ISSN 0494-7304 0207-4435

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

827

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1988

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1988

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 827 ВЫПУСК ОСНОВАНЫ В 1893.г

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1988

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1988

Kogumik "Kvantitatiivlingvistika ja tekstide automaatanalüüs" ilmub Tartu Riikliku Ülikooli rakenduslingvistika uurimisgrupi iga-aastase väljaandena alates 1985.a. (jätkates sarja "Töid keelestatistika alalt" I - X, mis ilmus 1976-1984). Käesolevas neljandas (14-ndas) väljaandes (1988) on avaldatud kõrgkoolide vahelise probleemrühma "Tekst interdistsiplinaarse uurimise objektina" liikmete artiklid.

Сборник "Квантитативная лингвистика и автоматический анализ текстов" публикуется Группой прикладной лингвистики Тартуского государственного университета начиная с 1985 г. (сборник является продолжением серии "Труды по лингвостатистике" I - X, 1976-1984 гг.). Настоящий 4-й (14-й) выпуск (1988 г.) содержит статьи членов Межвузовской проблемной группы "Текст как объект междисциплинарных исследований".

The collections "Quantitative Linguistics and Automatic Text Analysis" appears annually since 1985, edited by the members of the Applied Linguistics Group at Tartu State University (Estonia). The present issue No. 4 (1988) contains investigations by members of the All-Union Research Group "Text as an Object of interdisciplinary investigations".

ОБЪЕМ КОЛЛЕКТИВНОЙ СЛОВАРНОЙ ПАМЯТИ (ТРИ ПОДХОДА К ЕЕ ИЗМЕРЕНИЮ)[*]

М.В. Арапов

I. Одна из важнейших сторон культуры - накопление и передача в письменном или устном виде запаса текстов. Среди этих текстов есть художественные, юридические, мифологические и прочие. Мне представляется, что предельным случаем текста является отдельное слово. Статья посвящена вопросу, сколько таких предельно коротких текстов одновременно находятся в коллективной памяти носителей данной культуры. Прежде чем излагать свой подход к измерению ее объема, я должен уточнить принятое в этом докладе понимание текста и памяти.

Мне придется обратиться к более общему понятию, чем текст. Этим понятием будет сообщение.

Среди всего многообразия сообщений тексты выделяются не особой структурой, противопоставляющей их другим сообщениям на том же языке, а отношением к ним членов общества. Текст - это результат отбора. Данная культура отбирает сообщения, обладающие особой для нее ценностью, и присваивает им особый привилегированный статус, статус текстов.

2. В связи с изучением текстов возникает три группы вопросов. Прежде всего, это связь между ценностью данного текста и особенностями его структуры. Но я думаю, если мы будем развивать только это направление исследований, мы очень скоро зайдем в тупик. В частности, потому что, демонстрируя специфические особенности, например, лексической структуры "связного", "целостного", "замкнутого", наконец, просто "хорошего" текста, мы лишь соотносим текст с отметками на определенной шкале ценностей. Но недостаточно признать само ее существование, нужно понять ее природу и место в нашей культуре.

Философы, размышляя над корпусом сообщений, объявленных классическими текстами, пытаются вывести из них соответствующие этические и эстетические нормы. Историки искусства,

---

[*] Статья представляет собой переработанный вариант доклада на семинаре Группы по комплексному изучению текста 24.04.87 в г. Боржоми.

предполагая шкалу известной, стараются объяснить связь особенностей произведения с его местом на этой шкале. Критик пытается предсказать, какое место займет текст на этой шкале. Это вторая группа вопросов.

Но существует еще третья группа вопросов, на которой я собственно и предполагаю остановиться в данной статье.

3. Что значит, что культура отбирает сообщения? Я думаю, что сообщение становится текстом, когда размещается в коллективной памяти общества. Попав в эту память, сообщение приобретает целый ряд особых качеств.

Во-первых, очередное сообщение можно сличить с занесенным в память и решить, идентичны ли эти сообщения. Таким образом, тексты обладают индивидуальностью и образуют множество. По отношению к произвольным сообщениям это едва ли верно.

Во-вторых, превращение сообщения в текст - это событие, оно имеет свою дату, а текст тем самым - возраст. Текст может быть вычеркнут из памяти - и это тоже событие. Поток таких событий - это часть человеческой истории, сообщение же в общем случае внеположено истории.

В-третьих, коллективная память имеет вполне определенную конечную емкость. Поэтому ограничено как число хранимых текстов, так и размеры отдельного текста.

В-четвертых, кроме возраста и размера, текст характеризуется частотой обращения к нему - количественной мерой "цитируемости" данного текста.

Наконец, в-пятых, следует иметь в виду, что любая память - коллективная или индивидуальная - не есть нечто сотворенное раз и навсегда, а динамический процесс. Память не только пополняют, обновляют и извлекают из нее единицы. Требуется еще обеспечить к ней доступ. Поэтому хранимые объекты все время "тасуются": классифицируются, снабжаются новыми адресами, метками, именами, обеспечивающими систему перекрестных ссылок. Сообщение, в частности, становясь текстом, обретает (хотя и не всегда) заглавие. Но это не единственный вид меток, обеспечивающих идентификацию текстов. Вот обычный способ сослаться на текст, не цитируя его полностью. Привожу отрывок из газеты "Московские новости" (от 12.04.87): "Особенно досталось директору, кажется, из Сумгаита, которого из Москвы прилюдно, хотя и по радио, приводили в чувство известным в "определенных кругах" пассажем о роли скипидара в ускорении трудовой деятельности". Здесь "неудобный для печа-

ти" анекдот заменен его "синопсисом".

Итак, у текста есть по крайней мере пять специфических свойств, выделяющих из массы сообщений: индивидуальность, историчность, четкие пространственные границы, частота использования ("тираж") и система адресации.

4. Остановлюсь подробнее на одном из этих признаков, поскольку в рамках данного доклада он имеет особое значение. Речь идет о принципиальной ограниченности физических размеров текста и конечности запаса текстов. Сам этот тезис обычно не вызывает возражений; спорят с тем, специфично ли это именно для текстов или то же самое имеет место и для сообщений.

Каждое сообщение конечно в том смысле, что для него найдется такая величина N, что данное сообщение по своим размерам не превосходит N. Тонкость в том, что не существует единого N - некоторой универсальной константы, - которой не может превзойти ни одно сообщение. Для каждого конкретного N можно построить сообщение, превосходящее заданное N. Тексты как раз и отличаются от сообщений тем, что для них такая универсальная константа существует. Она различается для текстов разного типа, скажем, для отдельных слов и для романов. Мы не знаем точно ее значения, но она существует.

Здесь обычно возражают, что такая константа должна быть и для сообщений, так как по каналам коммуникации нельзя передать сообщение больше некоторого объема. Вот один из аргументов: слов произвольной длины не может существовать, так как у говорящего просто не хватит запаса воздуха в легких, чтобы произнести слово, превосходящее по своим размерам критические. Помеха возникает, таким образом, в самом начале канала коммуникации - голосовом тракте человека. Нетрудно, конечно, придумать и более изощренные доводы такого рода.

Мне кажется, однако, что такие аргументы не вполне убедительны. Они основаны на предположении, что передаваемое по каналу и воспринятое сообщения должны обязательно совпадать. Но почему? Передавать ведь можно не само сообщение, а закон, по которому оно построено, если, конечно, этот закон проще, чем само сообщение. Здесь есть много возможностей, но я укажу только на одну. Речь идет о многоточии и эквивалентных ему выражениях типа "и т.д.", "и т.п.".

В 40-50-х годах в США чрезвычайно популярным было полуфилософское, полуполитическое движение, называвшее себя general semantics, основной журнал которого так и назывался

"Etc". Адепты этого учения видели в наличии в языке конструкций, допускающих неоднозначное толкование одну из основных причин конфликтов в человеческом обществе. И прежде всего под подозрение попали конструкции, позволяющие заменить полное перечисление всех членов какой-то последовательности указанием такой ее части, которая достаточна для того, чтобы обнаружить общий закон и восстановить по нему всю последовательность. "Семантики" указывали, что восстановление может быть неоднозначным. Это действительно так. Но неоднозначность - умеренная плата за свободу от ограничений, связанных с физическими свойствами каналов коммуникации.

Хотел бы подчеркнуть, что любой метод сжатия сообщения, основанный на замене его описанием законом, по которому это сообщение построено, срабатывает лишь в том случае, если сообщение устроено относительно просто. Наоборот, сообщения, потенциально имеющие культурную ценность, обычно настолько формально сложны, что могут интерпретироваться как случайные последовательности знаков. Описание же случайной последовательности (в этом суть подхода А.И. Колмогорова к понятию вероятности) не может быть существенно короче, чем сама последовательность.

5. Некоторых пояснений требует понятие коллективной памяти. Мне кажется, что сегодня оно уже не звучит исключительно как метафора. Электронные модели коллективной памяти в виде распределенных банков данных существуют. Такой распределенный банк - сеть каналов связи, в узлах которой находятся ЭВМ, в их памяти записана часть информации, хранящейся в распределенном банке. Суть "распределенности" в том, что для внешнего пользователя, обращающегося к такому банку, "лоскутная" физически память представляется единой. Он даже не догадывается, что извлеченные для него сведения - это составленный программой "коллаж" из записей, хранящихся в удаленных друг от друга пунктах.

У распределенных систем есть важная особенность, которая быть может свойственна и коллективной памяти. При их проектировании приходится решать оптимизационную задачу, добиваясь компромисса между затратами на память и на связь между узлами. При минимальных затратах на коммуникацию все сведения приходится многократно дублировать в каждом узле. Стремление уменьшить затраты на память и сократить дублирование приводит к росту затрат на связь между узлами. Оптимальное решение может оказаться таким, что дублируются толь-

ко наиболее расхожие сведения.

Роль отдельной ЭВМ, через которую осуществляется вход в распределенный банк и которой поручается хранение части данных, в структуре коллективной памяти играет память отдельного носителя культуры. Поэтому так интересно выяснить предельные возможности индивида в отношении восприятия и запоминания текста. Можно даже сказать, что психофизиологические особенности человека - это основные факторы, лимитирующие возможности коллективной памяти, т.е. человеческой культуры.

Поэтому очень интересны последние результаты, полученные психологией памяти. А если сосредоточиться на письменной речи, то и такого специального раздела психологии, как психология чтения, который очень бурно развивался в последние 10-15 лет.

6. Чтобы сказать далее нечто конкретное, мне придется резко сузить класс рассматриваемых текстов. Я буду говорить лишь об одном классе текстов - об отдельных словах и эквивалентных им слитных речениях. Хорошо известно, что слово - это целый узел проблем. Я ограничусь только тремя, которые абсолютно невозможно обойти. Первая: что такое слово. Вторая: как соотносится слово с другими текстами. Третья: как устроен словарный раздел коллективной памяти.

По поводу первой проблемы. Я считаю, что слово - самое короткое культурно и социально значимое сообщение. Опираясь на внутренние законы языка, мы классифицируем некоторые осмысленные обобщения как слова. Но не с каждым, а с лишь ничтожной частью таких сообщений связан значимый исторический опыт общества. Слова, с которыми связан исторический опыт, имеют определенный возраст, ограниченные размеры, характеризуются определенными показателями употребительности. У этих слов есть условные метки - "адреса хранения". В качестве последних используется начальная форма или другой орфографически (орфоэпически) "правильный" его вариант.

В языке существуют осмысленные сообщения и более короткие, чем слова. Это - морфемы. Но морфемы - не единицы памяти данной культуры. Их выделение полезно с точки зрения исследования внутренней симметрии сообщений, составляющих язык. Для выделения морфем необходима операция, которую Л. Ельмслев называл катализом, имея в виду расслабление, развязывание субстанциальных связей между элементами сообщения. Если подвергнуть катализу такие, например, слова как

корм, кормить, кормушка, кормилица и проч., то морфемам корм-, -ушка, -ица удастся приписать смысл. Исходя из этого смысла и формальных правил комбинирования морфем, невозможно, однако, объяснить, почему кормушка - это устройство, в не лицо как болтушка, побирушка и т.п. Почему кормушка осмысляется как источник житейских благ в низком смысле, а кормилица - в высоком (земля-кормилица)? Почему вообще кормилица лицо, а не вещь, как падалица и виселица?

Эти вопросы носят, конечно, чисто риторический характер: процесс катализа как раз и состоит в том, что мы сознательно "забываем" ту конкретную историческую информацию, которая нужна для ответа на них.

С точки зрения потенциальных возможностей сделанный языком выбор - это чистая случайность. Только капризом истории можно объяснить, что от двух синонимов prison и jail с помощью одного и того же суффикса -er образованы английские слова с противоположными значениями: prisoner "заключенный" и jailer "тюремщик".

Заканчивая вступительную часть статьи, я хотел бы процитировать известного французского философа М. Мерло-Понти: "... современность растворяется в том, что прошло, а то, что было современностью, становится историей следующих друг за другом синхронных срезов, и случайный же характер языкового прошлого проявляется даже в системе синхронии" ("О феноменологии речи").

7. Чтобы сделать следующий шаг и перейти к отношениям между словом и другими типами текстов, мне нужно решить, хотя бы в рамках данной статьи, одну терминологическую проблему.

Я буду далее называть словом любое сообщение, удовлетворяющее неуточняемым здесь критериям правильности, безотносительно к тому, хранится ли оно в коллективной памяти или нет. Напомню, что классификация слов на общеупотребительные и особенные - это одна из первых в лингвистике классификаций. Уже Аристотель в "Поэтике" различал "лексы" и "глоссы". Общеупотребительным словом ("лексой") я называю то, - писал Аристотель, - которым пользуются все, а глоссой - то, которым пользуются немногие. Ясно, что глоссой и общеупотребительным может быть одно и то же слово, но не у одних и тех же. Например, дротик у жителей Кипра общеупотребительное, а у нас оно глосса" (Поэтика, гл. XXI).

Выделение глосс, т.е. элементов, специфических для данного текста (данного автора или памяти какого-либо микроколлек-

тива), их толкование было первой "поставленной на поток" задачей филологии. Непрерывная традиция соединяет здесь современную лингвостатистику с античной филологией, эллинистическими исследованиями текстов Гомера и Масорой - существовавшей на протяжении второй половины первого тысячелетия нашей эры школой еврейских схоластов, сформировавших иудаистский текст Ветхого завета. Только позднее, в конце эллинистического периода, была осознана задача каталогизации общезначных лекс, хранящихся в коллективной памяти.

Глосса - всегда часть какого-либо более обширного текста, в пределах которого ей только и можно приписать какое-либо определенное значение. Лекса же способна употребляться самостоятельно (в виде заглавия, вывески, названия и т.п.), но едва ли такое употребление есть первичная функция слова. В норме лекса "цитируется" в тексте. И наши муки слова - это муки поиска подходящей цитаты.

Огромное количество глосс в романах Джеймса Джойса "Улисс" и особенно "Поминки по Финнегану" обеспечило хлебом, по-видимому, целое поколение специалистов по английской литературе. В последнем из названных романов глоссы составляют около половины словаря. В одном месте словоуказателя к нему я нашел 65 глосс подряд, в алфавитном порядке (все на у). Типичный вопрос, который должны были решать литературоведы и текстологи в связи с глоссами, например, такой. Глосса состоит из четырех подряд следующих when, причем второе и третье автор дает курсивом, а последнее - заглавными буквами. В других местах повторяется та же цепочка when, но с другим распределением курсива и заглавных букв. Одно это слово или два разных? Один из первых интерпретаторов творчества Джойса Мартин Джус в послесловии к известному индексу Хенли (для романа "Улисс") показал, что этот вопрос осмыслен, важен для понимания текста и по отношению к роману Джойса может быть решен вполне определенно. Мысли героев Джойса постоянно возвращаются к одним и тем же темам, событиям, ассоциациям. Первый раз ассоциация может быть раскрыта и обозначена глоссой, дальше автор использует ее как значок, показывающий, что "поток сознания" героя снова вынес его на знакомое место.

Глоссы - это часть практически любого текста. Иноязычные вставки в "Войне и мире" - глоссы. Современный читатель, не слишком сведущий в языках, читает не текст Толстого, а результат трансформации этого текста. Не так давно дело о подобной трансформации разбирал суд. Известный английский пи-

сатель Энтони Берджесс, автор нашумевшего (и экранизированного) романа "Заводной апельсин", подал в суд на своего американского издателя, который снабдил текст романа словариком. В словарике были приведены переводы на английский язык слов придуманного автором молодежного сленга. Сленгом в этом футуристическом романе пользуются терроризирующие Англию молодежные шайки. Чтобы надежнее запугать читателя, Берджесс взял за основу сленга русский язык. Он перевел на русский современный английский молодежный каргон буквально, сохраняя внутреннюю форму слов. Издатель проделал обратную операцию. Эффект от включения глосс был снижен.

Примеры отнюдь не ограничиваются художественной литературой. Наоборот, глоссы преследуют нас в научной литературе и так называемой деловой прозе. Эти жанры текстов невозможно представить без окказиональных сокращений, которые, конечно, есть не что иное как глоссы.

Но как нельзя на плоскости совместить отпечаток левой руки с отпечатком правой, не выходя в третье измерение, так невозможно разграничить лексы и глоссы, оставаясь в рамках самого текста. Только память и история могут решить, что станет лексой, а чему суждено остаться глоссой.

8. Итак, статус текста имеет лишь часть слов - лексы, которые в речи используются наравне с глоссами и при статистическом обследовании учитываются вместе с ними. Правда, на заключительных этапах статистического обследования иногда делаются попытки (обычно непоследовательные) исключить глоссы. Изгоняются имена собственные, иноязычные вкрапления, цифры и прочие знаки. При этом вместе с глоссами изгоняется и то, что наверное глоссами не является и хранится в коллективной памяти. Например, в словаре Засориной мы найдем слова африканец, африканский, афро-азиатский, но ни Африки, ни Азии. Зато будет слово ахвицер.

Естественно предположить, что лексы и глоссы используются в речи по-разному, и статистическое исследование должно это рано или поздно выявить. Ведь глосса не воспроизводится (не припоминается), а "синтезируется" для данного текста. Или же она известна одному человеку, узкой группе лиц ("семейные языки"). Воспроизводятся лишь лексы, они и характеризуются относительно устойчивой мерой употребительности.

Однако на пути проверки высказанного предположения - очень серьезные препятствия. По оценкам психологов даже индивидуальная память человека хранит до 75 тыс. слов, в кол-

лективной их должно быть много больше. Доступных средств статистического исследования не хватит даже для того, чтобы "исчерпать" индивидуальную память, для этого пришлось бы обследовать 50-100 млн. словоупотреблений, если не больше.

Но если прямое исчерпывание невозможно, остаются методы, основанные на экстраполяции тенденций, заметных даже на материале, который удается собрать в ходе исследований. Успешность такой экстраполяции зависит, однако, от того, насколько удачно выбраны исходные предположения.

Я исхожу из следующих предпосылок:

А) Ценность любых двух лекс сравнима, т. е. для каждой пары всегда определено, какой из ее членов имеет большую ценность. Отношение старшинства является транзитивным, тем самым лексы можно представить точками на числовой оси, а за начало шкалы принять, скажем, единицу. В отношении произвольного текста такое предположение казалось бы мне чрезмерно сильным: ценность романа и телефонной книги заведомо несоизмеримы.

Б) Ценности конкретной лексы в различных узлах одной коллективной памяти и даже в различные моменты времени в одном узле могут отличаться, но должны быть согласованы. Ниже я кратко опишу конкретный механизм согласования.

В) Частота употребления лексы, ее физические размеры, возраст, шансы удаления из памяти и возможно другие измеримые параметры стохастически связаны с ее положением на ценностной шкале: с уменьшением ценности уменьшается ее употребительность и возраст, а физические размеры увеличиваются. Из Б и В вытекает, что количественные характеристики лекс в одном узле памяти стохастически связаны с характеристиками тех же единиц в другом узле.

Поскольку зависимости между перечисленными показателями монотонны, любую из них можно взять в качестве показателя ценности слова. По практическим соображениям удобнее всего работать с относительной характеристикой употребительности слова - его рангом в частотном словаре. В дальнейшем ранг рассматривается как эмпирический аналог ценности слова.

О состоянии памяти в фиксированные моменты времени можно судить по конкретным текстам. Но корпус текстов (скажем, литературных произведений) отражает не содержание отдельного узла памяти. Это результат усреднения по многим узлам и по многим моментам времени. Если корпус составлен разумно, то можно надеяться, что это близкие узлы и близкие моменты времени. Предпринимаемая в данном докладе попытка оценить разли-

ры коллективной лексической памяти частично будет основана на экстраполяции тенденций, прослеживаемых в таких "усредненных" словарях, частично на данных об отношениях словарей конкретных текстов. Критерием проверки полученных результатов будет близость оценок, полученных различными путями.

9. Итак, я попытался объяснить, что такое коллективная память и что она хранит. Теперь можно перейти к вопросу о ее объеме.

"Онтологическая неопределенность" объектов, число которых предполагается найти, - это один из недостатков многочисленных работ, где проблему объема словаря пытаются решить, если можно так сказать, кавалерийской атакой. Другой недостаток тех же работ - наивная доверчивость по отношению к существующим "инвентарным описям" коллективной памяти - словарям и письменным источникам вообще. Очевидно, что любые словари не полны, интересно, насколько они не полны.

Сама задача фиксации лексических единиц редко занимала составителей словарей. Например, при составлении первого в России академического словаря, того самого, в который любил заглядывать Пушкин, был подготовлен проект его словника. Это было сделано путем компиляции имеющихся печатных и рукописных словарей (в основном двуязычных). В сохранившихся копиях, которые были предназначены для раздачи академикам, насчитывается 100 тыс. вокабул. Из них в словарь попало не более 10-15%.

Далее. В прошлом каждая эпоха имела свои пристрастия в отношении того, что следует и чего не нужно фиксировать на письме. Поэтому кривые, отражающие рост со временем числа слов, использованных в памятниках письменности, практически не имеют никакого отношения к росту объема коллективной памяти, если такой рост вообще имел место в исторический период. Они отражают в лучшем случае изменение объема той части коллективной памяти, которая отражалась на бумаге, да и к этому нужно относиться с осторожностью: сколько бумаги погибло и сколько текстов мы еще не обследовали.

Если не учитывать работы, где ошибки указанного характера извращают суть дела, то окажется, что исследований, посвященных определению объема словаря, совсем немного. Мне здесь хочется упомянуть работу А. Эллергарда, известного шведского специалиста по статистической лингвистике, в которой впервые четко была разграничена проблема измерения коллективного и индивидуального словаря. Статья "Оценка объема словаря" была

опубликована в 1960 г. Но разграничив проблемы, автор предложил оценку лишь для индивидуального словаря, который он понимал как результат "регулярной трансформации" коллективного словаря.

Индивидуальным словарям посвящена большая психологическая литература. Но в ней решаются специальные вопросы. Например, соотношение активной и пассивной части словаря, отождествление звучащего и видимого слова, способы хранения и поиска в долговременной памяти и т.д. Это очень интересные вопросы, но к обсуждаемой проблеме они имеют косвенное отношение. Обзор психологических работ по лексической памяти можно найти в книге Л. Хендерсон "Орфография и распознание слов" (Лондон, 1982 г.).

Наконец, есть целый цикл работ (преимущественно французских), авторы которых хотели бы экстраполировать лексические спектры, описывающие лексическую структуру определенного текста с целью получить данные о числе слов, которые в данном тексте не используются. Экстраполяция осуществляется очень изобретательно, но когда задумываешься над смыслом полученных чисел, то вспоминаешь старый анекдот о продавце газированной воды, который спрашивает покупателя: "А вам без какого сиропа налить? Без малинового или без вишневого?".

Весьма невразумительно определен онтологический статус объектов, число которых тщательно вычисляется: "неиспользованная лексика данной ситуации". Желающие разобраться в самой технике экстраполяции могут обратиться к монографии Б. Дофен "Словарь и лексика" (Женева, 1979).

10. По-видимому, раньше других приходит в голову идея экстраполировать зависимость между частотой употребления и объемом словаря и таким образом установить предельный объем последнего. Предельным объемом словаря, отражающим объем коллективной памяти, можно было бы считать такой, которого достаточно, чтобы достоверно покрыть любой текст на $(100-\varepsilon)\%$, где ε - доля глосс в любом тексте. Примерно такую идею выдвинул в свое время Б. Кэрролл.

Но чтобы решить поставленную задачу, нужно выполнить следующие три условия:

Во-первых, необходим словарь, отражающий языковой опыт, усредненный по всему коллективу говорящих (альтернатива состоит в предположении, что зависимость между рангом и частотой употребления устойчива по отношению к способу усреднения).

Во-вторых, нужно знать вид зависимости между рангом и употребимостью и уметь достаточно точно оценить параметры этой зависимости.

В-третьих, нужно знать значение ε – долю глосс в тексте.

О зависимости "частота-ранг" мы знаем как раз столько, чтобы с уверенностью сказать, что ни одно из перечисленных предварительных условий не выполнено. Например, хорошо известно, что никакой простой формулой нельзя описать изменение частоты на всем диапазоне рангов. Эта зависимость не только не проста, но она еще и неустойчива: меняется метод обследования – изменяется и вид связи. Наоборот, я думаю, что если бы мы предварительно знали объем словаря, мы могли бы сказать что-нибудь более определенное о зависимости частоты и ранга.

Может быть, то, что я пытаюсь использовать для экстраполяции другие зависимости, – лишь оптимизм невежды: еще не известно, насколько сложны они.

II. Первый метод основан на экстраполяции зависимости ранга и возраста слова.

По отношению к этой зависимости с известной долей уверенности можно предполагать, что выполнено первое из трех названных выше условий. Эта зависимость имеет один и тот же вид при разных способах усреднения языкового опыта. Во всяком случае, начиная с 1971 года, когда мы с М.М. Херц впервые исследовали эту зависимость, я не обнаружил какого-либо конкретного частотного словаря, для которого найденная нами зависимость между возрастом слова и его рангом не выполнялась. Я специально исследовал с этой точки зрения словарь отдельного текста – романа "Дым отечества" К. Паустовского. Не только общий вид зависимости, но и параметры совпадали с величинами, найденными для частотных словарей русского языка, составленными на выборках из разных произведений. Во всяком случае с той точностью, которую может гарантировать используемый метод оценки параметров. Во всех случаях логарифм вероятности, что слово ранга i имеет возраст не меньше заданного, убывало пропорционально корню квадратному из ранга i. Коэффициент пропорциональности зависит от заданного возраста.

Надежность проверки обеспечивается двумя способами. Во-первых, для одного словаря можно варьировать заданный возраст и строить не одну кривую по ограниченному числу точек, а семейство кривых. При этом соответствие между эмпирически-

ми и теоретическими зависимостями проверяется для всех кривых семейства. Используются разные методы отбора материала, разные методы оценки параметров и разные тесты годности подбора.

Во-вторых, проверяются следствия из теории. Если оценить параметры кривой, описывающей связь возраста τ и ранга i для конкретного τ, то результаты, полученные при анализе части словаря, содержащей наиболее употребительные слова, можно дальше распространить на весь словарь, включая и столь редкие слова, которым нельзя даже на основе проведенного обследования приписать определенный ранг. Например, можно вычислить, сколько слов праславянского происхождения может сохраниться во всем словаре современного русского языка, а затем сопоставить этот прогноз с размерами праславянского фонда, выделенного методами компаративистики.

Наиболее сложный вопрос связан с параметрами модели. Идеально было бы, если бы коэффициент, связывающий логарифм вероятности сохранения слова и корень из ранга, был строго пропорционален возрасту слова. К сожалению, это не так. Темпы обновления словаря "пульсируют". Эта "пульсация" вполне интерпретируема. В истории русского языка (но отнюдь не только русского) пики приходятся на период становления литературного языка (для русского - на I8 в.). Забегая вперед, скажу, что ускорение обновления словаря эквивалентно сокращению объема коллективной памяти.

Но если исключить "выбросы", то окажется, что темпы изменения колеблются (по сравнению с точностью их измерения) незначительно. Если средние темпы известны, то можно оценить предельное число слов, которые хранятся в коллективной памяти по крайней мере сто, пятьдесят или десять лет. Или даже время достаточное лишь для записи данного слова.

Здесь я перехожу к проблеме "порога". Учитывая распределенный характер коллективной памяти, должен существовать какой-то минимальный срок, в течение которого слово, возникшее как глосса, запоминается в существенной части узлов памяти.

В I982 г. я выписал из журнала "Знание - сила" следующую цитату: "Эта книга посвящена активщикам, а можно поручиться, что почти ни один читатель не знает такого слова. Активщик - это инженер лаборатории активных воздействий Центральной аэрологической обсерватории под Москвой". Цитата взята из рецензии на книгу, опубликованную в I980 г. Но в лаборатории слово "активщик" использовали, наверное, и до публикации кни-

ги. Уверен, что никто из читателей и по сей день не столкнулся с этим словом. А, может, никогда и не столкнется, так как лабораторию к нынешнему дню могли закрыть, а отрецензированная книга явно не стала литературным событием. Во всяком случае, как вы наглядно убедились, семь лет, прошедших с появления этого слова в печати, - не слишком большой срок, если речь идет о распространении слова в коллективах говорящих.

В 1974 г. Р.Г. Пиотровский исследовал в количественном аспекте процесс распространения новаций в языке. Для описания этого процесса он предложил S-образную кривую. Аргумент функции - время, а значение - доля носителей языка, которые используют новацию в данный момент времени. Новация медленно берет "разгон", а потом или завоевывает весь коллектив, или исчезает.

Исследования Пиотровского продолжил Г. Альтманн, который пишет даже о "законе Пиотровского". Но в целом процесс распространения новаций изучен совершенно недостаточно. И я не ликвидировал этого пробела, хотя проследил за судьбой около 200 лексических новаций, первые письменные фиксации которых относятся к интервалу с начала 18 по начало 20 веков. Оказалось, что половина их ждала включения в словарь не менее 33 лет (т.е. примерно время смены одного поколения другим). Причем этот период очень мало менялся на протяжении 200 лет. Треть века - это тот "порог", который я предлагаю выбрать в качестве "характерного времени", нужного для занесения слова в коллективную память. Если слово появляется и исчезает за период, меньший трети века, его можно не считать лексой.

При темпах обновления словаря $\bar{\lambda} = 3{,}835 \cdot 10^{-3}$ 1/век и $\bar{L} = 0{,}33$ века объем словаря $V = \frac{2}{(\bar{\lambda}\bar{L})^2}$ составляет примерно 1,25 млн. слов.

Резкое увеличение темпов обновления словаря, ломка традиции, влечет значительное сокращение объема коллективной памяти, носителем которой становится небольшая часть населения. В 18 в. темпы обновления словаря возрастают примерно в 3 раза, следовательно, объем словаря должен был сократиться в 9 раз.

12. Используем второй метод. В этом случае мы будем помещать слова в коллективную память лишь в том случае, если их графические размеры не превышают критических. Известно, что размеры графического слова, измеренные числом букв или графических аналогов слога, растут пропорционально логарифму

ранга. Эта зависимость, насколько можно судить на основе сравнения уравнений регрессии, построенных по данным частотных словарей русского языка, довольно устойчива к изменению массива обследованных текстов.

На первый взгляд, задача сводится к тому, чтобы, подставив в уравнение регрессии критические размеры слова, найти ранг, при котором будут достигнуты эти размеры. На самом деле, ситуация чуть сложнее. Вместе с длиной слова растет дисперсия этой длины. Т.е. средняя длина слова может быть далека от критической, но некоторая доля слов с рангом, близким данному, может уже превысить эту величину.

Поэтому, кроме линии регрессии, нужно знать закон, по которому возрастает разброс длин слова, и характер распределения длин слов при фиксированном ранге. Я нашел, что дисперсия длины слова также растет пропорционально рангу, причем при больших значениях ранга распределение слов по длине приближается к нормальному.

Потребуем теперь, чтобы с заданной вероятностью длина слова не превосходила критическую. Используя выражение для средней длины слова, дисперсии этой величины и зная закон распределения, можно оценить то значение ранга, начиная с которого это требование будет нарушено. Это значение ранга и выбирается в качестве объема коллективной памяти. Вероятность превышения критического размера выберем равной, например, 10^{-3}, что соответствует, грубо говоря, трем квадратичным отклонениям. Сложная проблема - выбор критических размеров слова.

Были рассмотрены два подхода к определению этой величины, которые, к моему удивлению, дали весьма близкие результаты. Сначала я исходил из того, что неприемлемы слова, которые человеческий мозг не может воспринимать как целое, т.е. слова, для запоминания которых нужен объем памяти, превышающий объем оперативной памяти. В модели, предложенной А.Н. Лебедевым, объем оперативной памяти не может превышать 30 бит. В связном тексте, как показали исследования Р.Г. Пиотровского и его сотрудников, средняя информация, приходящаяся на одну букву, составляет около I бита. Таким образом, слова длиной в 30 букв и даже несколько меньше уже не проходят через "игольное ушко" оперативной памяти.

Далее я обратился к работам по психологии чтения. Как известно, при чтении глаз совершает саккадические движения от одной точки фиксации до другой. Установлено, что одному слову в тексте соответствует один скачок - одно саккадичес-

кое движение. Установлено также, что мы фиксируем взгляд где-то около центра слова и воспринимаем слово не меняя точки фиксации. Чтобы воспринять слово, его нужно увидеть под углом несколько меньше 5°, что при обычных условиях чтения и размерах шрифта соответствует 26–28 буквам. Такие слова, как дезаксирибонуклеиновая (22 буквы) или ямщикнегонилошадейство (22 буквы) приближаются, хотя и не достигают критических размеров.

Подставляя соответствующие величины в упомянутое выше уравнение, находим, что слово имеет критические размеры при ранге порядка 2 млн. Это еще одна оценка объема коллективной памяти.

I3. И наконец, последний, третий метод. Этот метод принципиально отличается от двух предшествующих. Он основан на предположении, которое, конечно, идеализирует реальную картину.

Мы принимаем допущение, что все индивидуальные словари, хранящиеся в узлах коллективной памяти, имеют один и тот же объем и мера сходства любых двух словарей одинакова. Тогда относительно просто вычислить объем объединенного словаря. Предполагается, слово попадает в него, если содержится не менее, чем в заданной доле индивидуальных словарей. Чтобы проделать необходимые вычисления, нужна модель, описывающая, как ранг слова в одном словаре связан с рангом того же слова в другом словаре. Полностью эта модель описана в моей книге "Квантитативная лингвистика" (М., I988); здесь я остановлюсь только на ключевых моментах.

Рассмотрим два индивидуальных словаря и отношение рангов, которые в этих словарях имеет одно и то же слово. Утверждается, что логарифм этого отношения есть случайная величина, подчиняющаяся нормальному закону распределения с некоторым средним μ и квадратичным отклонением σ. Таким образом, вероятность, что на месте с рангом, скажем, 5 в одном словаре и рангом 50 – в другом одно и то же слово, не зависит от значения рангов, но исключительно от их отношения. Та же самая вероятность будет для пар рангов 50 и 5,7 и 70, 700 и 7000. Среднее значение частного ранга и дисперсия этой величины характерны для данной пары словарей. Они определяют меру сходства словарей.

Сходство словарей Пушкина и Лермонтова значительно сильней, чем, скажем, для словарей, отражающих различные жанры современного русского языка. В свою очередь, словари драмы и

прозы похожи больше, чем словари драмы и, например, газеты.

Предположим, что в индивидуальном словаре 75 тыс. слов, что соответствует оценкам психологов. Предположим далее, что среднее значение и дисперсия близки к тем, которые наблюдаются для словарей различных жанров. При таких значениях параметров среди N первых по частоте слов двух словарей совпадает примерно $N/2$ слов (независимо от величины N). Зададим далее пороговое значение доли словарей, в которых должно присутствовать слово из коллективной памяти. Предположим, что оно составляет 5%. Тогда расчет показывает, что объем коллективной памяти должен составлять около 1,7 млн. единиц.

Но представим, что общество обнаруживает гораздо большее единообразие в использовании лексики. Например, среди первых по частоте слов в словарях Пушкина и Лермонтова совпадает не 50%, а 84-85%. Тогда при том же объеме коллективной памяти - 1,7 млн. - все хранящиеся в ней слова были бы известны уже не 5%, а по меньшей мере 58% говорящих.

14. Несколько слов в заключение. Все три использованных метода привели к оценкам объема коллективной памяти одного порядка - от 1,25 до 2 млн. Различие их не столь велико, так как по опытным данным приходится оценивать логарифм искомой величины. Так что заранее можно предвидеть, что оценить удастся лишь ее порядок. Но даже если ошибка велика, то и тогда игра стоит свеч, поскольку удается установить любопытные связи между параметрами, относящимися к совершенно разным "измерениям человека": между характеристиками его индивидуальных психических возможностей - объемом кратковременной памяти, особенностями зрительного восприятия, характеристиками его долговременной памяти - и характеристиками созданной им культуры - близостью индивидуальных словарей, скоростью распространения языковых новаций. Я думаю, что осознать существование таких связей значит уже дать стимул точнее определить эти характеристики и глубже понять их смысл.

VOLUME OF THE LEXICAL MEMORY OF A SPEAKERS' COLLECTIVE

Mikhail V. Arapov

S u m m a r y

The concept of lexical memory distribution of native speakers is discussed and some suggestions concerning its structure are made.

ПОИСК И КЛАССИФИКАЦИЯ НЕСОВЕРШЕННЫХ ПОВТОРОВ В МЕЛОДИЯХ ПЕСЕН

Бахмутова И.В., Гусев В.Д., Титкова Т.Н.

Введение

Повторы являются важными структурными элементами текстов различной природы: литературных произведений, первичных структур ДНК-молекул и белков, программ, написанных на алгоритмических языках. Велика роль повторов и в формировании музыкальных произведений, особенно песен. Повторяемость отдельных фрагментов (интонаций) мелодии способствует лучшему ее усвоению, а варьирование повторяемых фрагментов не только обогащает мелодию, но и является важным средством развития музыкальной темы.

Часто одинаковые (или близкие в определенном смысле) фрагменты обнаруживаются не только в отдельно взятой мелодии (повторы первого рода), но и при сопоставлении друг с другом различных мелодий (повторы второго рода), что позволяет говорить о наличии заимствований (возможно неосознанных). В [1] был описан аппарат для выявления и классификации совершенных (неискаженных) повторов произвольной длины. С его помощью был проведен анализ повторов первого и второго рода в 216 мелодиях советских композиторов, вошедших в сборник [2].

Целью данной работы является расширение предложенного в [1] аппарата путем добавления новых конструкций, основанных на понятии несовершенного повтора или повтора с искажениями. Привлечение несовершенных повторов позволяет строить гораздо более реалистичные (по сравнению с [1]) модели музыкальных текстов. В то же время существенно возрастают вычислительные трудности, преодоление которых является принципиальным моментом для получения значимых результатов по повторам второго рода.

Рассматривается класс несовершенных повторов, образуемых участками текста, близкими в смысле хэммингова расстояния. Ориентация на данный класс обусловлена тем, что варьирование фрагментов при их повторении чаще всего происходит путем замен символов в отдельных позициях и реже – за счет вставок, устранений символов или использования более сложных операций.

Анализ несовершенных повторов первого рода дает важную информацию о способах варьирования интонаций. Эта информация может быть использована при синтезе новых мелодий, а также в задачах идентификации мелодий, когда способ варьирования может служить одной из характеристик, отражающих индивидуальность композитора, особенности жанра и т.д. Выявление несовершенных повторов второго рода существенно облегчает задачу кластеризации мелодий по степени их похожести. Количественный анализ близости мелодий представляет интерес в плане исследования эволюции отдельных музыкальных произведений, установления авторства, создания вариаций на различные темы.

Эффективность методики иллюстрируется на примере анализа русских народных песен. Более богатые возможности, обеспечиваемые использованием новых структурных единиц (несовершенных повторов), позволяют сделать следующий шаг по сравнению с [I]- перейти от поиска сходных фрагментов к поиску сходных мелодий. Соответствующие примеры, не описанные, насколько нам известно, в литературе, приведены ниже.

I. Представление мелодий (первичное описание). Пусть $z_1, z_2 \ldots z_\ell$ - музыкальный фрагмент (интонация), состоящий из последовательности ℓ нот. Элементарными количественными характеристиками интонаций являются высотная, длительностная, метрическая, из которых формируются более сложные (производные) характеристики. Следуя Р.Х.Зарипову [3], для представления мелодий будем использовать одну из таких производных характеристик - интервально-метрическую. Под интервально-метрической характеристикой интонации $z_1, z_2 \ldots z_\ell$ понимается последовательность пар $I_k S_k$, $k = 1, 2, \ldots, \ell-1$, где I_k - количество ступеней между высотами k-го и $(k+1)$-го звуков интонаций, а S_k - метрическая характеристика, дающая представление об относительной силе этих звуков. Формально

$$I_k = W_{k+1} \ominus W_k ,$$

где W_k - высота k-го звука, а "$\ominus$" - символ специальной операции вычитания;

$$S_k = \begin{cases} \operatorname{sign}(p_{k+1} - p_k) & \text{при } p_{k+1} \neq p_k \\ -1 & \text{в противном случае,} \end{cases}$$

где p_k – относительная сила доли такта, соответствующей k-му звуку.

При перекодировке мелодии из нотной записи в (IS) - представление принимаем следующее соглашение: сначала указываем значение интервала (т.е. величину $|I|$), затем его знак ("+" соответствует движению звуковысотной линии вверх, "–" – вниз), потом S_k – знак стопы ("+" – соответствует переходу от сильного звука к слабому, "–" – наоборот). Если $I = 0$, знак интервала формально не определен, но для единообразия записи ставится "+". Код (2+–), к примеру, характеризующий переход от одной ноты к другой, трактуется как скачок на 2 ступени вверх с одновременным увеличением силы звука, а код (0++) – как сохранение высоты звука при уменьшении его силы.

Используемое описание обеспечивает, на наш взгляд, желательный компромисс между двумя противоречивыми требованиями. С одной стороны, оно достаточно полно, так что мелодия не утрачивает свою индивидуальность, остается узнаваемой; с другой стороны, оно не слишком детально, так что малосущественные факторы не затушевывают сходство между двумя мелодиями.

2. Описание мелодий в терминах "повторов" (вторичное описание). Пусть $T = a_1 a_2 \ldots a_N$ – музыкальный текст длины N ((IS) – характеристика мелодии). Элемент a_i ($1 \le i \le N$), стоящий в i-й позиции текста, есть тройка $\{|I|$, знак I, знак стопы $S\}$. Фрагмент текста T, расположенный в позициях с i-й по j-ю включительно, будем обозначать $T[i:j]$. Назовем (ℓ, k) – повтором текста T (или (ℓ, k) – повтором первого рода) пару произвольных фрагментов длины ℓ из T, отличающихся друг от друга по k ($k < \ell$) позициям. Два кода в описанной выше трехэлементной кодировке считаются отличающимися, если имеется различие хотя бы по одному из трех элементов.

