TARTU RIIKLIKU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

591

KVANTITATIIVSE LINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜSI AKTUAALSEID PROBLEEME

АКТУАЛЬНЫЕ ПРОБЛЕМЫ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ И АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

Töid keelestatistika alalt

VII

Труды по лингвостатистике

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 591 ВЫПУСК ОСНОВАНЫ В 1893.г.

KVANTITATIIVSE LINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜSI AKTUAALSEID PROBLEEME

АКТУАЛЬНЫЕ ПРОБЛЕМЫ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ И АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

Töid keelestatistika alalt

VII

Труды по лингвостатистике

ТАРТУ 1981

Ученые записки
Тартуского государственного университета.
Выпуск 591.
АКТУАЛЬНЫЕ ПРОБЛЕМЫ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ
И АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ.
Труды по лингвостатистике УII.
На русском языке.
Резюме на английском и немецком языках.
Тартуский государственный университет.
202 400, ЭССР, г.Тарту, ул.Юликооли, I8.
Ответственный редактор Я.Соонтак.
Сдано в печать I4.I2.I98I.
МВ I0952.
Формат 30х45/4.
Бумага печатная.
Машинопись. Ротапринт.
Учетно-издательских листов I0,2.
Печатных листов I0,25.
Тираж 400.
Заказ № I353.
Цена I руб. 50 коп.
Типография ТГУ, 202400, г.Тарту, ул.Пялсона, I4.

О КВАНТИТАТИВНОЙ ТИПОЛОГИИ ТЕКСТА

П.М. Алексеев

Возродившийся за последние годы интерес к изучению лингвистических свойств текста вызван потребностью в адекватных описаниях речевой деятельности (langage), которые используются при моделировании механизмов и процессов порождения и восприятия речи. Такие описания и модели необходимы для теории языка, лингводидактики, для теории и практики перевода, для работ по созданию систем автоматизированной переработки текста и систем искусственного интеллекта.

Возникающие при этом задачи решаются в современном языкознании коллективным, комплексным и многоаспектным изучением устройства и поведения сложных лингвистических объектов. Только в результате целенаправленных, системных исследований можно получить достоверные и развернутые сведения о языке и речи.

Системные исследования языка и речи можно осуществлять как по линии дедуктивных порождающих процедур, так и по пути индуктивного моделирования текста. В первом случае система языка, его "грамматика" рассматривается как конечное множество детерминированных правил, а реализация системы - как бесконечное число регулярных цепочек слов, построенных по этим правилам. Объяснение языковых объектов "точными" терминами тогда относится к компетенции "алгебраической" лингвистики, а количественными методами описываются лишь явления речи, текста.

Лингвисты, идущие по второму пути, считают, что система языка как результат коллективного опыта его носителей в действительности порождает не все "грамматически правильные" цепочки, но только те из них, которые являются "осмысленными". В порождении текста участвуют система и норма языка, узус и ситуация, поэтому естественные языки должны описываться контекстно-зависимыми грамматиками. Для выявления ограничений, накладываемых нормой и ситуацией на функционирование системы, используются методы индуктивной лингвистики текста, в том числе статистико-вероятностные методы. Лингвистические объекты могут, таким образом, описываться с помощью вероятностных оценок как на уровне речи, так и на уровне языка.

Индуктивное количественное исследование текста не сводится к накоплению фактического материала, но предусматривает его последовательное обобщение. В процессе количественного лингвистического анализа текста на каждом этапе обобщения выявляются характерные свойства текстов, идиолектов, подъязыков, функциональных стилей, языков. Полученные этим путем характеристики выступают в качестве типологических признаков соответствующего объекта, а лингвистика текста становится средством изучения типологии языка и речи. Текст как множество всех текстов на данном языке, уже созданных и тех, которые будут созданы, содержит полные сведения о системе и норме языка, обо всех лингвистических подсистемах. "Нет в языке ничего, чего не было бы ранее в речи" (Бенвенист, 1974, с. 140; ср. Соссюр, 1977, с. 57) и, следовательно, в тексте.

Поскольку при таком подходе важным инструментом являются количественные методы описания и анализа материала, можно утверждать, что в задачи лингвистики текста входит его квантитативно-типологическое описание, а сама эта дисциплина может приобрести статус квантитативной типологии текста. Этим термином здесь предлагается обозначать типологические исследования текста, которые исходят из статистико-вероятностных, системно-структурных и семиотико-информационных представлений о языке и речи и применяют соответствующий этим представлениям аппарат, сочетая его с собственно лингвистической методикой анализа.

В настоящее время накоплено большое число квантитативных исследований языка и речи. Однако обобщение, упорядочение и осмысление их материалов, выделение релевантных теоретических понятий и принципов происходит пока еще медленно. Причину отставания теории квантитативной лингвистики от ее практики следует, вероятно, искать в первую очередь в том, что основные методические и технологические понятия этой области пока еще слабо связаны с такими общими понятиями теоретического языкознания, как речевая деятельность, система и норма языка, узус, текст.

Ближайшая задача индуктивной количественной лингвистики текста видится поэтому в выделении и обобщении ее концептуально-методологической и методической базы, в экспликации этой базы на фоне общих понятий современного языкознания, соотносимых со схемой речевой деятельности.

Цель квантитативной типологии текста определяется как получение вероятностных и статистико-информационных моделей,

описывающих и объясняющих типологические особенности сложных лингвистических объектов, которые представлены текстом.

Предметом квантитативной типологии текста можно считать описание речевой деятельности на различных уровнях ее системно-структурной организации, начиная с исходной, текстовой репрезентации. Текст со всеми содержащимися в нем лингвистическими единицами является непосредственным объектом изучения как единственная данная в наблюдении реальность. Текст, будучи результатом речевого акта, включает в себя некоторую часть инвентаря языковых элементов. Они сочетаются в тексте в соответствии с грамматикой языка и речи, отбираются для использования в нем согласно этим правилам и в зависимости от условий внешней ситуации; их отбор регулируется и нормой языка. Таким образом, текст реализует и одновременно формирует систему языка, норму, речь, функциональные стили, подъязыки. Поэтому лингвистическая типология текста отражает типологию всех реализуемых в тексте лингвистических систем и подсистем. От наблюдений над текстом к выявлению его структуры, об обобщений на уровне текста к наблюдениям на уровнях узуса, нормы и системы языка - таков путь, который используется для индуктивного построения теорий речевой деятельности.

Лингвистика текста, следовательно, является главным инструментом для таких теорий. Она может тогда не ограничиваться рассмотрением индивидуальности отдельного текста, как это делается в стилистике художественной речи. Ее задачей будет выявление определенного стереотипа, лежащего в основе устройства усредненного текста, выявление лексико-фразеологических, морфологических, синтаксических, ситуативных и других "формул" построения текста (Пиотровский, 1975, с. 55-56). Понятие усредненного текста позволяет говорить о типологии текста вообще, а не только о типологии текстов, хотя не исключается рассмотрение и отдельного текста. Важно подчеркнуть, что единые, унифицированные процедуры могут применяться для описания устройства текста типового и текста конкретного.

Статистико-вероятностные приемы анализа речевого материала, сочетаясь с собственно лингвистическими приемами, образуют методику квантитативной типологии текста. Обсуждение проблем, возникающих в связи с лингвистическими, вероятностными, системно-структурными и семиотико-информационными представлениями о речевой деятельности, лежит в области теории квантитативной типологии текста.

Аксиоматическая часть этой теории может быть представлена рядом утверждений, принимаемых в качестве постулатов.

1. Язык и речь суть проявления языкового феномена.

2. Язык включает в себя систему и норму.

3. Речь рассматривается как узус и совокупность речевых актов, текстов.

4. Язык и речь системны.

5. Как язык, так и речь характеризуются и парадигматическими, и синтагматическими отношениями между лингвистическими элементами и между классами элементов.

6. Языковому феномену свойственны категории качества и количества: и язык, и речь подлежат количественным измерениям.

7. Язык и речь информационны, являясь средством коммуникации, средством передачи и приема сообщений, содержащих синтактическую, семантическую и прагматическую информацию.

8. Язык и речь, как проявления лингвистического поведения, вероятностны, как вероятностна любая форма поведения человека.

9. Язык и речь семиотичны, участвуя в семиотико-информационных процессах.

На базе этих постулатов можно строить схемы речевой деятельности, интерпретирующие и дополняющие исходную соссюровскую диаду язык-речь.

Особого внимания заслуживает вопрос о месте нормы и узуса в схемах речевой деятельности. Если норма - это свод правил реализации системы языка, то она не может быть полностью изолированной от условий существования языка; следовательно, она связана с внешней обстановкой. Но поскольку норма представляет собой лингвистические правила употребления лингвистического материала, она не может быть противопоставлена и языку. Значит, она входит в состав языка и связывает его с неязыковыми ситуациями. С учетом этих ситуаций норма регулирует использование элементов, структур, моделей системы языка в речи. Она, таким образом, служит фильтром, распределяющим возможности системы (обеспечивая при этом некоторый выбор в ограниченных пределах) по конкретным речевым актам в зависимости от конкретных типовых ситуаций.

Между нормой, выполняющей функцию такого фильтра, и речевыми актами должно находиться еще одно звено. Норма, существующая в осознании коллектива носителей языка, постоянно подкрепляется кодификацией не только в виде нормативных спра-

вочников, но и в виде множества высказываний, принимаемых коллективом за образцовые, за кодифицированную норму. Если перефразировать Э.Косериу (Косериу, 1963, с. 175), то система содержит то, что можно говорить, норма - то, что и как следует говорить, а речь - то, что и как говорится в действительности. Тогда будет определенное различие между тем,что и как говорится на самом деле, и тем, что и как принято говорить, обычно говорится. Очевидно, что есть разница также между понятиями "следует говорить" и "принято говорить". Это последнее ("принято говорить") и относится к промежуточному звену между нормой и речевым актом. Оно представляет собой узус, обобщение конкретных речевых актов и текстов.

Систему и норму объединяет их принадлежность к языку,узус и собственно речь - их принадлежность к речи, норму и узус - их "нормальность". Узус можно понимать как неосознанную и некодифицированную норму. Сама же норма выполняет роль фильтра не только при переходе от системы языка к речи; она фильтрует накопленные в речевых актах и текстах и обобщенные узусом изменения, прежде чем эти изменения попадут в систему.

В одной из наиболее полных сводок для статистических концепций языка и речи (Богданов, 1973) не нашли места существующие различия между парадигматикой и синтагматикой на речевом уровне. Между тем, если эти различия на уровне системы относятся пока к классу ненаблюдаемых, моделируемых объектов, то в речевом акте и в тексте эти отношения реальны и обнаруживаются в наблюдении. Следовательно, как язык, так и речь характеризуются и синтагматикой, и парадигматикой (Головин, 1969; Солнцев, 1971, с. 65).

Действительно, каждый текст (или каждая совокупность текстов) реализует какое-то ограниченное число разных языковых элементов, например, слов. Все эти слова обладают одинаковой парадигматической численной характеристикой. При группировке этих слов в классы, например, по частям речи, их равные доли в словаре текста будут давать в сумме разные численности классов. При группировке по другому признаку суммарные веса будут перераспределяться, образуя новые по объему классы.Соответственно парадигматические характеристики слов будут приобретать различное численное выражение. И естественно, далее, что синтагматические их признаки, такие, как способность к сочетанию с другими словами, также могут и будут приобретать различное количественное выражение. Численная мера парадигматики и синтагматики лингвистических явлений в тексте будет

определяться соответственно через парадигматические и синтагматические частоты.

Однако и такого представления о синтагматике и парадигматике в речи недостаточно для описания реальных, наблюдаемых в тексте употреблений лингвистических элементов. Сочетание слов в фразеологические образования, в синтагмы, в предложения реализует синтагматические характеристики слов. Такие сочетания, как "in order to" или "in accordance with" могут встретиться в тексте более чем однократно, скажем, по 5 раз. Тогда цифра 5 будет численной синтагматической характеристикой (частотой) слов "order" и "accordance", если иметь в виду их сочетаемость со словами "in", "to", "with" в данных трехсловных контекстах. Но слово "in" будет иметь еще и характеристику 10, показывающую его способность быть употребленным в словосочетании 10 раз; оно будет характеризоваться также цифрой 2, показывающей реализованную способность войти в 2 словосочетания. Оно будет характеризоваться и другим, гораздо большим числом, свидетельствующим о его способности встретиться в тексте и в других сочетаниях, в других контекстах, которая проявляется в его суммарной частоте в данном тексте. Очевидно, что речевая синтагматическая характеристика лингвистической единицы отражается не в одном виде частот. Следовательно, и синтагматические частоты нельзя однозначно квалифицировать только как суммарные частоты употребления лингвистического элемента в тексте. Все они должны занять место в статистической схеме речевой деятельности.

В этой связи необходимо найти более четкие определения парадигматики и синтагматики. Вот как предлагает понимать эти "две формы и два вида функционирования языковой структуры, всех ее единиц и категорий" Б.Н.Головин: "Парадигматика осознается... как область закономерного варьирования единиц и категорий языка в процессе их функционирования для построения речи. Синтагматика осознается как область закономерного сцепления единиц и категорий языка для построения речевой структуры" (Головин, 1969, с. 76). Синтагматика, далее, представляет собой двуединую сущность. С одной стороны, она включает в себя валентность (потенциальную способность лингвистического элемента сочетаться с другими элементами того же уровня), а с другой - она включает в себя сочетаемость - реализацию этой валентности в речевом потоке (Головин, 1969, с.79). Несколько сходное определение предлагает В.М.Солнцев: "Оба эти вида отношений присущи элементам языка и, следовательно,

характеризуют язык в целом... каждый элемент, или каждая единица языка потенциально может вступить в три вида отношений: парадигматические (или ассоциативные), синтагматические (отношения актуального взаимодействия или в абстрактной форме - отношения классов) и иерархические (отношения вхождения в более сложную единицу)" (Солнцев, 1971, с. 65-66).

Последнее понятие как будто исключает рассмотрение сложных лингвистических единиц в терминах синтагматики более простых единиц, составляющих сложные, однако, оно позволяет делить синтагматические отношения на одноуровневые и разноуровневые; последние могут связывать единицы не только соседних, но и более отдаленных уровней.

Лингвистику текста и квантитативную типологию текста интересуют прежде всего явления, образующие систему текста. Они составляют ее объект; однако, поскольку квантитативная типология текста рассматривается как часть языкознания, часть науки о языке, описание системы текста приводит к описанию на его базе системы языка.

Для квантитативной типологии текста инвентаризация лингвистических единиц и описание отношений между ними в пределах одного уровня и на разных уровнях системной иерархии является начальным, исходным этапом, что сопутствует в терминах семиотики синтактическому уровню описания речевой деятельности. Конкретные лингвистические единицы текста (употребления) и представители их в словаре текста обобщаются в понятия знакоупотребления и знакотипа. Это фундаментальное противопоставление знака и его употребления в тексте связывается в общем с соссюровской антиномией абстрактного-конкретного с тем уточнением, что "абстрактный" уровень системы текста (речи) не полностью тождественен абстрактному уровню системы языка и что антиномия абстрактного-конкретного в речи и, следовательно, бинарное противопоставление знак-знакоупотребление предстает как триада "знак языка - знак речи - знакоупотребление". Системно-структурный аспект речевой деятельности оказывается, таким образом, тесно связанным с семиотико-информационным аспектом[1].

---

[1] Другие вопросы системного подхода в квантитативной лингвистике обсуждаются в работе: (Алексеев, 1973), а проблемы вероятностной трактовки парадигматики и синтагматики в работе: (Алексеев, 1977).

Квантитативная типология текста имеет дело с лингвистическими знаками, но по необходимости в поле зрения попадают и элементы других семиотических систем, например, вторичных по отношению к языку, искусственных, если они используются в речевом сообщении. Кроме того, она может служить и служит своим материалом для создания вторичных систем знаков. Однако в этом случае, строго говоря, имеют место не семиотические, а кодовые системы, например, в стенографии, сурдо- и тифлопедагогике, в технике передачи сообщений. Рассмотрение неязыковых знаков, по-видимому, вообще лежит вне компетенции лингвиста и соответственно за пределами квантитативной типологии текста. Это, впрочем, не означает, что квантитативная типология текста не имеет никакого отношения к языковому знаку. Так же, как и в случае построения вторичных кодовых систем на основе ее данных, методы анализа языковых знаков могут использоваться при анализе квантифицированных последовательностей совокупностей элементов других, неязыковых "сообщений", например, произведений музыкального или изобразительного искусства (Моль, 1973, гл. I-IУ; Орлов, 1976).

В знаковом процессе, семиосисе, участвуют источник (отправитель сообщения), канал для передачи сообщения, знак, интерпретатор (приемник сообщения), интерпретант (готовность интерпретатора реагировать на означаемое знака и выдавать ответный эффект) и контекст (Nauta, 1972, с. 27-28; Пиотровский, 1975, с. 6-7). Семиотико-информационный процесс имеет три уровня, соответствующие формальному аспекту (синтактика), аспекту значения (семантика) и функциональному аспекту (прагматика). Систематический порядок в семиотике (синтактика-семантика-прагматика) не совпадает с эпистемологическим порядком (прагматика-семантика-синтактика). Лингвистическое исследование базируется на прагматических аспектах использования языка (Nauta, 1972, с. 35 и след.).

Здесь необходимо внести некоторое уточнение. В лингвистике, в частности в лингвостатистике и в квантитативной типологии текста, строгая последовательность эпистемологической иерархии уровней не выдерживается. Отбор текста, изучение его "ценности" соответствует прагматическому уровню, изучение его содержания в общем связано с семантическим, а исследование составляющих его единиц - с синтактическим уровнем. Однако это прагматика, семантика и синтактика всего текста, а не только входящих в него знаков. Собственно изучение знака начинается на синтактическом уровне информационной ор-

ганизации текста; более того, квантитативная типология текста сосредоточивает внимание на сегодняшнем этапе своего существования прежде всего на синтактике. Помня, что знак содержит в себе информацию всех трех типов, стараются извлечь из него максимум информации первого типа - синтактической информации. Извлечение и упорядочение семантической информации, содержащейся в знаке, в квантитативной типологии текста также пока еще происходит на "нижнем", фундаментальном уровне информационной организации текста - на синтактическом уровне.

Однако если нет никакого значения и никакой целенаправленности в коммуникативной ситуации или если абстрагироваться от аспекта значения, то останется только передача физических состояний и событий с некоторой степенью неожиданности. Можно отнести к ней некоторое количество потенциальной информации, если мы готовы в то же время приписать ей в соответствующей мере потенциальное значение и целенаправленность (Nauta, 1972, с. 62).

Для квантитативной типологии текста и для инженерной лингвистики отсюда следуют очень важные выводы. С одной стороны, полностью абстрагируясь от содержания носителя информации или пытаясь это сделать, как и бывает при составлении статистических инвентарей (например, частотных словарей) единиц текста без учета значений этих единиц, мы имеем дело преимущественно с потенциальной информацией, "прединформацией", которая еще не тождественна синтактической информации, хотя лежит в ее основе и включается в нее. С другой стороны, шенноновская информация уже используется для оценок актуальной информации, в том числе семантической и прагматической (Пиотровский, 1975, с. 150-206).

При анализе сообщений на естественном языке и инвентаря семиотических элементов, используемых в этих сообщениях, нельзя полностью исключить аспект значения; можно лишь отвлечься от него на том или ином (начальном) этапе исследования. Вероятностная лингвистика идет от наблюдения на досинтактическом (в понимании Д.Науты) уровне семиосиса через синтактический уровень к семантическому и прагматическому. Хотя на каждом уровне описания информация более высокого уровня может и не присутствовать эксплицитно, имплицитно она присутствует, и ее наличие не может не приниматься в расчет. Например, высокий процент (скажем, до 10 % всего текста) употребления определенного артикля в английском тексте уже говорит о том, что этот текст представляет письменную форму ис-

пользования языка, что скорее всего это научный или технический текст, что следующее по частоте слово будет вероятнее всего предлог of, что текст насыщен именами существительными, что сообщение в тексте обладает высокой степенью определенности, категоричности, описательности и т.д. Иными словами, при описании текста даже на досинтактическом уровне не удается полностью отвлечься от его содержания, но это не мешает, а помогает его квантитативно-типологическому изучению.

Для дальнейшего развития концептуальной базы квантитативной типологии текста необходима разработка проблем семантического и прагматического уровней семиосиса, поиски эффективных и достаточно простых оценок смысловой информации, содержащихся в языковых сообщениях. Стратегия таких работ сформулирована Р.Г.Пиотровским (Пиотровский, 1975, с. 150 и след.); она реализуется на практике его учениками в группе "Статистика речи".

Детальное обсуждение всех вопросов, возникающих в процессе исследований, обобщение достигнутого опыта, развитие идей и методов системно-вероятностного и инженерно-лингвистического подхода в языкознании позволит еще более приблизиться к решению задач, выдвигаемых перед лингвистами в период НТР.

ЛИТЕРАТУРА

Алексеев П.М. Квантитативные аспекты речевой деятельности. – В кн.: Языковая норма и статистика. М., 1977.

Алексеев П.М. О системном характере лингвостатистики. – В кн.: Теория языка и инженерная лингвистика. Л., 1973.

Бенвенист Э. Общая лингвистика. Русск. пер. М., 1974.

Богданов В.В. Статистические концепции языка и речи. – В кн.: Статистика речи и автоматический анализ текста – 1972. Л., 1973.

Головин Б.Н. К вопросу о парадигматике и синтагматике на уровне морфологии и синтаксиса. – В кн.: Единицы различных уровней грамматического строя языка и их взаимодействие. М., 1969.

Косериу Э. Синхрония, диахрония и история. Русск. пер. – В кн.: Новое в лингвистике. Вып. Ш. М., 1963.

Моль А. Социодинамика культуры. Русск. пер. М., 1973.

Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. - В кн.: Вычислительная лингвистика. М., 1976.

Пиотровский Р.Г. Текст, машина, человек. Л., 1975.

Солнцев В.М. Язык как системно-структурное образование. М., 1971.

Соссюр Ф. де. Труды по общему языкознанию. Русск. пер. М., 1977.

Nauta D. The Meaning of Information. The Hague-Paris, 1972.

ON QUANTITATIVE TYPOLOGY OF TEXT

Pavel M. Alekseyev

S u m m a r y

The term "quantitative typology of text" is offered to denote typological studies based on linguistic, statistical (probabilistic), systemic, structural and semiotic concepts of language and speech. The quantitative typology of text aims at obtaining probabilistic and informational models that could describe and explain typological features of complex linguistic objects (language, sublanguage, functional style, idiolect) presented in text. The subject of it is to describe language and speech at all possible levels of their structural organization.

The notion of an average text leads to the idea of typology of text in general, not only of texts, though it does not exclude considering an individual text.

The axiomatics of the quantitative typology of text formulated as a set of postulates represents its theory. Methods of observation borrowed from linguistics, mathematical statistics, information theory and theory of systems constitute its technique. Its empirical basis is provided with material by statistical lexicography and linguistic statistics as a whole.

Some problems of the statistical interpretation of paradigmatics and syntagmatics of language and speech, the type-token ratio, etc. are given a certain amount of attention in the article.

ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА КАК НАУЧНАЯ ДИСЦИПЛИНА

В.М. Андрющенко

Термин "вычислительная лингвистика" появился лет 15 тому назад в печально знаменитом докладе ALPAC (ALPAC, 1966) и с тех пор этим термином обозначают область научной и конструкторской деятельности, покрывающую собой автоматическую обработку данных на естественном языке. Тот, кто считает себя работающим в этой области, знает, насколько она разнородна по методам, целям, средствам и концепциям, и в то же время понимает, что дальнейшее развитие этих работ в организационном, методологическом и концептуальном отношении связано и зависит от усилий, направленных на систематизацию понятий, выработанных внутри данной области и формирующих саму эту научную дисциплину, и на рассмотрение связи ее со смежными дисциплинами.

Ниже предлагается один из возможных подходов. Он основан на том представлении, что предмет научной дисциплины образует структуры, являющиеся абстракциями отношений предметов реальной действительности и изучаемые в соответствии с природой самих предметов или их отношений. Если мы согласимся с тем, что предметом языкознания являются лингвистические структуры, образуемые единицами и конструкциями языка, так же как предметом математики являются математические структуры, т.е. произвольные множества с определенными на них отношениями, а предметом информатики, соответственно, - вычислительные структуры (структуры данных, алгоритмов, аппаратуры и реализованных в ней схем), то мы можем легко представить себе, что могут существовать дисциплины, предметом которых является систематическое изучение отношений между структурами различных дисциплин. Такой дисциплиной является, например, математическая лингвистика, занимающаяся изучением соотношений между языковыми и математическими структурами, или программирование, занимающееся соотношением математических и вычислительных структур. По аналогии мы можем определить в первом приближении вычислительную лингвистику как дисциплину, занимающуюся изучением соотношений между лингвистическими и вычислительными структурами.

Мы можем назвать три круга проблем, конституирующих проблемную область вычислительной лингвистики:

I) изучение соотношений лингвистических и вычислительных структур;

2) лингвистический подход к проблеме коммуникации человека и ЭВМ;

3) автоматизация лингвистических исследований.

Заметим, что в вычислительной лингвистике, таким образом, мы устанавливаем членение проблематики, аналогичное тому, которое мы имеем в программировании:

I) изучение соотношений математических и вычислительных структур (теоретическое программирование);

2) создание инструментов для решения задач (практическое программирование, создание программного интерфейса для общения человека и ЭВМ);

3) автоматизация программирования (системное программирование).

Поэтому вычислительную лингвистику мы можем определить также (в известной мере метафорически) как лингвистическое программирование, отметив, что в языкознании вычислительная лингвистика занимает приблизительно то же место, что и программирование в вычислительной математике.

Но если программирование в вычислительной математике опирается на "ладно скроенный и крепко сшитый" фундамент теории чисел, теории численных методов, теории алгоритмов и схем программ, то в вычислительной лингвистике положение иное. Ни так называемое "традиционное языкознание", ни традиционная структурная лингвистика, ни теория порождающих грамматик не представляет прочной основы ни для построения систем данных, ни для построения алгоритмов лингвистического анализа и синтеза, хотя без сомнения в каждой из лингвистических школ мы можем найти и действительно находим предпосылки для построения соответствующих систем.

Построение вычислительной лингвистики как научной дисциплины может быть начато с рассмотрения аналогий между известными вычислительными и лингвистическими структурами.

Первую группу таких аналогий мы находим в структурах известных языков программирования.

Языки программирования и - шире - другие средства взаимодействия с ЭВМ (языки управления заданиями, системы директив оператора, проблемно-ориентированные языки, диалоговые языки и т.п.) суть знаковые системы, созданные для специаль-

ного класса ситуаций общения, а именно, для общения человека (программиста, пользователя) с программным обеспечением с целью управления вычислительным процессом. Общение есть двусторонний акт и предполагает поток информации не только от пользователя ("говорящего") к ЭВМ ("слушателю"), но и обратно - от ЭВМ к пользователю. Это позволяет предположить, что к ситуации общения с ЭВМ приложимы те же методы анализа коммуникативных ситуаций и на основе этого - семиотических функций языка, познавательная эффективность которых была доказана для случая коммуникации в обществе, в особенности на естественных языках. В коммуникации "человек - ЭВМ" мы видим не только создание специфической сферы профессионального общения, но и специфических для этой сферы средств, аналогичных социальным диалектам естественных языков (см. Вандриес, 1937). Следовательно, можно ожидать, что эти средства не только наследуют структурные черты естественных языков, но и могут быть развиты в направлении их сближения с последними, и более того - они естественным образом развиваются в этом направлении, что проявляется в смене поколений языков, каждое из которых наследует структурные черты предыдущего и обогащает их новыми выразительными возможностями, аналоги которых имеются в естественных языках.

Хотя термину "естественный язык" трудно дать единственное определение, можно указать на ряд свойств естественных языков, отличающих их от других систем коммуникации, таких как чистые коды, "языки" животных, системы сигнализации, "языки" искусства, письменности, языки общения с ЭВМ, формально-логические пазиграфии. Такими свойствами являются:

-словесно-речевой характер естественно-языкового общения, использующего преимущественно и изначально вокально-слуховой канал; это свойство предопределяет краткое время существования языкового сигнала, высокую степень избыточности и надежности речи (Хоккетт, 1970);

- первично и принципиально необходимым является наличие адресатов ("говорящих") и адресатов ("слушающих") речи, осуществляющих рассеянную "передачу" и направленный "прием"; говорящие и слушающие чередуются ролями, но их отношение к акту речи несимметрично: говорящий является одновременно и слушающим, обратное же неверно; как структура сообщения, так и структурные свойства соответствующей языковой системы складываются как компромисс между интересами говорящих и слушающих (Успенский, 1967);

– естественно-языковая система не наследуется генетически, но овладение одной из систем предопределяет способность обучаться другим системам, причем овладение каждой новой системой повышает способность к обучению следующим системам (Хоккетт, I970);

– между континуумом содержания и континуумом речи лежит система дискретных и семантизированных языковых означающих, состоящая минимум из трех классов знаков: десигнаторов, форматоров и субститутов (Вейнрейх, I970);

– по отношению к внеязыковому содержанию языковой знак произволен, но внутри системы данного языка он, как правило, мотивирован (Хоккетт, I970);

– языковые знаки обладают семантическим тропизмом: возможны, а иногда и необходимы употребления знаков в переносном значении (Хоккетт, I970);

– среди знаков естественного языка находятся такие знаки, которые позволяют строить высказывания не только о самом данном языке, но и о других знаковых системах. Естественный язык обладает универсальной метаязыковой функцией (Хоккетт, I970);

– энергетические затраты на производство речи не зависят от важности сообщаемого для коммуникантов и нерелевантны для возможности общения. Означаемые внешнего мира могут как наличествовать в ситуации общения, так и отсутствовать в ней. Осмысленность выражений естественного языка не зависит от формальной логики (Хоккетт, I970);

– система естественного языка является открытой: в ней возможны и неограничены появления как новых означающих, так и новых значений, употребляемых и выражаемых в рамках данной грамматической системы. Грамматические системы со временем изменяются (Хоккетт, I970);

– структурная организация естественного языка семиотически дуальна: в ней взаимодействуют две подсистемы – кинематическая подсистема речедвигательных структурных единиц и плерематическая подсистема структурных единиц внутриязыкового плана содержания (Хоккетт, I970);

– каждая из взаимодействующих семиотических подсистем грамматически организована в плане парадигматики и синтагматики;

– парадигматический план образуется системой грамматических категорий, значения которых должны обязательно выражаться в речи;

– в синтагматическом плане действуют два главных механизма сочленения знаков – соединение и вставление (Вейнрейх, 1970). На число единиц соединения и вставления наложены статистические ограничения (Ингве, 1965), а на правила соединения и вставления – структурные ограничения (требование проективности структуры на вектор означающих). Длины подцепочек, образуемых проекцией структуры на вектор означающих, статистически согласованы по уровням вложения (Андрющенко, 1966);

– программа понимания языкового выражения выражена в нем самом – его структурой (правилами свертки) и значениями грамматических категорий (операциями свертки);

– в зависимости от способов задания правил и операций свертки языки образуют морфологические (лингвистические) типы. Существует такая последовательность типов языков, что каждый предыдущий может быть принят в качестве грамматического метаязыка для всех последующих. Грамматическим метаязыком для всех языков всех типов является аморфный язык (Успенский, 1965).

Языки программирования и взаимодействия с ЭВМ в большей степени, чем другие семиотические и коммуникативные системы наследуют указанные свойства естественных языков.

Перечисленные свойства естественных языков непосредственно связаны с основными объектами коммуникативной ситуации: говорящим, слушающим, текстом (сообщением, репликой), его содержанием, системой языка (кодом) и каналом связи. Считается, что активным элементом, возбудителем коммуникативной ситуации является говорящий, установка которого на выражаемой в сообщении его отношение к другим объектам коммуникативной ситуации, называется лингвистической (семиотической) функцией. Различаются следующие лингвистические функции: металингвистическая (установка на выражение отношения к системе языка), когнитивная (установка на выражение мыслительного содержания), символизирующая (установка на представление в речи объектов действительности или мысли), репрезентативно-номинативная (установка на именование объектов действительности), изобразительная (установка на непосредственное отражение в сообщении значимых свойств объектов содержания), общекоммуникативная (установка на связность акта коммуникации), частно-коммуникативные функции (установка на выражение эмоционального состояния, "настройку" адресата, поддержание контакта и т.п.) (Якобсон, 1960). Различные знаковые системы и системы коммуникации могут быть характеризованы наличием

специальных средств для выражения этих лингвистических функций. Так, в классификационно-индексационных языках информационно-поисковых систем нет средств для выражения металингвистической, изобразительной, общекоммуникативной и частно-коммуникативных функций. В системах сигнализации в языках животных нет средств для выражения металингвистической функции, а в первой лишь в слабой степени представлены средства для выражения репрезентативно-номинативной, изобразительной и коммуникативных функций. Языки жестов и искусств, системы письменности и естественные языки обладают всеми перечисленными функциями, причем первые - лишь в ослабленной степени металингвистической функцией. Формально-логические языки и пазиграфии не обладают коммуникативными и в очень малой степени - репрезентативно-номинативными и изобразительными средствами. Языки программирования и взаимодействия с ЭВМ наиболее близки к естественным языкам по набору средств для выражения всех перечисленных функций.

