических связей слов в минимальном интервале несомненно принесет много интересного как в языке вообще, так и при сравнении разных

---

[1] Конечно, эти группы не исчерпывают потенциал сочетаемости цветовых прилагательных, ср. blue/black bag, red/white curtains, green/red flag, white/black letters, Black/Blue Lion, blue/green umbrella, White Hart, white horse, mice, steps, sugar; black board, ladder, sheep, teapot; red brick, coals, ink, lamp, marks, rim, wheels; Blue Beard, Blue Boar, Blue Dragon, blue mug, veins; green chariot, copper, fan, gate, packet, parasol, spectacles, spot; brown paper; grey mare, pony. На фоне всей этой предметной лексики диккенсовский пейзаж кажется довольно бледным: black cloud, blue sky, grey mist. Впрочем, именно здесь сосредоточилась вся "зелень»: green corn, fields, lane, leaves, mounds, pastures, turf. Само собой разумеется, встречаются повторения цветовой метафорики и идиоматики: black humours, ingratitude; red tape, brown study.

[2] Любопытно, что в комедиях Шекспира именно в неопределенном артикле в максимальной степени проявилось влияние фактора социального положения. в группе А (короли, принцы, герцоги) относительная частота неопределенного артикля равна 0,015, в группе Б (средний слой) — 0,19, а в группе В (слуги, шуты, крестьяне) — 0,24. См. [Шайкевич, 1974, с. 160].

авторов (ср. выше с.32) и разных языков[1]. Этому аспекту ДСА предполагается уделить много места во втором томе данного труда.

Что касается малого интервала в ДСА, то он остается почти неизученным, если не считать публикации 1963 г. (см. п. 1.2.3). Можно полагать, что в нем откроются некоторые перспективы формального подступа к синтаксису.

1.5.2. Средний интервал

В 1992 году мне посчастливилось перейти в Институт русского языка в Отдел Машинного фонда русского языка, основанного и руководимого Владиславом Митрофановичем Андрющенко. Это было время смены технической базы; персональные компьютеры вытеснили старые громоздкие ЭВМ. Прогресс техники позволил небольшим коллективам и даже одиночкам осуществлять то, на что раньше уходили десятилетия. В отделе постепенно росли собрания русских текстов на электронных носителях — естественная эмпирическая база для ДСА. Совершенствовалась и техника самого ДСА. Теперь можно было тестировать разные алгоритмы с помощью программ, созданных Наталией Александровной Ребецкой. Русская проза XIX в. стала основным объектом работы нашего коллектива. Прежде всего по инициативе Ю. Н. Караулова (тогдашнего директора Института) мы обратились к Достоевскому.[2]

Первые результаты компьютерной обработки русской прозы были получены уже в 1994 г. В статье «Конкорданс к прозаическому тексту» [Шайкевич, 1995] уже удавалось статистически сравнивать отдельные слова (*как-то, какой-то, странно, особенно, чувство, ощущение*) в прозе Пушкина, Лермонтова, Гоголя, Гончарова, Тургенева, Достоевского и Толстого (частотные словари к текстам этих писателей уже были готовы).

Непосредственным поводом для статьи стало появление в Японии конкорданса к «Преступлению и наказанию»[3], поэтому именно непосредственный контекст, задаваемый конкордансом (практически совпадающий с нашим малым интервалом), служил лексической основой для сравнения авторов. Приведем фрагмент этой статьи:

Поиски авторских расхождений ведут нас к анализу лексической комбинаторики пар существительное-прилагательное. В качестве полигона выберем важнейшие обозначения всей сферы психологических явлений – слова *ощущение* и *чувство.* Соотношение этих синонимов в прозе русских писателей показано в табл. 1.14 (относительная частота на 100 тыс. словоупотреблений текста).

Таблица 1.14

	ЧУВСТВО	ОЩУЩЕНИЕ	совокупная частота	доля слова *ощущение*
Пушкин	49	1	50	2%
Лермонтов	86	2	88	2%
Гоголь	58	3	61	5%
Гончаров	50	5	55	9%
Толстой	92	4	96	4%
Тургенев	86	18	84	21%
Достоевский	47	14	61	23%
1845-1849	66	17	83	20%

[1] Только что продемонстрированная связь слов the и fire(place), обусловленная текстом, не найдет себе параллелей в русской прозе не только потому, что в русском языке нет артикля, но и в связи с тем, что камин — характерная английская реалия XIX века, не свойственная русской жизни. Точно так же русская крестьянская печь не найдет себе параллелей в английской прозе.

[2] Группа, возглавляемая Ю. Н. Карауловым, разработала Словарь языка Достоевского. М., 2001, затем Словарь языка Достоевского. Идиоглоссарий. М., 2008.

[3] A concordance to Dostoevsky’s Crime and Punishment (Ando Atsushi, Urai Yasuo and Mochizuki Tetsuo, eds.), Sapporo, 1994.

1857-1865	38	11	49	22%
«Пр. и нак»	36	25	61	41%
1866-1880	59	12	71	17%

Таблица эта весьма выразительна. Во-первых, совокупная частота двух слов демонстрирует две группы авторов — Пушкин, Гоголь, Гончаров, Достоевский, с одной стороны, и Лермонтов, Толстой и Тургенев — с другой; последние используют эти слова значительно чаще. Во-вторых, ясно видно, насколько привержены *ощущению* Тургенев и Достоевский, в некотором смысле, литературные антиподы. «Преступление и наказание», быть может, самый «психофизиологический» роман Достоевского (концентрация *ощущения* выше только в «Игроке» и в коротеньком «Сне смешного человека»). Анализируя ближайшие синтаксически связанные слова (обычно обнаруживаемые в конкордансах), можно подойти к выявлению авторских различий в употреблении лексем, к определению их функциональной значимости. Вот как в корпусе рассмотренных текстов выстраиваются списки глаголов, ассоциированных со словом *ощущение*, если их упорядочить по абсолютной частоте.

Достоевский						Тургенев	
овладеть	5	забыть	2	охватить	2	испытывать	5
испытывать	4	родиться	2	передать	2	предаваться	3
отразиться	3	мучить	2	поразить	2	волновать	2
чувствовать	3	ослабеть	2	пройти	2	исчезнуть	2
вздрогнуть	2	отозваться	2				

Как видим, списки не слишком специфичны. Более длинный список Достоевского, естественно, объясняется большим объемом текстов.<…>. Отметим, однако, *овладеть* и *охватить* у Достоевского с общей идеей неожиданности, насильственности и полноты. С другой стороны, у Тургенева примечательны *волновать* и *предаваться*, явно указывающие на положительные эмоции, предвкушаемые через ощущение. У обоих писателей слово *ощущение* выступает индикатором психологической динамики героев. С лингвистической точки зрения оно (как и слово *чувство*) есть гипероним целого семейства обозначений конкретных чувств. Поэтому оба синонима удобны в тех случаях, когда надо передать нечто, не поддающееся точному однословному обозначению. Быть может, *ощущение* несколько более физиологично, чем *чувство*. Ср. два характерных контекста:

— Так не можешь угадать-то, — спросил он [Раскольников] вдруг с тем ощущением, как бы бросался вниз с колокольни.
— Н-нет, — чуть слышно прошептала Соня.
— Погляди-ка хорошенько.
И как только он сказал это, опять одно прежнее, знакомое ощущение оледенило вдруг его душу: он смотрел на нее и вдруг, в ее лице как бы увидел лицо Лизаветы («Преступление и наказание», ч. 5, гл. 4).

Он [Литвинов] испытывал ощущение, подобное тому, которое овладевает человеком, когда он смотрит с высокой башни вниз: вся внутренность его замирала и голова кружилась тихо и приторно. Тупое недоумение и мышья беготня мыслей, неясный ужас и немота ожидания, и любопытство, странное, почти злорадное, в сдавленном горле горечь непролитых слез, на губах усилие пустой усмешки, и мольба, бессмысленная… нм к кому не обращенная… о, как это было жестоко и унизительно безобразно! («Дым», ч. 1, гл. 9).

При общности поэтического задания, непосредственно толкающего автора к использованию слова *ощущение* (или более избитого — *чувство*), все остальное может различаться кардинальным образом. Это ясно обнаруживается при анализе прилагательных, сопровождающих наше слово.

Достоевский				Тургенев	
странный	20	непосредственный	4	сладкий	4
какой-то	19	беспрерывный	3	приятный	3

новый	18	прежний	3	странный	3
болезненный	13	приятный	3	жуткий	2
сильный	8	тяжелый	3	неприятный	2
неприятный	7	ужасный	2	новый	2
мучительный	5			смутный	2
				собственный	2

Конечно, и здесь мы находим лексическое отражение психологической эволюции героев (слово *новый*, например), но различий здесь больше. Характерные эпитеты Достоевского — *болезненный*, *неприятный*, *мучительный*, *тяжелый*, *ужасный* (и далее — *скверный*, *злорадный*, *тягостный*, *горький*, *мстительный*, *зловещий*, *подавляющий*, *брезгливый*, *раздражительный*, *мрачный*, *нестерпимый*, *истерический дикий*, *злобный* и т. д.). Трудно представить себе, чтобы Достоевский мог использовать здесь тургеневский эпитет *сладкий*, даже более сдержанное *приятный* тут же сменяется чем-то мрачным.

Но скоро и эти новые, приятные ощущения перешли в болезненные и раздражающие. («Преступление и наказание», ч. 1, гл. 5).[1]

За пределами возможности конкорданса остается выявление большинства существительных, ассоциированных с нашим словом; они, как правило, располагаются в соседних строках. Если пользоваться конкордансом как словоуказателем, можно эту же работу проделать прямо по тексту. Вот каким оказывается этот список у обоих писателей при учете статистической значимости совместного появления слов [существительных] (в скобках приводятся слова, чье появление вместе со словом *ощущение* стоит на грани статистической значимости).

Достоевский: *мозг, мука, мысль, натура, прилив, риск, сердце, тело* (*боль, вихрь, горе, душа, желание, казнь, оттенок, отчет, предчувствие, радость, скука, сон, страх, тоска, холод, чувство*).

Тургенев: *грусть, презрение, сердце, тревога* (*душа, квартира, мысль, чувство*).

Загадочное для читателя появление слова *квартира* в списке Тургенева показывает, как подобный статистический анализ помогает выявить неожиданные идиосинкразии у какого-то одного писателя и тем ярче оттенить их отсутствие у другого. Рассмотрим три контекста:

Лаврецкий пришел к себе на квартиру и заперся. Он испытывал ощущения, едва ли когда-нибудь им испытанные. («Дворянское гнездо», гл. 30).

Санин вернулся домой — и, не зажигая свечи, бросился на диван, занес руки за голову и предался тем ощущениям только что сознанной любви, которые и описывать нечего: кто их испытал, тот знает их томление и сладость; кто их не испытал — тому их не растолкуешь. («Вешние воды», гл. 25). Примеру из «Дыма», приведенному выше, также предшествуют слова «и возвратился к себе на квартиру».

Эти три примера указывают на характерную ситуацию — тургеневский герой (после сильной эмоциональной встряски — то ли свидания, то ли важного сообщения) наедине с самим собой предается ощущениям, пытаясь в них разобраться. Рефлексия типична и для героев Достоевского, но там она скоротечна и импульсивна.

И у Достоевского, и у Тургенева *ощущение* и *чувство* оказались связанными друг с другом. Этот результат достигнут средствами ДСА, но теперь можно на время забыть о нашем обете формализма и обратиться к реальным контекстам, где оба слова появляются рядом. Эта работа отражена в публикации 1996 г., которая следует ниже с некоторыми сокращениями.

В русской прозе второй половины XIX в. отмечена феноменальная экспансия слова *ощущение*. <…> Новое слово пришло из языка психологии и физиологии и сначала сохраняло свою терминологическую и материалистическую окраску. Характерен контекст из «Отцов и детей»:

---

[1] Далее следует первый пример обращения к среднему интервалу. Правда, к тому времени соответствующая программа еще не была реализована; поиск и подсчеты можно назвать скорее автоматизированными, а не автоматическими.

— Ты говоришь, как твой дядя. Принципов вообще нет – ты об этом не догадался до сих пор! – а есть ощущения. Все от них зависит.

— Как так?

— Да так же. Например, я: я придерживаюсь отрицательного направления – в силу ощущения. Мне приятно отрицать, мой мозг так устроен – и баста! Отчего мне нравится химия? Отчего ты любишь яблоки? – тоже в силу ощущения. Это все едино. Глубже этого люди никогда не проникнут. Не всякий тебе это скажет, да и я в другой раз тебе этого не скажу.

— Что ж? и честность – ощущение?

— Еще бы!

Но как только этот психологический термин включается в прозаический текст, он начинает взаимодействовать со старым ключевым словом художественной литературы (давним философским термином) словом *чувство*.

Все эти чувства были в нем, но в виде ощущений – и то неясных. (там же).

Пример из «Первой любви».

Особенно я полюбил развалины оранжереи. Взберусь, бывало, на высокую стену, сяду и сижу там таким несчастным, одиноким и грустным юношей, что мне самому становится себя жалко, — и так мне были отрадны эти горестные ощущения, так упивался я ими!...

Вот однажды сижу я на стене, гляжу вдаль и слушаю колокольный звон… Вдруг что-то пробежало по мне – ветерок не ветерок и не дрожь, а словно дуновение, словно ощущение чьей-то близости.

Слово *чувство* здесь прямо не названо, но *несчастье, одиночество, грусть, жалость* могут считаться его гипонимами.<…> Отсюда один шаг до совместного употребления двух слов как контекстуальных синонимов.

Я шел, чтоб вам это все рассказать, как будто время для меня остановилось, как будто одно ощущение, одно чувство должно было остаться с этого времени во мне навечно, как будто одна минута должна была продолжаться целую вечность и словно вся жизнь остановилась для меня… («Белые ночи»).

Все во мне волновалось от какого-то нового, необъяснимого ощущения, и я не преувеличу, если скажу, что страдала, терзалась от этого нового чувства. Короче – и пусть простят мне мое слово – я была влюблена в мою Катю. («Неточка Незванова»).

я с неудержимым любопытством, в припадке страха и радости и какого-то особенного, безотчетного чувства, отворила первый шкаф и вынула первую книгу… Я унесла к себе книгу с таким странным ощущением, с таким биением и замиранием сердца, как будто я предчувствовала, что в моей жизни совершается большой переворот (там же).

— Во мне простое чувство справедливости заговорило, а вовсе не родственное, — возразил запальчиво Аркадий. – Но так как ты этого чувства не понимаешь, у тебя нет этого ощущения, то ты и не можешь судить о нем. («Отцы и дети»).

чувство снисходительной нежности к доброму и мягкому отцу, смешанное с ощущением какого-то тайного превосходства наполнило его душу (там же).

В лице раздраженной Зины показалось болезненное ощущение, как будто от острой пронзительной внутренней боли; но она перемогла свое чувство («Дядюшкин сон»).<…>

Не последнюю роль в сближении этих слов играет и глагол *чувствовать* (*почувствовать*).

Мигом одно темное, далекое воспоминание детства моего воскресло в моей памяти. Чтоб понятно было то странное ощущение, которое я почувствовала в эту минуту, я расскажу это воспоминание («Неточка Незванова»).

Что я чувствовал, было не то смутное, еще недавно испытанное ощущение всеобъемлющих желаний, когда душа ширится, звучит, когда ей кажется, что она все понимает и все любит… («Ася»).

Лаврецкий прижался в уголок: ощущения его были странны, почти грустны; он сам не мог хорошенько разобрать, что он чувствовал («Дворянское гнездо).

Странное дело: мне было о чем раздуматься, а между тем я весь погрузился в анализ ощущений моих чувств к Полине («Игрок»).

но что могла сделать сила страсти, то могло быть, наконец, побеждено чувством обязанности, ощущением долга, чина и значения и вообще уважением к себе… («Идиот»).<…>

Синонимия слов *чувство* и *ощущение, чувствовать* и *ощущать* оказывается удобной для писателя и потому закрепляется в языке прозы.<…>. Ниже приводятся некоторые связи рассматриваемых слов с их значениями S, определенными в семистрочном интервале в текстах Достоевского.

ЧУВСТВО (f=821, S>3)

сердце =17, слезы =16, любовь =12, самосохранение =10, смычок =9, благородство, боль, восхищение, гордость, грусть, движение, дружба, дыхание, жалость, жизнь, звук, избыток, миг, **минута**, мир, моление, мука, **мысль**, невинность, ***ощущение***, порыв, природа, радость, реалист, ревность, ресницы, рука, скрипка, **сознание**, сострадание, страсть, стремление, стыд, счастье, торжество, **тоска**, уважение, черта, человек, щека;

великодушный =13, глубокий =12, рыцарский =9, безвыходный, бездыханный, безотчетный, бесчеловечный, благородный, благоуханный, вечный, враждебный, высокий, горький, горячий, грубый, девственный, добрый, дурной, задушевный, искренний, кроткий, **маленький**, **мучительный**, **неведомый**, нежный, несчастный, неудержимый, прекрасный, **сильный**, сладкий, сладостный, собственный;

мочь =20, лишиться =8 взглянуть, влечь, возрастать, домогаться, заблудиться, знать, излить, исцелит, любить, мутить, мучиться, **овладеть**, оскорбить, основываться, **ощущать**, пересилить, пожать, **помнить**, помутиться, поплакать, потрясти, притворяться, разгореться, сдавить, спасти, страдать , удержать, упасть, **чувствовать**;

я =18, свой =9, в =8, без, будто, **бы**, всегда, **какой-то**, мой, наружу, не, но, себя, этот.

ОЩУЩЕНИЕ (f=234, S>2)

сердце =6, душа =5, мгновение =5, раздвоение =5, **сознание** =5, ***чувство*** =5, воспоминание, глаза, закат, колокольня, мечтание, **минута**, мозг, многосторонность, **мысль**, отвращение, побои, потребность, раздумье, рассуждение, риск, солнце, сон, страх, тело, **тоска**;

смутный =12, **мучительный** =9, болезненный =8, странный =8, новый =7, беспрерывный, внезапный, **маленький**, **неведомый**, недавний, неприятный, нервный, нестерпимый, плотоядный, приятный, свежий, **сильный**, ужасный, упорный, яркий;

испытывать =6, **мочь** =6, **овладеть** =6, подписаться =6, **чувствовать** =6, воротить, вспомнить, доходить, жечь, начинать, обливаться, останавливаться, отразиться, **ощущать**, побледнеть, подавлять, поклониться, **помнить**, поставить, припомнить, пройти, пронестись, родиться, ругаться, рушиться, сбивать, совершаться, шуметь;

какой-то =8, злее =8, **бы**, вдруг, порой. <…>

В данном случае конкретный анализ контекстуальных употреблений слов свидетельствовал об ослаблении семантической дискретности. Подобный же подход может оказаться полезным при поисках семантической дискретности внутри слова. Обратимся к слову *чувство*, в котором можно подозревать реальную или латентную полисемию. Сгруппируем контексты согласно элементарным синтактико-морфологическим свойствам – формам единственного или множественного числа,

наличию или отсутствию полнозначного прилагательного, наличию зависимого существительного в родительном падеже и т. п.