Совокупность всевозможных (ℓ, k) – повторов, содержащихся в тексте T, образует (ℓ, k)-характеристику текста, которую будем обозначать $\mathcal{P}_{\ell,k}(T)$.

Пример I. Пусть T= I++I+–2–+I++I+–3–+I++.

Тогда

$$\mathcal{P}_{3,1}(T) = \left\{ \begin{pmatrix} T[1:3] \\ T[4:6] \end{pmatrix}; \begin{pmatrix} T[2:4] \\ T[5:7] \end{pmatrix} \right\} = \left\{ \begin{pmatrix} I++I+–2–+ \\ I++I+–3–+ \end{pmatrix}; \begin{pmatrix} I+–2–+I++ \\ I+–3–+I++ \end{pmatrix} \right\}.$$

Здесь фрагменты, составляющие (ℓ, k)-повтор, расположены по вертикали, а совпадающие элементы помечены знаком " $\|$ ".

В худшем случае число (ℓ, k)-повторов в $\mathcal{P}_{\ell,k}(T)$ может иметь порядок N^2. Однако при работе с реальными текстами практически всегда удается подобрать параметры ℓ и k таким образом, чтобы число (ℓ, k)-повторов было невелико и сами эти повторы были функционально значимыми.

Пусть $\{T_1, T_2, \ldots, T_m\}$ - группа из m текстов ($m \geq 2$), $N_1, N_2, \ldots, N_m$ - их длины. Назовем совместным (ℓ, k)-повтором (или (ℓ, k)-повтором второго рода) пару произвольных фрагментов длины ℓ, разнесенных по разным текстам группы и отличающихся друг от друга по k позициям. Совокупность всевозможных совместных (ℓ, k)-повторов, содержащихся в группе текстов $\{T_1, T_2, \ldots T_m\}$ образует совместную (ℓ, k)-характеристику группы, которую будем обозначать

Пример 2. Пусть T_1 = I++I+-2-+I++I+-3-+I++;

T_2 = 4++I+-2-+2-+I+-;

T_3 = I+-0++2-+I+-.

Тогда

$$\mathcal{P}_{4,1}(T_1, T_2, T_3) = \left\{ \begin{pmatrix} T_1[2:5] \\ T_2[2:5] \end{pmatrix}; \begin{pmatrix} T_2[2:5] \\ T_3[1:4] \end{pmatrix} \right\} =$$

$$= \left\{ \begin{pmatrix} \text{I+-2-+I++I+-} \\ \text{I+-2-+2-+I+-} \end{pmatrix}; \begin{pmatrix} \text{I+-2-+2-+I+-} \\ \text{I+-0++2-+I+-} \end{pmatrix} \right\}.$$

Заметим, что пара фрагментов $T_1[1:4]$ и $T_1[4:7]$, также образующая (4,1)-повтор, не входит в $\mathcal{P}_{4,1}(T_1, T_2, T_3)$, поскольку оба фрагмента принадлежат одному тексту.

Прямой алгоритм вычисления $\mathcal{P}_{\ell,k}(T)$ путем попарного сопоставления всех фрагментов длины ℓ из T имеет трудоемкость $O(N^2 \cdot \ell)$. Рекуррентная схема вычисления со "скачком", изложенная в [4], позволяет избежать сопоставления части фрагментов, которые заведомо не могут образовать (ℓ, k)-повтор. Это снижает трудоемкость до $O(N^2/C(\ell, k))$, где величина делителя всегда больше единицы, а с уменьшением отношения k/ℓ может стать много больше единицы.

Алгоритм вычисления $\mathcal{P}_{\ell,k}(T_1, T_2, \ldots, T_m)$ является модификацией алгоритма из [4]. Сохраняется и приведенная выше оценка трудоемкости с той лишь разницей, что под N следует понимать сумму длин текстов $T_1, T_2, \ldots, T_m$. Очень часто обнаруженный (ℓ, k)-повтор допускает расширение: его длину можно увеличить до значения $\ell' > \ell$ при сохранении числа несовпадений k. Добавление

к алгоритму процедуры расширения позволяет в значительной степени избежать перебора по значениям k и обеспечивает компактность представления результатов. К примеру, близость мелодий "Вдоль по Питерской" и "Я вечор млада" выявляется по совместному (II,2)-повтору, расширение которого в обе стороны полностью покрывает обе мелодии.

Заметим, что выявление повторов второго рода на слух возможно лишь по очень ограниченной выборке текстов. Этот путь неперспективен, если мы заранее не знаем, что с чем следует сравнивать. Использование ЭВМ открывает принципиально новую возможность: создание машинного банка мелодий и поиск аналогов по всему банку. Но даже для ЭВМ эта задача весьма трудоемка и решение ее в полном объеме (поиск несовершенных повторов с учетом вставок и устранений символов) требует применения процедур типа "динамического программирования".

3. Описание эксперимента. Для анализа были взяты 172 мелодии русских народных песен, большая часть которых (131) опубликована в сборниках [5,6]. Длины мелодий изменялись в диапазоне от 20 до 100 символов (средняя длина - 44 символа).

Отыскивались все $(\ell, 0)$-повторы первого рода ($\ell = 1, 2, \ldots, \ell_{max}$), где ℓ_{max} - длина максимального повтора, присутствующего в мелодии), (7,1)- и (9,2) - повторы первого рода, а также (9,1) и (11,2) - повторы второго рода. Подобный выбор параметров ℓ и k в случае повторов первого рода был обусловлен тем, что при анализе отдельных мелодий по многим соображениям удобно иметь полный спектр повторов каждой мелодии (см. п.4.1). Выбор же больших значений ℓ и малых k при обнаружении повторов второго рода служит основанием для выдвижения гипотезы о неслучайном характере совпадений и обеспечивает возможность ее верификации путем прослушивания соответствующих фрагментов.

4. Основные результаты

4.1. Структура мелодий. Повтор (в том числе варьированный)-основной структурный элемент песен. Примерно 80% песен имеют повторы длины 4 и выше (иногда до нескольких десятков символов). Как правило, повторяющиеся фрагменты представляют собой те характерные интонации, которые определяют индивидуальность мелодии. Нередко повторяющиеся фрагменты покрывают всю мелодию. В этом случае ее удобно представлять в компактной форме, которую иногда называют схемой мелодии. Вычисление полного спектра пов-

торов - необходимый этап для получения схемы.

Пример 3. Схема мелодии "В темном лесе" имеет вид:

$$T = XXY(0+-)ZZZY,$$

где

$X = (0++0+-)0++I--;$
$Y = (0++0+-)0++I+-I++I+-2+-(I-+I--);$
$Z = (0++0+-)2++I--(I-+I--).$

Заметим, что фрагменты X, Y, Z содержат внутри себя более мелкие повторяющиеся интонации (выделены скобками). Их можно рассматривать как элементарные семантические единицы, из которых формируются более крупные. Формально такие единицы выявляются путем анализа изменений частот встречаемости фрагментов при их расширении. Если частота не изменяется или уменьшается, но не значительно, можно считать, что формирование семантической единицы еще не закончено. Момент скачкообразного изменения частоты сигнализирует об окончании процесса формирования.

В частности, элемент (0+-) встретился в мелодии T (пример 3) 9 раз, а (I-+) - 5 раз. Фрагмент (0++0+-), являющийся левосторонним расширением элемента (0+-), встретился 8 раз, т.е. частота изменилась всего на единицу. Соответственно фрагмент (I-+I--), являющийся правосторонним расширением элемента (I-+), встретился 5 раз, т.е. частота не изменилась вовсе. При дальнейшем расширении указанных фрагментов, частота падает скачком, т.е. фрагменты длины 2 можно считать элементарными семантическими единицами. Если продолжать процесс расширения, то вновь будут наблюдаться участки стабилизации частот, что свидетельствует о наличии более крупных семантических единиц и т.д.

Пример 4. Схема мелодии "Приданые-удалые" имеет вид: $T = X(2++7-+)X'$, где пара фрагментов X и X образует варьированный (I3,6)-повтор:

$X = 5+-0++I-+I+-I++I--I-+I--I++I++2--I++I+-,$
$X' = 5+-0++I--I++I++I--I-+I-+I+-I++2--2-+2+-,$

Примерно 20% мелодий можно отнести к категории "малоповторных": число повторов в них невелико и длины их не превышают трех символов. Как правило, это очень короткие мелодии (20-:30 символов). Типичный пример - мелодия песни "По диким степям Забайкалья" (N=33, два повтора длины 2).

Малоповторные мелодии требуют особого изучения. Следуют

учитывать, что малоповторная в I$ -кодировке мелодия может обладать ярко выраженными повторами по отдельным своим характеристикам: метрическим, ритмическим, знакам интервалов и т.д. В целом, ситуации, когда представление мелодий в терминах повторов оказывается малоинформативным, весьма редки. В связи с этим актуальной представляется задача автоматического построения и классификации схем мелодий.

4.2. Количественные характеристики варьирования интонаций. Каждая интонация есть цепочка из трехэлементных кодовых комбинаций. При варьировании интонации произвольная кодовая комбинация, в принципе, может быть заменена на произвольную. Анализ (ℓ, k)-повторов показывает, однако, что существуют как наиболее предпочтительные типы замен, так и маловероятные (практически, запрещенные).

Из трех элементов кодовой комбинации ($|I|$, знак I, знак стопы $) изменению чаще всего (примерно в 70% случаев) подвергается лишь первый элемент - величина интервала. Более чем в половине таких случаев ненулевое значение $|I|$ заменяется на нулевое (замены типа "0") или, наоборот. Примерно в 15% случаев изменению подвергается лишь второй элемент кодовой комбинации (знак I), а в 10% случаев - первый и второй элемент одновременно (замены типа "I+2"). Замены, искажающие метрические характеристики, довольно редки (в сумме порядка 5%).

Если рассматриваются (ℓ, k)-повторы со значением $k > 1$, наблюдается кластеризация замен внутри фрагментов, образующих повтор. К примеру, если $k = 2$, то примерно в половине случаев несовпадения расположены рядом друг с другом, образуя своего рода "тандем". Каждый такой тандем можно трактовать как зону неустойчивости внутри рассматриваемой интонации. Информация о зонах неустойчивости может оказаться полезной при членении мелодии на элементарные семантические единицы.

Количественные характеристики замен обнаруживают специфические различия в зависимости от обрабатываемого материала. Сопоставление русских народных песен [5,6] с песнями советских композиторов 50-х годов [2] показывает, к примеру, что замены типа "0" составляют в первом случае 48%, а во втором - лишь 28%. В то же время замены типа "I+2" довольно редки для русских народных песен (9,3%), более часты для песен советских компози-

торов (в среднем I7%) и могут быть еще более часты у отдельных композиторов (22% - 23% - Соловьев-Седой).

4.3. Заимствование музыкальных фрагментов и мелодий. В [I] была проведена классификация повторов 2 рода, выявленных в мелодиях песен советских композиторов. Неформально они были разделены на два класса: бытующие и оригинальные.

Бытующие повторы - это наиболее типичные интонации, широко используемые различными композиторами (например, гаммообразные, речитативные и др.). Они достаточно просты по структуре, занимают первые места в частотных словарях интонаций, упорядоченных по убыванию, в них, обычно, отсутствуют элементы со значениями $|I| > 1$. Существует мнение [3], что без использования таких интонаций трудно создать достаточно благозвучную мелодию. Наличие бытующих повторов мы не относим к категории заимствования.

Оригинальные повторы - это наиболее яркие, характерные интонации, определяющие индивидуальность мелодии. Они нерегулярны по своей структуре, в них обязательно присутствуют элементы со значениями $|I| > 1$, в частотных словарях интонаций они попадают на хвосты распределений. При достаточно большой длине ℓ ($\ell \geq 7$) сходство фрагментов, образующих оригинальный повтор, часто улавливается на слух. Термин "заимствование музыкального фрагмента" (возможно, неосознанное) мы относим к последнему случаю.

Термином "заимствование мелодии" мы характеризуем ситуации, когда в сравниваемых мелодиях имеется несколько достаточно длинных оригинальных повторов, причем составляющие их фрагменты расположены в обеих мелодиях в одинаковом порядке и на примерно одинаковом расстоянии друг от друга. Допускается существенное различие в длинах мелодий, но не за счет добавления новых интонаций, а за счет различия в числе повторений тех интонаций, на которых базируется сходство мелодий.

Наглядный способ представления близких мелодий основан на выравнивании их друг относительно друга по участкам-повторам, несущим одинаковую функциональную нагрузку. При этом одна мелодия записывается под другой, тождественные элементы связываются знаком " || ", близкие в функциональном отношении фрагменты помечаются пунктирными линиями, пробелы соответствуют устранению из одной мелодии части элементов. Формальной мерой близости

двух мелодий при таком представлении может служить число элементов, помеченных знаком " || ", отнесенное к средней длине двух мелодий.

Если мера близости превышает заданный порог, можно формально говорить о близости мелодий. Выбор порога осуществляется путем анализа по описанной схеме заведомо близких мелодий (например, разных вариантов одной и той же песни). Окончательное решение о близости (или различии) мелодий всегда принимается с учетом слухового восприятия.

Для иллюстрации на рис. I приведены результаты выравнивания двух известных вариантов мелодии песни "Вниз по матушке по Волге", а на рис. 2 демонстрируется обнаруженный с помощью описанной техники эффект заимствования на примере русских народных песен: "Пряха" и "Тонкая рябина".

Весьма неожиданным для авторов явился масштаб распространенности заимствований в народном песенном творчестве, причем не только на уровне отдельных интонаций, но и целиком мелодий. В довольно ограниченной по объему выборке из 172 мелодий наряду с множеством близких фрагментов обнаружилась не одна пара мелодий, при создании которых эффект заимствования сыграл, по мнению авторов, решающую роль. Это - "Вдоль по Питерской" и "Я вечор млада"; "Вот мчится тройка почтовая" и "Не слышно шуму городского"; "Окрасился месяц багрянцем" и "Эх ты, доля"; "Спускается солнце за степи" и "Замучен тяжелой неволей" и ряд других. Очевидно, что увеличение объема выборки будет приводить к пополнению этого списка и к эффекту объединения близких мелодий в классы, состоящие более чем из двух представителей.

Столь широкое распространение эффекта заимствования в народном песенном творчестве делает актуальным вопрос о его психологическом осмыслении, о роли его в других творческих процессах, об уточнении термина "заимствование", о разработке принципов моделирования творческой деятельности, учитывающих в явном виде данный эффект.

4.4. Роль неучитываемых факторов. Далеко не все повторы даже при благоприятном соотношении между параметрами ℓ и k (большое ℓ, малое k) воспринимаются на слух как похожие. При $k=0$ основной (но не единственной) причиной этого является ограниченность используемой нами кодировки, учитывающей лишь интервальные и метрические

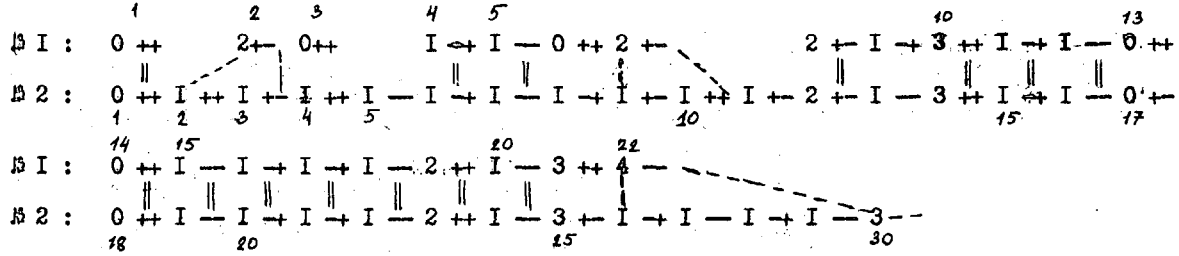


Рис.I. Выравнивание двух вариантов мелодии песни "Вниз по матушке по Волге" (№ I - вариант Прача (конец I8 века): № 2 - вариант Кашина - I8[illegible] г.)

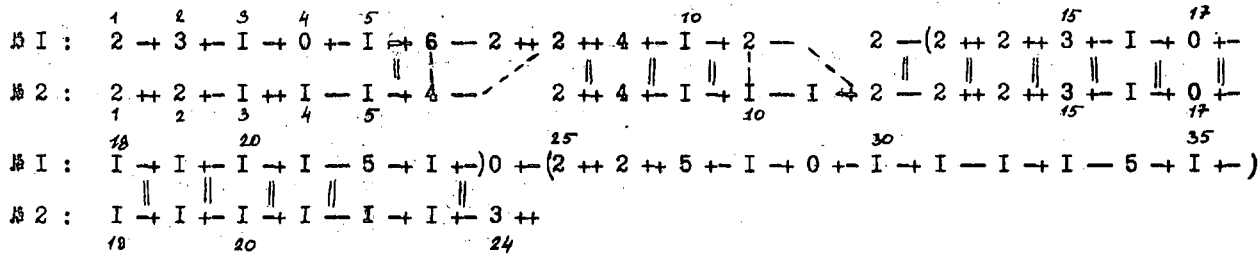
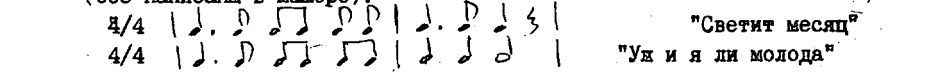


Рис.2. Выравнивание русских народных песен: "Пряха" (№ I) и "Тонкая рябина" (№ 2). Различие в длинах обусловлено повтором в мелодии № I параллельной интонации, заключенной круглыми скобками.

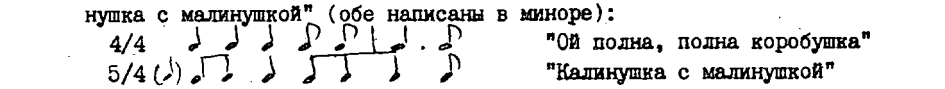
характеристики мелодии. При $k \neq 0$ возникают дополнительные искажения, обусловленные заменами. Не останавливаясь на этом подробно, отметим, что иногда даже единичная замена в интонации делает ее неузнаваемой. Ниже рассмотрим лишь основные факторы, неучитываемые кодировкой.

Общеизвестно влияние ладовой основы на восприятие музыки. Используемая нами кодировка игнорирует качественную (тоновую) характеристику интервала. Кодируя интервал между соседними звуками (пусть, например, это будет терция восходящая - (2+)), мы опускаем такие ее характеристики, как малая, большая, уменьшенная, увеличенная. Различие в них может привести к затушеванию сходства при слуховом восприятии. Примером этого является (I0,0) - повтор (2+I—I-+I—I-+I—2-+5+-I-+I—) из мелодий русских народных песен "Зачем тебя я, милый мой, узнала" (написана в ми миноре мелодическом) и "Не одна во поле дороженька" (в ре миноре натуральном). На слух явно улавливается различие, вызванное тем, что в первой мелодии элемент кодировки (2-) звучит как терция большая, а во второй - малая.

Совпадение метрических акцентов в интонациях не исключает различия их ритмических рисунков. Если мелодии написаны в одинаковом размере, небольшие различия в ритмических рисунках не препятствуют обычно выявлению сходства при слуховом восприятии. Примером может служить (8,0) - повтор из песен "Светит месяц" и "Ух и я ли молода" (обе написаны в мажоре):



Если размеры не совпадают, то различия в ритмических рисунках сказываются сильнее. Иллюстрацией может служить (6,0) - повтор, обнаруженный в мелодиях песен "Ой полна, полна коробушка" и "Калинушка с малинушкой" (обе написаны в миноре):



Интонации, образующие данный повтор, не воспринимаются на слух как сходные.

Значительное влияние на восприятие оказывает (при исполнении мелодии со словами) песенный ритм - соотношение длительностей, приходящихся на отдельные слоги текста. Совпадение песенных ритмов двух мелодий способствует обнаружению сходства интонаций при восприятии их на слух.

Заключение

Введено понятие несовершенного повтора и предложена схема

представления музыкальных текстов, закодированных специальным образом, в терминах повторов. Это представление обеспечивает широкие возможности для выявления структуры отдельных мелодий, получения количественных характеристик способов варьирования интонаций, установления сходства отдельных музыкальных фрагментов и целиком мелодий.

Методика анализа музыкальных текстов в терминах повторов реализована в Институте математики СО АН СССР в рамках модифицированной версии пакета прикладных программ "СИМВОЛ". Ее возможности иллюстрируются на примере обработки достаточно представительной выборки русских народных песен. Большой интерес, в частности, с психологической и методологической точек зрения представляет ранее отмечавшийся "эффект заимствования", играющий важную роль в понимании истоков песенного творчества и получивший новые яркие подтверждения в ходе наших экспериментов.

ЛИТЕРАТУРА

I. Бахмутова И.В., Гусев В.Д., Зарипов Р.Х., Титкова Т.Н. Выявление и анализ сходных фрагментов в музыкальных произведениях // Анализ символьных последовательностей. (Вычислительные системы, вып. II3). - Новосибирск, I985. - С. 5-43.

2. Русские песни. Песни советских композиторов. Вып. 3. - Л.: Музыка, I956.

3. Зарипов Р.Х. Построение частотных словарей музыкальных интонаций для анализа и моделирования мелодий // Проблемы кибернетики, вып. 4I. - М.: Наука, I984. - С. 207-252.

4. Гусев В.Д., Куличков В.А., Никулин А.Е. Алгоритмы поиска повторов в генетических текстах // Анализ символьных последовательностей. (Вычислительные системы, вып. II3). - Новосибирск, I985. - С. I07-I22.

5. Русские народные песни. - М.: Музыка, I985.

6. Русские народные песни. - М.: Советский композитор, I985.

SEARCHING AND CLASSIFICATION OF IMPERFECT REPETITIONS IN SONG MELODIES

I.V. Bakhmutova, V.D. Gusev, T.N. Titkova

S u m m a r y

The concept of the "imperfect" repetition is introducod and the scheme of musical texts representation in terms of repetitions, coded in a special way, is proposed. Such a representation provides wide possibilities for finding out the structures of separate melodies, receiving the quantitative characteristics of intonation varying methods, determining the similarity of separate musical fragments and melodies on the whole.

The method of musical texts analysis in terms of repetitions is suggested at the Institute of Mathematics of the Siberian Branch of Academy of Sciences, Novosibirsk as a modified applied program package "Symbol". The package possibilities are illustrated by the processing of the sufficiently large sample from the Russian folk songs. From the psychological and methodological points of view a great interest, in particular, presents the "adoption effect", as it was noted before, which plays very important role in understanding the song creation sources and receives new striking confirmations in our experiments.

К ПРОБЛЕМЕ КВАНТИТАТИВНОГО АНАЛИЗА ЯЗЫКОВОЙ ИНТЕРФЕРЕНЦИИ

М.Г.Борода

Проблематика языковой интерференции находится на скрещении интересов языкознания, психологии и методики преподавания языка. Обычно объектом исследования является влияние одного языка на другой в лексическом, фонетическом, синтаксическом плане. Заметно меньшее внимание уделяется общим квантитативным характеристикам, значения которых для данного языка достаточно стабильны, или медленно меняются, во времени - таким, например, как длина лексической единицы. Между тем, анализ влияния языковой интерференции на такие характеристики представляет значительный интерес, особенно при изучении индивидуальной речевой деятельности и существенном различии языков в плане данных характеристик. Априори совершенно неясно, существует ли такое влияние, а если да, то как оно сказывается на значениях фундаментальных квантитативных характеристик во взаимодействующих языках. Естественным материалом исследования оказываются здесь тексты, созданные, в условиях хотя бы частичного билингвизма, на разных языках и принадлежащие одному канровому кругу.

В данной работе описаны некоторые первые результаты, полученные автором в этом направлении. Исследуемыми языками были английский и русский. В качестве квантитативной характеристики рассматривалась длина слова в числе слогов. Изучалось распределение ее в массиве английских и русских поэтических текстов в жанре лимерика. В качестве материала использовались английские и русские лимерики автора, а также - лимерики из сборника (Демурова, 1978) и их русские переводы, представленные там же. (Ниже названные массивы обозначены как АА, РА, АД и РД соответственно). Массивы АА и РА, АА и АД, РА и РД существенно различались в тематическом плане, при сохранении традиционной формы лимерика: пятистрочности, рифмовки по типу "аабба", анапестической структуры, двухстопности 3-й и 4-й строк и трехстопности остальных[+].

Подмассив массива АА, объединивший тексты из разных тематических групп, приведен ниже (тексты даны в 4 "строки").

[+]За исключением I-го и 9-ти последних из приводимых ниже текстов (в I-м случае - структура "ааббвва", во 2-м - "ааббаа").

5

*An old bachelor named Sir John Gray
Got mad, being lonely, O.K.
He drank Sekt with an insect,
He made bows to his cows,
And was glad o'er his ears, as they say.

*There was an old sinner called Shell,
Who said, "I won't go to hell!"
When they asked him, "Well, why?"
He used to reply,
"It's not tortures I hate, it's the smell."

*A reckless assassin Jim Hogg
Decided to give up his job.
He said, "It's a great sin
To be an assassin!
And, besides, they don't pay enough bobs."

*An old preacher named Jonathan Hughes
Had died while expressing his views
On religion and ethics,
On art and esthetics,
Love and crime...He was a worker, poor nuis!

*There was a quick-minded young fellow,
Who, when asked why he used his umbrella
When at home, as a rule,
Cried, "You think me a fool!
To get wet my dear, costly umbrella!"

*The speaker had only one sin:
He was talkative up to the chin.
When they, nevertheless, tried
To stop him, he died,
With the words: "As could easily be seen..."

*There was a queer man from Peru,
Who believed: "All they say is true."
He was happy, till once
He learned - quite by sheer chance -
The "Do doubt!". And he died in a rue.

*Said a critical person from Hague:
"All they say in the New Test is vague."
When they asked him, "What's wrong?"
He replied, "Well, so long!"
Be so touchy! It's a sin, for God's sake!

*Said the picture of Dorian Gray,
"How dull it's to reflect one's decay!
To follow his soul
To its trivial goal...
Life is hell, art is double, I can say!"

*There was a strange person who had
An African cobra for pet.
"It is silent and tactful,
And - well, much more respectful
Than some people. And less poisonous," he said.

*A young handsome priest from Westminster
Fell in love with an ugly old spinster.
But she put him aside:
"We, spinsters, have pride!
Don't you know, I'm an Honorable Spinster!"

*A humorous man from Sri Lanka
Had married a pretty little monkey.
When they shamed him, he said,
Isn't wife just a pet?
As for women-pets... Well, I am funky."

*There once was a man of Cape Town,
Who sold his dear soul for a crown.
When they cried, "It's a sin!"
He replied with a grin,
"Cost me nothing. And a crown's a crown."

*A sprightly American miss -
Well, married a pious young Swiss.
But, after a while
He cried, "It's a trial!
She is shameless! She says, "Give me a kiss!"

*There once was a naughty young lad,
Who thought, "What if my sweet-heart is bad?
I should try her - why not? -
With cold water and hot..."
Well, she's gone. And he cried,"Life is sad."

*Cried King David, "O Lord, I'm in trouble!
Save me, Father! My life goes to rubble!"
He cried aloud and cursed. -
No reply. Universe
Whirled expanding, according to Hubble.

*There was an ambitious young pig,
Who thought, "If I wear a wig,
They would treat me so fine!
No one dares call me swine!" -
Well, they killed it, in spite of the wig.

*There was a great believer in God,
Who, but, could not keep boilin his pot.
He asked Elochim and Christ,
He asked Heiligen Geist.
But, alas! It was not even hot...
And he burnt himself under his pot.

*There was a wild person, who thought:
"Am I born for boiling my pot?!
To give my dear soul
For that trivial goal -
What a stupid idea, o God!"
And he kicked with disgust his cold pot.

*There was a strange man, who said, "What
Is the reason for boiling one's pot?
We are guests in God's house.
So, before he says "Raus!",
He must feed us, our Host. Well, why not?"
But he died hungry near his pot.

*There was a resourceful old hound,
Who used to tell fortunes for a pound.
Well, it had its pot hot,
Called them "darling" - why not?
And was never on slippery ground.

**A strange man on that humorous soil,
Idea-fixed on how keep his pot boiling,
Searched in East, North and West,
Did his utmost and best,
Got it hot... But his spirit was spoiled.
And he cried, "Oh, God, give me my rest!"

**There was a young scholar, who said,
"Boiling pots? It's a reason to get
Some development funds,
Not to mention sweet grants,
To explore, if hot water is wet."...
But he died of starvation, poor lad.

**Said an old-fashioned colonel, "Your Got
Ist ein Narr! - Want make boil your pot? -
Drang nach Osten! Go West!
North and South! Do your best!
Burn them down!! And your pot will be hot!"
But, they laughed, "What a bore, old coyote!"

*Said a yogi, "You came from the Star!
What's the reason to be taken so far
From your home and soul
By your trifle pots-goal,
To get ravaged your sweet Holy Jar?..."
But they cried, "Hush you! Go to your Star!"

*There was a strong person - why not? -
Who s u c c e e d e d in boiling his pot!
Why, his fire burnt well,
He could sniff some good smell...
But, what was this smell for? He forgot.
And he cried and blamed life, pot and God.

*Said Our Lord, "O ye handfuls of dust!
I am tired, I am sick with your trust,
With your pot-problems, cries,
Your histerical "Why?!"s!
Go to hell, if you don't like strong fast!"...
But he has been untrue and unjust.

В таблице I приведены абсолютные и относительные частоты встреч слов различной длины в каждом из исследованных массивов лимериков (АА и РА - английские и, соответственно, русские авторские тексты; АД и РД - лимерики английских поэтов и, соответственно, русские их переводы, приведенные в сборнике (Демурова, 1978)), а также в приведенном выше подмассиве массива АА (этот подмассив обозначен как $\acute{A}\acute{A}$).

Таблица I.

("дл."- число слогов в слове, "абс." и "отн." - абсолютная и относительная частота слова данной длины в массиве, N - число всех словоупотреблений в массиве).

	$\acute{A}\acute{A}$		АА		РА		АД		РД	
дл.	абс.	отн.	абс.	отн.	абс.	отн.	абс.	отн.	абс.	отн.
I	745	.8371	I637	.833I	402	.3847	I650	.8I56	97	.3356
2	I00	.II24	228	.II67	343	.3260	294	.I453	98	.339I
3	42	.0472	8I	.04I5	24I	.229I	63	.03II	66	.2284
4	3	.0034	II	.0056	60	.0570	I5	.0074	27	.0934
5	-	-	6	.003I	6	.0057	I	.0005	I	.0035
N	890		I953		I053		2023		289	

Как видно из таблицы, распределение длин слов в массиве АА близко к таковому для массива АД (т.е. массива лимериков английских авторов; о различиях по двухслоговым словам-ниже). Это свидетельствует о "настройке", в момент создания лимериков массива АА, на некоторые фундаментальные квантитативные характеристики английского языка, связанные с употреблением слов, особенно такие базовые, как длина слова (см.выше замечание о тематическом различии массивов АА и АД). О реальности такой настройки говорит и стабильность распределения длин слов в АА (ср.АА и А'А). Интересно в этом плане, что относительная частота двухслоговых слов в АА даже ниже (P=0.01)чем в АД: в АА происходит как бы отталкивание от квантитативных норм употребления слов в русском языке, где, например,в поэзии двухслоговые слова имеют существенно большую частоту, чем однослоговые (Златоустова, 1981). Важно отметить и качественное отличие распределения длин слов в массивах АР (русские лимерики) и ПД (переводы английских лимериков) от характерного для русской поэзии. Это не определяется нормами лимерика, где, в русском варианте, вполне допустимы многосложные слова (см.Демурова,1978:276, и др.). Повидимому,и здесь проявляется отмеченная "настройка" на квантитативные нормы английского языка. Характерно в связи с этим, что относительная частота однослоговых слов в АР (где, как и в АА,возможно непосредственное взаимодействие языков в условиях создания текстов фиксированного жанра) больше частоты двухслоговых.

Отмеченные факты показывают реальность межъязыковой интерференции,в плане фундаментальной квантитативной характеристики, в условиях текстотворчества. Активным в этом процессе оказывается язык, "родной" для данного типа текстов.

Л И Т Е Р А Т У Р А

Демурова Н.М. (сост.). (Topsy-Turvy World -М.: Прогресс, 1978.
Златоустова Л.В. Фонетические единицы русской речи. МГУ,1981.

TOWARDS A QUANTITATIVE STUDY OF LANGUAGE INTERFERENCE

M.G.Boroda

S u m m a r y

The paper deals with the problem of interplay between English and Russian in conditions of partial bilingualism. It is shown that this interplay reveals itself clearly in the generation of an artistic text. As a quantitative characteristic w o r d l e n g t h (in syllables) is used.

НЕКОТОРЫЕ АСПЕКТЫ ПРОБЛЕМЫ ОДНОРОДНОСТИ ЛИНГВОСТАТИСТИЧЕСКИХ ВЫБОРОК И "СЛОЖЕНИЯ" ЧАСТОТНЫХ СЛОВАРЕЙ

В.Н. Бычков

Лингвостатистику постоянно приходится решать вопрос об однородности экспериментального материала, который в различных случаях может формулироваться по-разному - включать или нет какую-либо совокупность в состав опытной выборки, объединять ли два и более частотных списка в единый частотный словарь, относятся ли экспериментальные тексты к одному подъязыку и т.д. Во всех случаях качественное сравнение выборок, частотных списков и словарей необходимо, но оно не может считаться достаточным, пока нет объективной количественной меры, а в идеале - совокупности мер, способных оценить одним числом такие многоаспектные качества, как однородность выборок, с одной стороны, и близость частотных словарей и списков, с другой. Необходимость иметь в своем распоряжении систему взаимодополняющих мер лексикостатистической однородности объясняется не только сложностью и полифункциональностью лингвистических системных объектов, но также и тем, что при проведении аналитических и сравнительных процедур лингвостатистик, как правило, располагает лишь частичной информацией о выборках и частотных словарях, составленных другими исследователями.

По своему онтологическому содержанию проблема однородности - это проблема существования самих лингвостатистических объектов - выборок, генеральных совокупностей, частотных словарей и подъязыков, последовательное формирование которых может происходить только при строгом соблюдении совокупности взаимосвязанных критериев однородности, достоверности, и эффективности. Эти критерии имеют содержательный лингвосемиотический и формально-математический смысл. В последнем случае конкретное решение вопроса однородности лингвостатистических макро- и микрообъектов зависит от того, как и в какой мере признается связь лингвостатистических частот и вероятностей. Если такая связь понимается как лингвосемиотическая реальность (Пиотровский, Турыгина, 1971; Богданов, 1972; Алексеев, 1975), то проблему однородности можно интерпретировать в терминах сводимости частот к вероятностям, частотных структур к вероятностным и наоборот. Поэтому с лингвостатистической точки зрения объединение только таких подвыборок и частотных списков оправдано, если при этом проис-

ходит улучшение или хотя бы сохранение вероятностных свойств частот и рангов, т.е. сохраняется их однородность. Случайное или намеренное нарушение этого принципа означает, что по какой-то причине допускается уменьшение однородности выборки, вероятностно-статистической достоверности частот и рангов, эффективности частотного словаря. Это в свою очередь означает, что изменяется вероятностная структура словаря и, соответственно, изменяется функционально-семиотическая ориентация этого словаря.

Сложность установления взаимосвязи между частотами и вероятностями объясняется тем, что на лингвистическом материале в разновеликих выборках не только абсолютные частоты, но и соответствующие им относительные частоты и ранги, а также их вероятности оказываются изменяющимися, скользящими величинами, связь между которыми обычно постулируется лишь исходя из принципа "интуитивной очевидности" и прецедента в классической статистике. Вероятность, если ее понимать как некоторое число, оказывается постоянно изменяющейся величиной. Такая подвижность и нефиксированность относительных частот и вероятностей, как может показаться, подрывает вероятностный фундамент статистической лингвистики. Парадоке подвижности частот и вероятностей частично снимается после установления закономерного характера их совместного смещения, но и в этом случае оказывается, что о лингвостатистических вероятностях не имеет смысла говорить как о "точечных" количественных выражениях. Такое положение в целом соотносится с принципами вероятностей. В лингвостатистике вероятность - это только достоверный интервал возможных значений частот, который в различных ситуациях имеет тенденцию к закономерному сужению или расширению и смещению относительно ранее зафиксированного уровня. Обычно в статистике этот интервал практически удается свести к точечному, дискретному значению в силу ограниченности состава и разнообразия частот статистического множества, его конечности и закрытости, что в случае необходимости достигается отсечением недостоверной, "неинтересной" хвостовой части распределения. Это создает видимость абсолютной однородности выборки и ее предела - генеральной совокупности. Но в любом случае точечная вероятность - это лишь центральное значение на интервале теоретически достоверных и эмпирически возможных количественных значений частот. Если же взять другую выборку, тем более иного размера и отличающуюся мерой однородности, то лингвистические единицы

будут "иметь право" на иной вероятностный удельный вес, но в уже несколько ином вероятностном интервале. Ниже будет рассмотрен вопрос об изменчивости интервала вероятности в зависимости от изменений основных условий.

Следует отметить, что в лингвостатистике понятие об интервальной вероятности имплицитно присутствует, когда говорится о верхних и нижних границах достоверности частот, надежности, или досоверной вероятности (Ешан, Алексеев, 1964; Фрумкина, 1964; Пиотровский, Бектаев, Пиотровская, 1977, Нелюбин, 1983). Но на практике принцип интервальности вероятности не находит последовательного и логического применения в отличие от статистической физики (см., например, Коломеец, 1976), где об интервальной вероятности говорят как о ее амплитуде и даже о волне вероятности. Далее величину достоверного вероятностного интервала будем определять как амплитуду вероятности A_P. Тогда решение проблемы однородности на вероятностно-статистическом уровне будет означать возможность сведения эмпирических частот f_i к амплитудам их вероятностей A_{P_i}. Для конкретных, прагматически заданных условий может оказаться достаточным или желательным знать и работать с точечным, среднеинтервальным значением амплитуды вероятности. Так, в дальнейшем при оценке абсолютной и относительной ошибок наблюдения, или точности совпадения частот f_i и вероятностей P_i среднеинтервальное значение амплитуды $\bar{A}_{P_i}$ вероятностей P_i ($\bar{A}_{P_i}$) будет определяться теоретически по одной из нелинейных ципфовских моделей, а сама ошибка будет вычисляться относительно срединного значения амплитуды вероятности $\bar{A}_{P_i}$, а не относительно верхней и нижней доверительных границ $P_{в,н}$, как это обычно делается в лингвостатистических исследованиях.

В нелинейных ципфовских моделях амплитуды вероятностей A_{P_i} представляют собой меру теоретически достоверного рассеяния частот около частотно-ранговой кривой, основные характеристики которой определяются эмпирическими значениями параметров, заложенных в структуру моделей. Этим обеспечивается единство вероятностного и частотного распределений в лингвостатистических выборках. Такую логико-математическую реконструкцию можно аргументировать также следующей причиной. Как известно, при определении верхней $P_в$ и нижней $P_н$ доверительных границ и относительной ошибки наблюдения ε в качестве исходных величин в расчетах выступают абсолютные эмпирические частоты F_i, а соответствующие им относитель-

ные частоты фактически принимаются за среднеинтервальное значение амплитуд вероятностей $\bar{A}p_i$ из того актуального для классической теории вероятностей условия, что в большой выборке $f_i = p_i$. В статистической лингвистике такое логическое предположение было бы оправданно, если бы выборочные относительные частоты f_i действительно принимали устойчивые или хотя бы некоторые усредненные значения. Но классической стабилизации частот f_i в лингвостатистических выборках не происходит. Поэтому, если мы априорно принимаем, что $f = p_i$, а затем без всяких поправок включаем это условие во все последующие рассуждения и расчеты, то это будет означать, что наш частотный словарь абсолютно точен. Т.е. $\varepsilon = 0$, а оценки доверительных границ вероятностей будут мерой теоретически допустимых флуктуаций эмпирических частот в других, аналогичных по объему и однородности выборках. Поэтому исходное утверждение о сходимости частот и их вероятностей ($f_i \longrightarrow p_i$) может использоваться только в качестве предварительного условия для обоснования логики допустимых рассуждений, но не для самих точных количественных измерений.

В настоящей работе верхняя $p_в$ и нижняя $p_н$ границы амплитуд вероятностей Ap_i соответствующих рангов i намеренно вычислялись по упрощенной формуле (Ешан, Алексеев, 1964)

$$p_{в,н} = \frac{F_i + 1,92 \pm 1,96 \cdot \sqrt{F_i + 0,96}}{N}, \quad (1)$$

так как ее более сложные варианты фактически приводят к сходным количественным результатам, а прозрачность математической структуры выражения (1) позволит в дальнейшем более четко и однозначно понять механизм изменения и смещения Ap_i.

В предложении об однородности соответствующих выборок рассмотрим прежде всего, как изменяется амплитуда вероятностей в зависимости от величины рангов ($i = 100$, $i = 500$, $i = 1000$) в различных частотных словарях из английских подъязыков: 1 - геологии нефти и газа, 2 - физики твердого тела, 3 - переработки нефти, 4 - электроники, 5 - космических кораблей. Выходные данные указанных частотных словарей и их некоторые основные параметры приводятся в работе (Алексеев, 1975).

Для доверительных границ амплитуд вероятностей Ap_i лингвистических единиц, как оказывается, характерно то, что они достаточно широки по абсолютной величине в верхней зоне частотных словарей, где расположены высокочастотные единицы, и значительно сужаются с увеличением ранга (см.: таблицы 1 и 2).