В языках программирования мы встречаем немало конструкций, аналогичных конструкциям естественных языков. Так, константы аналогичны именам собственным, атрибуты типов данных - показателям именных классов, встроенные функции - словообразовательным категориям и несинтаксическим падежам, размерности массивов - категории числа, параметры, переменные типа метки, указатели и имена собственные - местоимениям и некачественным наречиям, разделители - частицам естественных языков, операции - синтагматическим отношениям, изображения и форматы - иконическим знакам и т.д.

Естественный язык устроен таким образом, что между интересами говорящих и слушающих устанавливается компромисс (Успенский, 1967), результатом которого является выработка специфических конструкций, аналоги которых мы также встречаем в языках программирования. В целом языки программирования развиваются в направлении от "языков слушающих" к "языкам говорящих", однако, развитие специфических черт "языков говорящих" приводит к необходимости развивать коррелирующие с ними средства "языков слушающих". Так, увеличение количества типов данных коррелировано с увеличением количества встроенных функций (аналог словообразовательных категорий и несинтаксических падежей), увеличение количества элементарных операций (аналог синтаксических отношений) коррелировано с усложнением правил старшинства и уподобления типов данных (аналог синтагматического согласования).

Развитие в языках программирования аппарата передачи параметров, встроенных функций, согласования типов данных аналогично развитию в естественных языках системы местоимений, падежей и синтаксических связей – типичных средств компромиссов между интересами говорящих и слушающих (Успенский, 1967). Языки программирования могут быть приближены по своей структуре к естественным языкам путем отказа от противоречащих структуре естественных языков конструкций и развития средств, аналогичных средствам естественных языков. Следующие пути сближения языков программирования с естественными языками представляются реальными:

- развитие возможностей обходиться без явных присваиваний (Цейтин, 1974);
- развитие возможностей обходиться без явной передачи управления (Цейтин, 1974);
- развитие средств расширения языка посредством новых определений и введения метаязыкового уровня (Цейтин, 1974);
- расширение полисемии, синонимии, омонимии и семантического тропизма;
- развитие аппарата умолчаний;
- морфологизация структур идентификаторов;
- введение неопределенного типа данных и аппарата его динамического доопределения;
- включение в языки аппарата исчисления высказываний и предикатов;
- развитие изобразительных средств (расширение функций изображений);
- развитие более естественных выражений для категории времени;
- развитие средств установления и поддержания контакта (аналог конативной и фатической частично-коммуникативных функций);
- развитие неиндексных средств выбора из агрегата данных;
- развитие причинно-следственных конструкций;
- развитие средств модальности времени выполнения;
- развитие средств переключения с диалекта на диалект и др.

В языках программирования можно обнаружить сходные с естественными языками типологические соотношения. Аналогом аморфных языков является язык СИМПОЛИЗ-64, задуманный как метаязык (язык-посредник) для транслятора с АЛГОЛа-60. Машин-

ные языки аналогичны инкорпорирующим языкам, а процедурно-ориентированные языки типа АЛГОЛа, ФОРТРАНа, ПЛ/I, говоря лингвистическим языком, занимают промежуточное положение между инкорпорирующими и агглютинативными языками. Командные и процедурно-ориентированные языки приблизительно так же относятся к сентенциальным языкам, как аморфные и инкорпорирующие языки относятся к агглютинативным и флективным языкам. Развитие языков от аморфных по направлению к флективным, инициированное интересами говорящих, приводит к конструкциям, менее выгодным для окружающих в одних отношениях (усложнение парадигматики), но более выгодным в других (упрощение синтагматики), что и является следствием компромисса между интересами говорящих и слушающих (Успенский, 1965, 1967). Видимо, языки программирования "естественно" развиваются по законам естественных языков (Цейтин, 1974). Более поверхностные примеры этой же тенденции - диалектная дифференциация и противостоящая ей тенденция - формирование койне-образных языков (языковые стандарты и языки публикаций).

Лингвистическое изучение языков программирования помогает глубже понять и устройство естественных языков. Основным методом здесь, по-видимому, является проведение аналогии между языковыми и вычислительными структурами. Языковые структуры являются предельно сложными типами вычислительных структур. Поэтому "программистский" подход к изучению структур естественных языков может быть полезен для развития теории и практики программирования, в особенности в области обработки данных на естественном языке и реализации диалоговых систем.

Грамматическая структура предложения на естественном языке может мыслиться как программа понимания (= вычисления) его смысла относительно баз данных коммуникантов. В этой грамматической структуре роль переменных играют имена объектов, относительно которых строится высказывание, обладающее данной структурой; роль операций (вычисляемых функций) выполняется синтаксическими и лексическими функциями, выраженными в синтаксических связях, аффиксах и частицах; роль предикатов выполняют специальные имена, называющие качества, состояния, положения и другие классы характеристик объявленных переменных.

Вычисление смысла предложения в такой интерпретации состоит в восстановлении по тексту высказывания его грамматической структуры и в выполнении программы, генерируемой этой структурой.

Грамматическая структура закодирована формой языкового выражения, которое представимо в виде двунаправленного мультисписка. В этом мультисписке адреса связи кодируются грамматическими формантами, возможно, нулевыми. Автоматическая обработка текста на естественном языке может мыслиться как последовательное декодирование списковой структуры, восстановление по ней грамматической структуры, генерация программы понимания, выполнение этой программы, вычисляющей смысл.

Если естественным вычислительным представлением выражения на ЕЯ является список и, соответственно, основной типовой процедурой вычисления - преобразование списка (трансформация), то для единиц языкового уровня таким естественным представлением является программа (алгоритм), точнее набор алгоритмов, каждый из которых порождает структуру или семейство структур, включающих данную единицу в сеть заданных лингвистических отношений. Так, для лексемы могут быть предложены следующие типы алгоритмов ее вычислительного "описания":

а) алгоритм перечисления ее грамматических форм, причем параметрами такого алгоритма является набор морфологических признаков;

б) алгоритм перечисления синтагм, причем параметрами такого алгоритма является набор ее семантико-синтаксических признаков и признаков "целевой" синтагмы;

в) алгоритм построения семантического поля, причем параметрами этого алгоритма является набор семантических признаков лексем;

г) алгоритм построения словообразовательного поля, причем параметрами для такого алгоритма является набор лексико-семантических признаков лексемы.

Такой подход приводит к формулированию задачи автоматического словаря как его макрогенерации, где описание лексемы есть макрокоманда, расширяемая в программы, реализующие алгоритмы описания лексемы.

В данном случае устанавливается аналогия между схемой словарной статьи (схемой, точнее подсхемой - в терминологии баз данных - лингвистического описания) и макроопределением, состоящим из прототипа (перечня параметров) и макротекста, содержащего параметры, позволяющие выдавать объектный текст, образуемый путем подстановки параметров и комбинацией кусков макротекста в единый текст. Такой макротекст может описывать как структуру данных (например, в терминах зон и полей сло-

варя), так и структуру программы, генерирующей в месте ее вызова нужные сведения о лексеме.

ЛИТЕРАТУРА

Андрющенко В.М. О взаимозависимости между средними длинами лингвистических единиц разных уровней. – Тезисы конф. по проблемам изучения универсальных и ареальных свойств языков. АН СССР, Ин-т народов Азии. М., 1966.

Вандриес Ж. Язык. – М.: Гос. соц.-экон. изд-во, 1937, с.227-248.

Вейнрейх У. О семантической структуре языка. – В кн.: Новое в лингвистике. Вып. У. – М.: Прогресс, 1970.

Ингве В. Гипотеза глубины. – В кн.: Новое в лингвистике. – М.: Прогресс, 1965.

Успенский Б.А. Структурная типология языков. – М.: Наука, 1965.

Успенский Б.А. Проблемы лингвистической типологии в аспекте различения "говорящего" (адресанта) и "слушающего" (адресата). – В кн.: To Honor Roman Jacobson. – Mouton, The Hague, 1967.

Хоккетт Ч. Проблема языковых универсалий. – В кн.: Новое в лингвистике. Вып. У. – М.: Прогресс, 1970.

Цейтин Г.С. Черты естественных языков в языках программирования. – В кн.: Машинный перевод и прикладная лингвистика. Вып. 17. – М.: Изд-во МГПИИЯ им. М.Тореза, 1974.

ALPAC REPORT "A report by the Automatic Language Processing Advisory Committee", Language & Machines. Computers in translation and linguistics. Nat. Acad. of Sc., Washington, 1966.

Jacobson, R. Linguistics and Poetics. – In: Style in Language, ed. by A. Sebeok. Cambridge, Mass., 1960.

COMPUTATIONAL LINGUISTICS AS A SCIENTIFIC DISCIPLINE

Vladislav M. Andryushtshenko

S u m m a r y

Computational linguistics is defined as a scientific discipline studying relations between linguistic and computational structures and dealing with different representations of linguistic structures in a computer. Computational linguistics has three main aspects: theoretical (relations between linguistic and computational structures), experimental (automatisation of linguistic research) and applicational (a linguistic approach to the problem of communication between man and computer). In the present paper we discuss some linguistic features of programming analogies in the approach to language processing.

СИММЕТРИЯ В ПРЕДИКАТИВНЫХ ПАРАХ

Н.Н. Дарчук

В нашем исследовании принцип симметрии применен к свойствам соединения подлежащих и сказуемых в тексте. В этом случае симметрия проявляется в позиции сказуемого: в обобщенной форме она выражена пре-/постпозицией сказуемого по отношению к подлежащему и в способе выражения сказуемого.

Материалом для исследования послужили короткие рефераты научных текстов по кибернетике, состоящие из трех или четырех предложений (всего 229 рефератов).

Выявление симметрии осуществляется при помощи двух основных форм движения: не зеркального, реализующегося в виде разного рода поворотов или переносов, и зеркального движения, или отражения.

Например, если в реферате, состоящем из четырех предложений, в которых сказуемое в первом предложении находится по отношению к подлежащему в препозиции, во втором и третьем предложении - в постпозиции и в четвертом - в препозиции, обозначить препозицию сказуемого через I, а постпозицию - через II, то получим такой ряд: I II II I. В нем проведем ось через середину. При повороте левая от центра позиция II совпадает с правой от центра II, левая позиция I совместится с правой I: I II II I. Это зеркальный вид симметрии. Зеркальную симметрию мы получим также и в том случае, когда ось симметрии будет включать предикативную пару, являющуюся, таким образом, центром отсчета симметрии.

Реферат научного текста представлен у нас набором последовательностей пре- и постпозиции сказуемого по отношению к подлежащему во всех имеющихся в данном реферате предложениях.

Как правило, повторяются такие позиции:

I II - II I (5 раз), II I - I II (2 раза), I - II - I (9),

II - I - II (3), I II - II - II I (2), I I - II - I I (4),

II II - I - II II (2), I II - II I - I II (2).

Введем понятие такта, который может быть равен одной предикативной паре (I или II) или больше одной предикативной пары (I II), (I II I).

Такт создает ритмичную симметрию, которая может быть простой или непростой. При простом ритме возможны две ситуации:

I. Такт равен одной предикативной паре, что создает монотонную симметрию: например, I I I (f = 19), I I I I (f = 11), II II II (f = 4), II II II II (f = 4), II II II II II (f = 1), I I I I I (f = 8).

2. Такт равен нескольким предикативным парам. Например, I II - I II (f = 3), II I - II I (f = 5).

При повторении такта в границах реферата такт может расширяться, когда к нему прибавляется пре- (I) или постпозиция (II) сказуемого, или сужаться, когда такт уменьшается на пре- (I) или постпозицию (II) сказуемого.

Например,

I II - I I II (3), I II - I II II (3), I I II - I II (2).

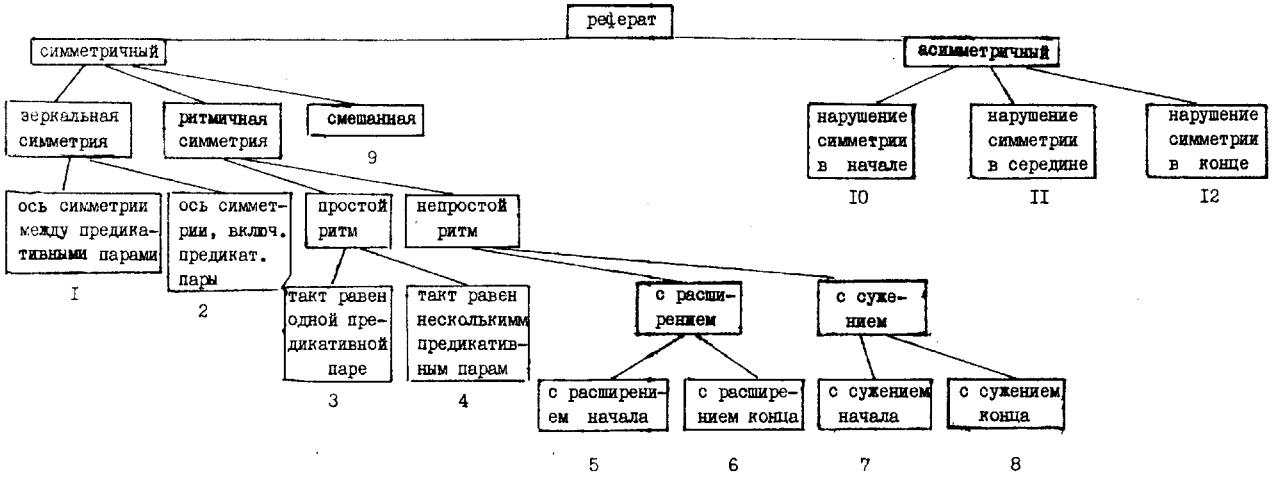
Следует остановиться еще на одном виде симметрии, который мы назовем смешанным. Этому виду свойственна и ритмичная симметрия (такт повторяется), и зеркальная (при повороте левая часть ряда совпадает с правой), и монотонная (где такт равен одной предикативной паре), причем один и тот же такт может участвовать в разных видах симметрии. Например, монотонно-ритмичная симметрия II I - II II II - II I : монотонная симметрия в середине ряда обрамляется ритмичной симметрией по краям. Или зеркально-монотонная: I II - II I - I I I : начинается ряд с зеркальной симметрии и заканчивается монотонным повторением препозиции (I).

Симметрии противопоставлена асимметрия предикативных пар, которая представлена у нас тремя видами, когда:

I. симметрия нарушена в начале I - I II I (пример зеркальной симметрии с осью через предикативную пару с нарушением симметрии в начале);

2. симметрия нарушена в середине I II I I I (пример монотонного повторения препозиции I с "вклиниванием" постпозиции II);

3. симметрия нарушена в конце ряда : I II II II I - I. Зеркальная симметрия, в которой ось проходит через предикативную пару, нарушена препозицией I.



Примечание: Порядковые номера обозначают группы симметрии.

Рис. I. Виды и группы симметрии (для позиции сказуемого и способа выражения).

Таблица I

Распределение позиций сказуемого по видам и группам симметрии

Кол-во предик.	симметричный								асимметричный			
	зеркальная симметрия	ритмичная симметрия							смешанная симметрия	с наруш. симметр. в начале	наруш. симметр. в серед.	наруш. симметр. в конце
	ось между предикат. парами	ось сим. включает предикат. пару	простой ритм		непростой ритм							
			где такт равен одной предик. паре	где такт равен нескольким пред. парам	с расширением		с сужением					
					начала	конца	начала	конца				
	I	2	3	4	5	6	7	8	9	I0	II	I2
3	–	I2	23	–	–	–	–	–	–	II	–	5
4	7	–	I5	8	–	–	–	–	6	I7	–	I3
5	–	8	9	–	3	3	3	–	4	8	3	8
6	3	2	–	–	I	I	I	I	I	4	4	8
7	–	–	–	–	2	–	–	–	8	–	2	4
8	–	–	–	–	–	–	–	–	I	–	I	3
9	–	–	–	–	–	–	–	–	2	–	–	–
I0	–	–	–	–	–	–	–	–	–	I	–	I
II	–	–	–	–	–	–	–	–	2	–	–	–
I2	–	–	–	–	–	–	–	–	–	–	–	–
I5	–	–	–	–	–	–	–	–	I	–	–	–
	I0	22	47	8	6	4	4	I	24	4I	I0	42

Предикативным парам исследуемых рефератов, в основном, свойственны два вида симметрии: зеркальная и ритмичная. По признаку наличия или отсутствия того или иного вида симметрии можно выделить 12 групп (см. рис. I). Дерево дает представление о них, а также показывает те зависимости, которые существуют между группами в пределах одного вида симметрии.

Теперь важно установить, какую долю составляют асимметрично построенные ряды среди всего их инвентаря. Определим, какой вид или группа симметрии наиболее характерны для нашего материала.

В результате анализа было установлено, что из 220 рефератов[I] 126 симметричны, что составляет 57,3 % от количества рефератов, и 94 - асимметричны (см. табл. I).

Количество предикативных пар в реферате из 3 или 4 предложений может быть большим, достигая 13 и образуя довольно длинные предикативные ряды.

Количество симметричных предикативных рядов зависит от их длины. Чем длиннее ряды, тем меньше среди них симметричных. Например, практически отсутствует симметрия у рядов длиной в 8, 9, 10, 11, 12, 13 предикативных пар. В менее длинных рядах преобладает, в основном, ритмичная симметрия с простым ритмом, где такт равен одной предикативной паре (монотонное распределение), таких рефератов 47, причем 23 из них состоят всего из трех предикативных пар. Все 47 рефератов, относящиеся к третьей группе, характерны монотонным повторением либо препозиции сказуемого (I), либо постпозиции (II) (соответственно, в 38 и 9 рефератах).

Среди симметричных на втором месте по численности рефераты с зеркальным видом симметрии, причем таких, в которых ось симметрии включает предикативные пары, в два раза больше, чем тех, где ось симметрии проходит между предикативными парами.

Симметричных, но с непростым ритмом, 15 рефератов, причем больше с расширением начала - таких шесть рефератов, с сужением конца - всего один.

Интерес представляет девятая группа со смешанным характером симметрии. Во-первых, потому, что в этой группе оказались рефераты с наибольшей длиной рядов пре- и постпозиции сказуемого. Например, реферат, состоящий из 4 предложений,

---

[I] Из 229 рефератов 9 не исследовались, поскольку в них отсутствовали либо подлежащее (в 7 случаях), либо сказуемое (в 2 случаях).

содержит 13 предикативных пар с таким рядом пре- и постпозиции сказуемого: I I I I – II II – I I II – I I II. В нем монотонное повторение препозиции сказуемого сменяется монотонным повторением постпозиции сказуемого и завершается повторением такта I I II. Это пример монотонно-ритмичной симметрии.

Во-вторых, рефераты этой группы состоят из двух видов тактов в монотонно-ритмичной симметрии: I I I II II II, II II I I I I I, I I II II II II (всего 14 рефератов); ритмичная с сужением конца I II II – I II – I II; или двух или даже трех видов симметрии. Например,

монотонно-ритмичная:

I I I I – II II II – I I II – I I II;

ритмично-монотонная с расширением конца такта:

I II – I II II – I I I;

зеркально-монотонная:

I II – II I – I I I, I II – I II – II II II,

II II I – II II II – I II II, I II – II I – I I I,

II I – II I – I II – II II II;

зеркально-ритмичная с расширением начала такта:

II II I – I II II – I II – I I II;

ритмично-зеркальная симметрия:

I II – I II – II II II – II I – I II.

Рефераты девятой группы со смешанным характером симметрии находятся на стыке между симметрией и асимметрией.

Значительную часть (около 43 % рефератов) составляют асимметричные ряды предикативных пар (10, 11, 12 группы), причем нарушается симметрия, как правило, либо в начале, либо в конце ряда. Чаще всего это ряды, состоящие всего из четырех предикативных пар, где нарушается ритмичность или зеркальная симметрия. Таких в десятой группе тринадцать из 17, например, I – I II I, II – II I II, II – I I I; в двенадцатой группе шесть из 13: I II I – I; II I II – II, в остальных нарушается монотонность.

Во всех трех асимметричных группах наблюдается отклонение от симметрично построенных рядов всего на одну пару (кроме двух случаев). Не исключено, что при больших размерах рефератов и соответственно при увеличении предикативных рядов некоторые из асимметричных могли бы стать симметричными и наоборот. Во всяком случае наличием довольно значительного

числа асимметрично построенных рядов язык рефератов избегает избыточной симметрии в соположении сказуемого и подлежащего, чтобы изобилие симметрии не сковывало (пусть подсознательно) внимание читателя.

Проанализировав симметрию в позиции сказуемого по отношению к подлежащему, на том же материале (220 рефератов) исследуем наличие симметрии в способе выражения сказуемого. Будем квалифицировать способ выражения сказуемого по основному семантическому компоненту. Так, простое и составное глагольное сказуемое обозначим – V, именное составное и сложное сказуемое, выраженное кратким причастием – P, именное составное сказуемое, выраженное наречием – D, именное составное и сложное, а также именное сказуемое с нулевой связкой, где присвязочное слово существительное – N, именное составное со связкой, где присвязочный член выражен кратким прилагательным – A.

Аналогично ряду сказуемых по позиции (пре- и пост-) составлен ряд сказуемых по способу выражения. Возьмем пример: "Излагаются принципы, положенные в основу разработки аппаратных компонентов и математического обеспечения проблемно-ориентированного процессора, работающего в комплексе с универсальной ЭВМ. Этот спецпроцессор предназначен для использования при решении широкого круга задач, реализация алгоритмов которых основана на цифровых методах анализа результатов наблюдений. К их числу относятся задачи экономики и планирования производства, астрофизики, сейсмической разведки полезных ископаемых: метеорологии, биомедицины и многие другие, а также класс проблем, сводящихся к задаче распознавания образов".

В этом реферате, состоящем из трех предложений, в первом и третьем предложении по одному сказуемому, которое находится в препозиции (I) к подлежащему и в обоих случаях выражено глаголом (V). Во втором предложении этого реферата два сказуемых, оба находятся в постпозиции (II – II) к подлежащему и выражены кратким страдательным причастием (P – P). Мы имеем два ряда: первый составлен по позиции сказуемого, второй по способоу выражения:

1. I II II I
2. V P P V

Симметрия в рядах сказуемых по способу выражения устанавливалась с помощью того же дерева, отражающего виды сим-

Таблица 2

Распределение способа выражения сказуемого по видам и группам симметрии

Кол-во предик. пар	симметричный								асимметричный			
	зеркальная симметрия	ритмичная симметрия							смешан. симмет-рия	наруш. симметр. в начале	наруш. симметр. в серед.	наруш. симметр. в конце
	ось между предикат. парами	ось. сим. включает предикат. пару	простой ритм		непростой ритм							
			где такт равен одной пред. паре	где такт равен нескольким предикат. парам	с расширением		с сужением					
					начала	конца	начала	конца				
	I	2	3	4	5	6	7	8	9	I0	II	I2
3	-	I0	25	-	-	-	-	-	-	4	-	I0
4	5	-	20	I	-	-	-	-	3	I0	-	I8
5	-	4	9	-	-	2	2	-	I3	9	I	7
6	2	-	5	I	-	-	-	-	4	3	4	3
7	-	I	I	-	-	-	-	-	4	2	-	3
8	-	-	3	-	-	-	-	-	-	-	-	-
9	-	-	I	-	-	-	-	-	I	-	-	-
I0	-	-	-	-	-	-	-	-	I	-	I	-
II	-	-	-	-	-	-	-	-	-	-	I	-
	7	I5	64	2		2	2		26	28	7	4I

метрии (рис. I), затем ряды были сгруппированы по признаку наличия или отсутствия в них симметрии (см. табл. 2).

Из 220 рядов более половины (53 %) оказались симметричными с явным преобладанием ритмичной симметрии с простым ритмом, где такт равен одной предикативной паре (третья группа - 54 %). Причем, чем короче ряд, тем чаще наблюдается в нем симметрия. В третьей группе замечена еще одна закономерность: ее ряды представляют собой монотонное повторение то ли глагола (V), то ли краткого причастия (P), причем 54 ряда из 64 полностью "глагольных", а I0 - "причастных", которые оказались только трех- или четырехчленными. "Глагольные" же ряды насчитывают 5, 6, 7, 8 и даже девять глаголов.

На третьем месте по частоте ряды с зеркальной симметрией, причем среди них больше трехчленных, например: V - P - V

P - N - P, V - D - V, P - V - P и т.д.

В отличие от рядов, составленных по позиции сказуемых, здесь отсутствуют примеры ритмичной симметрии с непростым ритмом с расширением начала (пятая группа) и с сужением конца (восьмая группа).

Примерно I2 % от общего количества рефератов составляет девятая группа, т.е. смешанная. Преобладает в ней зеркально-монотонная симметрия (I2 случаев):

P - V - P P P, V - D - V - V V, P V - V P - P P P P P P

или монотонно-зеркальная (7 случаев).

В четырех случаях наблюдается монотонно-монотонная, напр., P P P - V V V и в одном случае монотонно-ритмичная симметрия: V V V V V - P N - P N.

Частота асимметричных рядов почти такая же, как и симметричных. Полностью асимметричными оказались 26 рядов, которые не вошли ни в I0, ни в II, ни в I2 группы. Среди асимметричных наиболее многочисленной оказалась двенадцатая группа, в которой симметрия нарушалась в конце ряда (4I ряд). Например,

V V V V - P (на один предикат)

V P V - V P V - P (на один предикат).

Затем следует десятая группа (28 рядов), где нарушается симметрия в начале ряда. Например,

P - P - V - P (на один предикат)

P - V V V V (на один предикат).

Сравнительно небольшая одиннадцатая группа с нарушением

симметрии в середине ряда. Например,

V V V - A - V V V на один предикат,

V V V V P V V V V V на один предикат.

Сравнение рядов предикатов по их позиции и по способу выражения на материале 220 рефератов показало, что в 73 рефератах оба ряда симметричны, причем в 40 рефератах (33,I %) полностью совпадают группы семметрии. В 99 рефератах (45 %) симметрия частичная: либо по позиции, либо по способу выражения. 48 рефератов (2I,8 %) полностью асимметричны по позиции и по способу выражения.

Как показывает анализ, предикативным рядам свойственна ритмичная и зеркальная симметрия. При этом симметричны,в основном, рефераты с тремя-четырьмя сказуемыми, где преобладают простые предложения. Однако чрезмерность симметрии нивелировала бы экспрессивные свойства языка и сделала бы реферат монотонным для читающего. Поэтому даже среди коротких рефератов почти 22 % асимметричных. Т.е. соотношение между симметрией и асимметрией в синтаксисе является тем принципом, который лежит в основе функционирования элементов системы языка и речи (Хоккетт, I970).

Л И Т Е Р А Т У Р А

Хоккетт У.Ф. Проблема языковых универсалий. - В кн.: Новое в лингвистике, У. I970.

SYMMETRY IN THE ANALYSIS OF PREDICATES

Natalya P. Darchuk

S u m m a r y

In this study the principle of symmetry is applied to the analysis of predicates in scientific abstracts. Symmetry is revealed in both pre- or post-position of the predicate as to the subject and in the means of its expression.

Types and groups of symmetrical structures are established and statistically characterized.

It was noted that the amount of symmetrically or rhythmically organized predicate sequences lessens with the growth of their length.

АВТОМАТИЧЕСКИЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ПОЭТИЧЕСКОГО ТЕКСТА

А.В. Зубов

Полезность использования квантитативных методов и современных ЭВМ в стиховедении сейчас не вызывает сомнения (Гаспаров, 1974, 18-38; Григорьев (ред.), 1973, 23-25). С помощью машин составляются словари и конкордансы к отдельным произведениям и собраниям произведений, словари рифм; проводятся автоматический анализ ритмической организации стихов и анализ эмоционального фона стихотворного текста, и целый ряд других работ (см., например, зарубежные работы, указанные в исследовании (Григорьев (ред.), 1973, 23-25) и советскую библиографию: Гиндин, 1978; Ралько, 1977, 144-208).

Однако такие исследования проводятся, как правило, фрагментарно. Исследователь не получает полной информации о статистической структуре всех произведений конкретного автора на разных уровнях: синтаксическом, лексическом, морфологическом и ритмическом.

В нашей работе сделана попытка провести такой анализ на материале поэтических произведений С.Есенина (Есенин, 1966).

На первом этапе работы был проведен с помощью ЭВМ статистический анализ употребительности лексики в поэтических произведениях С.Есенина. Были получены алфавитный и частотный списки словоформ (Гайдукова, Зубов, 1975). В словарь вошло 16 786 словоформ, выбранных из текста длиной в 56347 словоупотреблений.

На следующем этапе работы каждой словоформе "вручную" присваивался код, содержащий русские буквы и десятичные цифры. Последние использовались для указания общего числа слогов и места ударного слога в словоформе. В процессе анализа различались 17 классов слов: существительное (код - С), глагол в личной форме (Г), глагол в инфинитиве (И), прилагательное (А), прилагательное в краткой форме (Л), наречие (Н), числительное количественное (Ч), числительное порядковое (Я), имя собственное (Б), причастие (П), деепричастие (Д), предлог (Р), частица (Ц), местоимение (М), междометие (Ж), союз (Ю), модальное слово (Ф).

Помимо этого, для каждой словоформы указывались грамматические значения рода, числа, падежа и времени.

В итоге каждая словоформа получала код, состоящий из 6 знаков. Например, существительное ВЕСНОЙ и глагол РАСЦВЕЛА получали следующие коды:

ВЕСНОЙ - СЖЕТ22, РАСЦВЕЛА - ГЖЕП33

Здесь: С - существительное, Г - глагол, Ж - женский род, Е - единственное число, Т - творительный падеж, П - прошедшее время. Число 22 у первого слова обозначает, что в нем 2 слога и ударение падает на второй слог. Аналогично расшифровывается и число 33 у второго слова.

В процессе расстановки ударений в словоформах алфавитного списка безударными считались все односложные служебные слова и односложные местоимения. Предлогам типа "В", "К" и им подобным словарным единицам, не образующим слогам, приписывался код "00".

В дальнейшем алфавитный словарь вместе с кодами перфорировался на перфокарты и вводился в память ЭВМ. Таким же образом в ЭВМ вводились и все поэтические произведения С. Есенина.

Все нижеописываемые результаты получены на ЭВМ ЕС-1020 с помощью одной основной программы, состоящей из 1117 операторов языка ПЛ/I и 10 вспомогательных программ, написанных на том же алгоритмическом языке.

Принцип действия основной программы весьма прост: наложение каждой строки произведений автора на закодированный алфавитный словарь автора и подсчет соответствующих буквенных и цифровых кодов слов[1] каждой строки.

Рассмотрим получаемые при этом результаты на примере первой строфы известного стихотворения С.Есенина (Есенин,1967, 161):

Каждый труд благослови, удача!
Рыбаку - чтоб с рыбой невода,
Пахарю - чтоб плуг его и кляча
Доставали хлеба на года.

В результате замены каждого слова строки кодом соответствующего класса слова, "отбираемого" машиной в алфавитном словаре автора, ЭВМ выдает следующие структурные строки этой строфы на уровне классов слов (I):

---

[1] Здесь и далее под словом понимается словоформа.

МСГ,С!		МСГС	
С-ЮРСС,	(I)	СЮРСС	(2)
С-ЮСМЮС		СЮСМЮС	
ГСРС.		ГСРС	

Рядом с каждой такой синтаксической формулой машина выдает структуру строки на уровне классов слов, но без знаков пунктуации (2).

Вслед за этим каждое слово строки заменялось информацией о количестве слогов в слове и месте ударного слога в нем. В итоге на печать выдавались структура строки по способу распределения в ней слогов всего и ударных слогов (3).

2I	II	44	32			
33	I0	00	2I	33		(3)
3I	I0	II	22	I0	2I	
43	2I	I0	22			

Далее, на основе информации (3) для каждой строки строилась ритмическая формула строки. Место каждого очередного ударного слога строки $R(k)$ подсчитывалось машиной по следующим формулам:

$$R(1) = U(1), \quad \text{если} \quad U(1) \neq 0$$

$$R(k) = \sum_{i>1}^{k-1} S(i) + U(k), \quad \text{при } k>1 \text{ и } U(k) \neq 0,$$

где $U(i)$ - место ударного слога в слове строки с номером i;
$S(i)$ - число слогов в слове строки с номером i.

Для рассматриваемого примера ритмические формулы строк имели следующий вид (4):

0I	03	07	09	
03	05	09		(4)
0I	05	07	09	
03	05	09		

Помимо этого, для возможности проведения в дальнейшем анализа связи длины строки (в слогах) и типа ритмических структур строк, для каждой строки выдавалась ритмическая формула строки в зависимости от длины строки (5):

I0	0I	03	07	09	
09	03	05	09		(5)
I0	0I	05	07	09	
09	03	05	09		

Здесь первые две цифры каждой строки обозначают длину строки.

Все пять типов информации (I), (2), (3), (4), (5) по всем строкам объединялись во внешней памяти машины. После окончания обработки всех I26I0 строк произведений С.Есенина, ЭВМ классифицировала и суммировала эту информацию по различным признакам.