В качестве первой группы такого типа рассмотрим контексты *без чувств* (а также формы род. п. мн. ч. при слове *лишиться*), у Достоевского находим 32 таких контекста. <…> Тривиальным оказывается присутствие здесь слов *без* (S=14), *лишиться* (S=31) и *очнуться* (S=8), намного интереснее глаголы с семантикой резкого движения: *упасть* (21), *броситься* (8), *оттолкнуть* (4) и, напротив, статичный глагол *лежать* (11) и слово *неподвижный* (4). Все это указывает на повторение одной и той же ситуации — потери сознания, сопровождаемой криком, о чем свидетельствуют *вдруг* (4), *вскрикнуть* (3), *закричать* (3), *звать* (6), *крик* (7), *пронзительный* (5), *раздаться* (18). Характерно указание на место – *столик* (5), *кровать* (5), *постель* (4), *кабинет* (4), *ковер* (3).

Он слабо вскрикнул и лишился чувств… («Хозяйка»).<…>

— Папочка! Папочка! – закричала я в последний раз, но вдруг поскользнулась на тротуаре и упала у ворот дома. Я почувствовала, как все лицо мое облилось кровью. Мгновение спустя я лишилась чувств… Очнулась я на теплой, мягкой постели… («Неточка Незванова»).<…>

Ламберт зашатался и упал без чувств; кровь хлынула из его головы на ковер («Подросток»). <…>

Сравнивая контексты с приведенными выше количественными данными о семантических связях, убеждаемся в том, что лексические связи пар слов выявляются статистическими методами достаточно полно, хотя восстановление ролевых функций при этом не гарантировано. Как бы то ни было, слово *чувство* в форме род. п. мн. ч. с предлогом *без* или с каким-то другим словом с приативной семантикой, совершенно оторвалось от остальных контекстов *чувства*. Приходится либо описывать семантику сочетания *без чувств*, либо реконструировать своеобразную семантическую окаменелость: ЧУВСТВА (мн.) — способность нормально функционировать психически: воспринимать мир, соображать.

Другой семантический процесс наблюдается в конструкциях с главным словом *чувство* и зависимым существительным в родительном падеже. Некоторые из таких сочетаний — несомненные фразеологизмы, например, *чувство собственного достоинства* (f=7) и чувство меры (f=5). В остальных случаях зависимое существительное обозначает конкретное чувство, а значит, главное слово чувство десемантизируется и выступает как формальный элемент, поддерживающий всю конструкцию. Здесь находим *чувство ревности* (7), *чувство долга* (7), *чувство сострадания* (6), *чувство дружбы* (5),*чувство жалости* (4), *чувство унижения* (4). Из этого ряда *чувство справедливости* (5), напоминающее упомянутые выше фразеологизмы. Своеобразным термином выступает у Достоевского *чувство самосохранения* (8), специфичность этого сочетания проявляется, в частности, в повторяемости контекстуально близких слов: *закон, идея, нормальный, человечество.*

Значительная часть контекстов слова *чувство* (около ста случаев) приходится на конструкцию *с чувством,* обычно сопровождающую реплики персонажей, их жесты и мимику. Своеобразие семантических связей в этих контекстах достаточно хорошо отражают слова с S>2:

поцелуй =7, ресницы =5, слезы -5, жар, идеал, память, рука, усмешка, шалунья;
побежденный =9, горький =6, гостеприимный =5, неудержимый =5, благородный, добрый, искренний, маленький, несчастный, робкий, сильный;
поцеловать =8, притворяться =7, благодарить =5, пожать -5, проговорить =5, благословить, покраснеть, прибавить, приезжать, сверкать, солгать, спохватиться;
с =9, ты =5, назавтра, неужто, я.

Характерные примеры подобного употребления дают следующие контексты.

...прибавил выразительно и с глубоким чувством герой наш, так что подбородок его запрыгал немножко и слезы готовы были опять навернуться. («Двойник»).

...с высоким чувством отвечала Марья Александровна («Дядюшкин сон»).

...проговорил он с чувством, пожимая мне руку. ...прибавил он, с чувством смотря на Сашу и украдкой на Настеньку. …сказал он, обнимая Сашеньку и с чувством смотря ей в глаза. …продолжал он с глубоким чувством (Дядя из «Села Степанчикова»).

Перечисленные группы контекстов исчерпывают конструктивные ограничения на использование слова *чувство*. Внимательный содержательный анализ остальной массы контекстов, может быть, позволит обнаружить еще какие-то центры семантической кристаллизации — латентные семантические варианты слова. Возможности формализованного дистрибутивно-статистического анализа здесь ограничены.

[Шайкевич. Оковы слова (или поиски дискретности в семантике), в сб. Словарь. Грамматика. Текст. М., 1996, с. 159–170]

Полная программа исследования среднего интервала была осуществлена на корпусе (точнее — трех корпусах) текстов Достоевского. Результат опубликован в «Статистическом словаре языка Достоевского».[1]

На с. XI этого издания дается краткое изложение техники анализа:

Формула оценки статистической значимости может быть использована для выявления текстуальных связей слов. Весь текст механическим образом членится на фрагменты равной длины (40 слов), а затем подсчитывается число фрагментов, в которых одновременно встретились слово X и слово Y. Если реальная частота совместной встречаемости статистически значима, делается вывод о текстуальной связи двух слов. Таким образом, в Словаре найдет отражение еще один лингвистический объект — **текстуальные связи** слов.

Так, редкое слово *агония* встретилось в жанре критики всего 4 раза, но показало текстуальные связи с пятью словами: *актер* (=18), *естественный* (=6), *зритель* (=10) *правда* (сущ.) (=2), *умирать* (=6). Из этих пяти связей одна (со словом *умирать*) может считаться общезначимой для русского языка, остальные — обусловлены конкретным текстом, где ведется речь об изображении агонии на сцене.

Надо иметь в виду, что основная формула ДСА дает тем больше статистически значимых результатов, чем больше анализируемый корпус текстов. Надежность и обилие результатов оборачивается кошмаром при публикации их в традиционном бумажном виде. Полученные текстуальные связи были опубликованы в виде компакт-диска, приложенного к словарю.

Ниже даются фрагменты публикаций, показывающие текстуальные связи двух важных слов.

Первый фрагмент взят из Введения к Словарю (с. XXXIII–XXXVIII).

Выше уже было введено понятие текстуальных связей. В электронной части Словаря содержатся обширные таблицы текстуальных связей слов отдельно по каждому макрожанру. Рассмотрим в качестве примера таблицы 1.15 и 1.16.

Таблица 1.15

Текстуальные связи слова РУССКИЙ

РУССКИЙ (прилагательное)

Художественная литература (f=459)			Публицистика (f=1420)		
S	f		S	f	
6	2	апофеоза	5	11	бессознательный
7	6	буква	10	15	«День"
6	6	бывать	8	89	дух
6	18	вообще	6	41	еврей
6	5	глупее	8	44	европеец
7	16	граница	4	67	европейский
6	8	дворянин	6	64	женщина
6	8	Европа	4	129	жизнь
17	14	европейский	7	71	земля

[1] Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А. Статистический словарь языка Достоевского. М., 2003. 832 с.

10	5	завет	7	32	интеллигентный
6	2	заморский	6	43	исторический
7	7	изящный	4	44	история
10	7	иностранец	11	104	литература
14	5	культурный	4	141	люди
13	9	либерал	5	6	местожительство
19	10	либерализм	21	447	народ
11	12	литература	12	47	народность
11	7	национальный	7	100	народный
11	84	наш	11	466	наш
7	12	немец	5	10	патриотизм
12	9	перевод	5	24	Петр
7	3	писатель	6	8	подняться
8	8	по-русски	5	65	понять
14	10	пословица	6	21	православный
6	8	поэт	5	8	пророческий
8	4	православный	7	19	простолюдин
6	3	просвещенный	6	73	Пушкин
14	27	Россия	4	138	Россия
10	13	русский (сущ.)	8	114	русский (сущ.)
6	3	Русь	4	27	семейство
6	3	светлость	5	16	слой
6	4	слой	5	6	стремительность
8	9	современный	6	43	тип
14	14	тип	5	28	царь
10	4	тысячелетний	6	17	чутье
6	10	француз	6	22	элемент
7	11	французский	4	35	явление
8	16	язык	12	82	язык

РУССКИЙ (существительное)

Художественная литература (f=107) Публицистика (f=390)

S	f		S	f	
4	3	Америка	5	12	англичанин
11	5	англичанин	4	5	атака
5	4	Астлей	4	7	болгарин
11	5	бывать	9	6	Греч
5	6	гораздо	4	6	дар
4	5	граница	11	28	еврей
8	5	Европа	7	62	Европа
11	3	европеец	15	35	европеец
3	2	жид	4	11	интеллигентный
3	2	забитый	7	5	коммунар
3	2	иностранец	5	7	край
3	3	лето	6	5	Крым
5	4	мистер	5	4	Лефорт
5	36	мы	5	3	Мериме
7	4	Наполеон	5	6	мусульманин
3	14	наш	6	223	мы
14	10	немец	5	3	нелиберальный
4	3	немецкий	8	23	немец
8	2	падкий	4	5	парадокс
4	4	петербургский	5	7	песня
5	3	поляк	4	9	племя
6	4	родиться	5	5	Польша
8	8	Россия	11	13	поляк
3	2	ругать	4	9	помогать
12	6	рулетка	4	5	приезжать
10	13	русский (прил,)	7	7	равный
8	3	славянофил	5	3	редут
3	4	способный	8	69	Россия
6	2	табльдот	8	114	русский (прил.)

3	2	Франция	8	9	серб
15	10	француз	4	6	Сербия
12	3	французик	9	34	славянин
6	2	цивилизованный	17	22	татарин
			7	23	турок
			7	4	удача
			4	5	цивилизованный
			5	22	язык

Табл. 1.15 подтверждает интуитивные представления читателей о Достоевском как философе русской идеи, однако во взаимодействии с таблицами лексических маркеров они могут обнаружить важные детали. Так, среди лексических маркеров «Дневника писателя» за 1876 год мы обнаруживаем слова *бескорыстие* (S=4), *Сербия* (S=4), *славяне* (S=7), *Россия* (S=6). Характерные контексты:

... Убеждение в бескорыстии России если придет когда-нибудь, то разом обновит и изменит весь лик Европы.
... Тем не менее честность, бескорыстие, прямота и откровенность демократизма в большинстве русского общества не подвержены уже никакому сомнению.
... Не служила ли она [Россия], напротив, в продолжение всей петербургской своей истории всего чаще чужим интересам с бескорыстием, которое могло бы удивить Европу, если б та могла глядеть ясно, а не глядела бы, напротив, на нас всегда недоверчиво, подозрительно и ненавистно. Да бескорыстию в Европе и вообще никто и ни в чем не поверит, не только русскому бескорыстию, — поверят скорее плутовству или глупости. Но нам нечего бояться их приговоров: в этом самоотверженном бескорыстии России — вся ее сила, так сказать, вся ее личность и всё будущее русского назначения.
... политика чести и бескорыстия есть не только высшая, но, может быть, и самая выгодная политика для великой нации потому, что она великая.
... Да если б Россия не только объявила, а и доказала бы даже, de facto, свое бескорыстие, то это, может быть, еще пуще смутило бы Европу. Ну, что ж такое, что мы ничего не возьмем себе, «облагодетельствуем» и уйдем назад, ничем не попользовавшись, а только лишь доказав Европе наше бескорыстие.

В следующем году начинается русско-турецкая война, и в «Дневнике писателя» за 1877 год среди лексических маркеров уже нет слов *бескорыстие* и *славянин*, но есть слова *Россия* (S=11), *турок* (S=13), *Турция* (S=6), *турецкий* (S=5), *болгарин* (S=6), *грек* (S=2), *Константинополь* (S=6), *народ* (S=4) и характерное *народец* (S=2), которое встретилось здесь 6 раз, а больше у Достоевского и не встречающееся. Этим словом обозначены балканские славяне. Снова обратимся к контекстам:

... Как может Россия участвовать во владении Константинополем на равных основаниях с славянами, если Россия им неравна во всех отношениях - и каждому народцу порознь и всем им вместе взятым? Великан Гулливер мог бы, если б захотел, уверять лилипутов, что он им во всех отношениях равен, но ведь это было бы очевидно нелепо. Зачем же напускать на себя нелепость до того, чтоб верить ей самому и насильно? Константинополь должен быть наш, завоеван нами, русскими, у турок и остаться нашим навеки.
... Федеративное же владение Константинополем разными народцами может даже умертвить Восточный вопрос, разрешения которого, напротив того, настоятельно надо желать, когда придут к тому сроки, так как он тесно связан с судьбою и с назначением самой России и разрешен может быть только ею. Не говорю уже о том, что все эти народцы лишь перессорятся между собою в Константинополе, за влияние в нем и за обладание им.
... Конечно, трудно устроить согласное и равное на правах владение Константинополем всех восточных народов и народцев, но ведь допускает же автор статьи [Данилевский], что Россия могла бы владеть Константинополем одна, пока, временно, так сказать, более охраняя его, чем смея владеть им, с тем, однако, чтоб после передать его на общее владение народцам (для чего? для чего передать?).

В «Дневнике писателя» на 1881 год на фоне побед русского оружия в Средней Азии Достоевский обращается к новым идеям — список лексических маркеров возглавляет слово *Азия*, есть здесь *азиаты*, *азиатский* и, парадоксальным образом, *цивилизаторский*. Писатель в обиде прощается с Европой:

... Они [в Европе] ни за что и никогда не поверят, что мы воистину можем участвовать вместе с ними и наравне с ними в дальнейших судьбах их цивилизации. Они признали нас чуждыми своей цивилизации, пришельцами, самозванцами. Они признают нас за воров, укравших у них их просвещение, в их платья перерядившихся. Турки, семиты им ближе по духу, чем мы, арийцы. Всему этому есть одна чрезвычайная причина: идею мы несем вовсе не ту, чем они, в человечество — вот причина!
... Но от окна в Европу отвернуться трудно, тут фатум. А между тем Азия — да ведь это и впрямь может быть наш исход в нашем будущем, — опять восклицаю это! И если бы совершилось у нас хоть отчасти усвоение этой идеи — о, какой бы корень был тогда оздоровлен! Азия, азиатская наша Россия, — ведь это тоже наш больной корень, который не то что освежить, а совсем воскресить и пересоздать надо!
... — В Европе мы были приживальщики и рабы, а в Азию явимся господами. В Европе мы были татарами, а в Азии и мы европейцы. Миссия, миссия наша цивилизаторская в Азии подкупит наш дух и увлечет нас туда, только бы началось движение.
... А главное — цивилизаторская миссия наша в Азии, с самых первых шагов (и это несомненно), поймется и усвоится нами. Она возвысит наш дух, она придаст нам достоинства и самосознания, — а этого сплошь у нас теперь нет или очень мало.

Если текстуальные связи слова *русский* часто вели нас к заветным идеям писателя, слово *глаз* (f=2383) с его 570 связями скажет нам больше о писательской технике.

Таблица 1.16

Текстуальные связи слова ГЛАЗ (S>5)
Художественная литература

S		S		S	
14	бледный	28	засверкать	7	посмотреть
8	блеск	7	засматривать	13	потупить
9	блеснуть	6	левый	12	пристально
8	блистать	7	лихорадочный	8	прищурить
7	брызнуть	20	лицо	6	проговорить
9	будто	7	личико	9	прямо
9	бывать	10	морщинка	6	раскрыть
9	вдруг	12	на	7	ресницы
13	взгляд	6	налитый	6	рост
9	взглянуть	7	него	9	рука
10	волосы	7	нее	31	сверкать
7	воспаленный	8	нежность	21	сверкнуть
9	впиться	9	неподвижно	6	свет
6	вскинуть	7	обвести	9	секунда
24	выпучить	10	обводить	6	серый
7	выступить	8	огонь	8	сиять
13	вытаращить	7	он	9	скосить
15	глядеть	12	она	19	слеза
12	голубой	6	опускать	6	смигнуть
6	Голядкин	23	опустить	21	смотреть
6	гореть	9	отвести	6	смыкаться
6	градом	10	отводить	28	спускать
9	грудь	7	отворотиться	7	стоять
16	губа	6	открывать	6	темный
13	его	6	открытый	6	Тихон
18	ее	11	открыть	8	ужас

9	заблистать	7	отрывать	6	хлопать
6	заглядывать	6	платок	6	хлынуть
10	загореться	6	побледнеть	6	цвет
9	зажмурить	13	поднять	14	черный
12	закрывать	6	подымать	15	щеки
15	закрыть	6	покраснеть		

Таблица ведет исследователя к признанию, по крайней мере, пяти сфер использования слова *глаза* в художественной литературе.

Во-первых, глаза — орган зрения, а отсюда такие текстуальные связи, как *смотреть, посмотреть, глядеть, спускать, обводить, обвести, отводить, отвести, отрывать, впиться* (а также с меньшей степенью значимости — *видеть, видать, всматриваться, глянуть, зоркий, искать, кругом, осмотреться, от, поглядеть, попадаться, провожать, разглядывать, рассматривать, сводить, скрыться, следить, толпа, увидеть, устремить*). Подобные текстуальные связи нейтральны в отношении субъекта и объекта зрения, хотя чаще ассоциируются с субъектом, т. е. наблюдателем, говорящим (именно на субъекта указывают связи с *защемить, зеленеть, померещиться, помутиться, потемнеть, темнеть*).

Вторая сфера — портрет, где глаза важный компонент всего изобразительного комплекса. Приметами этой сферы выступают такие связи: *морщинки, ресницы* (*большой, белки, близорукий, ввалиться, впалый, жадно, заплывший, навыкате, опухший, подслепый, очки, подбитый*); *цвет, голубой, серый, черный* (*карий, светло-карий, светло-серый, синий*); *волосы, темный* (*белокурый, брюнетка, всклоченный, густой, локон, проседь, седой, темно-русый*); *щека* (*румянец, румяный, красный, подбородок, рот, губки, зубы*); *лицо, личико* (*борода, бровь, нос, вострый, горбатый, продолговатый, сплюснутый, набеленный, голова, лоб, шея, кадык, уши*); *бледный* (*белый, бледно-желтый, желтизна, пожелтелый, смуглый, темно-желтый*); *рука* (*плечо, ручищи*); *рост* (*похудеть, одетый, стройный, сухощавый, худенький, худой*). Сюда же относится и эстетическая оценка (*красота, прекрасный, прелестный, чудно, чудный*). Эта сфера целиком связана с объектом наблюдения, хотя теоретически мыслимо совпадение объекта и субъекта (например, в зеркале), ср. такой контекст в «Бедных людях» —

... Я так похудела в последнее время; щеки и глаза мои ввалились, я была бледна, как платок...

Главное отличие второй сферы — статичность, именно этим отличается она от третьей сферы, хотя некоторые элементы ее могут участвовать и в динамике (лицо, щеки, руки).

Семантика третьей сферы формулируется принципом — глаза зеркало души, но не только души в ее статике, но и малейших движений души. Характерны такие связи: *взгляд, взглянуть* (*окинуть, уставиться*); *засверкать, сверкать, блеск, блеснуть, блистать, гореть, огонь, свет, сиять , заблистать, загореться* (*блестеть, искры, засветиться, засиять, затлеться, луч, лучистый, разгореться*); *воспаленный, лихорадочный; выпучить, вытаращить* (*выкатиться, вылупить*). Семиотическую значимость в литературном тексте приобретают любые движения глаз: *потупить, опускать, подымать; неподвижно* (*медленно*); *прищурить, смигнуть, хлопать, зажмурить* (*блуждать, замигать, мигать, моргать, щурить*).