Таблица I

Доверительные границы амплитуд вероятностей ΔP_i относительно рангов i = I00, i = 500, i = I000 в частотных словарях 5 английских подъязыков (ПЯ)

ПЯ	i = I00		i = 500		i = I000	
	$P_н$	$P_в$	$P_н$	$P_в$	$P_н$	$P_в$
I	0,00084	0,00II2	0,00019	0,00033	0,00008	0,00018
2	0,00084	0,00II2	0,00019	0,00033	0,00008	0,00018
3	0,00091	0,00II9	0,00019	0,00034	0,00008	0,00018
4	0,00098	0,00I27	0,00022	0,00037	0,00010	0,00020
5	0,00081	0,00II0	0,00022	0,00037	0,00010	0,00020

Таблица 2

Амплитуды вероятностей $\Delta P_i = P_в - P_н$ относительно рангов i = I00, i = 500, i = I000 в относительночастотном Δf_i и абсолютночастотном ΔF_i выражениях в частотных словарях 5 английских подъязыков (ПЯ)

ПЯ	i = I00		i = 500		i = I000	
	Δf_i	ΔF_i	Δf_i	ΔF_i	Δf_i	ΔF_i
I	0,00028	56	0,00014	28	0,0001	20
2	0,00028	56	0,00014	28	0,0001	20
3	0,00028	56	0,00015	30	0,0001	20
4	0,00027	54	0,00015	30	0,0001	20
5	0,00029	56	0,00015	30	0,0001	20

Уменьшение амплитуд вероятностей особенно хорошо видно, если для большей наглядности их выразить в абсолютных частотах, чтобы избежать оперирования с микродробями в низкочастотных зонах словарей. Из табл. I следует, что в равных по объему выборках (N = 200 тыс. словоупотреблений) при переходе от i = I00 к i = I000 при одно и том же доверительном уровне β = 0,95 амплитуда вероятностей уменьшается от ~ 54 - 58 до 20. Расчеты по формуле (I) показывают, что в конце частотного списка, где i = I0000 ÷ I2000, амплитуда уменьшается до ~ 5. Таким образом, можно сделать парадоксальный вывод о том, что значения эмпирических частот наиболее высокоранговых лексических единиц по модулю могут лишь

незначительно отличаться от центрального значения амплитуды вероятности $\bar{A}\,p_i$. На основе относительной оценки доверительных границ вероятностей $p_в$ и $p_н$ обычно делается диаметрально противоположный вывод (Бектаев, 1978, с. 42-43). В перечисленных выше английских подъязыках частный случай $\varepsilon \sim 3$ означает, что у лексических единиц с максисальными рангами при доверительном уровне 0,95, параметрически заданном в (I), эмпирические частоты в выборках с N = 200 тыс. с/у могут принимать значения в 3 раза больше или меньше центрального вероятностного, не нарушая общего уровня достоверности и вероятностно-статистической однородности как различные допустимые функционально-семиотические состояния элементов языковой системы, То, что эмпирические флуктуации частот могут составлять до 3ε от центрального амплитудного, не означает, что они действительно принимают такие значения. Поэтому относительная ошибка эмпирических наблюдений даже в самой низкочастотной зоне, как это будет показано ниже, может сказаться в несколько раз меньше, чем это допускается теорией, что, естественно, повышает общую валидность лингвостатистических исследований уже на выборках объемом 200 тыс. с/у.

Из структуры математического выражения (I) следует, что амплитуда вероятности одних и тех же i будет увеличиваться с уменьшением объема выборки N и наоборот, амплитуда будет уменьшаться при увеличении выборки и ее приближении к объему генеральной совокупности, что соответствует также повышению меры относительной однородности, или близости f_i и p_i. Например, расчеты по эмпирическим данным частотного словаря подъязыка космических кораблей показывает, что при переходе от выборки I00 тыс. к выборке 200 тыс. словоупотреблений амплитуда вероятности i = I00 уменьшается с 0,0004 до 0,00029, i = 500 - с 0,0002I до 0,0,000I5, i = I000 - с 0,0,000I7 до 0,000I. Такое положение объясняется самой логикой теории вероятности. В произвольно взятых малых подвыборках возможна значительная вариативность частот одних и тех же лексических единиц. В объединенной большой выборке дисперсия частот уменьшается, так как уклонения частот не накапливаются, а наоборот, взаимно погашаются, через относительное усреднение по статистическому закону больших чисел. Еще более специфичным для них является последовательный сдвиг центрального амплитудного значения по мере увеличения выборки, так как с увеличением выборочного объема происходит искривление лингвостатистического пространства соотношения рангов и частот,

что хорошо известно в статистической лингвистике. Основная особенность нелинейных ципфовских моделей в том и состоит, что в них воспроизводится такая вероятностная аномалия, встречающаяся лишь только в некоторых социальных и биологических статистиках. Поэтому если мы сравниваем вероятности и частоты употребления одних и тех же единиц, их рангов в разновеликих выборках, необходимо производить перерасчет амплитуд вероятностей и их центральных значений. Последние фиксируются с помощью нелинейных ципфовских формулировок. Это свидетельствует не о произволе в мире статистической лингвистики, но о более сложной по сравнению с "механическими" статистиками детерминации вероятностно-статистических распределений языковых единиц.

В общем если в результате сравнения эмпирических рядов распределения в разновеликих выборках окажется, что эти ряды укладываются в теоретический интервал амплитуд вероятностей, то можно сделать вывод о том, что они являются однородными статистическими ансамблями, относящимися к одной и той же вероятностной системе. Те элементы лексических систем, которые систематически уклоняются не только от центрального амплитудного значения, но и выпадают за пределы амплитуды вероятностей, являются инородными для данной вероятностно-статистической системы. Очевидно, что однородные в заданном отношении выборки, их частотные списки или их однородные фрагменты могут быть объединены в единую систему.

При анализе материала частотного словаря английского подъязыка космических кораблей оказалось, что абсолютное большинство единиц словаря полностью и с большим "запасом прочности" уложилось в пределы амплитуд соответствующих вероятностей $A p_i$, центральные значения которых $\bar{A} p_i$ определялись по нелинейной ципфовской модели (подробнее см.: Бычков, 1986). Экстрагированные данные приводятся в табл. 3. Будем оценивать относительную ошибку (точность) наблюдения частот по степени отклонения эмпирических частот от теоретически предсказанного центрального значения амплитуд вероятностей $\bar{A} p_i$, т.е.

$$\varepsilon = \frac{|A p_i - f_i|}{\bar{A} p_i}, \quad \text{или} \quad \varepsilon = \frac{|F_t - F_e|}{F_t}, \quad (2)$$

где $\bar{A} p_i$ и F_t – теоретически вычисленные вероятность и абсолютная частота i-ой единицы частотного словаря; f_i и F_e – соответствующие им эмпирические относительная и абсолютная

частоты. Для удобства наблюдения и большей сопоставимости данных относительным частотам, их вероятностям и амплитудам имеет смысл придать абсолютночастотное выражение, округляя, где это окажется целесообразным, дроби до целочисленного значения.

Из приведенной ниже таблицы следует, что вероятностно-статистическая достоверность частотного словаря, если судить по относительной шкале наблюдения, не превышающей в пределе 0,36 при доверительном уровне 0,95, и по амплитудам вероятностей, существенно превышает оценки по ранее принятой методике расчетов.

Таблица 3

Относительная ошибка ε наблюдаемых частот F_e по данным частотного словаря космических кораблей

i	F_t	F_e	ε	$A_{P_i}(P_б \div P_н)$
I00	I72	I85	0,07	I99-I48
500	48	50	0,04	63-36
I000	24	27	0,I2	29-20
5000	4,I	3	0,26	I0-I,6
I0000	I,57	I	0,36	6-0,4
I2686	I,48	I	0,32	5-0,3

Эмпирические частоты F_e отстоят далеко от предельных амплитудных значений A_{P_i}, группируясь около центрального амплитудного. Необходимо еще раз подчеркнуть, что высокая вероятностно-статистическая достоверность частотного словаря говорит прежде всего о высокой лексико-терминологической однородности подвыборок, на основе которых составляется этот словарь, что в общем и следовало ожидать, так как предварительная программа построения выборочной совокупности была нацелена на достижение максимальной однородности эмпирического материала.

Включение в базовую лингвостатистическую выборку инородной или отличающуюся мерой относительной однородности по лексическому составу подвыборки приводит к смещению и изменению амплитуд вероятностей и их центральных значений. Это можно показать аналитически, если в выражении (I) величину объема выборки N заменить на $N = aV^2$, как это показано в (Бычков, I986). Так как

$$\frac{F_{max}}{V} = \gamma, \quad V = \frac{F_{max}}{\gamma} = \frac{kN}{\gamma}, \quad (3)$$

амплитуды вероятностей имеют пределы

$$P_{в,н} = (F_i + 1,92 \pm 1,96 \cdot \sqrt{F_i + 0,96} \cdot \frac{\gamma}{a\sqrt{kN}} \cdot \quad (4).$$

Теперь делается понятной зависимость амплитуд вероятностей от основных ципфовских параметров, в том числе от ципфовского "оптимального" объема $V_z = \frac{1}{ak}$, постоянство которого и его составляющих в выборках различного объема также является признаком их лексико-терминологической однородности и принадлежности к одной генеральной совокупности. Объединение двух неоднородных выборок сопровождается более резким увеличением выборочного словаря V, чем при объединении однородных подвыборок. Формально это проявляется в уменьшении a-параметра и V_z-объема. Из (4) следует, что уменьшение a приводит к увеличению амплитуд вероятностей по $P_в$ и $P_н$. В двух равных по объему частотных словарях одного и того же подъязыка чем однороднее их базовые выборки (т.е. чем больше a), тем уже амплитуды вероятностей и тем ближе к центральному амплитудному значению располагаются эмпирические частоты.

По данным частотного словаря подъязыка космических кораблей в первой подвыборке объемом 50 тыс. с/у **а-параметр** равнялся I,42. Если бы все последующие подвыборки были лексико-терминологически абсолютно однородными ($a_t = const$), то в общей выборке 200 тыс. с/у встретилось II876 лексических единиц, но в действительности их оказалось I2586. Ниже, в табл. 4, показано, как изменяется однородность (по a_e-параметру) подвыборок, составляющих общую выборочную совокупность N = 200 тыс. с/у английских подъязыков космических кораблей и электроники. Здесь a_t и V_t - теоретические величины при условии постоянства однородности, a_e и V_e - эмпирические данные по названным частотным словарям, в которых однородность, хотя и незначительно, но тем не менее изменялась от подвыборки к подвыборке.

Отсюда видно, что эмпирическая однородность по a_e в первых двух подвыборках оставалась практически постоянной и даже несколько увеличилась на текстовых материалах по устройству космических аппаратов и ракетных двигателей (a_e = = I,42 - I,49). Затем она снизилась после включения тексто-

вых отрывков по динамике и навигации (a_e = I,38) и еще более резко снизилась после включения текстов по жизнеобеспечению и космической медицине (a_e = I,26). Иначе говоря, резкое расширение тематического круга выборок приводит, как в общем и можно было ожидать, к резкому понижению лексико-терминологической однородности.

Таблица 4

Изменение лексической однородности в подвыборках английского подъязыка космических кораблей

N тыс. с/у	50	I00	I50	200
a_t	I,42	I,42	I,42	I,42
V_t	5933	8391	I0277	II867
V_e	5933	8I92	I0425	I2686
a_e	I,42	I,49	I,38	I,26

Для сравнения приводятся данные, полученные или расчитанные на основе частотного словаря электроники.

Таблица 5

Изменение лексической однородности в подвыборках английского подъязыка электроники

N тыс. с/у	50	I00	I50	200
A_t	I,7I	I,7I	I,7I	I,7I
V_t	5399	7647	9365	I08I4
V_e	5399	7853	9361	I0582
a_e	I,7I	I,62	I,7I	I,78

В данном случае картина носит иной характер. Все подвыборки отличаются постоянством однородности (a_e-параметр изменяется незначительно). Во второй подвыборке отмечается некоторое снижение a_e-параметра, которое компенсировалось в следующей, а за счет последней общий уровень однородности превзошел первоначально заданный.

Для амплитуд вероятностей такая ситуация имеет следующий смысл. На подвыборках подъязыка космических кораблей амплитуды вероятностей будут сужаться медленнее, на подвыборках по электронике - наоборот, несколько быстрее первоначально зафиксированного уровня.

Изменение однородности проявляется и на поведении цип-

фовского γ - параметра. Здесь постоянно приходится учитывать, что γ - параметр закономерно изменяется в зависимости от изменений объема однородной выборки. При уменьшении однородности (по a_e -параметру) γ -параметр уменьшает свое количественное значение сверх **теоретически** предсказываемой нормы, что следует также из

$$\gamma = \frac{F_{max}}{V} = \frac{kN}{V} = \frac{ka\,V^2}{V} = ka\,V. \qquad (5)$$

В табл. 6 приводятся данные об изменении γ -параметра в нелинейной модели закона Ципфа на **материале** подвыборок подъязыка космических кораблей.

Таблица 6

Изменение γ -параметра в зависимости от увеличения объема выборки (γ_t) и изменения однородности (γ_e)

N тыс. с/у	50	I00	I50	200
a_e	I,42	I,49	I,38	I,26
γ_t	0,9606	I,0006	I,0226	I,0373
γ_e	0,9606	I,0032	I,0209	I,03II

Изменение γ -параметра в зависимости от объема выборки N и ее однородности по a играет важную роль в лингвостатистических выборках, так как за ним стоит закономерное смещение значений абсолютных частот и рангов, а также соответствующее смещение амплитуд вероятностей этих частот. В целом количественная репрезентация не только частот, но и их амплитуд вероятностей, центрального амплитудного значения вероятностей оказываются в лингвостатистических условиях относительными, проективными свойствами подъязыков как вероятностно-статистических коммуникативных систем. В зависимости от того, проектируются ли базовые выборочные вероятностно-статистические пропорции на подвыборку, на еще большую выборку или на генеральную совокупность, в числовые расчеты вероятностей и частот необходимо вводить поправки на объем выборки и на ее однородность. В случае нелинейной ципфовской модели это не представляет особой трудности, так как параметры объема и однородности наряду с некоторыми другими встроены в структуру самой модели. С общетеоретических **позиций такое** положение вещей в лингвостатистике означает, лингвистические вероятности со всей определенностью относятся к разряду ус-

ловных, зависимых не только от микроконтекстов непосредственного предшествования, но и от макроконтестуальных параметров лексико-грамматической системы конкретного языка или подъязыка.

В условиях относительно высокой лингвостатистической неопределенности для точной оценки однородности выборок и близости частотных словарей и списков необходима разнообразная информация о сравниваемых лексико-терминологических макрообъектах. Их реальная связь отличается сложностью и противоречивостью. Поэтому объединению частотных словарей и списков, построению больших и сверхбольших выборок должно предшествовать обоснование тех критериев и принципов, которые кладутся в основу комплексного словаря. Для этого необходимы исследования вероятностно-статистических структур различных типов толковых и терминологических словарей, различных типов текстов в соотнесении со структурой стандартных частотных словарей. Уже теперь становится очевидным, что вероятностно-статистическая структура больших сводных частотных словарей не просто суммирует количественную информацию малых или средних по объему словарей. Речь должна идти о диверсификации структуры макрословарей. Актуальна разработка собственных лингвостатистических методов оценки выборок и частотных словарей по мере их близости и изоморфности, в том числе упрощенных методов, необходима разноаспектная теоретическая и эмпирическая интерпретация и обобщение уже имеющихся методик.

Л И Т Е Р А Т У Р А

Алексеев П.М. Статистическая лексикография. - Л.: ЛГПИ, 1975. - 120 с.

Бектаев К.Б. Статистико-информационная типология тюркского текста. - Алма-Ата: Наука, 1978. - 184 с.

Богданов В.В. Статистическая концепция языка и речи,// Статистика речи и автоматический анализ текста. - Л.: Наука, 1972. - С. 9-19.

Бычков В.Н. Теоретические и практические аспекты проблемы выборки "оптимального объема". // Квантитативная лингвистика и автоматический анализ текстов. - Тарту: ТГУ, 1986. - С. 51-60.

Елизаветина Л.И., Алексеев П.М. Рец. на кн.: Э.А. Штейнфельдт. Частотный словарь современного русского литературного языка. // Вопросы языкознания, 1964, № 6.

Нелюбин Л.Л. Перевод и прикладная лингвистика. - М.: Высшая школа, 1983. - 207 с.

Пиотровский Р.Г., Турыгина Л.А. Антиномия "язык - речь" и статистическая интерпретация нормы языка. // Статистика речи и автоматический анализ текста. - Л.: Наука, 1971. - С. 5-46.

Фрумкина Р.М. Статистические методы изучения лексики. - М.: Наука, 1964. - 110 с.

SOME CONSIDERATIONS ON THE PROBLEM OF HOMOGENEITY OF LINGUO-STATISTICAL SAMPLES AND COMPILING FREQUENCY DICTIONARIES

Valery N. Bychkov

S u m m a r y

The paper deals with the problem of estimating homogeneity of lexical samples in statistical linguistics. Some new approaches to establishing the correlation between linguistic probabilities and frequencies are proposed which may be effective in compiling frequency dictionaries. Numerical calculations and examples are presented to support the theoretical cosiderations.

О СТАТИСТИКЕ СЛОВОФОРМ В ТЕКСТАХ ПО ДЕТСКИМ ИНФЕКЦИЯМ НА АНГЛИЙСКОМ ЯЗЫКЕ

Д.М. Вольфберг

Обследовался массив текстов по педиатрии длиной в 200 тыс. словоупотреблений. Приводятся результаты статистического обследования 200 текстовых выборок длиной по 1000 словоупотреблений на материале книг и журналов Великобритании и США за период с 1970 по 1980 годы. Весь материал распределён с учётом программ советских медицинских вузов по основным разделам педиатрии следующим образом: воздушно-капельные инфекции – 42,5%, кишечные инфекции – 31%, инфекционные болезни с многообразными механизмами передачи – 20,5%, кровяные инфекционные заболевания – 5%, инфекционные заболевания, передаваемые через наружные покровы,– 1%.

В выборке было обнаружено 12128 словоформ, лексические омонимы не разграничивались. Анализ полученного материала показывает, что первая тысяча словоформ покрывает 78,1% текстов, из них 39,9% составляют служебные слова. Отмечается номинативный характер английских текстов по педиатрии. Существительные, входящие в первую тысячу словоформ, покрывают 19,9% всего корпуса, их доля увеличивается по мере снижения частот в частотном списке. Ниже приводятся первые 1319 самых употребительных словоформ /табл.1/ и распределение словоформ по частотам при $F < 20$ /табл.2/.

Условные сокращения

adj	прилагательное	m v	модальный глагол
adv	наречие	n	существительное
aux	вспомогательный глагол	p	причастие
comp	сравнительная степень	part	частица
conj	союз	prep	предлог
imp	повелительное наклонение	pron	местоимение
int	вводное	v	глагол
inf	неопределённая форма глагола		
Q	числовое обозначение		

Таблица I

Частотный список словоформ подъязыка детских инфекций

i	Словоформы	F	i	Словоформы	F
1	the	12806	53	blood	390
2	of	9814	54	illness	364
3	and	6787	55-56	antibody, who	363
4	in	6404	57	only adv	354
5	a	3351	58	group	349
6	to prep	2513	59	infections	331
7	is	2460	60	also	330
8	with	2272	61	cells	324
9	were	2123	62	when	322
10	was	1985	63	infants	315
11	for prep	1697	64	more adv	310
12	by prep	1664	65	most adj	303
13	from	1600	66	period	292
14	or	1457	67-68	acute, infected pII	282
15	be	1321	69-70	months, usually	279
16	are	1268	71	respiratory	278
17	to part	1235	72	measles	276
18	infection	1117	73-74	fever, strains	270
19	may	1014	75	between	269
20	that conj	985	76	type	268
21	virus	968	77	three	264
22	as conj	955	78	they	262
23	at	916	79	isolated pII	257
24	children	887	80	vaccine	255
25	been	873	81	some	252
26	this	815	82-84	can, influenza, those	251
27	not	807	85	their	250
28	on prep	796	86	found pII	235
29	an	791	87	serum	234
30	it	787	88	case	229
31	had	780	89	hospital	227
32	disease	755	90-91	each, rubella	226
33	have	728	92	study n	215
34	but conj	721	93	viruses	214
35	cases	694	94-96	diagnosis, first adj, human	212
36	patients	609			
37	has	589	97	time	209
38	these	567	98	present adj	207
39	which	564	99-100	diarrhoea, patient	206
40	one num	532	101	evidence	205
41	there int	496	102	hepatitis	201
42-43	days, during	494	103-104	laboratory, many adj	200
44	after prep	476	105	if	199
45	all adj	467	106-107	reported pII, severe	195
46	than	455	108	culture n	194
47	clinical	450	109	associated	191
48	two	449	110-111	incidence, table	190
49	years	431	112	used pII	189
50	age	430	113-115	however conj, into, within	186
51	no pron	410			
52	other adj	405			

116	although	185
117	symptoms	184
118	common adj	183
119	skin	182
120	year	181
121-122	cultures, weeks	180
123	onset	176
124	normal	175
125-126	often, per(-)cent	173
127	over	172
128-130	any, four, rash	171
131	lesions	170
132	he	169
133	child	167
134-135	et al, hours	166
136	coli	165
137	antibodies	164
138	given	162
139-140	number, results n	159
141-142	its, tissue	158
143-145	day, seen, test n	155
146	described pII	154
147-149	five, outbreak, types	153
150	obtained pII	152
151	high	151
152	that pron	150
153	positive	149
154	small	147
155	sera	145
156	E (Escherichia)	144
157	several	143
158-159	shown, tests	141
160-163	cough n, occurred v, studies, such adj	140
164-166	among, methods, very	138
167-171	before prep, caused pII, early adj, infections, population	136
172	family	135
173	infant	134
174	six	133
175-176	same, we	132
177-179	man, outbreaks, presence	131
180-182	per prep, rate, varicella	130
183	treatment	129
184	primary	128
185-186	adults, congenital	127
187	groups	125
188-190	antigen, specific, through	124
191-194	CMV, meningitis, second, without	122
195-198	about prep, mumps, organisms, viral	121
199-203	both conj, fluid, his, isolation, then adv	119
204-205	persons, under	118
206	cell	117
207-208	examination, occur inf	116
209	specimens	115

210-211	mild, showed	114
212-213	being, less	113
214	pertussis	112
215	occurs	111
216-220	could, previously, though, throat vaccination	110
221	attack	108
222-225	as pron, immunity, titre, will aux	107
226-227	control n, pneumonia	106
228-230	both pron, most adv, well adv	105
231-235	data, due(to), more adj, similar tested pII	104
236-237	made pII, strain	103
238-240	even, organism, probably	102
241-244	admitted pII, health, known, liver	101
245-247	history, identified pII, urine	100
248-255	important, incubation, possible, should m v, taken, therapy, where, while conj	99
256	following	98
257-259	appears, large, women	97
260-262	contact, course, system	96
263-265	agent, followed pII, week	93
266-268	examined pII, manifestations, susceptible adj	92
269-270	exposure, general	91
271-277	areas, gastro(-)enteritis, herpes, index, streptococcal, temperature, young	90
278-282	animals, antigens, birth, developed v, school	89
283-289	because(of), cause n, fig., later adv, low, syndrome, use n	88
290-292	performed pII, seven, tract	87
293-296	approximately, either conj, frequently, titres	86
297-299	jaundice, negative, responsible	85
300-302	nursery, signs, such(as)	84
303-306	area, ill adj, stage, stools	83
307-311	detected pII, different, related pII, risk, vomiting	82
312-316	against, form n, involvement, method, red	81
317-320	factors, response, she, spread n	80
321-325	changes n, complications, deaths, did aux, occasionally	79
326-336	agents, bacteria, epidemics, generally, HSV, immune, level, rates, samples, significant, vaccines	78
337-343	admission, available, epidemic n, life, recent, stool, would	77
344-348	(a)few, since prep, swabs, therefore, transmission	76
349-352	central, especially, greater, using pI	75
353-361	another, bacterial, collected pII, effect, families, least n, parainfluenza, particularly, serotypes	74

362-369	because conj, considered pII, less, ml, report n, Salmonella, so adv, whooping	73
370-375	illnesses, prior, rare, secondary, staphylococcal, toxoplasma	72
376-384	compared pII, development, forms n, frequency, immunuzation, inoculation, major, mortality, various	71
385-387	characterised pII, initial, whom	70
388-392	babies, epidemic adj, lower adj comp, occurring, serotype	69
393	studied	68
394-400	addition, aetiology, distribution, eight, levels, occur inf, protein	67
401-402	faeces, surveillance	66
403-411	commonly, first adv, including, kidney, maternal, noted pII, poliomyelitis, reports n, until prep	65
412-418	acquired pII, appearance, neonatal, pregnancy, serious, single, third	64
419-424	became, food, heat, must, pain, treated pII	63
425-432	cause inf, Coxsackie, death, except, isolates, new, picture, rapidly	62
433-442	clinically, diseases, dose, further adj, now, observed pII, particles, total, typical, water	61
443-447	findings, peak, S(Salmonella), serologic, streptococci	60
448-457	febrile, her, live adj, mouth, one pron, out adv, production, since conj, source, tissues	59
458-465	containing, cultured, direct, dysentery, increased pII, membranes, result n, rise n	58
466-471	asymptomatic, childhood, HBeAg, host, investigation, recognised	57
472-483	defined pII, end n, higher, mean adj, medical, mother, nervous, old, parotitis, subsequently, ten, world	56
484-491	countries, encephalitis, epidemiology, nine, part, readily, throughout, upper	55
492-510	affected pII, almost, aureus, before conj, centre, certain, characteristic, died v, do aux, highest, likely adj, majority, malaria, mothers, newborns, plasma, recurrent, staphylococci, subsequent	54
511-518	antigenic, for conj, growth, included v, inoculated pII, long, rarely, scarlet	53
519-528	aged, diarrhoeal, duration, generalised, importence, members, monkey, others, prepared pII, transfusion	52
529-539	fatal, HBsAg, hospitals, late, left adj, natural, phase, proportion, reaction, remaining pI, technique	51
540-553	always, body, boys, controls, does aux, enterotoxin, much adj, older, produced pII,	50

	rapid, revealed v, shigella, them, toxin	
554-565	carriers, chronic, difficult, little adj, meningococcal, rabbit, secretions, severity, so conj, sometimes, surface, whereas	49
566-575	after conj, agar, bilirubin, cytomegalovirus, delivery, determined pII, eggs, individuals, serological, still adv	48
576-596	as adv, born, chest, develop inf, epidemiological, episodes, every, fixation, identification, medium, mm, occurrence, our, penicillin, physical, received pII, recently, units, up prep, variety, winter	47
597-608	again, complement, haemagglutination, inhibition, numbers, persistent, pneumoniae, recovery, remained v, techniques, times, Q-year-old	46
609-624	about adv, absence, appear inf, condition, damage, disseminated, features, later adj, might, nasal, range n, recovered pII, S(Staphylococcus), simplex, state n, toxoplasmosis	45
625-640	activity, and/or, attacks n, based, clear adj, contrast, diphtherie, established pII, exposed pII, mucosa, none, pathogenesis, person, prevalence, species, vesicles	44
641-650	chickenpox, confirmed pII, few adj, follow-up, healthy, increase n, milk, rabbits, series, thus	43
651-666	adult, ages, causes v, contacts, estimated pII, fact, foetal, frequent, glands, having, home n, involved pII, previous, subclinical, transmitted pII, white adj	42
667-681	adenovirus, care, complete, included pII, introduction, local, materials, mononucleosis, month, pattern, received v, relatively, size, vaccinated pII, whether	41
682-704	active, antisera, appear v, association, becomes, bowel, brain, conditions, demonstrated pII, inclusion, intestinal, mainly, mg, newborn, oral, pathology, problem, products, rather, respectively, significantly, standard adj, status	40
705-714	antibiotic, carrier, contaminated, flexnery, haemolysis, information, lymph, necrosis, produce inf, role	39
715-730	admissions, characteristics, donors, effective, electron, gland, highly, include v, January, July, malaise, membrane, microscopy, past adj, practice, sonnei	38
731-749	abdominal, apparent, cerebrospinal, diagnosed, difference, factor, figure,	37

	intervals, main, mg/Qml, multiple, parents, periods, recorded pII, relative, resistance, testing ger, typhoid, zoster	
750-770	additional, antibiotics, appeared v, cannot, convalescent, grown, initially, killed, lesion, material, nasopharyngeal, nature, paralytic, required pII, routine adj, site, streptococcus, suspected pII, Q-th, tularensis, widely	36
771-792	close adj, convulsions, December, developed, foetus, injection, immunofluorescence, June, localised, M(Mycoplasma), measured, neutralising, oedema, only adj, records, rotaviruses, sample, seems, shows, stable, tetanus, widespread	35
793-812	accompanied, colonisation, despite, diametre, drug, figures, follows, headache, instances, labile, measures n, much adv, next, October, patient's, remains, S(Shigella), subjects, swelling n, together	34
813-836	absent, assay n, B(Bordetella), boy, carried (out) pII, count, CSF, cytopathic, daily, detectable, enlargement, eye, great, minutes, mucous, nodes, parotid adj, parts, plates, prolonged, relationship, resistant, spleen, ward	33
837-862	according(to), analysis, basis, become v, department, develop v, enlarged, epidemiologic, failure, fluorescent, hand, HSV(-)Q, organs, parasite, pressure, prodromal, ratio, route, salive, SH(Shigella), short, summary, unknown, weight, whole adj, Z/Hong	32
863-885	administered, attenuated, causing pI, cervical, CNS, colonies, completely, country, endemic, example, experimental, face, feature, girls, inapparent, invasion, paediatric, populations, reactions, reovirus-like, show v, specimen, variable	31
886-914	adenoviruses, apparently, begins, classified, complement-fixing, easily, ECHO, e.g., egg, extensive, faecal, formed pII, function, globulin, identical, indirect, infectivity, leukocytes, marked, mice, observations, paired, pathogenic, rotavirus, slight, summer, swab, symptomatic, workers	30
915-949	alone, anterior, applied pII, become pII, began, bordetella, cellular, clinic, closely, communities, community, cord, determine inf, diagnostic, distinct, drugs, glucose, heart, herpesvirus, infantile, investigations, last adj, male, May, non-specific, particle, poliovirus, quite, regarded pII, rhesus, selected, sequelae, spots, uncommon, value	29
950-985	ability, baby, CF, embryo, elevated, enteritis, enteroviruses, female, gastro-	28

	intestinal, gestation, immunised, interval, laboratories, minor, neck, neutralisation, nm, P(Plasmodium), perhaps, pharynx, polioviruses, public, reinfection, repeated, research, salivary, sporadic, syncytial, tularaemia, University, up adv, usual, varies, whose, younger, β-haemolytic	
986-1013	already, antigenically, bacilli, dilution, A/England/Q/Q, enteropathogenic, focal, formation, haemolytic, hepatic, IgM, infancy, lung, mucus, persist inf, place n, polymorphonuclear, produces, receiving pI, right adj, spasms, special, survey, syndromes, thereafter, too adv, true, wide	27
1014-1045	administration, April, around, associates, below prep, bone, children's, complication, concentration, contained v, date, effects, enteric, enteroviral, essential, excluded pII, fall n, HI, I, increasing, lymphocytes, notified pII, pathogens, possibility, procedure, process, reported v, seronegative, significance, social, stored, until conj	26
1046-1088	acid, circulating, clearly, colour, commonest, coughing ger, detection, differences, difficulty, directly, discharge n, echovirus, F(Francisella), far adv, final, fresh, fully, good, hands, hospitalised, hour, humans, individual, involving, limited, longer adj, LT, name, nerve, occurred pII, out(of), over(-)all, paralysis, particular, possibly, prevent inf, proved pII, predominantly, ranged v, referred, review, September, viraemia	25
1089-1129	added pII, ADV-Q, defects, derived pII, develops, discussed, documented, efficacy, exchange, extent, HQNQ, household, incubated, induced, inflammation, intestine, just, muscle, national, newborn, nor, observation, off, orchitis, persisted v, physicians, preparation, presenting pI, prevalent, primarily, produce v, seasonal, shigellosis, sore, systems, ulcers, via, virulence, volume, wild, ZIP	24
1130-1158	bronchitis, cause v, combined, continued v, erythrocytes, excretion, flow n, following adj, fourth, genital, knowledge, media, nearly, nose, once adv, origin, paroxysmal, placenta, presumably, procedures, pulmonary, removed, SGOT; spinal, stages, staphylococcus, term n, woman, well adj	23
1159-1203	abnormalities, antitoxin, attempt, average, bacillus, become inf, bodies, bronchiolitis, consists, DNA, Escherichia, experience n, gamma, haemagglutinin, imported, increased v, indeed, intravenously, introduced pII,	22

investigated pII, lymphadenopathy, males, mononuclear, necessary, nurse, pandemic, passage, people, piglets, recipient, resevoir, rises, rural, separate, sepsis, septicaemia, simple, sources, suggests, thought pII, trunk, unit, worldwide, yet, /um

1204-1255 able, /accounted v, all pron, amount n, August, bacteriology, capable, chick, child's, cm, contains, Coxsackievirus, designated pII, developing pI, earlier adv, early adv, enterovirus, epithelial, eruption, expected, falciparum, found v, immediately, immunofluorescent, immunoglobulin, injected, isolations, Kong/Q, labour, latent, means n, member, minimal, muscles, notifications, November, obvious, pregnant, prevention, producing pI, ranging, see imp, serologically, sex, sheep, slightly, spreads, swine, upon, urticaria, vesicle, volunteers, way, Weil's, Q-year 21

1256-1319 above adv, above prep, agglutination, agona, air, appropriate, attempts, bilateral, calf, causative, centres, change n, Chinese, colonised, considerable, corticosteroids; distributed, Dr, doses, entry, epithelium, equally, evaluation, extremely, FA, fairly, form inf, four(-)fold, g, gave, give inf, haemorrhagic, HAI, here, herpetic, inflammatory, indicated pII, ingestion, intranuclear, leishmanua, leishmaniasis, M(matrix), maximum, meningococci, neutraminidase, never, normally, parasites, phage, protection, salmonelaae, screening, seldom, sensitivity, shortly, spring n, states, tender, tenderness, transfer, transient, useful, values, what 20

Таблица 2

Распределение частот после 20

i	m	F
1320-1386	67	19
1387-1448	62	18
1449-1523	75	17
1524-1594	71	16
1595-1672	78	15
1673-1779	107	14
1780-1897	118	13
1898-2018	121	12
2019-2143	125	11
2144-2293	150	10
2294-2485	192	9
2486-2745	260	8
2746-3017	272	7
3018-3363	346	6
3364-3835	472	5
3836-4469	634	4
4470-5439	970	3
5440-7235	1796	2
7236-12128	4893	1

ON THE STATISTICS OF WORD FORMS IN ENGLISH TEXTS ON CHILDREN'S INFECTIONS

Daniel M. Volfberg

S u m m a r y

The article presents a selective study of medical texts in English speaking countries. A list of word forms was compiled on the basis of a sample of 200 thousand word usages. The study proves a definite nominative character of English texts in paediatrics. The percentage of nouns increases along with rank numbers in the frequency list. The article provides statistical data dealing with distribution of word forms in the frequency list.

О НЕКОТОРЫХ КВАНТИТАТИВНО-ЛИНГВИСТИЧЕСКИХ ОСОБЕННОСТЯХ ОСНОВНОГО СЛОВАРНОГО ФОНДА АРМЯНСКОГО ЯЗЫКА

С.Григорян, Н.Манасян

Предметом работы являются разные по своему происхождению пласты армянской лексики+. "Важное место в лексикологических исследованиях занимает вопрос о генетическом составе лексики данного языка вместе с соответствующими квантитативными характеристиками. Подсчет различных слоев лексики: исконного наследия, заимствований из других языков и собственных образований позволяет сделать обоснованные выводы о происхождении и развитии языка, о культурных связях с другими народами в разные исторические периоды, об удельном весе различных генетических групп в современном языке. Особый интерес представляют древнейшие элементы лексики, которые передавались в течение столетий от одного поколения к другому и дожили до наших дней. Эти элементы принадлежат, как правило, к т.н. словарному фонду, на котором базируется вся лексико-семантическая система языка. В рамках квантитативно-системного исследования лексики возникает вопрос о закономерностях распределения древних слов в словаре современного языка, особенно в той его части, которая охватывает наиболее частотные лексические единицы и в определенной степени совпадает с основным словарным фондом языка. Совокупность наиболее частотных слов в представительном словаре именуется "базовым словарем" /Тулдава, с.136/.

В настоящей работе в качестве базового словаря были взяты первые самые частые 1400 слов/с частотой не менее 226/частотного словаря современного армянского языка / /Казарян/.

+Статья в основном построена согласно работе/Тулдава/,сделанной для эстонского языка на материале авторской речи в художественной прозе. В отличие от работы Ю.А.Тулдава, в данной статье рассматриваются различные пласты армянского словарного состава только на уровне словаря.

Краткие исторические сведения. Как известно, армянский язык составляет отдельную ветвь в индо-европейской семье языков. Армянский язык отпочковался от индоевропейской языковой общности в Ш тысячелетии до н.э. и начал свое существование как отдельный язык+. До этого периода его можно определить как один из индоевропейских диалектов. Народ, говоривший на этом языке, населял Армянское нагорье. Первые письменные памятники датируются У-ым веком н.э. Периодизация армянского языка имеет следующий вид.

Дописьменный период

I. Протоармянский язык /Ш тыс. до н.э. - ХШ в. до н.э./
2. Древнейший армянский /ХП в. до н.э. - IУ в. н.э./

Письменный период

I. Древнеармянский, известный под названием "грабар" /У - XI вв./
2. Среднеармянский /ХП - ХУI вв./
3. Новоармянский /ХШ в. - до наших дней/.

Ядро протоармянского и древнейшего армянского составляли индоевропейские слова, иначе говоря, исконно армянские слова. Впоследствии произошли заимствования из ряда языков, и уже в древнеармянском в большом количестве присутствуют заимствования, причем иранские заимствования составляют большинство.

Имеются многочисленные заимствования также из фракийского, арамейского, ассорийского, кавказских языков, греческого, древнееврейского, более поздние заимствования - из арабского, латинского и других языков. Несмотря на многочисленные заимствования ядро словарного состава армянского языка продолжают составлять слова индоевропейского происхождения.

Идентификация древних слов. В настоящей работе к древним словам относятся исконные слова древнейшего происхождения и древние заимствования. Вслед за Ю.А.Тулдава /Тулдава, с.I4I и дальше/ здесь постулируется принадлежность к древней лексике

+Краткая периодизация истории армянского языка приводится по работе Г.Б.Джаукяна "История армянского языка"/Ջաուկյան/.

следующих слов[+]:

I/ непроизводные слова с установленной древней основой и со значением или с одним из значений, предпочтительно равным или близким к старому значению или к одному из старых значений, например, կա "есть, имеется", գեղեցիկ "красивый", պատկեր "образ, изображение", գետ "река";

2/ производные или сложные слова, состоящие из древних основ и аффиксов, например, գոյություն "существование", կառուցվածք "структура", տեսակետ "точка зрения", բացատրել "объяснять".

Всего в базовом словаре встретилось I274 древних слова.

Следует отметить, что при идентификации древних слов учитывались регулярные фонетические изменения, происходившие в языке в ходе его развития. Сюда относятся, например, многие бывшие производные или сложные слова, которые в силу фонетических сокращений выступают в современном языке как простые слова, например, տեր տի այր .

В качестве примера приводится фрагмент базового словаря, в котором древние слова отмечены звездочкой /см. табл.I/.

Таблица I

I003	*գետ	"река"	977	սովետական	"советский"
I003	*ըստ	"по"	974	*թվական	"числительное"
997	*երիտասարդ	"молодой"	97I	*կանչել	"звать"
993	*կապ	"связь"	97I	*հատկություն	"свойство"
983	*պայման	"условие	967	*գիտություն	"наука"
983	*վիճակ	"состояние"	964	*աշակերտ	"ученик"

Анализ по частям речи. Итак, в базовом словаре, т.е. среди I400 наиболее частотных слов было установлено I274 слов древнего происхождения /0,9I%/. Наиболее обширную группу составля-

---

[+]Идентификация проводилась по Новоайказскому словарю /Նոր հայկազեան բառգրան / и по "Этимологическому коренному словарю армянского языка" Г. Ачаряна /Աճառյան /.

Следуя хронологизации Г.Е.Джаукяна /см. предыдущую страницу/, границей для новых и древних слов принимается XII в.

ют существительные - 554 слов древнего происхождения /43,5%/, далее следуют глаголы - 337 /26,5%/ и прилагательные - 186 /14,6%/, затем - наречия - 73 /5,7%/, местоимения 45 /3,5%/ и остальные части речи, которые вместе взятые составляют 74 слова /5,8%//см. табл. 2/.

Таблица 2

Часть речи	Число	%
Существительное	554	43,5
Глагол	337	26,5
Прилагательное	186	14,6
Наречие	73	5,7
Местоимение	45	3,5
Релятивные слова	29	2,3
Модальные слова	24	1,9
Союзы	20	1,6
Числительные	4	0,3
Междометия	2	0,2
Всего	1274	100,0

Если рассматривать распределение древних слов по частотным зонам базового словаря, разделяя словарь на 14 зон по 100 слов в каждой /по убывающей частоте/, то можно констатировать большие различия между отдельными частями речи /см. табл. 3/. Особенно наглядно проявляется динамика изменения удельного веса частей речи при рассмотрении накопленных частот. Сделав необходимые перерасчеты на основе данных табл.3, проиллюстрируем это на графике, откуда видно, что удельный вес вложенных групп существительных среди древних слов постоянно увеличивается по мере приближения к зонам меньшей частотности. Из графика видно также, что доля глаголов среди древних слов сначала чуть-чуть увеличивается, а потом остается более или менее стабильной. Из графика можно также заключить следующее -

Табл. 3. Распределение древних слов по частям речи в базовом словаре

Частотн. зона (ί)	сущ.	глаг.	прил.	нар.	мест.	рел. сл.	мод. сл.	союзы	числ.	межд.
I(I-I00)	27	22	5	8	I4	9	3	I0	I	
2 (I0I-200)	37	29	9	8	II	5	I	I	I	
3 (20I-300)	42	25	I2	5	9	3	I			I
4 (30I-400)	33	24	I9	2	I	4	3	3	I	I
5 (400-500)	40	I8	I8	7	I	2	I	2		
6 (50I-600)	4I	23	I3	6	2	I	2			
7 (60I-700)	43	22	I2	II	I		2		I	
8 (70I-800)	50	I9	I5	I	2	I	2			
9 (80I-900)	39	29	I4	3	I					
I0(90I-I000)	37	3I	I2	8		I	2			
II(I00I-II00)	48	I7	I7	I		I	5			
I2(II0I-I200)	42	26	I2	2	I		I			
I3(I20I-I300)	37	26	I4	2		2	I			
I4(I30I-I400)	38	26	I7	6						
Всего	554	337	I86	73	45	29	24			

удельный вес прилагательных к середине списка сначала увеличивается, а затем стабилизуется. Доля наречий сначала увеличивается, а затем также стабилизуется, в то время как доля всех других частей речи монотонно уменьшается /см. рис.I/.

Лексические группы древней лексики. Стопроцентно древними оказываются существительные, обозначающие части /или органы/ тела. Их общее число в базовом словаре - I8. К ним относятся աչք "глаз" /4I83/, ձեռք "рука" /39I3/, լեզու /3I60/ и др. Стопроцентно древними являются также слова, обозначающие животных-I0 слов, названия цветов - 7, свойства - I7, человеческие чувства - 9, слова, обозначающие меру - 9, слова, относящиеся к природе - 40. В группе названий людей насчитывается в базовом словаре 53 слова, из них 53 древних, например, մարդ "человек,мужчина" /7652/, տղա "мальчик" /2750/, մայր "мать" /2686/. Обширную группу составляют глаголы, обозначающие типичные действия человека. Из 229 глаголов в базовом словаре 220 являются древними : խոսել "говорить" /I2277/, գալ "приходить" /8360/, տեսնել "видеть"/678I/ и др. Большинство местоимений относятся к древней лексике.Числительные также представляют древнюю лексику.