Так, наиболее употребительными структурами строк на уровне классов слов, оказались следующие (табл. I):

Таблица I

Наиболее употребительные структуры строк в произведениях С.Есенина

№№ пп.	Структура строки	Абсолютная частота F	Относительная частота
I.	АС.	76	0,00603
2.	ГАС.	62	0,00492
3.	ГС.	59	0,00468
4.	РАС.	44	0,00349
5.	ГСС.	42	0,00333
6.	АС.	4I	0,00325
7.	РАС.	4I	0,00325
8.	ГРС.	40	0,003I7
9.	ГС.	40	0,003I7
I0.	РСАС.	34	0,00270

Всего таких структур оказалось 8890. Из них первые 50 структур (включая F = I3) покрывают I2I8 структур (9,6 %), первые I00 (включая F = 8) - I709 структур (I3,6 %).

Несколько отличающиеся данные получены для структур строк без знаков. Начало соответствующего частотного списка приведено в таблице 2.

Таблица 2.

Наиболее употребительные структуры строк без знаков в произведениях С.Есенина

№№ пп.	Структура строки	Абсолютная частота	Относительная частота
I.	АС	I67	0,0I324
2.	ГС	I5I	0,0II97
3.	ГАС	I33	0,0I055
4.	РАС	II6	0,00920
5.	ГСС	86	0,00682
6.	ГРС	83	0,00658

7.	САС	76	0,00603
8.	РСАС	7I	0,00563
9.	ГМС	70	0,00555
I0.	СС	55	0,00436

Общее число структур в этом случае оказалось гораздо меньшим - 664I. Из них первые 50 структур включают структуры с частотой $F \geqslant 2I$ и покрывают 2293 структуры (I8,2 %). Первая сотня этих структур включает 3053 структуры, что составляет 24,2 % общего числа структур.

Интересные результаты дает частотный список ритмических структур строк. Из общего числа I26I0 строк машиной зафиксировано 787 разных структур. Причем первые I0 структур (табл. 3) покрывают 32 % всех структур, первые 50-7I %, первые I00-80 %. Первые I60 структур (они включают все структуры до частоты 9) покрывают 89 % всего количества структур.

Таблица 3

Наиболее употребительные ритмические структуры строк в произведениях С.Есенина

№№ пп.	Ритмическая структура строки	Абсолютная частота	Относительная частота
I	03 06 09	670	0,053I3
2	03 05 07	462	0,03664
3	02 05 08	450	0,03569
4	02 04 08	402	0,03I88
5	03 07	394	0,03I25
6	03 05 09	38I	0,0302I
7	03 06 08	380	0,030I3
8	02 05 07	293	0,02324
9	03 05	287	0,02276
I0	0I 03 07	258	0,02046

При анализе распределения структур по количеству слогов в строке оказалось, что 88,5 % всех строк имеют длину от 6 до I2 слогов. Строк длиной в 8 слогов оказалось 20 % от общего числа строк, в 9 слогов - 2I,6 %, в I0 слогов - I5,7 % (это - самые употребительные типы ритмических структур).

Достаточно интересные результаты можно наблюдать, анализируя частотные списки употребляемости строк по распределению в них общего числа слогов в словах строки и мест ударных слогов (табл. 4).

Таблица 4

Наиболее употребительные структуры строк по числу слогов в словах и месту ударного слога в них

№№ пп.	Структура строки	Абсолютная частота	Относительная частота
I	2I 2I 33	29	0,00230
2	22 22	29	0,00230
3	2I	27	0,002I4
4	2I 43	25	0,00I98
5	43 33	25	0,00I98
6	I0 22 22 22	24	0,00I90
7	I0 22 22 32	23	0,00I82
8	2I 2I 2I 2I	23	0,00I82
9	33 22 22	23	0,00I82
I0	43 2I 3I	22	0,00I74

Всего таких типов оказалось 693I. Первые 50 из них покрывают 95I структуру (7,5 %), а первые I00-I603 структур (I2,7 %).

Большой объем информации получен при изучении взаимосвязи длины строки в слогах и типа ритмических строк. Оказалось, что в исследуемом массиве стихотворений присутствует I386 подобных типов зависимостей (табл. 5). На долю первых 50-ти зависимостей приходится 6I75 или 49 % от общего I26I0 структур. Первые I00 структур покрывают 8I82 структуры (64,9 %).

Таблица 5

Наиболее употребительные ритмические типы строк в зависимости от длины строки в слогах

№№ пп.	Типы строк	Абсолютная частота	Относительная частота
I.	09 03 06 09	356	0,02823
2.	08 02 05 08	234	0,0I856
3.	I0 03 06 09	233	0,0I848
4.	09 02 05 08	I99	0,0I578
5.	09 02 04 08	I95	0,0I546
6.	08 02 04 08	I95	0,0I546
7.	09 03 05 09	I86	0,0I475
8.	08 03 06 08	I83	0,0I45I

9.	I0 03 05 09	I82	0,0I443
I0.	08 03 05 07	I76	0,0I396

Наконец, машина выдала данные, показывающие вышеуказанную зависимость для всех длин строк. Например, наиболее употребительные типы ритмических структур для строк длиной в 9 слогов выглядят так (табл. 6).

Таблица 6

Наиболее употребительные ритмические типы строк для строк длиной в 9 слогов

№№ пп.	Типы строк	Абсолютная частота	Относительная частота
I.	09 03 06 09	356	0,02823
2.	09 02 05 08	I99	0,0I578
3.	09 02 04 08	I95	0,0I546
4.	09 03 05 09	I86	0,0I475
5.	09 0I 03 06 09	I43	0,0II34
6.	09 03 06 08	I37	0,0I086
7.	09 03 05 07	I33	0,0I055
8.	09 02 04 06 08	9I	0,00722
9.	09 02 06 08	89	0,00706
I0.	09 03 07	74	0,00587

Следующий тип информации, выдаваемой компьютером, связан со статистикой грамматических особенностей текстов С. Есенина.

После окончания обработки каждого стихотворения машина выдавала информацию о распределении в этом стихотворении основных классов слов. Такая же информация выдавалась и после окончания обработки всех произведений автора (табл. 7).

В ходе программного наложения строк произведений на словарь автора проводился подсчет употребительности в текстах С.Есенина грамматических значений рода, числа, падежа и времени. Все соответствующие данные выдавались машиной в виде таблиц, указывающих распределение данных значений по различным классам слов.

Так, например, грамматические значения мужского, женского и среднего рода употреблены С.Есениным в целом по текстам соответственно в 49,9%, 42,5%, 7,6% всех случаев. Для существительных же эти цифры таковы: 45,3%, 42,5% и I2,2% (табл.8).

Таблица 7

Распределение классов слов в поэтических произведениях С.Есенина

№№ пп.	Класс слов	Абсолютная частота	Относительная частота
I.	Существительное	I6235	0,2882
2.	Имя собственное	982	0,0I75
3.	Местоимение	7933	0,I408
4.	Числительное количественное	30I	0,0054
	Итого:	25209	0,4474
5.	Глагол	7485	0,I329
6.	Глагол в инфинитиве	I203	0,02I4
7.	Прилагательное в краткой форме	496	0,0088
	Итого:	9I84	0,I630
8.	Прилагательное	4674	0,0830
9.	Причастие	673	0,0I20
I0.	Числительное порядковое	83	0,00I5
	Итого:	5430	0,0964
II.	Наречие	3700	0,0657
I2.	Деепричастие	466	0,0083
	Итого:	4I66	0,0740
I3.	Другие классы слов	I2358	0,2I94
	Всего:	56347	I,0002

Таблица 8

Употребительность категории рода в поэтических произведениях С.Есенина

	Код и частота					
	мужской		женский		средний	
	абсол.	относ.	абсол.	относ.	абсол.	относ.
I	2	3	4	5	6	7
Существительное	735I	0,4528	6904	0,4253	I980	0,I220
Имя собственное	665	0,6772	3I0	0,3I57	7	0,0072

I	2	3	4	5	6	7
Местоимение	3929	0,4953	3632	0,4579	372	0,0469
Числительное количественное	229	0,7608	58	0,I927	I4	0,0466
Глагол	4040	0,5398	32II	0,4290	234	0,03I3
Прилагательное в краткой форме	354	0,7I37	89	0,I795	53	0,I069
Прилагательное	2283	0,4885	2I60	0,4622	23I	0,0495
Причастие	477	0,7088	I48	0,2200	48	0,07I4
Числительное порядковое	76	0,9I57	6	0,0723	I	0,0I2I
Итого:	I9404	0,4993	I65I8	0,425I	2940	0,0757

Единственное число употреблялось С.Есениным в 76,7% случаев, множественное - в 23,3 %. Для существительных эти цифры соответственно равны 73,5 % и 26,5 % (табл. 9).

Таблица 9

Употребительность категории числа в поэтических произведениях С.Есенина

	Число и частота			
	единственное		множественное	
	абсол.	относит.	абсол.	относит.
Существительное	II926	0,7346	4309	0,2655
Имя собственное	975	0,9929	7	0,007I
Местоимение	6309	0,7953	I624	0,2047
Числительное количественное	265	0,8804	36	0,II96
Глагол	5888	0,7867	I597	0,2I33
Прилагательное в краткой форме	4II	0,8287	85	0,I7I3
Прилагательное	3492	0,747I	II82	0,2529
Причастие	477	0,7088	I96	0,29I2
Числительное порядковое	80	0,9639	3	0,036I
Итого:	29823	0,7674	9039	0,2326

Для глаголов и причастий даются таблицы распределения грамматических значений времени. Так, 43,8 % глаголов в текстах употреблены в настоящем времени, 4I,7 % - в прошедшем и I4,5 % - в будущем. Для причастий соотношение настоящего и

прошедшего времени определяется соответственно цифрами 20,7% и 79,3 % (табл. I0).

Таблица I0

Употребительность категории времени в поэтических произведениях С.Есенина

	Класс слов и частота				Итого	
	Глагол		Причастие			
	Абсол.	Относ.	Абсол.	Относ.	Абсол.	Относ.
Настоящее	2908	0,4380	I39	0,2066	3047	0,4I67
Прошедшее	2767	0,4I68	534	0,7934	330I	0,45I5
Будущее	964	0,I452	-	-	-	-
Повелительное наклонение	846	-	-	-	-	-
Инфинитив	I203	-	-	-	-	-

Компьютер выдал также распределения падежных форм для существительных, имен собственных, местоимений, порядковых и количественных числительных. Однако в виду большой омонимии падежных форм эти данные требуют дополнительных уточнений.

Полученная информация может быть с успехом использована не только в стиховедении, но и в литературоведении вообще, а также в стилистике и в лингвистике.

ЛИТЕРАТУРА

Гаспаров М.Л. Современный русский стих. Метрика и ритмика. М., I974.

Гайдукова З.С., Зубов А.В. Частотный словарь поэтических произведений С.Есенина. - В кн.: Вопросы общей и прикладной лингвистики. Минск, I975, с. I65-I86.

Гиндин С.И. Общее и русское стиховедение. Систематический указатель литературы, изданной в СССР на русском языке с I958 по I974 гг. - В кн.: Исследования по теории стиха. Л., I978, с. I52-222.

Григорьев В.П. (ред.) Поэт и слово. Опыт словаря. М., I973.

Есенин Сергей. Собрание сочинений в пяти томах. М., том I, I966; том II, I966; том Ш, I967.

Есенин Сергей. Собрание сочинений в пяти томах. М., I967, том, Ш.

Ралько І.Д. Вершаскладанне. Мінск, I977.

AUTOMATIC STATISTICAL ANALYSIS OF POETICAL TEXTS

Alexander V. Zubov

S u m m a r y

In this work, an attempt was made to carry out an analysis of syntactic, lexical, morphological and rhythmical peculiarities of texts by S. Esenin.

The alphabetic and frequency lists of the forms taken by words were obtained. Each dictionary comprised 16,786 forms of words taken from the text, 65,347 usages of words in length. The following data were calculated: the usage of 17 word classes, the usage of grammatical meanings of gender, number, case and tense. Information was obtained about the distribution of structure of lines on the level of word classes, and the distribution of rhythmic structures of lines.

These results were obtained on an EC-1020 electronic computer by means of one master program comprising 1117 statements of PL-I and 10 auxiliary programs, written in the same algorithmic language.

ЗАМЕЧАНИЯ О ПРИМЕНЕНИИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ ИЗУЧЕНИЯ ЗАВИСИМОСТЕЙ И СВЯЗЕЙ МЕЖДУ ХАРАКТЕРИСТИКАМИ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

Ю.И. Левин

0. Применение статистических методов в изучении литературы (и, шире, языка) имеет ряд специфических черт, отличающих этот случай от применения статистики, скажем, в производстве. Специфика эта связана, прежде всего, со значительно большей неоднородностью рассматриваемых совокупностей (даже в наиболее "автоматизированной" области - фонологии), чем это бывает обычно в других областях применения статистики[1]. Пусть, например, сравниваются по каким-либо параметрам два текста. Прежде, чем говорить о различиях между ними, мы должны убедиться в том, что сами эти тексты "внутри себя" однородны, а не состоят из гетерогенных частей, "усреднение" между которыми резко нарушает истинную картину.

С другой стороны, именно эта неоднородность часто позволяет вообще обходиться без применения методов математической статистики (статистических критериев), ограничиваясь описательной статистикой и прямым "визуальным" сопоставлением числовых данных или графиков, - ибо различия между текстами часто оказываются статистически очевидными.

Если же расхождения между текстами сравнительно невелики (скажем, сравнивается ямб "Полтавы" и "Медного всадника"), то применение статистических критериев становится необходимым, но возникает новая методологическая трудность, связанная с тем, что рассматривать здесь как генеральную совокупность (ГС). Если считать, что множество всех строк "Полтавы" ("Медного всадника") составляет ГС, и вести подсчеты по всему корпусу, как это обычно делается, то результаты сравнения по какому-либо критерию не имеют смысла, поскольку сравнение ГС

---

[1] Кроме, пожалуй, социологии. Это сходство не случайно, ибо и в литературе, и в социальной жизни мы имеем дело со своеобразным сочетанием целеполагания и спонтанности, связанным с человеческим творчеством (в широком смысле слова), что делает объекты изучения (художественные тексты, человеческое поведение) гораздо более сложными, а их свойства менее предсказуемыми, чем в других областях, от техники до биологии. Можно сказать также, что такие объекты, как художественные тексты, близки по своим свойствам к "субъектам".

статистически бессмысленно: с помощью критериев сравниваются выборки, и критерий позволяет судить о том, можно ли рассматривать эти выборки как извлеченные из одной и той же ГС (гипотеза об однородности). Выход состоит в том, чтобы рассматривать множество строк "Полтавы" как выборку из гипотетической ("идеальной") бесконечной ГС, которую можно интерпретировать как множество всех текстов (строк), которые автор мог бы создать в том же "состоянии", в котором был создан рассматриваемый текст. Существование аналогичной ГС надо предположить и для "Медного всадника", и теперь проверять гипотезу об однородности (то есть о том, что на деле эти две "идеальные" ГС совпадают; иначе: о том, что обе поэмы созданы в одинаковом "состоянии")[2].

Эта заметка посвящена, прежде всего, особенностям устройства ГС, их элементов и выборок в рассматриваемой области. Нас интересует выявление различных типов ГС и заданных на них характеристик. При этом мы ограничиваемся рассмотрением случаев, когда статистика применяется не в чисто описательном плане, а для выявления зависимости каких-либо параметров текстов от "типа" этих текстов (разные авторы, жанры, направления, эпохи, метры и т.д.) или же корреляции между различными параметрами одного текста (корпуса текстов).

1. С р а в н е н и е р а з л и ч н ы х т е к с т о в.

10. Речь может идти о сравнении

а) для одного автора: "однотипных" текстов; текстов разных жанров; разных периодов; разных по тематике;

б) текстов разных авторов;

в) текстов разных литературных направлений и/или разных периодов;

г) текстов разной тематики, разных жанров или функциональных стилей;

д) стихотворных текстов разных размеров (одного автора или множества авторов);

е) различных (в композиционном или тематическом отношении) частей одного текста (например, "военные", "мирные" и "философские" части "Войны и мира"; реплики разных персонажей в драме и т.д.) и т.д.[3]

---

[2] В этой необходимости конструировать "идеальные" ГС нет еще особой специфики. Если мы решаем вопрос о симметрии монеты, подбрасывая ее 100 раз, то эти 100 бросаний надо рассматривать как выборку из "идеальной" бесконечной ГС, состоящей из "всевозможных" бросаний этой монеты.

[3] Все примеры, связанные с различными типами текстов и

Мы имеем, таким образом, k "текстов" (в широком смысле слова), которые сравниваются по интересующим нас характеристикам. Если при этом выделенный для подсчетов объем текста n (в соответствующих единицах) мал по сравнению с объемом всего "текста" N (скажем, $n < 0{,}1N$), то весь "текст" можно рассматривать как ГС (характер элементов которой еще подлежит уточнению в зависимости от рассматриваемых характеристик), из которой извлечена выборка объема n (сюда же относится случай, когда весь "текст" незамкнут и/или необозрим; например, такие "тексты", как "газетная публицистика" или "русский рассказ 70-х годов XX века"); если же обследуется весь реальный текст или значительная его часть, то обследуемый объем следует считать выборкой из "идеальной" (гипотетической) ГС - см. п.0. Цель сравнения - проверка гипотезы о том, что наши выборки можно рассматривать (относительно данных характеристик) как извлеченные из одной и той же ГС, или, иначе, о том, что тип текста не влияет на значения (распределения) изучаемых характеристик.

Сравнение может идти по одному или нескольким параметрам (количественным характеристикам). Эти характеристики можно разделить на два класса.

11. Первый класс: характеристика по самой своей природе относится к определенному формально выделяемому и имеющему "лингвистический смысл" сегменту текста: слову (число графем или слогов), предложению (число слов), стихотворной строке (число слогов, число ударений), стихотворению[4] (число строк) и т.д., - причем весь текст членится без остатка[5] на такие непересекающиеся сегменты.

ГС в этом случае представляет собой "идеальный" или реальный текст как множество сегментов (слов, предложений, строк, стихотворений и т.д.) соответствующего вида. В случае $k = 2$ сравнение проводится - если по характеру данных в этом есть необходимость - по известным критериям однородности, используемым в "задаче двух выборок", - наиболее известен крите-

---

их характеристиками, приводимые в этой заметке, являются чисто иллюстративными, и соответствующие перечни ни в коем случае не претендуют на полноту.

[4] Если в качестве текста взято множество стихотворений, каждое из которых рассматривается как "единица" текста.

[5] Это требование не является обязательным: "остаток" (скажем, внеметрические строки в стихотворении данного метра или сложные предложения, если нас интересуют только длины простых) мы можем просто игнорировать.

рий Стьюдента (применимый, когда распределения хотя бы приближенно нормальны и обладают одинаковыми дисперсиями), но предпочтительнее "свободные от распределения" критерии типа критериев Смирнова (где сравниваются выборочные функции распределения), Вилкоксона, χ^2 (для двух выборок) и т.д. В случае, когда $k > 2$, прибегают либо к "нестрогому" приему визуального сравнения графиков (например, гистограмм или полигонов частот), либо к дисперсионному анализу (что возможно при тех же предположениях, которые относятся к критерию Стьюдента), либо, наконец, к попарным сравнениям с помощью указанных выше критериев[6].

12. Второй класс: характеристика не привязана к определенному типу сегментов и может вычисляться и для текста в целом и для любого его сегмента.

121. Рассмотрим вначале случай, когда именно тексту как целому приписывается одна числовая характеристика. В этом случае существенно, имеет ли эта величина выборочный характер, то есть может ли эта характеристика рассматриваться как реализация некоторой случайной величины, распределение которой в принципе известно (см. п. 1211) - или же она является существенно невыборочной (см. п. 1212).

12111. К первому классу относятся, прежде всего, частоты тех или иных элементов текста. Текст при этом мыслится разбитым[7] на сегменты (элементы) типов A_1, A_2, , A_S (в частности, A и $\bar{A}$), или, общее, обладающие свойствами A_1, A_2,..., A_S, - что позволяет отнести сюда и случай, когда события A_i и A_j ($i \neq j$) совместимы[8]. Здесь возможны два эквивалентных подхода. Либо весь текст длины n (единица - сегмент) рассматривается как выборка объема 1 из идеальной ГС[9], реализующая для каждого A биномиальную случайную величину с испытнаниями и с (неизвестной) вероятностью "успеха" (то есть наступления события A) . Либо текст рассматривается как выборка объема из "идеальной" ГС, причем элементами выборки (и ГС) служат рассматриваемые сегменты, каждый из которых

---

[6] Отметим, что при сравнении только средних происходит потеря информации, связанной с характером распределения (средние могут совпадать и в случае значимого различия распределений данной характеристики в текстах).

[7] Возможно, "с остатком" - см. сн. 5.

[8] Например, при рассмотрении профиля ударности, где сегмент - строка, A_i - ударен i-ый икт.

[9] Которая в данном случае мыслится как состоящая из (гипотетических) текстов длины n каждый.

обладает или не обладает свойством A_i. При любом подходе выборочные данные сводятся к набору $(n_1, n_2, \ldots, n_s)$ частот элементов со свойствами A_1, A_2,..., A_s. Такими сегментами могут быть

а) фонемы (графемы): A_i – некоторая фонема или группа фонем;

б) словоформы: A_1 – существительные, A_2 – прилагательные и т.д.; A_1 – усеченные, A_2 – полные формы прилагательных; A_1 – сущ. им., A_2 – сущ. род. и т.д.; A – слово, входящее в метафору, $\overline{A}$ – остальные; A – словоформы частоты 1, $\overline{A}$ – остальные; A – словоформы, принадлежащие к числу 20 наиболее частых, $\overline{A}$ – остальные[10] и т.д.;

в) предложения: A_1 – простые, A_2 – сложносочиненные, A_3 – сложноподчиненные и т.д.;

г) стихотворные строки: A_1 – неточная рифма, A_2 – приблизительная рифма, A_3 – точная рифма; A_1, A_2,... – ритмические формы данного размера и т.д.;

д) стихотворения (см. сн. 4): A_1, A_2,... – различные размеры и т.д.

Наиболее распространенным критерием для проверки гипотезы об однородности (т.е. о том, что все выборки извлечены из одной и той же ГС) или, что то же, о независимости (вероятности сегментов A_1, A_2,..., A_s не зависят от типа текста), является критерий χ^2 для выборок (Крамер, 1975, 482). Рассмотрим вначале "полиномиальный" случай, когда A_i и A_j $(i \neq j)$ несовместимы. Если выборочные данные по k текстам сведены в таблицу

тип сегмента \ тип текста	1	...	j	...	k	суммы по строкам
A_1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1\cdot}$
⋮						⋮
A_i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i\cdot}$
⋮						⋮
A_s	n_{s1}	...	n_{sj}	...	n_{sk}	$n_{s\cdot}$
суммы по столбцам	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot k}$	n

[10] Последние два примера – особые: отнесение к классу здесь зависит от текста в целом (а соответствующая вероятность – от его длины – см. п. 1212).

где n_{ij} - частота сегмента A_i в j-ом тексте; $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$; $n_{\cdot j} = \sum_{i=1}^{s} n_{ij}$, $n = \sum_{i,j} n_{ij}$,

то значение критерия вычисляется по формуле

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}} , \qquad (*)$$

причем эта величина, если гипотеза верна, имеет распределение χ^2 с $(k-1)(s-1)$ степенями свободы.

В частном случае $k = 2$ (сравнение двух текстов) таблица принимает вид

тип сегмента \ тип текста	1	2	суммы по строкам
A_1	m_1	n_1	$m_1 + n_1$
⋮			
A_i	m_i	n_i	$m_i + n_i$
⋮			
A_s	m_s	n_s	$m_s + n_s$
суммы по столбцам	m	n	$m + n$

а формула (*) может быть приведена к виду

$$\chi^2 = mn \sum_i \frac{1}{m_i + n_i} \left(\frac{m_i}{m} - \frac{n_i}{n}\right)^2 = \frac{1}{\omega(1-\omega)} \left(\sum_i m_i \omega_i - m\omega\right),$$

где

$$\omega_i = \frac{m_i}{m_i + n_i}, \qquad \omega = \frac{m}{m + n} .$$

В частном случае $s = 2$ (сравнение по альтернативному признаку) имеем:

признак \ тип текста	1	j	k	суммы
A	m_1	m_j	m_k	$\sum_i m_i$
$\bar{A}$	$n_1 - m_1$	$n_j - m_j$	$n_k - m_k$	$n - \sum_i m_i$
всего	n_1	n_j	n_k	n

$$\chi^2 = \sum_i \frac{(m_i - n_i p)^2}{n_i p q} = \frac{1}{pq} \sum_i \frac{m_i^2}{n_i} - \frac{np}{q},$$

где $$p = \frac{1}{n} \sum_i m_i \ , \quad q = 1 - p \ .$$

Случай совместимых A_i покажем для большей ясности на примере сравнения текстов по профилю ударности (A_i – ударен i-ый икт):

тексты / A_i	1 ...	j ...	k	суммы по строкам
A_1	n_{11} ...	n_{1j} ...	n_{1k}	m_1
⋮	⋮	⋮	⋮	⋮
A_i	n_{i1} ...	n_{ij} ...	n_{ik}	m_i
⋮	⋮	⋮	⋮	⋮
A_s	n_{s1} ...	n_{sj} ...	n_{sk}	m_s
число строк в j-м тексте	n_1 ...	n_j ...	n_k	n

где $m_i = \sum_j n_{ij}$ – общее число строк с ударным i-м иктом, n_j – число строк в j-м тексте (в отличие от полиномиального случая, вообще говоря, $n_i \neq \sum_j n_{ij}$), $n = \sum_j n_j$. Здесь

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{m_i n_j}{n}\right)^2}{\frac{m_i n_j}{n}}$$

с $s(k-1)$ степенями свободы.

I2II2. Иногда в качестве характеристик целого текста используются функции (а именно, отношения) частот отдельных элементов текста. Таковы, например, "меры", предложенные Б.Н. Головиным (Головин, I97I, I43-I5I). С точки зрения структуры и вероятностных свойств можно выделить три типа таких мер.

а. Типа "меры связанности" $= \frac{\text{число предлогов и союзов}}{3 \cdot \text{число предложений}}$.

Такая величина совпадает (с точностью до множителя) со средним числом элементов данного типа в предложении – см. п.I2II3.

б. Типа $\frac{\text{число простых предложений}}{\text{число сложных предложений}}$, то есть типа $\frac{m}{n-m}$,

где n – общее число элементов в выборке, m – число "успехов". Такие меры несущественно отличаются от относительных частот

$\frac{m}{n}$ $\left(= \frac{\text{число простых предложений}}{\text{общее число предложений}}\right)$, и притом отличаются невыгодно, поскольку относительные частоты можно сравнивать по известным критериям (например, χ^2), а отношения типа $\frac{m}{n-m}$, хотя тоже в принципе сравнимы, однако требуют построения новых, более сложных критериев.

в. Типа $\frac{\text{число прилагательных}}{\text{число существительных}}$. В отличие от случая б прилагательные и существительные не покрывают весь текст, и мы имеем здесь дело со следующей ситуацией: имеются два полиномиальных распределения, каждое с тремя исходами - A'(прилагательные), A'' (существительные), A''' (прочее). По тексту вычисляется отношение $\frac{m'}{m''}$ по другому, сравниваемому, отношение $\frac{n'}{n''}$. Для статистического сравнения текстов надо проверить гипотезу о равенстве отношений двух вероятностей в двух "триномиальных" распределениях: $\frac{p_1'}{p_1''} = \frac{p_2'}{p_2''}$ (где p_i' (p_i'') - вероятность прилагательного (существительного) в i-м тексте). Критерии для проверки такой гипотезы неизвестны, хотя в принципе могут быть построены.

12113. К числовым характеристикам целого текста относятся, далее, любые усредненные по всему тексту величины класса 1 (см. п. 11), характеризующие те или иные естественно выделенные сегменты текста (например, среднее число графем в слове, слов в предложении, ударений в стихотворной строке и т.д.). Здесь нет ничего нового по сравнению с ситуацией п.11 (только мы отказываемся от рассмотрения распределений, ограничиваясь средними). Более интересны такие "меры", как дисперсия той или иной количественной характеристики (например, длин предложений), или как коэффициент корреляции между какими-либо характеристиками текста. В этих случаях (как и в случае средних или частот) изучаемые тексты следует рассматривать как выборки (обычно из "идеальных" ГС), и использовать критерии для сравнения соответствующих выборочных характеристик (для дисперсий - критерий дисперсионного отношения для проверки гипотезы $\sigma_1^2 = \sigma_2^2$ в случае $k = 2$ и критерии Кокрена или Бартлета для проверки гипотезы $\sigma_1^2 = \ldots = \sigma_k^2$ в случае $k > 2$; для коэффициентов корреляции - критерий равенства коэффициентов корреляции (Кендалл и Стьюарт, 1973,395).

1212. Иная ситуация возникает, когда рассматриваются "глобальные" характеристики целого текста, не имеющие выборочного характера, то есть такие, при рассмотрении которых нельзя рассматривать текст как выборку. Речь идет о таких "мерах" текста, которые аналогичны, скажем, длинам или весам предметов. Сюда относятся такие характеристики, как "мера компактности" распределения в тексте тех или иных элементов (Левин, 1967; Веденина и Шор, 1973), а также различные меры, связанные с лексической структурой текста (отношение объема словника к числу словоформ; доля текста, покрываемая словами частоты 1 или s наиболее частыми словами; параметр закона Ципфа для данного текста и т.д.). В силу невыборочного характера этих величин статистическое сравнение текстов здесь невозможно - возможно лишь сравнение "на глаз", а также ранжирование текстов по значениям рассматриваемой меры.

Отметим, что меры, связанные с лексической структурой текста, упомянутые выше, обладают важной особенностью: они существенно зависят от длины текста[11]. Поэтому по таким характеристикам могут сопоставляться только тексты равного (или близкого) объема (или же равные по объему выборки из разных текстов)[12].

122. Поскольку характеристики рассматриваемого в п. 12 класса могут вычисляться для любого отрезка текста, часто используется следующий прием: текст разбивают на непересекающиеся отрезки равной "длины" (по числу графем, слов, предложений, строк и т.д. - в зависимости от того, о какой характеристике идет речь), скажем, 100, - причем эти отрезки могут и не покрывать весь текст, - и характеристика вычисляется для каждого из этих отрезков. Такой подход превращает рассматриваемую характеристику в характеристику класса 1 (п. 11), только сегменты, к которым относится характеристика, оказываются здесь не "естественными" (как, например, предложение), а "искусственными", - и с ней можно обращаться как описано в п. 11, в частности, рассматривать и сравнивать как распределения этой характеристики в различных текстах, так и средние значения. ГС в этом случае представляет собой текст

---

[11] То же верно, конечно, и для частоты того или иного элемента в тексте; но здесь эта зависимость - дело поправимое, поскольку частота пропорциональна (в статистическом смысле) длине текста, что дает возможность рассматривать относительную частоту.

[12] Все, сказанное до сих пор в п. 12 применительно к це-

(может быть "идеальный") как множество отрезков длины 100.

Такой подход имеет смысл при следующих обстоятельствах:

а) когда происходит не сравнение текстов, а проверка однородности данного текста;

б) когда вычисляемая характеристика является невыборочной (п. 1212) и/или зависящей от длины текста и, вычисленная для целого текста, не дает возможности применять статистические критерии для сравнения текстов;

в) когда неизвестны (или слишком сложны) критерии для сравнения значений характеристики по целым текстам (см. п. 12112, в).

В остальных случаях - то есть при сравнении частот, средних значений, дисперсий, коэффициентов корреляции в текстах, каждый из которых внутренне однороден, - применение этой "многовыборочной" методики не дает никаких выгод по сравнению с рассмотрением этих характеристик по целым текстам (просто вместо одной "большой" выборки рассматривается несколько "малых").

2. Внутритекстовые корреляции.

20. Здесь, в отличие от п. 1, мы имеем дело с одной ГС, на элементах которой заданы две характеристики (будем рассматривать только случай парной корреляции), каждая из которых может быть как качественной, так и количественной (в том числе частотной). Цель исследования - выявление определенного статистического свойства ГС в целом, состоящего в том, что наличие у некоторого элемента данного значения одной из характеристик влияет (или не влияет) на вероятность того, что вторая характеристика принимает то или иное значение.

21. Случай, когда одна из характеристик является качественной (могущей принимать взаимоисключающие значения $A_1, \ldots, A_k$, в частности, A и $\bar{A}$), а элементами ГС служат законченные тексты (загадка, пословица, стихотворение, рассказ и т.д.), может рассматриваться как с точки зрения "внутритекстовых" корреляций - если не множество текстов (может быть, "идеальное") берется как единая ГС, - так и с точки зрения п. 1, если множества текстов, обладающих различными A_i, мы рассматриваем в качестве различных ГС и сравниваем их по второй характеристике.

лому тексту, относится, конечно, и к любой (достаточно большой) выборке из него - в предположении об однородности текста.

Пусть, например, ГС - множество (реальное или идеальное) русских загадок, выборка - некоторое случайно взятое его подмножество, характеристика А - наличие (Ā - отсутствие) рифмы, В - наличие (B̄ - отсутствие) антитезы. С помощью таблицы сопряженности признаков (в данном случае 2 х 2) - (Рао 1968, 355) - и критерия χ^2 может проверяться гипотеза о зависимости между этими характеристиками (или же может вычисляться выборочный коэффициент корреляции качественных признаков и устанавливаться значимость его отличия от нуля). С другой точки зрения (п. 1) мы можем рассматривать две ГС (два "жанра") - скажем, рифмованных и нерифмованных загадок - и сравнивать их по частоте антитез, как в п. 12111, пользуясь, например, критерием χ^2 для двух выборок.