Функционально здесь же оказываются мимика, жесты, любые неконтролируемые симптомы чувств: *побледнеть, покраснеть* (*краска, побагроветь, побелеть; кривиться, кривой, измениться*); *губы; дрожать, задрожать, вздрагивать, вздрогнуть, судорожный*.

Естественно, что в этой сфере оказываются как внешнее выражение мыслей и эмоций, так и сами эти мысли и эмоции. Ср. *нежность* (единственный пример положительной эмоции); *изумление; наивно, кроткий, стыдливо; лукаво, насмешка; бессилие, потеряться, смешаться; ужас; боль, мучительный, невыносимый; грозный, злобно, зловещий, негодование, ненавистный; бесстыжий, нахальный*. Очень характерен для Достоевского некоторый семиотический мост между выражением и содержанием, который призван «реалистически» отразить некоторую неуверенность наблюдателя в своей интерпретации взглядов и мимики объекта наблюдения: *будто* (*видимо, казалось, словно, точно, выражение, как-то, какой-то, что-то*).

В литературном тексте третья сфера почти целиком связана с описанием объекта наблюдения (лишь глаголы *потупить, подымать* и т. п. распространяются на субъект наблюдения).

Четвертая сфера, крайне характерная для Достоевского, основана на тривиальном факте: глаза — источник слез. Свидетели этой сферы — текстуальные связи со словами *слеза, брызнуть, выступить, градом, хлынуть* (*влажные, заплаканный, катиться, омочить, накипать, плакать, платок, повлажнеть, потечь, сквозь, слезинка, утирать*). Пятая (менее важная) сфера связана со сном и просыпанием: *открытый, открыть, смыкаться* (*закрывать, закрытый, закрыть, проснуться, утирать*). На долю этих пяти сфер приходится более 90% всех текстуальных связей слова *глаза*.

Привязанность второй и третьей сфер к наблюдаемому объекту поразительным образом может исчезать у Достоевского в тех текстах, где повествование ведется от первого лица, претензия на реализм (вряд ли здесь сознательный прием, скорее — небрежность издержки автоматизма писательского пера).

... я бледнела, краснела и сидела потупив глаза, боясь шевельнуться, дрожа всеми членами.
... — Нет, — отвечала я, посмотрев на нее ясными глазами.
... Он еще раз шагнул ко мне, но, взглянув на меня, увидел в глазах моих столько решимости, что остановился, как будто в раздумье.
(«Неточка Незванова»)
... Я нагло обвел их всех осоловелыми глазами.
(«Записки из подполья»)
... Глаза мои налились кровью. На окраинах губ запекалась пена.
... Губы у меня побелели, и я стал дрожать.
(«Игрок»)
... — Я не знаю, что выражает мое лицо, но я никак не ожидал от мамы, что она расскажет вам про эти деньги, тогда как я так просил ее, — поглядел я на мать, засверкав глазами.
...Я смотрел, выпуча глаза.
... — Если я выражался как-нибудь дурно, — засверкал я глазами, — то
... — Версилов не любил вас, оттого и не понял вас, — вскричал я, сверкая глазами.
... Я выпучил на нее глаза...
... Это — неправда! — вытаращил я глаза.
... И что ж, что я ее люблю, продолжал я вдохновенно и сверкая глазами...
(«Подросток»)

Фрагмент второй публикации [Шайкевич, 2007]посвящен слову ДЕНЬГИ:

Обращаясь к Достоевскому, рассмотрим текстуальные связи слова *деньги* отдельно по трем основным жанрам. Начнем с публицистики, где частота этого слова минимальна (36 на 100 тысяч словоупотреблений). Здесь находим около сотни текстуальных связей нашего слова. Приведем некоторые из них в сопровождении показателя (S) статистической значимости данной связи: *бюджет 4, взять 5, доставаться 4, достать 5, красть 5, купить 5, нуждаться 4, оружие 4, получение 4, помогать 5, помощь 6, порядочный 4, потребный 8, предложить 5, просить 5, разорение 4.* Вот характерный пассаж из «Дневника писателя» 1873.

Но оставим западников и положим, что деньгами все можно сделать, даже время купить, даже самобытность жизни воспроизвести как-нибудь на парах; спрашивается: откуда такие деньги достать? Чуть не половину теперешнего бюджета нашего оплачивает водка, то есть по-теперешнему народное пьянство и народный разврат, — стало быть, вся народная будущность. Мы, так сказать, будущностью нашею платим за наш величавый бюджет великой европейской державы. Мы подсекаем дерево в самом корне, чтобы достать поскорее плод. И кто же хотел этого? это случилось невольно, само собой, строгим историческим ходом событий. (П.С.С., т. 21, с. 91)

Однако стройной картины вокруг слова *деньги* не видно. Слишком велика в публицистике роль отдельных статей, где риторика требует бесконечных

повторений, благодаря чему аномально усиливается текстуальная связь слов, многие из которых человек признал бы случайными.

В письмах видим максимум частоты слова *деньги* (282 на 100 тысяч словоупотреблений), среди существительных оно уступает лишь слову *письмо*. Здесь у слова *деньги* обнаружено более двухсот текстуальных связей (при S>2). За вычетом имен собственных список выглядит так:[1]
банкир 3, без 4, билет 3, брать 3, ВЕКСЕЛЬ 10, взаймы 3, взять 7, возвращение 3, ВПЕРЕД* 10, выдавать 3, выдать 7, выдача 3, выкупить 3, ВЫСЛАТЬ 8, давать 3, дать 6, для 3, добывать 4, добыть 5, ДОЛГ* 10, ДОСТАТЬ 9, достать* 3, забрать 3, заем 4, заклад 4, занимать* 3, занять* 5, ЗАПЛАТИТЬ 10, зарез 3, издержать 3, иметь 4, истратить 5, копейка 7, крайне 3, месяц 3, надобность 3, наличные 3, напечатание 3, насчет 3, необходимость 3, нет 3, нужда 6, нуждаться 7, НУЖЕН 10, нужно 4, обязываться 3, остальной 3, отдавать 3, ОТДАТЬ 9, охота 3, печатать 3, платить 4, погибнуть 3, подписной 4, положение 3, получать 4, ПОЛУЧЕНИЕ 9, ПОЛУЧИТЬ 10, помочь 5, помощь 3, предлагать 3, принужден 3, прислать 7, присылать 5, присылка 3, продавать 3, продать 4, просить 4, просьба 4, процент 4, разом 3, расписка 5, распорядиться 3, рассчитывать 4, расход 5, РУБЛЬ 16, сверх того 4, серебро 7, сколотиться 3, случай 3, сумма 3, трата 6, требование 3, тысяча 5, удовлетворить 3, уплата 6, условие 3, хватит 6, часы 3, через 3, честный 3, чрезмерно 4, чтобы 3, 50 4, 100 7, 175 5, 200 4, 400 3, 500 3, 1000 5, 2000 5.

Просмотрев этот список, читатель сразу заметит лейтмотив писем Достоевского: крайняя нужда заставляет автора просить деньги в долг с обещанием вернуть их через какое-то время. Самые необычные слова подтверждают эту доминанту. Вот три контекста употребления слова *сколотиться*:

(из письма к брату 1844 г.)
— Спросишь, где достали деньги; я сколочусь и дам 500. Паттон — 700; у него они есть; и маменька Паттона 2000. Она дает сыну деньги по 40 процентов. Этих денег вельми довольно для печатания. Остальное в долг.

(из письма к брату 1858 г.)
Пишешь, что вышлешь фрак и одни брюки. По-моему, лучше бы сюртук. Ведь он всегда полезнее. Как-нибудь сколочусь и сделаю здесь, хотя в деньгах у меня большая крайность.

(из письма к Чумикову 1865 г.)
Сколько я ни бился, как ни разрывался, чтоб доставить Вам обещанные 500 р. к сроку, но ничего не мог сделать. До самого последнего дня надеялся, что сколочусь и достану их; ничего не вышло...

Это редкое слово еще дважды встретилось в художественных текстах. В «Двойнике» (передача речи Голядкина 2-го):
...что... даже на сапожишки не мог сколотиться и что вицмундир взят им у кого-то на подержание на малое время.

В «Преступлении и наказании» (слова Мармеладова):
И откуда они сколотились мне на обмундировку приличную, одиннадцать рублей пятьдесят копеек, не понимаю?

Слово *деньги* очень важно и в художественных произведениях Достоевского (95 на 100 тысяч словоупотреблений, 15-е по частоте среди существительных). В «Бедных людях», «Записках из мертвого дома», «Братьях Карамазовых» относительная частота превышает 140, а в «Игроке» составляет 247 (6-е место среди существительных после слов *бабушка, генерал, Полина, де Грие* и *раз*). Громадный объем корпуса (более 1800 тыс. словоупотреблений) обеспечивает

[1] Звездочкой отмечены слова-значения, так или иначе связанные с денежной сферой — *долг, занимать, занять, достать (достанет=хватит), вперед* (в виде аванса). В данном списке хорошо видна изосемантичность однокорневых слов, о которой шла речь выше, ср. *выдавать, выдать, выдача; получать, получение, получить.*

богатый, разносторонний и легко интерпретируемый список более трехсот текстуальных связей (при S>2). Назовем важнейшие:
антрепренер 4, билет 4, БРАТЬ 16, бросать 4, бумажник 7, вексель 6, вещь 4, ВЗАЙМЫ 13, ВЗЯТЬ 16, вор 5, выдавать 6, вынуть 5, грабеж 6, гривна 4, грош 5, гульден 5, давать 12, дарить 5, даром 4, ДАТЬ 10, двести 8, двугривенный 4, десять 5, добывать 5, добываться 5, добыть 9, долг* 4, доставать 4, достать 12, заклад 4, залог 9, заложить 4, занять* 7, заплатить 5, заработать 4, истратить 5, казенный 4, калач 4, капитал 4, карман 11, карманный 5, КОПЕЙКА 17, копить 6, кредитка 8, купить 9, кутеж 4, кутить 5, куча 6, лежать 5, ломбард 5, медный 8, миллион 4, Митя 4, нажить 5, наследник 4, наследство 4, нужда 5, нуждаться 5, нужен 10, нужно 4, обеспечение 4, обложка 4, ограбить 6, окровавленный 4, отдавать 5, ОТДАТЬ 15, откуда 4, отобрать 6, отсчитать 5, ПАКЕТ 14, пан 4, паспорт 5, ПАЧКА 16, ПЛАТИТЬ 13, подсудимый 11, полтора 5, получать 5, получение 4, получить 9, портмоне 4, посылать 4, похваляться 5, принести 4, присвоить 6, прислать 4, продавать 4, продажа 4, продать 8, прожить* 7, проиграть 9, проигрывать 4, пропить 6, процент 12, пятипроцентный 6, пятьдесят 5, пятьсот 8, разбросать 4, расписка 8, растратить 6, расчет 4, РУБЛЬ 22, рулетка 4, сдача 5, семьсот 5, серебро 4, сказывать 4, скопить 7, Смердяков 4, сосчитать 6, спрятать 6, сто 5, сторублевый 5, сумма 11, счет 4, считать 4, торговать 4, ТРАТИТЬ 13, требовать 4, три 9, тридцать 4, триста 8, ТЫСЯЧА 21, уберечь 5, убить 5, украсть 8, унести 8, уплатить 7, франк 7, целовальник 5, чтобы 5, шкатулка 8.

Важная черта семантики денег в художественной литературе — их овеществление. Оно проявляется в обозначениях денег как вещи (*билет, бумажка, монета, кредитка, медный, сдача*), в существительных, обозначающих место хранения денег (*карман, портмоне, бумажник, шкатулка, пакет, пачка, обложка, конверт*), а также в глаголах, обозначающих конкретное действие (*вынуть, выложить, вытащить, выхватить, отсчитать, разбросать, швырнуть, топтать, спрятать, принести, унести*). Не столь значимы, но характерны обозначения того, на что тратятся деньги (*вино, калач, пряник, сахар, платье* «одежда»). Текстуальные связи с обозначениями конкретных сумм денег — общее свойство литературных текстов и писем. Как мы видели выше в Русском ассоциативном словаре, на слово-стимул *деньги* 11 раз последовала реакция *зло*. Конечно, у писателя Достоевского такой банальной текстуальной связи мы не обнаружим, даже в публицистике, где максимально проявляется Достоевский-морализатор, мы находим лишь две-три текстуальные связи, замешанные на морали (*золотой мешок, рабство, фанатизм*). В литературных текстах мы наблюдаем явную связь *денег* со словами, ассоциируемыми с преступлением — *вор, воровать, грабеж, ограбить, подсудимый, окровавленный, убийца, убить*. Преступление, в том числе корыстное преступление, — важная черта сюжетов зрелого Достоевского.

Семантический круг, очерчиваемый вокруг какого-то конкретного слова, может быть расширен за счет связей второго порядка, т. е. связей каких-то слов Y с рассматриваемым словом X через третье слово Z. Чем больше таких путей от Y к Хтем больше оснований соединить Y и X даже при отсутствии прямой текстуальной связи между ними. Например, слово *ставка* связано со словом *деньги* через такие слова: *бабушка, выиграть, выигрывать, гульден, десять, игра, проиграть, расчет, рулетка, тысяча, флорин, фридрихсдор*. Благодаря связям второго порядка круг слов Y, связанных с X, разрастается (в том числе в результате усиления прежде слабых прямых текстуальных связей).

К рассмотренному выше списку теперь добавляются: *ассигнация, аукцион, банк, банковый билет, барыш, букинист, бумажка, вернейший, взятка, владение, возвратить, вол, воровать, восемьдесят, восемьсот, вперед*, вручить, выдавать, выдать, выиграть, выигрывать, выигрыш, выложить, вынимать, выправить, вытащить, выхватить, вычет, двугривенничек, десятирублевый, десяток, до зарезу, должен*, ежемесячно, ералаш, жалование, жид, завернутый, загребать, задаток, заем, заемное письмо, закладная, зачет, золото, золотой, игорный, игра, имущество, исправно, колесо, конфетка, копеечка, кошелек, кредитный билет, крупер, лавка, лавочка, луидор, медь, меняльный, миллион, монета, монетка, наличный, наменять, недодать, обыскать, отыграться, пенсион, пересчитать, плата, подарить, подписывать, пожертвование, покупать, покупщик, полтина, полуимпериал, приданое, прииск, продаваться, проиграться, прокутить,

пряник, пятак, радужный, развернуть, разменять, расплатиться, расход, ростовщик, с лишком, сверток, серебряный, синенький, содержание, состояние*, ссудить, ставить, ставка, стоить, талер, товар, торговаться, тысячка, тяпнуть, флорин, фридрихсдор, целковый, цена, цыганка, ящик, zero.*

То или иное слово, входящее в определенный таким образом круг, изредка этим кругом и исчерпывает свои текстуальные связи. Таковы, например, *двугривенничек, до зарезу, закладная*. Однако чаще какие-то текстуальные связи (и даже их большинство) ведут наружу. Доля внутригрупповых связей в общем числе текстуальных связей слова варьирует в широких пределах. Вот несколько примеров — *недодать* 75%, *банковый билет* 70%, *зачет* 67%, *взаймы* 67%, *пятак* 53%, *медь* 50%, *ассигнация* 50%, *банк* 46%, *полтина* 43%, *взятка* 38%, *вексель* 36%, *вор* 27%, *заем* 25%. По-видимому, где-то между 10% и 20% можно провести условную границу сферы слова *деньги*, тогда за пределами этой границы окажутся, например, *бумага, ключ, пять, девять, брать, взять, убить, поставить*. Впрочем, рассматриваемая сеть никак не свидетельствует о существовании более или менее замкнутого поля ДЕНЬГИ. Картинка ПРОСТРАНСТВА кажется более адекватной сути дела.

Итак, сравнив психологические ассоциативные эксперименты и методику ДСАТ, мы убеждаемся в том, что оба пути ведут к построению большой сети семантических отношений. Преимущества ДСАТ, на мой взгляд, заключаются в следующем. 1. Несравнимо меньшие трудовые затраты (особенно при учете постоянного расширения электронных корпусов текстов). 2. Возможность получения дифференцированных результатов (подобно жанрам Достоевского). 3. Возможность возвращения к контексту при поисках отдельных значений многозначного слова.

Последнее требует некоторого пояснения. Слово *заклад* показывает исключительно значимые текстуальные связи со словами *побиться* (S=42) и *биться* (S=13). Ясно, что (*по*)*биться об заклад* следует считать отдельными единицами, их частоты придется вычесть из частоты слова *заклад*, заново получить текстуальные связи. При этом *побиться об заклад* покинет сферу денег, а слово *заклад* укрепит свое место в этой сфере. Сложнее обстоит дело с текстуальной связью *ни* и *копейка* (S=11), *ни* и *грош* (S=7), *ни* и *единый* (S=15), *ни* и *капля* (S=10), *ни* и *малейший* (S=38), *ни* и *слово* (S=28). Конечно, можно признать самостоятельность соответствующих словосочетаний и выявить их текстуальные связи, но операцию пересчета вряд ли стоить производить. Ведь в этом случае произойдет общее снижение значений S, при том, что *копейка* и *грош* несомненно принадлежат к сфере денег, а у слова *ни* чрезвычайно информативна связь со словами с общим смыслом «бесконечно малый».

Возврат к реальным контекстам позволил выделить отдельные слова-значения, отмеченные звездочкой в вышеприведенных списках. По-видимому, такую операцию следует применить и к некоторым другим словам (*играть, ставить, заложить*). Например, у слова *заложить* обнаружены связи со словами *книга, крючок, рука, спина*, никак не соотнесенные со словами сферы денег — *вещь, деньги, заклад, пистолет, рубль, часы*.

Анализ текстуальных связей может быть использован при межъязыковом сравнении текстов. Первый опыт такого рода был доложен на XIII Международном съезде славистов в 2003 г. [Шайкевич, 2003] ДСА был применен здесь к поэтическим текстам Пушкина и Мицкевича. Длина фрагмента составляла 30 графических слов.

Сеть семантических текстуальных связей в поэзии Пушкина и Мицкевича

В отличие от ДСА прозаических текстов анализ поэтических текстов обнаруживает совместную встречаемость, объясняемую эффектом рифмы. Сравнительно редко семантическая интерпретация рифмующихся слов самоочевидна: dźwięk ‘звук’ — jęk ‘стон’; *упоенье – наслажденье*. В большинстве случаев, однако, расстояние в 1-2 строки оставляют поэту достаточную свободу «семантического маневра» для соединения двух смыслов:

droga – noga	f=18 S= 4	любовь – вновь	f=20 S= 3
oko – głęboko	f=16 S= 6	очи – ночь	f=37 S= 8
człowiek -- wiek	f=15 S= 6	поэт – свет	f=24 S= 5

męka – ręka	f=13 S= 5	любовь – кровь	f=20 S= 3
katusza – dusza	f=13 S=11	человек – век	f=15 S=12
droga – trwoga	f=12 S= 6	мука — рука	f=11 S= 4
grzech – uśmiech	f=10 S=13	веселье — похмелье	f= 7 S= 13
brzeg – bieg	f= 8 S= 9	брег – бег	f= 6 S= 8
gwiazda – jazda	f= 7 S= 7	Киприда — обида	f= 4 S= 8
trup – słup	f= 5 S= 7	роза – мороз	f= 4 S= 7

Можно полагать, что наличие потенциально рифмующихся слов способствует усилению некоторых семантических связей и ведет к соответствующей окраске фрагментов текста (ср. katusza ‘мука’ –– dusza, *младость — радость* f=17 S=17, *радость — сладость* f=5 S=6, *сладость — младость* f=4 S=7). Тем не менее пары рифмующихся слов (даже при очевидной семантической связи) были изъяты из окончательной сети связей.