Возраст и частотность слов. "... если говорить о связи между возрастом и частотностью слов, то приходится, по сути дела констатировать наличие очень сложной сети взаимосвязей и взаимообусловленностей не только непосредственно между упомянутыми свойствами, но и между ними и особенностями общественного развития, а также особенностями языковой структуры, в том числе связь с такими моментами, как многозначность и длина слова. Следовательно, связь между возрастом и частотностью слов предстает как проявление сложного взаимодействия компонентов системы и общностью их поведения, причем обнаруживаемые связи в системе подчиняются законам развития и функционирования языка" /Тулдава, с.I5I/. Для выявления связи между частотой слова и его возрастом , вслед за Ю.А.Тулдава, мы воспользовались методикой М.В.Арапова И М.М.Херц /Арапов, Херц/, которые рассматривают связь возраста и частотности слова как основу для построения математической модели эволюции словаря. По

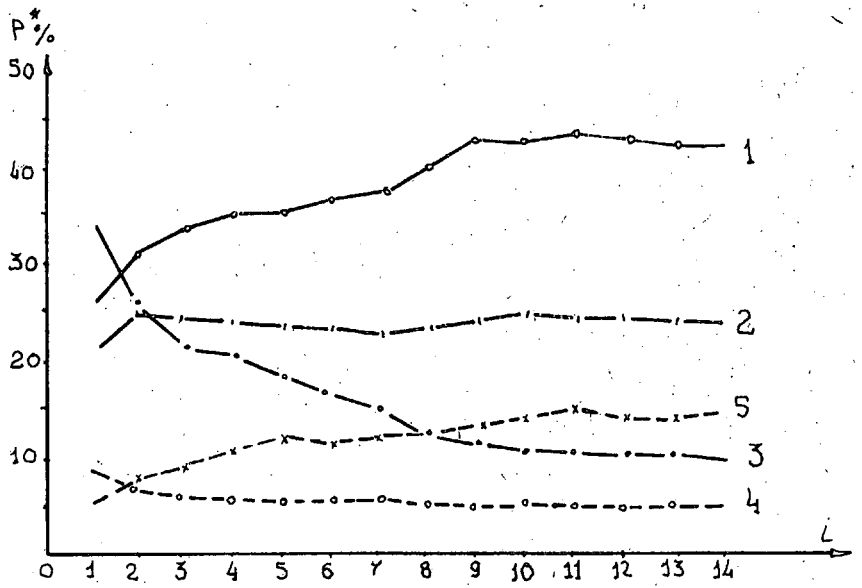


Рис.I. Распределение древних слов в базовом словаре: связь между вложенными группами слов частотных зон (i) и накопленных частот (p^* %) различных частей (I - существительные; 2 - глаголы; 3 - прилагательные; 4 - наречия; 5 - остальные вместе взятые).

этой методике строится распределение между указанными двумя признаками. Слова объединяются в частотные зоны по 100 слов в каждой зоне. Ранжированным зонам приписываются номера или ранги /i/. В каждой зоне выявляется абсолютная частота древних слов $\mathcal{F}(i)$ и соответствующие относительные частоты

$$\rho(i) = \frac{\mathcal{F}(i)}{n} \quad (n = 100).$$

По данным рассматриваемого базового словаря армянского языка является картина, представленная на табл.4 и рис.2. В отличие от распределений по другим языкам /см. Тулдава, с,160/ распределение возраст-частота по армянскому базовому словарю имеет вид неопределенных флуктуаций. Чтобы усилить достоверность результатов, была определена доля древних слов в первой самой частой тысяче частотного списка слов в художественной прозе того же частотного словаря. Частотные данные были извлечены из третьей части словаря, алфавитно-частотного списка слов с частотами не менее 50 /см. Манасян, с. 149/. Всего было рассмотрено 1057 слов. Однако и в этом случае кривая распределения имела вид флуктуаций /см. рис. 2/. Этот факт дает основание предположить, что такой характер зависимости является универсальным для армянского словаря.

Чтобы составить представление о характере кривой по всей протяженности частотного списка, была посчитана доля древних слов в последней зоне общего списка частотного словаря с частотой не менее 3 /частота 2 и 1 в этом списке не представлены/, также были просчитаны дополнительно по три сотни слов на протяжении всего словаря. Эти группы слов по сто слов в каждой отстояли друг от друга на равное количество слов.

Полученные данные /см. рис. 2/ позволяют сделать вывод о том, что убывание количеств древних слов имеет место и в частотном словаре армянского языка. Но этот процесс происходит гораздо медленнее, чем в словарях, составленным по другим исследованным языкам, получая монотонное убывание только где-то в последней четверти всего частотного списка. Типичной кривой такого типа является сглаженное распределение по эстонскому

Табл. 4. Распределение слов древнего происхождения по частотным зонам армянского словаря

i				
I	I00	I00	I,00	I,000
2	I00	200	I,00	I,000
3	99	299	0,99	0,997
4	96	395	0,96	0,988
5	87	482	0,87	0,964
6	88	570	0,88	0,950
7	92	662	0,92	0,946
8	90	752	0,90	0,940
9	89	84I	0,89	0,934
I0	9I	932	0,9I	0,932
II	89	I02I	0,89	0,928
I2	84	II05	0,84	0,92I
I3	82	II87	0,82	0,9I3
I4	87	I274	0,87	0,9I0

Табл.5. Распределение слов древнего происхождения по частотным зонам армянского словаря по всему списку

I	I00	I00	0,I0	0,I0
2	73	I73	0,73	0,865
3	72	245	0,72	0,8I7
4	68	3I3	0,68	0,783
5	33	346	0,33	0,692

$\sum = 346$ $n = 500$

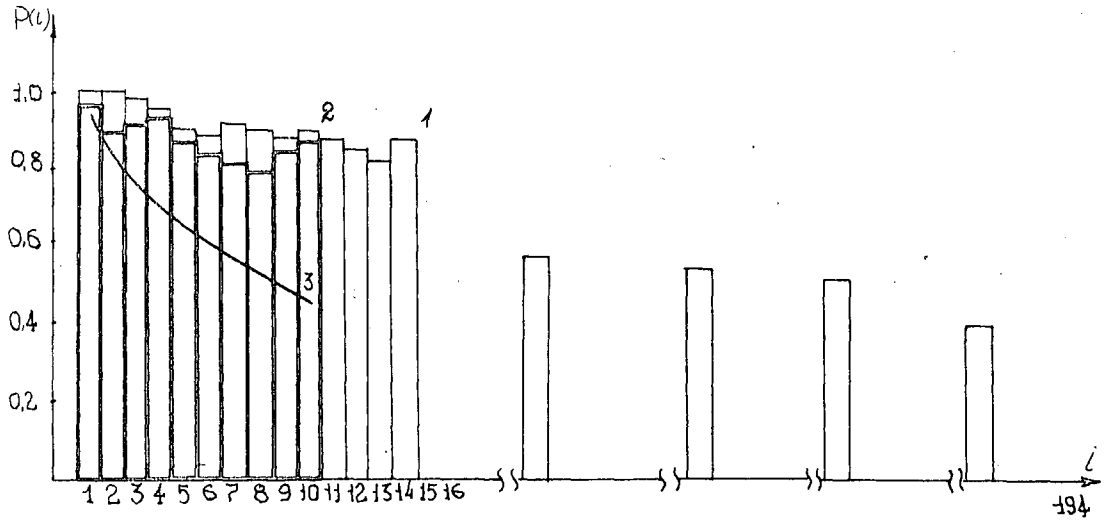


Рис.2 Распределение древней лексики в частотных словарях: I - армянского языка (частотный список слов с частотой не менее 3), 2 - армянского языка (художественная проза), 3 - эстонского языка (авторская речь).

словарю /рис. 2/.

Таким образом, мы убедились в том, что армянский язык характеризуется наличием большого количества слов древнего происхождения в своем словаре по сравнению с другими языками. Возможно признак насыщенности древними словами можно рассматривать как типологизирующий. Для этого надлежит провести аналогичный анализ частотных словарей других современных языков, например, греческого, хинди /предшественниками были рассмотрены эстонский, русский, чешский, английский, немецкий и французский частотные словари/.

ЛИТЕРАТУРА

Арапов М.В., Херц М.М. Математические методы в исторической лингвистике.-М.: Наука, 1984.

Казарян Б.К. Частотный словарь современного армянского языка. -Ереван: Изд-во АН АрмССР, 1982.

Манасян Н.С. Рец. на: Казарян Б.К. Частотный словарь современного армянского языка.//ВЯ, 1985, № 4.

Тулдава Ю. Квантитативное исследование генетического состава лексики эстонского языка. // Лингвостатистика и вычислительная лингвистика. Труды по лингвостатистике. Тарту: Изд-во ТГУ, 1982. -С.136-166.

Ա՛ճառյան Հր. Հայոց լեզվի պատմություն, մաս 1-2, Երևան, Պետ. համալսարանի հրատ, 1940-1951:

Ա՛ճառյան Հր. Հայերեն արմատական բառարան, հ. Ա-Փ, Երևան, Պետ. համալսարանի հրատ, 1971-1979:

Նոր բառգիրք հայկազեան լեզուի, Երևան, Պետ. համալսարանի հրատ, 1979:

Ջահուկյան Գ.Բ. Հայոց լեզվի պատմություն, Երևան, ՀՍՀ ԳԱ հրատ, 1987:

ON SOME QUANTITATIVE AND LINGUISTICAL PECULIARITIES OF THE ARMENIAN BASIC VOCABULARY

Susanna Grigoryan,
Narine Manasyan

S u m m a r y

In this article the various "strata" of different origin in the Armenian vocabulary are examined as constituent parts of the lexicon. The analysis is carried out on the basis of the core of a frequency dictionary (1,400 most frequent words in the modern language) which is considered as the quantitative model of the text reflecting the use of the most important and characteristic words in the living language. A short historical survey of the development of the Armenian language is given and the principles of the identification of the ancient words in modern language are stated. Then follows the quantitative analysis of ancient words according to their division into parts of speech and lexical groups. The problem of the relation between the age of the words and their frequency is discussed. It is stated that the age-frequency relation in the Armenian language is not similar to that in a number of languages investigated by other authors.

СТРУКТУРИЗАЦИЯ ХУДОЖЕСТВЕННОЙ ПРОЗЫ С ИСПОЛЬЗОВАНИЕМ ЭВМ
1. ФОРМАЛЬНО-ПУНКТУАЦИОННЫЙ МЕТОД СТРУКТУРИЗАЦИИ

О.Н. Гринбаум

В статье описывается формально-пунктуационный метод структуризации художественной прозы, базирующийся на алгоритмическом подходе к основным правилам русской пунктуации. Метод разработан, реализован на ЭВМ и используется для автоматической разметки художественных текстов (выделение глав, страниц, абзацев, предложений, словоупотреблений) и маркировки абзацев по типам письменной речи. Проверка представленного здесь метода структуризации велась на материале полного текста романа М.Ю.Лермонтова 'Герой нашего времени' и показала его высокую эффективность.

В первой части статьи обсуждается проблема структуризации художественной прозы применительно к формированию текстовой части Машинного Фонда Русского Языка, описан метод структуризации и его программно-алгоритмическое ядро.

Во второй части описаны алгоритм маркировки абзацев по типам письменной речи, алгоритм детализации видов прямой речи, а также система диалоговой корректировки структурированного текста. Представлены основные результаты структуризации полного текста романа 'Герой нашего времени'.

Общие положения. Концепция Машинного Фонда Русского Языка (МФРЯ) предусматривает (МФРЯ,1986) разработку, создание и развитие двух основных частей Фонда - словарно-грамматической и текстовой; их гармоничное сочетание или 'сосуществование' (Шайкевич А.Я.,1986) должно обеспечить функционирование Фонда как многоцелевой научно-исследовательской и информационной системы. Создание текстовой части Фонда, т.е. представительного корпуса русских текстов различных типов : литературно-художественного, научно-технического, публицистического, фольклорного, организационно-распорядительного и др.(Караулов Ю.Н.,1986; Герд А.С.,1986а) за последнее время из проблемы дискуссионного плана (Гиндин С.И.,1986) все более становится осознанной необходимостью и превращается в одну из аксиом компьютеризации филологических исследований. Об актуальности создания текстовой части МФРЯ свидетельствуют и многочисленные зарубежные работы в

области компьютерной лингвистики (подробнее см.:Тулдава Ю.А. 1986; Розенцвейг В.Ю.,1986).

Утвердившийся в компьютерной лингвистике постулат о примате текста при своей практической реализации встречает ряд сложных вопросов[+], среди которых далеко не последним является вопрос о структуризации текстов. Потребность в проведении филологических, в особенности лингво-статистических исследований диктует необходимость включать в состав текстовой части Фонда такие произведения художественной прозы, для которых выполнены - как минимум - следующие два условия :

1. Любой текст, хранимый в памяти ЭВМ,должен быть структурно-фрагментирован[++] с выделением томов, глав, страниц, абзацев, предложений и других структурных элементов.

2. Абзацы художественно-прозаических текстов должны быть разделены минимум на две категории : абзацы авторского повествования и абзацы с элементами прямой речи.

Первое условие существенно, например, при создании машинного фонда лексических единиц (Рогожникова Р.П.,1985) или в разного рода филолого-статистических исследованиях (Тулдава Ю.А.,1977). Действительно, применяемое ныне правило определения контекста как определенного числа символов вправо и влево от ключевого слова (Рогожникова Р.П.,1985; Азарова И. В. и др.,1983) не может считаться безупречным и используется только потому, что текст, хранимый в памяти ЭВМ для получения карточек-цитат, не структурирован. При наличии же структурно-фрагментированного текста понятие 'контекст' даже в формальных (компьютерных) системах имело бы неформальное (смысловое) значение, а сам контекст мог бы по желанию исследователя включать различное число предложений, абзацев или еще более крупных образований разной или одинаковой длины вправо и влево от ключевого слова.

Второе условие является следствием того, что авторское повествование в художественно-прозаическом тексте распадается на авторскую речь и речь персонажей. Без учета этого обстоятельства можно прийти к ошибочным выводам в любых фи-

---

[+] См.например : Андрющенко В.М.,1986; Котов Р.Г., 1986; Пиотровский Р.Г.,1986; Герд А.С., 1986б.

[++] Мы исходили из того факта, что полные авторские тексты вводятся и хранятся в памяти ЭВМ в форме, идентичной печатному оригиналу. Это тем более справедливо для систем, сопряженных с фотонаборными автоматами (Андрющенко В.М.,1985).

лологических исследованиях, особенно в тех, которые основаны на использовании статистических методов (Лесскис Г.А.,1962; Мартыненко Г.Я.,1983).

Если быть более точным, то следует отметить, что условие, указанное нами вторым (разграничение собственно авторских и не авторских фрагментов повествования) с необходимостью требуют предварительного решения первой задачи (структуризации текста). И это понятно, ибо создавая программу автоматической структуризации, можно было бы задать перечень формальных признаков, указывающих на начало прямой речи, но решить, где же она заканчивается - без указания,как минимум, границ абзацев - попросту невозможно. Также невозможно, в общем случае, отделить заключенную в кавычки прямую речь от предложений с цитатами, речениями, различными названиями и т.п., если отсутствуют сами границы предложений. Структурирование же вручную произведений художественной прозы - даже в режиме диалога с компьютерной системой - занятие архисложное, особенно если учесть масштабы этой работы. И, напротив, при наличии предварительно структурированного текста работа исследователя по его корректировке, проводимая в режиме диалога с ЭВМ, будет иметь достаточно высокую практическую эффективность.

Таким образом, решение проблемы структуризации прозаических художественных текстов, ввиду своей сложности и чрезвычайной трудоемкости, не может рассматриваться иначе,чем в ракурсе автоматической обработки.

Формально-пунктуационный метод структуризации прозаических текстов. Содержательно задача структуризации прозаических текстов формулируется (Гринбаум О.Н.,Мартыненко Г.Я.,Фитиалов С.Я.,1985) следующим образом : текст, введенный в память ЭВМ в виде, идентичном печатному оригиналу, должен быть структурирован формальным способом с указанием :

а) границ страниц, абзацев, предложений, словоупотреблений;

б) принадлежности структурных элементов текста (абзацев и предложений) к одному из типов письменной речи (авторская или прямая речь).

Минимальное требование к введенному в ЭВМ тексту - выделенность прописных букв, поскольку большинство современных ЭВМ еще не позволяют одновременно пользоваться строчными и прописными буквами.

Задача структуризации включает в качестве самостоятель-

ных следующие подзадачи :

а) разработка прграммно-алгоритмического комплекса для структурной фрагментации текста;

б) разработка алгоритмов и программы формального разграничения и классификации абзацев по типам письменной речи - авторской и прямой;

в) разработка алгоритма и программы детализации абзацев прямой речи с выделением абзацев собственно прямой речи, абзацев 'смешанной' речи (содержащих и авторскую, и прямую речь) и абзацев 'вложенной прямой' речи (прямая речь внутри абзаца собственно прямой речи);

г) построение эффективной диалоговой системы, позволяющей легко проверять правильность и - при необходимости -корректировать структурированный художественный текст.

Для решения первых трех подзадач был разработан и реализован формально-пунктуационный метод, основанный на :

а) детальном анализе правил русской пунктуации;

б) трансформации этих правил в формально-логическую схему распознающего автомата -алгоритмическое ядро всего метода структуризации[+].

Разработка формально-пунктуационного метода структуризации художественной прозы натолкнулась на ряд существенных противоречий между нечеткостью (размытостью) пунктуационной системы, с одной стороны, и строгостью формализованной записи алгоритма -с другой. Нечеткость использования знаков препинания, отсутствие обязательной регламентации и почти стихийный процесс исторического развития пунктуации (Шапиро А. Б.,1974) - все это, видимо, тормозило развитие работ по формальной структуризации текстов. Тем более, что художественные тексты, увидевшие свет в разное время (например, даже с прошлого века и по сегодняшний день) несут в себе не только характерные для своей эпохи пунктуационные особенности, авторские предпочтения и вольность при употреблении тех или иных знаков, но и редакторскую правку последующих изданий.

Такое наслоение многих 'мешающих' факторов резко усложняет процесс формализации пунктуационных правил. Но, поскольку главной задачей при построении формально-логической схемы структуризации была задача вычленения предложений, из

[+]В нашем исследовании отсутствует сопоставительный анализ при выборе наилучшего метода решения, поскольку альтернативные методы (кроме ручных) нам неизвестны.

рассмотрения были исключены многие знаки, не относящиеся к разграничительным знакам препинания, в том числе запятые, двоеточия, точки с запятой. Однако некоторые из них были впоследствии использованы нами при разработке процедуры маркировки абзацев и детализации видов прямой речи; это – тире, двоеточие, кавычки, скобки.

Выделение предложений построено на анализе правого и левого окружения знаков препинания, потенциально выступающих в качестве разграничителей предложений : точки, вопросительного и восклицательного знаков, многоточия. Необходимость анализа окружения таких знаков диктуется тем обстоятельством, что в художественных текстах разграничительные знаки '.', '?' и '!' помимо своего традиционного места в конце предложения часто встречаются и в середине предложений, тогда как для других типов текстов это практически исключено.

Итак, построение формально-логической схемы структуризации есть – в самом общем виде – задание последовательного перечня ситуаций (состояний распознающего автомата), каждая из которых предопределяет один из возможных результатов анализа (элемент множества возможных решений). Общая схема принятия решения относительно выделения из текста очередного предложения выглядит следующим образом :

а) распознается символ, принадлежащий множеству потенциальных разграничителей предложений;

б) производится анализ правого или левого окружения выделенного символа (поиск на дереве альтернатив пунктуационных ситуаций);

в) решение о выделении очередного предложения принимается лишь в том случае, если поиск на дереве альтернатив завершился в терминальной вершине, соответствующей одной из решающих ситуаций.

Для примера рассмотрим две простейшие ситуации (значения двух терминальных вершин) :

< знак + пробел + прописная буква > (1)

и < знак + пробел + буква(не прописная) > , (2)

где 'знак' есть либо '.', либо '?', либо '!', а '+' используется для обозначения операции сцепления (конкатенации). Ситуация (1) является решающей, а ситуация (2) таковой не является. Иллюстрацией к (1) может служить следующая строка текста :

ции, нам отвели ночлег в дымной сакле. Я пригласил (3)

(Лермонтов М.Ю. Герой нашего времени)

и к ситуации (2) – другая строка из того же произведения :

чем у Бэлы; а какая сила! скачи хоть на пятьдесят (4)

После распознавания в строке (3) разграничительного знака '.', к очередному i-ому выделяемому предложению будет отнесена вся левая часть строки (3) вплоть до пробела перед буквой 'Я'; этой буквой начнется следующее $(i+1)$-ое предложение, в которое будут включены два слова строки (3), расположенные правее знака '.', и все последующие слова вплоть до выявленного центра $(i+1)$-ой решающей ситуации. В строке (4) решающей ситуации нет, следовательно эта строка целиком войдет в очередное выделяемое предложение.

Маркировка абзацев и предложений по типам письменной речи, а также детализация видов прямой речи производятся аналогичным образом, хотя и требуют других решающих правил для других пунктуационных ситуаций.

Анализ и формализация правил русской пунктуации. Формально-пунктуационный метод структуризации художественной прозы базируется на алгоритмическом подходе к основным правилам русской пунктуации. К этим правилам мы, в первую очередь, относим такие правила, которые позволяют проводить разграничение (выделение) предложений в тексте, исходя только из их графического (печатного) изображения.

Известные правила пунктуационного оформления конца предложения были подвергнуты нами тщательному изучению на предмет составления перечня всех возможных и встречающихся в реальных художественных текстах ситуаций – таких, в которых символы окончания предложения ('.', '?', '!') действительно означают конец предложения, а также ситуаций, в которых это места не имеет. В результате анализа выявлено три варианта ситуаций, безусловно фиксирующих конец предложения :

А. *ZAG* – ситуация : знак перед заголовком или подзаголовком.

Б. *AC* – ситуация : знак перед началом нового абзаца (перед появлением в тексте абзацного отступа).

В. *FIN* – ситуация : знак внутри абзаца в окружении других знаков или символов, образующих вместе с ним решающую ситуацию.

В а р и а н т А очевиден : заголовок (подзаголовок) не может разрывать предложение на части, следовательно, умение формально выявлять в тексте заголовки (подзаголовки) позволит безошибочно фиксировать конец предшествующих им предложений.

В а р и а н т Б тоже достаточно прозрачен, но требует некоторой детализации. Дело в том, что предложения, состоящие из прямой речи и предшествующих ей вводящих слов автора, нередко размещаются на разных строчках печатного текста; в таких случаях вводящие слова заканчиваются двоеточием,а прямая речь оформляется абзацным отступом (АО) , за которым следует тире. Например :

Она вдруг почувствовала себя оскорбленной и сказала холодно :

- Нам нужно расстаться на некоторое время, а то от скуки мы можем ...

(Чехов А.П. Попрыгунья)

Наличие абзацного отступа, таким образом, сигнализирует о конце предложения лишь в том случае, если новая строка (включающая этот АО) не начинается с тире. Если же после абзацного отступа следует тире, а предыдущая строка заканчивается двоеточием, то АО-ситуация

< двоеточие + АО + тире + ПБ > ,

где ПБ - прописная буква, решающей не является. Если двоеточия в конце строки, предшествующей АО, нет, то ситуация :

< АО + тире + ПБ >

является решающей.

В а р и а н т В из всех трех является наиболее сложным, поскольку выявление FIN-ситуации производится внутри абзаца и, следовательно, может базироваться только лишь на анализе взаимного расположения знаков и символов вокруг потенциального центра решающей ситуации. Эти центры,как мы уже отмечали, образуются знаками '.', '?' и '!'. Обозначим множество таких знаков препинания через Z, тогда $Z = (z_1, z_2, z_3)$ и z_1 = '.', z_2 = '?', z_3 = '!' .

Любая FIN-ситуация рассматривается нами как последовательность трех более простых ситуаций F, I и N ($FIN = F+I+N$) :

1. F-ситуация ('конец предложения') фиксируется при наличии одного из элементов множества $F = Z+T$, где T = (пусто, кавычки, закрывающая скобка). Иначе, $F = (Z, Z'', Z)$), что соответствует любому набору символов, существенному для формального выявления в тексте ситуации 'конец предложения'.

2. I-ситуация ('интервал') определяется теми символами, которые используются для заполнения интервала между двумя предложениями, расположенными внутри одного абзаца: $I = (i_1, i_2)$ и i_1 = 'пробел', i_2 = 'тире'.

3. N-ситуация ('начало предложения') фиксируется при наличии одного из элементов множества $N = \rho + E$, где ρ = (пусто, кавычки, открывающая скобка), а E – множество всех прописных букв. Иначе, $N = (E, "E, (E)$, что соответствует набору символов, используемому для графического изображения начала любого предложения, находящегося в середине абзаца.

Приведем примеры, поясняющие введенные формализмы.

Пример 1.

> Вдоль оврага - по одной стороне опрятные амбарчики, клетушки с плотно закрытыми дверьми; по другой стороне пять - шесть сосновых изб с тесовыми крышами. - Над каждой крышей высокий шест скворечницы...
>
> (Тургенев И.С. Стихотворения в прозе. 'Деревья')

В этом отрывке тире встречается три раза : в первых двух случаях – в середине предложения, а в последнем случае – между двумя предложениями. FIN-ситуация для последнего случая выглядит следующим образом : f ='точка', i ='тире', n ='Н'; тогда в целом fin = < точка + тире + 'Н' >.

Пример 2.

> Придет, бывало, к нам в деревню; я то ему племянником довожусь. - "Что, брат, Вася, - скажет, - приди-ка, брат, переночуй-ка у меня!"
>
> (Тургенев И.С. Три встречи)

В этом случае -ситуация формируется так : f ='точка', i ='тире', ρ ='кавычки', e ='Ч'. Следовательно : fin = < точка + тире + кавычки + 'Ч' >.

Пример 3.

> "А своими наемными работниками ты доволен?" - "Да, - процедил сквозь зубы Николай Петрович, - ..."
>
> (Тургенев И.С. Отцы и дети)

Здесь FIN-ситуация образована следующей последовательностью символов : z ='?', t ='кавычки', i ='тире', ρ ='кавычки', e ='Д'. Следовательно : fin = < ? + кавычки + тире + кавычки + 'Д' > .

Пример 4.

> Да ко всему-то впридачу, кроме позора-то, ненавистную жену ввести в дом! (Потому что ведь ты меня ненавидишь, я это знаю!)
>
> (Достоевский Ф.М. Идиот)

Здесь : fin = < ! + пробел + (+ 'П' >.

Итак, любая FIN-ситуация представима последовательностью трех более простых ситуаций F ('конец предложения') , I ('интервал') и N ('начало предложения'). Этот факт отражает

таблица формализованного представления правил выявления ситуации 'конец предложения (табл.1), где по вертикали расположены F, I и N - ситуации, а по горизонтали - их возможные значения.

Таблица 1

Формализованное представление правил выявления ситуации 'конец предложения в середине абзаца

F - ситуация (конец предложения)		Значения F - ситуации								
		Z			Z''			$Z)$		
I-ситуация (интервал) и ее значения	пробел	+	+	+	+	+	+	+	+	-
	тире	+	+	-	+	+	-	+	-	-
N - ситуация (начало предложения)		Е	"Е	(Е	Е	"Е	(Е	Е	"Е	(Е
		Значения N - ситуации								

В табл.1 зафиксировано всего восемнадцать возможных значений FIN-ситуации. Знаком '+' отмечены такие последовательности символов, которые, реально встречаясь в художественных текстах, позволяют с гарантией фиксировать в середине абзаца ситуацию 'конец предложения'. Знаком '-' отмечены те ситуации, которые в реальных текстах практически не встречаются; они, тем не менее, тоже являются решающими FIN-ситуациями.

Потенциальные возможности графических средств оформления конца - начала предложения оказываются, как мы видим, избыточными : только 70% этих возможностей используется в графике русского языка. В более общем плане следует сказать, что пунктуационная избыточность графических средств русского языка, подтвержденная в настоящем исследовании при рассмотрении формальных способов структуризации художественной прозы, представляет собой еще один пример избыточности выразительных средств языка в целом.

Подведем некоторые итоги. Рассматривая варианты фиксации в тексте ситуации 'конец предложения' и имея в своем распоряжении только изобразительные средства письменной речи, т.е. буквы алфавита и знаки препинания, мы выделили три воз-

можных случая : ситуация 'конец предложения' фиксируется перед заголовком (*ZAG*-ситуация), перед началом абзаца (AO-ситуация) и внутри абзаца (*FIN*-ситуация). Каждый из этих случаев сравнительно легко формализуем : надо позаботиться лишь о том, чтобы иметь доступ одновременно к двум смежным строчкам анализируемого текста; для *ZAG* - и AO-ситуаций - просто потому, что заголовки (и начала абзацев) всегда расположены ниже текущей (анализируемой) строки текста; для *FIN* -ситуаций - потому, что внутри абзаца текущее и новое предложения могут располагаться на разных строчках, а знание пунктуационно-символьной обстановки на границе двух предложений обязательно.

Важной представляется не только сама возможность формализации правил выявления конца предложений, но и надежность метода, т.е. полнота охвата (фиксации) реальных текстовых ситуаций.

Проверка возможностей формально-пунктуационного метода структуризации прозаических текстов проводилась на материале романа М.Ю.Лермонтова 'Герой нашего времени'. Структуризация всего произведения (150 страниц текста), выполненная в автоматическом режиме на ЭВМ ЕС-1035 , дала ничтожный процент ошибок : из 3090 выделенных предложений лишь в двух случаях зафиксирована излишняя детализация текста. В обоих случаях причиной явилось наличие знака препинания 'точка' перед именем собственным :

Его звали... Григорьем Александровичем Печориным. (4)

... на кресте написано крупными буквами, что он поставлен по приказанию г. Ермолова, а именно в 1824 году. (5)

Оба случая соответствуют одной и той же *FIN* -ситуации : <знак + пробел + прописная буква>. Эта ситуация для большинства предложений является решающей, что и отмечено в табл.1; в предложениях же (4) и (5) после разграничительного знака 'точка' следует пробел и имя собственное, что формально (но здесь - ошибочно!) воспринимается как ситуация 'конец предложения'. Тем не менее, второй случай в принципе исправим : для этого в алгоритм структуризации следует ввести блок анализа стандартных сокращений, что несколько увеличит время работы программы, которая реализует этот алгоритм, но позволит избежать излишней ручной корректировки структурированного текста. Предложение (4) для своего формального выде-

ления из текста требует не только дополнительного использования программы синтаксического анализа, но и специального решателя, указывающего, какие выделенные из текста и следующие друг за другом предложения составляют единое целое, а какие - нет. Затраты на проведение такого анализа (а его надо будет проводить для каждых двух смежных предложений текста!), а также нерешенность проблемы автоматического синтаксического анализа предложений сложной структуры лишний раз убеждают в правильности выбранного нами подхода к этому вопросу : исправление ошибок, допущенных при автоматической структуризации художественных текстов, должно производиться в режиме диалога 'человек - ЭВМ' с помощью диалоговой системы корректировки структурированног текста.

Алгоритмическое ядро метода структуризации. Формально-пунктуационный метод, основу которого составляет алгоритм вычленения предложений, базируется на автоматическом поиске одной из трех решающих ситуаций : ситуации 'конец предложения' перед заголовком (*ZAG*-ситуации), ситуации 'конец предложения' перед новым абзацем (AD-ситуации) и ситуации 'конец предложения' в середине абзаца (*FIN*-ситуации).

Попытка отнести текущую ситуацию в строке к одной из этих трех решающих ситуаций производится последовательно : сначала делается попытка отнести ее к *ZAG*-ситуации (т.е. определить, не является ли очередная строка заголовком), затем - к AD-ситуации; если же обе эти попытки окажутся неудачными, то уточняется принадлежность текущей ситуации к одной из *FIN*-ситуаций.

П о и с к *ZAG*- ситуации осуществляется просмотром двух строчек текста - текущей, в которой зафиксирован знак 'Σ' (либо '.', либо '?', либо'!'), и последующей (иногда - двух последующих) строки, смежной с текущей строкой. Выявление заголовков опирается на ряд особенностей их графического начертания и расположения печатного материала на страницах текста. Сюда, в первую очередь, следует отнести :

а) сдвиг начала заголовка относительно левого поля текста на величину, большую, чем абзацный отступ;

б) оформление начала заголовка прописной буквой;

в) отсутствие (как правило) в заголовке разграничительных знаков препинания;

г) оформление следующей за заголовком строки в виде строки с абзацным отступом;

д) уменьшенное (как правило) значение длины текущей

строки по сравнению со средней длиной строки печатного текста (для данного издания) в том случае, если эта строка предшествует заголовку.

Совокупность таких особых примет, сопровождающих появление в тексте заголовков, вполне достаточна для их формального распознавания в художественных текстах. Это, конечно, не означает, что разработанный нами алгоритм безошибочен при идентификации всех вариантов заголовков и подзаголовков в любых прозаических текстах; существуют, однако, два способа преодоления возможных ошибок :

а) корректировка структурированного текста в режиме диалога 'человек - ЭВМ';

б) пополнение списка условий для формального выявления заголовков, такими,например, как наличие в строке только цифровых символов или наличие в строке только слова 'Глава' с последующим набором цифровых символов.

Введение новых признаков для идентификации заголовков может быть сделано достаточно просто ввиду используемого нами модульного принципа построения всего программно-алгоритмического комплекса.

П о и с к АО - ситуации (абзац) не представляет особой трудности, поскольку наличие в строке текста абзацного отступа является и необходимым, и достаточным для этого условием.

П о и с к *FIN* - ситуации ('конец предложения' в середине абзаца) осуществляется согласно решающим правилам, представленным в табл.1 : считывается очередная строка текста и, если в ней обнаруживается одна из F-ситуаций ('конец предложения'), то уточняется значение I-ситуации ('интервал' между словами), а затем производится поиск одного из возможных значений N-ситуации ('начало предложения'). Если для найденной последовательности символов, образующих $(f+i)$-ситуацию, не подтвердится ни одно из возможных значений N - ситуации, то *FIN*-ситуация не фиксируется; в противном случае производится формирование очередного выделяемого предложения и затем - несколько упрощая реальную работу алгоритма - оно записывается во внешнюю память ЭВМ. Оставшаяся часть текущей строки - справа от i-символа в *FIN*-ситуации - переписывается в поле памяти, отведенное для хранения очередного предложения, и вновь подвергается анализу на наличие *FIN*-ситуации.

Таким образом, алгоритм формально-пунктуационного метода включает две циклически выполняемые операции : поиск ре-

шающей ситуации для каждой строки текста (внешний цикл) и – при наличии в данной строке хотя бы одной *FIN*-ситуации – поиск других возможных в этой строке *FIN*-ситуаций (внутренний цикл).

Программа, реализующая этот алгоритм, составлена и отлажена на языке ПЛ/1 в среде ОС ЕС. Программа содержит пять основных блоков : блок выделения страниц, блок выделения абзацев, блок поиска конца предложения, блок исключения неинформационных строк, блок записи предложений во внешнюю память ЭВМ. Общий об'ем программы *AGRIS* составляет около 500 операторов языка ПЛ/1; среднее время на вычленение одного предложения – 2,4 секунды; структурирование полного об'ема романа М.Ю.Лермонтова 'Герой нашего времени' (150 страниц текста) заняло – с учетом работы программ маркировки абзацев и детализации видов прямой речи – 2,1 часа работы ЭВМ ЕС-1035.

Заключение. Вернемся к проблеме организации работ при формировании текстовой части МФРЯ. На наш взгляд, структуризация художественных прозаических текстов, включаемых в состав Фонда, не только желательна, но и – с разработкой машинного метода фрагментации – попросту необходима. Текст, перенесенный на машиночитаемый носитель информации в виде, идентичном печатному оригиналу, должен быть подвергнут структуризации (либо методом, представленным в этой работе, либо каким-то другим – но не ручным! – методом), после чего этот текст следует записать во внешнюю память ЭВМ. Далее обязательным является этап диалоговой корректировки структурированного текста : филолог-исследователь работает с ЭВМ в режиме диалога и , читая на экран дисплея в любом удобном ему порядке фрагменты текста, проверяет правильность выполненных ЭВМ операций и тут же исправляет возможные ошибки структуризации. Лишь после завершения этапа корректировки текст может считаться подготовленным к включению в состав текстовой части Фонда.

Что же касается самого представленного здесь метода структуризации, то – несмотря на его высокую эффективность (всего 0,06% ошибок на 150 страниц текста) – работы по его совершенствованию продолжаются как в плане улучшения скорости вычленения предложений, так и с целью расширения его 'лингвистических' возможностей.

Л И Т Е Р А Т У Р А

Азарова И.В., Горохова С.И., Григорьев Г.Г., Кузнецова Е.А. Разработка автоматических словоуказателей и конкордансов для художественных текстов. – В кн.: Структурная и прикладная лингвистика. Вып. 2. – Л.: Изд-во ЛГУ, 1983, с. 187–190.

Андрющенко В.М. Машинный фонд русского языка: постановка задачи и практические шаги. – ВЯ, 1985, № 2, с. 54–64.

Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка. – В кн.: МФРЯ: идеи и суждения. М.: Наука, 1986, с. 26–43.

Бусленко В.Н. Автоматизация имитационного моделирования сложных систем. М., 1977, с.

Герд А.С. Типы русских текстов и организация Машинного фонда русского языка. – В кн.: МФРЯ: идеи и суждения. М.: Наука, 1986, с. 67–74.

Герд А.С. Русская морфология и Машинный фонд русского языка. – ВЯ, 1986, № 6, с. 90–96.

Гиндин С.И. – В кн. : МФРЯ: идеи и суждения. М.: Наука, 1986, с. 176–179.

Гринбаум О.Н., Мартыненко Г.Я., Фитиалов С.Я. Об одном эвристическом алгоритме структуризации художественных текстов. – В кн.: Автоматизированные системы переработки текстовой информации (АСПТИ). – Тезисы докл. – Львов: УНИИПП, 1985, с. 52–54.

Ершов А.П. Машинный фонд русского языка (Внешняя постановка вопроса). – ВЯ, 1985, № 2, с. 51–53.

Караулов Ю.Н. Методология лингвистического исследования и Машинный фонд русского языка. – В кн.: МФРЯ: идеи и суждения. М.: Наука, 1986, с. 13–25.

Котов Р.Г. Машинный фонд русского языка и автоматизация словарно-терминологического обеспечения информационной службы. – В кн.: МФРЯ : идеи и суждения. М.: Наука, 1986, с. 84–89.

Лесскис Г.А. О размере предложений в русской научной и художественной прозе 60-х годов XIX века. – ВЯ, 1962, № 2, с. 78–95.

Мартыненко Г.Я. Многомерный синтактико-статистический анализ художественной прозы. – В кн.: Структурная и прикладная лингвистика. Вып. 2. – Л.: Изд-во ЛГУ, 1983, с. 58–72.

Машинный фонд русского языка : идеи и суждения. М. : Наука, 1986,240 с.

Рогожникова Р.П. Машинный фонд русского языка и словарное дело. – ВЯ,1985, N 4,с.54-60.

Розенцвейг В.Ю. Опыт создания национальных лексикографических служб за рубежом. – В кн.: МФРЯ:идеи и суждения.М.:Наука,1986,с.75-83.

Тулдава Ю.А. О квантитативных характеристиках богатства лексического состава художественных текстов. – Учен.зап. Тарт.ун-та,вып.437. *LINGUSTIKA* .Тарту,1977,с.159-175.

Тулдава Ю.А. О машинных фондах русского языка за рубежом. – В кн.:МФРЯ:идеи и суждения.М.:Наука,1986.с.141-142.

Шайкевич А.Я. – В кн.: МФРЯ : идеи и суждения.М.:Наука,1986, с.226-230.

Шапиро А.Б. Современный русский язык.Пунктуация.М.,1974,287с.

BELLES - LETTRES STRUCTURIZING BY MEANS OF COMPUTER

1.FORMAL-PUNCTUATOIN METHOD OF STRUCTURIZING

OLES N.GRINBAUM

S U M M A R Y

THE ARTICLE DESCRIBES SOME METHOD OF BELLES-LETTRES STRUCTURIZING IN ORDER TO FORM A TEXT BULK OF THE MACHINE FUND OF THE RUSSIAN LANGUAGE. THE METHOD IS BASED ON AN ALGORITHM APPROACH TO THE RULES OF THE RUSSIAN PUNCTUATION AND IS BEING APPLIED FOR AUTOMATIC MARKING-OUT OF BELLES-LETTRES (MARKING-OUT OF CHAPTERS, PARAGRAPHS, SENTENCES, USES OF WORDS) AND MARKING-OUT OF PARAGRAPHS ACCORDING TO THE WRITTEN-FORM TYPES OF SPEECH. THE METHOD WAS CHECKED ON THE PROSE OF M.LERMONTOV BY MEENS OF A COMPUTER.

РАСПРЕДЕЛЕНИЕ РЕАКЦИЙ В ЭКСПЕРИМЕНТАХ ПО ВЕРБАЛЬНЫМ АССОЦИАЦИЯМ

В.А. Долинский

Среди семантических характеристик лексики одной из наименее изученных и наиболее верифицируемых является ассоциативная способность слов. Исследование ассоциативности опирается на хорошо известный в психологии эксперимент, впервые проведенный Ф. Гальтоном (Galton) в 1879 г. Он состоит в том, что испытуемому предъявляется некоторое слово-стимул и предлагается отреагировать на это слово первым "пришедшим в голову" словом или словосочетанием. В ассоциативном эксперименте, проводимом с группой испытуемых, частота появления той или иной реакции отражает общеизвестность этой ассоциации и свидетельствует о существовании таких же связей между словом-стимулом и словом-реакцией у данной группы. Разнообразие, ассортимент ассоциаций, полученных от группы, свидетельствует о широте или узости ассоциативных связей у слова-стимула. Собственно лингвистический характер вербальных ассоциаций вытекает из двух обстоятельств. С одной стороны, установлены различия в реакциях испытуемых на предметы, их изображения и названия, с другой стороны, наблюдаются различия результатов ассоциативных экспериментов по языкам. Путем повторных экспериментов и статистического анализа получаемых данных становится возможным объективно выявить семантические поля ассоциирования. Так возникает представление о вероятностном семантическом пространстве слова, об ассоциативном значении, как о мере предпочтения для ценностных представлений, упорядоченных на шкале ассоциаций, или как о мере потенциальной возможности раскрытия смысла слова. Ассоциативное значение - это потенциальное распределение ответов различных информантов на данное слово-стимул. Такое распределение нельзя получить от одного индивида. Только массовый эксперимент, который можно планировать и повторять, с получением одной реакции на один стимул от большого числа испытуемых дает распределение, приближающееся к "потенциальному" с ростом числа информантов. Если мы имеем, скажем, 1000 испытуемых, то ясно, что каковы бы ни были их индивидуальные предпочтения, общая тенденция "пробьется" в наиболее частых ответах, а сколь велик бы ни был вес самой популярной ассоциации (primary), ассортимент различных реакций на данное слово-стимул останется специфическим. Предъявляя группу стиму-

лов группе испытуемых, мы можем получить семейство распределений реакций для данных стимулов, сопоставимых благодаря нормировке.