Другой пример: ГС - некоторое множество стихотворений; выявляется связь между тематикой (любовное, медитативное и т.д.) и длиной стихотворения (количественный признак) или размером (качественный признак) или частотой той или иной группы фонем (частотный признак) и т.д. Здесь также возможен альтернативный подход, когда каждая тематическая группа рассматривается как ГС, и эти ГС сравниваются по второй характеристике (в качестве различных ГС можно брать и, например, множества стихотворений, написанных различными размерами).

22. В других случаях - когда элементы ГС не являются законченными текстами и/или когда обе характеристики не являются качественными - альтернативный подход типа п. 1 невозможен. Приведем несколько примеров.

а. Корреляции между частотой двух типов элементов А и В (например, существительных и прилагательных или неточных рифм и метафор) в текстах. ГС здесь - некоторое (реальное или идеальное) множество текстов, характеристики - относительные частоты элементов А и В в тексте. По выборке из ГС может быть вычислен выборочный коэффициент корреляции r и установлена значимость его отличия от нуля[13].

б. Корреляция между количественными (например, стопность

[13] В этом и других случаях, когда может вычисляться r, на него можно смотреть с двух точек зрения: 1) r как оценка истинного коэффициента корреляции ϱ, вычисляемая для проверки гипотезы $\varrho = 0$; 2) r как "мера связанности" данных двух характеристик для данной совокупности текстов. В этой второй ипостаси r может вычисляться для разных совокупностей текстов (разных авторов, жанров и т.д.) и использоваться как количественная характеристика этих ГС для их сравнения в духе п. 12113.

ямба [14]) и частотным (например, доля текста, покрытая метафорами) признаками. Здесь также может вычисляться r.

в. Корреляция между качественными признаками, носителями которых являются не (как в п. 21) целые тексты, а элементы текста, например, корреляция между метафорическим/неметафорическим употреблением глагола и его видом или залогом или временем. Здесь ГС - реальный или идеальный текст как множество вхождений в него глагольных словоформ. Для выявления связи могут использоваться таблицы сопряженности признаков и критерий χ^2 (в случае таблиц 2 х 2 может также вычисляться r).

23. Кратко резюмируем то, что уже говорилось об употребляемом при изучении зависимостей в статистическом аппарате.

Если изучается зависимость между признаками, хотя бы один из которых является качественным, то универсально применимый аппарат - таблицы сопряженности признаков и критерий χ^2 [15]. В частном случае, когда каждый признак имеет два значения, получается таблица 2 х 2 и можно использовать не критерий χ^2 (являющийся приближенным), а пользоваться точным критерием для таблиц 2 х 2 (Большев и Смирнов, 1968, табл. 5.6), что особенно существенно в случае малых выборок, - или же вычислять выборочный коэффициент корреляции для качественных признаков (приписывая А и В значение 1, а $\bar{A}$ и $\bar{B}$ - значение 0) и устанавливать значимость его отличия от нуля.

Если же оба признака - количественные (в частности, частотные), то наиболее естественный путь - вычисление обычного выборочного коэффициента корреляции и установление значимости его отличия от нуля.

---

[14] Отметим, что в этом случае (и многих других) количественный признак может рассматриваться как качественный, принимающий значения, скажем, A_2, A_3, A_4, A_5, A_6 (A_i - i-стопный ямб). При таком подходе r вычисляться уже не может, а используются таблицы сопряженности признаков ("стопность" - "тропность") и критерий χ^2 (при этом второй признак должен быть дискретизирован путем разбиения множества его возможных значений на интервалы). Наконец, можно рассматривать A_i как "типы текстов", то есть как различные ГС, и использовать критерий χ^2 как критерий однородности.

[15] Если при этом второй признак - количественный (и притом непрерывный), то необходима его дискретизация - см. сн. 14. То же, разумеется, можно проделать и при изучении зависимости между количественными признаками.

3. Подведем итоги, касающиеся характера ГС и их элементов. Во всех случаях ГС выступает как текст или корпус текстов, реальный - или же идеальный, гипотетический (в случае, когда обследованию подвергается - то есть в выборку попадает - весь реальный текст или значительная его часть), рассматриваемый как множество элементов либо определенной "естественной" лингвистической (или "поэтической") природы - от фонемы (графемы) и слова до пословицы, стихотворения или другого законченного в себе текста (в последнем случае ГС является "текстом" в обобщенном смысле слова[16]), либо искусственной природы - скажем, отрезков фиксированной длины. Элементы эти могут и не покрывать весь текст как таковой - в таком случае оставшаяся часть текста мыслится несуществующей. При этом каждый элемент является носителем той или иной количественной или качественной характеристики (в случае изучения корреляций - двух характеристик). Такой качественной характеристикой является обладание хотя бы одним из непересекающихся (тогда - ровно одним) или, реже, пересекающихся свойств A_1, ..., A_S (часто $s = 2$: случай A и $\bar{A}$). При этом иногда (см. п. 12111) возможен альтернативный подход: частота элементов со свойством A_i может рассматриваться как количественная характеристика текста, который при таком подходе предстает уже не как состоящий из элементов, образующих выборку объема n, а как выборка объема 1 из (идеальной) ГС.

Заметим в заключение, что выходя за пределы статистического изучения текстов в общую лингвистику, психо- и социолингвистику и т.д., мы встречаемся с ГС совершенно другого характера. Так, при изучении корреляции между числом фонем в языке и средней длиной морфемы (два количественных признака) элементами ГС служат языки; при изучении зависимости употребления тех или иных фонетических, морфологических и т.д. вариантов от возраста, социального статуса и других характеристик (качественный - количественный или качественный признаки) элементами ГС являются носители языка (а элементами выборки - информанты) и т.д.

---

[16] В этом случае особенно ясно, что ГС выступает как неупорядоченное множество своих элементов, но это же верно и в других случаях: линейная упорядоченность "настоящего" текста при тех статистических рассмотрениях, о которых шла речь, во внимание не принимается.

Л И Т Е Р А Т У Р А

Большев Л., Смирнов Н. Таблицы математической статистики. М., 1968.

Веденина Л., Шор Е. Некоторые приемы стилистического исследования текста. М., 1973.

Головин Б.Н. Язык и статистика. М., 1971.

Кендалл М., Стьюарт А. Статистические выводы и связи. М., 1973.

Крамер Г. Математические методы статистики. М., 1975.

Левин Ю. О количественных характеристиках распределения символов в тексте. - Вопросы языкознания, 1967, № 6.

Рао С. Линейные статистические методы. М., 1968.

NOTES ON APPLICATION OF MATHEMATICAL STATISTICS TO INVESTIGATION OF DEPENDENCES AND RELATIONS BETWEEN PARAMETERS OF LITERARY TEXTS

Yuri I. Levin

S u m m a r y

This paper deals with the specific traits of populations, samples, and their elements arising from statistical studies of literary texts.

A classification of various real and possible types of statistical approaches to texts depending on the aims pursued by the investigator is proposed and statistical tests corresponding to each type are indicated.

О РАСПРЕДЕЛЕНИЯХ ТЕРМИНОВ В АНГЛИЙСКОМ НАУЧНО-ТЕХНИЧЕСКОМ ТЕКСТЕ (ПОДЪЯЗЫК КВАНТОВЫХ ГЕНЕРАТОРОВ)

Н. Манасян

Проблема распределения терминов в научно-техническом тексте становится весьма актуальной для функциональной стилистики, инженерного языкознания и информатики. Распределения в наиболее обобщенной, абстрагированной форме представляют количественную организацию текста и его словаря и могут служить выявлению типологии научно-технического стиля, его подстилей, подъязыков. Различные лексико-семантические группы словарных единиц описываются количественно различными статистическими распределениями их в тексте (Бектаев, Лукьяненков, 1971; Каширина, 1973). Если установить в совокупности текстов распределения для единиц, например, таких, как общеупотребительные, терминологические единицы, общенаучные, отраслевые и др. термины, то можно формальным путем классифицировать такие единицы в текстах аналогичного содержания, исходя из их распределений.

Обычно в таких работах составляется частотный словарь (вручную или с помощью ЭВМ) в с е х единиц текста (совокупности текстов) с фиксацией частот каждой единицы в каждой из равных по длине (1 тыс. словоупотреблений) минимальных выборок При расписывании текста не регистрируется никакая содержательная информация о его единицах. В результате в один и тот же класс единиц в соответствии с типом их распределения могут попасть омонимичные друг другу единицы, например, нетерминологическая единица и терминологическое употребление этой же единицы.

Еще одна особенность предшествующих исследований состоит в том, что в них рассматривается ограниченное число единиц частотного словаря — 200-300 слов словоформ, то есть не более 1 - 3 % всего инвентаря или не более 10 % тех единиц инвентаря, которые по своим количественным характеристикам (суммарная частота и число минимальных выборок) могли бы описываться распределениями.

Наконец, в самих описаниях процедур расчетов, связанных с анализом распределений, либо отсутствует существенная информация, либо имеют место определенные неточности.

Исследование распределений терминов в текстах по квантовым генераторам на английском языке отличается от других работ прежде всего тем, что сам исходный материал, а именно употребления терминов, извлекался из текста "нетрадиционным" путем. При расписывании выборки объемом 200 тыс. словоупотреблений регистрировались только термины, включая терминологические употребления тех единиц, которые в другом контексте могли встретиться в общеупотребительном значении (то есть в этом втором случае они, естественно, не регистрировались). Другое отличие состоит в том, что из этой же текстовой выборки извлекались и терминологические сочетания. Следовательно, получены два частотных словаря — словарь однословных терминов и словарь терминосочетаний. Кроме того, из обоих частотных словарей для анализа отобраны в с е единицы, количественные характеристики которых (суммарная частота, частоты в минимальных выборках и число этих выборок) удовлетворяют требованиям анализа распределений. Таких единиц оказалось 1025. Количество вариант для каждой из них не менее 4; исключение составляют два термина **"replace, v"** и **"summarise, v"**,, имеющие по два значения частоты.

Первоначально за основу алгоритма анализа была взята схема, описанная в (Бектаев, Лукьяненков, 1971). Однако в ходе работы над алгоритмом и программой расчета на ЭВМ пришлось отказаться от строгого следования этой схеме, поскольку

а) характер данных и величина выборки внесли изменения в вышеупомянутый алгоритм (у К. Б. Бектаева и К. Ф. Лукьяненкова выборка равна 400 тыс. словоупотреблений),

б)в указанной схеме обнаружились некоторые неточности. Например, при сравнении эмпирических рядов частот с теоретическими при помощи критерия χ^2 авторы статьи берут для закона Пуассона (1-1) степеней свободы, а для нормального и логарифмически-нормального (1-2), где 1 — количество интервалов частот после укрупнения малочисленных интервалов. Это представляется некорректным -- в литературе по математической статис-

тике, включая также ту, на которую ссылаются авторы статьи, число степеней свободы для законов Пуассона и Гаусса нормального закона составляет (1-2) и (1-3) соответственно. Это естественно вытекает из понятия "степени свободы" и его определения, которое гласит, что при подсчете числа степеней свободы надо из числа разрядных частот вычесть число линейных связей, налагаемых на распределение. В линейные связи входят неизвестные параметры и сумма наблюденных частот (так как величина выборки всегда фиксирована).

Анализ распределений относительно трех законов, Пуассона, Гаусса и логнормального, описываемый в настоящей статье, выполнен на ЭВМ ЕС-1020. Распределение каждой терминологической единицы проверялось по 12 вариантам объемов текста при четырех разных объемах внутрисерийных выборок.

Для сравнения построенных эмпирических рядов с теоретическими использовались критерий Пирсона χ^2 и критерий Колмогорова.

Критерий χ^2 широко применяется как в статистике вообще, так и в лингвистической статистике в частности. Следует отметить, что его применение считается обоснованным только в том случае, если ни одна из разрядных частот не очень мала. Если же крайние частоты меньше пяти, их предлагают объединять так, чтобы суммы частот были не меньше пяти. Вследствие этого критерий χ^2 нельзя было применить ко всем нашим данным, и в тех случаях, когда в силу указанных условий его применение было невозможным, был использован критерий Колмогорова. Описание критериев см. в (Пиотровский, Бектаев, Пиотровская, 1977; Митропольский, 1971; Вентцель, 1969).

Ниже прилагается список анализируемых терминов. Степень приближения характеризуется уровнем 0,05.

Условные обозначения и сокращения: № — машинный номер единицы; П — закон Пуассона; Н — нормальный закон; Л — логарифмически-нормальный закон; + — соответствие эмпирического распределения теоретическому; - — несоответствие эмпирического распределения теоретическому; **F** — частота единицы; a — имя прилагательное; adv — наречие; g — герундий; n — имя существительное; num — имя числительное; PII — причас -

тие прошедшего времени; pn(attr) — имя собственное в атрибутивной функции; v — глагол.

Список анализируемых терминов

№	термин	F	П	Н	Л	№	термин	F	П	Н	Л
1	laser, n	1640	-	+	-	50	number, n	201	-	-	-
2	mode, n	634	-	-	-	51	measure, v	197	-	-	-
3	frequency, n	604	-	-	-	52	emission, n	189	+	-	-
4	beam, n	599	-	-	-	53	threshold, n	187	-	-	-
5	Fig.,	576	-	-	-	54	absorption, n	186	-	-	-
6	energy, n	566	-	-	-	55	experiment, n	186	+	-	-
7	field, n	561	-	-	-	56	pump, n	186	-	-	-
8	power, n	507	-	+	+	57	distribution, n	185	-	-	-
9	wave, n	496	-	-	-	58	photon, n	183	-	+	+
10	system, n	460	-	+	-	59	range, n	183	-	-	-
11	output, n	442	-	-	-	60	resonator, n	181	-	-	-
12	pulse, n	415	-	-	-	61	condition, n	177	-	-	-
13	radiation, n	399	-	-	-	62	operation, n	172	-	-	-
14	state, n	378	-	-	-	63	material, n	171	-	-	-
15	gain, n	376	-	-	-	64	coupling, a	169	-	-	-
16	level, n	376	-	-	+	65	maximum, n	169	-	-	-
17	light, n	353	-	-	-	66	amplifier, n	167	-	-	-
18	equation, n	343	-	-	+	67	term, n	164	-	-	-
19	optical, a	343	-	-	-	68	waveguide, n	162	-	-	-
20	cavity, n	342	-	-	-	69	structure, n	161	-	-	-
21	current, n	330	-	-	-	70	electric(al), a	159	-	-	-
22	electron, n	327	-	-	-	71	method, n	159	-	-	-
23	effect, n	321	-	+	+	72	efficiency, n	156	-	-	-
24	value, n	319	-	-	-	73	determine, v	155	-	-	-
25	result, n	296	+	-	-	74	noise, n	153	-	-	-
26	time, n	288	-	-	-	75	theory, n	149	-	-	-
27	mirror, n	287	-	-	-	76	coefficient, n	147	+	-	-
28	wavelength, n	279	-	-	-	77	process, n	147	+	+	-
29	crystal, n	278	-	-	-	78	direction, n	146	-	-	-
30	maser(MASER),n	277	-	-	-	79	quantum, a	144	-	-	-
31	rate, n	259	-	-	-	80	Eq., n	142	-	-	-
32	measurement, n	248	-	-	-	81	constant, n	141	-	-	-
33	density, n	247	-	-	-	82	tube, n	140	-	-	-
34	atom, n	246	-	-	-	83	surface, n	139	+	-	-
35	line, n	245	-	-	+	84	magnetic, a	138	-	-	-
36	loss, n	244	-	-	-	85	model, n	137	-	-	-
37	transition, n	228	-	-	-	86	spectrum, n	135	-	-	-
38	device, v	226	-	-	-	87	solution, n	134	-	-	-
39	function, n	225	-	-	-	88	band, n	132	-	-	-
40	intensity, n	225	-	-	-	89	medium, n	132	-	-	-
41	high, a	218	-	-	-	90	low, a	130	-	-	-
42	signal, n	217	-	-	-	91	technique, n	130	+	-	-
43	phase, n	216	-	-	-	92	axis, n	129	-	-	-
44	region, n	216	-	-	-	93	dye, n	129	-	-	-
45	temperature, n	212	-	-	-	94	experimental, a	128	+	-	-
46	produce, v	211	-	+	-	95	amplitude, n	125	-	-	-
47	obtain, v	207	-	-	-	96	index, n	125	-	-	-
48	gas, n	204	-	-	-	97	order, n	125	+	-	-
49	length, n	201	-	-	-	98	reduce, v	125	+	-	-

99	pressure, v	123	-	-	-	155	vary, v	89	-	-	-
100	source, n	123	+	-	-	156	zero, n	89	-	-	-
101	component, n	122	-	-	-	157	transverse, a	88	-	-	-
102	oscillator, n	122	-	-	-	158	contain, v	87	+	-	+
103	problem, n	121	-	-	-	159	electromagnetic,a	87	-	-	-
104	active, a	120	-	-	-	160	saturation, n	87	+	-	-
105	analysis, n	119	-	-	-	161	data, n	86	-	-	-
106	velocity, n	119	-	-	-	162	laboratory, n	86	-	-	-
107	coherent, a	118	+	-	-	163	peak, n	86	-	-	-
108	percent, n	118	-	-	-	164	configuration, n	85	+	-	-
109	voltage, n	115	-	-	-	165	variation, n	84	+	-	-
110	junction, n	114	-	-	-	166	layer, n	83	-	-	-
111	nonlinear, a	114	-	-	-	167	radius, n	82	-	-	-
112	factor, n	113	-	-	-	168	behavior, n	81	+	-	-
113	discharge, n	112	-	-	-	169	pump, v	81	-	-	-
114	modulation, n	112	-	-	-	170	radio, n, a	81	-	-	-
115	angle, n	110	-	-	-	171	thickness, n	81	+	-	-
116	glass, n	110	-	-	-	172	equal, a	80	+	-	-
117	interaction, n	110	-	-	-	173	particle, n	80	-	-	-
118	observe, v	110	-	-	-	174	film, n	79	-	-	-
119	plane, a	110	-	-	-	175	pumping, a	79	-	-	-
120	element, n	109	-	-	-	176	section, n	78	-	-	-
121	injection, n	109	-	-	-	177	apply, v	77	+	-	-
122	parameter, n	109	-	-	-	178	ion, n	77	-	-	-
123	microwave, a	108	-	-	-	179	pattern, n	77	-	-	-
124	calculation, n	107	-	-	-	180	property, n	77	-	-	-
125	oscillation, n	107	-	-	-	181	provide, v	77	+	-	-
126	calculate, v	106	+	-	-	182	upper, a	77	-	-	-
127	cell, n	105	-	-	-	183	bandwidth, n	75	-	-	-
128	molecule, n	105	-	-	-	184	polarization, n	75	-	-	-
129	form, n	103	-	-	-	185	cross, n	74	-	-	-
130	increase, v	103	+	-	-	186	emit, v	74	-	-	-
131	paper, n	102	+	-	-	187	excitation, n	73	-	-	-
132	figure, n	101	-	-	-	188	pulsed, a	73	-	-	-
133	reflection, n	101	-	-	-	189	atomic, a	72	-	-	-
134	detector, n	98	-	-	-	190	develop, v	72	+	-	-
135	application, n	97	+	-	-	191	diffraction, n	72	-	-	-
136	guide, n	97	-	-	-	192	result (in), v	72	+	-	-
137	spectral, a	96	-	-	-	193	ratio, n	71	+	-	-
138	stimulated, a	96	-	-	-	194	unit, n	71	-	-	-
139	width, n	95	-	-	-	195	approximation, n	70	-	-	-
140	diode, n	94	-	-	-	196	change, n	70	-	-	-
141	input, n	94	-	-	-	197	effective, a	70	+	-	-
142	lasing, g	94	-	-	-	198	shift, n	70	-	-	-
143	total, a	94	+	-	-	199	development, n	69	+	-	-
144	generate, v	93	+	-	-	200	lamp, n	69	-	-	-
145	grating, g	93	-	-	-	201	minimum, n	69	+	-	-
146	thermal, a	92	-	-	-	202	profile, n	69	-	-	-
147	excite, v	91	-	-	-	203	single, a	69	+	-	-
148	operate, v	91	+	+	-	204	amplification, n	68	-	-	-
149	resonance, n	91	-	-	-	205	CW (cw), a/n	68	+	-	-
150	matrix, n	90	-	-	-	206	rotational, a	68	-	-	-
151	point, n	90	-	-	-	207	size, n	67	+	-	-
152	diameter, n	89	-	-	-	208	vector, n	67	-	-	-
153	population, n	89	+	-	-	209	dielectric, a	66	-	-	-
154	propagation, n	89	-	-	-	210	aperture, n	65	-	-	-

211	circuit, n	65	-	-	-	268	tunable, a	53	-	-	-
212	curve, n	65	+	-	-	269	characteristic, n	52	+	-	-
213	mechanism, n	65	-	-	-	270	define, v	52	+	-	-
214	parametric, a	65	-	-	-	271	distance, n	52	-	-	-
215	report, v	65	-	-	-	272	external, a	52	-	-	-
216	design, n	64	-	-	-	273	set, n	52	+	-	-
217	linear, a	64	+	-	-	274	decay, n	51	-	-	-
218	spontaneous, a	64	-	-	-	275	mechanical, a	51	-	-	-
219	flow, n	63	-	-	-	276	relative, a	51	+	-	-
220	dependence, n	62	+	-	-	277	sample, n	51	-	-	-
221	parallel, a	62	-	-	-	278	chemical, a	50	+	-	-
222	resonant, a	62	-	-	-	279	fraction, n	50	+	-	-
223	space, n	62	+	-	-	280	fringe, n	50	-	-	-
224	target, n	62	-	-	-	281	generation, n	50	-	-	-
225	electrode, n	61	-	-	-	282	heat, n	50	-	-	-
226	reflect, v	61	+	-	-	283	predict, v	50	+	-	-
227	scatter, v	61	-	-	-	284	ruby, n	50	-	-	-
228	expression, n	60	+	-	-	285	study, n	50	+	-	-
229	phenomenon, n	60	-	-	-	286	flux, n	49	-	-	-
230	volume, n	60	-	-	-	287	normal, a	49	+	-	-
231	complex, a	59	-	-	-	288	satisfy, v	49	-	-	-
232	operator, n	59	-	-	-	289	scattering, n	49	-	-	-
233	rod, n	59	-	-	-	290	couple, v	48	+	-	-
234	stability, n	59	-	-	-	291	nitrogen, n	48	-	-	-
235	vapor, n	59	-	-	-	292	semiconductor, n	48	-	-	-
236	communication, n	58	-	-	-	293	substrate, n	48	-	-	-
237	plate, n	58	-	-	-	294	hologram, n	47	-	-	-
238	reflection, n	58	-	-	-	295	molecular, a	47	+	-	-
239	spin, n	58	-	-	-	296	perturbation, n	47	-	-	-
240	absorb, v	57	+	-	-	297	pumping, g	47	-	-	-
241	spatial, a	57	-	-	-	298	unstable, a	47	-	-	-
242	theoretical, a	57	-	-	-	299	water, n	47	-	-	-
243	amplify, v	56	+	-	-	300	disk, n	46	-	-	-
244	approximately,adv	56	+	-	-	301	fluorescence, n	46	-	-	-
245	lense, n	56	-	-	-	302	infrared, a	46	-	-	-
246	mixing, n	56	-	-	-	303	portion, n	46	-	-	-
247	proportional, a	56	+	-	-	304	propagate, v	46	+	-	-
248	work, n	56	+	-	-	305	solid-state, a	46	-	-	-
249	area, n	55	-	-	-	306	stable, a	46	+	-	-
250	concentration, n	55	-	-	-	307	excited, a	45	-	-	-
251	inversion, n	55	-	-	-	308	hole, n	45	-	-	-
252	magnitude, n	55	+	-	-	309	increase, n	45	+	-	-
253	motion, n	55	-	-	-	310	long, a	45	+	-	-
254	plasma, n	55	-	-	-	311	amount, n	44	-	-	-
255	position, n	55	+	-	-	312	continuous, a	44	+	-	-
256	action, n	54	+	-	-	313	longitudinal, a	44	-	-	-
257	average, a	54	-	-	-	314	modulator, n	44	-	-	-
258	cyclotron, n	54	-	-	-	315	numerical, a	44	-	-	-
259	dc, a/n	54	-	-	-	316	refractive, a	44	-	-	-
260	delay, n	54	-	-	-	317	torr, n	44	-	-	-
261	Raman, pn(attr)	54	-	-	-	318	agreement, n	43	+	-	-
262	vibrational, a	54	-	-	-	319	mixture, n	43	-	-	-
263	Table (table),n	53	-	-	-	320	modulate, v	43	-	-	-
264	investigate, v	53	+	-	-	321	strength, n	43	-	-	-
265	liquid, a	53	-	-	-	322	sum, n	43	+	-	-
266	quantity, n	53	+	-	-	323	efficient, a	42	-	-	-
267	relaxation, n	53	-	-	-	324	equilibrium, n	42	-	-	-

325	highly, adv	42	+	-	-	382	shock, n	36	-	-	-
326	intence, a	42	+	-	-	383	spot, n	36	-	-	-
327	lifetime, n	42	-	-	-	384	computer, n	35	-	-	-
328	limit, n	42	+	-	-	385	gap, n	35	-	-	-
329	relation, n	42	+	-	-	386	interference, n	35	+	-	-
330	tune, v	42	-	-	-	387	scheme, n	35	-	-	-
331	constant, a	41	-	-	-	388	Stark, pn(attr)	35	-	-	-
332	decrease, v	41	+	-	-	389	superconductor, n	35	-	-	-
333	dimention, n	41	+	-	-	390	axial, a	34	-	-	-
334	electronic, a	41	+	-	-	391	dipole, n	34	-	-	-
335	feedback, n	41	-	-	-	392	duration, n	34	-	-	-
336	FIR, a/a	41	-	-	-	393	electrooptic, a	34	-	-	-
337	fluctuation, n	41	-	-	-	394	experimentally, adv	34	+	-	-
338	Gaussian, a	41	+	-	-	395	fiber, n	34	-	-	-
339	geometry, n	41	+	-	-	396	normalize, v	34	-	-	-
340	information, n	41	+	-	-	397	orientation, n	34	-	-	-
341	potential, n	41	-	-	-	398	reaction, n	34	-	-	-
342	reflectivity, n	41	-	-	-	399	receiver, n	34	-	-	-
343	relativistic, a	41	-	-	-	400	solve, v	34	+	-	-
344	short, a	41	+	-	-	401	strong, a	34	+	-	-
345	cathode, n	40	-	-	-	402	transmit, v	34	+	-	-
346	estimate, v	40	+	-	-	403	vacuum, n	34	+	-	-
347	filter, n	40	+	-	-	404	bias, n	33	-	-	-
348	finite, a	40	+	-	-	405	charge, n	33	+	-	-
349	formula, n	40	-	-	-	406	consist, v	33	-	-	-
350	macroscopic, a	40	-	-	-	407	error, n	33	-	-	-
351	neon, n	40	-	-	-	408	face, n	33	-	-	-
352	transfer, n	40	-	-	-	409	interferometer, n	33	-	-	-
353	visible, a	40	-	-	-	410	narrow, a	33	+	-	-
354	center, n	39	+	-	-	411	observed, a	33	+	-	-
355	frame, n	39	-	-	-	412	oscillate, v	33	+	-	-
356	mount, v	39	+	-	-	413	prism, n	33	-	-	-
357	probability, n	39	-	-	-	414	TE, a/a	33	-	-	-
358	product, v	39	+	-	-	415	window, n	33	-	-	-
359	copper, a	38	-	-	-	416	analyze, v	32	+	-	-
360	DFB, PII/n	38	-	-	-	417	applied, a	32	+	-	-
361	helium, n	38	+	-	-	418	assumption, n	32	+	-	-
362	observation, n	38	+	-	-	419	boundary, a	32	-	-	-
363	optimum, n	38	+	-	-	420	chromium, n	32	-	-	-
364	physical, a	38	+	-	-	421	Josephson, pn(attr)	32	-	-	-
365	probe, a	38	-	-	-	422	rectangular, a	32	-	-	-
366	transmission, n	38	-	-	-	423	research, n	32	+	-	-
367	aluminium, n	37	-	-	-	424	symmetric(al), a	32	+	-	-
368	control, v	37	+	-	-	425	use, v	32	-	-	-
369	detect, v	37	-	-	-	426	air, n	31	-	-	-
370	incident, a	37	-	-	-	427	conventional, a	31	+	-	-
371	radar, a	37	-	-	-	428	coupler, n	31	-	-	-
372	travel, v	37	-	-	-	429	diffusion, n	31	-	-	-
373	treatment, n	37	-	-	-	430	Doppler, pn(attr)	31	-	-	-
374	accurate, a	36	+	-	-	431	equivalent, a	31	+	-	-
375	carrier, a	36	-	-	-	432	flashlamp, n	31	-	-	-
376	coherence, n	36	+	-	-	433	interval, n	31	-	-	-
377	image, n	36	-	-	-	434	locking, g	31	-	-	-
378	optics, n	36	-	-	-	435	negative, a	31	-	-	-
379	performance, n	36	+	-	-	436	polarized, a	31	-	-	-
380	ray, n	36	-	-	-	437	reduction, n	31	+	-	-
381	ring, n	36	-	-	-	438	accuracy, n	30	+	-	-

439	arrangement, n	30	+	-	-	497	approximate, a	25	- - -
440	boundary, n	30	-	-	-	498	decrease, n	25	+ - -
441	circular, a	30	-	-	-	499	direct, a	25	+ - -
442	growth, n	30	-	-	-	500	limit, v	25	+ - -
443	instability, n	30	-	-	-	501	non-linearity, n	25	- - -
444	quality, n	30	-	-	-	502	polarize, v	25	+ - -
445	spacing, n	30	+	-	-	503	solid, a	25	+ - -
446	broadening, n	29	-	-	-	504	trap, v	25	- - -
447	channel, n	29	-	-	-	505	vibration, n	25	- - -
448	cutoff, n	29	-	-	-	506	atmosphere, n	24	- - -
449	front, n	29	-	-	-	507	atmospheric, a	24	- - -
450	Hamiltonian, n	29	-	-	-	508	cool, v	24	+ - -
451	internal, a	29	+	-	-	509	deposition, n	24	- - -
452	move, v	29	+	-	-	510	instrument, n	24	- - -
453	optically, adv	29	+	-	-	511	mean, a	24	+ - -
454	path, n	29	-	-	-	512	momentum, n	24	- - -
455	radial, a	29	-	-	-	513	overall, a	24	- - -
456	scattering, a	29	-	-	-	514	photomultiplier, n	24	- - -
457	attenuation, n	28	+	-	-	515	plot, n	24	+ - -
458	collision, n	28	+	-	-	516	radiate, v	24	+ - -
459	contact, n	28	-	-	-	517	reference, n	24	- - -
460	heating, g	28	-	-	-	518	reservoir, n	24	- - -
461	mercury, n	28	-	-	-	519	resistance, n	24	+ - -
462	operating, a	28	+	-	-	520	RF (r-f, r.f.), a/n	24	- - -
463	pass, v	28	-	-	-	521	segment, n	24	- - -
464	perpendicular, a	28	+	-	-	522	single-mode, n	24	- - -
465	pure, a	28	+	-	-	523	speed, n	24	- - -
466	pyroelectric, a	28	-	-	-	524	steady, a	24	+ - -
467	refraction, n	28	-	-	-	525	test, n	24	- - -
468	responce, n	28	-	-	-	526	wire, n	24	- - -
469	shape, n	28	+	-	-	527	zero-order, a	24	- - -
470	yield, v	28	+	-	-	528	adjust, v	23	+ - -
471	arc, n	27	+	-	-	529	analytical, a	23	+ - -
472	argon, n	27	-	-	-	530	buffer, a	23	- - -
473	combination, n	27	+	-	-	531	degradation, n	23	- - -
474	dispersion, n	27	-	-	-	532	detection, n	23	- - -
475	focal, a	27	-	-	-	533	earth, n	23	- - -
476	focus, n	27	-	-	-	534	evaluate, v	23	- - -
477	high-power, a	27	+	-	-	535	expansion, n	23	+ - -
478	integrated, a	27	+	-	-	536	illuminate, v	23	- - -
479	interface, n	27	-	-	-	537	moving, a	23	- - -
480	investigation, n	27	+	-	-	538	pair, n	23	- - -
481	regime, n	27	+	-	-	539	paramagnetically, adv	23	- - -
482	resulting, a	27	+	-	-				
483	static, a	27	-	-	-	540	focus, v	23	+ - -
484	stationary, a	27	-	-	-	541	scattered, a	23	- - -
485	susceptibility, n	27	-	-	-	542	second, n	23	- - -
486	UV (uv), a	27	-	-	-	543	situation, n	23	- - -
487	coupled, a	26	-	-	-	544	stored, a	23	- - -
488	force, n	26	+	-	-	545	temporal, a	23	- - -
489	injected, a	26	-	-	-	546	xenon, a	23	- - -
490	measure, n	26	+	-	-	547	wavefront, n	23	- - -
491	moment, n	26	-	-	-	548	ammonia, a	22	- - -
492	procedure, n	26	+	-	-	549	collisional, a	22	- - -
493	recombination, n	26	-	-	-	550	confocal, a	22	- - -
494	semiclassical, a	26	-	-	-	551	exposure, n	22	- - -
495	steady-state, a	26	-	-	-	552	fixed, a	22	+ - -
496	angular, a	25	+	-	-	553	integral, n	22	+ - -