Текстуальные связи, полученные на двух корпусах текстов, могут совпадать во многих деталях, ср.:

берег – вода	6	brzeg – woda	3	тело – труп	4
берег – волна	9	brzeg – fala	3	слеза – плакать	4
берег – дно	3	brzeg – dno	3	плакать – рыдать	3
берег – озеро	2	brzeg – jezioro	8	человек – зверь	3
берег – скала	3	brzeg -- skała	4		
берег – тихий	2	brzeg – cichy	3		
берег – челн	5	brzeg – łódka	5		

ciało – trup	5		
łza – płakać	5		
płakać – szlochać	11		
człowiek – zwierzę	3		

Нас, однако, будут занимать более или менее обширные секторы СТС, как совпадающие, так и различающиеся у двух поэтов.

Совершенно очевидно, что во многих случаях текстуальные связи мотивируются особенностями сюжета. Прежде всего это касается имен собственных.<…>

Влияние отдельных сюжетов очевидно в связях многих других слов. Так у слова *бес* (f=49) с его связями (*ад* 6, *Балда* 4, *веревка* 4, *вниз* 4, *вылезти* 6, *Гавриил* 11, *дух* 4, *Иерусалим* 5, *кобыла* 7, *лукавый* 7, *оброк* 5, *проклятый* 8, *сатана* 13, *соблазн* 4, *тащить* 4) явно проступает привязанность к трем текстам — «Монах», «Гавриилиада» и «Сказку о Балде». Так, у слова *ад* (f=22) находим связи — *архангел* 4, *бес* 6, *дорога* 3, *клобук* 5, *Лета* 3.

В поэзии Мицкевича piekło (f=55) имеет текстуальные связи, прямо указывающие на семантическое поле веры, — bestyja 3, bezkarnie 3, dusza 4, łotr ‘подлец’ 3, Lucyper 4, niebo 3, przesąd ‘суеверие’ 3, raj 4, zbrodnia ‘злодеяние’.3 Единство поля веры и религии у Мицкевича демонстрируется такими словами с их связями:

Bóg (f=214) ‘Бог’ anioł 3, cierpieć 4, człowiek 4, grzech 4, grzesznik 3, ludzie 3, mędrzec ‘мудрец’ 4, rodzica ‘матерь’ 3, stwórca 3, stworzyć 4, swobodnie 4, wiara

wiara (f=59) ‘вера’ bóg 5, kościół 3, ksiądz 3, moc 3, święty 4, ustawa ‘закон’ 3

kościół (f=54) ‘церковь’ cmentarz ‘кладбище’ 3, człowieczy 3, lud 4, mnich ‘монах’ 3, msza ‘месса’ 4, papież ‘папа’ 5, pobożność ‘набожность’ 3, święty 4, wiara 3

anioł (f=74) ‘ангел’ bóg 3, dusza 3, modlić 7, niebieski 3, niebo 3, pański 5, święty 4, sąd ‘суд’ 3, tron 3, wieczny 3

święty (f=120) ‘святой’ anioł 4, cnota ‘добродетель’ 5, duch 7, kościół 4, krzyż ‘крест’ 3, litanija 5, modlić 5, obrazek 3, obrząd ‘обряд’ 3, patron 7, piekło 5, ratować ‘спасать’ 6, syn 4, w imię 5, wiara 4, wszechmocny 3

Подобных текстуальных связей мы совсем не найдем в поэзии Пушкина. Здесь мы видим одно из важных различий двух семантических систем, топика серьезной веры, столь характерная для Мицкевича, у Пушкина не выявляется статистическими средствами.

Рядом с полем веры располагается еще один фрагмент СТС, в высшей степени характерный для Мицкевича, это поле страдания и терпения:

cierpieć (f=53) ‘терпеть’ bóg 4, ból 5, gardzić ‘презирать’ 4, katusza 6, męczarnia ‘мучение’ 4, męka 3, potrafić ‘смочь’ 3, stukać ‘стукнуть’ 4, znosić ‘сносить’ 1 5

ból (f=36) ‘боль’ cierpieć 5, cierpienie 7, czoło ‘лоб’ 3, gorzeć ‘гореть’ 3, męka 3, namiętny ‘страстный’ 4, obraza ‘обида’ 4, obrazić 5, okupić ‘искупить’ 13, płochy ‘пугливый’ 5, przeminąć ‘минуть’ 6, rajski 3, rana 10, śmierć 5, szaleć ‘сойти с ума’ 3, trucizna ‘отрава’ 3, ukoić ‘утешить’ 6, wycierpieć 4, zadać 3, zapał ‘воодушевление’ 3

katusza (f=25) ‘мука’ choroba ‘болезнь’ 3, cierpieć 6, cierpienie 6, 4, dręczyć ‘терзать’ 3, przeklęty ‘проклятый’ 3, sprawca ‘виновник’ 8, wieczny 5, wpleciony ‘вплетенный’ 11, zbawić ‘спасти’ 4

męka (f=39) ‘мука’ ból 3, cierpieć 3, duchy ‘духи’ 3, dusza 3, lękać się ‘бояться’ 4, męczyć 3, obrońca ‘защитник’ 4, srogi ‘жестокий’ 3, syn 3, twarz ‘лицо’ 3, tyran 3, ulżyć ‘облегчить’ 5, wezwać ‘вызвать’ 3, zbawiciel 4

Аналогичные слова у Пушкина не образуют единства и семантически сдвинуты в сторону темы любви, ср. **мука** (f=47) душевный 5, испытать 3, позабыть 3, скорбь 3, слезы 4, согбенный 4, сочетать 3, тайный 6, услаждать 3 и **страдать** (f=36) бешеный 3, затрепетать 5, изобразить 3, любовь 3, мрамор 3, мученик 4, погрузиться 4, способность 9.

То или иное семантическое поле языка, постулируемое на логических («тезаурусных») основаниях, может сохраняться в СТС, а может и совершенно исчезнуть. Противопоставления четырех времен года хорошо представлены в тексте:

весна (f=61)	4	лето (f=20)	wiosna (f=45)	8	lato (f=17)
	4		12		10
осень (f=18)	8	зима (f=40)	jeseń (f=12)	4	zima (f=28)

Напротив, единство животного мира не находит подтверждения в текстуальных связях (оно, возможно, сохранилось бы в зоологических текстах). Такой же вывод придется сделать и в отношении семантического поля цвета. С другой стороны, семантическое поле лица и тела довольно хорошо подтверждается средствами ДСА. Об этом можно судить по связям, представленным на рис. 10 и 11 (в круглые скобки взяты слова, чьи текстуальные связи сосредоточены за пределами рассматриваемой группы).

В обоих рисунках проглядывает общая внутренняя структура — в верхней части сосредоточены слова, так или иначе связанные с выражением эмоций.

Сеть текстуальных связей показывает некоторые комплексы слов, совершенно отсутствующие в тезаурусах. Речь идет о группах слов, связанных с определенной темой. Характерной иллюстрацией может служить тема застолья у Пушкина (рис. 11). Эта же тематическая группа хорошо представлена у Мицкевича, хотя и с меньшим числом членов и с менее сильными связями (рис. 11).

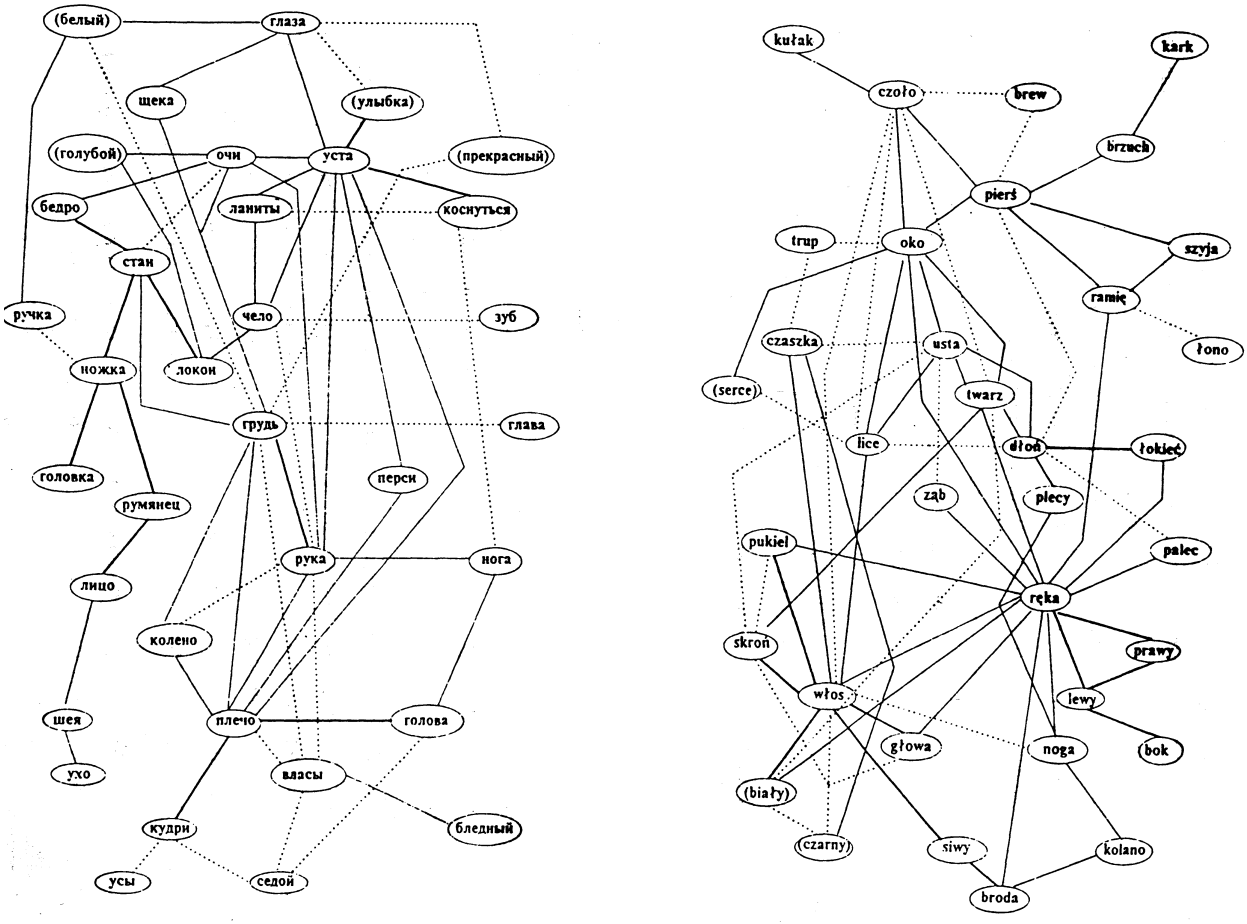

Рис. 10

Две части рис. 11 дают сходную, но все же не идентичную картину. У Мицкевича группа включает слова, обозначающие еду; в социальном отношении интересно здесь и слово wódka.

Во многом совпадают у двух поэтов и наиболее важное поле лирики с тремя центральными словами kochać — любить, miłość — любовь, serce — сердце. Вот некоторые текстуальные связи:

kochać (f=121) czuły ‘нежный’ 3, kochanek 3, kochanie 5, miłość 4, nienawiśdć 4, ona 4, serce 3, szał ‘безумие’ 5, szczęśliwy 3, ty
miłość (f=94) kochać 4, ona 3, serce 6, wesele ‘свадьба’ 3
serce (f=373) czuły 8, ja 3, kochać 3, kochanie, łza ‘слеза’ 3 miłość 6, namiętność ‘страсть’ 3, płomień 3, twój 4

любить (f=328) душа 3, жажда 3, любовь 3, сердце 3, твой 3, ты 3, я 5.
любовь (f=451) волненье 4, восторг 4, дружба 6, душа 5, желанье 6, жизнь 5, лобзанье 3, любить 3, наслажденье 5, нега 3, нежный 6, объятие 3, поцалуй 4, пылать 3, радость 5, ревнивый 4, сердце 3, страстный 6, счастливый 6, счастье 4, твой 3, томный 3, упоенный 3, упоенье 3, я 4
сердце (f=426) жар 3, желанье 3, любить 3, любовь 6, наслажденье 3, сладость 3, страстный 4, страсть 7, чувство 4.

Бросаются в глаза не только громадные различия в частоте слов *любить* и *любовь* в сравнении с польскими kochać и miłość, но и ярко выраженная чувственность лирики русского поэта. Это эротическое окружение трех центральных слов подтверждается такими текстуальными связями:

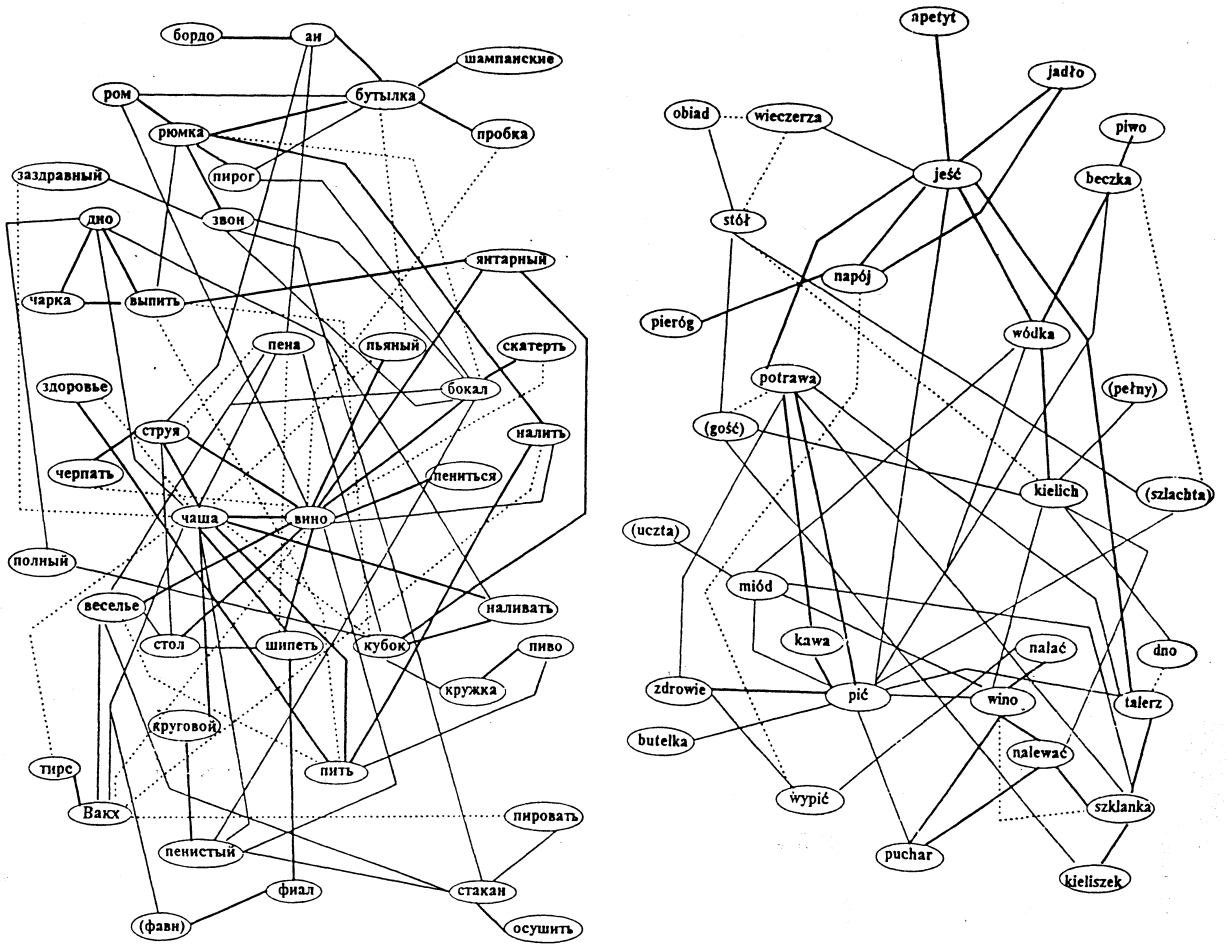

Рис. 11

лобзанье (f=27)	лобзать 4, нежный 4, страстный 4, томительно 8, уста 5;
лобзать (f=10)	лобзанье 4, желать 3, коснуться 4, уста 5;
объятие (f=35)	желанье 3;
страсть (f=113)	восторг 3, жар 6, нежный 3, пылать 4, упоенье 4;
восторг (f=97)...	жар 3, желанье 4, наслажденье 8, пламенный 5, пылкий 3, счастье 3, упоенный 6, упоенье 7;
желанье (f=94)	восторг 4, наслажденье 3, нега 5, объятие 3, томить 5, упоенье 3;
нега (f=88	желанье 5, сладострастный 3, томный 3;
томный (f=88	нега 3, стыдливость 3, упоенье 3;
наслажденье (f=76)	восторг 8, желанье 3,;
упоенье (f=38)	восторг 7, желанье 3, страсть 4, стыд 4,томный 3.

Тематическая группа «поэзия» — одна из самых важных в СТС Пушкина. Сюда входят слова *поэт* (f=188), *стих* (f=157), *муза* (f=132), *петь* (f=124), *лира* (f=118), *певец* (f=112), *писать* (f=79), *дар* (f=68), *вдохновенье* (f=57), *гений* (f=52), *перо* (f=38), *Феб* (f=37), *рифма* (f=36), *Аполлон* (f=30), *парнасский* (f=26), *ода* (f=25), *воспеть* (f=23), *Парнас* (f=21), *пиит* (f=17), *бумага* (f=15), *Геликон* (f=11), *Пегас* (f=11), *воспевать* (f=11), *марать* (f-11), *Аонида* (f=7).

Автономное положение в этой группе занимает слово *петь*, чьи текстуальные связи ведут нас и к поэзии, и к музыке (*стих* 3, *Торквато* 4, а с другой стороны — *звонкий* 3, *песенка* 11, *песнь* 7, *пир* 3, *соловей* 5, *хор* 4).

Обратное соотношение представлено у Мицкевича — группа поэзии включает всего четыре слова wiersz (f=18), rym (f=39), muza (f=25), jamb (f=12). Большую группу пения и музыки возглавляют слова:

śpiewać (f=76) cicho 4, hymn 3, kościół 3, lutnia 3, minstrel 4, muzyka 4, nastroić 5, pieśń 10, piosenka 9;
pieśń (f=83) brzmieć ‘звучать’ 3, chór 3, lutnia 3, muzyczny 6, nucic ‘напевать’ 9, pienie 3, słuchać 3;
piosenka (f=79) d**ź**więk ‘звук’ 7, mistrz ‘мастер’ 3, nucić 7, trubadur 5, zadzwonić 6, zanucić 7, zaśpiewać 11;
muzyka (f=27) brzmieć 3, grać ‘играть’ 3, nuta ‘нота’ 5, śpiew ‘пение’ 3, słuchacz ‘слушатель’ 3, taniec 3.

Сюда же входят d**ź**więk (f=42), brzmieć (f=8), śpiew (f-15), nucić (f=29), lutnia (f=21), nuta (f=9), słowik ‘соловей’ (f=9), lira (f=6).