Материалом исследования послужили распределения реакций в экспериментах по свободным вербальным ассоциациям с регистрацией первого ответа (дискретные ассоциации). Использованы данные четырех экспериментов, проводившихся с американцами (2) и русскими (2), причем данный язык является родным для испытуемых. Приводим описания четырех экспериментальных массивов.

I. Массив J. "Миннесотские нормы" (Jenkins,1970). Количество испытуемых - в среднем, 1008. Стимулы - 61 существительное.

2. Массив P. "Калифорнийские нормы" (Postman, 1970). Количество испытуемых - 1000. Стимулы - 96 существительных.

3. Массив Л. (Словарь ассоциативных норм русского языка. Под ред. А.А. Леонтьева, 1977). Количество испытуемых - - в среднем, 700. Стимулы - 55 существительных.

4. Массив Д. Эксперимент проведен автором. Количество испытуемых - в среднем, 205. Стимулы - 12 прилагательных.

Общее число реакций, по всем массивам, - 198448. Число стимулов и, соответственно, распределений реакций, подвергнутых обработке, - 224. Из них - 157 по американским массивам и 67 - по русским. По каждому массиву параметры распределений были пересчитаны для равных объемов (1000 реакций). Определены следующие параметры.

1. Ассортимент реакций V, а также lgV.

2. Частота главного ответа (primary) F_1, а также lgF_1.

3. Уклон распределения γ, отношение логарифма частоты главного ответа к логарифму ассортимента:

$$\gamma = \frac{lgF_1}{lgV}.$$

4. Относительная неоднородность (вогнутость) распределения C:

$$C = \frac{lgF_1 \cdot lgV}{2} - lgN,$$

где N - число реакций (равное везде 1000).

Средние параметры распределений реакций по массивам (табл. I) достаточно устойчивы, что позволяет сделать ряд важных выводов. Все распределения могут быть аппроксимированы формулой, выражающей обратно пропорциональную зависимость рангов и частот реакций с переменным параметром γ, известной как закон Ципфа в четвертом приближении:

$$F_i = F_1 i^{-(\gamma + c \lg i)}$$

где i - ранг реакции, F_i - частота i-го ассоциата. Интегрированием получим:

$$\lg N = \lg F_1 \lg V - \gamma \frac{\lg^2 V}{2} - c \frac{\lg^3 V}{3}$$

где $V = i_{max}$. При $C > 0$, кривая вогнутая.

Среднее значение параметра γ для русского (флективно-синтетического) языка - 0,919, для американского варианта английского (аналитического) языка - I,098. В русском языке также выше ассортимент ассоциаций, в английском - выше доля самой распространенной реакции. Все распределения демонстрируют относительную однородность ($C < 0$), причем однородность выше - по американским массивам.

По каждому массиву рассчитывались так называемые интегральные распределения реакций. Для этой цели каждое из 224 распределений "свертывалось" в точку с координатами $\lg V$; $\lg F_1$, и все точки каждого массива выносились на диаграмму рассеяния в новой системе координат. Интегральная система координат отличается тем, что вместо переменных рангов и частот переменными здесь выступают концевые точки распределений $\lg V^j$; $\lg F_1^j$ - ассортимент и частота главного ответа каждого j-го распределения. Диаграммы рассеяния интегральных распределений были подвергнуты регрессионному анализу по массивам. Таким образом, "облако точек" каждого массива аппроксимировалось отрезком прямой, проходящей через "центр массива" с тангенсом наклона β (рис. I, табл. 2 и I). На графике видно, что с ростом ассортимента резко снижается частота главного ответа, соответственно возрастает однородность распределений, что не позволяет интегральному распределению сохранить вид гиперболы и оно описывается прямой; коэффициент корреляции находится в пределах $[0,49 \div 0,77]$ (см. табл. 2). Тангенс наклона β лежит в границах $[2,20 \div 4,63]$, что существенно отличается от тангенса наклона γ для отдельных распределений $[0,60 \div I,75]$, в среднем равного I,044. На графике нанесены линии, соответствующие $\gamma = I$ (биссектриса координатного угла) и $C = 0$ (отрезок гипербола), что позволяет наглядно представить динамику взаимозависимости параметров ассоциативных распределений. Обратно пропорциональная зависимость рангов и частот предстает в значительно более сложном виде, когда берется в расчет изменение кривизны бла-

годаря равенству числа реакций N для всех распределений). Широкое варьирование частоты главной реакции F_1 и умеренное варьирование ассортимента реакций V имеет эффект в виде изменения однородности (кривизны) распределений (расстояние до линии $C = 0$). Чем больше ассоциаций вызывает слово-стимул, тем ниже частота самой популярной ассоциации, тем выше однородность распределения в целом (более выпуклый график), то есть больше ассоциаций распространенных и меньше редких и уникальных. Другими словами, рост численности различных реакций сопровождается обычно большей прочностью ассоциаций.

Если ассоциативная способность слова связана с другими важными его характеристиками, то частота и полисемия слова-стимула не может не отразится на "профиле" распределения реакций в ассоциативном эксперименте. Была поставлена задача исследовать возможную функциональную связь между параметрами распределений реакций и общеязыковыми характеристиками стимула; при этом значения частоты и полисемии, хотя и тесно связанные (Zipf, 1936, 1945), должны быть рассмотрены независимо друг от друга.

Каждому стимулу был придан индекс частоты, для американских массивов был использован частотный словарь Kučera-Francis (1967), откуда индекс частоты стимулов был получен суммированием частот соответствующих словоформ; для русских массивов был взят частотный словарь Засориной (1977) с частотами лексем непосредственно. Так как объемы выборок обоих словарей равны (Кучера - 1 014 232, Засорина - 1 056 382), стало возможным разделить частотную шкалу на сопоставимые интервалы. Было выделено шесть таких интервалов с логарифмическим шагом: первый интервал - "нуль-частотные" слова (не вошедшие в словарь) и слова с частотой 1 - 3; второй интервал - слова с частотой 4 - 10; третий - с частотой 11 - 31; четвертый - 32 - 100; пятый - 101 - 316; шестой - 317-1000. Все стимулы со своими частотными индексами оказались в рамках соответствующих интервалов. Английские стимулы достаточно равномерно размещены в шести интервалах (см. табл.3). К сожалению, русские стимулы не столь разнообразны по частоте, тем не менее, они вошли в три интервала - 4, 5 и 6. Для каждого частотного интервала были найдены средние значения параметров распределений реакций на соответствующие слова-стимулы. Из всех параметров распределений наиболее "чувствительным" к общеязыковой частоте стимула оказался уклон

распределения γ. С ростом частоты стимула наблюдается падение уклона распределения γ, то есть рост ассортимента реакций (рис. 2, табл. 3). Наиболее употребительные слова, таким образом, вызывают повышенный ассортимент ассоциаций, понижая удельный вес самой распространенной реакции. Подобная зависимость характеризует все стимулы, кроме малознакомых для испытуемых (редких и "нуль-частотных"): так, в первом интервале наблюдается падение γ и "стохастические" реакции, характеризующиеся большим разнообразием.

Для выявления связи ассоциативного значения с полисемичностью слова каждому стимулу был придан индекс полисемии, определенный по количеству значений в толковых словарях. Английским стимулам присваивался индекс по словарю Webster (1978), для русских слов использован Словарь русского языка в 4 тт. (1957-1961). Слова-стимулы по количеству значений заняли широкий диапазон от 1 до 21 значения в английском языке и от 1 до 12 значений в русском языке. Шкала полисемии была также разделена на сопоставимые интервалы с логарифмическим шагом. Было выделено пять полисемических интервалов как для английского, так и для русского языка: первый интервал – слова с одним значением; второй – с двумя значениями; третий интервал – слова с тремя и четырьмя значениями; четветрый – слова с числом значений от 5 до 8; пятый – все слова с девятью и бо́льшим числом значений.[+] Все стимулы со своими индексами полисемии были размещены в соответствующих интервалах (табл. 4). Все интервалы как для английских, так и для русских стимулов были заполнены. Для каждого интервала были найдены средние значения параметров распределений реакций на слова-стимулы соответствующих полисемических интервалов. Кроме четырех основных параметров F_1, V, γ, C, на связь с полисемией испытывались такие параметры как число одноразовых ассоциаций m_1 и другие. Наиболее зависимым от полисемичности стимула оказался параметр относительной неоднородности распределения C. Чем больше значений у слова-стимула, тем выше отрицательная величина параметра C, то есть тем более однородно (выпукло) распределение реакций. Наиболее многозначные слова, таким образом, вызывают ассоциации, более однородные по своему распределению, повышая удельный вес частых ассоциаций и понижая долю редких ассоциаций. Эта зависимость (рис. 3, табл. 4) характеризует все стимулы кро-

[+] См. Поликарпов, 1987, с. 139

ме однозначных: в первом интервале наблюдается повышение однородности в обоих языках.

Следует отметить, что группы стимулов, вошедшие в частотные интервалы, и группы стимулов, вошедшие в полисемические интервалы, не совпадают. Не обнаружено значимых зависимостей между частотой стимула и однородностью распределений реакций (-C), между полисемией стимула и уклоном распределений реакций (γ). Анализ связи между частотой и полисемией стимулов для массивов J, P и Л показал, как и ожидалось, устойчивую корреляцию между ними. Становится, таким образом, понятной связь параметров распределений с характеристиками частоты и полисемии стимулов. Более частые в языке слова являются, как правило, в среднем, и более многозначными. Следовательно, более "пологое" (с более низким уклоном γ) распределение должно быть одновременно и более однородным (выпуклым, с более отрицательным значением C). Именно это и наблюдалось в динамике взаимозависимости параметров распределений, представленных в интегральной системе координат (см. рис. I): степень однородности распределения и ассортимент ассоциаций, уклон распределения - параметры взаимосвязанные, но и относительно свободные. Точно так же, как существуют слова с "плюсовой" и "минусовой" полисемией (Тулдава, 1979), отклоняющейся от средней для данного частотного интервала величины, существуют ассоциативные значения "плюсового" и "минусового" характера, более или менее однородные (выпуклые), чем у стимулов той же частоты, но со "стандартной" полисемией. Исключение составляют слова редкие и/или однозначные.

Значимыми являются различия межъязыкового характера. Для массивов русского языка характерно большее разнообразие ассоциаций и, соответственно, меньшая доля частоты главной ассоциации, чем для массивов английского языка, что связано с типологическими особенностями языка более синтетического и особенностями организации лексической памяти его носителей. То же более низкое значение γ в распределениях реакций на отдельные стимулы (а не средних), отмечено для русских массивов, когда стимул являлся переводом соответствующего стимула в американских экспериментах (table - стол, girl - девочка).

Что касается различий между результатами разных экспериментов, как о них можно судить по интегральным распределениям по массивам (рис. I), то, хотя массивы расположены на

некотором расстоянии друг от друга, закономерности в них остаются повторяющимися, общие тенденции проявились во всех четырех массивах: падение γ с ростом частоты стимула, рост однородности (падение C) с ростом полисемии стимула. Отличия же распределений в целом отчасти объясняются разными условиями проведения экспериментов. Так, большее разнообразие ответов в эксперименте Постмэна по сравнению с экспериментом Дженкинса могло быть вызвано тем, что в первом случае испытуемыми являлись студенты-лингвисты. Вариативность результатов экспериментов по вербальным ассоциациям возрастает с повышением частоты слова-стимула. Так, в самых частых стимулах наблюдается наибольшее разнообразие отклонений при повторных экспериментах.

В результате анализа ранговых распределений вербальных ассоциаций установлена обратно пропорциональная зависимость рангов и частот реакций со средним значением параметра γ = = I,04. С ростом ассортимента ассоциаций существенно понижается доля самой популярной ассоциации (primary), при этом, как правило, повышается их прочность (однородность). Установлено наличие связей между ассоциативной способностью (ассоциативным значением) слов, их общеязыковой частотой и полисемичностью. Чем чаще слово употребляется в речи, тем больше ассоциативных связей оно имеет с другими словами. Чем выше полисемичность слова, тем выше прочность (однородность) его ассоциативных связей.

Ассоциативная способность слов является важнейшей языковой характеристикой, отражающей глубинные закономерности человеческого мышления, познания, общения. Ассоциативные нормы, благодаря своей статистической надежности, поддаются математической обработке. Психосоциальная природа ассоциаций, выявляемых только в массовом эксперименте, требует вероятностного подхода в квантитативной лингвистике. Ассоциативный потенциал (смысловая наполненность, meaningfullness) элементов лексической системы определяется их местом, рангом в пространстве семантического континуума. В эксперименте происходит оценивание смысла - задание на семантическом континууме весовой функции, придающей различные веса разным участкам семантического поля, разным ассоциациям. Закономерности, открываемые в результате статистического анализа ранговых и интегральных распределений ассоциатов, являются выражением системного вероятностного характера языка, накладывающего свои связи на устройство лексической системы и ее

единиц. Связи, ассоциации между лексическими единицами, возникая и повторяясь в специфических контекстах и конситуациях, закрепляются в процессах коммуникации, познания, средоформирующей деятельности и превращаются в сознании говорящих в значения. Значения сохраняют и связи с внеязыковой реальностью, денотатами, референтами, и с лексическими единицами, которые с ними ассоциируются. Значение - не ассоциация, но з н а н и е а с с о ц и а ц и й. Семантика слов служит банком социальной памяти. Семантическому уровню языка наиболее адекватно ассоциативное значение слов, состоящее из с у м м ы а с с о ц и а ц и й языковых представлений с внеязыковыми (ассортимент), свойственных индивиду и социуму (частота). Эта сумма ассоциаций не есть общее значение слова, ни главное его значение, ни набор значений или лексико-семантических вариантов. Эта сумма ассоциаций есть о б р а з с л о в а. Неопределенность и реальность образа слова осознавались многими. Приведем две цитаты. "В значительном числе случаев лексическое значение слова невозможно охарактеризовать с полной определенностью" (Шмелев, 1973, с. 21). "Не бывает ни малейшей трудности осознать слово как психологически нечто реальное" (Сепир, 1934, с. 27). Приближенное представление об образе слова дает статистическое распределение ассоциаций, полученных экспериментально. Здесь слово может рассматриваться в координатах "знак-денотат" (когнитивная ось рангов), "индивид-социум" (коммуникативная ось частот).

В лексической системе действуют принципы намекания, наименьшего усилия и согласованного оптимума. Полисемия слова отражает его ассоциативные свойства, она есть функция от числа его наиболее популярных ассоциаций, от количества наиболее известных контекстов. Сеть социально значимых ассоциаций отражается в речи, где действует принцип н а м е к а н и я, связывающий языковые представления - с внеязыковыми, социальные - с индивидуальными. Если повышенная частота отдельного слова связана с сокращением артикуляторных программ, укорочением длины слова, то высокая частота вербальных ассоциаций связана с сокращением затрат нервной энергии, укорочением длин нервных путей (Аминев, 1972). И здесь, и там проявляется принцип н а и м е н ь ш е г о у с и л и я. Психосоциальное единство образа слова отражается в распределении реакций на произнесение этого слова в эксперименте. Это распределение описывается законом Ципфа. Так действует принцип с о г л а с о в а н н о г о о п т и м у м а (equi-

librium).

Вероятностный подход в квантитативной лингвистике позволяет описывать в рамках единой модели, с одной стороны, дискретные миры лингвистических единиц и их связей, с другой стороны, континуальные миры экстралингвистической реальности и сознания говорящих. Стабильность относительных частот элементов, взаимозависимость характеристик ранговых и интегральных распределений разных уровней, интерпретируемость их параметров - все это должно послужить основой естественнонаучного подхода к феномену естественных языков, к психосоциальному миру и требует философского осмысления на основе широких междисциплинарных исследований.

ЛИТЕРАТУРА

Аминев Г.А. Вероятностная организация центральных механизмов речи. - Казань: Изд. Казанск. ун-та, 1972.

Поликарпов А.А. Полисемия: системно-квантитативные аспекты. // Уч. зап. Тартуск. ун-та, вып. 774. - Тарту, 1987. - С. 135-154.

Сепир Э. Язык. Введение в изучение речи. - М.; Л., 1934.

Словарь русского языка в четырех томах. - М.: ГИС, 1957-1961.

Словарь ассоциативных норм русского языка. / Под ред. А. А. Леонтьева. - М.: Изд-во МГУ, 1977.

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. // Уч. зап. Тартуск. ун-та, вып. 502. - Тарту, 1979. - С. 107-141.

Частотный словарь русского языка. / Под ред. Л.Н. Засориной. - М.: Русский язык, 1977.

Шмелев Д.Н. Проблемы семантического анализа лексики. - М.: Наука, 1973.

Jenkins J.J. Minnesota Word Association Norms. // Norms in Word Association. - L.: Academic Press inc., 1970.

Kučera H., Francis W.N. Computational analysis of Present-Day American English. - Providence: Brown Univ. Press, 1967.

Postman L. California norms: association as a function of word frequency. // Norms in Word Association. - L.: Academic Press inc., 1970.

Webster's 3'd New International Dictionary of the English Language. - Springfield, Mass, 1978.

Zipf G.K. The psycho-biology of Language: An introduction to dynamic philology. - L.: Routledge, 1936.

Zipf G.K. The meaning-frequency relationship of words. //The Journal of General Psychology, 1945, v. 33, N 2, p.251-255.

Таблица I

СРЕДНИЕ ПАРАМЕТРЫ РАСПРЕДЕЛЕНИЙ РЕАКЦИЙ ПО МАССИВАМ

Массив	$\lg V$	$\lg F_1$	γ		C	
J	I,998	2,483	I,243	I,098	- 0,520	- 0,453
P	2,2II	2,343	I,006		- 0,4I0	
Л	2,350	2,207	0,939	0,989	- 0,406	- 0,383
Д	2,57I	2,I23	0,826		- 0,27I	

Таблица 2

СРЕДНИЕ ПАРАМЕТРЫ ИНТЕГРАЛЬНЫХ РАСПРЕДЕЛЕНИЙ РЕАКЦИЙ ПО МАССИВАМ

Массив	β*		Наклон	r**
J	2,92	2,48	7I°20'	0,65
P	2,20		65°33'	0,69
Л	4,63	4,24	77°49'	0,49
Д	2,43		67°36'	0,77

*β - тангенс наклона интегрального распределения
**r - коэффициент корреляции

Таблица 3

ЗАВИСИМОСТЬ ПАРАМЕТРА γ РАСПРЕДЕЛЕНИЙ РЕАКЦИЙ ОТ ОБЩЕЯЗЫКОВОЙ ЧАСТОТЫ СЛОВ-СТИМУЛОВ В ДВУХ ЯЗЫКАХ

Частота стимула		0 - 3	4 - I0	II - 3I	32 - I00	I0I - 3I6	3I7 - I000
Интервал		I	2	3	4	5	6
Англ. язык (J,P)	γ	I,048	I,I73	I,I98	I,I79	I,I42	I,I3I
	n*	39	I9	2I	3I	3I	I6
Русск. язык (Л,Д)	γ				0,949	0,929	0,888
	n*				9	45	I3

*n - число стимулов.

Таблица 4

ЗАВИСИМОСТЬ ПАРАМЕТРА C РАСПРЕДЕЛЕНИЙ РЕАКЦИЙ И ПОЛИСЕМИИ СЛОВ-СТИМУЛОВ /ПО ТОЛКОВЫМ СЛОВАРЯМ/ В ДВУХ ЯЗЫКАХ

Полисемия стимула		I	2	3 - 4	5 - 8	9<
Интервал		I	2	3	4	5
Англ. язык (J,P)	C	- 0,403	- 0,376	- 0,386	- 0,405	- 0,502
	n	25	38	43	4I	I0
Русск. язык (Л,Д)	C	- 0,389	- 0,25I	- 0,283	- 0,395	- 0,557
	n	I4	9	26	I4	4

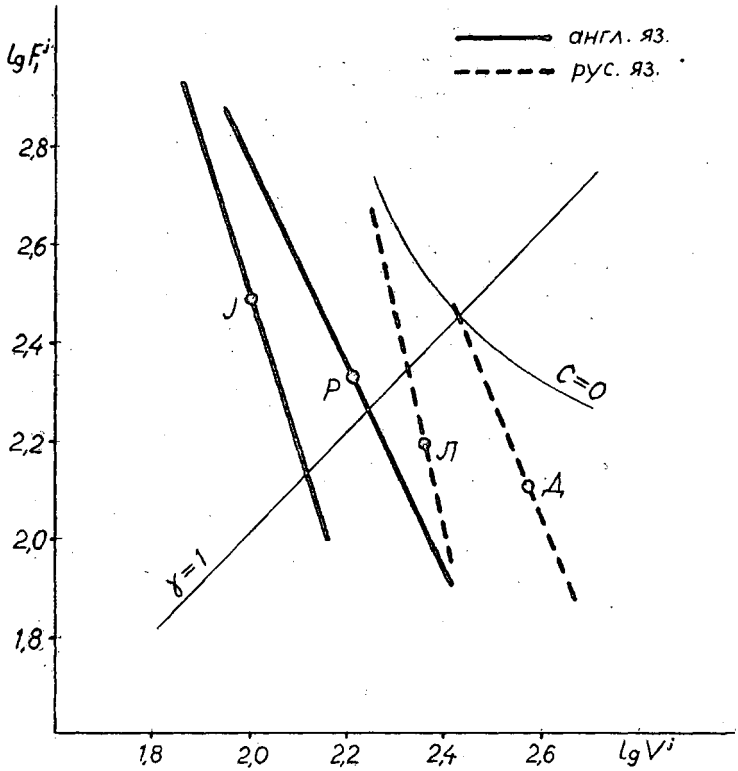


Рис.I Интегральные распределения реакций.
V^j, F_i^j - концевые координаты отдельных распределений, соответствующие ассортименту реакций и частоте главной реакции на стимул j . Отрезки J,P,Л,Д аппроксимируют семейства точек для распределений соответствующих массивов.

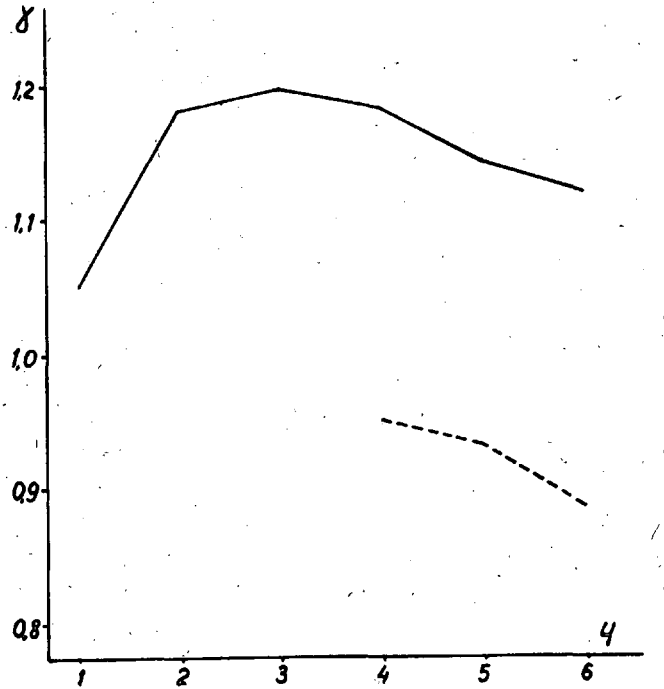


Рис.2 Связь между между частотой /Ч/ слов-стимулов и уклоном /γ/ распределений реакций в двух языках.
По оси Ч отложены частотные интервалы.

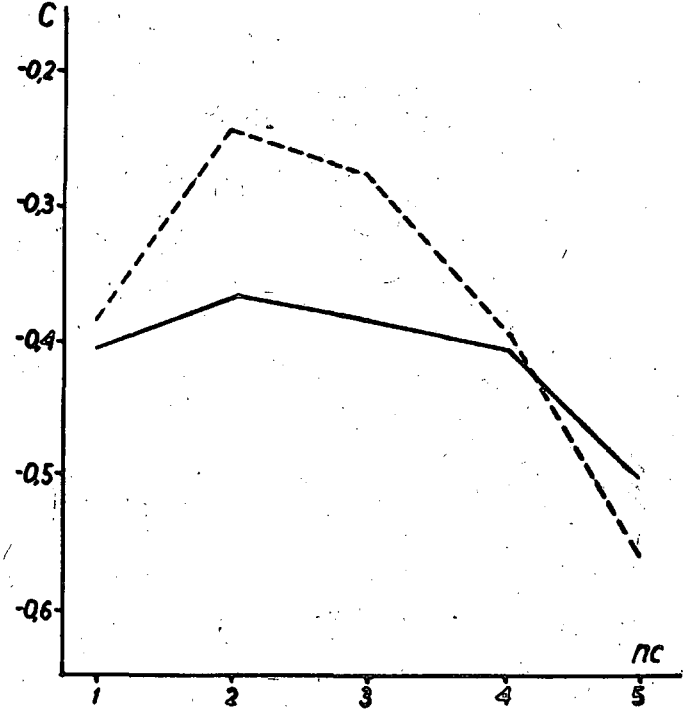


Рис.3 Связь между полисемией /ПС/ слов-стимулов и относительной неоднородностью /С/ распределений реакций в двух языках.
По оси ПС отложены полисемические интервалы.

DISTRIBUTION OF RESPONSES IN WORD ASSOCIATION TEST

Vladimir A. Dolinskiy

S u m m a r y

In this article are described the results of the investigation of quantitative characteristic features of associative capacity of 363 words of English and Russian languages. Rank-frequency counts of word associations yields Zipf-type curve with a slope (on the average) of 1,098 (English); 0,919 (Russian). It is significantly reduced the share of the most popular responses ("primary") with the responses availability growth and simultaneously the responses homogeneity as a rule is increased.

There are established relations of dependency between associative capacity and linguistic frequency and polysemy. The more frequently a word is used in speech the more it has associative relations with other words. The higher polysemy of a word the higher strength of its associations.

КОММУНИКАЦИЯ "ЧЕЛОВЕК-ЧЕЛОВЕК" и "ЧЕЛОВЕК-ЭВМ": СУЩНОСТЬ И ОБЕСПЕЧИВАЮЩИЕ СРЕДСТВА. КОНЦЕПТУАЛЬНОЕ МОДЕЛИРОВАНИЕ КАК ЛИНГВИСТИЧЕСКАЯ ЗАДАЧА

Р.Ю. Кобрин

Одной из центральных проблем современного языкознания является проблема роли языка в общении, шире - в процессах коммуникации. Все процессы научной и технической коммуникации, происходящие в мире, основываются в конечном счете на использовании естественного языка. Исследование роли языка в процессах коммуникации "неизбежно требует выхода за поверхностные формы языка в тот уровень, который определяет информационную ценность языковых единиц" (Принципы и методы ..., 1976, с. 3).+

Под коммуникацией обычно понимается обмен информацией между социальными и материальными объектами, осуществляемый при помощи семиотической (знаковой) системы, в которой знак выступает как материальный носитель социальной информации (Cornforth M., 1973; Hoffman L., 1976; Merten K., 1977; Steinmüller U., 1977; Meier G.F., Meier B., 1979-1980; Meggle G., 1981).

Одними из главных функций языка являются:

участие в формировании мысли;

выражение информации о работе мысли;

возбуждение мысли участника коммуникации, более или менее адекватной мысли автора.

Функция выражения информации о работе мысли имеет колоссальное значение в процессе общения и является одной из основ и условий успешного осуществления этого процесса. Воспринимая выраженную средствами знаковой системы информацию о деятельности мысли автора, участник коммуникации получает возможность воссоздать эту деятельность - разумеется, всегда с известными потерями и наслоениями (Березин Ф.М., Головин Б.Н., 1979, с. 75; Новиков Л.А., 1982, с. 42-47). Об этом писал А.А. Потебня, считавший невозможным одинаковое понимание у говорящего и слушающего потому, что разный воз-

+ Ср. замечания проф. Г.В. Колшанского: "Коммуникация и информация - две стороны одного и того же явления ... Номинация и коммуникация - две стороны одного и того же нераздельного процесса, в котором ведущей является коммуникация" (Колшанский Г.В., 1976, с. 3, 22).

раст, социальные условия, различный жизненный опыт людей обусловливают индивидуальное восприятие предметов действительности: "Люди ... понимают друг друга не таким образом, что действительно передают один другому знаки предметов ... и не тем, что взаимно заставляют себя производить одно и то же понятие, а тем, что затрагивают друг в друге то же звено цепи чувствительных представлений и понятий, ... вследствие чего в каждом восстают соответствующие, но не те же понятия" (Потебня А.А., 1862). Это высказывание, казалось бы, противоречит представлению о жесткой логико-понятийной организации терминов, обеспечивающих специально-профессиональное общение. Но анализ конкретных актов научной и учебной коммуникации показывает, сколь затруднительно одинаковое понимание специальных понятий.

Под коммуникативной функцией знаковых единиц будем понимать способность знаков вызывать в сознании приемника-участника коммуникации содержание (образы, представления в виде значений знаков), соответствующее содержанию передаваемого знака в сознании передатчика - участника коммуникации (Принципы и методы ..., 1976).

Таким образом, обычно употребляемый в литературе термин п е р е д а ч а и н ф о р м а ц и и не характеризует точно существо вопроса, так как информация, строго говоря, не передается, но воссоздается участником коммуникации. Более приемлем термин о б м е н и н ф о р м а ц и е й, который мы и будем использовать при рассмотрении типов коммуникации.

Можно выделить следующие типы коммуникации и - соответственно - обеспечивающей ее семиотической системы.

I. Общение "человек-человек" - обмен социальной информацией между людьми.

Обеспечивающее средство - естественный язык. Синоним - "человеческая коммуникация", "межчеловеческое общение".

2. Общение "человек-машина" - обмен информацией между человеком и машиной.

Частный вид - общение "человек-ЭВМ-человек".

Обеспечивающие средства - системы машинных команд, языки программирования, формализованные фрагменты естественного языка в диалоговых информационных системах. Синонимы - "человеко-машинное общение", "человеко-машинная коммуникация".

3. Общение "машина-машина" - обмен информацией между взаимодействующими техническими устройствами.

Частный вид – общение между ЭВМ, объединенными в вычислительные комплексы. Обеспечивающее средство – система машинных команд для передачи информации между ЭВМ вычислительного комплекса. Синонимы – "машинное общение", "машинная коммуникация", иногда – "техническая коммуникация" (Андреев Н. Д., 1967, с. 157-158).+

Эпоха НТР привела к серьезным структурным и содержательным изменениям в коммуникативных процессах человеческого общества.

Можно выделить два принципиальных момента, определяющих современную коммуникацию:

1. В условиях НТР наиболее динамичной частью лексико-семантической системы языка становится научно-техническая терминология (Азимов П.А. и др., 1975, с. 5).

Новые технологии, информатизация и компьютеризация общества приводят к появлению новых понятийно-терминологических систем, обеспечивающих коммуникацию.

Бурный количественный рост числа терминов в различных областях знания, активное проникновение терминов в "общенародный" язык ведут к актуализации терминоведческих проблем.++

+ Справедливо утверждение Н.Д. Андреева о том, что "содержание междумашинной коммуникации вполне подобно смыслу действий станка с программным управлением, между тем "общение" "человек-машина" и "человек-человек" существуют только через человека и для человека" (Андреев Н.Д., 1967, с. 197).

++ См. замечание Ф.П. Филина: "Кроме общеупотребительных слов имеется неисчислимое множество специальных терминов и значений слов, которые известны только тем, кто работает в разных отраслях хозяйства, науки, техники и культуры, – таких терминов в русском языке миллионы" (Филин Ф.П.,1980, с. 82).

Интересные данные, характеризующие рост числа публикаций по терминоведению, приведены Н.П. Романовой. В 30-е годы в СССР было выпущено 11 работ по терминоведению, в 40-е – 11, в 50-е – около 50-ти, в 60-е – более 400 (Романова Н.П., 1976, с. 96-100).

В то же время нельзя не согласиться с мнением Б.Н. Головина, который писал: "Хотя в последнее десятилетие возросло число публикаций в области терминоведения, в самой лингвистической науке, кажется, все еще не пробудился достаточно глубокий интерес к взаимосвязям лингвистических терминов и лингвистических идей, несмотря на обилие требующих этого пробуждения фактов" (Головин Б.Н., 1976, с. 20).

Непосредственной причиной этого является необходимость разработки оптимальных терминологических средств номинации новых объектов, а также эффективного терминологического обеспечения процессов научного, технического и учебного общения.

2. Получает широкое распространение человеко-машинная коммуникация, суть которой состоит в обеспечении диалогового взаимодействия человека и автоматизированных информационных систем.

Организация эффективного взаимодействия человека и ЭВМ требует решения ряда задач лингвистики и информатики, в том числе и задачи построения языков общения – языков программирования и языков диалоговых информационных систем (ДИС).

Языки программирования и языки ДИС – это разные по своей природе и функциям семиотические системы.

Языки программирования – искусственные семиотические системы, цель которых обеспечить управление вычислительными средствами, поэтому их структуры определяются технической и технологической организацией вычислительных средств. В 60-е – 70-е годы языки программирования были ориентированы на профессиональных программистов. В начале 80-х годов была поставлена задача максимального приближения языков программирования к естественному языку.

Количество языков программирования быстро растет и в настоящее время исчисляется сотнями. Современные языки программирования представляют собой системы, сопоставимые по сложности с естественными языками. Они обладают развитым "словарем", включающим как математические символы, так и слова и выражения естественного языка, а также располагают достаточно сложным "синтаксисом" – правилами формирования команд и составления текстов-программ (Андрющенко В.М., 1980).

Качественная эволюция языков программирования в последние два десятилетия увеличивает их функционально-структурное сходство с естественными языками. Для обозначения основных понятий все шире используются слова и словосочетания естественного языка, то есть естественно-языковые терминологии.

Отмечается возникновение и развитие естественно-искусственного двуязычия, обусловленного увеличением числа программистов (Лингвистические вопросы ..., 1983).

Языки диалоговых систем решают две задачи:

– формализацию предметной области,

– общение человека и ЭВМ на естественном языке.

Любая формализация включает определенные содержательные описания моделируемых фрагментов реальной предметной области в терминах естественного языка.

Языки ДИС реализуют модель предметной области, представленную в терминах естественного языка. Эта модель предметной области лежит в основе современных информационных систем – банков данных.

Банк данных – высокопроизводительная информационная система, реализованная на современных средствах вычислительной техники с помощью прогрессивных средств математического обеспечения – СУБД (систем управления базами данных), содержащая базы данных в виде формализованных описаний и обеспечивающая работу пользователей в диалоговом режиме на естественном языке.

Принципиальные различия между банками данных и информационными системами 60-х – 70-х годов заключаются в реализации однократного ввода информации, достижении независимости процессов описания данных от их машинного представления ("физического" представления в терминологии банков данных), обеспечения возможности эксплуатации банка многими пользователями.

Проектирование банков данных является новым этапом в развитии информационных систем. В настоящее время стало очевидным, что для всех систем обработки данных, включая автоматизированные системы управления (АСУ предприятиями, АСУ отраслями народного хозяйства, АСУ технологическими процессами, АС научных исследований и др.), эффективное функционирование возможно лишь на основе единой крупной информационной системы – информационного банка данных. В этом банке должны храниться как формализованные фактографические данные, так и неформализованные текстовые сообщения на ЕЯ. Кроме того, в памяти ЭВМ должна храниться математическая модель объекта и его описание на ЕЯ. Общение пользователей также должно вестись в терминах ЕЯ (Брябрин В.М. и др., 1983, с. 3).

Центральная проблема проектирования банка данных – конструирование базы данных, представляющей собой фиксированную в ЭВМ совокупность данных, отображающую структуру объекта и отношения между объектами (Системы управления ..., 1984, с. 5).

Базы данных основываются на концептуальных (инфологических) моделях данных, которые содержат полное информационное описание объектов и отношений между ними и определяются ло-

гико-математическим или логико-лингвистическим структурированием предметной области.

Конкретным выражением концептуальной модели может быть: I) логико-математическая модель объекта, 2) логико-лингвистическая модель объекта в виде его языкового представления (словаря, тезауруса, сетевой или фреймовой структуры).

Концептуальное моделирование требует построения математических моделей предметных областей. Известно, что информационные процессы, в отличие от энергетических, дискретны, и для их описания целесообразны дискретные математические модели. Однако реальные фрагменты действительности часто несводимы к их формальным математическим моделям. Поэтому "любая формализация проектирования банка данных явно или неявно включает в себя определенные содержательные описания моделируемых фрагментов реальной, исследуемой предметной области" (Кокорева А.В., Малащинин И.И., 1984, с. I2).

При проектировании концептуальной модели определяются языковые единицы, называющие объекты и их составные части, а также отношения между ними. Таким образом, для построения концептуальной модели необходимо разработать перечень лексических единиц и конкретизировать отношения между ними. Так как концептуальные модели представляют предметные области науки, техники, управления, социальной жизни общества, эти модели основываются на понятийно-терминологических системах соответствующих отраслей деятельности и должны использовать лингвистические методы описания и представления терминологий.

Модель должна быть динамичной, т.е. предусматривать описание цели действия и всех возможных информационных ситуаций. Модель может быть либо тезаурусной, либо сетевой и ориентироваться на семантическое представление предметной области в целом, либо фреймовой и ориентироваться на адекватное представление конкретных актуальных информационных ситуаций.

Современная терминология и прикладной характер решаемых задач не могут затемнять очевидный теоретико-лингвистический характер концептуальных моделей.

Языковой основой семантических полей является естественно-языковая терминология, поэтому терминологические исследования определяют процедуры концептуального моделирования предметных областей.

Между тем современные банки данных реализуют концепту-

альные модели, построенные прагматическим путем: без серьезных исследований понятийно-терминологических систем постулируется схема предметной области и определяется структура данных. Для описания данных и манипуляций с ними пытаются также использовать различные алгебраические методы, прежде всего методы аппликативных вычислительных систем, широко применяются реляционные модели. Эксплуатация БД показывает, что реальные фрагменты действительности: I) должны быть описаны в терминах ЕЯ, 2) не могут быть сведены к их формальным моделям. В последние годы все чаще высказывается мнение, что возможность задания концептуальных моделей в современных БД отсутствует (Дейт Р., I980; Будзко В.И., I983).

Одной из актуальных задач современной лингвистики оказывается, таким образом, разработка методов терминологической работы при создании концептуальных моделей данных.

Итак, необходимость эффективного терминологического обеспечения научного, технического, управленческого и учебного общения, включая терминологическую работу при концептуальном моделировании, требует создания и конкретизации теории и практики работы с терминами и терминологиями.

Л И Т Е Р А Т У Р А

Азимов П.А. и др. Научно-техническая революция и язык // Вопросы языкознания, I975, № 2.

Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языкознании. – Л.: Наука, I967.

Андрющенко В.М. Лингвистический подход к изучению языков программирования и взаимодействия с ЭВМ // Проблемы вычислительной лингвистики и автоматической обработки текста на естественном языке. – М.: МГУ, I980.

Атре Ш. Структурный подход к организации баз данных. – М.: Финансы и статистика, I983.

Березин Ф.М. История русского языкознания. – Высшая школа, I979.

Березин Ф.М., Головин Б.Н. Общее языкознание. – М.: Просвещение, I979.

Брябрин В.М. и др. Диалоговые системы в АСУ. – М.: Энергоатомиздат, I983.

Будзко В.И. Предисловие к (Атре Ш., I983).

Головин Б.Н. Лингвистические термины и лингвистические идеи // Вопросы языкознания, I976, № 3.

Дейт Р. Введение в системы баз данных. - М.: Наука, 1980.
Кокорева Л.В., Малашинин И.И. Проектирование банков данных. - М.: Наука, 1984.
Колшанский Г.В. Некоторые вопросы семантики языка в гносеологическом аспекте // Принципы и методы семантических исследований. - М.: Наука, 1976.
Лингвистические вопросы алгоритмической обработки информации. - М.: Наука, 1983.
Мартин Дж. Организация баз данных в вычислительных системах. - М.: Мир, 1980.
Новиков Л.А. Семантика русского языка. - М.: Высшая школа, 1982.
Принципы и методы семантических исследований. - М.: Наука, 1976.
Потебня А.А. Мысль и язык. Харьков, 1862. Цит. по (Березин Ф.М., 1979, с. 126).
Романова Н.П. О типологии терминов // Актуальные проблемы лексикологии. Доклады III Межвузовской конференции 3-7 мая 1976 г. - Новосибирск, 1976.
Системы управления базами данных для ЕС ЭВМ. - М.: Финансы и статистика, 1984.
Филин Ф.П. К глубинам русского слова // Вестник АН СССР, 1980, № 7.
Cornforth M. Labour, language and thought. - In: Der Mensch-Subjekt und Objekt. Festschrift für Adam Schaff. Wien, 1973.
Hoffman L. Kommunikationsmittel Fachsprache. Berlin, 1976.
Meggle Georg. Grundbegriffe der Kommunikation. - In: De Gruyter, 1981 - XIV.
Meier G.F., Meier B. Handbuch der Linguistik und Kommunikationswissenschaft. B. 1,2. - Berlin: Akad.-Verl., 1979-1980.
Merten K. Kommunikation: Eine Begriffs- und Prozessanalyse. - Oplanden: Westdentschen Verl., 1977.
Steinmüller U. Kommunikationstheorie. Eine Einführung für Literatur- und Sprachwissenschaftler. - Stuttgart, 1977.

THE "MAN-MAN" AND "MAN-COMPUTER" COMMUNICATION: ITS ESSENCE AND SUPPORTING FACILITIES. CONCEPTUAL SIMULATION AS A LINGUISTIC PROBLEM

Rafail Yu. Kobrin

S u m m a r y

This paper deals with types of communication and with those of supporting semiotic systems. It is underlined that in the time of scientific and technical revolution there has become widely applied a man-computer communication the essence of which lies in providing the dialog interface between a man and computerized information systems.

The structural arrangement and functions of programming languages and languages for dialog information systems have been studied. The languages of dialog information systems are considered to implement the model of this or that subject area represented in terms of some natural language or to realize the mathematical problem being solved at the computer. These models created as a foundation for contemporary information systems, i.e. data banks have been called "conceptual" or "infological". The models contain complete information description of the objects and relations between them, and these models are defined by logical and mathematical or logical and linguistic structuring of the subject area.

Conceptual models represent the subject areas of science, engineering, control, social life of the society and are based upon conceptual and terminological systems of the related fields of human activities. The principles of conceptual simulation realizing the linguistic approach proposed are also considered.