554	ionization, n	22	-	-	-	612	estimate, n	19	-	-	-
555	monochromator, n	22	-	-	-	613	fabricate, v	19	-	-	-
556	object, n	22	-	-	-	614	flow, v	19	-	-	-
557	oxide, s	22	+	-	-	615	gate, n	19	-	-	-
558	periodic, a	22	-	-	-	616	geometric(al), a	19	-	-	-
559	plot, v	22	-	-	-	617	HF, a/n	19	-	-	-
560	stream, n	22	-	-	-	618	lock, v	19	-	-	-
561	stress, n	22	-	-	-	619	parasitic, a	19	-	-	-
562	submillimeter, n	22	-	-	-	620	prediction, n	19	+	-	-
563	telescope, n	22	-	-	-	621	production, n	19	-	-	-
564	use, n	22	-	-	-	622	quartz, a	19	+	-	-
565	Bragg, pn(attr)	21	-	-	-	623	resultant, a	19	-	-	-
566	calculated, a	21	+	-	-	624	screen, n	19	-	-	-
567	characteristic, a	21	-	-	-	625	splitter, n	19	-	-	-
568	conduction, n	21	-	-	-	626	statistical, a	19	-	-	-
569	conductivity, n	21	-	-	-	627	uncertainty, n	19	+	-	-
570	continuously, adv	21	+	-	-	628	uniform, a	19	-	-	-
571	cylindrical, a	21	+	-	-	629	watt, n	19	-	-	-
572	dispersive, a	21	-	-	-	630	adjacent, a	18	-	-	-
573	disturbance, n	21	-	-	-	631	approach, n	18	-	-	-
574	drift, n	21	-	-	-	632	burst, n	18	-	-	-
575	eigenvalue, n	21	-	-	-	633	compound, n	18	-	-	-
576	Fourier, pn(attr)	21	+	-	-	634	cylinder, n	18	-	-	-
577	gallium, n	21	-	-	-	635	depth, n	18	-	-	-
578	metal, a	21	+	-	-	636	display, v	18	-	-	-
579	nozzle, n	21	-	-	-	637	examine, v	18	-	-	-
580	pass, n	21	+	-	-	638	formation, n	18	-	-	-
581	positive, a	21	-	-	-	639	linearly, adv	18	-	-	-
582	potential, a	21	-	-	-	640	kinetic, a	18	-	-	-
583	radiative, a	21	+	-	-	641	nature, n	18	-	-	-
584	replace, v	21	+	-	-	642	nuclear, a	18	-	-	-
585	rotation, n	21	-	-	-	643	physics, n	18	-	-	-
586	superconducting, a	21	-	-	-	644	planar, a	18	-	-	-
587	trace, n	21	-	-	-	645	remove, v	18	-	-	-
588	ultraviolet, a	21	-	-	-	646	resolution, n	18	-	-	-
589	analyzer, n	20	-	-	-	647	satellite, n	18	-	-	-
590	curvature, n	20	+	-	-	648	slit, n	18	-	-	-
591	differential, a	20	+	-	-	649	spike, n	18	-	-	-
592	coordinate, n	20	+	-	-	650	stage, n	18	-	-	-
593	Fabry-Perot, a	20	-	-	-	651	TEM, a/a	18	-	-	-
594	harmonic, a	20	-	-	-	652	traveling, a	18	+	-	-
595	magnification, n	20	-	-	-	653	treat, v	18	-	-	-
596	maximize, v	20	+	-	-	654	absolute, a	17	-	-	-
597	metastable, a	20	-	-	-	655	aircraft, n	17	-	-	-
598	silicon, n	20	-	-	-	656	arsenide, a	17	-	-	-
599	standard, n	20	-	-	-	657	concept, n	17	+	-	-
600	symmetry, n	20	-	-	-	658	convert, v	17	+	-	-
601	transformation, n	20	-	-	-	659	core, n	17	-	-	-
602	transparent, a	20	+	-	-	660	diagonal, a	17	-	-	-
603	acceptor, n	19	-	-	-	661	displacement, n	17	+	-	-
604	array, n	19	-	-	-	662	drop, v	17	-	-	-
605	basis, n	19	+	-	-	663	enhancement, n	17	-	-	-
606	bend, n	19	-	-	-	664	Fresnel, pn(attr)	17	-	-	-
607	construct, v	19	-	-	-	665	FWHM, a/n/n/n	17	-	-	-
608	DF , PII/n	19	-	-	-	666	irradiation, n	17	-	-	-
609	divergence, n	19	+	-	-	667	millimeter, n	17	-	-	-
610	dynamic(al), a	19	-	-	-	668	nonequilibrium, n	17	-	-	-
611	eigenmode, n	19	+	-	-	669	periodicity, n	17	+	-	-

670	potassium, n	17 - - -	727	create, v	14 - - -
671	pulsewidth, n	17 - - -	728	curved, a	14 - - -
672	rise, n	17 - - -	729	depend, v	14 - - -
673	sink, n	17 - - -	730	derivative, n	14 - - -
674	splitting, n	17 - - -	731	discrete, a	14 - - -
675	square, a	17 - - -	732	doped, a	14 - - -
676	step, n	17 - - -	733	evolution, n	14 - - -
677	stripe, a	17 - - -	734	incoherent, a	14 - - -
678	terminal, a	17 - - -	735	limitation, n	14 - - -
679	time-dependent,a	17 - - -	736	loop, n	14 - - -
680	traveling-wave,a	17 - - -	737	manifold, n	14 - - -
681	zone, a	17 - - -	738	neodymium, n	14 - - -
682	charge, v	16 - - -	739	net, n	14 - - -
683	classical, a	16 - - -	740	publication, n	14 - - -
684	decay, v	16 - - -	741	Q-switched, a	14 - - -
685	degeneracy, n	16 - - -	742	Q-swiching, n	14 - - -
686	distortion, n	16 - - -	743	record, v	14 - - -
687	first-order, a	16 - - -	744	rhodamine, a	14 - - -
688	free-space, a	16 - - -	745	scientist, n	14 - - -
689	induce, v	16 - - -	746	summarize, v	14 - - -
690	induced, a	16 - - -	747	ADP, a/n/n	13 + - -
691	integration, n	16 - - -	748	ambient, a	13 + - -
692	matching, g	16 - - -	749	Brewster,pn(attr)	13 - - -
693	millimeter-wave,a	16 - - -	750	center, v	13 - - -
694	mixer, n	16 - - -	751	central, a	13 - - -
695	oscillating, a	16 - - -	752	coaxial, a	13 - - -
696	phosphate, a	16 - - -	753	composition, n	13 - - -
697	sodium, n	16 - - -	754	computation, n	13 - - -
698	substance, n	16 - - -	755	control, n	13 - - -
699	TEA, a/a/a	16 - - -	756	damping, g	13 - - -
700	three-level, a	16 - - -	757	envelope, n	13 - - -
701	unity, n	16 - - -	758	fabrication, n	13 - - -
702	variable, a	16 - - -	759	Fermi, pn(attr)	13 - - -
703	calibration, n	15 - - -	760	focusing, g	13 - - -
704	cascade, n	15 - - -	761	germanium, n	13 - - -
705	circuitry, n	15 - - -	762	instantaneous, a	13 - - -
706	express, n	15 - - -	763	match, v	13 - - -
707	hollow, a	15 - - -	764	metal, n	13 - - -
708	homogeneous, a	15 - - -	765	mode-locked, a	13 - - -
709	illustrate, v	15 - - -	766	monitore, v	13 - - -
710	impurity, n	15 - - -	767	multiline, a	13 - - -
711	integral, a	15 - - -	768	notation, n	13 - - -
712	ir, a	15 - - -	769	repetition, n	13 - - -
713	lattice, n	15 - - -	770	small-signal, a	13 - - -
714	nanosecond, n	15 - - -	771	species, n	13 - - -
715	selection, n	15 - - -	772	stabilization, n	13 - - -
716	sensitivity, n	15 - - -	773	tensor, n	13 - - -
717	slope, n	15 - - -	774	wall, n	13 - - -
718	spherical, a	15 - - -	775	beamwidth, n	12 - - -
719	absorber, n	14 - - -	776	blocking, a	12 - - -
720	absorbing, a	14 - - -	777	carbon, a	12 - - -
721	additive, a	14 - - -	778	cladding, n	12 - - -
722	adequate, a	14 - - -	779	connect, v	12 + - -
723	birefringence, n	14 - - -	780	enhanced, a	12 - - -
724	boson, n	14 - - -	781	flat, n	12 - - -
725	build, v	14 - - -	782	four-level, a	12 - - -
726	capacitance, n	14 - - -	783	hydrogen, n	12 - - -

784	incoming, a	12	- - -	840	quantum-mechanical, a	10	- - -
785	infrared, n	12	- - -				
786	interacting, a	12	- - -	841	sequence, n	10	- - -
787	lossless, a	12	- - -	842	sputtering, g	10	- - -
788	nonradiative, a	12	- - -	843	strike, v	10	- - -
789	pipe, n	12	- - -	844	altitude, n	9	+ - -
790	sapphire, n	12	- - -	845	domain, n	9	- - -
791	shot, n	12	- - -	846	dynamics, n	9	- - -
792	shutter, n	12	- - -	847	electron-hole, a	9	- - -
793	TM, a/a	12	- - -	848	flash, n	9	- - -
794	transform, n	12	- - -	849	gaseous, a	9	- - -
795	amplifying, a	11	+ - -	850	green, a	9	- - -
796	analog(ue), n	11	+ - -	851	groove, n	9	- - -
797	ASE, PII/a/n	11	- - -	852	calcium, n	9	- - -
798	background, n	11	- - -	853	conclusion, n	9	- - -
799	Boltzmann, n	11	- - -	854	low-power, a	9	- - -
800	Brillouin, pn(attr)	11	- - -	855	multilayer, a	9	- - -
801	chamber, n	11	- - -	856	non-uniform, a	9	- - -
802	collimate, v	11	- - -	857	off-diagonal, a	9	- - -
803	creation, n	11	- - -	858	reflectance, n	9	- - -
804	depletion, n	11	- - -	859	root, n	9	- - -
805	dissipation, n	11	- - -	860	Shroedinger, a	9	- - -
806	electrostatic, a	11	- - -	861	strip, n	9	- - -
807	fluid, n	11	- - -	862	tapered, a	9	- - -
808	heater, n	11	- - -	863	body, n	8	- - -
809	Lorentzian, n	11	- - -	864	bound, n	8	- - -
810	low-noise, a	11	- - -	855	chloride, n	8	- - -
811	outer, a	11	- - -	866	class, n	8	- - -
812	photodiode, a	11	- - -	867	collection, n	8	- - -
813	phototube, n	11	- - -	868	demonstrate, v	8	- - -
814	prove, v	11	- - -	869	formulation, n	8	- - -
815	red, a	11	- - -	870	lase, v	8	- - -
816	streaming, g	11	- - -	871	megacycle, n	8	- - -
				872	miniature, a	8	- - -
817	superconductivity, n	11	- - -	873	photoelectron, n	8	- - -
818	superposition, n	11	- - -	874	picture, n	8	- - -
819	switch, n	11	- - -	875	quantization, n	8	- - -
820	train, n	11	- - -	876	quasi-Fermi, a	8	- - -
821	transmitter, n	11	- - -	877	radiating, a	8	- - -
822	tunability, n	11	- - -	878	saturated, a	8	- - -
823	tunneling, a	11	- - -	879	stainless-steel, a	8	- - -
824	two-photon, a	11	- - -	880	synchronism	8	- - -
825	uniformity, n	11	- - -	881	tunneling, n	8	- - -
826	valence, n	11	- - -	882	Bloch, pn(attr)	7	- - -
827	asymmetry, n	10	- - -	883	column, n	7	- - -
828	bent, n	10	- - -	884	constituent, n	7	- - -
829	bond, n	10	- - -	885	diffracted, a	7	- - -
830	carry, v	10	- - -	886	garnet, n	7	- - -
831	elastic, a	10	- - -	887	lithium, n	7	- - -
832	epitaxial, a	10	- - -	888	quantized, a	7	- - -
833	height, n	10	- - -	889	self-focusing, g	7	- - -
834	inhomogeneous, a	10	- - -	890	third-order, a	7	- - -
835	irradiance, n	10	- - -	891	translational, a	7	- - -
836	luminescence, n	10	- - -	892	vanish, v	7	- - -
837	near-field, a	10	- - -	893	algorithm, n	7	- - -
838	nonresonant, a	10	- - -	894	alkali, a	6	- - -
839	passive, a	10	- - -	895	antiparallel, a	6	- - -

896	bonding, g	6	-	-	-	950	power output	16	-	-	-
897	cadmium, n	6	-	-	-	951	in phase	16	-	-	-
898	coated, a	6	-	-	-	952	input power	16	-	-	-
899	collinear, a	6	-	-	-	953	excited atom	16	-	-	-
900	energetic, a	6	-	-	-	954	laser light	15	-	-	-
901	exit, n	6	-	-	-	955	TEM mode	15	-	-	-
902	increment, n	6	-	-	-	956	light pulse	15	-	-	-
903	jet, n	6	-	-	-	957	ground state	15	-	-	-
904	reconstruction, n	6	-	-	-	958	laser transition	15	-	-	-
905	transmittance, n	6	-	-	-	959	reflection coeffi- cient	14	-	-	-
906	waveguiding, a	6	-	-	-			14			
907	stimulated emission	54	-	-	-	960	conversion effi- ciency	14	-	-	-
908	electrical field	54	-	-	-	961	active material	14	-	-	-
909	magnetic field	54	-	-	-	962	transverse mode	14	-	-	-
910	for example	53	+	-	-	963	integrated optics	14	-	-	-
911	in terms of	53	+	-	-	964	unstable resonator	14	-	-	-
912	output power	37	-	-	-	965	electron beam	13	-	-	-
913	laser beam	35	+	-	-	966	coupling coeffici- ent	13	-	-	-
914	wave function	35	-	-	-						
915	injection laser	35	-	-	-						
916	density matrix	34	-	-	-	967	laser diode	13	-	-	-
917	energy level	32	-	-	-	968	longitudinal mode	13	-	-	-
918	active region	31	-	-	-	969	unit volume	13	-	-	-
919	laser system	31	-	-	-	970	light beam	12	-	-	-
920	of the order of	28	+	-	-	971	electrooptic cry- stal	12	-	-	-
921	radiation field	26	-	-	-						
922	power level	25	-	-	-	972	electron densuty	12	-	-	-
923	cross section	25	-	-	-	973	wave equation	12	-	-	-
924	refractive index	23	+	-	-	974	Doppler shift	12	-	-	-
925	coherent light	23	-	-	-	975	beam splitter	12	-	-	-
926	current pulse	23	-	-	-	976	absorption band	11	-	-	-
927	laser action	22	+	-	-	977	optical cavity	11	-	-	-
928	population inver- sion	22	+	-	-	978	coupling loss	11	-	-	-
						979	gain profile	11	-	-	-
929	laser('s) output	22	-	-	-	980	frequency range	11	-	-	-
930	gas laser	21	-	-	-	981	steady-state so- lution	11	-	-	-
931	laser pulse	21	+	-	-						
932	boundary condition	20	-	-	-	982	coherent state	11	-	-	-
933	spontaneous emis- sion	20	+	-	-	983	semiclassical the- ory	11	-	-	-
934	ruby laser	20	-	-	-	984	incident wave	11	-	-	-
935	quantum state	20	-	-	-	985	valence band	10	-	-	-
936	electromagnetic wave	20	-	-	-	986	absorption coeffi- cient	10	-	-	-
937	matrix element	19	-	-.-		987	threshold current	10	-	-	-
938	current density	18	-	-	-	988	noise figure	10	-	-	-
939	equation of motion	18	+	-	-	989	TE mode	10	-	-	-
940	rate equation,	18	+	-	-	990	ruby rod	10	-	-	-
941	excited state	18	+	-	-	991	macroscopic quan- tum state	10	-	-	-
942	dye laser	18	-	-	-						
943	cavity length	17	-	-	-	992	responce time	10	-	-	-
944	optical maser	17	-	-	-	993	millimeter wave- length	10	-	-	-
945	pump power	17	-	-.-							
946	above threshold	17	-	-	-	994	resonance condi- tion	9	-	-	-
947	conduction band	16	-	-	-						
948	laser energy	16	-	-	-	995	threshold condi- tion	9	-	-	-
949	index of refrac- tion	16	-	-	-						

996	in the laboratory frame	9	-	-	-	1011	absorbing molecule	8	-	-	-
997	Josephson('s) junction	9	-	-	-	1012	relaxation rate	8	-	-	-
998	lower laser level	9	-	-	-	1013	liquid nitrogen	7	-	-	-
999	solid-state maser	9	-	-	-	1014	optical pulse	7	-	-	-
1000	relative phase	9	-	-	-	1015	pumping pulse	7	-	-	-
1001	time scale	9	-	-	-	1016	spectral region	7	-	-	-
1002	phase shift	9	-	-	-	1017	time rate	7	-	-	-
1003	shock wave	9	-	-	-	1018	heat sink	7	-	-	-
1004	parametric amplifier	8	-	-	-	1019	transition temperature	7	-	-	-
1005	Shrodinger('s) equation	8	-	-	-	1020	electronic transition	7	-	-	-
1006	upper energy level	8	-	-	-	1021	optical axis	6	-	-	-
1007	absorption loss	8	-	-	-	1022	probability density	6	-	-	-
1008	ammonia maser	8	-	-	-	1023	Boltzmann distribution	6	-	-	-
1009	phase matching	8	-	-	-	1024	quasi-Fermi level	6	-	-	-
1010	waveguide mode	8	-	-	-	1025	spot size	6	-	-	-

Анализ распределений терминов в английском подъязыке квантовой электроники показал, что распределение большинства терминологических единиц, 817 из 1025 обследованных, что составляет 79,71 %, не подчиняются ни одному из вышеназванных теоретических законов. Из оставшихся 208 терминов только законом Пуассона описываются 196 единиц, только нормальным, только логнормальным, нормальным и логнормальным одновременно -- по 3 единицы, законом Пуассона и нормальным законом -- 2, Пуассона и логнормальным -- 1.

Полученные результаты вполне согласуются с мнением о том, что неподчинение законам Пуассона и нормальному свидетельствует о терминологичности лексических единиц (ср., Бектаев, Лукьяненков, с.109; Пиотровский, с.119).

Следует отметить, что границы терминологичности/нетерминологичности, выраженные в соответствии/несоответствии исследуемым теоретическим законам, не имеют четко обозначенного характера. Это говорит о том, что термины, как практически любое лингвистическое множество, является множеством с размытыми краями (ср., например, Пиотровский, с.208).

Л И Т Е Р А Т У Р А

Бектаев, К. Б., Лукьяненков, К. Ф. О законах распределения единиц письменной речи. – Статистика речи и автоматический анализ текста. Л., 1971.

Вентцель., Е. С. Теория вероятностей. М., 1969.

Каширина, М. Е. О типах распределения лексических единиц в тексте. – Статистика речи и автоматический анализ текста. Л., 1974.

Митропольский, А. К. Техника статистических вычислений. М., 1971.

Пиотровский, Р. Г. Текст, машина, человек. Л., 1975.

Пиотровский, Р. Г., Бектаев, К. Ф., Пиотровская, А. А. Математическая лингвистика. М., 1977.

ON DISTRIBUTION ANALYSIS IN ENGLISH SCIENTIFIC TEXTS (SUBLANGUAGE OF ACTIVE OSCILLATORS)

Narinay Manasyan

S u m m a r y

The article presents a description of a distribution analysis experiment (English texts on active oscillators). For the experiment a frequency dictionary on active oscillators was used. During compilation of the frequency dictionary only terms were registered. The terms were lemmatized. For the analysis 1,025 units were chosen from the frequency dictionary. The terminological units analysis showed that most terms, 817 (79.71 %), were not governed by any of the theoretical laws. Of the remaining 208 units, 196 terms were governed only by Poisson's Law, only by Normal Law - 3, by Lognormal - 3, by Normal and Lognormal at the same time - 3, by Poisson's and Normal - 2, by Poisson's and Lognormal - 1.

ОБ ИЗМЕРЕНИИ СВЯЗИ ОТРАСЛЕВЫХ ТЕРМИНОСИСТЕМ С ПРИМЕНЕНИЕМ ЭВМ

М.А. Марусенко

Измерение связей реально функционирующих терминосистем на лексическом уровне является одним из важнейших элементов системного подхода к анализу терминосистем. В настоящее время не существует сколько-нибудь разработанной лингвистической теории, позволяющей более или менее объективно производить выделение подъязыков и терминосистем. На практике для этого применяются наукометрические критерии: сколько наук, столько и "языков", сколько разделов имеет наука, столько у нее и "подъязыков". Однако такой подход не позволяет производить объективное членение, так как классификационное деление наук имеет очень много уровней и может вестись практически бесконечно в соответствии с углубляющимся членением предмета науки и с процессами интеграции и дифференциации наук, интенсивность которых особенно усилилась на настоящем этапе научно-технической революции. Свидетельством этого могут служить "Универсальная десятичная классификация" и "Рубрикатор изданий СССР", непрерывно пополняющиеся новыми рубриками. Кроме того, такое членение зависит и от национальных традиций, сложившихся в процессе развития науки в тех или иных исторических условиях.

В статье критерии обоснованности выделения терминосистем рассматриваются на материале французской радиотехнической терминологии. С точки зрения наукометрии существование подъязыка и, соответственно, терминосистемы радиотехники не вызывает никаких сомнений. Это подтверждается существованием радиотехнических ВУЗ-ов, факультетов, кафедр, специальностей в области научно-технической информации - большим числом специальных отраслевых изданий, монографий, в области терминоведения - наличием специальных радиотехнических словарей и т.д. С другой стороны, радиотехника, наряду с электроникой, входит в состав науки более высокого яруса - радиоэлектроники. Кроме того, сама радиотехника состоит из значительного числа разделов, представляющих собой целые направления современной науки и техники, таких как телевидение, радиолокация, радиоастрономия, радиосвязь и т.д.

Учитывая все это, более конструктивным представляется подход к выделению терминосистем, основанный на анализе системных характеристик и собственно лингвистических особенностей реально функционирующих совокупностей научно-технических терминов.

Для каждой отрасли науки и техники можно построить иерархию подъязыков, причем термины, употребляющиеся во многих предметных областях, образуют терминосистемы, находящиеся на вершине иерархии, а термины, употребляющиеся в отдельных областях, образуют терминосистемы отдельных подъязыков. Такие лексические единицы, которые принадлежат одной области и не воспроизводятся во всех остальных, образуют элементарную терминосистему.

В этих условиях изучение связей терминосистем приобретает особое значение, так как неконтролируемое изменение условий может вывести объект за границы допустимого размытия функциональных состояний. Предметом изучения должны стать пределы изменения условий, при которых объект остается идентичным самому себе (Мельников Г.П., с. 55).

Если имеются две совокупности терминов, то между ними могут существовать следующие отношения:

I. $M_i \subset M_j$

2. $M_i \cap M_j \neq \emptyset$

3. $M_i = M_j$

4. $M_i \cap M_j = \emptyset$

I. Одна терминосистема может включаться в другую; можно говорить о терминосистемах разных иерархических уровней (например, радиоэлектроника - радиотехника).

2. Обе терминосистемы могут пересекаться; в этом случае можно говорить о двух терминосистемах одного уровня (например, радиолокация - радионавигация) либо о терминосистемах разных уровней (например, радионавигация - астрономия).

3. Обе терминосистемы тождественны; в этом случае выделение двух объектов по эмпирическим основаниям может быть ошибочным.

4. Две терминосистемы не имеют общих элементов, находятся в альтернативном отношении; в этом случае они не принадлежат одному и тому же естественному языку.

Таким образом, изучая связи элементов двух терминосистем и давая им количественную оценку, можно устанавливать иерархию отношений между этими системами и, тем самым, оценивать смысловую связанность предметных областей, послуживших объектами терминирования.

Необходимо отметить, что такой способ оценки смысловой связи является в настоящее время единственным, так как нахождение такой оценки смысловой связанности двух предметных областей, которая имела бы "подходящую метрику для оценки такой связанности, либо вообще недостижимо, либо в лучшем случае представляет собой дьявольски трудную задачу, причем ни один из известных подходов не имеет шансов привести к определенной функции, которая дает нам полезную меру для оценки смыслового расстояния между темами..." (Bar-Hillel, I.,1964, p. 351).

Наиболее простым методом оценки близости двух классов является использование объемных оценок, например, числа элементов, принадлежащих пересечению двух классов. Такая оценка носит абсолютный характер, однако может рассматривать и оценку, носящую относительный характер, если взять отношение числа элементов, входящих в пересечение двух классов, к общему числу элементов, содержащихся в обоих классах. В том олучае, если неизвестен состав классов, то кроме объемных оценок приходится вводить и содержательные, которые, как было показано выше, трудно поддаются формализации (Балашов Л.А. и др.,1973, с. 6). Составы классов могут быть определены двумя способами: либо задаваться в виде инвентаря элементов (словари научно-технических терминов, информационно-поисковые тезаурусы, ГОСТы), либо в виде списка элементов с указанием удельного веса, значимости каждого элемента в системе (частотные и распределительные словари).

В зависимости от способа задания состава классов определяются и способы количественной оценки их близости. В лингвостатистике использовались различные эмпирические величины для определения взаимодействия лексических систем, такие, например, как приводимые Ш.Мюллером (Muller Ch., 1968) и Ю.А. Тулдава (1974) индексы лексической связи. Однако применение эмпирических индексов не нашло широкого применения и, по ряду причин, не является перспективным. Во-первых, они не дают возможности осуществить статистическую оценку существенности результатов. Целесообразно применять для этого такие величины, для которых известен закон распределения и существуют разработанные методы определения значимости результатов. С этой точки зрения гораздо более перспективным является применение методов корреляционного анализа, предложенное Ю.А. Тулдава.

Кроме того, известно, что согласно общему закону распре-

деления элементов сложных систем, который в лингвистике известен как закон Ципфа (см. Мартыненко Г.Я., 1978), небольшое число элементов системы образует ее ядро и употребляется значительно чаще, чем большое число элементов, составляющих периферию системы. Поэтому при сравнении систем недостаточно учитывать только лишь факт наличия или отсутствия какого-либо элемента в той или иной системе, но важно учитывать также и значимость, удельный вес каждого элемента в каждой лексической системе, образуемой данным текстом или словарем. Такая задача требует выполнения большого числа расчетов и может быть реализована только с помощью ЭВМ.

Любой текст может быть представлен в виде частотного словаря образующих его лексических единиц. Поэтому измерение лексической связи двух частотных словарей может производиться непосредственно, а для текстов необходим дополнительный этап приведения текста к частотному словарю лексем или словосочетаний, в зависимости от цели исследования.

В качестве меры лексической связи предлагается использовать широко применяемую в математической статистике величину - коэффициент корреляции Пирсона:

$$r = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}} \qquad /1/$$

Заменив в формуле /1/ значения x_i и y_i на значения относительных частот какой-либо лексической единицы в том или ином словаре, обозначив f_{i1}^* и f_{i2}^* относительные частоты i-той единицы в словарях 1 и 2 и приняв Σf_{i1}^* и Σf_{i2}^* равными единице, получим:

$$r = \frac{N\Sigma f_{i1} f_{i2} - 1}{\sqrt{N\Sigma f_{i1}^{*2} - 1}\sqrt{N\Sigma f_{i2}^{*2} - 1}} \qquad /2/$$

При равенстве относительных частот всех единиц словаря $r = 1$, в прочих случаях эта величина лежит практически в пределах $0 < r < 1$.

В описываемом эксперименте вычисление коэффициентов лексической связи производилось для словарей статистически устойчивых научно-технических терминов 4 терминосистем подъязыка радиотехники (Марусенко М.А., 1981). Относительные частоты f_{i1}^* и f_{i2}^*, а также коэффициенты лексической связи вычислялись на ЭВМ. Формула /2/ применялась для определения лексических связей между совокупностями статистически устойчивых терминов с учетом жанровой дифференциации. Для каждой пары терминосистем составлялись словари следующего вида:

Таблица I

Фрагмент сопоставительного словаря для определения коэффициента лексической связи

Термины	Частота в т/системе I (f_{i1})	Частота в т/системе 2 (f_{i2})
.....		
capacité	6	46
capacité de circuit	0	2
capacité de couplage	0	2
caractéristique	I3	24
cavité	I06	0
.....		

Результаты вычисления коэффициентов лексической связи могут сведены в таблицы:

Таблица 2

Коэффициенты лексической связи терминосистем подъязыка радиотехники

	Радиопередатчики	Антенны	Радиолокация	Радиоприемники
Радиопередатчики		0,3649	0,7I70	0,67I5
Антенны			0,490I	0,373I
Радиолокация				0,6956

Таблица 3

Коэффициенты лексической связи терминосистем, зафиксированных в монографиях

	Р/перед.	Антенны	Р/лок.	Р/прием.
Р/перед.		0,2583	0,6352	0,4582
Антенны			0,2949	0,2333
Р/лок.				0,58I3

Таблица 4

Коэффициенты лексической связи терминосистем, зафиксированных в журнальных статьях

	Р/перед.	Антенны	Р/лок.	Р/прием.
Р/перед.		0,3068	0,5604 0,4372	0,6016 0,3632 0,6800

Таблица 5

Коэффициенты лексической связи терминосистем разных жанров (монографии - журнальные статьи) внутри одной терминосистемы

Терминосистема	Коэффициент лексической связи
Р/перед.	0,6413
Антенны	0,6504
Р/лок.	0,7816
Р/прием.	0,4674

Поскольку все рассматриваемые терминосистемы принадлежат одному естественному языку, пересечение двух множеств должно быть не пустым: $M_1 \cap M_2 \neq \emptyset$. Если рассматривать эти терминосистемы в их отношениях к надсистеме - терминосистеме радиотехники и между собой, то одинаковое отношение терминосистем к надсистеме и друг к другу должно выражаться одинаковой мерой лексической связи. Отклонение меры лексической связи от средней величины в ту или иную сторону будет соответствовать тенденции либо к тождеству словарей ($M_1 = M_2$), либо к включению одного словаря в другой ($M_1 \subset M_2$). В обоих случаях будет иметь место нарушение иерархической структуры по отношению к надсистеме. Кроме того, различие коэффициентов лексической связи терминосистем разных жанров может послужить еще одним подтверждением неоднородности терминосистем, эмпирически включаемых в "подъязык радиотехники".

Коэффициенты лексической связи двух терминосистем представляют собой случайные величины, распределенные по нормальному закону (см. Митропольский А.К., 1961, с. 273). Поэтому для определения вышеуказанных отношений необходима проверка

статистических гипотез о равенстве или разности двух средних для величин, представляющих собой средние коэффициенты лексической связи для различных терминосистем (таблица 2), для различных жанров внутри одной терминосистемы (таблица 5),для одинаковых жанров разных терминосистем (таблицы 3-4). Средние коэффициентов лексической связи и средние квадратичные отклонения составляют $\bar{x}_1$ = 0,552 и σ_1 = 0,I63I, $\bar{x}_2$= 0,6352 и σ_2 = 0,I289, $\bar{x}_3$= 0,4I02 и σ_3= 0,I73I, $\bar{x}_4$= 0,49I5 и σ_4 = 0,I455 для таблиц 2, 5, 3 и 4 соответственно. Проверка по U-критерию:

$$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n^2}}} \qquad /3/$$

показывает, что нулевая гипотеза H_0: $a_I = a_2$ /9/ отклоняется только для $\bar{x}_2$ и $\bar{x}_3$ /U= 2,3522), т.е., что коэффициенты лексической связи различных терминосистем в целом и в отдельности по жанрам существенно не различаются, в то время как для словарей разных жанров внутри одной терминосистемы они представляют существенно большие величины. Таким образом,результаты статистического анализа показывают, что рассматриваемые совокупности терминов находятся в одинаковой связи друг с другом, т.е. представляют собой подсистемы одного иерархического уровня.

Л И Т Е Р А Т У Р А

Балашов Л.А., Гуськов А.А., Махотенко Ю.А., Смолянов О.Г. Некоторые критерии оценки классификационных систем. - Научно-техническая информация, сер. 2, № 7, I973.

Мартыненко Г.Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах. - В кн.: Структурная и прикладная лингвистика, вып. I. - Л.: Изд-во ЛГУ, I978, с. 63-79.

Марусенко М.А. Системно-сопоставительный анализ функционирования терминосистем (на материале французской радиотехнической терминологии). АКД. Л., I98I.

Митропольский А.К. Техника статистических вычислений. - М.: Физматгиз, I96I.

Мельников Г.П. Системология и языковые аспекты кибернетики.- М.: Советское радио, I978.

Тулдава Ю.А. Об измерении лексической связи текстов на уровне словаря. - В кн.: Вопросы статистической стилистики. - Киев: Наукова думка, 1974, с. 35-42.