Мы рассмотрели несколько полей или групп слов, однако следует помнить, что СТС не распадается на вполне автономные группы. Над отдельными группами могут располагаться слова с очень широким, а иногда и неопределенным спектром текстуальных связей. Таковы, например, у Пушкина прилагательные широкой семантики (*милый* f=371, *нежный* f=153, *прекрасный* f=122, *сладкий* f=90, *прелестный* f=80, *волшебный* f=54, *сладостный* f=44, *дивный* f=37, *чудесный* f=32, *пленительный* f=19) и некоторые абстрактные существительные (краса f=93, *красота* f=94).

Подведем итог. Чем обширнее эмпирический полигон исследования, тем более надежные результаты дают статистические методы анализа. С этой точки зрения рассмотренные корпусы (примерно по 200 тысяч слов текста) не очень велики. Тем не менее, приложение ДСА к поэтическим текстам позволило нащупать некоторые характерные черты искомого семантического пространства двух корпусов.

Дальнейшая проверка возможностей ДСА на среднем уровне связана 1) с переходом к анализу связей второго порядка, 2) с уточнением лексических единиц (их слиянием или расщеплением), 3) с итерацией процедуры обнаружения связей, на этот раз — над текстом, в котором отражены результаты предыдущих этапов ДСА.

Иллюстрацией связей второго порядка могут служить русские этимологические дублеты *берег* (f=124) и *брег* (f=91). Эти два слова не имеют взаимных текстуальных связей, но в СТС мы находим 17 третьих слов, с которыми связаны *берег* и *брег*, это —

6 вода 2; 9 волна 4; 3 дикий 2; 4 море 2; 2 Нева 3; 2 плыть 5;
4 ручей 4; 4 рыбак 5; 3 скала 4; 2 там 2; 2 Терек 2; 2 тихий 2;
4 утес 3; 5 хлынуть 4, 4 холм 5; 5 челн 6; 4 челнок 4.

На основе данных связей с третьими словами можно признать наличие связи второго порядка между *берег* и *брег*, конституируя тем самым новый лексический объект *берег/брег*. С меньшим основанием можно объединить дублеты *голос* (f=98) и *глас* (f=88) с их общими связями — 7 *звонкий* 2, 2 *звук* 3, 5 *лира* 3.

Переход к связям второго порядка поможет объединять не только дублеты, но и синонимы, такие как koń (f=170) и rumak (f=42), имеющие 11 общих связей с третьими словами — 9 cwał ‘галоп’ 6, 3 czarny 2, 11 je**ź**dziec 2, 4 kopyto 2, 5 lecieć ‘лететь’ 3, 3 na 2, 7 ostroga ‘шпора’ 3, 6 siodło 3, 2 trąba 2, 3 wodze 10, 5 wstrzymać 2.

Слова chart ‘борзая’ (f=32) и pies ‘собака’ (f=80) напрямую связаны текстуальной связью (S=7), кроме того у них обнаружено 10 общих связей с третьими словами – 9 asesor 3, 14 kot 4, 10 Kusy 6, 12 obróż ‘ошейник’ 3, 7 psarnia 3, 9 rejent 3, 6 smyk ‘поводок’ 3, 2 Sokół, 3 szarak ‘заяц-русак’ 7, 12 szczwacz ‘доезжачий’ 3. Все это позволяет свести два этих слова в единый лексический объект.

С другой стороны, необходимо преодолеть отрицательный эффект, порождаемый слишком большой концентрацией слова или словосочетания в отдельном тексте. Если, например, отдельно учесть связанные друг с другом ksiądz и Robak в «Пане Тадеуше», остальные случаи появления слова ksiądz яснее обнаружат его связь с полем религии. Концентрация слова *голова* в «Руслане и Людмиле», где оно выступает как обозначение действующего лица, ослабляет его связи с остальными словами поля лица и тела. Слишком большое давление жанра сказок с их многократными повторениями искажало подлинную семантическую картину в случае таких слов, как *ветер* (f=73) — *бежать 4, Буян 5, весело 4, гулять 3, забиться 3, кораблик 8, оборотиться 3, остров 8,*

подгонять 9, щель 10. Учет подобных влияний усилит системность конечной сети текстуальных связей.

Интригующими кажутся перспективы ДСА в исследовании корпусов параллельных текстов. Первый шаг в этом направлении был сделан в дипломной работе Г. Ю. Лавриной «Опыт дистрибутивно-статистического анализа параллельных текстов (на примере драматических произведений Шекспира)» [РГГУ, Институт лингвистики, 2009 г.]. Оригинал Шекспира сравнивался здесь с русскими переводами XIX–XX вв. (Е. Н. Бируковой, Т. Гнедич, М. А. Кузьмина, А. И. Курошевой, Л. С. Некора, Б. Л. Пастернака, А. Д. Радловой, Т. Л. Щепкиной-Куперник и др.). В первой части работы осуществлялся поиск переводческих эквивалентов на материале комедий Шекспира. Естественными фрагментами для ДСА служили: 1) реплика героя и 2) поэтическая строка. Как и следовало ожидать, в интервале «строка» значения S оказались выше, но число кандидатов на точный переводческий эквивалент — меньше, чем в интервале «реплика». В случае имен персонажей англо-русские пары с максимальными значениями S неизменно оказываются истинными переводческими эквивалентами: S=104 ford — форд, S=103 malvolio — мальволио, S=102 toby тоби, hector — гектор и т. д.[1] В парах англо-русских топонимов большие значения S обнаружены как у русских существительных, так и у прилагательных: brentford — брентфорд, брентфордский, troy — троя, троянский, rousillon — русильон, русильонский, Milan — милан, миланский и т. д.Нелемматизированные пары английских и русских графических слов (в интервале «строка») демонстрируют гибкость переводческой техники, например, у английского thee максимальные значения S находим у *тебя* (S=59), *тебе* (43), *тобой* (18), *тобою* (13), *ты* (13), *твою* (8), *твое* (6). Лемматизация в обоих интервалах обычно дает максимальное S подлинному эквиваленту, ср. (в интервале «реплика») английское lark и русское *жаворонок* (S=81), *щегленок* (25), *дрозд* (17), *воробей* (15), *запеть* (13), *кукушка* (13), *соловей* (12); ape — *мартышка* (70), *обезьяна* (43),bed — *ложе* (61), *постель* (32), fool — *шут* (76), *дурак* (48) Конечно, многозначность (или омонимия) английского слова отражается и в величинах S у русских слов, ср. bark и лаять (S=29), *кора* (29), art — *ты* (20), *искусство* (11), back — *спина* (24), *обратный* (14), bay — *затравить* (42), *порт* (23), bear — *медведь* (37), *рождаться* (25), case — *падеж* (45), *футляр* (21), crown — *крона* (71), *корона* (32), fine — *тонкий* (19), *штраф* (9), lead — *свинец* (20), *вести* (10), light — *свет* (19), *легкий* (19), lock — *замок* (43), *локон* (14), make — *заставлять* (11), *сделать* (7). Явные ошибки такой процедуры в списке из 5000 слов не превышают 4%.

Вторая часть работы Лавриной посвящена сравнению сети текстуальных связей в английском (длина фрагмента — 50 слов) и русском (длина фрагмента — 40 слов)корпусе. Добавим к этому французский перевод Шекспира, сделанный Франсуа Гизо в 1820-х гг. (длина фрагмента — 50 слов).

Примером хорошего структурного подобия небольшой группы слов могут служить три кластера вокруг значений «ад» и «дьявол». Вот как выглядят связи соответствующих слов (с их значениями S):

```
devil f=241   hell 15, dam 12, Lucifer 8, fiend 7, incarnate 5,
              damned 4,black 4 (+ tempt 7, evil 4, mad 4, torment 3);
hell f=150    fiend 14, damned 11, heaven 9, Lucifer 3 (+soul 5, torment 5,
              black 4, Lucifer 3 (+ hot 4, torture 4, witch 4);
damned f=104  hell 11, devil 4, black 3 (+ villain 7, conspire 6, wicked 5,
              dog 4, hellish 4, kill 4, sin 4)
fiend f=69    hell 14, devil 7, Lucifer 5 (+ foul 11, darkness 9, conscience
              6, hideous 6, Jew 5, poison 5);
Lucifer f=7   devil 8, fiend 5, hell (+ cuckold 8).

diable f=196  Lucifer 13, emporter 10,incarné 7, enfer 6, damner 4,
              démon 4, (+ conscience 4, juif 4, proverbe 4);
enfer f=168   démon 23, damner 9, noir 7, diable 6, maudit 5 horrible 3
              (+ ciel 8, âme 6, brûler 5, malédiction 5, Pluton 5);
démon f=144  enfer 23, incarné 9, horrible 7, diable 4 satan 3
             (+ rouler 6, suggérer 6, ange 5, effréné 5, oh 5, poison 5,
              tenter 5, abîme 40;
```

[1] Напомним, что в программах ДСА снимается различие прописных и строчных букв.

```
maudit f=133  démon 7, enfer 5;
damner f=62   enfer 9, satan 5, diable 4;
satan f=9     damner 5, horrible 4, démon 3.

черт f=214    дьявол 5 (+ побрать 33, взять 15);
дьявол f=152  ад 11, Люцифер 11, воплощенный 7, бес 6, преисподняя 6 черт 5
              (+ совесть 6, ангел 5б искушать 5, плоть 5
ад f=130      рай 14, дьявол 11, демон 8 бездна 4, бес 3 (+ кромешный 9,
              заклятье 6, небо 6, попасть 5, пытка 5, карта 4, Плутон 4,
              проклятый 4);
бес f=35      дьявол 6, ад 3 (+ совесть 8, жид 5);
рай f=26      ад 14.
бездна f=22   преисподняя 8, ад 4;
демон f=15    ад 8;
преисподняя f=9 бездна 8, дьявол 6;
чертовка f=9  черт 5, ад 3 (+ заклясть 20);
Люцифер f=7   дьявол 11.
```

Вполне замкнутый семантический кластер, заданный оригиналом, имеет выходы наружу в антонимичную сферу («рай», «небо», «ангел»), к значению «совесть». Отметим и некоторые различия — *побрать* и *взять* в русском языке и emporter — во французском, чему нет соответствия в английском.

Зависимость (до полного тождества) английских и французских текстуальных связей от шекспировского оригинала наблюдаем в названиях трех (!) частей года. Три смысла текстуально связаны друг с другом.

```
summer f=73    winter 21, spring 10, bud 8, warm 7, flower 6, fly 5, bee 4,
               hot 4, April 3;
winter f=68    summer 21, spring-time 7, spring 6, autumn 5, melt 5, weather
               5, April 4, cold 4, snow 4 (+ flaw 8, reap 6, shrink 6,
               sun 6);
spring f=47    summer 10, winter 6, flower 5, April 4, herb 4, shower 3
May f=14       April 18;
April f=13     May 18, shower 6, spring 4, summer 3.
spring-time f=4 shower 12, melt 8, winter 7
```

Обратим внимание на то, что смысл «зима» вводит необычайно развитую идею холода, и далее ведет к теплу и жару:

```
cold f=193     warm 11, raw 9, hot 8, freeze 6, heat 6, fire 4, winter 4,
               zeal 4 (+ bite 5, catch 5);
hot f=144      cold 8 fire 5, summer 4, cool 3 snow 3 (+ blood 9, drop 4,
               hell 4);
heat f=69      cool 12, cold 6;
melt f=59      snow 11, shower 9, spring-time 8 fire 5, thaw 5 winter 5
               summer 3 (+ drop 6, tears 6);
warm f=54      cold 11, summer 6, cool 4 (+ knead 8, youthful 6, bleed 5);
cool f=48      heat 12, sprinkle 9, warm 4 hot 3;
snow f=34      white 13, melt 11, rain 5 shower 4 winter 4(+ Alps 10, pure 7,
shower f=26    spring-time 12, melt 9 flower 4 snow 4 (+ rain 16, tears 7,
               tempest 6)
thaw f=10      melt 5, fresh 3 (+ fire 6).
```

Смысл «весна» ведет к миру растительности, особенно же к смыслам «цветок» и «роза» (с притаившейся червоточиной -- canker), а от них к цветовому спектру:

```
fresh f=99     dew 7, lily 7, revive 5, bud 4, flower 4 thaw 3 (+ salt 9
               arrow 5, disperse 5, garment 5, new 5, perfume 5);
flower f=95    weed 11, dew 10, juice 10, herb 9, purple 9, grass 8,
               nettle 6, summer 6, violet 6, blue 5, spring 5, bud 4,
               fresh 4, green 4, shower 4 (+ strew 21, bank 9, fairy 6,
               sweet 6, flock 5, grave 5);
root f=62      branch 11, leaf 7 (+ hew 7, spray 7)
rose f=60      red 23, thorn 14,white 13 pluck 8, damask 7, bud 6, pale 6,
```

```
                wither 6
weed f=56       herb 12, flower 11, garden 8, garland 6;
wither f=53     branch 6, rose 6, sap 5(+ old 6)
dew f=38        flower 10, fresh 7, brier 5, cloud 4, morning 4, rose 4
                (+ tears 6;
leaf f=38       blossom 9, branch 7, root 7 brier 5 (+ caterpillar 8, storm 7,
                tragic 7)
garden f=32     weed 8, herb 5 (+ brick 12, art 6)
branch f=30     lop 19, root 11, leaf 7, sap 7(+ cedar 20, swift-winged 16,
                witting 11)
bud f=25        canker 10, summer 8, blow 6, rose 6, flower 4, fresh 4
                (+ wont 6, eat 5, sweet);
canker f=21     bud 10, thorn 5 (+ discard 9)
thorn f=21      brier 14, rose 14, canker 5 (+ prick 14, sharp 8)
blossom f=17    leaf 9;
brier f=15      thorn 14, bush 13, dew 5 leaf 5 red 4 rose 4 (+ scratch 6)
herb f=15       weed 12, flower 9, garden 5, spring 4, wither 4 (+ grace 5);
sap f=10        branch 7, wither 5
```

В русских переводах основное отличие от оригинала состоит в появлении прилагательных:

```
суровый f=86    леденить 5, лето 4, зима 3 (+ приговор 5, произнести 4,
                сбросить 4, смириться 4);
зима f=47       осень 24, лето 20, летний 12, ледяной 10, весна 6, зной 5;
лето f=36       зима 20, весна 15, цвести 5, мороз 4, суровый 4(+ бабий 10);
весна f=26      лето 15, дождь 9, апрель 7, зима 6, мороз 5, цвести 5,
                цветок 5;
мороз f=19      трава 7, лед 6, весна 5, холод 5, лето 4,зима 3;
летний f=17     зной 18, зима 12;
ледяной f=12    зима 10 (+ сосулька 20);
май f=11        апрель 11 (+ декабрь 21);
апрель f=10     май 11, весна 7,(+ дождь 5);
зной f=10       летний 18, зима 5;
ненастье f=6    осень 22, зима 7 (+ лесной 14 навести 10);
леденить f=5    суровый 5
```

В целом кластер кажется более компактным, чем в тексте оригинала. В этом отношении показательны связи слова *мороз*.

```
холодный f=98    лед 13
снег f=27        знойный 11, холод 9, зима 3, (+ белый 11, огонь 7, зуб 5)
холод f=24       тепло 11, снег 9, зимний 6, мороз 5;
тепло f=18       холод 11;
зимний f=16      вьюга 21, холод 6
лед f=13         холодный 13, мороз 6, жаркий 4
```

В связях с частями года мир растений ограничен здесь только цветами:

```
цветок f=73      роза 7, сок 7, сорвать 7, роса 6, червь 6, весна 5 трава 5
                 (+ белый 12, девственный 12, мята 10, сосать 7, цвет 6);
роза f=56        бутон 18, увянуть 13, распуститься 12, сорвать 10, шип 10,
                 колючий 7, цветок 7, червь 5 (+ алый 31,
                 белый 15, щека 13, душистый 8, девственный 7, маска 6);
сорвать f=56     роза 10, цветок 7 (+ шлем 12, герб 11);
трава f=48       сорный 26, роса 8, мороз 7, цветок 5 (+ сад 6);
червь f=38       изъеденный 8, колючий 8, бутон 7 цветок 6 роза 5 (+ прах 8,
                 пожрать 6);
роса f=36        трава 8, цветок 6;
цвести f=16      весна 5 зима 4 лето 4, сад 4;
увянуть f=13     роза 13;
бутон f=7        роза 18, распуститься 7, червь 7
распуститься f=7 бутон 17, роза 12, цветок 5
колючий f=5      червь 8, роза 7
```

В рассматриваемом кластере единственная лингвистическая трудность для переводчика заключена в слове canker, этимологическом дублете слова cancer. Французская и следующая за ней русская традиция переводов Шекспира использует здесь слово ver (червь), усиливая метафоричность оригинала. Вот три примера переводов:

«Два веронца» 1.1

Yet writers say “As in the sweetest bud
The eating canker dwells, so doting love
Inhabits in the finest wits of all.”

And writers say “As the most forward bud
Is eaten by the canker ere it blow,
Even so by love the young and tender wit
Is turned to folly, blasting in the bud,
Losing his verdure even in the prime,
And all the fair effects of future hopes.”

Les auteurs disent cependant que l'amour habite dans les esprits les plus élevés, comme le ver dévorant s'attache au bouton de la plus belle rose.

Et les auteurs disent aussi que, comme le bouton le plus precoce est rongé intérieurement par un ver avant qu'il s'épanouisse, de même l'amour porte à la folie les esprits jeunes et tendres; qu'ils se fanent dans la fleur, perdent la fraîcheur de leur printemps, et tout le fruit des plus douces espérances.

Читал я в книгах, что бутон нежнейший
Таит червя порою, - так любовь
Вселяется в тончайшие умы.

А я читал: как полный уж бутон.
Источенный червем, не расцветет,
Так и любовью юный, нежный ум
В безумье обращается. Он сохнет,
Теряется вся свежесть раньше срока
И все приметы будущих надежд.

«Сон в летнюю ночь» 2.2

Come? now a roundel and a fairy song,
Then for the third part of a minute hence:
Some to kill cankers in the musk-rose buds,

Составьте круг теперь и спойте песню!
Потом на треть минуты - все отсюда:
Кто - убивать червей в мускатных розах,

Allons, un rondeau, et une chanson de fées; et ensuite, partez pour le tiers d'une minute, que les unes aillent tuer le ver caché dans le bouton de rose;

The First Part of Henry the Sixth 2.4
«Генрих шестой» (часть 1) 2.4

Hath not thy rose a canker, Somerset?

Hath not thy rose a thorn, Plantagenet?

Ay, sharp and piercing, to maintain his truth,
Whiles thy consuming canker eat his falsehood.

Somerset, ta rose n'a-t-elle pas un ver qui la ronge?

Plantagenet, ta rose n'a-t-elle pas une épine?

Oui, une épine aiguë et piquante, propre à défendre la vérité; tandis que ton ver rongeur détruit son mensonge.

Не червь ли в вашей розе, Сомерсет?

Не шип ли у твоей, Плантагенет?

Колючий, острый, чтоб стоять за правду.
Твою ж неправду пожирает червь.