РАНГОВЫЕ ПОЛИСЕМИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ ЛЕКСИКИ ТОЛКОВЫХ СЛОВАРЕЙ РУССКОГО И АНГЛИЙСКОГО ЯЗЫКОВ

А.В. Малов

Особое теоретическое значение в современном количественном языкознании наряду с варьированием языковых единиц по длине и частоте употребления приобретает исследование их типовой семантической вариативности. Существование такой вариативности, многозначности, отображает факт неравномерного распределения типовой семантической нагрузки по разным знаковым единицам, разный статус единиц данного уровня в их иерархии.

Наиболее важна и доступна для такого рода рассмотрения характеристика количественно-полисемической нагрузки единиц лексической системы. Совместное исследование этой характеристики с другими, формальными типами варьирования, надо ожидать, даст важный выход в широкую сферу базовых представлений об устройстве языка [Поликарпов, 1979; 1987].

Количественные исследования лексической полисемии должны основываться в первую очередь на данных одноязычных толковых словарей как наиболее представительных собраний общеязыковой лексики, служащих главным источником информации о многозначности слова. Несмотря на субъективизм, который может проявляться составителями словарей при отображении фактов лексической системы языка, на основании многочисленных данных можно сделать вывод, что, в основном, лексическая система получает в словаре вполне адекватное отображение. Это связано с тем, что языковая система, включая лексический уровень, имеет в языковом сознании своих носителей вполне объективное существование. В своей работе лексикограф руководствуется теми требованиями, которые предъявляются к словарю данного объема и типа в отношении состава словника и набора лексико-семантических вариантов у каждого слова. При этом последовательность подхода во многом обеспечивается лексикографической традицией, а также взаимной перепроверкой внутри авторского коллектива лексикографов. В результате деятельность составителей словаря оказывается достаточно строго детерминированной, а отраженный в словаре срез языка - достаточно представительным, чтобы по нему судить о языке в целом.

Ранжирование лексики толкового словаря по убыванию степени полисемичности дает ранговое полисемическое распределение, которое в целях наглядности удобно представлять в виде графика "степень полисемичности слова - его ранг" в билогарифмическом масштабе. Такие распределения были получены для нескольких толковых словарей русского и английского языков[+]. Анализ графиков позволяет сделать следующие наблюдения:

1. кривые, проведенные по правым концам "ступенек" /горизонтальных участков, соответствующих группам слов с одинаковым количеством значений/, имеют убывающий и, как правило, последовательный характер;

2. наклон кривых отражает относительную синтетичность - аналитичность языка и для словарей одного языка в целом постоянен;

3. кривые характеризуются выпуклостью, которая увеличивается с ростом объема словаря[++].

Под последовательным характером поведения кривых мы понимаем наличие выраженной тенденции, свойственной всем кривым. При определении наклона кривых локальные отклонения от общей тенденции не учитывались. Как и следовало ожидать, для близких по объему словарей русского и английского языков максимальная полисемичность оказывается выше в случае более аналитического английского языка. Это находит отражение в более крутом наклоне графика /см. рис.1/. Для удобства сравнения графиков между собой было решено рассматривать их в унифицированном виде, получаемом в результате выравнивания диапазонов "аргумента" и "функции" путем пропорционального ужатия оси ранга. При этом утрачивается возможность сравнивать графики по наклону, но облегчается анализ других особенностей. Было обнаружено, что

4. в унифицированном виде графики в целом симметричны относительно биссектрисы координатного угла /см. рис. 2/. Такая симметрия свидетельствует о выполнении в данном масштабе условия

$$y = f(x) <=> x = f(y) \qquad /1/$$

[+] Данные по русским словарям, а также по словарю "Шортер" предоставлены в наше распоряжение А.А. Поликарповым. С ним же обсуждались некоторые принципиальные вопросы анализа полисемических распределений, за что ему моя искренняя благодарность.

[++] См. также [Поликарпов, 1987; 1987а.]

Если выпуклость кривых можно проинтерпретировать как аналог "криволинейного" характера частотного рангового распределения лексических единиц в больших текстах [Алексеев, 1978], то сохранение симметричности при разных объемах словаря позволяет предположить, что в качестве некоего "нулевого" уровня выпуклости может выступать прямолинейность, описываемая формулой Ципфа

$$F_i \cdot i^{\gamma} = C \qquad /2/$$

и удовлетворяющая условию /1/. Эта особенность позволяет описать зависимость "степень полисемичности слова- его ранг" с помощью по крайней мере двух формул, сохраняющих генетическую связь с формулой Ципфа:

$$(P_i + A)(i^{\gamma} + A) = C; \qquad /3/$$

$$\lg P_i + \lg^{\alpha} i^{\gamma} = K, \qquad /4/$$

где P_i - степень полисемичности слова с рангом i,

A, α, γ, C, K - константы. В билогарифмическом масштабе за наклон графика отвечает γ, а за его выпуклость A и α. Формулы удовлетворяют условию /1/ и обращаются в формулу Ципфа при $A=0$ и $\alpha=1$.

В формуле /3/ используются две равные поправки, механизм действия которых такой же, как у поправки B в формуле Ципфа-Мандельброта [Тулдава, 1985]

$$F_i \cdot (i + B)^{\gamma} = C \qquad /5/$$

с тем отличием, что поправки одинаково действуют с обоих концов распределения, причем не локально, а по всей его длине, что обеспечивается несколько другим положением A в формуле /3/ по сравнению с формулой /5/. Генетическая связь формулы /3/ с законом Ципфа проявляется, в частности, в том, что в билинейном масштабе после выравнивания диапазонов графики, соответствующие этим двум формулам, выглядят как равносторонние гиперболы, сдвинутые относительно друг друга по вертикали и горизонтали на величину A.

Формула /4/, в отличие от /3/, обеспечивает геометрическое подобие унифицированных графиков в билогарифмическом масштабе при одинаковом α и различающихся других показателях. Она была сконструирована для проверки того, действительно ли возрастает выпуклость кривой при увеличении объема словаря, т.к. для отображения одинаковой выпуклости при различном объеме словаря формула /3/ в принципе может потребовать различных A. Однако показатели A и α, полученные на

реальных распределениях, никогда не противоречили друг другу, что свидетельствует о том, что различия между формулами /3/ и /4/ в данном случае не имеют большого значения.

Помимо полисемических распределений лексики, полученных на материале толковых словарей, рассматривались аналогичные распределения на текстах ряда художественных произведений[+] и т.н. политематическое распределение лексики в тезаурусе Роже[++]. Если ранговое полисемическое распределение лексики всего словаря Пушкина характеризуется "русским" наклоном графика, то наклон графиков для отдельных произведений Пушкина, Чехова и Горького более крутой /"завышена" максимальная полисемия/, что объясняется, по-видимому, другим соотношением в словарях такого рода между информативной и строевой лексикой. Для распределения лексики в тезаурусе Роже характерны значительно более высокие, чем у других словарей, показатели A и α. Поскольку тезаурус дает информацию не о том, какие именно лексико-семантические варианты есть у того или иного слова, а о том, в какие тематические группы это слово входит, то в данном случае более корректно нужно было бы говорить об отображении в тезаурусе не полисемии, а политематичности слова. Специфика тезауруса, организующей единицей которого является не слово, а тематическая группа, по-видимому, предопределила акцент на словах, входящих сразу в большое число тематических групп, что и привело к значительной выпуклости графика политематического распределения по сравнению с графиками полисемического распределения, полученными на материале толковых словарей. Несмотря на такое различие наклон графика в случае тезауруса такой же, как у графиков полисемического распределения лексики английского языка.

Обработка данных эмпирических распределений производилась на ЭВМ ЕС-1061. Подбор трех параметров аппроксимирующей формулы /см. табл.1/ осуществлялся методом сканирования. В качестве минимизируемого показателя близости эмпирического и теоретического распределения использовалась площадь фигуры между двумя соответствующими графиками в билогарифмическом масштабе.

[+] Данные предоставлены А.А. Поликарповым.

[++] Данные по: [Вишнякова, 1976]

Обе формулы хорошо аппроксимируют эмпирические данные, расхождения, как правило, не носят систематического характера. В связи с этим можно предположить, что данный вид зависимости имеет достаточно широкое распространение среди распределений рассматриваемого класса.

ЛИТЕРАТУРА

Алексеев П.М. О нелинейных формулировках закона Ципфа. – В кн.: Вопросы кибернетики, вып.41, М.; Л.,1978,с.53–65.

Вишнякова С.М. Опыт статистического исследования многозначности слов в английском языке. –В кн.: "Вычислительная лингвистика".-М., "Наука", 1976, с. 168-178.

Поликарпов А.А. Элементы теоретической социолингвистики. М., изд-во Московского ун-та, 1979.

Поликарпов А.А. Полисемия: системно-квантитативные аспекты. – Учен. зап. Тартуского ун-та, 1987, вып. 774, с. 135-154.

Поликарпов А.А. Полисемические спектры. – В сб.: Квантитативные аспекты системной организации текста /Материалы межвузовского семинара/. Тбилиси, 1987, с. 92-96.

Поликарпов А.А. Словарь и текст. 1987 /рукопись/.

Тулдава Ю.А. Частотная структура текста и закон Ципфа. – Учен. зап. Тартуского ун-та, 1985, вып. 711, с.93–116.

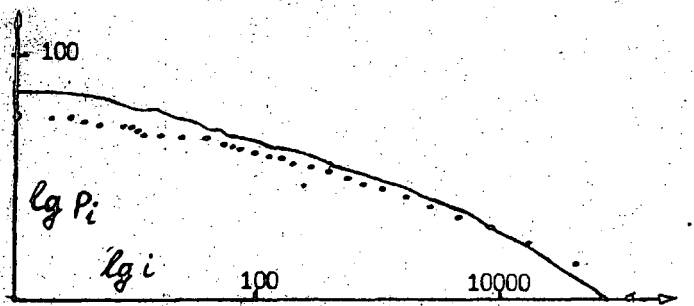


Рис.1 Графики полисемических ранговых распределений для 17-томного "Словаря современного русского литературного языка" /отдельные точки/ и словаря английского языка "Шортер" /сплошная линия/

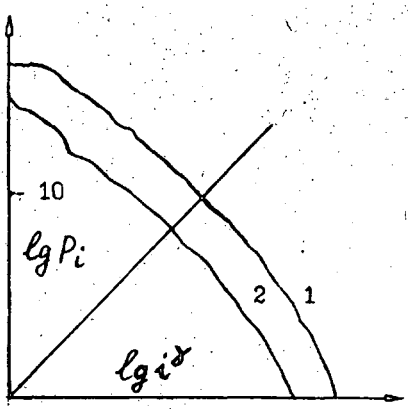


Рис.2 Унифицированные графики полисемических ранговых распределений для словаря "Шортер" /1/ и Малого Академического словаря /2/

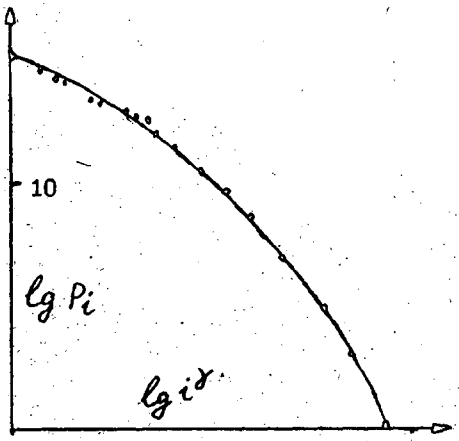


Рис.3 Унифицированный график рангового полисемического распределения для 17-томного "Словаря современного русского литературного языка" /отдельные точки/ в сопоставлении с графиком, полученным по аппроксимирующей формуле /сплошная линия/

Таблица 1

Показатели формулы /4/ для полисемических ранговых распределений лексики, представленной в словарях русского и английского языков

Словари	γ	α	Макс. полисемия	
			теоретич.	эмпирич.
СО	0,29	1,21	24	26
МАС	0,30	1,26	30	34
ССРЛЯ	0,30	1,36	33	33
Хорнби	0,33	1,21	31	30
Шортер	0,35	1,34	52	51
Роже	0,35	1,58	33	33

СО - "Словарь русского языка" С.И. Ожегова/1973-10-е изд./, МАС - "Словарь русского языка" под ред. А.П. Евгеньевой /1957-1961/, ССРЛЯ - "Словарь современного русского литературного языка" в семнадцати томах /1948-1965/, Hornby - A.S. Hornby "Oxford Advanced Learner's Dictionary of Current English" /1982/, Shorter - "Shorter Oxford English Dictionary" /1962/, Roget - Roget's International Thesaurus", 4th ed., 1975.

Таблица 2

Полисемические ранговые распределения лексики, представленной в словарях русского и английского языков в сопоставлении с данными аппроксимирующей формулы /4/

Словарь Ожегова						Тезаурус Роже					
p	$i_{эмп}$	$i_{теор}$	p	$i_{эмп}$	$i_{теор}$	p	$i_{эмп}$	$i_{теор}$	p	$i_{эмп}$	$i_{теор}$
1	57000	57457	14	17	18	1	25946	21807	18	66	71
2	12837	13400	15	11	14	2	13299	13258	19	51	56
3	4447	4386	16	8	10	3	7726	7651	20	41	44
4	1901	1835	17	–	8	4	4739	4694	21	37	35
5	941	894	18	6	6	5	3077	3035	22	28	28
6	515	483	19	–	4	6	2053	2044	23	20	22
7	284	281	20	3	3	7	1380	1421	24	19	18
8	168	173	21	–	2,5	8	956	1013	25	12	14
9	126	111	22	–	2	9	711	738	26	9	11
10	75	74	23	–	1,5	10	520	546	27	5	9
11	45	51	24	–	1	11	406	410	28	4	7
12	31	36	25	–	1	12	325	312	29	–	5
13	18	25	26	2	1	13	250	240	30	–	4
						14	190	185	31	–	3
						15	143	144	32	2	2
						16	110	113	33	1	1
						17	87	89			

p – степень полисемичности слова

$i_{эмп}$ – эмпирический ранг

$i_{теор}$ – теоретический ранг

POLYSEMIC RANK DISTRIBUTIONS OF RUSSIAN AND ENGLISH DICTIONARIES

Alexei V. Malov

S u m m a r y

The paper deals with problems of description of polysemic rank distributions. Bilogarithmic curves for several Russian and English monolingual dictionaries are considered and conclusions are drawn about how the typological differences in language systems affect the slope of the graph. The curve being convex, an attempt is made to adequately represent the rank-polysemy relationship in formulae which would allow an acceptable approximation of empirical data without over-complicating the original formula of Zipf's law.

КЛАССИФИКАЦИОННЫЕ ЗАДАЧИ СТИЛЕМЕТРИИ

Г.Я.Мартыненко, С.В.Чебанов

Стилеметрия – прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью упорядочивания и систематизации (атрибуции, диагностики, типологии, таксономии и т.п.) текстов и их частей.

Термин стилеметрия, по-видимому, был изобретен немецким филологом В.Диттенбергером, который в конце прошлого столетия решал задачу атрибуции и датировки диалогов Платона (Dittenberger W., 1880). Для этого он использовал частоты слов, реализация которых не зависит от тематики текста. В дальнейшем пионерские исследования Диттенбергера были продолжены В.Лютославским, Ф.Чадой, Ц.Риттером и другими авторами (Lutosławski W., 1887; Čáda F., 1901; Ritter C., 1903).

В России первым исследователем, применившем в целях атрибуции вероятностно-статистический метод, был Н.А.Морозов. Для отличения плагиата от подлинного произведения он использовал частотные распределения (спектры) служебных слов (Морозов Н.А., 1915).

К настоящему времени количественная атрибуция как филологическая задача стала хрестоматийной (Vašák P.,1980). Более того, даже в прикладной статистике эта задача фигурирует как типовая, иллюстрируя совместно с другими задачами широту приложений статистики как метода научного познания (Кокс Д., Снелл Э., 1982; Кимбл Г., 1984). В последние годы круг решаемых стилеметрией задач и репертуар применяемых ею методов существенно расширился. Практической повседневностью стала количественная таксономия текстов (Шайкевич А.Я.,1980; Тулдава Ю.А., 1981; Бородкин Л.И., 1983; Мартыненко Г.Я., 1983), относительно самостоятельное направление образовала квантитативная типология текста (Алексеев П.М., 1981), стилистическое приложение нашли дешифровочные модели (Тимофеева М.К., 1983), начала формироваться стилистическая диагностика (Севбо И.П., 1981) и некоторые другие направления.

Однако логический статус перечисленных задач, а также их соотношение практически не выявлены.

В самом общем виде можно говорить о том, что во всех

этих задачах мы имеем дело с различными совокупностями единиц и их систематизацией. Чаще всего эти задачи квалифицируются как классификационные. В данной работе показано, что это не совсем так и предпринята попытка адекватно квалифицировать упомянутые и другие способы стилеметрической работы.

В основу предлагаемого анализа положено различение собирательных и разделительных понятий и связанных с ними процедур членения и разбиения, а также выявление функционального назначения этих процедур в различных формах систематизации и упорядочивания.

Объект стилеметрии. Объектом стилеметрии является текст, но для нее это не текст вообще, для стилеметрии это всегда конкретный текст, т.е. текст, созданный конкретным автором, в конкретное время, в конкретной ситуации. При этом текст с формально-логической точки зрения мыслится как конкретное собирательное понятие, в котором отражены признаки группы предметов, образующих единое целое. Понятие "текст", впрочем, как и многие другие понятия, может быть и разделительным и собирательным, в зависимости от того, как оно рассматривается. Так "текст" есть собирательное понятие относительно входящих в него словоупотреблений, но он же есть разделительное понятие относительно различных видов текстов (научных, художественных, деловых и других текстов).

Как собирательное понятие, текст в терминах теории множеств может быть представлен как собирательное множество, а с точки зрения теории систем текст принадлежит к классу внутренних систем, являющихся целостными образованиями, к которым можно применять процедуры членения, представляя эти системы в виде некоторой структуры составляющих их частей (Шрейдер Ю.А., Шаров А.А., 1982).

С точки зрения статистики текст может рассматриваться как реальная совокупность. Эта категория была введена А.А.Чупровым применительно к различного рода реально существующим, а не создаваемым мыслью исследователя сообществам, представляющим собой целостное образование сосуществующих и взаимодействующих единиц, не обязательно однородных (Чупров А.А., 1910). Основным признаком таких совокупностей А.А.Чупров считает их устойчивость во времени: способность сохранять в течении более или менее длительного

времени свой состав и свои характерные черты. С этой точки зрения текст сверхустойчив: ни одно слово, ни одна фраза из текста удалена быть не может без нарушения его целостности или индивидуальности.

Собирательным понятиям в логике противопоставляются разделительные понятия, в математике — разделительные множества, а в теории систем — внешние системы. Соотношение между собирательными и разделительными категориями показано в таблице, которая приводится ниже.

Собирательные и разделительные категории

Категории	Собирательные	Разделительные
Понятие	собирательное	разделительное
Объем понятия	x)	класс референтов
Соотношение объема и содержания понятия	x)	обратное
Процедура выделения частей	членение	разбиение
Задача индукции по неполному основанию	реконструкция	экстраполяция
Характер составляющих	разнородные	однородные
Представление понятия в виде системы	внутренняя система	внешняя система
Характер системы	пространственно-временная организация	абстрактно-логическая конструкция
Совокупность составляющих	собирательное множество	разделительное множество
Модель	мереология Лесневского	множества Кантора
Тип статистической совокупности	реальная совокупность	искусственная совокупность
Филологические примеры	текст русская литература XIX века	словник список произведени

x) В классической логике не определен

Предмет стилеметрии. В типологических представлениях структура объекта описывается как архетип (Мейен С.В., Шрейдер Ю.А., 1976). В последнем можно вычленить мероны, которые осознаются как признаки, среди которых можно выделить существенные. Под существенными признаками в логике понимаются такие, после изъятия которых предмет перестает быть самим собой. Существенным признакам будут сопутствовать собственные и несобственные неотделимые признаки. Все они отражают обязательные, характерные мероны архетипа, логически (включая вероятностную логику) выводимые из существенных признаков. Кроме того, объект характеризуется и несобственными отделимыми признаками, которые практически независимы от существенных меронов и отвечают меронам, факультативно присутствующим в архетипе. Совокупность последних можно назвать периферией характеристики объекта. Тогда под элементами стиля будем понимать особенности периферии характеристики объекта. Другими словами, стиль может быть описан через факультативные признаки текста (в том числе количественные), которые не явным образом затрагивают его глубинные характеристики. Задаваясь разными уровнями достоверности, можно относить к периферии характеристики разные наборы признаков литературного произведения, так что к ней будут относиться в одном случае только характеристики языка, в другом - характеристики композиции и, наконец, при каком-то уровне достоверности в нее войдут и характеристики содержания. Такая позиция позволяет соотнести разные уровни стилевой организации с разными уровнями достоверности выводимости признаков из существенных. Тем самым можно оставить в стороне споры о том, что относится, а что не относится к стилю произведения, и, соответственно, снять вопрос о соотношении формы и содержания.

Более того, архетип можно интерпретировать как привилегированный стиль, при этом в качестве архетипа рассматривается новый объект - стилистическая организация, с которой связано особое членение литературного произведения или литературного процесса. Занимаясь количественной диагностикой, атрибуцией, типологией, классификацией, мы как раз и исследуем такие группировки, таксоны, архетипами которых являются стили.

Отталкиваясь от приведенной интерпретации понятия "стиль", можно более широко рассмотреть вопрос о художествен-

ной форме произведения как предмете эстетики. В этом случае уместно рассматривать художественную форму как своеобразную "ловушку" для смысла. При этом естественно раскрывается природа слова как строительного материала словесного произведения: слово обладает и определенными формальными характеристиками, и несет определенный смысл. В результате взаимодействия плана выражения и плана содержания формируются последовательности единиц, распределение которых в тексте и слагает то, что является предметом стилистики. Таким образом, мы уходим от противопоставления формы и содержания, и начинаем рассматривать форму, как нечто предполагающее смысл, как то, что является объектом эстетики в разных аспектах (смысловом и структурном), т.е. интерпретируем форму как нечто глубоко содержательное (Виноградов В.В., 1961).

Упорядочивающе-систематизирующая деятельность в стилеметрии. Всю совокупность основных видов стилеметрической работы можно обозначить как упорядочивающе-систематизирующую деятельность. Она распадается на две большие области: упорядочивание и систематизацию (Чебанов С.В., 1983 а).

Формы упорядочивания. Когда мы имеем дело с упорядочиванием, нас интересует только содержание соответствующих понятий, соотношение этих понятий друг с другом, а также вопрос о том, соотносима ли характеристика данного конкретного референта с содержанием соответствующего понятия. При этом мы можем отвлечься от того, что такой референт не единственен, а существует большее или меньшее множество аналогичных референтов. Такая деятельность, в свою очередь, многообразна и можно говорить о разных ее типах.

Во-первых, это будет задача диагносцирования. Занимаясь ею, мы должны перечислить существенные свойства того или иного объекта, обладание которыми позволит нам идентифицировать объект с рассматриваемым понятием. Для этого необходимо описать и охарактеризовать исследуемое явление, выявить закономерности, характеризующие его структуру. Часто сделать это строгими методами классической логики не удается, ввиду того, что рассматриваемое явление оказывается достаточно многообразным, и в нем практически невозможно выявить какие-нибудь устойчивые признаки, среди которых будут находиться и претендующие на то, чтобы быть существенными. В этом случае удобно характеризовать такое понятие некоторым диагностическо-

ким синдромом – совокупностью симптоматических признаков, которые характерны для рассматриваемого явления. Статистическим инструментом диагностики являются методы отбора диагностирующих признаков (регрессионный анализ, факторный анализ, метод корреляционных плеяд, метод главных компонент и др.). В качестве примера здесь можно привести работу Г.Ф. Мальцевой, в которой решается задача свертки признакового пространства с целью выделения в нем наиболее информативных стилистических признаков (Мальцева Г.Ф., 1969).

Важно обратить внимание на то, что при решении диагностических задач работа производится с признаками и их связями без учета (в первом приближении) того, какие именно объекты и их совокупности являются носителями конкретных сочетаний признаков. Таким образом, работа в данном случае является сугубо мерономической.

На основе решения диагностической задачи мы можем идентифицировать тот или иной конкретный объект, отождествить его с диагнозированным понятием. Пусть, например, мы установили систему признаков, характеризующих художественную прозу и нам предъявлен текст, ранее нам неизвестный. Мы начинаем исследовать вопрос о том, обладает ли данный текст тем же набором признаков. Ответив на этот вопрос, мы можем квалифицировать данный текст как относящийся или не относящийся к художественной прозе.

Возможна и другая ситуация. Допустим, мы имеем дело с некоторым важным явлением, но не можем его соотнести с каким-либо понятием, охарактеризовать его через набор явных признаков. При этом нам нужно взаимодействовать с этим явлением, исследовать его. Но сделать это будет трудно, если не дать этому явлению некоторое имя. Тогда отождествить это явление с тем же явлением, встретившимся повторно, можно путем использования того же имени, т.е. квалификация в данном случае осуществляется через именование.

Такой способ работы особенно важен при исследовании малоизвестных литературоведам или лингвистам явлений, когда исследователь выступает в качестве эксперта, скорее чувствующего то или иное явление, чем способного его явно описать и объяснить (Полани М., 1986). Вместе с тем результат такой экспертной оценки может быть предметом статистической обработки (Бешелев С.Д., Гурвич Ф.Г., 1973). К числу таких наи-

менований можно отнести, например, качественные характеристики стиля (выразительность, эмоциональность, напряженность и т.п.), с помощью которых экспертами оцениваются достоинства литературных произведений. Эти оценки могут сравниваться с объективными характеристиками стиля (Кэрролл Дж., 1972; Тулдава Ю.А., 1977).

Результаты диагностики и наименования часто используются не изолированно, а каким-то образом соотносятся друг с другом. Это находит свое отражение в построении шкал. Важно отметить при этом, что осуществляя шкалирование, мы также продолжаем работать не с конкретными объектами, совокупности которых являются разделительными множествами, а с содержаниями (интенсионалами) разных понятий. Начиная соотносить эти содержания друг с другом, мы в каком-то смысле обращаемся с ними как с индивидами, но способ существования таких индивидов - индивидов-архетипов совсем иной, чем способ существования индивидов-эмпирических референтов (Шрейдер Ю.А., 1984). В частности, в то время как эмпирических референтов может быть много, архетип существует в единственном "экземпляре". Следовательно, такая работа и здесь остается в пределах работы с собирательными понятиями.

Шкалы обычно разделяют на три типа: шкалу наименований, шкалу порядковую и шкалу количественную. Включая название в шкалу, давая объекту наименование, мы находим ему место в номинальной шкале. Ранжируя характеристики, фигурирующие в диагнозе, мы получаем порядковую шкалу, а переводя их в числовое выражение, мы получаем возможность измерять их и найти место объекта на количественной шкале. Важно отметить, что положение на шкале может быть зафиксировано для каждого отдельно взятого объекта. Например, взяв любое предложение из рассказов А.П.Чехова, мы можем найти место этого предложения на количественной шкале размера предложения.

Итак, различные виды упорядочивания - шкалирование, именование, диагностика позволяют нам по-разному квалифицировать (измерить, ранжировать, наименовать, идентифицировать, назвать) отдельно взятый объект, получив в результате то, что в дальнейшем понадобится для упорядочивания их совокупностей.

Формы систематизации. Рассмотрим теперь типы деятельности, в которых работа с собирательными понятиями переплетается тем или иным способом с работой с разделительными поня-

тиями. Здесь имеется в виду следующая ситуация: даже тогда, когда мы работаем, ориентируясь на результат, в основном формулируемый с помощью собирательных понятий, нужно все время иметь в виду и работу, сопряженную с разделительными понятиями. К рассматриваемой сфере деятельности можно отнести прежде всего некоторые обще-логические структуры. Таковыми будут структуры-иерархии, комбинативные (фасетные) структуры, структуры соположения составляющих. Эти структуры могут относиться как к собирательным, так и к разделительным понятиям.

Когда мы рассматриваем текст как собирательное понятие, то в нем можно выделить куски, соответствующие авторской речи и речи персонажей, и далее в каждом из полученных фрагментов выделить стилистически окрашенные и стилистически нейтральные фрагменты. Можно произвести то же самое расчленение в обратном порядке и получить аналогичное итоговое расчленение. Точно так же все произведения русской художественной литературы как разделительное понятие можно разбить на произведения, относящиеся к разным эпохам и разным жанрам, в любой последовательности. Эту форму систематизации обозначим как фрагментирование.

Следующий тип систематизации – морфологизация. В этом случае мы вычленяем и описываем обобщенное устройство (структуру) объектов некоторой группы (не интересуясь вообще говоря тем, как организованы сами эти группы). Описывать организацию можно по-разному. Так, если мы описываем организацию в открытых параметрах и реконструируем латентные характеристики как некоторые инварианты, то морфология выступает как структурализация; если же морфология описывается более богато, но менее конструктивно, то выделяются обобщенные типы организации – архетипы, исследованием которых занимается мерономия (Гете И.В., 1957; Мейен С.В., 1975). Обращаясь к собственно форме, понимаемой как "ловушка" смысла (см.выше), мы имеем дело с собственно морфологией, которая, правда, в большой мере лишается конструктивности. Занимаясь морфологией в том или ином смысле, мы выявляем совокупности частей и их взаимосвязей, присущих всем объектам данной группы.

Применительно к языкознанию структурализация может быть соотнесена, например, с грамматикой непосредственно состав-

лящих или с грамматикой зависимостей, мерономия – с традиционным синтаксисом, а морфология – с грамматическими учениями античных философов.

По характеру работы морфология сходна с упорядочиванием. Однако между ними есть и принципиальные различия. Так, при упорядочивании не принимается во внимание множество существующих референтов. Для морфологии это существенно. Кроме того, упорядочивание ориентировано на выделение жестких границ между значениями признаков, в то время как для морфологии актуален факт полиморфизма референтов. Это различие существенно, в частности, для выделения текстов как индивидов. В большинстве случаев, особенно при работе с современными текстами, эта задача для текстолога-эмпирика является тривиальной, однако при разыскании формальных критериев, позволяющих квалифицировать конкретную последовательность предложений как текст, возникают серьезные затруднения. Так, в лингвистике текста изучаются маркеры начала и конца текста, статистические методы используются для установления целостности текста, его отличия от фрагмента текста или конгломерата нескольких текстов. Например, Ю.К.Орлов, изучая роман Л.Н.Толстого "Война и мир" на предмет соответствия распределения словоформ закону Ципфа, приходит к выводу, что текст этого романа следует рассматривать как совокупность нескольких текстов, каждый из которых удовлетворяет условию ципфовости, в то время как текст всего романа этому условию не удовлетворяет (Орлов Ю.К., 1970). Выделение индивидов – важнейший вид работы по морфологизации и одновременно один из наиболее сложных компонентов предклассификационной деятельности. Принципиально, что при этом выделяется не один индивид, а сразу класс индивидов.*

Основным видом задач индукции по неполному основанию применительно к морфологии является задача реконструкции – выявление элементов и связей между ними, которые не даны исследователю в наблюдении.** Филологическим вариантом такой задачи являются дешифровочные алгоритмы, извлекающие

---

*Эта деятельность описывается той частью мереологии Лесневского, которая не охватывается мерономией Мейена-Шрейдера.

**При упорядочивании такой задачи не возникает.

грамматику из корпуса текстов. Результаты работы такого алгоритма могут быть использованы для стилистической диагностики текстов (Тимофеева М.К., 1983).

Особым случаем морфологизации является районирование. Занимаясь районированием, мы мерономизируем (расчленяем) пространство, причем в некоторых случаях (типологическое районирование) такое расчленение осуществляется на основании принципов, применимых к расчленению нескольких однотипных территорий. Специфика районирования как формы мерономизации заключается в том, что разные части формы мы привязываем к разным частям пространства, внеположенного, вообще говоря, этой форме (таким образом, привязку к карте при районировании можно сравнить с привязкой к часам в случае периодизации). Классической лингвистической задачей на районирование является выделение ареалов в лингвогеографии, в том числе с использованием статистических методов (Кузнецова Е.Л., 1981).

Как районирование в пространстве признаков можно рассматривать некоторые подходы к классифицированию. В этом случае классифицируемый объект мы представляем как набор значений признаков, задающих точку в многомерном фазовом пространстве признаков. Районы в пространстве признаков задают соответствующие таксоны. Дискриминантный и кластерный анализ как раз и занимаются выделением таких районов. Дистрибутивно-статистический метод решает задачу такого же типа (Шайкевич А.Я., 1976).

Особым весьма специфическим типом районирования является районирование во времени – периодизация и связанная с последней датировка. Очевидно, что периодизация также является формой работы с собирательными понятиями. Это своего рода временная морфологизация.

В связи с проблемами временного (исторического) расчленения ведущими задачами индукции по неполному основанию являются исторические реконструкции, причем ввиду направленности времени можно говорить о двух типах реконструкций: ретрогнозах и прогнозах. При решении этих задач существенно выявить, о каком времени идет речь. Так, поток речи существует в одном времени, и в этом случае перед нами может стоять задача восстановления предшествующих или предсказания последующих отрезков текста. В.В.Налимов в связи с этим, предлагает байесовский подход к интерпретации смысла последующего потен-

ционально многозначного текста: предшествующая ему часть является фильтром, наложение которого делает текст однозначно понимаемым (Налимов В.В., 1974). Другой смысл имеет время, в котором происходит эволюция индивидуального стиля. В этом случае ретрогноз и прогноз будут соотноситься с уже написанными произведениями и теми, которые еще предстоит написать данному автору.

Говоря о периодизации, важно отличать ее от временных классификаций. Это связано с тем, что произведения одного и того же автора, написанные в конкретный период, могут иметь сходство друг с другом и принадлежать к некоторому таксону. В этом случае мы рассматриваем совокупность произведений как разделительное понятие, в то время как соответствующий ему период творчества – понятие собирательное.

Следующий тип работы по систематизации – группирование. В этом случае основной задачей является разнесение каких-то объектов по группам, причем исследователя интересуют только такие группы, которые включают в себя объекты, сходные в каком-то нетривиальном отношении. Последнее утверждение равносильно тому, что с данной группой соотносится достаточно богатый архетип. Таким образом, задача группирования неотделима от задачи морфологизации и по характеру последней, с учетом особенностей самих групп, можно произвести различение разных типов группирования.

Прежде всего можно говорить о параметризации. В этом случае архетип задается фиксацией одного или нескольких параметров, и все объекты, удовлетворяющие фиксированному набору параметров, относятся к данному классу. Очевидно, что архетип в этом случае достаточно беден. Однако, работая с измеряемыми параметрами, можно в принципе разбить генеральную совокупность на систему непересекающихся классов. Примером такой работы будет разнесение словоупотреблений какого-то текста на классы по размеру (числу графем, слогов и т.п.).

Собственно классификация – это такой способ систематизации (группирования), когда класс выделяется на основе обладания объектами некоторыми предварительно выделенными существенными признаками. При этом богатство архетипа зависит от степени удачности выбора существенных признаков. Для того, чтобы охватить все множество исследуемых объектов, при классифицировании (в узком смысле) целесообразно следовать

закону единства основания деления понятий. Такой принцип строго выдерживается в исключительных случаях, например, в перечислительных документных классификациях иерархического типа. Для лингвистики и литературоведения более характерны типологии, являющиеся еще одним типом группирования. В основу выделения типов кладется выделение важных для конструкции целого аспектов организации и конструирования из них некоторых идеальных образов - типов как способов представления содержания конструируемого понятия. Сконструированный таким образом тип может реализоваться в относительно чистом виде, и тогда мы говорим о типических представителях, или же один объект может сочетать в себе свойства нескольких типов. В последнем случае, если свойства каждого типа выражены неявно, мы говорим, что имеем дело с переходными формами; в том же случае, когда в одном объекте выражены свойства нескольких типов, мы имеем дело со смешанным (сложным) типом. Очевидно, что чистые и сложные типы являются вариантами архетипа в собственно типологии. Спецификой типа в интенсиональном аспекте является то, что это смысловой, семантический инвариант, а не количественный (как в параметризации) или структурный (как в классификации). Кроме того, для типологии очень характерно то, что в интенсиональном аспекте мы рассматриваем не только тип, но и возможные его варианты, и в идеале, описывая тип, мы должны описать и совокупность его вариантов, а также пределы их варьирования. Однако это удается далеко не всегда.[и] Особенностью экстенсионального аспекта типологии, как следует из сказанного, является то, что группа, ассоциируемая с данным типом, принципиально неоднородна - в ней есть как характерные, типичные представители, так и периферические, в чем-то отличные от типа. Наконец, в ней могут быть элементы, которые вообще нельзя связать с каким-либо типом. Так, стиль романов Андрея Белого ("Петербург", "Серебряный голубь", "Котик Летаев" и др.) не может быть отнесен к какому-либо стилевому типу; стиль Белого в русской художествен -

---

[и] Так, акустические характеристики разных вариантов одной и той же фонемы в разных стилях произношения описывают пределы варьирования фонемы. Однако переход акустических характеристик через определенное пороговое значение превращает эти варианты в звук, репрезентирующий другую фонему.

ной прозе аналогов не имеет, хотя сам Андрей Белый любил сравнивать свой стиль со стилем Гоголя, Достоевского, Сологуба (Белый Андрей, 1934).

Итак, занимаясь типологией, мы в принципе не производим разбиения универсума объектов на классы, а получаем, скорее, частично пересекающиеся созвездия наряду с изолированными элементами. Примером типологизации может служить разделение совокупности текстов (или их авторов) на группы по критерию стилистической близости: большая часть текстов образует несколько многоэлементных групп, часть из которых пересекается, а некоторые тексты остаются изолированными (те, которые отличаются оригинальностью стиля) (Мартыненко Г.Я., 1983).

Для всех форм группирования основной задачей индукции по неполному основанию является задача таксономической (типологической) экстраполяции. Суть ее заключается в том, что результаты, полученные на основе изучения выборки класса, переносятся на весь класс, а результаты, измеренные на одной части таксономического универсума, переносятся на другие его части или на универсум в целом.

Взаимодействие форм упорядочивания и систематизации. Описанные выше формы упорядочивания и систематизации на практике никогда не встречаются в изолированном виде: существование каждой из них невозможно без существования других. Более того, порою даже нельзя отделить друг от друга операции разбиения и расчленения. Так мифологема раскола мирового яйца основана на "операции" раскалывания, которая одновременно является и расчленением, и разбиением, а яйцо, соответственно, выступает одновременно и как собирательное, и как разделительное понятие. Подобные ситуации связаны с символизацией (Чебанов С.В., 1983б).

Рассмотрим теперь несколько примеров взаимодействия форм упорядочивания и систематизации в стилеметрических исследованиях. Пусть мы имеем какую-нибудь удачную и содержательную типологию функциональных стилей. Изучив некоторое произведение, мы должны отнести стиль этого произведения к какому-то типу упомянутой типологии. Для этого мы должны располагать некоторой процедурой отождествления стиля данного произведения с одним из выделенных ранее типов стилей, т.е. осуществить идентификацию. Но идентификация, как говорилось выше, основана на диагностике как форме упорядочивания, т.е. идентификация связывает систематизацию и упорядочивание. В филологии такая за-

дача структурируют как атрибуция. Если мы имеем дело с группировкой стилей на основе некоторых измеримых параметров (т.е. занимаемся параметризацией), то атрибуция будет осуществляться с помощью измерения, а сами универсумы интенсионалов групп, выделяемых при параметризации, будут слагать шкалу. Такой вид работы относится к компетенции стилометрии.

Иллюстрацией другого типа работы будет квантитативная типология текста по П.М.Алексееву (Алексеев П.М., 1981). При таком способе работы практически последуются морфология текста определенного вида литературы с соответствующим набором канров. При этом морфология выступает в виде структурализации, причем особенности структуры (преимущественно лексико-семантической) описываются параметрами, что позволяет использовать методы квантитативной лингвистики. Таким образом, такой способ работы полностью относится к оперированию с собирательными понятиями. Однако для того, чтобы строить структуру абстрактного текста, приходится пользоваться некоторой типологией текстов, используя в качестве генеральной совокупности группу текстов одного типа (например, научно - технические, публицистические, эпистолярные и др. тексты). Таким образом, осуществление морфологизации без хотя бы тривиального группирования невозможно.

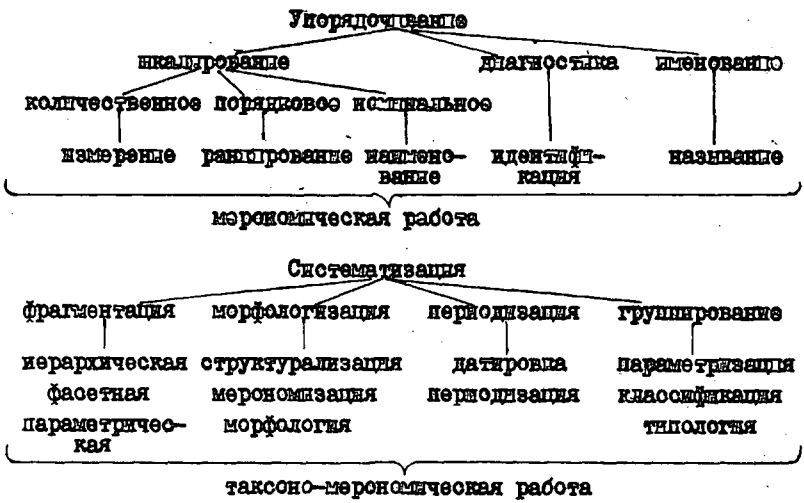
Производя подобную работу, П.М.Алексеев описывает количественные особенности устройства текста, т.е. характеризует мероны (например, лексические единицы) текста. При этом характеристики меронов получают через измерение количественного выражения, так что для них может быть найдено место на определенной количественной шкале. Результатом такой работы являются обобщенные образы (архетипы) текстов, представленные в виде усредненных структур текстов определенного класса. При таком способе работы фактически конструируются типы текстов. Это своего рода "типоСТРОЕНИЯ", а не типология. Подобным же образом можно заниматься и квантитативным "типостроением" (не типологией) стилей. Построив такие типы стилей, можно заниматься типологией стилей конкретных произведений. При этом мы по существу "перетипологизируем" ранее построенные типологии. Таким образом, квантитативная типология текста является одним из средств итеративной связи таксономии и мерономии (Мейен С.В., Шрейдер Ю.А., 1976). Здесь уместно обратить внимание на то, что в науке нет проблемы классификации, а есть

проблема переклассификации, улучшающей предшествующую. Квантитативная типология как раз и является одним из средств такой переклассификации.

Итак, у стилометрии есть область соприкосновения с квантитативной типологией текста в смысле П.М.Алексеева в области изучения мерономических структур.