Bar-Hillel, I. Language and information. - London: Addoson-Wesley, 1964.

Muller, Ch. Initiation à la statistique linguistique. - Paris: Larousse, 1968.

COMPUTER MEASURING OF LEXICAL CONNECTION OF TERMINOLOGICAL SYSTEMS RELATING TO A CERTAIN BRANCH OF INDUSTRY

Mikhail A. Marusenko

S u m m a r y

Consideration is given to methods of measuring the interconnection of complex linguistic systems as exemplified by four radio-engineering sub-language terminological systems. An approach is adopted based on registering the value of every lexical unit in the respective system.

The new quantitative criterion for measuring the connection is worked out, and a linguostatistical experiment is realized on the computer.

The experimental data obtained have been evaluated and tested.

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ФОНОЛОГИЧЕСКОЙ СТРУКТУРЫ СЛОВА (НА МАТЕРИАЛЕ ОДНОСЛОЖНЫХ СЛОВ РЯДА ИНДОЕВРОПЕЙСКИХ И КАЗАХСКОГО ЯЗЫКОВ)

Г.Я. Панкрац

Из комплекса вопросов о фонологической структуре слова в статье разбираются следующие: типы слов по количеству фонем (по длине), среднее количество фонем в словах (средняя длина слов), модели дистрибуции гласных и согласных фонем в словах, место гласной фонемы в словах, консонантные части слов, количество согласных фонем в консонантных частях слов, средняя консонантность консонантных частей слов.

Речь пойдет о фонологической структуре односложных слов, т.е. слов, состоящих из одной гласной и одной, двух и более согласных фонем. Материалом исследования послужили односложные слова ряда индоевропейских языков (русского, белорусского, немецкого, нижненемецкого диалекта в СССР, английского) и одного тюркского (казахского)*.

Ниже дается описание проанализированных слов по указанным семи параметрам.

I. Типы слов по количеству фонем (по длине)

По полученным данным (см. таблицу I) в отношении количественных показателей тех или иных типов слов наиболее близки русский и белорусский языки. В индоевропейских языках односложные слова распределяются по длине по шести типам (одно-, двух-, трех-, четырех-, пяти-, шестифонемные), в казахском языке - по пяти (нет шестифонемных). В индоевропейских языках количество слов от двухфонемного к трехфонемному типу резко возрастает и от четырехфонемного к пятифонемному снова резко падает. В русском, белорусском и немецком языках максимальное количество составляют четырехфонемные слова, в нижненемецком диалекте количество трех- и четырехфонемных слов одинаковое, а в английском языке преобладающее количество слов составляют трехфонемные.

---

* Слова отбирались из разных словарей в их исходной форме (слов русских, белорусских, немецких и нижненемецких примерно по 1000, английских 1400 и казахских 400).

Таблица I

Типы слов по длине (%)

№№ п/п	Типы слов	Языки					
		русский	белорусский	немецкий	нижненемецкий	английский	казахский
I.	Однофонемные	0,7	0,5	0,I	0,3	0,8	0,5
2.	Двухфонемные	3,6	3,9	7,3	5,6	I2,9	I6,2
3.	Трехфонемные	38,8	38,8	40,6	4I,8	49,4	73,6
4.	Четырехфонемные	45,7	45,9	40,9	4I,8	3I,9	9,4
5.	Пятифонемные	I0,3	I0,2	I0,I	I0,I	4,5	0,3
6.	Шестифонемные	0,9	0,7	I,0	0,4	0,5	-

В казахском языке количество слов от однофонемного типа к двухфонемному и к трехфонемному резко возрастает, достигает максимума на трехфонемных словах и от них к четырехфонемным резко падает.

На таблице хорошо виден переход максимума от русских и белорусских слов, у которых четырехфонемный тип слов значительно превосходит все остальные, через немецкий язык, в котором количество трех- и четырехфонемных слов почти одинаковое, и нижненемецкий диалект, в котором эти два типа по количеству равные, к английскому языку, в котором трехфонемные слова составляют около половины слов, и к казахскому, в котором трехфонемные слова составляют около трех четвертей всех слов.

Однофонемных слов во всех шести языках мало, а в казахском также мало пятифонемных.

2. Среднее количество фонем в словах (их средняя длина)

Средняя длина односложных слов в фонемном выражении составляет: в русском языке - 3,64, в белорусском - 3,63, в немецком - 3,62, в нижненемецком диалекте - 3,57, в английском языке - 3,28, в казахском - 2,93.

Средняя длина односложных слов шести языков слева направо, таким образом, постепенно уменьшается. Английское слово на I/I0 короче русского, белорусского и немецкого, а казах-

ское слово короче слова в индоевропейских языках на I/5*.

В литературе имеются указания на связь длины слов с их частотностью в тексте в том смысле, что слова, обладающие большей частотностью, в среднем более короткие, а слова, обладающие меньшей частотностью, в среднем более длинные. Так, Частотный словарь русского языка (I977) начинается с коротких слов (в, и, не, на и т.д.), и средняя длина первой сотни слов составляет 3,53 фонемы, второй сотни - 4,82, третьей - 5,86 и т.д. Эти факты подтверждают положение о связи, существующей между длиной и частотностью слов.

В литературе имеются также указания на связь длины слов с количеством их значений, с их многозначностью. Эту связь можно сформулировать следующим образом: короткие слова в среднем имеют большее количество значений, более многозначны, а более длинные слова менее многозначны (Плоткин, I969).

Вероятность появления слов в тексте с увеличением их длины резко уменьшается. Г.Аренс указывает, что в выборке в II млн. немецких слов восьмисложные слова встретились 5038 раз, девятисложные - I225 раз, слова в I0 слогов - 46I раз, в II слогов - 59 раз, в I2 слогов - 35 раз, в I3 слогов - 8 раз, в I4 слогов - 2 раза, в I5 слогов - I раз (Arens, I965, с.73).

В литературе далее приводятся данные о связи между длиной слова и длиной предложения. Г.Аренс указывает, в частности, на следующую закономерность в немецком языке: чем выше в каком-либо стиле длина слова, тем выше в данном стиле и средняя длина предложения; точнее - с увеличением средней длины слова на 0,05 слога средняя длина предложения увеличивается на 5 слов (Arens, I965, с. 7).

---

* В.А.Никонов в статье "Длина слова" (I978), сравнивая длину слова в русском, грузинском и казахском языках в четырех видах речи (разговорной речи, художественной речи, научной прозе, публицистике), приходит к выводу о большей длине слова в казахском, чем в русском в разговорной речи и художественной прозе, а при сравнении обращения В.И.Ленина "К населению" I8 ноября I9I7 г. (Полн. собр. соч., 35, с.55-57) к выводу о том, что средняя длина слова в казахском самая большая, в грузинском она меньше и в русском самая маленькая, "так как казахское слово и грузинское слово вобрали в себя предлоги и некоторые частицы, существующие в русском языке отдельно". Таково положение в тексте. Мы же берем односложные слова в их исходной форме, т.е. в изолированном виде.

В литературе имеются, наконец, указания на то, что выяснение вопросов архитектоники слова, в том числе их длины, представляет определенный интерес для построения типологической классификации естественных языков, а также для ряда областей прикладной лингвистики.

3. Модели дистрибуции гласных и согласных фонем в словах

В индоевропейских языках (см. таблицу 2) односложные слова разбросаны по 18 моделям дистрибуции фонем (однофонемные в русском и белорусском языках - по двум, в остальных - по одной, двухфонемные - по двум, трехфонемные - по трем, четырехфонемные - по трем, четырехфонемные - по четырем, пятифонемные - по четырем, в остальных - по трем, шестифонемные в русском, белорусском и немецком - по трем, в нижненемецком и английском - по двум).

В казахском языке моделей дистрибуции в два раза меньше - всего восемь (однофонемные слова составляют всего одну модель, двухфонемные - две, трехфонемные - две, четырехфонемные - две, пятифонемные - одну).

Самая распространенная модель дистрибуции во всех словах - это модель СГС; ее удельный вес слева направо нарастает (38,8% - 71,3%). В казахском языке ее удельный вес почти в два раза больше, чем в славянских и немецких языках.

Для двухфонемных слов в индоевропейских языках характерна модель СГ (она охватывает в них около 2/3 всех двухфонемных слов), в казахском - ГС (она также охватывает около 2/3 слов), для трехфонемных во всех языках - СГС (она охватывает 92,3% всех индоевропейских и 96,9% всех казахских трехфонемных слов), для четырехфонемных в индоевропейских языках - ССГС (58,2% всех четырехфонемных слов) и в казахском - СГСС (96,8% четырехфонемных слов), для пятифонемных в индоевропейских языках - ССГСС (64,3% пятифонемных слов) и в казахском языке - СГССС (100% пятифонемных слов), для шестифонемных в индоевропейских языках - СССГСС (54,3% шестифонемных слов).

4. Место гласной фонемы в словах

В двухфонемных словах индоевропейских языков (см. таблицу 3) гласная фонема в два раза чаще занимает конечную позицию, чем начальную; в казахском языке, наоборот, в два раза чаще начальную позицию, чем конечную.

В трехфонемных словах всех языков гласная фонема занимает преимущественно медиальную (среднюю) позицию - в индоев-

Таблица 2

Модели дистрибуции фонем (%)

№№ п/п	Модели дистрибуции	Языки					
		русский	белорусский	немецкий	нижненемецкий	английский	казахский
I.	Г	0,6	0,4	0,I	0,3	0,8	0,5
2.	С	0,I	0,I	-	-	-	-
3.	ГС	I,2	I,2	2,8	I,6	4,0	II,I
4.	СГ	2,4	2,7	4,5	4,0	8,9	5,I
5.	ГСС	0,2	0,2	2,I	I,I	0,8	2,3
6.	СГС	36,8	36,9	36,4	39,2	44,2	7I,3
7.	ССГ	I,8	I,7	2,I	I,5	4,4	-
8.	ГССС	0,2	0,2	0,5	0,3	-	-
9.	СГСС	I5,2	I3,9	20,4	I8,8	I6,I	9,I
I0.	ССГС	30,I	3I,5	I9,9	22,5	I5,6	0,3
II.	СССГ	0,2	0,3	0,I	0,2	0,2	-
I2.	ГСССС	-	-	0,I	-	-	-
I3.	СГССС	I,2	I,I	I,7	I,8	0,5	0,3
I4.	ССГСС	7,0	6,3	6,6	6,7	2,6	-
I5.	СССГС	2,I	2,8	I,7	I,6	I,4	-
I6.	СГСССС	0,I	0,I	0,4	-	-	-
I7.	ССГССС	0,35	0,25	0,2	0,I	0,I	-
I8.	СССГСС	0,45	0,35	0,4	0,3	0,4	-

Таблица 3

Место гласной фонемы в словах (%)

№№ п/п	Место гласной фонемы	Языки					
		русский	белорусский	немецкий	нижненемецкий	английский	казахский
I.	В начальной позиции	I,7	I,6	5,5	3,I	4,9	I3,4
2.	В интерконсонантной позиции	94,0	93,7	87,8	9I,2	8I,4	8I,4
3.	В конечной позиции	4,3	4,7	6,7	5,7	I3,7	5,2

ропейских языках такое положение в 92,3%, в казахском языке - в 96,9% всех трехфонемных слов. Это симметричное расположение согласных фонем в односложных трехфонемных словах можно считать общей для данных языков универсалией.

В четырехфонемных словах в индоевропейских языках гласная фонема расположена большей частью на третьем месте (58,2%), т.е. входит в состав второй половины слов; в казахском же гласная фонема занимает обычно второе место (96,8%), т.е. расположена в первой половине слов.

В пятифонемных словах индоевропейских языков налицо обычно симметричное расположение согласных фонем (64,6% всех пятифонемных слов). В небольшом вообще количестве пятифонемных слов в казахском языке гласная фонема расположена на втором месте, т.е. в первой половине слова (100%).

Для индоевропейских слов с нечетным количеством фонем (тремя, пятью) универсалией является симметричное расположение согласных фонем. В казахском языке такое расположение распространяется лишь на трехфонемные.

В словах с четным количеством фонем заметна тенденция к помещению гласных фонем в индоевропейских языках во второй половине слов, а в казахском языке - в первой половине.

Во всех языках гласная фонема чаще всего занимает интерконсонантную позицию. В индоевропейских языках гласная фонема чаще занимает конечную позицию, чем начальную (в два с лишним раза); в казахском языке наоборот - чаще начальную, чем конечную (также в два с лишним раза).

5. Консонантные части слов

В индоевропейских языках доля слов с предвокальными консонантными частями больше, чем доля с поствокальными консонантными частями; в казахском языке дело обстоит наоборот - слов с поствокальными консонантными частями больше, чем с предвокальными (см. таблицу 4).

Таблица 4

Слова с консонантными частями (%)

№№ п/п	Консонантные части	Языки					
		русский	белорусский	немецкий	нижненемецкий	английский	казахский
I.	Предвокальные	97,7	97,9	94,3	95,6	94,3	86,1
2.	Поствокальные	95,0	94,8	93,2	94,0	85,7	94,3

6. Количество согласных фонем в консонантных частях слов

Во всех языках (см. таблицу 5) предвокальная консонантная часть состоит в большинстве случаев из одной согласной фонемы (доля этих случаев слева направо возрастает). В индоевропейских языках эта часть довольно часто состоит из двух согласных (их доля слева направо падает) и изредка из трех. В казахском языке изредка бывают две согласные, трех не бывает.

Поствокальная консонантная часть во всех языках также состоит в большинстве случаев из одной фонемы (особенно велика доля однофонемных частей в казахском языке). Во всех языках в значительно меньшем количестве случаев в этой части слов встречаются две фонемы, редко три фонемы, а в славянских и немецком языках совсем редко четыре фонемы.

Таблица 5

Длина консонантных частей (%)

№№ п/п	Количество согласных в консонантных частях	Языки					
		русский	белорусский	немецкий	нижненемецкий	английский	казахский
I.	В предвокальных						
	одна согласная фонема	55,8	54,6	63,4	63,7	69,6	85,8
	две согласные фонемы	39,2	39,9	28,7	30,8	22,7	0,3
	три согласные фонемы	2,7	3,4	2,2	2,I	2,0	-
2.	В поствокальных						
	одна согласная фонема	70,3	72,4	60,8	64,9	65,I	82,7
	две согласные фонемы	22,8	20,8	29,4	26,9	I9,8	II,3
	три согласные фонемы	I,8	I,5	2,5	2,2	0,8	0,3
	четыре согласные фонемы	0,I	0,I	0,5	-	-	-

7. Средняя консонантность консонантных частей слов*

В индоевропейских языках (см. таблицу 6) средняя консо-

* Средняя консонантность каких-либо единиц получается в итоге деления всех согласных фонем во всех единицах (в дан-

нантность предвокальных консонантных частей больше поствокальных консонантных частей, в казахском языке и здесь дело обстоит наоборот - консонантность поствокальных частей больше.

Таблица 6

Средняя консонантность консонантных частей

№ п/п	Консонантные части	Языки					
		русский	белорусский	немецкий	нижнемецкий	английский	казахский
I.	Предвокальные	I,46	I,48	I,35	I,36	I,28	I,0
2.	Поствокальные	I,28	I,25	I,38	I,33	I,25	I,I3

Отобранный для исследования тюркский язык, казахский, отличается от индоевропейских языков самым значительным преобладанием трехфонемных односложных слов, самым маленьким средним количеством фонем, самым большим удельным весом модели дистрибуции фонем СГС, преимущественным помещением гласной фонемы в словах с четным количеством фонем в первой половине слова, бóльшим числом случаев нахождения гласной фонемы в начальной позиции, чем в конечной, большей средней консонантностью поствокальной консонантной части, наличием в предвокальной части не больше двух и в поствокальной части не больше трех согласных.

Для индоевропейских языков по сравнению с казахским, характерны бóльшая доля четырехфонемных слов (слева направо она падает и в английском языке уступает доле трехфонемных слов), большее среднее количество фонем, помещение гласной фонемы в словах с нечетным количеством фонем в середине слов (симметричное расположение согласных фонем по обе стороны гласной) и в словах с четным количеством фонем во второй половине слов, меньшее число случаев нахождения гласной фонемы в начальной позиции, чем в конечной, бóльшая средняя консонантность предвокальной консонантной части, наличие в предвокальной консонантной части трех согласных и в поствокальной четырех.

ном случае, во всех консонантных частях) на количество единиц.

Л И Т Е Р А Т У Р А

Никонов В.А. Длина слова. - Вопросы языкознания, 1978, № 6, с. 104-111.

Плоткин В.Я. О взаимоотношениях между фонетической и семантической структурой слова. - В кн.: Актуальные проблемы лексикологии. Выпуск II, часть I. Новосибирск, 1969, с. 79-81.

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

Arens, H. Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute. Düsseldorf, 1965.

STATISTISCHE UNTERSUCHUNG DER PHONOLOGISCHEN STRUKTUR DES WORTES (ANHAND DER EINSILBIGEN WÖRTER EINIGER INDOEUROPÄISCHER SPRACHEN UND DES KASACHISCHEN)

Heinrich Pankratz

R e s ü m e e

In der vorliegenden Arbeit wird die phonologische Struktur des einsilbigen Wortes (der Umfang des Wortes, die Modelle der Distribution der Vokale und Konsonanten im Wort, die Stelle des Vokals im Wort, die konsonantischen Teile des Wortes und ihre mittlere Konsonantität) in einigen indoeuropäischen Sprachen (im Russischen, Belorussischen, Deutschen, Plattdeutschen, Englischen) und im Kasachischen statistisch untersucht.

РАСПРЕДЕЛЕНИЕ ГЛАГОЛОВ В НАУЧНО-РЕФЕРАТИВНОМ ТЕКСТЕ

В.И. Перебейнос

Требование научно-технической революции к повышению эффективности обмена научной информацией привело к росту веса научно-информационных изданий, особенно реферативных журналов, призванных собирать публикации в каждой отрасли науки и в компактной форме информировать специалистов о их содержании. Не случаен поэтому интерес к изучению лингвистических характеристик научно-реферативного текста.

В данной работе анализируются рефераты из РЖ "Кибернетика" лишь одной тематической группы ("Программирование и теория математических машин"), что, как представляется, обеспечивает однородность исследуемых текстов.

Изучение закономерностей структурной организации реферата может осуществляться на основе двух подходов: а) анализ закономерностей построения реферата как разновидности научного текста, б) выявление особенностей строения реферата в сопоставлении с реферируемым текстом. В нашем исследовании принят первый подход. Мы считаем, что изучение закономерностей строения реферата важно и в теоретическом, и в практическом плане: анализ структуры реферата как разновидности коротких текстов даст возможность разработать методику и уточнить проблематику лингвистики текста, а также построить типологию научных текстов. В практическом плане оно обеспечит базу для рекомендаций по составлению и редактированию, в том числе и автоматизированию рефератов.

Мы исходили из того, что структура текста раскрывается не только в сверхфразовых единствах или более длинных единицах текста, но и в характере отбора и расположения в тексте грамматических классов слов.

Одним из возможных подходов к исследованию распределения в тексте грамматического класса слов является формальный анализ расстояний между двумя соседними вхождениями данного класса слов, без учета длины текста или каких-либо синтаксических или текстовых единиц.

Такой анализ распределения единицы в тексте целесообразно осуществлять для единиц высокочастотных. Так, цепочки классов слов между двумя вхождениями в текст существительного в ряде случаев показывают характер синтаксического окру-

жения существительного и соотносимы со словосочетаниями, как простыми, так и сложными. Глагол также можно отнести к высокочастотным грамматическим классам слов: он занимает 4-е место в ранговом списке грамматических классов слов в реферативных текстах*.

Анализ распределения расстояний между двумя соседними глаголами показал, что оно характеризуется тремя незначительными пиками, почти одинаковыми по высоте, на расстояниях в нуль, четыре и восемь слов (см. табл. I).

Таблица I

Распределение длины расстояний между двумя соседними глаголами в тексте реферата

Длина расстояния	Ее доля (в %)	Длина расстояния	Ее доля (в %)	Длина расстояния	Ее доля (в %)
0	9,4	I4	4,2	28	0,2
I	2,8	I5	2,0	29	0,8
2	2,6	I6	2,2	30	0,2
3	3,8	I7	I,4	3I	0,4
4	8,0	I8	2,4	34	0,4
5	5,6	I9	I,6	35	0,6
6	7,6	20	0,8	36	0,2
7	5,2	2I	0,8	4I	0,2
8	8,4	22	I,0	48	0,2
9	6,2	23	0,2	49	0,2
I0	5,2	24	0,2	5I	0,2
II	3,8	25	I,0	55	0,2
I2	4,4	26	0,8		
I3	4,0	27	0,6		

Контактное расположение глаголов указывает либо на модальное сказуемое (может быть, может показать), либо на сочетания глагола-сказуемого с зависимым инфинитивом в роли дополнения (позволяет вычислить).

* Сопоставить частоты глагола в рефератах и иных научных текстах затруднительно, так как обычно в класс глаголов включают причастия и деепричастия. Данные, приводимые С.И. Кауфманом /Кауфман, I970, 283/, позволяют вычислить частоту личных форм глагола вместе с инфинитивом, т.е. тех форм, которые включаются нами в грамматический класс глагола. Она составляет всего 7,9 %, т.е. существенного расхождения между нею и частотой глагола в реферативных текстах (7,I %) не наблюдается.

Пики на расстояниях в четыре и восемь слов, видимо, находят объяснение в построении предложения.

Как представляется, данные таблицы I свидетельствуют о целесообразности рассматривать не расстояния между соседними глаголами, а распределение глаголов по предложениям в реферате. Возможно, расстояния между соседними глаголами зависят от позиции глагола в предложении и определяются распределением длины предложения.

Таблица 2

Распределение глаголов по позициям в предложении

№ позиции	Абсолютн. частота	Относит. частота (в %)	№ позиции	Абсолютн. частота	Относит. частота (в %)
I	946	23,65	I8	I4	0,35
2	537	I3,42	I9	I8	0,45
3	70I	I7,52	20	9	0,23
4	448	II,20	2I	II	0,27
5	303	7,57	22	9	0,23
6	227	5,67	23	6	0,I5
7	I65	4,I2	24	6	0,I5
8	I20	3,00	25	2	0,05
9	92	2,30	26	5	0,I3
I0	87	2,I7	27	3	0,08
II	77	I,93	28	2	0,05
I2	43	I,07	29	2	0,05
I3	50	I,25	35	I	0,03
I4	36	0,90	36	I	0,03
I5	24	0,60	38	I	0,03
I6	28	0,70	45	I	0,03
I7	25	0,62			

Изучение распределения 4000 глаголов по позициям в предложении показывает, что более 54 % глаголов занимают I-3 позицию в предложении, хотя максимальная удаленность глагола от начала предложения составляет 45 позиций. 73,9 % глаголов расположены не далее пятой, а 9I,3 % - не далее десятой позиции в предложении. Почти четверть всех глаголов находится на первом месте в предложении.

Однако данные таблицы 2 не раскрывают зависимости расстояния между соседними глаголами в тексте от позиции глагола в предложении. Вероятно, имеет значение не только позиция

глагола в предложении, но и линейная последовательность предложений, что составляет отдельный предмет исследования. Пока же следует отметить действие широко распространенного в языке и в речи закона предпочтения: наибольшее количество единиц (в данном случае - номеров позиций глагола в предложении) встречается с высокой частотой и составляет большинство в исследуемом массиве, а большое количество единиц имеет низкую частоту и составляет незначительный процент массива /**Dewey, 1923; Meier,** 1964; Перебийніс, 1970, 157/.

Таблица 3

Распределение глаголов в рефератах различной **длины**

Кол-во глаголов	Количество предложений в реферативном тексте																Всего
	I	2	3	4	5	6	7	8	9	I0	II	I2	I3	I4	I5	20	
0	I0	I0	I0	I	I	-	-	-	-	I	-	-	-	-	-	-	33
I	42	26	9	4	-	I	I	-	-	-	-	-	-	-	-	-	83
2	6	50	I6	9	3	I	-	-	-	-	-	-	-	-	-	-	85
3	5	I6	22	I5	4	-	2	-	-	-	-	-	-	-	-	-	64
4	3	I3	26	I9	9	3	-	-	-	-	-	-	-	-	-	-	73
5	I	3	23	8	I6	4	2	-	-	I	-	-	-	-	-	-	58
6	-	I	8	2I	9	3	4	-	I	-	-	-	-	-	-	-	47
7	-	I	9	9	6	I6	7	3	I	-	-	-	-	-	-	-	53
8	-	-	6	3	8	5	8	2	-	-	I	-	-	-	-	-	34
9	-	I	2	4	4	6	8	5	2	-	-	I	-	-	-	-	33
I0	-	-	-	-	6	I	2	4	4	I	I	-	-	-	-	-	I9
II	-	-	-	I	6	4	I	I	-	-	-	I	I	-	-	-	I5
I2	-	-	-	I	3	2	5	3	I	3	-	I	-	-	-	-	I9
I3	-	-	-	I	3	I	2	3	3	I	2	-	-	-	-	-	I6
I4	-	-	-	-	I	2	-	I	2	I	-	-	I	-	-	-	8
I5	-	-	-	-	-	-	I	3	2	I	2	-	-	-	-	-	9
I6	-	-	-	-	-	-	I	I	2	2	I	-	-	-	-	-	7
I7	-	-	-	-	-	-	-	I	I	3	4	-	-	-	-	-	9
I8	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	2
I9	-	-	-	-	-	-	-	-	-	-	-	-	-	I	I	-	2
20	-	-	-	-	-	-	I	-	-	-	-	-	-	-	-	-	I
2I	-	-	-	-	-	-	-	-	-	I	I	-	-	-	-	-	2
23	-	-	-	-	-	-	-	-	-	-	-	-	-	I	-	-	I
24	-	-	-	-	-	-	-	-	I	I	-	-	-	-	-	-	2
25	-	-	-	-	-	-	-	-	-	-	-	I	-	-	-	I	2
Всего	67	I2I	I3I	96	79	49	44	27	20	I6	I4	4	2	2	I	I	550

Если глагол, как предикативный центр предложения, равномерно распределен по тексту реферата, то, во-первых, в каждом предложении должен быть глагол, во-вторых, количество глаголов в тексте реферата будет увеличиваться с увеличением длины реферата, исчисляемой количеством входящих в него предложений.

Таблица 3 показывает, что количество глаголов в реферате действительно коррелирует с длиной текста. Для рефератов, состоящих из одного или двух предложений, как будто довольно хорошо просматривается и равномерность распределения глаголов по тексту: в этих массивах больше всего рефератов, количество глаголов в которых равно количеству предложений в реферате. Однако в рефератах длиной в три и четыре предложения максимум сдвигается в сторону большего количества глаголов, хотя сумма рефератов, включающих меньше глаголов, чем предложений, в этих рефератах достаточно велика и составляет соответственно 26,7 % и 30,2 %.

В более длинных рефератах количество текстов с малым количеством глаголов уменьшается, каждая длина текста имеет свои особенности группировки текстов по признаку количества глаголов в них. Таблица 3 дает основания для предположения: среднее количество глаголов в предложении зависит от длины реферата и увеличивается с ее ростом.

Распределение всех исследуемых рефератов на массивы в зависимости от длины реферата и определение среднего количества глаголов в предложении отдельно для каждого массива показало более сложную зависимость между длиной реферата и средним количеством глаголов в предложении. Все рефераты образовали три группы по признаку средней частоты глагола в предложении: полоса колебания средней частоты глагола в предложении от 0,94 до I,I4 характерна для рефератов, состоящих из двух и из тринадцати предложений; от I,20 до I,32 наблюдается в рефератах, состоящих из I, 3, 4, 6, 7, I2, I5 и 20 предложений; от I,36 до I,42 свойственна рефератам из 5, 8, 9, I0, II и I4 предложений.

Как видим, в каждой группе есть и короткие, и длинные рефераты, а наиболее длинные рефераты оказались в группе средней частоты глагола, а не в самой высокой, как предполагалось. Причины такой группировки рефератов требуют специального исследования, выходящего за рамки настоящей работы.

Исследование распределений глагола в реферате, представленном в виде последовательности предложений, требует вычис-

лить среднее количество глаголов в предложении, занимающем определенную позицию в реферате. Для этого каждый реферат представим в виде схемы, отражающей последовательность абсолютных частот глаголов в предложении. Так, схема I-2-I показывает, что в первом предложении реферата, состоящего из трех предложений, есть один глагол, во втором - два, в третьем - один, т.е. в реферате всего четыре глагола. Такие схемы дают возможность определить среднюю частоту глагола в любом предложении реферата.

Таблица 4

Зависимость частоты глагола от позиции предложения в реферате

Количество предложений в реферате	Позиция предложения в реферате							
	I	2	3	4	5	6	7	8
2	I,02	I,08						
3	I,20	I,30	I,34					
4	I,03	I,43	I,2I	I,20				
5	I,25	I,38	I,54	I,46	I,30			
6	I,04	I,29	I,45	I,33	I,20	I,4I		
7	I,00	I,29	I,45	I,39	I,32	I,29	I,00	
8	I,4I	I,33	I,35	I,44	I,26	I,07	I,22	I,48

Полученные средние частоты в большинстве случаев не показывают существенных расхождений, особенно в тех случаях, когда количество рефератов недостаточно для получения статистически достоверных данных. Но все же представляет интерес тенденция к самой высокой частоте в третьем и к самой низкой - в первом предложении реферата. Низкая частота глагола в первом предложении может быть объяснена тем, что это предложение является повторением заглавия, а предикат выражен чаще всего кратким причастием, составляющем в нашем исследовании самостоятельный грамматический класс слов. Для выяснения причин других особенностей распределения средней частоты глагола в предложениях подробнее рассмотрим схемы, отражающие последовательность количества глаголов в линейной цепи предложений реферата.

Как показывает таблица 3, реферат любой длины имеет разное количество глаголов, поэтому все схемы линейного расположения глаголов следует сгруппировать в массивы с одинаковым количеством глагола в рефератах одной и той же длины и рас-

сматривать как каждый массив отдельно, так и все рефераты данной длины вместе.

Распределение глаголов в рефератах из двух предложений описывается 17-тью схемами. Максимальное количество глаголов в одном предложении - пять - встречается только во втором предложении. Все схемы по характеру распределения глаголов образуют три группы: а) одинаковое количество глаголов в обоих предложениях (равномерное распределение) - 51,3 %; 3)увеличение количества глаголов во втором предложении - 25,6 % рефератов; в) уменьшение количества глаголов во втором предложении - 23,1 %. Так как схемы второй группы незначительно превышают по количеству схем третьей группы, среднее количество глаголов во втором предложении реферата несколько выше, хоть и не выходит за пределы статистически допустимых колебаний. Больше всего (30 %) рефератов описываются схемой I-I, т.е. по одному глаголу в каждом предложении.

В рефератах из трех предложений отмечено 52 схемы распределения глаголов, их можно сгруппировать следующим образом:

а) монотонные, в которых глаголы распределены либо равномерно (I-I-I или 2-2-2), либо с повышением их количества в каждом последующем предложении (0-I-2), либо с понижением (3-2-I). Такие схемы составляют 29,8 %;

б) симметричные, в которых крайние предложения имеют либо больше, либо меньше глаголов по сравнению со средним (38,2%);

в) асимметричные, в которых количество глаголов в предложениях не упорядочено, их 32 %.

Самая частая схема (I-I-I) составляет I2,2 %.

Таким образом, более двух третей рефератов из трех предложений представляют собой некоторое гармоничное целое относительно распределения в них глаголов.

Рассмотрим с этих же позиций рефераты из четырех предложений. Естественно, схемы расположения глаголов в них сложнее и их больше (60 схем). Самая частая - монотонная схема I-I-I-I (I2,5 %). Возможности симметричного построения схемы в рефератах этой длины увеличились. Если в рефератах из трех предложений мы имели дело лишь с одним видом симметрии - зеркальная симметрия с осью, проходящей через средний элемент схемы, - то в рефератах из четырех предложений есть симметричные схемы, в которых ось симметрии проходит между двумя средними элементами (например, 2-I-I-2), есть схемы ритмичные (типа 2-I-2-I), представляющие простой ритм или ритм с

расширением, при котором один элемент такта остается постоянным, а второй увеличивается (2-I-3-I или 0-I-0-2); ритм с сужением, при котором один элемент уменьшается, а второй остается постоянным (0-2-0-I или I-2-3-2); ритм с двойным расширением, когда оба элемента увеличиваются на одинаковое число (0-I-I-2), или с двойным сужением, как в 2-I-I-0 (уменьшение на I каждого элемента второго такта).

Анализ показывает, что монотонные схемы составляют I5,6%, симметричные с зеркальной симметрией - II,5 %, ритмичные - 52,I %, т.е. и в этом массиве рефератов неупорядоченное расположение глаголов встретилось лишь в 20,8 % текстов.Как видим, для рефератов из 4-х предложений наиболее характерно ритмичное расположение глаголов в тексте. Ритмичная структура, в отличие от зеркальной, характеризуется открытостью ряда: ритм можно повторять многократно, предел такого повторения и ограничение длины ритмической последовательности в ритме не обозначены.

Основой ритма в реферате из четырех предложений служит отрезок из двух предложений. Рассмотрим характер распределения глаголов в первом и втором такте (см. табл. 5). Возможно, именно в нем кроется признак завершенности текста, содержится некоторое ограничение, указывающее на нее.