Для этого кластера во французских переводах характерно появление общего смысла «сезон», а, с другой стороны, сдвиг текстуальных связей в сфере растительности.

```
hiver f=73     été 11, froid 8, glace 7, saison 7, geler 6, neige 6,
               saison 5, soleil 5 stérile 4 (+ conte 8);
été f=42     hiver 11, rose 4, bouton 3 (papillon 6, richesse 6);
printemps f=37 fleur 8, bouton 7, hiver 6, avril 5 (+ source 6);
saison f=31    hiver 7, fleur 5 (+ année 7);
avril f=13     mois 10, hiver 3, printemps 3;

froid f=140     rechauffer 11, chaud 9, geler 8, glacé 8, hiver 8, neige 7,
                chaleur 6, glace 5 (+ transi 11, pâle 7, soupe 7,  marbre);
chaud f=71      froid 9;
chaleur f=58    froid 6, glace 5, glacé 4 (+ fiole 8, impression 6);
glacé f=50     froid 8, neige 5, chaleur 4 (+ transi 13, sang 5,vent 5)
neige f=40      blanc 12, glace 10, froid 7, hiver 6, geler 5, glacé 5
                (+ chaste 11, Alpes 9, pur 6)
glace f=33      neige 10, hiver 7, chaleur 5, froid 5, geler 5;
geler f=15      froid 8, hiver 7, fleur 5, glace 5, neige 5

fleur f=141    herbe 13, joncher 12, semer 11, épanouir 10, guirlande 8,
               suc 8, bouton 7 violette 7, feuille 5, flétrir 5 geler 5
               parfum 6, plante 6, pourpre 6, rosée 6, moisonner 5,
               prairie 5, saison 5, gazon 4, hiver 4 (+ doux 8, fée 6,
               virginal 6);
fruit f=97     tige 9, plante 6, rameau 5(+ arbre 14, recueillir 14);
flétrir f=73   fraîcheur 11, faner 9, fleur 5
rose f=68      bouton 16, épanouir 15 joue 14 blanc 13, rouge 13 teindre 10,
               ver 10, vermeil 8, couleur 7 (+ cueillir 22, églantier 17,
               épine 12)
racine f=39    branche 10, feuille 6;
plante f=34    stérile 7, fleur 6, fruit 6, herbe 6, ortie 6 vigne 7
               (+ jardinière 8, sol 7);
feuille f=33   racine 6, fleur 5 (+ abriter 11, papier 10, zéphyr 9);
stérile f=32   plante 7, hiver 4 (+ caillou 5, montagne 5 reconnaissance 5)
ver f=32       bouton 20, rose 10, épanouir 5 (+ ronger 17, luire 10,
               dévorer 6);
branche f=31   racine 10, rameau 9, souche 9;
tige f=25      fruit 9 tronc 5, écorce 6, rose 5 (+ rejeton 17,
               sauvageon 14, race 9);
bouton f=21    épanouir 22, ver 20, rose 16, fleur 7, printemps 7;
rameau f=21    cèdre 15, tronc 14, branche 9, olivier 9, vigne 9
écorce f=17    sève 10, tronc 7, tige 6 (+ arbre 33, graver 13)
épanouir f=16  bouton 22, rose 15, fleur 10, ver 5;
tronc f=15     rameau 14, sève 10, cèdre 9, renaître 8, écorce 6,
               tige 6(+ arbre 7);
cèdre f=12     rameau 15, tronc 9 (+ renaître 18, réunir 9);
faner f=9      flétrir 9;
sève f=9       écorce 10, tronc 10;
vigne f=8      rameau 9, plante 7;
```

joncher f=6 fleur 12;
souche f=6 branche 9.

Некоторые кластеры оригинала довольно хорошо сохраняются в переводах. Таково, например, собрание ругательных эпитетов и некоторых других слов, им сопутствующих.[1] Три нижеследующих списка слов дают примерную картину довольно размытой семантики этой группы:

villain f=274
knave f=179
base f=170
dog f=166
slave f=154
coward f=110
plague f=102
rascal f=172
vile f=128
rogue f=90
foolish f=87
whore f=55
sack f=52
cur f=42
whoreson f=37
peasant f=29
villainous 27
scurvy f=25
filthy f=20
arrant f=16
beggarly f=10
rascally f=11
drab f=10
bawd f=8
lousy f=8

lâche f=188
coquin f=181
pendre f=177
esclave f=142
vil f=142
drôle f=111
bâtard f=83
peste f=66
villain f=59
maraud f=48
paysan 47
fripon f=46
poltroon f=42
vulgaire f=37
gredin f=25
menteur f=24
sale f=23
ignoble f=21
filou f=19
abject f=13
fieffé f=12
pendard f=9
fripponnerie f=6

негодяй =163
жалкий =139
трус =118
гнусный =102
плут =97
подлый =88
раб =86
мерзавец =85
подлец =82
чума =56
грязный =55
мерзкий =52
презренный =51
херес =39
подлость =35
негодный =28
паршивый =20
низость =19
отъявленный =18
поганый =12
вшивый =8

В английском списке отметим два слова dog и cur с общим смыслом «собака». В остальном же списки хорошо совпадают друг с другом. Кроме слов, применявшихся к людям, здесь присутствуют междометные обороты (plague, peste и *чума*) и слова с семантикой пьянства (sack и *херес*).

Три списка в основном совпадают и в кластере вокруг центрального слова «слезы»:

[poor f=606]
tears f=260
weep f=260
grief f=241
sorrow f=217
alas f=212
joy f=202
woe f=147
drop f=135
sigh f=111
[wash f=88]
grieve f=85
groan f=71
wretched f=67
misery f=65
[shed f=55]
mourn f=50
lament f=44
[salt f=41]
unhappy f=40

pauvre f=534
pleurer f=338
larme f=328
hélas f=312
douleur f=297
[joie f=246]
malheur f=244
malheureux f=234
misérable f=227
chagrin f=219
triste f=183
pousser f=162
[verser f=143]
[couler f=116]
cri f=108
soupir f=93
somber f=83
plaindre f=78
gémir f=75
misère f=64

[o! f=1388]
слезы f=360
горе f=234
беда f=211
плакать f=205
скорбь f=179
увы f=135
печаль f=109
печальный f=87
страдать f=78
вздох f=75
поток f=66
вздыхать f=61
страданье f=59
тоска f=58
скорбеть f=57
рыдать f=39
оплакать f=37
стон f=34
скорбный f=26

[1] Опустим перечисление текстуальных связей.

miserable f=34	tristesse f=63	влага f=19
woful f=33	gémissement f=56	горестный f=19
wail f=26	essuyer f=52	горесть f=19
piteous f=23	mélancolie f=48	плач f=14
moan f=19	soupirer f=46	рыдание f=12
rue f=18	[humide f=42	стенанье f=11
lamentable f=17	infortuné f=40	слезинка f=8
wet f=15	douloureux f=32	унынье f=8
	deuil f=31	горючий f=4
	déplorer f=24	
	attendrir f=23	
	lamentable f=19	
	lamenter f=13	
	sangloter f=7	

Само собой разумеется, локальные участки семантической сети текстуальных связей могут сильно отличаться в языке оригинала и в языке перевода. Хороший пример такой ситуации видим в семантическом кластере ума и глупости в текстах Шекспира:

fool f=404	motley 14, wise 10, folly 9, wit 8, mad 7, ass 6, foolery 6, wisdom 5, jester 4 (+ laugh 6);
wit f=284	wise 9, fool 8, witty 4, folly 3 (+ hovel 6, jest 6, flout 5, blunt 4, writer 4);
mad f=219	madness 8, brain-sick 7, fool 7
wise f=200	fool 10, wit 9, folly 7, witty 6, foolery 3 (+ judgment 6, virtuous 5, gravity 4, ignorant 4, worse 4)
wisdom f=106	fool 5
ass f=87	fool 6;
folly f=83	fool 9, wise 7, wit 3(+ commit 4, guide 4)
madness f=64	mad 8
madman f=23	lunatic 6, mad 4
foolery f=18	fool 6, wise 3
witty f=17	wise 6 wit 4
motley f=9	fool 14
jester f=8	fool 4
lunatic f=6	madman 6, jealous 5

В русских переводах эти смыслы распределены по двум кластерам — ум/глупость:

ум =283	умный 6, глупец 3, осел 3, острота 3 (+ тонкий 7, хватит 5, отзыв 4, повредить 4, природа 4, служенье 4
дурак =152	дурость 11, умный 10, шут 8, осел 7, умник 7, дурачество 6, остроумие 4 (+ полезть 6, болтать 5, действительно 5, круглый 5, скотина 5);
глупый =142	глупец 5, умный 4, дурак 3
шут =107	пестрый 17,[1] дурак 8, шутовство 6, дурачиться 4, умник 4 колпак 3 (дяденька 12, ростовщик 5);
умный =95	дурак 10, ум 6, умник 5, глупый 4, глупость 3 (+ добродетельный 6, человек 6, судить 4, считать 4);
осел =75	дурак 7, ум 3 (ноша 11б вол 10б кабы 6, лошадь 4);
глупец =61	бесчувственный 7, глупый 4 ум 3 (+ придать 5, подобать 4)
глупость =47	умник 14, умный 3
дурацкий =25	дурак 3
дурачество =10	дурак 6, умный 3;

и разум/безумие:

мысль =292	безумье 3, бессмыслица 3, разум 3, смысл 3….
безумный =111	безумье 8, безумец 6, рассудок 5, хитрец 5

[1] Результат уловок переводчика. У Шекспира motley — наряд шута, в словарях также «пестрый».

безумье =94 безумный 8, разум 7, бессмыслица 6, рассудок 4, безумец 3, мысль 3 (+ прописать 8, сравниться 6, усвоить 5, утеха 5, впасть 4, совет 4, явный 4);
разум =92 безумье 7, безумный 4, смысл 4, мысль 3 ум 3 (+ страсть 6);
смысл =91 здравый 23, бессмыслица 6 бред 6 разум 4 (+ понять 6)
рассудок =57 безумный 5, безумье 4 (+ вопреки 6, слепой 5);
безумец =47 безумный 6, безумье 3 (+ прикрыться 5);
бред =15 смысл 6

Во французском переводе глупость в собственном смысле слова (sot, imbécile, âne) отделена от комплекса «ум/глупость/шутовство», выраженных четырьмя словами:

fou f=413 bigarré 12, folie 12, sage 9
folie f=175 fou 12, sagesse 6
sage f=174 fou 9, sagesse 5 folie 4 (+ conseil 6, discuter 4, juge 4, vertueux 4);
sagesse f=88 folie 6, sage 5;

Французское bigarré, как уже упомянутое *пестрый,* вводится для передачи безэквивалентного английского слова:

«Как вам это понравится» 2.7

A fool, a fool! I met a fool I’ th’ forest,
A motley fool. A miserable world!
As I do live by food, I met a fool,
Who laid him down and bask’d him in the sun,
And rail’d on Lady Fortune in good terms,
In good set terms – and yet a motley fool.

Un fou! un fou!... J'ai rencontré un fou dans la forêt, un fou en habit bigarré. O misérable monde! Comme il est vrai que je vis de nourriture, j'ai rencontré un fou qui s'était couché par terre, se chauffait au soleil, et invitait dame Fortune, mais en bons termes et bien placés, et cependant un vrai fou qui en portait la livree.

Шут! Шут! Сейчас в лесу шута я встретил!
Да, пестрого шута! О жалкий мир!..
Вот как живу я, - пищею шута!
Лежал врастяжку и, на солнце греясь,
Честил Фортуну в ловких выраженьях,
Разумных, метких - этот пестрый шут.

When I did hear
The motley fool thus moral on the time,
My lungs began to crow like chanticleer
That fools should be so deep contemplative;
And I did laugh sans intermission
An hour by his dial. O noble fool!
A worthy fool! Motley’s the only wear

Quand j'ai entendu ce fou bigarré moraliser ainsi sur le temps, mes poumons se sont mis à chanter comme le coq, de voir des fous si profonds en morale; et j'ai ri sans relâche, pendant une heure entière à son cadran. O noble fou! un digne fou! Oh! un habit bigarré est le seul que l'on doive porter.

……. Когда я услыхал,
Как пестрый шут про время рассуждает,
То грудь моя запела петухом
О том, что столько мудрости в шутах;
И тут смеялся я без перерыва
Час по его часам. О славный шут!
Достойный шут! Нет лучше пестрой куртки!

O that I were a fool!
I am ambitious for a motley coat.

Oh! si je pouvais être un fou! J'aspire à porter un habit bigarré.

... О! Будь я шутом! Я жду как чести пестрого
камзола!

Invest me in my motley;

Revêtissez-moi de mon habit bigarré,

Оденьте в пестрый плащ меня! Позвольте
3.3

Will you be married, motley?

Voulez-vous être marié, fou?

Ты хочешь жениться, пестрый шут?

«Двенадцатая ночь» 1.5

Lady, “Cucullus non facit monachum”; that’s as much as to say as I wear not motley in my brain. Good Madonna, give me leave to prove you a fool.

Madame, cuclus non facit monachum; c'est comme qui dirait, je ne porte pas d'habit de fou dans le cerveau. Bonne madonna, donnez-moi la permission de prouver que vous êtes une folle.

Величайшее недоразумение! Госпожа, cucullus non facit monachum;
другими словами - в мозгу у меня не пестрые тряпки. Добрейшая мадонна,
разрешите мне доказать, что вы глупое создание.

«Генрих восьмой» pr.

Only they
That come to hear a merry bawdy play,
A noise of targets, or to see a fellow
In a long motley coat guarded with yellow,
Will be deceiv’d.

Ceux-là seulement qui viennent pour entendre une pièce gaie et licencieuse,
et un bruit de boucliers, ou pour voir un bouffon en
robe bigarrée, bordée de jaune, seront trompés dans leur attente;
.
. И только разве тот,
  Кто ради сальностей сюда придет,
  Или боев с мечами и щитами,
  Иль сцен забавных с пестрыми шутами,
  Обманется.
«Король Лир» 1.4

The sweet and bitter fool
Will presently appear;
The one in motley here,
The other found out there.

  Le fou débonnaire et le fou mordant
  Seront aussitôt en présence:
  L'un ici en habit bigarré,
  Et on trouvera l'autre là

И тот и тот - дурак:
 Тот горек, сладок тот;
 Один нашел колпак,
 Другой еще найдет.

В отличие от поля «лица и тела», хорошо представленного в поэзии Пушкина и Мицкевича (см. рис. 10), в драмах Шекспира (и в переводах) оно распадается на части. Соответствующие существительные теряют взаимные текстуальные связи; теперь они соединяются теснейшим образом:
с прилагательными цвета (hair, beard — white, brown; cheveu, barbe — blanc, brun; *волосы*, *борода* — *седой*; cheek — red, white, pale; joue — rose, blanc, couleur; *щека* — *бледный*;),
с семантически связанными глаголами (smell — nose; *нюхать*, *чуять* — *нос*; see, look, gaze — eye; yeux — voir; *видеть*, *глядеть* — *глаза*; hear — ear; entendre, prêter — oreille; *слышать* — *ухо*; frown — brow; *хмурить* — *бровь*; speak – tongue; parler, prononcer – langue; kiss – lip, hand; baiser – lèvre, main; *целовать* – *уста*, *губы*, *рука*; shake, hold – hand; serer – main; tread – foot; *ступать* – *нога*; stop – mouth, ear; *заткнуть* – *рот*, *ухо*; fermer, ouvrir -- .yeux, bouche; *разинуть* – *рот*; *сомкнуть* – *глаз*, *очи*; wink – eye; *кусать* – *губы*, *зубы*; *пронзить* – *грудь*, *сердце*;pendre – cou; *сломать*, *свернуть* – *шея*; качать, отрубить – голова; скалить, грызть – зубы;),
с существительными из других кластеров (music – ear; musique – oreille; wrinkle – brow, eye; rides – front; ring – finger; *кольцо*, *перстень* – *палец*; love – heart; *любовь* – сердце; *румянец* – *щека*).

Прямо противоположную тенденцию демонстрируют громадный «топический» кластер бурного моря (в английском — 70 разных слов, во французском — 47 слов, в русском — 65 слов):

sea f=255	mer f=260	море f=232
wind f-217	vent f=205	ветер f=177
water f=150	eau f=154	корабль f=139
drop f=135	souffler f=136	вода f=132
blow f=134	vaisseau f=135	берег f=120
drown f=95	tempête f=90	буря f=110
ship f=92	flot f=97	волна f=80
shore f=84	ravage f=79	грома f=63
loud f=77	voile f=79	морской f=62
storm f=68	bord f=78	капля f=57
sail f=67	noyer f=76	дождь f=52
dreadful f=60	goutte f=66	дно f=51
thunder f-57	pluie f=55	плыть f=40
flood f=56	rocher f=50	река f=40
rain f=55	océan f=49	скала f=40
rock f=54	port f=48	туча f=39
swell f=54	orage f=47	парус f=38
threaten f=50	tonnerre f=47	нестись f=35
bark f=49	engloutir f=46	океан f=33
sink f=49	vague f=46	ручей f=32
tempest f=49	sable f=44	суша f=32
flow f=48	côte f=43	бурный f=31
sky f=45	flotte f=38	поглотить f=31
tide f=43	naufrage f=34	судно f=31
wreck f=41	vapeur f=34	теченье f=28
calm f=39	embarquer f=32	реветь f=27
wave f=38	barque f=29	дуб f=27
perish f=37	courant f=27	флот f=27
vessel 37	matelot f=26	прилив f=26
aboard f=33	navire f=26	дуть f=25
weather f=29	torrent f=25	гавань f=23
bank f=28	marée f=24	вихрь f=22
fleet f=28	onde f=24	разбойник f=22
sand f=28	Neptune f=23	тонуть f=22
ocean f=26	foudre f=22	песок f=21

deck f=25
harbour f=25
split f=25
board f=24
ragged f=24
Neptune f=23
coast f=22
ebb f=22
port f=20
rib f=20
swim f=20
lightning f=19
vanish f=18
boat f=17
sailor f=17
pirate f=16
anchor f=16
shelter f=15
whistle d=15
surge d=13
gust f=11
mast f=11
pilot f=11
billow f=10
navy f=10
drench f=9
flash f=9
mariner f=8
tackle f=8
reave f=8
beach f=7
sloth f=6
boatswain f=5
keel f=5
rig f=5

mât f=17
mugir f=17
rivière f=17
pirate f=16
gouffre f=14
pilote f=14
ancre f=13
cordage f=11
détresse f=11
tourbillon f=11
baie f=10
ouragan f=9

ураган f=21
молния f=20
бушевать f=19
Нептун f=19
отплыть f=19
пучина f=19
матрос f=17
моряк f=17
попутный f=17
утонуть f=17
якорь f=17
утес f=16
крушенье f=15
пират f=15
греметь f=14
мачта f=14
вал f=13
лодка f=13
порт f=11
борт f=10
канат f=10
хлестать f=10
мель f=9
палуба f=9
снасть f=9
прибой f=8
пристань f=7
кормчий f=6
отмель f=5
стихнуть f=5

Итак, средний интервал дает нам весьма разнообразную информацию. Частично эта информация наследуется от меньших интервалов. Чем шире семантика слова: чем ближе оно к статусу грамматического слов, тем меньше у него шансов получить новые текстуальные связи в среднем интервале. Рассмотрим несколько примеров из подкорпуса «Федеральная политика» (3,1 млн словоупотреблений) комплекта «Независимой газеты за 1996–2000 гг. У слова *ведь* (f=545) обнаружилась всего одна связь — со словом *именно* (S=5), у слова *возможно* (f=1435) со словом *вполне* (S=31). Целиком от минимального интервала унаследованы связи слова *внимание* (f=1096): *обратить, обращать, уделять, уделить, пристальный, привлечь, привлекать, акцентировать, заслуживать, особый, сосредоточить, читатель, обойти, повышенный, приковать, принимать* (S>4).Впрочем, даже текстуальные связи предлогов могут приобретать здесь весьма специфический (и интересный) характер в зависимости от специфики корпуса. Так, среди связей слова *вокруг* (f=908) легко выделяются три вполне определенные группы: а) *сплотить(ся), сплочение, группироваться, консолидировать, консолидация , концентрироваться, объединить(ся);*
б) дискуссия, развернуться, разгореться, спор;
в) ажиотаж, скандал, шумиха;
к ним примыкает менее определенный фон: *борьба, дебаты, интрига, палец, полемика, ситуация, складываться, фигура.*

Лексические маркеры[1] подкорпуса дают максимальное число текстуальных связей[2], тем большее, чем более специфичен данный маркер. Ср.
власть (f=8509, n=64) исполнительный (S=83), ветвь (44), орган (40), законодательный (26), партия (22), местный (21), оппозиция (19), представительный, вертикаль, эшелон, приход, федеральный, разделение,

[1] См. ниже (с. 110).
[2] Обозначены символом n в нижеследующих примерах при S>3.