В заключение обратим внимание на то, что определенным образом представив и проинтерпретировав материал, можно заменить один вид упорядочивающе-систематизирующей деятельности другим: классификацию можно представить как районирование в пространстве признаков, периодизацию как районирование во времени, систему классов можно представить как номинальную шкалу и тогда атрибуцию интерпретировать как наименование и т.д. Осуществляя подобного рода операции, нужно ясно отдавать себе отчет о том, какие операции мы осуществляем и каков их логический статус. Сказанное в полной мере относится и к стилометрическим задачам.

Основные виды упорядочивающе-систематизирующей работы, в том числе и те, которые относятся к компетенции стилометрии, показаны на двух схемах, которые приводятся ниже.



ЛИТЕРАТУРА

Алексеев П.М. О квантитативной типологии текста. – Уч.зап. Тартуск.ун-та, вып.591.Труды по лингвостатистике.Тарту: 1981, с.3-13.

Белый Андрей. Мастерство Гоголя. – М.-Л.:ОГИЗ, 1934.

Бешелев С.Д., Гурвич Ф.Г. Экспертные оценки.– М.:Наука,1973.

Бородкин Л.И. Математические методы классификации древних текстов. – В кн.: Методы количественного анализа текстов нарративных источников. – М.: Наука, 1983, с.8 – 30.

Виноградов В.В. Проблема авторства и теория стилей. – М.: Изд-во АН СССР, 1961.

Гете И.В. Избранные сочинения по естествознанию. – М.-Л: Изд-во АН СССР, 1957.

Кимбл Г. Как правильно пользоваться статистикой. – М.:"Статистика и финансы", 1982.

Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. – М.: "Мир", 1984.

Кузнецова Е.Д. Применение ЭВМ в лингвистической географии. Автореф. канд. дисс. – Л.: 1981.

Кэрролл Джон Б. Факторный анализ стилевых характеристик прозы. – В кн.: Семиотика и искусствометрия. – М.: "Мир", 1972, с.183-197.

Мартыненко Г.Я. Системно-статистический анализ синтаксических структур. – В кн.: Семиотика и информатика. – М.: ВИНИТИ, 1983, № 21, с.139-172.

Мальцева Г.Ф. Некоторые количественные приемы описания индивидуального авторского стиля. – В кн.: Статистика текста. Сборник статей. Том I.– Минск: Изд-во БГУ, 1969, с.206-248.

Мейен С.В. Систематика и формализация. – В кн.: Биология и современное научное познание. Ч.I. – М.: Изд-во Ин-та философии АН СССР, 1975, с.32-34.

Мейен С.В., Шрейдер Ю.А. Методологические аспекты теории классификации. – Вопросы философии, 1976, № 12, с.67-79.

Налимов В.В.Вероятностная модель языка. – М.:Наука, 1974.

Ордов Ю.К. О статистической структуре сообщений, оптимальных для человеческого восприятия. – Научно-техническая

информация. Серия 2, 1970, № 2, с.11-16.
Полани М. Личностное знание. – М.: Прогресс, 1986.
Себбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. – Киев: Наукова думка, 1981.
Тимофеева М.К. О методике индуктивной реконструкции грамматик. – Новосибирск: Ин-т математики СО АН СССР, 1983.
Тулдава Ю.А. К проблеме сопоставления субъективных и объективных характеристик стиля. – Уч.зап.Тартуск.ун-та, вып. 420. Studia metrica et poetica II, Tartu 1977. – С. 82-93.
Тулдава Ю.А. Опыт классификации текстов с помощью кластер-анализа. – Уч.зап.Тартуск.ун-та, вып.658. Труды по лингвостатистике. Тарту: 1981, с.136-157.
Чебанов С.В. Системный и комплексный подход к экономическим классификациям. – В кн.: Экономика и совершенствование управления на базе системного подхода. – Волгоград: Обл. совет НТО, 1983 а, с.98-101.
Чебанов С.В. Единство теоретизирования в способах упорядочивания. – В кн.: Теория и методология биологических классификаций. – М.: Наука, 1983 б, с.18-28.
Чупров А.А. Очерки по теории статистики. – С.-Петербург: Издание М.иС.Сабашниковых, 1910.
Шайкевич А.Я. Гипотезы о естественных классах и возможность количественной таксономии в лингвистике. – В кн.:Гипотеза в современной лингвистике. М.: Наука, 1980, с.319-357.
Шайкевич А.Я.Дистрибутивно-статистический метод в семантике. – В кн.: Принципы и методы семантических исследований. – М.: Наука, 1976, с.353-378.
Шрейдер Ю.А., Шаров А.А. Системы и модели. – М.:Наука, 1982.
Шрейдер Ю.А. Многоуровневость и системность реальности, изучаемой наукой. – В.кн.: Системогенез и эволюция. – М.: Наука, 1984, с.69-82.

Čáda F. Datování Platónova Faidra.– Listy filologické 28, 1901, 173-193, 342-359, 401-439.

Dittenberger W. Sprachliche Kriterien für die Chronologie der platonische Dialoge.– In: Hermes 16. Berlin, 1880, 321-345.

Lutosławski W. Theorie der Stylometrie auf die platonische Frage angewendet.– In: Zeitschrift für Philosophie und philologische Kritik. Leipzig, 1887, 110-121.

Ritter C. Die Sprachstatistika in Anwendung auf Platon und Goethe.- Neue Jahrbücher für das klassische Altertum. 1903, 241-261, 313-325.
Vašák P. Metody určování autorství.- Praha: Academia, 1980.

CLASSIFICATIONAL TASKS OF STYLOMETRICS

Grigory Ya. Martynenko, Sergei V. Chebanov

S u m m a r y

From the standpoint of general theory of classification the paper considers the forms of arrangement and systematization of texts and their parts (attribution, dating, periodization, taxonomy, typology and diagnostics) which stylometrics an applied philological discipline related to textology, stylistscs and quantitative linguistics deals with. In the context of collective and distributive categories the authors interpret the object and the subject of stylometrical analysis.

К ВОПРОСУ О СТАТИСТИЧЕСКОМ АНАЛИЗЕ СООТНОШЕНИЯ МНОГОЗНАЧНОСТИ И КОНТЕКСТУАЛЬНЫХ ПРИЗНАКОВ СЛОВ +

А.А. Поликарпов, О.В. Бунуева

I. Одним из основных компонентов ряда автоматизированных информационных систем (АИС) является блок разрешения лексической многозначности при обработке поступающих на их вход текстов. К числу таких систем относятся системы машинного перевода, автоматического контент-анализа, автоматического аннотирования и реферирования, автоматического индексирования документов, автоматического построения и ведения тезаурусов в ИПС, автоматизированные вопросно-ответные и интеллектуальные диалоговые системы. Без снятия этого рода неоднозначности функционирование подобных систем либо вообще теряет смысл, либо сопровождается весьма значительным уровнем шума.

Вследствие этого разработке блока разрешения лексической многозначности при построении современных АИС необходимо уделять весьма серьезное внимание. Причем, удельный вес этих проблем должен постоянно возрастать ввиду тенденции перехода все большего числа АИС к активному использованию на их входе текстов естественного языка в их обычном, полном, нередуцированном виде.

Вместе с тем, основные проблемы, встающие при этом, не выявлены в достаточном объеме и не оценены в должной степени, в том числе, и с количественной стороны. В частности, автоматизированные ИПС во многих случаях строятся без учета факта степени развитости многозначности ключевых для данной предметной области лексических единиц. Даже для систем МП, имеющих с самого начала своего развития дело с естественно-языковыми текстами во всей полноте их специфики, проблематика многозначности и ее разрешения не исследована в достаточной степени ни с качественной, ни с количественной стороны. Это связано со слабым еще теоретическим осознанием в лингвистике значимости явления лексической (и, в целом, языковой) многозначности в языке и с недостаточной до последнего времени экспериментальной проработкой вопроса о распространенности этого явления в разных языках и в отдельных их под-

---

+ Настоящая статья написана на основе доклада, прочитанного на международной конференции по научно-техническому переводу (2-4 декабря 1985 г., Москва) (Поликарпов, Бунуева, 1985).

системах, стилях, панрах.

Только исследования последних лет (Папп, 1969; Вишнякова, 1976; Поликарпов, 1976; 1979; 1981; 1987; 1988; Поликарпов, Каримова, 1988; Борода, Поликарпов, 1984; Крылов, Якубовская, 1977; Андрукович, Королев, 1977; Тулдава, 1979; 1987) позволили собрать достаточно представительные и надежные экспериментальные данные о представленности лексической многозначности в словаре и в текстах разных языков, о соотношении полисемической и омонимической многозначности с частотными и некоторыми другими характеристиками употребления слов.

2. Если вопрос о степени распространенности многозначности в словаре и текстах естественного языка за последнее время прояснился, то вопрос о каких-либо закономерностях системно-количественного плана для формальных контекстуальных характеристик, сопровождающих (а, точнее, детерминирующих) реализацию каждого определенного значения слова, об их закономерных качественных и количественных соотношениях с полисемическими, категориальными и прочими характеристиками слов практически еще не ставился. В частности, как для теории, так и для практики контекстуальной детерминации значений слов важно знать, каков для словаря данного языка общий круг формальных контекстуальных признаков (КП), набор их устойчивых комбинаций (контекстуальных детерминант - КД) при детерминации отдельных значений, каковы их качественные и количественные характеристики (в частности, разнообразие), степень продуктивности разных КП в образовании различных КД, зависимость качественных и количественных характеристик КП и КД от типологического статуса данного языка, широты или узости тематической сферы, для которой рассматривается вопрос о контекстуальной детерминации значений, зависимость качественных и количественных характеристик КП и КД слов от их категориальной (в том числе, частиречной) принадлежности, вхождения в ту или иную полисемическую категорию и т.п. Знание этих характеристик и зависимостей, по всей видимости, позволило бы развить и экспериментально обосновать теорию контекстуальной детерминации значений слов. +

+ В число проблем теоретического и экспериментального изучения контекстуальной детерминации лексической многозначности должно, видимо, входить и изучение связи семантического объема слова с такой характеристикой, как политекстия (Köhler, 1986) - мерой распределенности употреблений слова по выделенным отрезкам текста.

3. Наиболее серьезная трудность для экспериментального исследования указанных проблем - крайняя немногочисленность полных описаний действующих блоков разрешения лексической многозначности в тех или иных АИС. Известны лишь два достаточно информативных описания контекстуальных характеристик слов и правил разрешения их многозначности, полученные на материале английской лексики текстов смешанного-бытового и общественно-политического характера (Kelly and Stone, 1975) и текстов общественно-политической тематики (Марчук, 1976). В качестве объекта исследования нами был взят "The Disambiguation Dictionary", представленный в упомянутой работе (Kelly and Stone, 1975), как наиболее полный и многосторонний источник данных о частотах слов и их значений, продуктивности КП и КД при детерминации значений слов.

Словари подобного типа - пример аспектуального словаря, т.е. словаря, ориентированного на описание какого-либо одного из многочисленных аспектов семантики и употребления слов. Более того, контекстуальная детерминация в подобных словарях рассматривается лишь со стороны вероятностных корреляций между типовыми, узуальными смыслами (значениями) и формально-знаковыми характеристиками контекстов, в которых реализуются эти значения. Подобного рода специализированные словари - исходный материал создания синтетических машинных словарей (Поликарпов, 1979а) и для перехода в дальнейшем к действующим машинным моделям человеческого языка, речепорождения и речепонимания, способным имитировать реальную смысловую контекстность человеческого общения, реальную эвристику глубокого смыслового понимания. Изучение закономерных соотношений между типовыми смыслами и формально-контекстуальными характеристиками, сопровождающими их, - тоже один из этапов на пути к познанию полной картины закономерностей контекстуальной детерминации человеческого общения.

4. Представление в словаре значения слов и их контекстуальные признаки определялись на основе текстового массива объемом около 511 000 словоупотреблений. Та многозначность, которой обладают здесь слова, - это реализованная в данном массиве текстов многозначность слов (в среднем, ок. 3 зн.). Из 1790 слов словаря 580 - однозначные (т.е. это те слова, которые реализовали лишь одно значение). Остальные 1210 слов являются многозначными. Поскольку многозначные слова, как правило, и более частотны, чем малозначные, то средняя величина многозначности, которая должна быть разрешена любым по

текстовых словоупотреблений, равна, по крайней мере, 6,4 значения.+

Из 1210 многозначных слов правила разрешения их многозначности содержатся в словарных статьях 1014 слов, которые и явились для нас источником информации для анализа зависимостей между количественно-полисемическими и количественно-контекстуальными признаками слов.

При анализе контекстуальных признаков различались лексические и грамматические, левые и правые контекстуальные признаки, усредненные для всех слов данного словаря и для каждой полисемической зоны в отдельности, для слов всех частей речи вместе и для каждой части речи в отдельности.

5. Выявлены следующие факты.

5.1. Для слов любой части речи при разрешении лексической многозначности ведущая роль принадлежит лексическим контекстуальным признакам. Причем, левый контекст в целом (для всех частей речи вместе) количественно преобладает над правым (в среднем на 1 значение 1,75 КП слева и 1,39 КП справа).

5.2. Служебные слова (предлоги и союзы), знаменательные слова в целом и местоимения по контекстуальной детерминации разграничивались следующим образом:

- знаменательные слова на 1 значение требуют существенно меньшего среднего количества контекстуальных признаков, чем служебные слова (знам. - 4 КП; служ. - 5,4 КП);

- местоимения, которые выделялись в отдельную категорию, требуют в среднем на 1 значение еще большего числа КП (8,4);

- более дифференцированное рассмотрение знаменательных слов показывает, что на каждое из значений глагольных слов требуется в среднем большее число КП (6,5), чем для остальных слов (для существительного - 4,08; для прилагательного - 3; для наречия - 2). То есть, глагол в наибольшей степени приближается по этой, как и по многим другим характеристикам, к служебным частям речи. В том числе, как и у всех служебных слов (и у местоимения), у глагола преобладает правая контекстуальная детерминация в отличие от других знаменательных частей речи (существительного, прилагательного, наречия).

5.3. Важным для теории и практики является вопрос о ко-

+ Оговорка "по крайней мере" здесь связана с тем, что это не полная многозначность, а лишь та, что успела накопиться на данном текстовом материале.

личественном соотношении контекстуальных и полисемических характеристик слов. Чтобы проанализировать этот вопрос, нами был предпринят дифференцированный анализ по лексическим и грамматическим, левым и правым контекстуальным признакам, по общему количеству контекстуальных признаков и по величине разнообразия их на I значение для слов разных полисемических зон. При этом обнаруживаются две существенные тенденции, различающие слова разных полисемических зон.

Во-первых, для зоны статистической достоверности результатов, т.е. для слов, имеющих от 2-х до 9-ти значений, при движении от малозначных слов к многозначным обнаруживается тенденция к увеличению общего числа КП, приходящихся на I значение и необходимых для разрешения многозначности (например, если для 2-значных = 3,36 КП, то для 3-значных - 3,99 КП и т.д. вплоть до 6,23 КП - у 9-значных). Эта тенденция увеличения среднего количества КП при движении от малозначных к многозначным в равной степени свойственна как для левых, так и для правых контекстуальных окрушений слов. (см. таблицу I и рис. I).+

Эти данные указывают на то, что процедура разрешения многозначности для более многозначных слов нуждается в большем объеме контекстуальной информации на I значение, чем соответствующая процедура для менее многозначных слов. То есть, видимо, и при реальной коммуникации получателю для понимания более многозначных слов требуется бо́льшая поддержка контекстом, чем для понимания менее многозначных слов. Этот внутриязыковой вывод, полученный на нашем материале, хорошо соответствует ранее полученным данным в ходе межъязыковых сопоставлений английских и русских параллельных текстов (Поликарпов, 1976; 1979). Параллельные тексты английского языка (языка с более развитой лексической полисемией) оказывались более длинными в количестве словоупотреблений в сравнении с русскими, причем, в той же степени более длинными, чем более семантически неопределенными ++ были в среднем лексические единицы английского текста в срав-

+ В зоне самых многозначных слов, как кажется, появляется обратная тенденция, что требует еще своего истолкования. На данном материале ввиду его статистической недостаточности (см. табл. I) это не представляется возможным.

++ Семантическая неопределенность (H сем.) - величина, производная от количества значений у слова (V сем.). Связаны они следующим соотношением: H сем. = $\log V$ сем. (Поликарпов, 1976).

нении с русским.

С другой стороны, на нашем материале обнаруживается, что при движении от малозначных слов ко все более многозначным среднее число различных контекстуальных признаков приходящихся на I значение в слове, довольно устойчиво уменьшается (см. рис. 2). Это, видимо, связано с тем, что для детерминации значений более многозначных слов степень повторного использования одних и тех же контекстуальных признаков увеличивается в сравнении с менее многозначными словами.

Таблица I

Среднее количество контекстуальных признаков, приходящихся на I значение у слов данной полисемической группы

Кол-во знач.	КП независимо от позиции	КП в левой позиции от слова	КП в правой позиции от слова
24 /I сл./	3,38; лекс.- 2,83 грамм.- 0,54	I,25; лекс.- I,2I грамм.- 0,04	2,I3; лекс.-I,63 грамм.-0,50
I9 /I сл./	3,63; лекс.- 2,95 грамм.- 0,68	2,2I; лекс.- 2,2I грамм.- -	0,84; лекс.-0,74 грамм.-0,II
I7 /I сл./	6,53; лекс.- 3,7I грамм.- 2,82	0,I8; лекс.- 0,I2 грамм.- 0,06	6,I8; лекс.-3,59 грамм.-2,59
I5 /I сл./	2,20; лекс.- I,98 грамм.- 0,27	I,93; лекс.- I,87 грамм.- 0,07	0,20; лекс.-0,I3 грамм.-0,07
I4 /4 сл./	3,50; лекс.- 3,I3 грамм.- 0,38	I,04; лекс.- I,02 грамм.- 0,02	2,63; лекс.-2,50 грамм.-0,I3
I3 /4 сл./	4,58; лекс.- 3,67 грамм.- 0,90	2,7I; лекс.- 2,46 грамм.- 0,25	I,27; лекс.-I,23 грамм.-0,04
I2 /3 сл./	6,33; лекс.- 4,06 грамм.- 2,28	2,53; лекс.- 2,39 грамм.- 0,I4	2,22; лекс.-I,78 грамм.-0,44
II /5 сл./	5,98; лекс.- 4,56 грамм.- I,42	I,46; лекс.- I,29 грамм.- 0,I6	4,09; лекс.-3,42 грамм.-0,67
I0 /5 сл./	3,22; лекс.- 2,26 грамм.- 0,96	I,40; лекс.- I,34 грамм.- 0,06	I,I0; лекс.-I,04 грамм.-0,06
9 /I6 сл./	6,23; лекс.- 3,90 грамм.- 2,33	2,4I; лекс.- 2,20 грамм.- 0,2I	2,26; лекс.-2,00 грамм.-0,26
8 /22 сл./	5,54; лекс.- 3,43 грамм.- 2,II	2,I5; лекс.- I,94 грамм.- 0,2I	I,84; лекс.-I,59 грамм.-0,26
7 /30 сл./	4,I9; лекс.- 2,95 грамм.- I,24	I,99; лекс.- I,8I грамм.- 0,I9	I,5I; лекс.-I,29 грамм.-0,22
6 /86 сл./	4,38; лекс.- 2,79 грамм.- I,60	I,75; лекс.- I,57 грамм.- 0,I9	I,57; лекс.-I,35 грамм.-0,22
5 /II4 сл./	3,92; лекс.- 2,56 грамм.- I,36	I,7I; лекс.- I,49 грамм.- 0,22	I,39; лекс.-I,I7 грамм.-0,22
4 /I88 сл./	3,74; лекс.- 2,3I грамм.- I,43	I,69; лекс.- I,4I грамм.- 0,28	I,I9; лекс.-I,0I грамм.-0,I8
3 /269 сл./	3,99; лекс.- 2,34 грамм.- I,65	I,75; лекс.- I,53 грамм.- 0,22	I,06; лекс.-0,87 грамм.-0,I9
2 /264 сл./	3,36; лекс.- 2,08 грамм.- I,28	I,50; лекс.- I,30 грамм. -0,20	0,95; лекс.-0,82 грамм.-0,I3

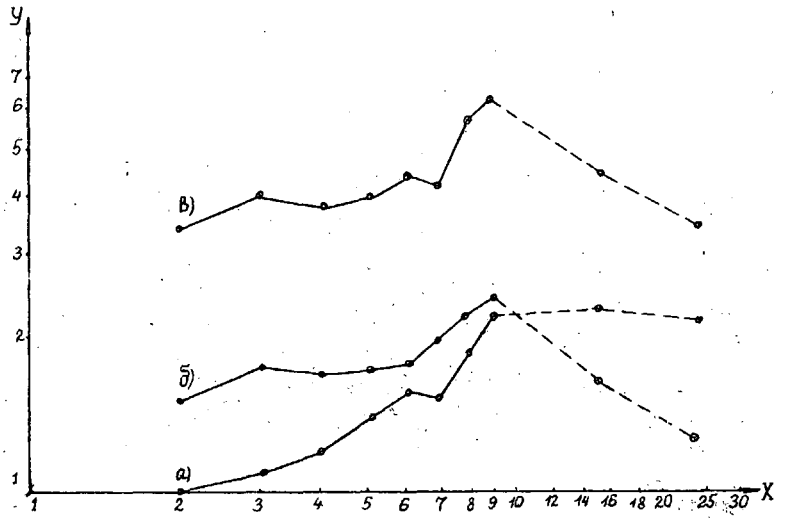


Рис. I. Зависимость среднего количества КП, приходящихся на I значение слов данной полисемической группы (У), от количества значений слов (Х).

а) КП в позиции справа от слова;
б) КП в позиции слева от слова;
в) вне зависимости КП в предложении.

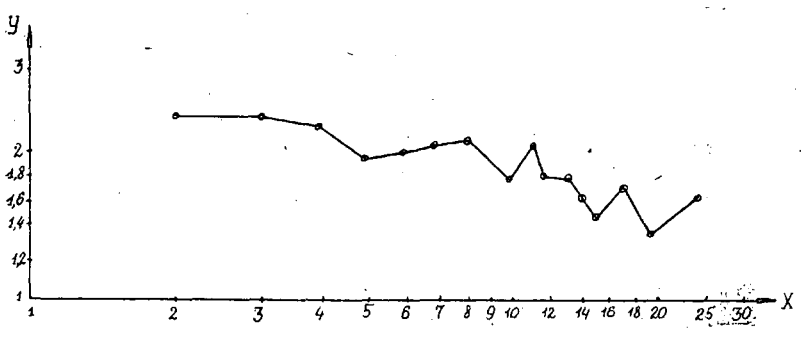


Рис. 2. Зависимость среднего количества различных КП, приходящихся на I значение слов данной полисемической группы (У), от количества значений слов (Х).

Л И Т Е Р А Т У Р А

Андрукович П.Ф., Королев Э.И. О статистических и лексико-грамматических свойствах слов // НТИ, сер. 2, 1977, № 4.

Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единицы различных уровней организации текста // Квантитативная лингвистика и автоматический анализ текстов (Ученые записки Тартуского ун-та, вып. 689). - Тарту, 1984.

Вишнякова С.М. Опыт статистического исследования многозначности слов // Вычислительная лингвистика. - М., 1976.

Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова // НТИ, сер. 2, 1977, № 3.

Марчук Ю.Н. (составитель). Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. - М.: ВЦП, 1976.

Папп Ф.О. О некоторых количественных характеристиках словарного состава языка // Slavica, т. УП. - Дебрецен, 1967.

Поликарпов А.А. Факторы и закономерности аналитизации языкового строя. Канд. дисс. - М., 1976.

Поликарпов А.А. Элементы теоретической социолингвистики. - М., 1979.

Поликарпов А.А. От банка словарей - к синтетическому машинному словарю // Международный семинар по машинному переводу. - М., 1979. - Тезисы докладов. М.: ВЦП, 1979а.

Поликарпов А.А. Квантитативная универсалия удельной семантической специфичности // Количественные методы в гуманитарных науках. - М.: МГУ, 1981.

Поликарпов А.А. Полисемия: системно-квантитативные аспекты // Квантитативная лингвистика и автоматический анализ текста. (УЗ Тартуского ун-та, вып. 774). - Тарту, 1987.

Поликарпов А.А. К теории жизненного цикла лексических единиц // Прикладная лингвистика и автоматический анализ текста: Тезисы докл. научн. конф. 28.01-30.01 1988. - Тарту, 1988.

Поликарпов А.А., Каримова Г.О. Словарь новых слов и значений (СНСЗ-1984) как источник данных об эволюции языка. - Там же.

Поликарпов А.А., Бушуева О.В. О закономерных системно-ко-

личественных соотношениях полисемических и контекстуальных признаков слов // Международная конференция "Теория и практика научно-технического перевода". - М., 1985. - Тезисы докладов. М.: ВЦП, 1985.

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии // Ученые записки Тартуского ун-та, вып. 502, (Linguistica, XI). - Тарту, 1979.

Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики. - Таллин: Валгус, 1987.

Kelly E.F., Stone Ph.J. Computer Recognition of English Word Senses. - Amsterdam; Oxford, 1975.

Köhler R. Zur linguistischen Synergetik. - Bohum: Brockmeyer, 1986.

STATISTICAL ANALYSIS OF THE RELATION BETWEEN POLYSEMY AND CONTEXTUAL FEATURES OF WORDS

A.A. Polikarpov, O.V. Busuyeva

S u m m a r y

Some problems of contextual determination of lexical polysemy from the quantitative point of view are discussed.

ОБ ИЗМЕРЕНИИ СВЯЗИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ В ЛИНГВИСТИКЕ (I): СОПРЯЖЕННОСТЬ АЛЬТЕРНАТИВНЫХ ПРИЗНАКОВ

Ю. А. Тулдава

Для выявления и измерения связи (зависимости) между качественными признаками объектов изучения во многих науках широко используются т.н. коэффициенты сопряженности (контингенции).[+] На возможность их применения в лингвистике указывалось в ряде работ более раннего периода развития квантитативной лингвистики (например, Herdan G., 1964; Супрун А.Е., 1964; Андрюшенко В.М., 1966; Шайкевич А.Я., 1969). В настоящее время имеются первые опыты практического применения коэффициентов сопряженности на обширном лингвистическом материале (работы Г.Г. Сильницкого, С.Н. Андреева, В.В. Левицкого, Б.И. Барткова и др.; обзор см. Андреев С.Н. и др., 1983, с. 110 и след.; см. также Tuldava J., 1975 и 1976). Однако до сих пор нет методических работ, в которых более подробно анализировались бы обоснованность, условия и ограничения применения коэффициентов сопряженности, в частности на лингвистическом материале. Настоящая серия статей призвана восполнить этот пробел. В первой статье рассматривается одна из разновидностей связей между признаками – сопряженность альтернативных признаков и возможности ее измерения и интерпретации.

Таблица 2 x 2. При анализе сопряженности альтернативных признаков мы ограничимся двумя переменными, которые обозначим A и B, и предположим, что они являются дихотомически расчлененными, т.е. принимают по два различных значения: A_1 и A_2 (или A и не-A) и B_1 и B_2 (или B и не-B) соответственно. Следовательно, имеется четыре возможных варианта их совместного появления: (A_1, B_1), (A_1,B_2), (A_2,B_1) и (A_2,B_2). Частоты этих вариантов обозначим через a, b, c, d соответственно.

[+] Иногда методы измерения сопряженности относят к методам корреляционного анализа. Это оправдано, по-видимому, в таких науках, как биология, социология, психология, лингвистика и т.д., в рамках которых изучается содержательный аспект корреляционных зависимостей и понятие "корреляция" трактуется как соотношение, взаимозависимость, связь в широком смысле слова (см. Методика и техника ..., 1968, с. 17). В более узком смысле корреляция определяется в математической статистике как линейная связь в условиях нормального распределения (в т.ч. частная и множественная корреляция), а также как ранговая корреляция и нелинейная корреляция, измеряемая т.н. корреляционным отношением.

На основании этих данных строится следующая четырехпольная таблица размером 2 x 2:

Признаки	B_I	B_2	$\sum_A$
A_I	a	b	a+b
A_2	c	d	c+d
$\sum_B$	a+c	b+d	n

где A_I и A_2 (B_I и B_2) - альтернативные признаки;
a, b, c, d - частоты сопоставляемых признаков;
$\sum_A$ и $\sum_B$ - маргинальные частоты (суммы);
a + b + c + d = n - общее число наблюдений.

Числа (частоты) в таблице сопряженности можно представить весьма разными способами. Например,

- мы рассматриваем n объектов и классифицируем их в зависимости от наличия и отсутствия у них признаков A и B; например, n глаголов и наличие и отсутствие семантического признака каузативности и словообразовательного признака производности;

мы рассматриваем n_1 объектов с признаком A_I и n_2 объектов с признаком A_2 (причем $n_1 + n_2 = n$) и классифицируем их в зависимости от их отношения к признаку B; например, глаголы и не-глаголы (т.е. остальные части речи) и какой-нибудь фонетический признак вроде односложность/неодносложность.

Для таблиц сопряженности существуют два принципиальных вопроса: какова сила связи между переменными (т.е. признаками A и B) и статистически значима ли эта связь.

Показатели связи. К настоящему времени известно множество различных показателей связей между качественными альтернативными признаками (см., например, Елисеева И.И., Рукавишников В.О., 1977; Миркин Б.Г., 1980; Аптон Г., 1982). Причина такого множества заключается в том, что различные меры (коэффициенты) меряют несколько различные аспекты связи. В настоящее время наиболее употребительными коэффициентами сопряженности альтернативных признаков являются коэффициент связи Юла (Q)+ и коэффициент взаимной сопряженности Бернулли (Φ = греч. "фи")++, которые вычисляются по следующим

+ См. Юл Дж.Э., Кендэл М.Дж., 1960, с. 58 и след.

++ Название "коэффициент Бернулли" встречается у Г.Хердана (Herdan G., 1966, с. 421). Известно, что К. Пирсон связал этот коэффициент с χ^2 по формуле $\Phi^2 = \chi^2/n$ (см. Елисеева И.И., Рукавишников В.О., 1977, с.87), но собственно "коэффициентом Пирсона" называется универсальный показатель взаимной сопряженности $C = \sqrt{\frac{\chi^2}{n + \chi^2}}$ (для таблиц размером $r \times s$).

формулам:

$$Q = \frac{ad - bc}{ad + bc}; \qquad (1)$$

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (2)$$

Как видно, формулы коэффициентов Q и ϕ походят друг на друга в том смысле, что числители у них полностью идентичны и лишь знаменатели несколько различны. Диапазон значений обоих коэффициентов простирается от -I до +I, причем крайние точки соответствуют полной связи (отрицательной или положительной), а 0 (нуль) означает отсутствие связи (т.е. независимость признаков А и В). Сходство Q и ϕ проявляется еще в том, что их значения не изменяются, если все частоты в таблице размером 2 x 2 умножить или разделить на одно и то же число. Поэтому Q и ϕ можно считать по процентам, вычисленным на базе общего числа наблюдений n.

Различие между коэффициентами Q и ϕ состоит в том, что они измеряют различные аспекты сопряженности в четырехпольной таблице: Q отражает наличие односторонней направленности связи (когда, например, признаку А приписывается воздействие на признак В, но не наоборот), в то время как ϕ позволяет отразить степень взаимосвязи между признаками в обоих направлениях (Методика и техника ..., I968, с. 233 и 237). Различие между коэффициентами может проявляться еще в том, что коэффициент Q оценивает только один аспект отношения между признаками: как наличие у объекта признака А способствует появлению у этого объекта признака В, а коэффициент ϕ оценивает сразу два аспекта, т.е. дополнительно еще то, как отсутствие у объекта признака А способствует отсутствию у этого объекта признака В (Миронов Б.Н., Степанов З.В., I975, с. I30). Во многих случаях между переменными наблюдается не зависимость типа причина - следствие, а взаимодействие и взаимовлияние. Соответственно проводится интерпретация, которая всецело зависит от конкретного материала, условий и задач исследования.

Сравнение значений Q и ϕ. Учитывая различие в конструкции коэффициентов, можно ожидать, что по одному массиву данных Q и ϕ имеют различные численные значения, причем всегда $Q \geqslant \phi$. Их значения совпадают тогда, когда существует полная д в у с т о р о н н я я взаимосвязь, т.е. ког-

да частоты в таблице 2 х 2 расположены в диагональных клетках, например:

	B	не-B	
A	30	0	30
не-A	0	70	70
	30	70	I00

($Q = \Phi = I$).

В данном случае каждый признак полностью "объясняет" другой, т.е. каждый объект наблюдения, обладающий признаком А, будет обладать также и признаком В, и наоборот.

С другой стороны, полная н е з а в и с и м о с т ь проявляется при равномерном распределении частот между признаками, и оба коэффициента получают нулевое значение, например:

	B	не-B	
A	30	30	60
не-A	20	20	40
	50	50	I00

($Q = \Phi = 0$).

Различие между значениями коэффициентов проявляется при всех других конфигурациях. Концентрация частот в любых трех клетках всегда свидетельствует о полной о д н о с т о р о н н е й связи ($Q = I$), в то время как двусторонняя связь может быть умеренной ($\Phi < I$), например:

	B	не-B	
A	30	0	30
не-A	20	50	70
	50	50	I00

($Q = I$; $\Phi = 0{,}65$).

Полная односторонняя связь означает здесь то, что все объекты, обладающие признаком А, обладают также и признаком В. Однако двусторонняя связь является неполной, так как в данном примере только часть объектов, обладающих признаком В, обладают и признаком А.

Приведем еще один пример для сравнения значений коэффициентов Q и Φ.

	B	не-B	
A	30	I0	40
не-A	20	40	60
	50	50	I00

($Q = 0{,}7I$; $\Phi = 0{,}4I$).

Последние примеры показывают, что при умеренной (или даже слабой) двусторонней связи односторонняя (однонаправленная) связь все же может быть значительной. Поэтому, в практической работе можно рассчитывать оба показателя Q и ϕ, хотя ϕ предпочтительнее вследствие большей теоретической обоснованности формулы (Методика и техника ..., 1968, с. 240). Коэффициент ϕ является по существу простой модификацией строгой в математическом отношении формулы парного коэффициента корреляции Пирсона-Бравэ

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}, \quad (3)$$

если в четырехпольной таблице обозначить наличие/отсутствие признака через I и 0. Приведем простой пример:

		(У) I	(У) 0	
(Х)	I	5	I	6
	0	2	2	4
		7	3	I0

Для большей наглядности вычисляем r_{xy} с помощью следующей вспомогательной таблицы:

№ пп.	x	y	x^2	y^2	xy
I.	I	I	I	I	I
2.	I	I	I	I	I
3.	I	I	I	I	I
4.	I	I	I	I	I
5.	I	I	I	I	I
6.	I	0	I	0	0
7.	0	I	0	I	0
8.	0	I	0	I	0
9.	0	0	0	0	0
I0.	0	0	0	0	0
n =I0	6	7	6	7	5

Следовательно, по формуле (3):

$$r_{xy} = \frac{I0 \cdot 5 - 6 \cdot 7}{\sqrt{(I0 \cdot 6 - 6^2) \cdot (I0 \cdot 7 - 7^2)}} = 0{,}356.$$

Точно такой же результат дает вычисление ϕ по формуле (2):

$$\phi = \frac{5 \cdot 2 - 2 \cdot 1}{\sqrt{7 \cdot 3 \cdot 6 \cdot 4}} = 0{,}356.$$

Рассмотренные коэффициенты сопряженности (корреляции) измеряют силу связи между переменными, но надо еще

проверить, являются ли конкретные численные значения коэффициентов статистически значимыми. Это можно сделать двояким путем: вычислением стандартного отклонения (и доверительных границ) или проверкой на значимость с помощью какого-нибудь критерия независимости, например, критерия "хи-квадрат".

Процедуру вычисления коэффициентов сопряженности и применения критериев значимости мы проиллюстрируем на конкретном примере из практики квантитативно-лингвистических исследований последнего времени.

Пример № I. В работе О.Ю. Головинской (I987)[+] исследуются взаимоотношения между разными аспектами словообразования, синтаксиса и семантики аффиксальных глаголов английского языка. В частности рассматривается взаимосвязь между способом образования аффиксальных глаголов и частью речи производящей основы (по данным Краткого Оксфордского словаря английского языка). В результате статистического исследования выявилось следующее перекрестное распределение частот аффиксальных глаголов (табл. Ia).[++]

Таблица Ia

Способ деривации \ Основа	Суб-стантив	Не-суб-стантив	Всего
Суффикс	498	440	938
Не-суффикс (префикс)	477	I007	I484
Всего	975	I447	2422

I) Вычисление коэффициента Q дает результат:

$$Q = \frac{498 \cdot I007 - 440 \cdot 477}{498 \cdot I007 + 440 \cdot 477} = 0{,}4I.$$

В предположении, что величина Q имеет нормальное распределение (при достаточно большом n), стандартное отклонение вычисляется по формуле

[+] Автор выражает искреннюю благодарность Ольге Юрьевне Головинской за любезное разрешение использовать фактический материал из ее исследования в качестве иллюстративного примера в данной работе.

[++] Небольшое число префиксально-суффиксальных глаголов причислено к группе суффиксальных глаголов. Всего их 48 (32 с субст. основой и I6 с не-субст. основой) (Головинская О.Н., I987, с. 54).

$$\sigma_Q = \frac{1-Q^2}{2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \quad (4)$$

Доверительные границы коэффициента Q определяются выражением $Q \pm k\sigma_Q$, где k зависит от выбранного уровня значимости (α), например:

α	k
0,05	I,96
0,045	2,0
0,0I	2,58
0,0027	3,0
0,00I	3,29

Статистическую значимость коэффициента Q можно считать установленной (на данном уровне значимости), если $|Q| > k\sigma_Q$.

В настоящем случае

$$\sigma_Q = \frac{I-0,4I^2}{2}\sqrt{\frac{I}{498}+\frac{I}{440}+\frac{I}{477}+\frac{I}{I007}} = 0,036.$$

Так как эмпирическое значение Q (0,4I) превосходит трехкратное стандартное отклонение ($3 \cdot 0,036$), то можно считать Q статистически значимой на уровне $\alpha < 0,0027$. Можно сделать вывод о достаточно сильной зависимости между способом образования аффиксального глагола и частью речи производящей основой (по условиям эксперимента - между суффиксацией и субстантивной основой). Связь положительная, т.е. суффиксация и субстантивные основы "тяготеют" друг к другу. Одновременно должна быть отрицательная связь между признаками "не-суффиксация" (т.е. префиксация) и "субстантивная основа". Это будет видно при перестановке признаков в таблице сопряженности (табл. Iб).

Таблица Iб

Способ деривации \ Основа	Субстантив	Не-субстантив основа	Всего
Не-суффикс (префикс)	477	I007	I484
Суффикс	498	440	938
Всего	975	I447	2422

По этой таблице $Q = -0,4I$.

2) Стандартное отклонение коэффициента Φ (в предположении о нормальности распределения) вычисляется по следующей формуле:

$$\sigma_\phi = \frac{1 - \phi^2}{\sqrt{n}}, \qquad (5)$$

где n - общее число наблюдений. Доверительные границы и статистическая значимость эмпирического ϕ определяются по аналогии с Q.

По данным таблицы (Ia) коэффициент ϕ, вычисляемый по формуле (2), получает значение

$$\phi = \frac{498 \cdot 1007 - 440 \cdot 477}{\sqrt{975 \cdot 1447 \cdot 938 \cdot 1484}} = 0{,}208.$$

Стандартное отклонение

$$\sigma_\phi = \frac{1 - 0{,}21^2}{\sqrt{2422}} = 0{,}019.$$

И здесь эмпирическое значение ϕ (0,208) намного превосходит трехкратное стандартное отклонение ($3 \cdot 0{,}019$), и статистическая значимость может считаться установленной на уровне значимости $\alpha < 0{,}0027$. Доверительные границы на этом уровне значимости определяются выражением

$$\phi \pm 3\sigma_\phi = 0{,}208 \pm 0{,}057 = 0{,}151 \ldots 0{,}265.$$

Коэффициент ϕ дает более умеренную оценку связи (взаимной зависимости) между признаками "способ образования" и "часть речи производящей основы", чем коэффициент Q. Это объясняется различием в конструкции и в содержательном смысле коэффициентов, о чем шла речь выше.

Так же как и при вычислении Q, при перестановке признаков в таблице 2 x 2 мы можем получить отрицательное значение ϕ, т.е. констатируется отрицательная корреляция между признаками "субстантивная основа" и "не-суффикс".

3) Другим способом проверки статистической значимости коэффициентов сопряженности является применение критерия "хи-квадрат" (χ^2). Если известно эмпирическое значение ϕ, то хи-квадрат вычисляется по формуле

$$\chi^2 = n \cdot \phi^2, \qquad (6)$$

где n - общее число наблюдений. Очевидно, что имеет место и обратное соотношение

$$\phi = \sqrt{\frac{\chi^2}{n}}.$$

При вычислении хи-квадрата надо иметь в виду, что число n берется в абсолютном значении (общее число наблюдений по данным конкретного опыта), а не из таблицы с процен-

тами (где n = 100). Например, по данным первого примера (где n = 2422 и ϕ = 0,208):

$$\chi^2 = 2422 \cdot 0{,}208^2 = 104{,}79.$$

Статистическая значимость по таблице 2 x 2 определяется теоретическими (критическими) значениями χ^2 при одной степени свободы на разных уровнях значимости, в частности

$$\chi^2_{0{,}05;1} = 3{,}84; \quad \chi^2_{0{,}01;1} = 6{,}64; \quad \chi^2_{0{,}001;1} = 10{,}83.$$

Если эмпирическое (наблюдаемое) значение χ^2 равно или больше теоретического значения (на данном уровне значимости), то зависимость между признаками A и B по таблице сопряженности можно считать статистически значимой. Поскольку эмпирическое значение (104,79) значительно превосходит теоретический 0,001 (0,1%) - уровень χ^2-распределения (10,83), то зависимость между признаками A и B можно считать статистически значимой.

4) Для вычисления хи-квадрата имеются еще другие возможности. Основной формулой, применяемой для любых таблиц размером r x s, является следующая:

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - m'_i)^2}{m'_i}, \qquad (7)$$

где m_i - наблюдаемая частота в ячейке i, m'_i - ожидаемая частота, k - число ячеек (полей). Ожидаемые частоты для каждой ячейки в таблице 2 x 2 вычисляются следующим образом:

$$a' = \frac{(a+c)(a+b)}{n}; \quad b' = \frac{(b+d)(a+b)}{n};$$

$$c' = \frac{(a+c)(c+d)}{n}; \quad d' = \frac{(b+d)(c+d)}{n}.$$

По данным примера № 1 (табл. 1а) можно составить следующую таблицу для вычисления хи-квадрата:

m_i	m'_i	$m_i - m'_i$	$\frac{(m_i - m'_i)^2}{m'_i}$
498	377,6	+120,4	38,39
440	560,4	-120,4	25,87
477	597,4	-120,4	24,27
1007	886,6	+120,4	16,35
			χ^2 = 104,88

Результат тот же, что и при вычислении по формуле (6). Малое различие (I04,79 и I04,88) объясняется ошибками округления.