Таблица 5

Характер распределения глаголов в первом и втором такте реферата из четырех предложений

№ такта / Характер расположения	I	II
Равномерное	40,6 %	39,6 %
С повышением	38,5 %	27,I %
С понижением	20,8 %	33,3 %

Проверка на существенность расхождения процентных показателей между первым и вторым тактом свидетельствует о том, что в первом такте существенно более высокую частоту имеет расположение глаголов с повышением их количества к концу такта, а во втором - с понижением (t равно 2,74 и I,99 соответственно, при табличном I,98).

Таким образом, текст реферата, состоящего из четырех предложений, стремится к обозначению завершенности текста,к строению зеркально симметричному, т.е. закрытому. Но образ-

цом закрытости текста следует считать рефераты из трех предложений с симметричным расположением глаголов.

В рефератах из пяти предложений насчитывается 67 схем распределения глаголов, из которых монотонных 7,8 %, симметричных - 9,I %, ритмичных - 29,9 % и асимметричных - 53,2 %. Как видим, удельный вес симметрии и ритма в этих рефератах значительно ниже. Меняется и характер ритмических тактов: появилась новая разновидность ритма - ритм с удлинением (I-2-I-I-2) или с укорочением, как в I-I-2-I-2 или I-2-2-I-2.

Рассмотрение первых и последних пар предложений показывает ту же тенденцию, что и в рефератах из 4-х предложений: количество глаголов в начальной паре предложений повышается к концу пары, а в последней - понижается. Иными словами, строение текста показывает стремление к симметричности.

Можно предположить, что тексты рефератов этой длины строятся, как из блоков, из последовательностей 2+3 или 3+2 предложения. Если это так, то вес асимметричных структур в блоках, состоящих из трех предложений, должен быть примерно равным весу этих схем в текстах из трех предложений. Подсчеты показывают, что количество асимметричных структур в этих блоках составляет 40,2 % в начальных и 44,2 % - в конечных. Возможно, это свидетельствует о предпочтительности деления на блоки 3+2, а не 2+3, но окончательно решить, какое членение на блоки предпочтительнее (если оно вообще целесообразно), может только анализ сверхфразовых единств в сопоставлении со схемами расположения глагола.

Но несомненно то, что рефераты длины в два и три предложения - это образцы коротких законченных текстов, первый из которых служит основанием ритма в более длинных текстах, а второй, помимо этого, является еще и примером симметричности и структурного оформления законченности текста. Очевидно, не случайно, что среди рефератов больше всего рефератов в два и три предложения.

С увеличением длины текста вес симметричных расположений глагола снижается, но возрастает возможность членить тексты на блоки и анализировать их взаимодействие.

Так, рефераты длиной в шесть предложений при разделении их на два блока по три предложения в каждом показывают, что в начальном блоке вес асимметричных структур составляет 30,6%, а в конечном - 44,9 %, т.е. начало текста организовано более четко, чем его конец. О том же свидетельствуют рефераты длиной в семь предложений: при выделении в них начального и ко-

нечного блока по три предложения в каждом первый из них содержит 34,I % асимметричных схем расположения глаголов, а последний - 45,4 %.

Таким образом, анализ схем расположения глаголов в линейной последовательности текста помогает раскрыть особенности его строения как целого и указывает на такие закономерности его структурной организации, которые нельзя обнаружить иным путем.

ЛИТЕРАТУРА

Кауфман С.И. Из курса лекций по статистической стилистике. М., Московский обл. пед. институт им. Н.К.Крупской, 1970.

Перебийніс В.С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. - К.:Наукова думка, 1970.

Dewey G. Relative Frequency of English Speech Sounds. London, 1923.

Meier H. Deutsche Sprachstatistik. Hildelsheim, 1964.

DISTRIBUTION OF VERBS IN SCIENTIFIC ABSTRACTS

Valentina Perebeynoss

S u m m a r y

The regularities of text structural organization can be traced in the distribution of grammatical word classes in the text considered as a whole. The distribution of verbs (finite forms and the infinitive) in scientific abstracts in the field of cybernetics is represented by means of patterns showing the number of verbs in each sentence of an abstract. Three groups of these patterns are established: monotonous (1-1-1, 0-1-2-3. 2-1-0 etc), symmetric, including mirror (0-1-0, 2-1-2 etc) and rhythmic (1-2-1-2 etc) patterns, and asymmetric (1-1-2, 2-1-0-1).

Mirror symmetry is characteristic mostly of three-sentence abstracts, while four-sentence ones are mainly rhythmically built. The quota of symmetrically structured abstracts drops with the growth of abstract length, but abstracts of any length show a tendency to mark the beginning and the end of the text by lessened mean frequency of verbs in the first and the last sentences.

"ПРОЛЕГОМЕНЫ" К СТАТИСТИЧЕСКОЙ ТЕОРИИ ТЕКСТА

Б.Я.Слепак

I. Общие замечания*

Лингвостатистика начиналась с изучения отдельных текстов (см. Марков, 1916; Yule, 1938; Williams, 1940). Середина нашего столетия была отмечена резким повышением интереса к количественным исследованиям речевого материала, что было обусловлено, в первую очередь, усилившейся в этот период тенденцией к расширению взаимосвязи общественных, естественных и технических наук. Широкому введению статистической методологии в языкознание способствовало интенсивное развитие кибернетики, и не только потому, что электронная техника заметно уменьшала вычислительные затраты - результаты лингвостатистических изысканий были настоятельно необходимы для удовлетворения разнообразных практических задач, связанных с использованием ЭВМ для автоматической переработки текстового материала, в частности для машинного перевода, информационного поиска, автоматического аннотирования, реферирования, индексирования и др.

Все это и предопределило в значительной мере смену магистральной лингвостатистической ориентации - количественное изучение отдельных текстов явно и безоговорочно уступило место квантитативному описанию совокупного продукта коммуникативно-речевой деятельности - речевого материала (речи). Текст сам по себе теперь начал рассматриваться как своеобразный "черный ящик", его внутреннее статистическое устройство, как правило, не служило объектом самостоятельного исследования.

Сейчас, с нашей точки зрения, настало время - в соответствии с диалектикой развития статистической лингвистики - возвратиться от широких количественных обобщений к тому, с чего начиналась эпоха квантификации языкознания, что объективно должно было предшествовать "статистике речи" - к статистике текста (ср. лингвистика текста, грамматика текста).

Если в области статистики речи и статистики стиля уже накоплены данные, позволяющие делать обобщения, формулировать

* Несколько нарочито звучащим словом "пролегомены" автору статьи хотелось лишь подчеркнуть предварительный, во многом зондирующий характер изложенных в ней мыслей и результатов.

закономерности (хотя в общем еще далеко неокончательного характера), то в сфере статистики текста обобщать и теоретизировать практически нечего.*

Основная цель предлагаемой статьи – привлечь внимание языковедов к этому "забытому" региону стилостатистики, который в результате парадокса развития квантитативной лингвистики остается практически "белым пятном" на лингвостатистической карте.

2. Некоторые теоретические предпосылки

В качестве исходного, рабочего принимается следующее определение текста: текст – это однозначно локализуемая в пространстве совокупность предложений**, образующая сообщение, которое имеет самостоятельную коммуникативно-смысловую ценность.

С целью уточнения терминологического контекста, в котором используется понятие текста, отметим, что под речью (устной, письменной) понимается деятельность, связанная с реализацией способности человека производить и регистрировать семиотически значимые звуки, сочетания звуков (слова) и сочетания слов (предложения). Укажем также на некоторые соответствия между используемыми в настоящей статье лингвистическими понятиями и понятиями метаязыка математической статистики: коммуникативно-речевая деятельность $\Longleftrightarrow$ эксперимент; единицы (системы) языка $\Longleftrightarrow$ единицы подсчета; текст $\Longleftrightarrow$ единица наблюдения; речевой материал и его разновидности $\Longleftrightarrow$ генеральная и выборочные совокупности; стиль $\Longleftrightarrow$ способ организации генеральной совокупности (функция распределения, закон распределения).

---

* Ярким подтверждением тому может служить одна из последних работ К.Б.Бектаева (Бекьаев, 1978), в которой – вопреки тому, что можно ожидать по названию – статистика текста явно подменена статистикой текстов ("речи").

** Точнее, по всей видимости, говорить о том, что текст может состоять из $1 \ldots n + 1$ предложений. Примером "вырожденного" текста, состоящего из одного предложения, служит "поэма": "О, закрой свои бледные ноги!", имевшая распространение в начале XX века.

Представляется удачным термин "этноязык", предложенный А.И.Горшковым (см. Березин, Головин, 1979, 59). Данный термин позволяет избежать интерференции двух значений слова "язык": язык - важнейшее средство человеческого общения, комплексное явление, характеризуемое тремя основными аспектами - структурным, материальным и функциональным (в таком значении и употребляется термин "этноязык"); язык - семиологическая структура, совокупность знаков.

Из текстов извлекаются частоты употребления языковых единиц и другие выборочные характеристики, которые объективизируют понятие узуса; другой основной аспект коммуникативно-речевой деятельности - коммуникативно-стилевая норма - конституируется вероятностями, распределениями вероятностей и корреляционными отношениями.

3. Квантификация понятия "текст"

3.1. Статистическая идентификация текста

Проблема формулируется следующим образом: релевантно ли в статистическом плане само понятие текста как самостоятельной единицы наблюдения? Для решения этого вопроса в плане внешних сопоставлений используем однофакторный дисперсионный анализ, с помощью которого можно одновременно сопоставить (на достоверностной основе) любое количество средних частот. Для корректного применения указанного метода необходимо, чтобы генеральная совокупность была распределена близко к нормальному и чтобы групповые дисперсии существенно не различались.

Применим дисперсионный анализ для сопоставления частот употребления сложноподчиненных предложений в авторской речи романов С.Льюиса: "Главная улица" (ГлУ), "Бэббит" (Бэб), "Эроусмит" (Эр), "Кэсс Тимберлейн" (КТ), "Кингсблад, потомок королей" (КПК). Процедура расчета критерия Бартлета B, служащего для сравнения групповых дисперсий и различных показателей, фигурирующих в однофакторном дисперсионном анализе, подробно описана во многих работах по математической статистике (см., например, Пустыльник, 1968; Урбах, 1964). В описываемом случае $B = 2{,}464 < \chi^2_{0,95} = 9{,}5$ (при четырех степенях свободы), что позволяет гипотезу о равенстве дисперсий в сопоставляемых текстах Льюиса признать справедливой. Машинный эксперимент по аппроксимации эмпирических распределений теоретическим путем использования критерия хи-квадрат показал, что распределение сложноподчиненных предложений в авторской

речи романов С.Льюиса подчиняется нормальному закону при уровне существенности $P_{0,05}$ ($A_S = +0{,}203$; $E_x = -0{,}089$).

Исходные и кодированные данные для однофакторного дисперсионного анализа приведены в таблицах 3.I и 3.2 (величина микровыборки здесь и в дальнейшем изложении – I00 самостоятельных предложений). В соответствии с этими данными

$$S_a = 4I46{,}8;\quad S_x = \frac{368^2}{36} = 376I{,}8;\quad S_{a_i} = 6^2 + 2^2 + \ldots + 20^2 + I3^2 = 4482;\quad S_A = 4I46{,}8 - 376I{,}8 = 385{,}0;$$

$$S_Z = 4482 - 4I46{,}8 = 335{,}2;\quad f = 5 - I = 4;\quad f_Z = 36 - 5 = 3I;$$

$$S_A^2 = \frac{385{,}0}{4} = 96{,}3;\quad S_Z^2 = \frac{335{,}2}{3I} = I0{,}8;\quad F_{A/Z} = \frac{96{,}3}{I0{,}8} = 8{,}9.$$

Таблица 3.I.

Исходные данные для однофакторного дисперсионного анализа

Тексты	Номера микровыборок									
	I	2	3	4	5	6	7	8	9	I0
ГлУ	24	26	22	29	26	30	3I	26	23	36
Бвб	28	25	26	26	30	28				
Эр	25	28	29	29	30	32	34	30	33	
КТ	34	36	33	38	27	33				
КПК	36	35	37	40	33					

Таблица 3.2.

Однофакторный дисперсионный анализ

I	2	3	4	5	6	7	8	9	I0	n_a	x_a	x_a^2	x_a^2/n_a
I	6	2	9	6	I0	II	6	3	I6	I0	73	5329	5329
8	5	6	6	I0	8					6	43	I849	308,2
5	8	9	9	I0	I2	I4	I0	I3		9	90	8I00	900
I4	I6	I3	I8	7	I3					6	8I	656I	I093,5
I6	I5	I7	20	I3						5	8I	656I	I3I2,2
					Сумма					36	368		4I46,8

Поскольку число 8,9 больше, чем $F_{0,05} = 2{,}9I$ для $f_1 = 4$, $f_2 = 3I$ (F определяется по таблице критических значений критерия Фишера), есть все основания считать, что фактор текста

(в аспекте времени написания) существенно влияет на частотность употребления сложноподчиненных предложений в авторской речи романов С.Льюиса.

Резкую противопоставленность текстов одного автора, разных авторов, одного жанра, разных жанров вскрывают и попарные сопоставления. Так, между романами Т.Драйзера "Сестра Керри" и "Финансист" (авторская речь) обнаружено 77 % существенных различий по I3 синтаксическим явлениям (типы и виды предложений, партиципные конструкции, инверсия, начала предложений и т.п.). Показатель размежеванности для "Американской трагедии" и "Кэсса Тимберлейна", вычисляемой как отношение числа значимых различий к общему количеству сопоставлений, равен 0,85; для эссе Т.Драйзера "Трагическая Америка" (ТрА) и "Америку стоит спасать" (АСС) - 0,62; для "Трагической Америки" и "Титана" (Тит) - 0,70 (максимально возможное значение коэффициента размежеванности равно единице).

Попарные сопоставления текстов производились с помощью параметрических и непараметрических (см. подробнее: Урбах, I964; Сепетлиев, I968) критериев различия. Последние более предпочтительны, поскольку их применение не предусматривает необходимости решения вопроса о характере распределения вариант в генеральной совокупности (проблемы чрезвычайно сложной и малоисследованной в лингвостатистике) и в то же время обеспечивает достоверностную оценку наблюдаемых различий.

Проиллюстрируем применение непараметрических методов анализа на примере серийного критерия. При сопоставлении вариационных рядов, репрезентирующих функционирование сложносочиненных предложений в авторской речи романов "Кэсс Тимберлейн" (число вариант $n_x = 32$) и "Американской трагедии" (АТр) ($n_y = 66$), было обнаружено 2 серии ($S = 2$), каждая из которых представляет собой непрерывную последовательность вариант, принадлежащих только к одному из двух рядов.

Расчет серийного критерия U_S не требует сложных вычислений. В нашем случае (число вариант обоих рядов уменьшено на 5 - количество пар с нулевой разницей между вариантами):

$$a = 2 \cdot 27 \cdot 6I = 3294; \quad b = 27 + 6I = 88;$$

$$\hat{S} = \frac{3294}{88} + I = 38,43; \quad S_S^2 = \frac{3294\ (3294 - 88)}{88^2 \cdot (88 - I)} = I5,67;$$

$$U_S = \frac{38,43 - 2 - 0,5}{I5,67} = 9,39.$$

Поскольку $U_S = 9{,}09 > 2{,}58$ (теоретическое критическое значение), нулевую гипотезу следует отвергнуть - между текстами "Кэсс Тимберлейн" и "Американская трагедия" (авторская речь) по частотам сложносочиненных предложений наблюдаются существенные различия.

Применение критерия Вилкоксона позволяет уточнить это заключение: рассмотренные вариационные ряды значимо различаются по центральной тенденции, т.е. по средним частотам ($\bar{X}$ соответственно I2,25 и 4,56).

Многообещающие возможности для статистической идентификации текста открывает использование аппарата многомерной статистики, специализированного для одновременного сопоставления сложных объектов по комплексу признаков. Стилостатистические работы многомерного характера немногочисленны (см., например, Kraus, Vašák, I967; Мальцева, I969; Слепак,I975). Чрезвычайно перспективным в этой связи представляется метод дискриминантного анализа, который дает возможность обобщенно и, что очень важно, достоверностно оценивать характер наблюдаемых различий в сопоставляемых текстах по величине нескольких признаков.

Картина, получаемая при анализе результатов попарных межтекстовых сопоставлений, как правило, мозаична и не всегда позволяет выявить единые и устойчивые качественно-количественные тенденции. Правомерно ожидать, что разнообразные стилостатистические задачи, такие, например, как количественная типология текстов, стилевая дифференциация, периодизация творческого пути писателя, будут объективно решены на базе именно многомерного, комплексного подхода.

3.2. Статическая статистическая структура текста

Под статической статистической структурой текста понимается набор количественных характеристик, описывающих вариационные ряды, которые репрезентируют функционирование языковых единиц в тексте, и предназначенных главным образом для оценки и сравнения средних величин и показателей рассеивания, показателей взаимосвязи и взаимообусловленности разных языковых средств, а также для аппроксимации эмпирических распределений теоретическими.

Сопоставление попарно сопряженных вариант в замкнутом тексте может быть произведено с помощью непараметрического критерия знаков. В принципе, видимо, все языковые единицы в связном тексте в той или иной мере сопряжены, взаимосвязаны,

что обусловлено системным характером речевой организации текста (одно из объективных подтверждений правомерности системной трактовки текста см. в 3.3).

Необходимо, например, определить, достоверна ли разница между частотами сложносочиненных и сочинено-подчиненных предложений в авторской речи романа "Кингсблад" (см. табл. 3.3.).

Таблица 3.3.

Внутритекстовые сопоставления посредством критерия знаков

Сложносочиненные	I6	I4	II	I7	8	8	I0	5	7	I3
Сочинено-подчиненные	I7	I9	I3	I4	I4	I2	I7	I9	2I	I3
	+	+	+	−	+	+	+	+	+	0
Продолжение										
Сложносочиненные	II	8	I3	7	I2	7	II	9	9	I2
Сочинено-подчиненные	I4	I6	20	I3	II	I0	I8	I4	I0	I3
	+	+	+	+	−	+	+	+	+	+
Продолжение										
Сложносочиненные	I4	I0	I9							
Сочинено-подчиненные	I5	I3	I8							
	+	+	−							

Сопоставление пар показывает, что в I9-и случаях в авторской речи романа "Кингсблад" обнаруживается преобладание сочинено-подчиненных предложений над сложносочиненными, в 3-х случаях чаще используются сложносочиненные предложения и, наконец, в одном случае разница между вариантами равна нулю. Сравнение полученного эмпирического числа реже встречающихся знаков $Z = 3$ с критическими табличными значениями $Z_{05} = 6$ и $Z_{01} = 5$ (при $n = 22$; количество вариант обоих сопоставляемых рядов уменьшается на число, соответствующее числу пар вариант, между которыми обнаружена нулевая разница; нулевая гипотеза принимается при $Z \geqslant Z_{05}$ и отвергается при $Z < Z_{01}$) показывает, что расхождения частот сложносочиненных и сочинено-подчиненных предложений в тексте "Кингсблад" не являются случайными. В авторской речи названного романа сложносочиненные предложения употребляются реже.

Как видно из приводимого ниже фрагмента таблицы (см.табл. 3.4), каждый замкнутый текст характеризуется своеобразным набором отношений сопряженных языковых единиц.

Интерес, на наш взгляд, представляет попытка обобщенной оценки синтаксической структуры текста с помощью энтропии - количественной меры информации, предложенной К.Шенноном.

Таблица 3.4.

Сопоставления сопряженных синтаксических явлений с помощью критерия знаков

Исследуемые тексты	Сопряженные синтаксические явления							
	Простые предложения	Сложные предложения	Сложно-сочин. предл.	Сочинен-но-подч. предл.	Конструкции с причаст. I	Конструкции с причастием II	NP-начала предложений	NNP-начала предложений
СК	=	=	+		+		+	
Фин		+		+	+		+	
Тит		+	=	=	+		–	–
Ст		+		+	+			+
АТр		+		+	+			+
ГлУ	+		=	=	+		+	
Бэб	=	=	=	=	+		+	
Эр		+	–	–	+		+	
КТ		+		+	+		+	
КПК		+		+	+		+	•
ТрА		+	=	=	–	–		+
АСС		+	=	=	=	=	=	=

Примечание: + означает наличие существенных расхождений с более высокими частотами у того явления, в колонке которого данный знак помещен;
= означает отсутствие существенных расхождений;
– означает неопределенность вывода о характере наблюдаемых расхождений. СК – "Сестра Керри", Фин – "Финансист", Ст – "Стоик",
NP – noun phrase, NNP – non-noun phrase.

В роли символов алфавита в описываемом эксперименте выступают отдельные виды предложений; буквами алфавита являются элементарные предложения, объединенные в рамках пунктуационного единства посредством сочинительных или подчинительных связей. Набор выделенных видов предложений (алфавит) состоит из I6 символов (табл. 3.5). Теоретически число видов предложений, выделяемых согласно используемым нами критериям, бесконечно, но, поскольку вероятности появления предложений с очень большим числом элементарных (порядка шести и более) крайне низки, этот набор можно представить в виде условно-конечной совокупности символов.

Для измерения энтропии текста (в ее первом приближении - при условии независимости вероятности появления отдельного символа в определенной точке сообщения от вероятности предшествовавшего символа или символов) применим формулу, введенную К.Шенноном:

$$H_1 = [P(A_1)\lg_2 P(A_1) - P(A_2)\lg_2 P(A_2) - \ldots - P(A_N)\lg_2 P(A_N)],$$

где $P(A)$ - вероятности встречаемости символов в сообщении.

Из формулы Шеннона следует, что энтропия тем выше, чем больше символов в алфавите и чем равномерней распределены вероятности их употребления в реальном сообщении. Отсюда ясно, что максимальной энтропия будет в том случае, когда все символы равновероятны:

$$H_0 = \lg N ,$$

где N - общее число символов алфавита.

Чтобы определеить избыточность сообщения (R), вначале высчитывают относительную энтропию:

$$H_{отн} = \frac{H_1}{H_0} .$$

Избыточность определяется по формуле:

$$R = 1 - H_{отн} .$$

В контексте описываемого исследования показатели энтропии (табл. 3.6) рассматриваются в качестве мерила а) разнообразия синтаксиса (само собой разумеется, что текст, в котором встречаются предложения только одного вида, например, простые, характеризуется нулевой энтропией на одно предложение и абсолютно однообразной - в плане использования разных моделей предложений - структурой); б) усложненности речи (чем выше энтропия, тем чаще употребляются в тексте разные виды предложений, в том числе предложения с высокой насыщенностью

Таблица 3.5.

Вероятности символов алфавита видов предложений

Тексты	Символы алфавита видов предложений															
	I	2	3	4	5	6	7	8	9	I0	II	I2	I3	I4	I5	I6
СК	5038	2620	0733	07I4	0397	0070	0I59	0I23	0022	00I0	0038	0037	00I0	000I	00I3	00I5
Фин	4348	2I86	0877	0874	0583	0I02	0268	0270	0064	0008	0II5	0II4	002I	000I	0063	0I06
Тит	4706	2328	0725	0880	04I0	0079	03I6	0I90	0062	00I2	0095	0077	00I8	0003	0042	0057
АТр	4036	2318	0379	I323	0344	0064	06I5	0I79	0036	00II	0252	0082	0023	0003	0242	0093
Ст	4I98	279I	0400	II48	0354	0046	0482	0I7I	003I	0006	0I40	0068	00I4	0003	0075	0073
АСС	378I	2492	0586	I247	0364	0I03	0525	02I4	0033	0003	0300	0I03	00II	0003	0I53	0082
ТрА	4595	2602	0402	I095	0262	005I	0457	0I08	002I	002I	0I69	0059	0005	0007	0087	0059
ГлУ	5677	I883	0655	0537	0378	0I25	0I6I	0I53	007I	0043	005I	0039	004I	00I0	0042	0098
Вэб	5II7	I932	0782	0450	0507	0229	0I93	0I93	0I2I	0050	007I	0086	006I	00I4	0058	0I36
Эр	4286	I798	0928	0723	058I	0240	0295	0284	0I49	0049	0I09	0I40	0I02	0028	0054	0234
КТ	3738	20I3	I003	0809	0659	0I8I	0344	044I	0II9	00I9	0I00	0I44	0069	0022	0087	0252
КШК	3748	2I04	086I	0900	066I	0I57	0439	026I	0I39	0044	0I6I	0I83	0083	00I7	0074	0I68

Примечания: I. Нуль везде опущен.

2. Виды предложений: I – простые; 2 – сложноподчиненные с I-м придаточным; 3 – сложносочиненные с 2-мя элементарными; 4 – сложноподчиненные с 2-мя придаточными; 5 – с 2-мя сочиненными и I-м придаточным; 6 – сложносочиненные с 2-мя элементарными; 7 – сложноподчиненные с 3-мя придаточными; 8 – с 2-мя сочиненными и 2-мя придаточными; 9 – с 3-мя сочиненными и I-м придаточным; I0 – сложносочиненные с 4-мя элементарными; II – сложноподчиненные с 4-мя придаточными; I2 – с 2-мя сочиненными и 3-мя придаточными; I3 – с 3-мя сочиненными и 2-мя придаточными; I4 – сложносочиненные с 5-ю и более элементарными; I5 – сложноподчиненные с 5-ю и более придаточными; I6 – сочиненно-подчиненные с 6-ю и более элементарными.

элементарными); в) стандартности текста (чем ниже энтропия, тем больше обусловленность выбора из всего комплекта видов предложений ограниченного числа моделей, тем выше зафиксированность правил построения текста и, значит, его стандартность).

Таблица 3.6.

Значения энтропии H_I и избыточности R

Тексты	H_I	R
СК	2,088I	0,4780
ГлУ	2,I533	0,46I7
ТрА	2,3034	0,424I
Тит	2,3655	0,4086
Ст	2,3908	0,4023
Бэб	2,4I57	0,396I
Фин	2,54I9	0,3645
АТр	2,595I	0,35I2
АСС	2,6429	0,3393
Эр	2,7685	0,3079
КПК	2,8555	0,286I
КТ	2,8575	0,2856

Примечания: I. $H_{макс.} = 4$.

2. Тексты расположены в порядке увеличения энтропии.

Самая высокая "структурность", если вслед за В.Ингве считать "показателем структурности" "любые отклонения от равновероятностного исхода" (цитируется по: Ревзин, I958), свойственна авторской художественной речи последних по времени написания романов обоих авторов и художественно-публицистической разновидности жанра эссе, представленной текстом "Америку стоит спасать" (см. табл. 3.6.). При этом достаточно четко прослеживается закономерность: чем позднее написан текст, тем выше энтропия на один вид предложения, тем разнообразней его статистический "портрет".

На основании приведенных выше иллюстративных данных можно сделать общий вывод: и внутритекстовые, и межтекстовые сопоставления неоспоримо свидетельствуют о том, что каждый замкнутый текст имеет своеобразную статическую статистическую структуру, благодаря которой он выделяется на фоне других текстов как релевантная единица количественного наблюдения.

3.3. Динамическая статистическая структура текста

Количественная объективизация одного из основных качественных дифференциальных признаков понятия "текст" – связности может быть осуществлена в двух направлениях. Прежде всего необходимо ответить на вопрос: как ведут себя на разных участках речевого континуума текста различные языковые средства? Для решения задач такого рода непараметрическая статистика предлагает метод последовательных серий (см. Сепетлиев, 77-79).

Необходимо, например, проверить, как распределяются в отдельных, последовательно сменяющих друг друга "кадрах" авторской речи романа "Финансист" сложноподчиненные предложения – равномерно или с тенденцией к чередованию "скоплений" высоких и низких частот. Технология применения метода последовательных серий проста. После ранжирования числовых значений в восходящей градации определяют медиану $\tilde{X}_{me}$, находят разности конкретных значений и медианы, принимая во внимание только знаки этих разностей, устанавливают число серий с одинаковыми знаками и сравнивают его с теоретическим значением. В нашем случае (при числе наблюдений $n = 60$) число серий, которые бы могли появиться случайно, равно $22 \leq R \leq 39$. Поскольку полученное эмпирическое число серий $R = 20$ меньше нижней границы табличного диапазона, правомерно заключить, что налицо систематические колебания ("скачки" и "падения") частот сложноподчиненных предложений в авторской речи романа "Финансист".

После того как установлено, что "непрерывность" текста может сопровождаться неравномерным, "тенденциозным" употреблением языковых средств, естественно возникает другой вопрос: каковы величина и направление динамических изменений, наблюдаемых при функционировании языковых единиц, вовлеченных в сюжетно-композиционное развертывание текста?

Одним из метаязыков, позволяющих описывать связный текст "как динамическую, закономерно организованную структуру" (В.В.Виноградов), является статистический аппарат динамических рядов (см. Сепетлиев, 176-208), который служит для выявления основной тенденции развития изучаемого явления, для отражения того типичного (не затемненного действием второстепенных факторов), что характеризует исследуемый процесс. Другими словами, посредством аппарата динамических рядов ва-

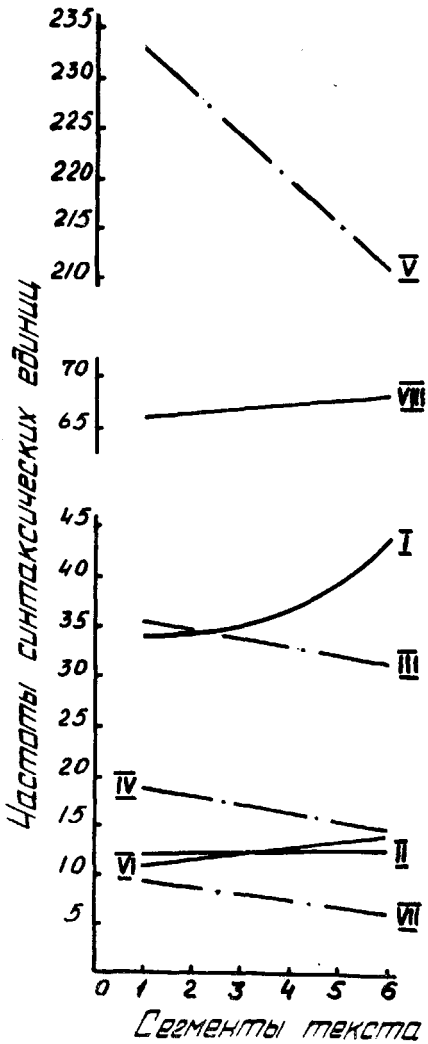


Рис. 3.I. Динамика функционирования синтаксических единиц в романе "Кэсс Тимберлейн"

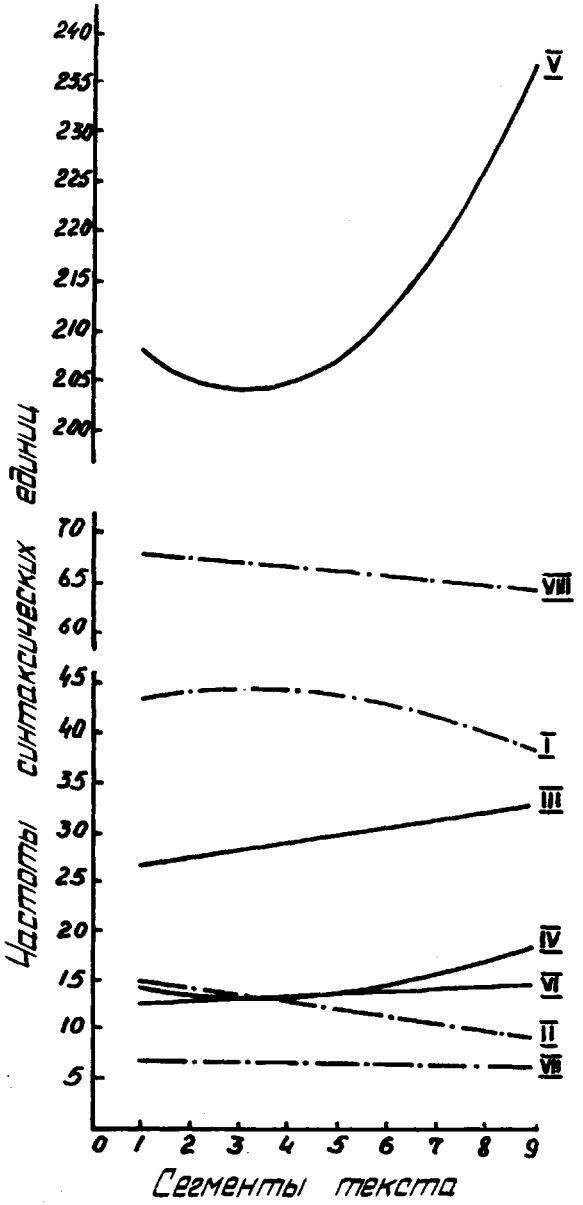


Рис. 3.2. Динамика функционирования синтаксических единиц в романе "Эроусмит"

риативность речи может изучаться не только в статической форме (в виде вариационных рядов), но и в динамике, как процесс (в виде динамических рядов).