субъект, предержащий, общество, имущий, самоуправление…
Ельцин (f=8344, n=66) президент (40), встреча (19), Кремль (18), окружение (18), отставка 16, преемник (15), Зюганов (13), импичмент (13), президентский, пресс-секретарь, здоровье, победа, указ, вчера, глава, команда, 1996, болезнь, Ястржембский, отрешение, тур, уход, Черномырдин…
партия (f=4905, n=110) движение (40), аграрный (33), КПРФ (28), съезд (27) коммунистический (26), Лысенко (26), партийный (26), демократический, лидер, политический, власть, организация, Лапшин, объединение, социал-демократический, выборы, НДР, социалистический, Яблоко, АПР…
Зюганов (f=3670, n=49) КПРФ (48), лидер (31), коммунист 29, ЦК (24), господин (20), Явлинский (19), Жириновский (18), компартия (17), народно-патриотический, Купцов, НПСР, оппозиция, Ельцин, левый, пленум, Селезнев, коммунистический, кандидат, победа, тур, Тулеев, Подберезкин…
Черномырдин (f=3189, n=35) премьер 40, Чубайс 31, НДР 26, правительство (20), дом (19), премьер-министр (18), экс-премьер (15), кабинет, кандидатура, пост, отставка, ЧВС, Лужков, Рыжков, Ельцин, отпуск, Гайдар, Гор, ямало-ненецкий, Аяцков, возвращение, встреча…
Путин (f=3023, n=39) и.о. (13), президент (11), ФСБ (11), премьер (10), премьер-министр (8), читатель (8), преемник (7), тур (7), победа (6), исполнять, обязанность, популярность, встреча, директор, президентский, Ельцин, Кремль, Лубянка, назначение, опрос, Патрушев, пост, рейтинг…
Лебедь (f=2377, n=36) генерал (52), красноярский (24), Лужков (21), Красноярск (19), КРО (19), РНРП (18), край (17), Явлинский (16), секретарь, Жириновский, честь, безопасность, генерал-губернатор, губернатор, родина, Скоков, народно-республиканский, Рогозин…
Примаков (f=2346, n=32) иностранный (16), МИД 16, Козырев (15), Лужков (15), дипломатия (11), правительство (11), премьер (11), премьер-министр (10), дело, кабинет, НАТО, разведка, министр, ОВР, внешнеполитический, внешний, Маслюков, пост, восток, СВР, экс-премьер, Явлинский…
Лужков (f=2312, n=38) мэр (78), Москва (33), московский (32), отечество (28), столичный (23), Лебедь (21), Примаков (15), столица (15), градоначальник, Севастополь, москвич, левоцентристский, хозяйственник, кепка, Черномырдин, Зюганов, Строев, Кремль, кандидат, метро, политик…
бюджет (f=2238, n=131) расход (50), доход (38), бюджетный (34), расходный (34), налог (33), рубль (30), доходный (29), триллион (29), федеральный, чтение, налоговый, правительство, год, поступление, согласительный, финансирование, дефицит, долг, ВВП, секвестр, миллиард…
право (f=2232, n=68) гражданин (29), свобода (22), конституция 19, устанавливать (18), закон (17), конституционный (17), уполномоченный, защита, нарушать, норма, собственность, субъект, ущемлять, человек, самоопределение, избирательный, суд, иметь, гарантия, нарушение…
Гайдар (f=838, n=82) ДВР (51), Чубайс (35), Демвыбор (30), Бурбулис (23), Немцов (22), демократический (19), выбор (17), 1992 (15), демократ (15), выборосс, Федоров, Явлинский, либеральный, КПСС, Мурашев, правый, партия, демвыборосс, дословный, коалиция, Сакс, Хакамада, Кириенко, либерал…

В подкорпусе «Северный Кавказ» (1,5 млн словоупотреблений) отметим два самых частых слова:
боевик (f=2824, n=106) район (23), удар (18), отряд (17), населенный (15), пункт (15), бой (14), войска (13), село (13), селение (13), федеральный, авиация, захватить, скопление, командование, подразделение, уничтожить, артиллерия, Бамут, Грозный, группа, блокировать, база, вертолет…
Масхадов (f=2673, n=152) президент (34), Басаев (25), встреча (25) переговоры (24), лидер (18), Яндарбиев (14), полевой (13), чеченский (13) встретиться, Ельцин, заявление, командир, Примаков, Чечня, Лебедь, Москва, Арсанов, окружение, Грозный, контакт, покушение, шура, намерен…

В подкорпусе «Страны СНГ» (3,2 млн словоупотреблений) характерны слова:
Кучма (f=2233, n=51) Украина (53), президент (40), украинский (32), Киев (28), рада (25), Мороз (23), Ивженко (22), Марчук (21), Лазаренко, верховный, президентский, Тимошенко, Ельцин, администрация, выборы, Квасьневский, Кравчук, окружение, победа, премьер, Симоненко…
Крым (f=1319, n=81) крымский (67), полуостров (54), Лебедев (44), автономия (41), Симферополь (41), ВС (38, Севастополь (38), Грач (35),

Украина, татарин, Франчук, автономный, крымскотатарский, крымчане, верховный, президиум, конституция, депутат, русский, рада, сессия…

В подкорпусе «История» (800 тыс. словоупотреблений) интересны:

Сталин (f=1459, n=132) товарищ (23), Молотов (19), Гитлер (17), Ленин (14), Жуков (13), Косолапов (12), Аргентина (11), Хрущев (11), Ворошилов (10), Каганович, культ, браво, вождь, Германия, Мухин, Рузвельт, сталинский, 1953, Берия, маршал, Резун, Троцкий, упреждать, диктатор, доложить…

Ленин (f=396, n=63) Сталин (14), ленинский (13), Ульянов (13), Маркс (11), Плеханов (11), вождь (10), диплом (10), Ильин (10), Красин (10), Троцкий, послеоктябрьский, революция, Энгельс, большевик, Крупская, Мартов, НЭП, партия, Совнарком, социализм, 1922, Арманд, Горки

Горбачев (f=370, n=71) Ельцин (20), КПСС (18), Буш (16), Шульц (15). Шеварднадзе (14), генсек (13), Лигачев (13), перестройка (13), Лукьянов, политбюро, Рейган, Рыжков, Яковлев, ЦК, Бейкер, Болдин, президент, реформировать, секретарь, демплатформа, 1990, союзный, Черняев…

1941 (f=217, n=97) июнь (26), упреждать (15), май (14), удар (13), Барбаросса (13), война (11), генштаб (11), 1942 (10), войска (10), готовность, директива, нанесение, округ, боевой, гитлеровский, Резун, армия, Германия, Мухин, приграничный, фронт, бомбить, ВМФ, военный…

Наконец, рассмотрим текстуальные связи трех слов из подкорпуса «Эссе».[1]

государственный (f=2141, n=33) дума (32), орган (21), устройство (21), строительство (18), регулирование (17), власть (14), собственность, управление, институт, высший, федерация, переворот, РФ, самоуправление, экономика, государство, депутат, машина, система, совет, функция…

русский (f=1870, n=179) националист (18), язык (16), национализм (14), татарин (14), государствообразующий (13), культура (13), нация (13), православный, интеллигенция, украинец, возрождение, эмиграция, еврей, литература, мыслитель, православие, Русь, этнический, великий…

православие (f=114, n=40) а) религия (27), христианство (20), ислам (18), конфессия (18), протестантизм (18), католицизм (16), православный (15), церковь, буддизм, русский, секта, доминанта, Византия, духовный, исповедовать, корень, культура, знаменитый, религиозный, традиционный…
б) самодержавие (45), народность (43), триада (23), уваровский (18),Уваров (15), формула (6).

1.5.3. Большой и максимальный интервалы

Из общих соображений можно предположить, что текстуальные связи слов в большом интервале будут определяться тематическими и/или жанрово-стилистическими факторами. Единственный (и для меня — первый машинный) эксперимент был осуществлен на материале атрибутивных слов в английских текстах XVI–XVII вв. [Шайкевич, 1969]. Фрагмент длиной в 1000 слов здесь создавался путем объединения десяти фрагментов по 100 слов.[2] На рис. 12 представлена большая группа слов, возглавляемая словами great и other. Из этой группы были отобраны самые частые слова таким образом, чтобы самое редкое из отобранных слов составляло не менее 1% всей суммы частот отобранных слов. Всего таких репрезентирующих слов оказалось 15: great f=1548, other 687, same 234, whole 224, small 191, said 182, certain 154, chief 123, diverse 121, English 115, ancient 104, present 97, former 92, like 86, sundry 82. Этот набор хорошо согласуется со списком диагностирующих слов делового функционального стиля (см. ниже п. 1.6).

---

[1] Статьи на общие темы (часто — историософского характера, 1,5 млн словоупотреблений).

[2] Строго говоря, это было некоторое нарушение основной идеи интервала, предполагающей членение связного текста на фрагменты. Это отступление от общего принципа мешает выявлению текстуальных связей, а значит, результаты эксперимента, полученные вопреки действию этого фактора, становятся тем более убедительными.

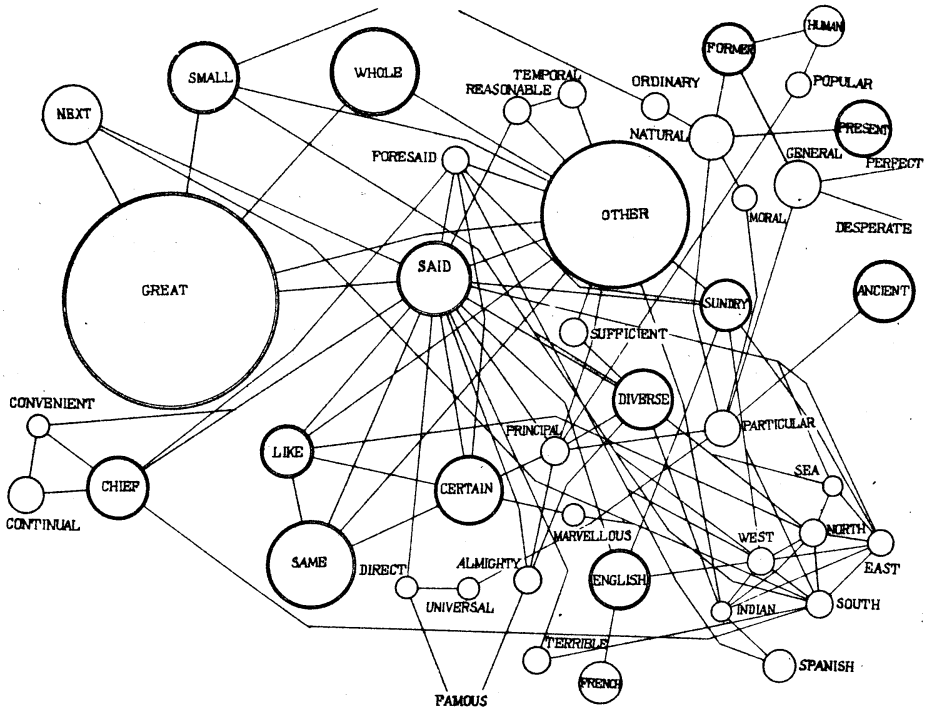

Рис. 12

Доля этих «репрезентантов» в общей сумме атрибутивных слов особенно высока в сорока текстах, среди которых находим «Очерки» Бекона, Конституционнвные документы, «Путешествие Дженкинсона в Бухару», «Три путешествия Хокинса в Вест-Индию», «Письма Томаса Вуда, пуританина», но также Новый Завет и (еще более любопытно) «Процесс дьявола» Вебстера.

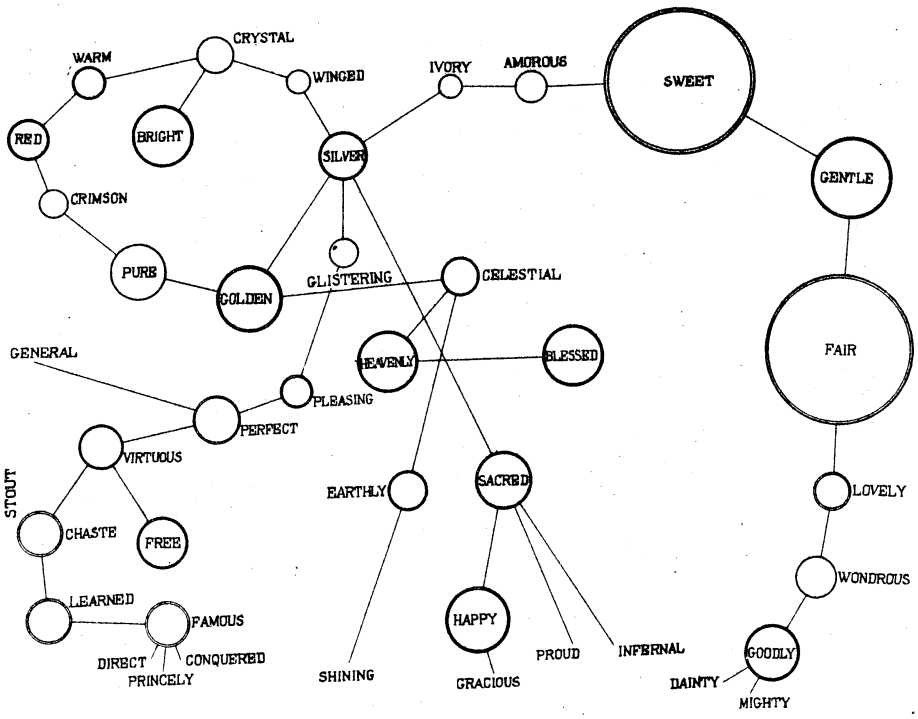

Рис. 13

На рис 13 можно выделить цепочку от amorous до goodly. Эту группу репрезентируют пять слов: fair 743, sweet 715, gentle 205, goodly 91, lovely 59 (группа 2а).

На том же рисунке представлена еще одна группа (группа 2б), которую репрезентируют следующие слова:

golden	147	sacred	98	perfect	64	earthly	43
happy	138	pure	95	silver	63	celestial	33
bright	122	free	78	virtuous	62	pleasing	32
heavenly	116	famous	65	chaste	59	warm	32
blessed	115	learned	64	red	44		

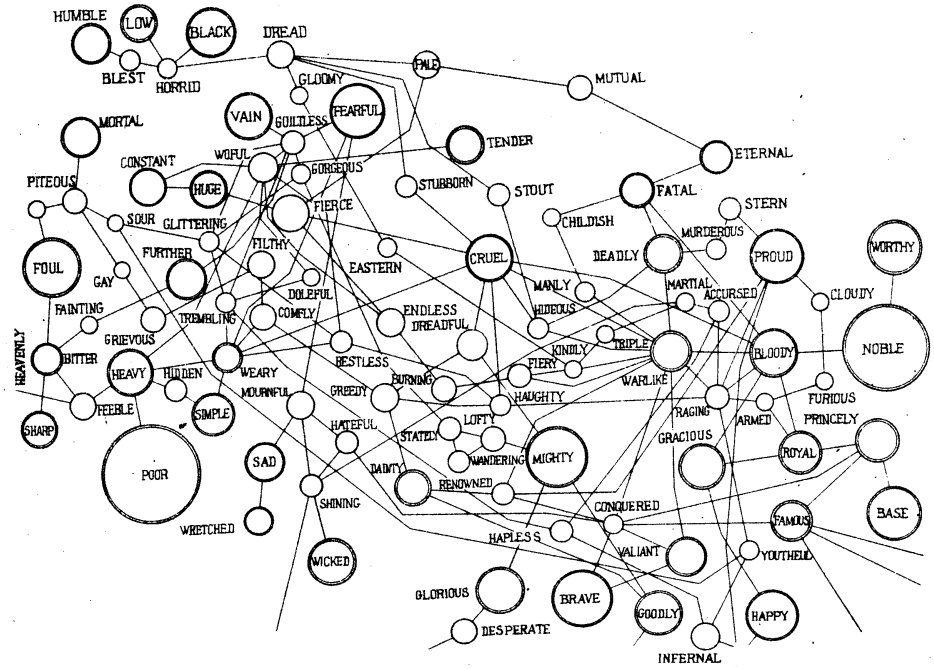

Рис. 14

Сложнее выделить группы на рис. 14. Однако и здесь проглядывают две группы. Справа видна группа (группа 3а), репрезентируемая словами:

noble	375	proud	123	gracious	79	deadly	58
brave	188	bloody	109	princely	77	dainty	48
mighty	167	cruel	109	famous	65	fatal	47
happy	138	royal	104	valiant	63	dreadful	44
base	127	glorious	94	warlike	61		
worthy	126	goodly	91	eternal	58		

Наконец, в левой части рис. 14 появляется последняя группа (группа 3б). Ее репрезентируют такие слова:

poor	479	wicked	93	wretched	66	sharp	58	dreadful	44
foul	149	heavy	92	low	64	humble	57	weary	43
sad	143	vain	84	further	62	bitter	53		
black	115	simple	77	fearful	60	huge	50		
cruel	107	mortal	71	tender	60	constant	49		

Содержательно интерпретируя последние четыре группы, присвоим ярлыки четырем компонентам смысла — 1) «лирический» (группа 2а); 2) «возвышенный» (группа 2б); 3) «гражданский» («военно-феодальный») (группа 3а) и 4) «минорный». Теперь для отдельных текстов можно подсчитать долю каждого компонента (в процентах), опираясь на частоту соответствующих слов. Как пример приведем некоторые тексты Шекспира:

	Компоненты			
	1	2	3	4
«Напрасные усилия любви»	69	16	12	3
«Сон в летнюю ночь»	56	13	14	17
«Как вам это понравится»	44	12	19	25
«Венера и Адонис»	28	26	6	40
«Лукреция»	19	23	12	46

«Ромео и Джульетта	29	18	22	31
«Сонеты»	40	15	17	28
«Троил и Крессида»	35	4	43	17
«Тит Андроник»	35	-	51	14
«Король Джон»	21	5	63	16
«Ричард III»	19	7	48	26
«Генрих V»	22	6	47	25
«Юлий Цезарь»	9	-	61	30
«Антоний и Клеопатра»	16	9	69	6
«Макбет»	6	18	52	24
«Мера за меру»	12	7	36	45
«Король Лир»	13	5	18	64

Как видим, данное исследование привело к разбиению атрибутивных слов, которое зависит от самых широких семантических факторов. Не следует забывать при этом, что изучаемая популяция на 70% состояла из поэтических и драматических Переход к корпусам другого типа привел бы к семантической классификации иного рода.