В таблице величины m'_i указывают на наиболее вероятную частоту совместного появления признаков в ячейке i в предположении, что признаки А и В на самом деле независимы. Если А и В не независимы, а связаны тем или иным образом, то наблюдается различие между m_i и m'_i. Общую статистическую оценку этим различиям дает критерий хи-квадрат.

5) К таблице сопряженности размером 2 x 2 можно применить и следующую формулу для вычисления хи-квадрата:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} . \qquad (8)$$

По содержанию эта формула в точности совпадает с формулой (6). Она удобна в том смысле, что ее легко можно модифицировать для применения с поправкой Йэйтса (Yates F., I934; ср. Антон Г., I982, с. 20):

$$\chi^{2*} = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} . \qquad (9)$$

Дело в том, что критерий хи-квадрат, строго говоря, подходит к измерениям на непрерывной шкале, а в таблице сопряженности мы имеем дело с дискретными величинами. Применение хи-квадрата к дискретным величинам приводит к некоторой неточности вычислений, а модификация формулы, по мнению Йэйтса, дает результаты, лучше согласующиеся с распределением χ^2.

Значения хи-квадрата, вычисленные по формуле с поправкой Йэйтса, всегда немного меньше, чем значения критерия без поправки. По данным нашего примера $\chi^{2*} = I04,0$ (без поправки $\chi^2 = I04,9$). Здесь разница ничтожна, однако при малых значениях χ^2 применение поправки может повлиять на оценку статистической значимости.

Точный критерий Фишера. Если ожидаемые частоты в ячейках таблицы 2 x 2 оказываются совсем малыми (5 или меньше), то эффективность критерия хи-квадрат уменьшается. Для более точной оценки статистической значимости можно применить "точный критерий Фишера" (Fisher R.A., I934; ср. Антон Г.,I982, с. 2I), который вообще требуется применить к малым выборкам (если общее число наблюдений меньше 50, т.е. $n < 50$). Для таблицы

		B	
A	a	b	n_1
	c	d	n_2
	n_3	n_4	n

критерий Фишера (P) вычисляется по формуле

$$P = \frac{n_1!\ n_2!\ n_3!\ n_4!}{a!\ b!\ c!\ d!\ n!}, \qquad (10)$$

где $m!$ означает m-факториал (причем $0! = 1$). Величины n_i представляют собой маргинальные частоты $(n_1 + n_2 + n_3 + n_4 = n)$ в таблице 2 x 2.+

Критерий P дает нам вероятность (уровень значимости) в предположении о независимости переменных. Например, если $P > 0{,}05$, то нулевая гипотеза о независимости принимается, если же $P \leq 0{,}05$, то отвергаем нулевую гипотезу о независимости на уровне значимости 0,05 и делаем вывод о зависимости (связи) рассматриваемых признаков A и B. Однако надо учесть, что точный критерий Фишера – это о д н о с т о р о н н и й т е с т (Аптон Г., 1982, с. 22). Поэтому для получения двусторонней вероятности приходится удвоить одностороннюю вероятность. Сравним два примера:

5	I	6
4	8	I2
9	9	I8

7	I	8
4	8	I2
II	9	20

Для первого примера получаем

$$P = \frac{6!\ I2!\ 9!\ 9!}{5!\ I!\ 4!\ 8!\ I8!} = 0{,}06I.$$

Так как $P = 0{,}06I > P_{теор} = 0{,}05$, то нуль-гипотеза о независимости не отвергается на уровне значимости $\alpha_1 > 0{,}05$ при одностороннем тесте и на уровне значимости $\alpha_1 > 0{,}I0$ при двустороннем тесте.

Для второго примера вычисление дает результатом $P = 0{,}024$, т.е. нуль-гипотеза отвергается на уровне значимости 0,024 при одностороннем и на уровне значимости 0,048 при двустороннем тесте. Зависимость между переменными (признака-

---

+ Для удобства вычислений можно обратиться к таблицам логарифмов факториалов. Формула (I0) принимает тогда вид:

$$P = \exp\{[\ln(n_1!) + \ln(n_2!) + \ln(n_3!) + \ln(n_4!)] - [\ln(a!) + \ln(b!) + \ln(c!) + \ln(d!) + \ln(n!)]\}.$$

ми) можно считать статистически значимой ($\alpha_2 < 0{,}05$ при двустороннем тесте).

Нормированный коэффициент Φ. В практической работе иногда появляется надобность сравнения численных значений коэффициента взаимной сопряженности Φ по данным из разных экспериментов. Прямое сравнение значений коэффициента было бы некорректным, так как в силу особенностей конструкции формулы, коэффициент Φ может иметь разные экстремальные значения в зависимости от структуры таблицы 2 х 2. Крайние значения (+I или -I) достигаются только в том случае, если в таблице размером 2 х 2 два диагональных поля заняты нулями, т.е. если $a = d = 0$ или $b = c = 0$. В других случаях экстремальные значения для конкретной таблицы оказываются меньше единицы. Для выявления этих конкретных экстремальных значений - Φ_{max} и Φ_{min} - существует определенная методика (см. ClauB G., Ebner H., I970, с. 255). Для их вычисления можно применять и готовые формулы:

$$\Phi_{max} = \frac{n \cdot min(b,c) + (ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad \text{(II)}$$

если $\Phi \geqslant 0$ (т.е. $ad \geqslant bc$);

$$\Phi_{min} = \frac{n \cdot min(a,d) - (ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad \text{(I2)}$$

если $\Phi < 0$ (т.е. $ad < bc$).

Экстремальные значения выполняют функцию нормирования (корригирования) коэффициента Φ таким образом, что значение Φ, полученное в исследовании, выражается через отношение к максимально (минимально) возможному значению:

$$\Phi_{норм} = \frac{\Phi}{\Phi_{max}} \quad \text{если } \Phi > 0, \quad \text{(I3)}$$

$$\Phi_{норм} = \frac{\Phi}{|\Phi_{min}|} \quad \text{если } \Phi < 0. \quad \text{(I4)}$$

По данным нашего примера (табл. Ia) $\Phi = 0{,}208$ (т.е. $\Phi > 0$) и, следовательно, надо применять формулу (II):

$$\Phi_{max} = \frac{2422 \cdot 440 + (498 \cdot I007 - 440 \cdot 477)}{\sqrt{975 \cdot I447 \cdot 938 \cdot I484}} = 0{,}969$$

и формулу (I3):

$$\phi_{\text{норм}} = \frac{0{,}208}{0{,}969} = 0{,}215.$$

Если бы значение ϕ было отрицательным, например, $\phi = -0{,}208$ (по табл. Iб), то вычисляем по формуле (I2):

$$|\phi_{min}| = \frac{2422 \cdot 440 - (477 \cdot 440 - 1007 \cdot 498)}{\sqrt{975 \cdot 1447 \cdot 1484 \cdot 938}} = 0{,}969$$

и по формуле (I4):

$$\phi_{\text{норм}} = \frac{-0{,}208}{0{,}969} = -0{,}215.$$

Нормированные значения ϕ можно также вычислить прямо по следующим формулам Коула (Cole L.C., 1949):

$$\phi_{\text{норм}} = \frac{ad - bc}{n \cdot min(b,c) + (ad - bc)}, \quad \text{(I5)}$$

если $\phi \geq 0$ (т.е. $ad \geq bc$);

$$\phi_{\text{норм}} = \frac{ad - bc}{n \cdot min(a,d) - (ad - bc)}, \quad \text{(I6)}$$

если $\phi < 0$ (т.е. $ad < bc$).

Например, по данным примера (табл. Iа) вычисляем по формуле (I5):

$$\phi_{\text{норм}} = \frac{498 \cdot 1007 - 440 \cdot 477}{2422 \cdot 440 + (498 \cdot 1007 - 440 \cdot 477)} = 0{,}215,$$

что в точности совпадает с результатом вычисления по формуле (I3).

Для того, чтобы наглядно продемонстрировать важность нормирования ϕ при сравнении результатов из разных опытов, приведем следующий пример:

Опыт I

60	I8	78
I2	I0	22
72	28	I00

$\phi = 0{,}207$;

$\phi_{max} = 0{,}852$;

$\phi_{\text{норм}} = \frac{0{,}207}{0{,}852} = 0{,}243.$

Опыт II

60	I0	70
2I	9	30
8I	I9	I00

$\phi = 0{,}I84$;

$\phi_{max} = 0{,}739$;

$\phi_{\text{норм}} = \frac{0{,}I84}{0{,}739} = 0{,}249.$

На основе значений ненормированных ϕ может показаться, что связь признаков в первом опыте сильнее (0,207), чем во втором опыте (0,184). Однако в результате нормирования коэффициента оказывается, что в действительности нельзя говорить о существенном различии между результатами (нормированное значение ϕ по данным второго опыта получается даже немного выше $\sim$ 0,25 против 0,24).

Коэффициент ϕ как мера близости. Вполне возможно использовать коэффициент сопряженности ϕ не по прямому назначению (т.е. для выявления внутренней взаимосвязи, взаимозависимости, взаимообусловленности признаков), а для измерения "связи" в смысле внешней близости или сходства объектов. Такой пример можно найти, например, у Г. Хердана, который сравнивает лексический состав публицистических текстов шести авторов (Herdan G., 1966, с. 151 и след.). При попарном сравнении авторов устанавливается общая часть лексики для каждой пары авторов (ячейка a в таблице 2 x 2), специфичная для данного автора лексика (ячейки b и c) и лексика, которая не встречается у сравниваемых двух авторов, но встречается у других авторов группы (ячейка d). Общее число наблюдений (n) указывает на "общий словарь" данной группы авторов (текстов).

Например, перекрестное распределение лексики в двух текстах А и В (авторы: Черчилль и Галифакс) представлено в следующей форме (Herdan G., 1966, с. 156):

	В	не-В	Всего
А	197	296	493
	(a)	(b)	(a+b)
не-А	259	775	1034
	(c)	(d)	(c+d)
Всего	456	1071	1527
	(a+c)	(b+d)	(n)

Таким образом, общая лексика двух текстов составляет 197 единиц (слов), в тексте А встречаются 296 слов, которые отсутствуют в тексте В, а в тексте В - 259 слов, которых нет в тексте А. Число 775 указывает на те слова, которые не встречаются ни у А, ни у В, но встречаются в каком-нибудь из остальных четырех текстов. Общий словарь всех шести текстов составляет 1527 слов. Кроме того, из таблицы видно, что объем словаря текста А - 493 слова, объем словаря текста В - 456 слов (объемы всех шести текстов были приблизительно равны -

около 1200 словоупотреблений).

На основе приведенной таблицы 2 х 2 вычисляется коэффициент ϕ по формуле (2).

$$\phi = \frac{197 \cdot 775 - 296 \cdot 259}{\sqrt{456 \cdot 1071 \cdot 493 \cdot 1034}} = 0{,}152.$$

Таким же образом сравниваются все пары текстов, и на основе меры ϕ оценивается попарная близость ("лексическая связь") шести текстов между собой.

При сравнении и измерении близости объектов с помощью формулы ϕ выдерживается принцип системности в отношении данной конечной совокупности объектов. В приведенном примере лексика двух сравниваемых словарей соотносится с лексиками всех других словарей рассматриваемой группы текстов. Это полезный прием при изучении лингвистических объектов, относящихся к определенной предполагаемой (или постулируемой) закрытой системе. Однако, если мы рассматриваем изучаемую совокупность как открытую систему и мы не можем определить ни величины n, ни величины d по таблице 2 х 2, то приходится как-то модифицировать формулу ϕ. Удобно представить эти величины асимптотически бесконечными, например, модифицируя формулу следующим образом:+

$$\phi^* = \lim_{d \to \infty} \phi = \lim_{d \to \infty} \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{a}{\sqrt{(a+b)(a+c)}}. \quad (17)$$

Как известно, такую формулу вывел и использовал А. Эллегорд для измерения близости генетически родственных языков, причем a означало количество совпадающих корней слов в двух сравниваемых языках (Ellegård A., 1959). Эту же формулу в виде $R = \frac{C}{\sqrt{A \cdot B}}$ мы рекомендовали для измерения лексической близости словарей и текстов в синхронии и диахронии (Тулдава Ю.А., 1974; 1987, с. 157-158). По своему содержанию эта формула выражает отношение общей части C к геометрической средней объемов A и B сравниваемых двух словарей. На примере Г. Хердана вычисление дает:

+ В действительности мы предполагаем, что величины n и d являются достаточно большими. Например, в данном случае (в примере Г. Хердана) уже при $d = 50\,000$ показатель ϕ^* вплотную приближается к показателю ϕ (их значения 0,410 и 0,415, соответственно).

$$\phi^* = R = \frac{197}{\sqrt{456 \cdot 493}} = 0{,}4155.$$

Применение коэффициента R теоретически обосновано при попарном сравнении словарей особенно в тех случаях, когда объемы сравниваемых словарей сильно отличаются друг от друга (чтобы уменьшить излишнее влияние более крупного словаря). Если же объемы словарей равны или близки друг другу, то можно использовать более простую формулу - $R' = \frac{2C}{A+B}$, где вычисляется отношение общей части (C) к арифметической средней объемов A и B. При равных объемах словарей, т. е. когда $A = B$, формулы R и R' дают одинаковые результаты ($R = R'$). В вышеприведенном примере объемы словарей близки друг к другу, и значения коэффициентов R и R' практически совпадают:

$$R' = \frac{2 \cdot 197}{456 + 493} = 0{,}4152.$$

Оба коэффициента, R и R', изменяются в пределах от 0 до +1.

ЛИТЕРАТУРА

Андреев С.Н. Исследование языковой системы при помощи ЭВМ (на материале деривационных и морфемных классов английских глаголов). - Смоленск: Изд-во Смол. пединститута, 1987. - 88 с.

Андреев С.Н., Кузьмин Л.А., Луговской В.П., Сильницкий Г.Г. Исследование связи деривационных и синтаксических характеристик английских глаголов методом корреляционного анализа // Исследование деривационной подсистемы количественным методом. - Владивосток: ДВНЦ АН СССР, 1983. - С. 108-124.

Андрющенко В.М. Некоторые проблемы анализа лингвистических данных // Иностр. яз. в школе, 1966, № 3. - С. 76-84.

Аптон Г. Анализ таблиц сопряженности. / Перев. с англ. и предисловие Ю.П. Адлера. - М.: Финансы и статистика, 1982. - 144 с.

Бартков Б.И. Корреляционный анализ в дериватологии // Дериватология и дериватография литературной нормы и научного стиля. - Владивосток: ДВНЦ АН СССР, 1984. - С. 3-27.

Головинская О.Ю. Производные аффиксальные глаголы в современном английском языке: Дис. ... канд. филол. наук. Минск, 1987. - 180 с.

Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. - М.: Статистика, 1977. - 144 с.

Левицкий В.В., Быстрова Л.В., Гинка Б.И. Изучение лексической сочетаемости с помощью статистических методов // Функционирование языковых единиц в речи и в тексте. - Воронеж: Изд-во Воронежского ун-та, 1987. - С. 48-59.

Методика и техника статистической обработки первичной социологической информации. - М.: Наука, 1968. - 327 с.
Миркин Б.Г. Анализ качественных признаков и структур. - М.: Статистика, 1980.
Миронов Б.Н., Степанов З.В. Историк и математика. - Л.: Наука, 1975. - 184 с.
Ольницкий Г.Г., Андреев С.Н., Кузьмин Л.А., Луговской В.П. О некоторых количественных методах определения мер связи между языковыми уровнями // Исследование деривационной подсистемы количественным методом. - Владивосток: ДВНЦ АН СССР, 1983. - С. 40-61.
Супрун А.Е. Заметки о частях речи // Вопросы лексики и грамматики русского языка. Вып. 2. - Фрунзе, 1964.
Тулдава Ю.А. Об измерении лексической связи текстов на уровне словаре // Вопросы статистической стилистики. Киев: Наукова думка, 1974. - С. 35-42.
Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики. - Таллин: Валгус, 1987. - 204 с.
Шайкевич А.Я. Корреляционный анализ в лингвостатистике, и понятие интервала текста // Вопросы лингвостатистики и автоматизация лингвостатистических работ. Вып. 2. - М.: ЦНИИПИ, 1969. - С. 254-271.
Юл Дж.Э., Кендэл М.Дж. Теория статистики. 14-е изд. / Перев. с англ. - М.: Госстатиздат ЦСУ СССР, 1960. - 779 с.
ClauB G., Ebner H. Grundlagen der Statistik fur Psychologen, Pädagogen und Soziologen. - Berlin: Volk und Wissen, 1970.
Cole L.C. The Measurement of Interspecific Association // Ecology, vol. 30, 1949.
Ellegård A. Statistical Measurement of Linguistic Relationship // Language, vol. 35, 1959. - P. 131-156.
Fisher R.A. Statistical Methods for Research Workers. - London: Oliver & Boyd, 1934.
Herdan G. Quantitative Linguistics. - London: Butterworths, 1964.
Herdan G. The Advanced Theory of Language as Choice and Chance. - Berlin; Heidelberg; New York, 1966.
Langley R. Practical Statistics for Non-Mathematical People. - London: Pan Books Ltd., 1968.
Tuldava J. Statistiline sõltuvus keeleteaduses (1,2) = Статистическая сопряженность в языкознании (1,2) // Linguistica VI. - Tartu, 1975, lk. 169-187; Linguistica VII. - Tartu, 1976, lk. 190-202. (На эст. яз.; рез.рус.)
Yates F. // Journal of the Royal Statistical Society. Suppl. 1, 1934. - P. 217-235.

ON THE MEASUREMENT OF CORRELATION BETWEEN QUALITATIVE FEATURES IN LINGUISTICS (1): CONTINGENCY OF ALTERNATIVE FEATURES

Juhan Tuldava

S u m m a r y

This article deals with the so-called 2 x 2 contingency tables for measuring mutual connection of qualitative linguistic characteristics. Some special problems concerning the use of Yule's Association Coefficient (Q) and the Bernoullian Correlation Coefficient (Φ) are discussed.

КВАНТИТАТИВНЫЙ АНАЛИЗ ТЕКСТА И ЯЗЫКА: ПЕРСПЕКТИВЫ КОМПЛЕКСНОГО ПОДХОДА

(по материалам семинаров группы "Текст как объект междисциплинарных исследований")

М.Г.Борода, А.А.Поликарпов

Межвузовская проблемная группа "Текст как объект междисциплинарных исследований" была организована в 1985 г. по инициативе ряда исследователей с целью интеграции усилий различных специалистов в области системно-квантитативного анализа текста и языка. Работа группы, объединяющей лингвистов, математиков, музыковедов, психологов из разных ВУЗов страны, организуется бюро (Ю.А.Тулдава, председатель, Тартуский гос. университет, М.Г.Борода, Тбилисская гос. консерватория, А.А.Поликарпов, МГУ, В.К.Детловс, Латвийский гос. университет). Дважды в год группой проводятся семинары по основным направлениям ее работы (к настоящему времени проведено 6 семинаров). В данном обзоре рассмотрена работа III-VI семинаров (о работе I-го и II-го см. Борода, Долинский, 1986).

III семинар группы - "Квантитативные аспекты системной организации текста" (28.10-1.11.86, Тбилиси)-был проведен на базе проблемной лаборатории физической кибернетики Тбилисского гос. университета и кафедры эстетики и искусствоведения Тбилисской гос. консерватории. Были заслушаны доклады по проблемам математических моделей организации текста, квантитативных универсалий в тексте и языке, и др. (см. Борода, Гачечиладзе, Кокочашвили, Цилосани, 1987).

В докладах пленарного заседания была показана связь друг с другом различных квантований лингвистического пространства (П.М.Алексеев, В.Н.Бычков), актуальность в статистической лингвистике понятия нечеткого случайного события (Т.Г.Гачечиладзе, Т.В.Мандхапарашвили), обсуждена проблема минимальных единиц текста (В.И.Перебейнос), показана необходимость сохранения множественности подходов в описании частотной организации текста, (Ю.А.Тулдава).

В докладах секций "Автоматический анализ текста", "Лингвистические распределения" и "Структурные единицы текста" основное внимание было уделено моделированию процесса понимания текста, формирования его базовых единиц, автоматическому анализу лексики, квантитативным характеристикам стиля (докладчики:

А.А.Поликарпов, Ю.К.Крылов, М.Д.Якубовская, А.Н.Лебедев, Н.П. Дарчук, Т.П.Цилосани, Т.Г.Кокочашвили, Г.Ш.Беришвили,В.М.Григорян, Л.Н.Беляева, Т.Я.Аполлонская, Л.И.Колодяжная, Ю.Б.Сафронова, Б.Я.Слепак, В.А.Отлыгин, и др.). В докладах были продемонстрированы новые подходы к проблемам (в частности, в докладе Ю.К.Крылова и М.Д.Якубовской о статусе единиц в связном тексте, А.А.Поликарпова о моделировании рангово-полисемических распределений, Л.И.Колодяжной и А.А.Поликарпова об исследовании семасиологических характеристик в машинных версиях синонимических словарей, Н.П.Дарчук о значимости высокочастотной лексики в авторском стиле, Т.П.Цилосани, Т.Г.Кокочашвили,Г.Ш.Беришвили о моделировании распределений словоформ в тексте).

Серьезное обсуждение вызвали доклады секции "Квантитативная организация текста: общие принципы, математические модели". Заслушанные доклады (М.В.Арапова о лексической согласованности текстов естественного языка, Ю.К.Крылова о моделировании статистической структуры текста на базе вариационных принципов, М.Г.Бороды и А.А.Поликарпова о семиотических универсалиях в музыкальном и естественном языке, А.О.Поликарповой и А.А.Поликарпова об изучении индивидуального знания лексики, и др.) были связаны с новыми подходами, новым аналитическим аппаратом.

Значительный интерес вызвали доклады секции "Количественный анализ музыкального текста", посвященные проблемам квантитативных универсалий музыкального мышления (М.Г.Борода), интонационной организации мелодии (И.В.Бахмутова, В.Д.Гусев, Т.Н. Титкова, А.Рохлин, Т.Н.Овсянникова), моделированию психологически значимых характеристик музыкального произведения (А.Г. Гейн, В.М.Цеханский, Ю.В.Кац, В.П.Овчаренко, Г.Л.Русин), фундаментальным характеристикам ритмики и звуковысотности в народной песне (Я.Росс, И.Рюител), представлению ее в ЭВМ (Э.Г. Гаспарян). Была отмечена органичность разработки этих проблем в рамках развиваемого группой подхода.

IУ семинар группы - "Структурные единицы языка и текста: проблемы междисциплинарных исследований" (23-29.IУ.87,Боржоми -Тбилиси) - был проведен на базе кафедры эстетики и искусствоведения Тбилисской гос. консерватории. Рассматривались проблемы "Структурные единицы и математические модели текста","Ста-

тус единицы в тексте и словаре", "Психологические аспекты проблемы единиц в языке и тексте", "Единицы музыкального текста".

В докладе М.В.Арапова "Объем коллективной словарной памяти" были рассмотрены принципы формирования "коллективного словаря" носителей данного языка и описаны три метода оценки его объема (приводящие к значениям 1.25-2.00 млн.слов) на базе ранее предложенных автором соотношений, связывающих возраст слова и его ранг, и оценки максимальной возможной длины слова. В докладе А.А.Поликарпова "О связи между полисемическими и частотными характеристиками слов" было произведено различение полной (потенциальной) и реализованной на определенном текстовом массиве полисемией; рассмотрена динамика накопления семантического объема слова с ростом его частоты, с ростом объема текстовой выборки. В докладе Ю.К.Крылова "Стохастические модели порождения текста на базе квантово-механических представлений" была предложена гипотеза, что любому лингвостатистическому распределению можно поставить в соответствие некоторый линейный оператор, спектр собственных значений которого однозначно определяет частотную структуру (набор частот) текста. Теория, построенная на этом, и некоторых других допущениях, позволяет построить различные модели порождения текста, уже простейшие из которых дают возможность объяснить ряд наблюдаемых в лингвистических распределениях эффектов (напр., синонимии, "сгущений" в словаре, и др.). В докладе Ю.К.Крылова и М.Д.Якубовской "О тематических единицах связного текста" была показана актуальность использования в квантитативном анализе текста единиц, границы которых совпадают с естественно выделяемыми в тексте тематическими фрагментами, и предложен формальный подход к выделению в тексте таких единиц, связанный с изучением статистических характеристик, динамика флуктуаций которых была наиболее чувствительной к изменению локальной тематики текста. Актуальности формирования общего подхода к проблеме единиц текста был посвящен доклад М.Г.Бороды "Базисные единицы текста: проблемы определения", где была подчеркнута кардинальность выделения спектра однозначно определенных и естественных единиц для всего системно-квантитативного подхода к языку и тексту, важность исследования процессов расчлененности в речи и музыке с некоторых единых позиций, предложены "ритмический" подход к этой проблеме, связанный с опытом выделения музыкальных единиц.

В докладе А.Н.Лебедева "Единицы памяти-единицы языка-единицы текста" проблема единиц была рассмотрена с позиций развиваемой автором теории циклической актуализации образов памяти и, в частности, гипотезы о существовании некоторого нижнего предела для относительной частоты появления доминантного образа, а также - предположения об изменчивости ранга данного слова в последовательных отрезках текста. "Психологическим аспектам проблемы единиц были посвящены доклады В.Я.Ляудис "Функционирование языковых единиц в тексте и проблемы памяти" и Т.Н. Ушаковой "Единицы текста и проблема "внутренней речи"".

Различные аспекты проблемы функционирования базовых единиц в музыкальном тексте были рассмотрены в докладе В.К.Детлова (где был предложен подход к проблеме оценки "эффективности" единиц), Э.Г.Гаспарян, Г.М.Гварджаладзе (рассмотревшей специфику расчленённости в грузинской народной песне), сотрудников кафедры эстетики и искусствоведения Тбилгосконсерватории М.В. Натенадзе, М.В.Цхакая, А.М.Курдашвили.

Заключительное заседание семинара состоялось 29.04.87 на кафедре эстетики и искусствоведения ТГК. В выступлении зав.кафедрой, проф.И.А.Урушадзе была подчеркнута необходимость комплексного, интердисциплинарного, подхода к проблеме единицы, рассмотрения квантитативных закономерностей текста в плане отражения в них закономерностей творческого процесса.

У семинар группы - "Системно-квантитативные проблемы исследований языка и текста" (3-8.X.1987, Звенигород) был организован на базе кафедры общего, сравнительно-исторического и прикладного языкознания МГУ. В числе основных методологических докладов были доклад Ю.А.Тулдава "Методологические проблемы математических моделей в лингвистике", М.В.Арапова "Об основных проблемах квантитативной лингвистики", Р.Ю.Кобрина "Методологические проблемы изучения терминологической лексики" и др.

В докладах Б.И.Барткова "О мерах синонимичности слов" и "О градациях морфем по степени их аффиксальности" были поставлены вопросы квантификации семасиологических и грамматических отношений в языке. В докладе А.А.Поликарпова и Г.О.Карпиевой "К вопросу об определении гиперлексемы" были предложены критерии выделения вариантов и инвариантов лексических единиц, в частности, гиперлексемы как центральной единицы лексического сос-

тава языка. Проблема определения статуса ассоциативных характеристик слов как характеристик его системной организации была рассмотрена в докладе В.А.Долинского. В докладе В.Я.Курлова рассматривалась зависимость качественных и количественных характеристик стилистических помет слов от их принадлежности к определенному полисемическому диапазону. А.М.Малов предложил свою интерпретацию параметров полисемических распределений лексики английского языка. Проблеме моделирования распределений словоформ в научно-технических текстах с помощью известных в лингвостатистике законов был посвящен доклад Т.П.Цилосани и Т.Г.Кокочашвили. Особый слой разговорной лексики - лексика т.н. пьющих лиц, связанная с их функционированием в этом плане - рассматривался в докладе А.П.Василевича.

Ряд докладов семинара был посвящен анализу системных характеристик музыкального текста и языка. Проблеме моделирования ряда важных характеристик музыкального восприятия (в частности, ладовых) был посвящен доклад Ю.В.Кац. Эффективность аппарата кластерного анализа в задаче типологической классификации мелодики эстонских рунических песен была показана в докладе И.Рюйтел, где был предложен новый подход к проблеме. Общим принципам эволюции "музыкальной лексики" в европейской музыке (в аспекте предложенной автором гипотезы постепенной "аналитизации" европейского музыкального языка) был посвящен доклад М.Г.Бороды (проблематика данного и предыдущего докладов в более расширенном виде была представлена на конференции "Прикладная лингвистика и автоматический анализ текста"-см.ниже).

VI семинар группы, проведенный в рамках конференции "Прикладная лингвистика и автоматический анализ текста" (28-30.I. 88, Тарту), был организован Группой прикладной лингвистики Тартуского гос. университета. В докладах были предложены новые методы квантитативного изучения словаря и текста с точки зрения их системных характеристик, рассмотрены различные аспекты систем человеко-машинного диалога, автоматического анализа текста и создания машинных фондов языка, компьютеризации преподавания языка, квантитативного музыковедения.

В докладе М.В.Арапова был предложен оригинальный метод сравнения словарей, основанный на введении автором некоторой теоретически обоснованной меры их сходства (нормы различия) и, затем, исследовании групп слов, отклоняющихся в своем употре-

блении от этой нормы. Предложенный метод показал свою эффективность при сравнении словарей Пушкина и Лермонтова (полные корпусы текстов). В докладе Ю.А.Тулдава было предложено решение актуальной для квантитативного анализа текста проблемы измерения взаимной сопряженности лингвистических объектов, был предложен усовершенствованный метод такого измерения, приведены примеры его эффективного применения. "Синергетическому" подходу к анализу лексики, проблеме измерения взаимосвязи различных лингвистических признаков друг с другом были посвящены доклады Г.Г.Сильницкого, Г.В.Сильницкой, С.Н.Андреева, в которых было показано, что степень многозначности глагола связана с его структурной сложностью, степенью валентности, отражается в диахроническом аспекте, и др. - в целом, выявлены связи семантических групп глаголов с 11 признаками фонетического, морфологического, внешнедеривационного и синтаксического уровней. Проблеме измерения семантической информации в слове был посвящен доклад С.В.Райтар. В докладе А.Арольд был изложен опыт классификации и расчленения семантического поля методами корреляционного и факторного анализа. Новый (эргодический) подход к выводу закона Ципфа был предложен Л.Блюмовым.

Значительный интерес вызвали доклады А.А.Поликарпова - "Логическое пространство единиц лексической подсистемы языка и квантитативные соотношения между ними" и "К теории жизненного цикла слова". В первом из них был выделен спектр лексических и грамматических признаков плана выражения и плана содержания лексических единиц и описана, на базе их комбинирования, логическое пространство классификации лексических единиц языка и их типы (словоформы, ЛСВ, лексема, гиперлексема, и т.д.). Во втором докладе был сделан ряд допущений, позволяющих построить модель жизненного цикла слова, начинающегося с его рождения, как однозначной единицы, обладающей высокой степенью конкретности, образности, ассоциативности значения. По мере образования новых значений порождающие потенции исходного значения слова исчерпываются, а новые значения оказываются все менее конкретными, менее способными к дальнейшему развитию-делению. Усиливается степень вхождения слова в устойчивые сочетания слов, часть из которых превращается в единое производное слово. Прежде самостоятельные слова превращаются в нем в морфемы, в результате дальнейшего опро-

щения слова "рассасывающиеся", завершая жизненный цикл слов. В докладе А.В.Зубова была предложена новая концепция порождения текста как психолингвистического объекта. Проблеме валентностных свойств текста был посвящен доклад В.М.Григоряна.

На секциях "Автоматический анализ текста" и "Машинные фонды языка" были заслушаны доклады о комплексных проектах в этом плане (проект ЛИНДА, представленный в докладе О.Н.Гринбаума, Г.Я.Мартыненко, С.Я.Фитиалова; проекты машинных фондов латышского и грузинского языка, а также фонда русского и национальных языков были обсуждены в докладах С.Н.Клявиня, Ю.К.Орва, В.М.Андрющенко). Проблеме автоматического морфологического анализа текста был посвящен доклад Н.П.Дарчук. На секциях, посвященных проблемам диалога человека и ЭВМ и компьютерного преподавания языков центральное место заняли сообщения центрального место заняли сообщения исследователей из Тартуского университета - Х.Я.Нымма, М.Э.Салувеэр, Ю.А.Тулдава, Х.Нымма, Я.А. Маика, а также - А.И.Мордынской, О.П.Крюковой, и др. На секции "Вопросы квантитативного музыковедения" обсуждались проблемы автоматического распознавания мелодических типов, выявления их взаимосвязей (доклад И.Рэйтел и К.Хаугаса), формирования комплексных программ исследования музыки с помощью ЭВМ (И.В.Бахмутова), теория сегментации музыкального текста (В.К.Детлове), анализа повторов (А.Н.Мурадян), внутренних факторов эволюции музыкального языка (М.Г.Борода).

Значительный интерес вызвал проведенный в рамках конференции круглый стол (руководитель - А.А.Поликарпов) по проблеме системных взаимосвязей характеристик лексических единиц. В выступлениях была отмечена значимость комплексного изучения этих взаимосвязей, естественно формирующихся в коммуникативном процессе под влиянием фундаментального "принципа намеканий".

QUANTITATIVE TEXT ANALYSIS: PERSPECTIVES OF INTERDISCIPLINARY STUDY

M.G.Boroda, A.A.Polikarpov

S u m m a r y

The survey covers 346 seminars of the research group "Text as an object of interdisciplinary study". The group has been set up in 1985; it unites specialists in quantitative linguistics, mathematics, musicology and psychology from various cities of the USSR. Its organizing committee includes J.A.Talova, chairman (Tartu), M.G.Boroda (Tbilisi), A.A.Polikarpov (Moscow), V.K.Detlovs (Riga).

Р Е Ц Е Н З И Я

Э.М. Добрускина, В.Е. Берзон. Синтаксические сверхфразовые связи и их инженерно-лингвистическое моделирование / Отв. редактор д.ф.н. Р.Г. Пиотровский. - Кишинев: Штиинца, 1986. - 167 с.

(Книга состоит из "Введения", 5 глав, "Заключения". В приложении содержатся результаты работы системы автоматического реферирования на ЭВМ, исчерпывающий список литературы по проблеме - 178 наименований, списки текстов и условных обозначений).

В рецензируемой книге содержатся результаты многоцелевого исследования единиц и связей сверхфразовой структуры специальных текстов. В отличие от многих работ по лингвистике текста (ЛТ) преимущественно описательного характера труд Э.М. Добрускиной и В.Е. Берзона изначально был ориентирован на создание конструктивно-технологических приемов анализа научных текстов с использованием теоретической базы инженерной лингвистики, разработку и машинную реализацию алгоритмов формального выделения сверхфразовых связей и сверхфразовых единиц для прагматических задач (автоматическое реферирование, индексирование, фрагментирование, машинный перевод текстов).

Аналитическим разделам книги предпослана теоретическая часть (глава I), где рассматриваются современное состояние и основные понятия лингвистики текста, в частности, специфика собственно текстовых единиц (сверхфразовых) и системных отношений за пределами предложений. В центре внимания авторов понятия и реалии сверхфразового уровня, сверхфразовой структуры, сверхфразовых отношений, связей и сверхфразовых единиц.

Важнейшим объектом изучения и моделирования в работе являются сверхфразовые связи как грамматические механизмы объединения предложений в смысловые блоки (глава 2). На основе изучения сходства и различия синтаксических связей на сверхфразовом (СУ) и внутрифразовом (ВФУ) уровнях делается вывод о большей "синтаксичности", т.е. четкой выраженности внутрифразовых связей. Но хотя межфразовые связи пока осознаются как смысловые, а не грамматические, авторы не без оснований настаивают на самостоятельности сверхфразового синтаксиса, поскольку существует только ему свойственная об-

ласть описания таких, например, явлений, как прономинализация, явление анафорики, категорий неопределенности - определенности и пр.

Одна из сильных сторон работы - попытка подойти вплотную к созданию теории сверхфразового синтаксиса, ведущее место в которой занимает понятие синтаксической сверхфразовой связи (ССС). Эту теорию предлагается строить на основе:

1) теории валентности, допускающей существование и обнаружение для каждого предложения набора межфразовых валентностей (что определяется наличием повторов);

2) теории насыщения предложений в связном тексте (позволяет различать предложения автосемантичные (насыщенные), синсемантичные (ненасыщенные) и изолированные).

Но представления авторов о предмете синтаксиса достаточно спорны. Настаивая на полноправном существовании сверхфразового синтаксиса, что в целом вполне логично, З.М. Добрускина и В.Е. Берзон неожиданно интерпретируют онтологию синтаксиса. В их трактовке "предметом синтаксиса являются синтаксические единицы различных уровней языка (подчеркнуто мной - Л.С.) в их синтагматических и парадигматических отношениях ..." (с. 31). Если синтаксис - это языковый уровень, как до сих пор было принято думать, то, судя по данному определению, он не имеет четких очертаний, так как его единицы - элементы и других уровней языка? Авторская позиция в этом вопросе еще более определенна на стр. 52, где синтаксисом объявляется "аппарат для объединения единиц языка одного уровня в единицах более высокого уровня".

По-видимому, все же целесообразно сохранить за синтаксисом его сложившуюся в лингвистике "уровневую" определенность, ограничив предмет синтаксиса именно синтаксическими единицами (словосочетание, предложение, наконец, СФЕ) и синтаксическими категориями. Кроме того, в данном случае помогает все расставить на свои места идущая от Ф. де Соссюра и сложившаяся в науке тенденция различения синтагматики и парадигматики как разных форм реального существования языковых единиц (В.М. Солнцев, Б.Н. Головин). В свете этой тенденции то, что для авторов книги является синтаксисом, есть не что иное как область синтагматики - одной из форм функционирования, определяющей возможности линейного сочетания и реализации объединения единиц разных языковых уровней (фонем в морфемы - синтагматика фонетическая, морфем в слова - синтагматика морфем и т.п.). Вероятно, сверхфразовый уровень также

может рассматриваться как одна из реализаций синтаксической синтагматики предложений, законы которой и стали предметом новаторского изучения в рецензируемой монографии.

Развивая теорию сверхфразового синтаксиса, авторы активизируют внимание на классификации сверхфразовых связей, построении их типологии и выделении сверхфразовых единиц (главы 3 и 4). На материале английских и русских научных текстов выявляются эксплицитные маркеры синтаксических сверхфразовых связей. На базе классификации маркеров СС разработана экспериментальная типология синтаксических сверхфразовых связей, используемая при создании алгоритмов сегментации текста на СФЕ.

Процессу выделения СФЕ должна, по убеждению авторов, предшествовать процедура алгоритмического построения сверхфразовой структуры текста. Используя математический аппарат алгебры отношений и теории графов, авторы наглядно демонстрируют, как с помощью ряда формальных процедур эксплицируется структура текста в виде связного ориентированного графа. Реализуя идею "нисходящего" подхода (от целого текста к его единицам). Э.М. Добрускина и В.Е. Берзон предлагают формальный аппарат вычленения из текста разных модификаций СФЕ на базе графа сверхфразовой структуры. Правило выделения СФЕ несложно: первое предложение СФЕ должно быть автосемантичным, другие - синсемантичны, непосредственно или косвенно связаны с автосемантичным. Учитывается и ситуация невыраженности маркера связи, что обусловило различение эксплицитных (ЭСЕ) и имплицитных сверхфразовых единиц (ИСЕ).

На основе моделей текстовой структуры разработаны сложные процедуры алгоритмического анализа семантики текста, часть которых реализована на ЭВМ. В книге ретроспективно освещены принципы программного обеспечения системы автоматизированной обработки текста, которое включает общее программное обеспечение (экспликации на ЭВМ синтаксической сверхфразовой структуры) и частное программное обеспечение подсистем (автоматического реферирования, автоматического индексирования, фрагментирования и пр.).

Вместе с тем некоторые теоретические положения работы недостаточно обоснованы. Такова, в частности, концепция сверхфразового уровня как языкового, аналогичного другим уровням языка. В основе тезиса о языковом статусе сверхфразовых явлений лежат возможные, по мнению авторов, доказательства неисчислимости (конечности) связей в текстах, числа

СФЕ. Сами СФЕ представляются авторам единицами дискретными, регулярными и воспроизводимыми. Показанные же в книге алгоритмы проверены только на материале научных текстов, где высока степень стандартизованности, схематичности синтактико-смысловой организации. Но и там они не всегда применимы и работают только при наличии эксплицитных маркеров связности. Разработка же трансформационных процедур перехода от имплицитных сверхфразовых связей и единиц к эксплицитным связям и единицам чрезвычайно сложна (и, видимо, не всегда возможна). Очевидно, что всеобъемлющих и строгих формальных правил извлечения любых СФЕ из текста пока (а может, и вообще) не существует, проблема объективного вычленения этих единиц не решена. И утверждения, что СФЕ относятся к разряду дискретных, регулярных, воспроизводимых единиц, и, следовательно, структурируют особый языковый уровень, весьма гипотетичны.

Спорность и неустойчивость концепций естественна при становлении новых теорий, а в разработке теории сверхфразового синтаксиса делаются лишь первые шаги. Несомненно, что книга Э.М. Добрускиной и В.Е. Берзона вызовет интерес у специалистов по аналитическому описанию и моделированию текстовой информации и создателей информационных систем машинной обработки текстов.

Серебрякова Л.А.
(г. Горький)

Review of the monograph: E.M. DOBRUSKINA, V.E. BERSON. SYNTACTIC SUPERPHRASE RELATIONS AND THEIR COMPUTERIZED AND LINGUISTIC SIMULATION. - Kisinev: Štiintsa, 1986. - 167 p.

L.A. Serebrjakova

S u m m a r y

The monograph under review presents the results of multipurpose research of units and relations of superphrase structures in scientific and technical texts. Basic foundations of a new theory, the theory of superphrase syntax, are described, and the principles for classifying the superphrase relations and for recognising the superphrase units are also proposed.

On the basis of text structure simulation there have been developed and described in detail the complex procedures of the algorithmic analysis of text semantics, these procedures being considered essential for automatic abstracting, indexing and text segmentation. A part of these procedures has been realized on digital computers.

С О Д Е Р Ж А Н И Е

Хроника:

Рецензия:

SUMMARIES - RESÜMEES

Ученые записки Тартуского государственного университета.
Выпуск 827.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ
АНАЛИЗ ТЕКСТОВ 1988.
QUANTITATIVE LINGUISTICS AND AUTOMATIC
TEXT ANALYSIS.
На русском языке.
Резюме на английском языке.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор Ю. Туддава.
Подписано к печати 28.06.1988.
Формат 60x90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 10,88. Печатных листов 11,0.
Тираж 500.
Заказ № 622.
Цена 2 руб. 20 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.