Какое-то представление о том, чего при этом можно ожидать, дает ответвление дистрибутивно-статистического анализа, направленное на открытие лексических маркеров (keywords) подкорпуса на фоне более широкого корпуса текстов.

Пусть все, что написал Достоевский, составит общий корпус текстов. Тогда подкорпусами могут считаться 1) проза, 2) критика и публицистика, 3) письма. Зная частоту слова в общем корпусе и частоту слова в подкорпусах, а также зная долю каждого подкорпуса в общем корпусе, легко подсчитать математическое ожидание этого слова в подкорпусе (в предположении, что это слово равномерно представлено во всех трех подкорпусах). Затем снова применим нашу основную формулу. Частота слова *деньги* в текстах Достоевского равна 2097, в письмах оно встретилось 822 раза; доля писем в общем корпусе составляет 0,183, m=284. Подставляя величины 822 и 284 в нашу формулу, получаем S=31 (величину исключительно высокую), отсюда вывод — слово *деньги* очень характерно для писем Достоевского. Механика вычисления очень проста, для каждого отдельного слова задача решается тривиальным образом. Но что можно сказать о совокупности лексических маркеров? Образуют ли они хоть в какой-то степени систему? Сравним списки лексических маркеров с максимальными значениями S в трех «макрожанрах» Достоевского [ССЯД, с. XIII].

Таблица 1.17

Художественная литература	S	Критика и публицистика	S	Письма	S
он	44	всё	74	письмо	72
она	28	народ	62	писать	70
-с	25	мы	59	Достоевский	59
князь	24	наш	58	мой	58
вдруг	22	русский	50	роман	53
был	21	Европа	49	написать	52
я	20	народный	37	рубль	48
Алеша	17	Франция	36	Ф.	47
Лиза	16	Россия	35	получить	46
лицо	16	война	34	я	45
рука	16	европейский	31	ты	42
глаз	15	лишь	30	Аня	41
так	15	русский (сущ.)	29	Петербург	38
дверь	14	славянин	29	твой	38
комната	14	идея	23	будет	37
проговорить	14	литература	23	Паша	37
вскричать	13	новый	23	выслать	35
да (утверждение).	13	общество	23	прислать	34
Митя	13	политический	23	ваш	33
смотреть	13	статья	23	лист	33
что-то	13	турок	23	многоуважаемый	33

будто	12	восточный	22	Катков	32
весь	12	господин	22	год	31
как	12	искусство	22	деньги	31
как будто	12	они	22	если	30

Верхушка списка маркеров писем дает некоторое представление о строении семантического пространства этого макрожанра. Эти маркеры легко группируются по общей роли в коммуникации:

мой, я, ты, твой, ваш, Достоевский, Ф., Аня, Паша, Анна Григорьевна, + Федор, Федя, брат, Лиля, Люба, Миша;[1]

письмо, писать, написать, получить, выслать + адрес, высылать, доставить, ответ, переслать, получать, послать, посылать, почта, присылка, строка, телеграмма, P.S, poste restante;

многоуважаемый + целовать, до свидания, обнимать ангел, ангел мой, бесценный, глубокоуважаемый, голубчик, дорогой, жать, искренний, кланяться, любезнейший, любезный, милый, милый друг, преданный, принять уверения, прощай;

роман, лист, Катков + журнал, редакция, «Русский вестник», «Заря», Стелловский, занят, «Идиот», издание, «Карамазовы», книга, книгопродавец, контракт, корректура, Майков, напечатать, Некрасов, печатать, печатный, печатный лист, Плещеев, повесть, подписчик, «Преступление и наказание», работа, работать, сочинение, срок, статья, Страхов, типография;

рубль, деньги + вперед (врем.), заплатить, кредитор, марка, серебро, счет, талер, уплата, франк;

Петербург + Старая Руса, Барнаул, Дрезден, Женева, Москва, Семипалатинск, Тверь, Флоренция, Эмс

будет, если + во всяком случае, дай Бог, надеяться, ;

год, август, июль + воскресенье, вторник, декабрь, день, июнь, лето, май, март, месяц, неделя, октябрь, осень, понедельник, пятница, сентябрь, среда, суббота, теперь, февраль, четверг, число, январь

+ просьба, просить, ради Бога, помочь;

+ здоров, здоровье, лечение, нервы, падучая, припадок

Характерные лексические маркеры публицистики сводятся к одной главной идее — МЫ.vs ОНИ

народ, мы, наш, русский, народный	Европа, Франция, европейский,
Россия, русский (сущ.), славянин,	они, Германия, + Австрия, Англия,
идея, нация, идеал, исторический,	европеец, католический,
великий, народность, почва, принцип	католичество, папа, революция
сила, славянский, христианство, царь	римский

За пределами этой главной оппозиции остается еще много — и текущая политика, и Восточный вопрос, и война, и литература.

Что касается подкорпуса художественной литературы, то и там, конечно, можно было бы обнаружить какие-то магистральные линии, но материала Достоевского здесь явно недостаточно для серьезных утверждений. Необходимо учесть еще одно важное обстоятельство. Чем больше доля данного подкорпуса в общем корпусе, тем более прихотливыми и случайными становятся лексические маркеры, тем они менее системны, тем меньше их число. Проблема обсуждается в публикации 2001 года «Contrastive and Comparable Corpora: Quantitative Aspects» [Shaikevich, 2001]. Рассмотрим три примера из этой статьи.

В общей совокупности текстов лондонской «The Times» за первый квартал 1995 г. (около 10 млн словоупотреблений) письма читателей составляют 3,2%. Тот факт, что этот подкорпус представляет собой отклик на самые разные темы, обсуждавшиеся в газете, объясняет относительно малое число лексических маркеров (всего 1554 при S>2) и удивительную бедность начала списка маркеров. Ср.

I S=34, my 24, am 18, our 33, we 18;

[1] Знак + вводит дополнительные слова, не вошедшие в малый список таблицы, но обладающие высокими значениями S>12.

sir 108, your 22;
yours 196, faithfully 179, sincerely 89, truly 11;
letter 44, letters 30, article 45, report 38, editorial 11, public 10;
February 75, January 77, March 60, December 17;
Road 35, Nr 21, Lane 19, Avenue 17, Street 17, House 14;
Surrey 19, Sussex 17, Oxfordshire 15, Dorset 14, Kent 14, Hampshire 13;
SW 36, EC 19, WC 17, NW 14, W1 14, Eć 12, NW 11, SW10 12, WC2 11…

Две последние строки многое скажут о географическом распределении круга читателей газеты — южные графства и запад Лондона.

Рубрика «Court Circular», посвященная королеве, составляет лишь 0,2% общего корпуса. Около 700 разных слов показали S>2, начало списка маркеров демонстрирует исключительно высокие значения S, а сам список естественным образом членится на осмысленные группы:

patron 156, royal 130, queen 106, princess 100, prince 75;
duke 173, countess 53, baroness 29, lady 27, duchess 20, lord 18, earl 16;
highness 260, majesty 165, majesty's 144, excellency 144, hon 89, hmy 63;
Buckingham 161, Palace 154, James' 153, Kensington 56;
attendance 255, attended 145, accompanied 27, lady-in-waiting 12;
lord-leutenant 169, governor-general 53, counselor 47;
insignia 44, appointment 42, investiture 44, award 20;
thanksgiving 24, preached 21, commemorative 19, sermon 16;
plenipotentiary 84, ambassador 34, attaché 28, credence 24;
afternoon 118. luncheon 87, morning 84, evening 75;
visited 127, received 107, reception 48, opened 36, bade 34;
deafblind 34, Fund's 20, fellowship 24, memorial 22, trust 21.

Церемониальные функции монарха описываются лексическими маркерами с исчерпывающей полнотой.

Футбольный подкорпус составляет 5% от объема общего корпуса. Лексические маркеры полностью включают все специфические слова этого популярнейшего вида спорта. Ср. начало списка маркеров:

goal	123	goalkeeper	72	0	57	team	46
ball	95	league	72	keeper	56	half-time	44
football	86	referee	72	shot	55	kick	42
game	82	manager	71	defender	54	cross	41
minute	81	midfield	71	season	54	equaliser	41
player	79	header	68	scored	50	penalty	41
cup	73	club	62	match	46	supporter	41

и далее:
1, 2, after, almost*, angle, area, attack, bench, blocked, bookable, booked, but*, byline, chance, cleared, corner, cross-shot, cross-bar, crossed, crowd, cup-tie, curle, defeat, deflected, division, down, draw, dribble, fan, FIFA, first-half, flank, flick, foot, forward, foul, free, free-kick, from, goalless, goalmouth, half, hat-trick, headed, home, hooligan, inside, into, just, kick-off, lads*, left, legs, matches, midfielder, moment, net, pass, play, police, post, rebound, replay, right, right-back, right-foot, rush, save, saw*, scoring, sent, shirt, shoot, side, stadium, tackle, their*, throw-in, tie, towards, transfer, victory, volley, was*, whistle, won, yellow.

Большинство перечисленных маркеров прямо передают важнейшие смыслы этой игры, но некоторые из них объясняются скорее жанром газетного репортажа (отмечены * в списке) или связаны с внешними социальными обстоятельствами (hooligan, police).

Совместное появление слов в тексте (максимальный интервал) или шире — в текстах и даже в корпусах текстов, вплотную подводит нас к обсуждению проблемы соотношения лингвистических единиц.

1.6. Лингвистические единицы и тексты

В рамках дистрибутивно-статистического анализа исходными данными являются тексты, к исходным данным относятся также графические слова — прообразы собственно лингвистических единиц. По мере работы ДСА, в ходе уточнения репертуара лингвистических единиц, формирования дистрибутивных классов и т. п. возникает возможность поставить обратную задачу — использовать накопленный лингвистический материал для классификации текстов, для выявления их внутренней структуры.

Первый опыт такого подхода был проделан в середине 1960-х гг., результаты опубликованы в статье «Опыт статистического выделения функциональных стилей» [Шайкевич, 1968]. Повторим начало той публикации, в котором ставится задача и описывается состав корпуса текстов:

Излагаемая далее процедура выделения функциональных стилей строилась «формальным» образом, т. е. без учета значения языковых единиц. (Случаи отступления от формальной процедуры будут особо оговариваться). Такая процедура должна быть пригодной и для неизвестного языка. Однако для первого опыта желательно взять известный язык, чтобы полученные результаты сопоставить с нашими интуитивными знаниями о стилях языка и проверить, насколько правдоподобны эти результаты. Таким языком послужил здесь английский язык XVI–XVII вв. (времен Елизаветы и первых Стюартов).

Общий план процедуры таков: используя статистический корреляционный анализ, попытаться разбить тексты выборки на более или менее автономные группы. Этим группам текстов и будут предположительно соответствовать отдельные стили.

Выборка производилась следующим образом. Было взято 307 разных текстов, из каждого текста путем механической выборки отбирались куски по 100 слов в каждом. Таким образом было выбрано 10000 кусков, т. е. общий объем выборки — 1 миллион слов текста.

Выборка включала: драматические произведения Гаскойна, Кида, Лили, Марло, Шекспира, Марстона, Бена Джонсона, Форда, Мессинджера, Деккера, Бомонта, Флетчера, Вебстера (всего 161 текст, или 50% всего объема выборки), поэтические произведения Гаскойна, Сидни, Чэпмена, Даниеля, Марло, Спенсера, Шекспира, Рэли, Бена Джонсона, Дрейтона, Херберта, Крешо, Донна (76 текстов, или 20% объема выборки); Библия (8% объема выборки); прозаические романы Лили и Делони (5 текстов, или 4,5% объема выборки); произведения Бекона (11 текстов, или 6% объема выборки – научные сочинения, юридические, публицистические); описания путешествий (32 текста, или 4% объема выборки); проповеди Донна; частные письма (Уоттона и Вуда); королевские указы и т. д.

Вручную было бы очень трудно обработать статистически миллион слов текста, поэтому пришлось ограничиться только одним классом слов — атрибутивными словами, стоящими в препозиции к существительному, т. е. словами, которые могут занимать положение между артиклем и существительным. В основном — это прилагательные, но к ним присоединяются причастия и существительные. (Это первое отступление от формальной методики. Для английского языка, однако, выделение подобных классов, вероятно, можно легко формализовать). Обращение к особому классу слов допустимо по следующему содержательному соображению: если стиль характеризуется разносторонними лингвистическими признаками, можно надеяться на то, что даже по части признаков удастся выделить этот стиль.

Всего в выборке оказалось 41966 атрибутивных слов, которые и послужили материалом для всех статистических выкладок. На основе статистического словаря к этой выборке была составлена матрица, столбцами которой служат тексты (234 текста), а строками — «признаки» (30 «признаков»). «Признаками» называются отдельные слова (good, great, fair, own, sweet, old, poor, noble, true, dear, little, fine, foul, hole, strange), классы слов (слова, оканчивающиеся на -ed, -y, -ing, -ous, -al, -ly, -nt, -ful, -an, -less, -ble, -ic и начинающиеся на un-, сложные слова)и общий класс атрибутивных слов… Числа на пересечении столбцов и строк указывают на относительную частоту данного признака в данном тексте.

Не будем подробно описывать дальнейший алгоритм обработки матрицы. Скажем только, что для пар столбцов подсчитывался коэффициент корреляции, на основе которого определялись расстояния между текстами. Затем строился граф, вершинами которого были тексты, а ребрами – расстояния. В дальнейшем к этому

графу применялись довольно сложные правила, направленные на получение автономных кластеров.

В результате были обнаружены три большие группы текстов.

Группа *A* включает 52 текста, из которых 34 отнесены к поэзии, а остальные – к драме (3 текста Кида, 4 Марло, 7 Шекспира, 3 Марстона). Группа *B* объединяет 45 текстов, среди которых преобладают комедии. В эту же группу входят романы Лили «Эвфуес» и «Эвфуес и его Англия». В группу *C* входит 24 текста – большинство произведений Бекона, 6 описаний путешествий, письма Вуда, проповеди Донна, конституционные документы, из драматических произведений 3 текста — «Генрих VIII» Шекспира, «Томас Вайят» и «Процесс дьявола» Вебстера.

Можно полагать, что полученная классификация текстов носит стилистический (или жанрово-стилистический) характер. Группа *A* -- поэзия (собственно поэзия и высокая трагедия), группа *B* -- отражение в письменной речи особенностей диалогической речи, группа *C* соответствует недифференцированному деловому стилю.

Принимая каждую из полученных групп за ядро соответствующего стиля можно получить списки 20 положительных и 10 отрицательных диагностирующих признаков (ДП) и подсчитать коэффициент («вес») каждого из них. Полученные ДП показаны в табл. 1.18. (Attrib. — общее число атрибутивных слов на 1000 слов текста; все остальные ДП — относительные частоты на 1000 атрибутивных слов, Deriv. — все производные атрибутивные слова).

Таблица 1.18

Индекс *A*		Индекс *B*		Индекс *C*	
признак	вес	признак	вес	признак	вес
		Положительные признаки			
Attrib.	1/3	un-	4	in-	5
Deriv.	1/3	-y	2	-al	2
-y	1	-nt	5	-nt	3
-ed	1	-ing	1	-ble	4
-ing	2	-ble	6	-te	4
-ly	3	dear	4	-ry	8
-ful	2	excellent	8	-ar	5
-less	5	fine	6	ancient	10
bloody	12	gentle	6	certain	8
cruel	16	good	4	diverse	9
deadly	18	honest	15	divine	13
fair	4	ill	10	excellent	10
foul	10	mad	22	great	2
goodly	12	noble	8	other	3
heavy	13	old	9	particular	15
lovely	10	own	3	present	10
mighty	9	poor	8	said	7
proud	11	pretty	19	same	7
warlike	17	sweet	3	small	8
weary	20	young	9	whole	7
		Отрицательные признаки			
-nt	3	Attrib.	0,2	Deriv.	0,5
-ble	6	Deriv.	0,5	-y	2
good	4	certain	12	-ed	2
great	4	English	16	-ing	2
honest	16	greaat	4	-less	6
old	6	high	11	dear	10
other	8	other	7	fair	6
own	6	said	23	gentle	14
said	19	same	10	poor	7
whole	14	small	7	sweet	6

С помощью ДП для всех текстов были подсчитаны три индекса. Группа с преобладанием индекса *A* охватывает теперь 82 (23% всего объема выборки); сюда, например, относятся:

	Индекс *A*	Индекс *B*	Индекс *C*
Спенсер «Amoretti”	180	-10	-40
«Царица фей»	160	-10	-20
Сидни «Сонеты»	160	-30	-30
Даниель «Сонеты к Делии»	160	50	-60
Марло «Тамерлан Великий», ч.1	140	-40	-40
Шекспир «Венера и Адонис»	140	30	-60
«Ричард II»	90	20	-20
Дрейтон «Матильда	140	0	-40
Лили «Женщина на луне»	120	-20	-40

Довольно часто высокими оказываются сразу два индекса — *A* и *B*, что позволяет говорить о существовании большой группы *AB* (27 текстов, 9% всего объема выборки) стилистически сложных текстов:

	Индекс *A*	Индекс *B*	Индекс *C*
Дрейтон «Нимфида»	90	130	-70
Марло «Эдуард II»	50	60	-50
Шекспир «Ричард III»	100	70	-20
«Троил и Крессида»	70	100	-10
«Сонеты»	60	50	-50
«Сон в летнюю ночь»	50	30	-40
Флетчер и Мэссинджер «Мальтийский рыцарь»	80	90	-40

Самую большую группу образовали тексты с явным преобладанием индекса *B* (105 текстов, 34% всего объема выборки:

	Индекс *A*	Индекс *B*	Индекс *C*
Лили «Матушка Бомби»	-110	160	0
«Эвфуес»	-70	90	10
Шекспир «Много шума из ничего»	-100	180	10
«Укрощение строптивой»	20	140	-10
«Виндзорские кумушки»	-150	130	0
«Король Лир»	10	100	-30
«Ромео и Джульетта»	20	80	-20
«Гамлет»	-40	70	0
Бен Джонсон «Всяк в своем нраве»	-60	130	10
Форд «Ведьма Эдмонтона»	-80	130	-30
Деккер «Праздник сапожника»	-40	230	-40
Бомонт и Флетчер «Высокомерная леди»	-70	170	-20
Флетчер «Жена на месяц»	-100	130	-20
Вебстер «Лекарство для рогоносца»	-100	130	10

Область пересечения групп *B* и *C* очень невелика (всего 5 текстов):

	Индекс *A*	Индекс *B*	Индекс *C*
Гаскойн «История Гемета»	-50	50	70
Лили «Кампаспе»	-30	40	30
Вебстер Процесс дьявола»	-40	40	30

Группа с преобладанием индекса *C* охватывает значительную часть выборки (61 текст, 16% всего объема выборки):

	Индекс *A*	Индекс *B*	Индекс *C*
Бекон «Максимы закона»	-110	-180	140
«Новая Атлантида»	-70	-140	130
конституционные документы	-170	-250	120
Вилас «Об острове Япония»	-120	-290	90
Рэли «Открытие Гвианы»	-160	-240	90
Письма Томаса Вуда, пуританина	-100	-40	60

А. Я. Шайкевич