В описываемом эксперименте в роли независимой переменной (аргумента X) выступает последовательное сегментирование связного текста (ось абсцисс); в качестве переменной зависимой (функции У) используется частотность синтаксических единиц в отдельных сегментах текста (ось ординат; см. рис. 3.I и 3.2).* Отметим, что на представленных рисунках не изображены эмпирические кривые, а даны уже выравненные теоретические линии (прямые и параболы второй степени). Выравнивание осуществлялось с помощью метода наименьших квадратов.

Для сохранения единства расчетов и облегчения графической репрезентации произведена "компрессия" – на оси абсцисс откладывались сегменты текста длиной в 500 самостоятельных предложений, на оси ординат – средние частоты синтаксических явлений на I00 самостоятельных предложений.

Из анализа приведенных графиков видно, что динамическая структура связного текста может быть представлена в виде специфичной только для него системы тенденций развития, отражающих характер функционирования отдельных языковых явлений (синтагматику текста), их взаимосвязь и взаимообусловленность (парадигматику текста) и различающихся по величине и направлению динамических изменений.

Объективное подтверждение, таким образом, находит мысль советского психолога Л.С.Выготского о том, что "новым для искусства фактором" являются не элементы художественного произведения, которые "существуют до него", а "способ построения" (подчеркнуто нами – Б.С.) этих элементов (см.Выготский, I962).

Упорядоченность (сбалансированность) текста по парадигматической и синтагматической осям, которая проявляется в своеобразной для каждого текста динамической статистической структуре, следует рассматривать как один из признаков, отличающих "текст" от "нетекста" – совокупности грамматически правильных высказываний.

---

* Примечания к рис. 3.I и 3.2. Синтаксические единицы: I – простые предложения; 2 – сложносочиненные предложения; 3 – сложноподчиненные; 4 – сочиненно-подчиненные; 5 – общее количество элементарных предложений на I00 самостоятельных предложений; 6 – конструкции с причастием I; 7 – конструкции с причастием II; 8 – N^{β}-начала предложений.

По величине динамических изменений синтаксические средства можно разделить на две основные группы: с ярко выраженными тенденциями развития и, соответственно, с малозаметными тенденциями к изменению. Другими словами, различные синтаксические единицы неодинаково варьируют на протяжении всего текста (от его начала к концу), одни - больше, другие - меньше. Из рассмотренных синтаксических явлений ярко выраженными тенденциями развития в большинстве текстов отличаются простые и сложноподчиненные предложения, *NP*-начала предложений, показатели средней насыщенности сложного предложения элементарными, общее количество элементарных предложений на 100 самостоятельных предложений.

Отметим, что одно и то же синтаксическое явление может функционировать с ярко выраженными тенденциями развития в одном тексте и с малозаметными - в другом. Так, из обследованных текстов наименее заметные динамические изменения прослеживаются при функционировании синтаксических единиц в текстах "Титан", "Главная улица" и "Трагическая Америка".

К числу явлений, частоты которых малозаметно изменяются на протяжении всех обследованных текстов, необходимо отнести сложносочиненные предложения и конструкции с причастием II (их можно назвать "стационарными"). Использование сочиненно-подчиненных предложений и конструкций с причастием I характеризуется несколько более высокой внутритекстовой динамикой.

С точки зрения направления динамических изменений функционирование синтаксических единиц в связных текстах характеризуется тенденциями или I) к более частому использованию от начала до конца текста (сопоставляются первая и последняя точки теоретической линии) или 2) к менее частому употреблению.

Исследование внутритекстовой динамики функционирования синтаксических явлений может послужить основой для количественной типологии текстов. Так, например, четко противопоставляются тексты с тенденцией к более частому использованию к концу текста простых и сложноподчиненных предложений текстам с противоположной по направлению тенденцией, тексты с тенденцией к увеличению к концу текста общего количества элементарных предложений на 100 самостоятельных и средней насыщенности сложного предложения элементарными текстами с противоположными по направлению тенденциями и т.д.

В связи с проблемой внутритекстовой вариативности нами был также рассмотрен и вопрос о взаимоотношении средней час-

тоты выборки из художественного текста со средней частотой для авторской речи всего текста, обследованного вполнъ и рассматриваемого в качестве генеральной совокупности. Применение критерия t, вычисляемого по формуле:

$$t = \frac{\bar{x}_{в.с} - \bar{x}_{г.с}}{\sqrt{\frac{\sigma^2_{г.с}}{n_{в.с}}}},$$

показывает, что выборочное обследование художественного текста, если его осуществлять по принципу "spread sampling" (см. Yule), дает очень надежные результаты. Само собой разумеется, что объем выборки должен зависеть, в первую очередь, от величины вариации рассматриваемого синтаксического явления в тексте.

3.4. Тексты и стиль

Стили этноязыка как устойчивые, функционально-целесообразные способы отбора и сочетания языковых средств формируются в коммуникативно-речевой деятельности и реализуются в текстах. Взаимоотношения текстов и стиля строятся на принципе сложного динамизма, который заключается в следующем: несмотря на наличие внутристилевой вариативности выделяются интегративно-динамические тенденции, объединяющие тексты одного стиля и противопоставляющие тексты разной стилевой отнесенности. Пример такой тенденции - высокие частоты сложносочиненных предложений в информативных журнальных научно-технических текстах ($\bar{X}$ = 20,I4; I2,34; I4,I7; I0,I7) и очень низкие частоты этого вида предложений в публицистических текстах ($\bar{X}$ = 3,40; 5,54; 4,80; 6,94).

4. Заключение

Основываясь на вышеизложенном и осознавая его во многом предварительный характер, попытаемся предложить определение текста, уточненное за счет введения квантифицируемых дифференциальных признаков.*

Текст - это однозначно выделяемая, письменно зафиксированная, функционально-целесообразная, непрерывная, системно организованная последовательность слов или предложений, образующая сообщение, характеризующаяся композиционной целост-

* Ср., например, с определением, предложенным И.Р. Гальпериным (см. Гальперин, I974).

ностью, имеющая общий модальный характер, обладающая самостоятельной коммуникативно-информационной ценностью, соотносимая с определенным стилем на основе принципа сложного динамизма и отличающаяся от других текстов своеобразной статической и динамической статистической структурой.

Л И Т Е Р А Т У Р А

Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата: Изд-во Наука Казахской ССР, 1978.

Выготский Л.С. Психология искусства (Анализ эстетической реакции). – В кн.: Симпозиум по структурному изучению знаковых систем. – М.: Изд-во АН СССР, 1962, с. 118–122.

Гальперин И.Р. О понятии текст. – Вопросы языкознания, 1974, с. 68–77.

Мальцева Г.Ф. Некоторые количественные приемы описания индивидуального авторского стиля. – В кн.: Статистика текста, т. I. – Минск: Изд-во Белорусского государственного университета, 1969, с. 206–247.

Марков А.А. Об одном применении статистического метода. – В кн.: Известия императорской Академии наук, 6-я серия, IX, № 4, 1916.

Пустыльник Е.И. Статистические методы анализа и обработки наблюдений. – М.: Наука, 1968.

Ревзин И.И. О соотношении структурных и статистических методов в современной лингвистике. – В кн.: Вопросы статистики речи (Материалы совещания).–Л.: ЛГУ, 1958, с.45–56.

Сепетлиев Д. Статистические методы в научных медицинских исследованиях. – М.: Медицина, 1968.

Слепак Б.Я. Попытка индуктивного выделения функционально-стилевых разновидностей английского языка. – В кн.:Вопросы романо-германского языкознания, вып. 4. – Саратов: СГУ, 1975, с. 74–78.

Урбах В.Ю. Биометрические методы. – М.: Наука, 1964.

Kraus Jiří - Vašák Pavel. Попытка количественной типологии текстов. – In: Prague Studies in Mathematical Linguistics. – Praha: Academia, 1967, p. 77–88.

Williams C.B. A Note on the Statistical Analysis of Sentence-length as a Criterion of Literary Style. – In: Biomet-

rica, vol. XXXI, pt. 3-4, Cambridge University Press, March, 1940, p. 356-361.

Yule G.U. On Sentence-length as a Statistical Characteristic of Style in Prose: with Application to Two Cases of Disputed Authorship. - In: Biometrica, vol. XXX, 1938, p. 363-390.

A "PROLEGOMENA" TO THE STATISTICAL THEORY OF TEXT

Boris Slepack

S u m m a r y

An attempt has been made based on preliminary statistical observations to define more exactly the notion of text. The problem has barely been touched upon in modern stylostatistics. Mainly nonparametric methods of analysis have been employed. Text is treated as an intermittent sequence of words or sentences having integral communicativity - informational value as well as compositional and modal integrity, organized as a system, associated with an appropriate style due to integratively-dynamic speech trends and possessing compared to other texts a distinctive static and dynamic statistical composition.

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ МЕЖСЕГМЕНТНЫХ ГРАНИЦ В ДИАЛОГЕ

Д.И. Сливняк

В этой работе, как и в предшествующих (Сливняк, 1977, 1979), статистическими методами исследуется функционирование в диалоге трех коммуникативно важных грамматических категорий предложения: лица, коммуникативной целевой установки (утверждение/вопрос) и времени. Целью является изучение закономерностей, характеризующих "поверхностный уровень" организации диалога. В этом - существенное отличие данной работы от распространенного подхода, при котором основным объектом исследования является глубинный, семантический аспект коммуникативных процессов. Для осмысления получаемых результатов статистические закономерности сопоставляются с содержательными соображениями, относящимися к свойствам диалогической коммуникации. Тем самым, в какой-то мере, затрагивается и "глубинный уровень" диалога. Заметим, однако, что используемые при этом рассуждения никоим образом не претендуют на доказательную силу, - скорее их можно рассматривать как правдоподобные гипотезы.

Поведение лица, целевой установки и времени рассматривается в данной работе для предложений, соседних с границами сегментов определенного рода, на которые разбивается диалогический текст. Такими сегментами могут быть, например, тематические блоки (Теплицкая, 1975), сверхфразовые единства и т.п. Как будет показано, искомые закономерности в известной степени независимы от процедуры членения текста.

Исследование состоит из двух этапов. На первом текст членится интуитивным образом на отрезки, соответствующие "разговору на одну тему" (ТЕ-членение). Выделяется "оценочная" позиция, в которой наблюдаемые явления наиболее доступны для содержательного истолкования. А именно, рассматриваются предложения, расположенные в конце сегмента и одновременно - в конце реплики. Для различных типов предложения вводится особая характеристика - импульс, с помощью которой удается с единой точки зрения охватить совокупность наблюдаемых фактов. Исследуются также ситуации на других, интуитивно менее прозрачных позициях. Показано, что установленные для них

факты тесно связаны с закономерностями, имеющими место в оценочной позиции. Этой связи дается содержательное истолкование, опирающееся на понятие импульса, что, на наш взгляд, является основным результатом работы. На втором этапе интуитивная сегментация текста заменяется другим членением, основанным на весьма простой формальной процедуре ($-членение). Несмотря на то, что новое разбиение заметно отличается от предыдущего, характер наблюдаемых явлений в основном сохраняется. В то же время оно приводит и к некоторым новым интересным закономерностям.

Статистической обработке подверглись I2 пьес на трех языках: I. А.П.Чехов, Дядя Ваня (русск. яз.). 2. А.С. Пушкин, Борис Годунов (русск. яз.). 3. Б.Брехт, Матушка Кураж (нем. яз.). 4. Н.В.Гоголь, Ревизор (русск. яз.). 5. А. Корнейчук, Калиновая роща (русск. перев.). 6. Г.Тер-Григорян, Поговорим начистоту (русск. перев.). 7. Ж.Ромен, Кнок (франц. яз.). 8. А.Штейн, Флаг адмирала (русск. яз.). 9. С.Беккет, В ожидании Годо (франц. яз.). I0. В.Билль-Белоцерковский, Шторм (русск. яз.). II. Ж.Ануй, Антигона (франц. яз.). I2. Г. Сундукян, Пэпо (русск. перев.).

Кроме того, использовались записи русской разговорной речи из книги "Русская разговорная речь (тексты)", М., I978. Результаты приведены в табл. I-3, где строки I-I2 соответствуют отдельным пьесам, а строки PPI, PP2 - записям разговорной речи, разбитым на две части по I000 предложений в каждой. Сравнивая какие-либо две характеристики текста, считаем неравенство между ними установленным, если оно выполнено по крайней мере для девяти из I2 пьес.

Нашей целью является установление закономерностей, не связанных с особенностями конкретных языков. Поэтому и предлагаемая методика обработки данных должна быть достаточно общей, не зависящей от этих особенностей. При обработке статистических данных предложению приписываются лицо и время его сказуемого. С категорией времени были связаны определенные трудности, так как языки, на материале которых выполнялась работа, существенно различаются устройством этой системы. Указанная категория была подвергнута следующей унификации: видо-временные формы различных языков, обозначающие действие, заканчивающееся до момента речи, объединяются в унифицированное "прошедшее"; начинающиеся после момента речи - в "будущее", а те формы, для которых момент речи заключен между началом и концом действия, - в "настоящее",

При обработке текстов применялись следующие правила. Та или иная грамматическая категория приписывается предложению на основании формальных признаков. Например, praesens historicum рассматривается как настоящее, а не прошедшее; риторический вопрос - как вопрос, а не утверждение, и т.п. В сложных предложениях каждая самостоятельная предикация рассматривается как отдельное предложение; придаточные опускаются. Неполные предложения разного рода, эллипсисы, предложения с глаголом-сказуемым в повелительном или сослагательном наклонениях объединяются в специальный класс "нулевых предложений", не подвергаемый анализу. Исключение составляют эллипсисы, для которых все три параметра однозначно восстанавливаются из контекста.

Под типом предложения понимается класс предложений, задаваемый полным или неполным набором указанных параметров. При записи типа лицо обозначается цифрами 1, 2, 3; время - буквами П, Н, Б; целевая установка - буквами У, В. Примеры типов: БВ, 2У, 1.

Поведение трех рассматриваемых категорий исследуется сперва в позициях, на которых происходит смена темы диалога, - на тематических границах (ТГ). Поясним, что под этим понимается. Рассмотрим пример.*

В о й н и ц к и й. Я молчу. Молчу и извиняюсь. (Пауза).

Е л е н а А н д р е е в н а. А хорошая сегодня погода...

Здесь диалог резко переходит с одной темы на другую, которая с прежней не имеет ничего общего. Отчетливо ощущается обрыв связности текста. В таких случаях будем говорить о тематической границе 1-го рода. Другой пример.

Е л е н а А н д р е е в н а... И сегодня за завтраком вы опять спорили с Александром. Как это мелко!

В о й н и ц к и й. Но если я его ненавижу!

Е л е н а А н д р е е в н а. Ненавидеть Александра не за что, он такой же, как и все.

Здесь переходная фраза - реплика Войницкого связывает прежнюю тему разговора - характеристику поведения Войницкого с новой - характеристикой Серебрякова. Благодаря этому между двумя частями текста ощущается некоторая связь. Такие плавные переходы назовем границами 2-го рода. Границы 1-го рода

---

* Здесь и далее примеры взяты из пьесы А.П.Чехова "Дядя Ваня".

делят текст на отрезки, которые, в свою очередь, могут делиться границами 2-го рода.

Подобные разбиения уже неоднократно рассматривались. Так, отрезки, порожденные границами I-го рода, соответствуют "темам" в терминологии Е.М.Розенбаума (Розенбаум, 1975, с. 14), а более мелкие отрезки, связанные с границами 2-го рода, - его "подтемам". К последним близки "тематические блоки" (Теплицкая, 1975).

Мы не будем делать различия между тематическими границами I-го и 2-го рода. Отрезок текста, заключенный между соседними ТГ, назовем тематическим единством (ТЕ), а членение текста, произведенное указанным образом, - ТЕ-членением.

Будем относить к диалогическим все характеристики и представления, относящиеся к позициям на стыках реплик, а к монологическим - внутри реплик. В частности, будем говорить о диалогических и монологических ТГ. В обоих приведенных выше примерах ТГ были диалогическими. Приведем пример монологической ТГ.

А с т р о в... А как рассветет, ко мне поедем. Идеть? | У меня есть фельдшер, который никогда не скажет "идет",а "идеть". Мошенник страшный...

Разбиение текста на ТЕ - задача трудно формализуемая, в виду чего мы предпочли исходить из содержательного анализа текста. По отношению к каждой ТГ выделяются левая и правая позиции, в которых предложение непосредственно предшествует ТГ или следует за нею. В зависимости от того, где происходит смена темы, - внутри реплики или между ними, - получаем четыре позиции: левую и правую монологические (ЛМ и ПМ) и левую и правую диалогические (ЛД и ПД). Выясним, как меняются количественные соотношения между различными типами предложений при переходе от всего текста к каждой из четырех указанных позиций. Начнем с ЛД-позиции, которую назовем оценочной, так как в ней яснее всего видна природа изучаемых явлений. Закономерности, наблюдаемые в этой позиции, послужат ключом для интерпретации фактов, установленных для других трех позиций.

Будем исходить из следующего представления: предложение, расположенное в конце реплики, может, в зависимости от своего типа облегчить или затруднить смену темы. Второе должно иметь место, если последняя фраза говорящего побуждает слушателя к продолжению разговора в прежнем направлении. Например, если реплика заканчивается вопросом, маловероятно,

чтобы слушатель перешел к новой теме, так как вопрос обычно требует реакции-ответа, иначе говоря, является диалогическим стимулом. Вообще, чем больше для предложения, завершающего реплику, шансы оказаться диалогическим стимулом, тем менее вероятно для него завершить ТЕ, то есть оказаться в ЛД-позиции*.

Введем, на интуитивном уровне, соответствующую характеристику типа предложения - его импульс, под которым будем понимать относительную способность предложений данного типа выступать в качестве диалогического стимула. Импульс является количественной величиной, но в наши цели не входит его вычисление (и, тем самым, точное количественное определение). В то же время, мы будем сравнивать импульсы разных типов.

Возвращаясь к приведенному примеру, можно сказать, что импульс вопроса больше, чем импульс утверждения, так как вопрос, оставшийся без ответа, - аномалия, а утверждение, замыкающее собою ТЕ, - обычное явление. Итак, чем выше импульс у предложений данного типа, расположенных в конце реплики, тем меньше для них шансы оказаться в ЛД-позиции, и наоборот. Иначе говоря, если ввести меру "притяжения" того или иного типа к ЛД-позиции, то всякое неравенство, связывающее значения этой меры для двух типов, можно интерпретировать как обратное соотношение между соответствующими импульсами. Указанным способом и будет получена система импульсных соотношений, играющая основную роль в данной работе.

Перейдем к определению упомянутой выше меры. Пусть Q_X - массив всех предложений некоторого типа X, расположенных слева от границы реплики, а Q_X^T - совокупность тех из них, которые замыкают ТЕ. Пусть Q_X и Q_X^T обозначают также число предложений в соответствующем массиве. Тогда отношение

$$B_X^д = Q_X^T : Q_X \qquad (I)$$

есть доля массива Q_X^T по отношению к Q_X, так что сравнивая значения $B_X^д$ при различных X, можно судить об относительной связи ЛД-позиции и типа X. Индекс "д" указывает, что речь идет о диалогической характеристике, описывающей ситуацию на стыках реплик.

---

* Здесь мы ограничиваемся ситуациями, в которых диалогическим стимулом нагружено предложение, замыкающее реплику, причем сам стимул индуцирует продолжение разговора на ту же тему. Как видно из дальнейшего, такое упрощение допустимо.

Приведем некоторые результаты, полученные для B^{g}_{x}. Ограничимся наиболее важными и простыми типами предложений, задаваемыми путем фиксации только одного параметра: X = В, 2, П и т.д. Начнем с оппозиции "утверждение-вопрос". Как уже отмечалось, кажется очевидным, что вопрос обладает бóльшим импульсом, чем утверждение. Соответствующее неравенство между B^{g}_{x} должно иметь вид

$$B^{g}_{в} < B^{g}_{у}. \qquad (2)$$

И действительно, как видно из таблицы I, оно всегда выполняется.

Рассмотрим с точки зрения понятия импульса и другие две категории - лица и времени. При этом, говоря об истолковании соответствующих неравенств, будем пренебрегать наличием в тексте вопросов и ограничимся одними утверждениями - доля вопросов в тексте обычно весьма мала, а в рассматриваемой позиции, согласно (2), и подавно. Для категории лица установлены неравенства

$$B^{g}_{2} < B^{g}_{3}, \qquad (3)$$

$$B^{g}_{3} < B^{g}_{1}, \qquad (4)$$

согласно которым наибольшим импульсом обладает второе лицо, а наименьшим - первое. Это можно объяснить следующим образом. Из двух типов бóльшим импульсом должен обладать тот, в рамках которого слушающий получает от говорящего более актуальную, важную для себя информацию и, следовательно, испытывает бóльшую потребность в реакции на нее. В частности, для слушателя более актуально высказывание о нем самом (2 лицо), чем о говорящем (I лицо), а 3 лицо занимает промежуточное положение.

Для категорий унифицированного времени получены неравенства

$$B^{g}_{п} < B^{g}_{б}, \quad B^{g}_{н} < B^{g}_{б}. \qquad (5)$$

Как видим, будущее имеет наименьший импульс, что, возможно, связано с наименьшей осведомленностью о нем участников диалога. Для $B^{g}_{н}$ и $B^{g}_{п}$ устойчивого неравенства получить не удалось.

Итак, с помощью величин B^{g}_{x} получена некоторая система соотношений между импульсами. Выясним, насколько она согласуется с фактами, наблюдаемыми для остальных позиций ПД, ЛМ и ПМ. Соответствующие величины, аналогичные B^{g}_{x}, обозначим

$C^д_х, B^м_х, C^м_х$, где буква "С" относится к позиции справа от ТГ, а индексы "д" и "м" – к диалогическим и монологическим величинам. Определяются они аналогично $B^д_х$.

Сделаем это для $C^м_х$. Введем монологический аналог массива $Q_х$. Для этого вместо пар предложений, расположенных на стыках реплик, рассмотрим пары соседних предложений, принадлежащих одной и той же реплике. Пусть $S_х$ – массив предложений типа $х$, являющихся правыми членами таких пар, то есть не являющихся начальными в какой-либо реплике; $S^т_х$ – совокупность тех из них, которые начинают ТЕ. Тогда

$$C^м_х = S^т_х : S_х . \qquad (6)$$

На определении величин $C^д_х, B^м_х$ останавливаться не будем.

Рассмотрим результаты подсчета $B^м_х, C^м_х, C^д_х$ (табл. I,2). Как видно из таблицы I, величина $B^м_х$ ведет себя подобно $B^д_х$, хотя и в менее четко выраженной форме. Именно, для $B^м_х$ выполняются монологические аналоги неравенств (2), (3), (5):

$$B^м_в < B^м_у, B^м_2 < B^м_3, B^м_п < B^м_б, B^м_н \le B^м_б, \qquad (7)$$

но первое из них выражено слабее, чем для $B^д_х$. Кроме того, выполняется неравенство

$$B^м_2 < B^м_1 \qquad (8)$$

– монологический аналог вытекающего из (3), (4) соотношения

$$B^д_2 < B^д_1 . \qquad (9)$$

Однако (4) перестает быть верным. Иными словами, картина для предложений, замыкающих ТЕ, примерно одинакова как внутри, так и на стыках реплик, но на последних выражена резче.

Для $C^м_х$ выполняются неравенства, обратные (2), (3), (9):

$$C^м_в > C^м_у, C^м_2 > C^м_3, C^м_2 > C^м_1 . \qquad (10)$$

Но аналог (4), как и для величин $B^м_х$, уже не имеет места.

При переходе от $C^м_х$ к $C^д_х$ общая картина остается прежней – выполняются диалогические аналоги соотношений (10):

$$C^д_в > C^д_у, C^д_2 > C^д_3, C^д_2 > C^д_1 , \qquad (11)$$

причем первое из них выражено слабее, чем в (10).

Таким образом, картина для предложений, открывающих ТЕ, противоположна картине для предложений, замыкающих ТЕ: при переходе от $B_х$ к $C_х$ знаки неравенства меняются на обратные, а первое место с точки зрения информативности переходит от диалогических позиций к монологическим. Заметим, что для позиций ПД, ЛМ и ПМ не выполняются аналоги (4). Как будет по-

Таблица I

	$\frac{B^{э}_{в}}{B^{э}_{у}}$	$\frac{B^{э}_{2}}{B^{э}_{3}}$	$\frac{B^{э}_{3}}{B^{э}_{1}}$	$\frac{B^{э}_{п}}{B^{э}_{б}}$	$\frac{B^{э}_{н}}{B^{э}_{б}}$	$\frac{B^{м}_{в}}{B^{м}_{у}}$	$\frac{B^{м}_{2}}{B^{м}_{3}}$	$\frac{B^{м}_{п}}{B^{м}_{б}}$	$\frac{B^{м}_{н}}{B^{м}_{б}}$	$\frac{B^{м}_{2}}{B^{м}_{1}}$
I	0,20	0,34	0,90	0,7I	I,05	0,00	0,64	0,8I	I,03	0,64
2	0,II	0,84	0,90	0,56	0,35	0,96	0,46	0,54	0,55	0,32
3	0,52	0,7I	I,04	I,08	I,06	0,87	0,75	0,59	0,76	0,72
4	0,I8	0,55	0,78	0,78	0,62	0,30	0,6I	0,57	0,64	0,27
5	0,2I	0,4I	0,92	0,58	0,77	0,43	0,60	0,28	0,64	0,70
6	0,39	0,63	0,9I	0,87	I,20	0,39	0,52	0,50	0,5I	0,89
7	0,I4	0,33	0,87	0,7I	0,7I	0,54	0,I9	3,02	3,02	0,23
8	0,33	0,6I	0,79	0,77	0,70	0,56	0,98	0,58	0,57	I,I2
9	0,45	0,70	0,73	0,75	0,70	0,86	0,83	0,36	0,44	0,59
I0	0,26	0,40	I,20	0,60	0,72	0,I3	I,35	0,60	I,0I	I,37
II	0,I3	0,22	0,96	I,24	0,87	0,50	0,33	I,55	I,50	0,20
I2	0,34	0,58	0,70	0,70	0,84	0,96	0,87	0,56	0,78	0,78
PPI	0,I7	0,40	0,89	I,93	I,34	0,36	0,22	I,65	2,29	0,27
PP2	0,36	0,00	0,82	0,45	0,78	0,67	I,95	0,50	0,95	3,43

казано, поведение I-го лица вообще является в некотором смысле аномальным.

Сходство неравенств для $C^{э}_{х}$, $B^{м}_{х}$, $C^{м}_{х}$ с неравенствами для $B^{э}_{х}$ делает естественным расширение понятия импульса на рассматриваемые три позиции. Условимся смотреть на импульс как на характеристику, присущую данному типу независимо от места предложения в речевой цепи. Покажем, что при этом содержательный характер импульса сохраняется и в ЛМ, ПМ и ПД-позициях.

Рассмотрим сперва ЛМ-позицию. В условиях спонтанного диалога любое предложение внутри реплики рискует оказаться в ней последним из-за вступления в разговор слушающего. В частности, когда говорящий заканчивает ТЕ внутри реплики, он, по существу, еще не знает, будет ли смена темы монологической или диалогической. Поэтому его речевая стратегия в конце такого ТЕ имитирует ситуацию слева от диалогической ТГ. Этим и можно объяснить сходство (7), (8) с (2) - (5). Однако продолжение монолога по другую сторону от ТГ, как правило, все же прогнозируется говорящим. Поэтому он обладает некоторой свободой в формировании перехода через ТГ, что делает для него менее обязательными соотношения между импульсами, присущие

Таблица 2

	$\frac{C_8^м}{C_4^м}$	$\frac{C_2^м}{C_3^м}$	$\frac{C_2^м}{C_1^м}$	$\frac{C_6^э}{C_4^э}$	$\frac{C_2^э}{C_3^э}$	$\frac{C_2^э}{C_1^э}$		$\frac{C_8^м}{C_4^м}$	$\frac{C_2^м}{C_3^м}$	$\frac{C_2^м}{C_1^м}$	$\frac{C_6^э}{C_4^э}$	$\frac{C_2^э}{C_3^э}$	$\frac{C_2^э}{C_1^э}$
I	2,90	I,94	I,53	I,2I	0,86	I,I2	8	I,53	I,30	I,3I	I,I7	2,I0	I,95
2	I,85	I,55	I,75	I,92	2,28	4,70	9	3,0I	2,25	2,32	2,I0	I,27	I,43
3	3,27	I,95	I,78	I,6I	I,95	I,66	I0	2,77	2,36	I,96	I,89	I,04	I,03
4	2,I6	3,00	I,55	0,9I	0,87	3,05	II	2,36	I,79	2,25	2,22	I,43	3,I4
5	3,05	I,46	I,02	I,88	I,27	I,77	I2	I,39	0,96	0,97	I,53	I,4I	I,55
6	I,94	2,37	I,24	0,92	I,07	0,75	PPI	I,80	I,II	I,4I	2,35	2,44	3,I0
7	3,76	2,54	I,76	2,24	2,II	I,69	PP2	5,I2	2,77	3,34	I,90	2,26	2,64

диалогическим ТГ. В результате неравенства (7), (8) оказываются "ослабленным" вариантом неравенств (2) - (5).

Перейдем к ПМ-позиции. Как было сказано, относящиеся к ней неравенства (10) являются обращением (2) - (5). Таким образом, с возрастанием импульса величина $C^{м}_{х}$ не убывает, подобно $B^{д}_{х}$, а, наоборот, возрастает. Но $C^{м}_{х}$ является мерой притяжения предложений типа X к ПМ-позиции. Следовательно, в монологе прослеживается весьма четкая тенденция: говорящий предпочитает начинать новое ТЕ с предложений сильного импульса. Можно предположить, что такие предложения мобилизуют внимание слушателя для восприятия нового ТЕ, являясь своеобразными "красными строками".

Остается обсудить ПД-позицию. Связи между $C^{д}_{х}$ аналогичны связям между $C^{м}_{х}$, но выражены менее четко (табл. 2). Это может быть объяснено следующим образом. Диалогическая связность, поскольку она выходит за рамки индивидуальной речевой деятельности, является значительно более структурированной и четкой, чем монологическая. Поэтому ее отсутствие (то есть отсутствие стимула слева) на ТГ I-го рода - достаточно сильный сигнал для смены ТЕ, уменьшающий необходимость в специальных сигналах справа от ТГ. Присутствия же ее, хотя и в ослабленной форме, на ТГ 2-го рода опять-таки достаточно для того, чтобы заметно исказить картину, присущую ПД-позиции. Из этих же рассуждений следует, что в ЛД-позиции соотношения между импульсами должны быть выражены резче, чем в ПД-позиции. Как видно из анализа таблиц I, 2, это действительно имеет место.

Резюмируя изложенное, можно сказать, что

- независимо от расположения ТЕ относительно границ реплик существует тенденция начинать его предложениями сильного и заканчивать предложениями слабого импульса;

в позиции слева от ТГ соотношения между импульсами выражены сильнее для диалогических, а в позиции справа - для монологических ТГ;

- для диалогических ТГ ведущей является позиция слева от ТГ.

Эти факты дают основание рассматривать импульс как относительную способность предложений данного типа вызывать внутреннюю реакцию слушателя. В каждой из рассмотренных позиций эта способность реализуется по-своему. В частности, в ЛД-позиции она выступает как мера "внешнего" диалогического стимула.

Определим теперь формальную процедуру членения текста, которой посвящена вторая часть работы. В лингвистике сейчас известны весьма совершенные процедуры такого рода, например, метод Б.М.Гаспарова (Гаспаров, 1975). Однако именно в силу своих достоинств они трудоемки, что нежелательно при обработке больших массивов текста. В то же время, можно не предъявлять особенно высоких требований к сегментации, если целью является установление сильно выраженных статистических закономерностей.

В основе предлагаемого членения текста лежит явление субституции (повтор, анафора и т.д.), широко использовавшееся в подобных процедурах на начальном этапе развития лингвистики текста. Будем считать, что имеет место отношение субституции, в котором элемент (знаменательное слово или словосочетание) В является замещающим, а элемент А - замещаемым, если А и В обозначают один и тот же денотат, причем предложение, содержащее В, следует непосредственно за предложением, содержащим А, и выполняется одно из следующих условий:
1) А и В принадлежат одной лексеме; 2) элемент В - личное местоимение, денотат которого совпадает с денотатом А либо включает его в себя, или наоборот; 3) В - дейктический элемент (местоимение или местоименное наречие), отсылающий к А.

Предполагается, что в эллипсисах, поддающихся однозначной расшифровке, недостающие члены восстановлены. Легко видеть, что отношение субституции в этой трактовке весьма упрощено и "грамматикализовано": сюда не включены способы выражения одного и того же денотата при помощи синонимов, перифраз и т.д., а также субституция для предложений, не расположенных рядом.

Двигаясь от любого замещающего элемента по цепочке субституций в обе стороны до конца цепочки, получаем сегмент текста, связанный с данным элементом. Разбиение текста на такие сегменты и есть искомое $ -членение текста. Мы ограничимся сегментами, содержащими более одного предложения и не лежащими целиком внутри другого сегмента (очевидно, сегменты могут перекрываться).

Введем аналоги позиций, левых и правых по отношению к ТГ. При этом не будем различать монологических и диалогических позиций, так как для ТЕ-членения между ними не было обнаружено существенных различий. Назовем начальными (НП) предложения, начинающие сегмент, и свободными (СП) - предложения, находящиеся вне сегментов.

Кажется естественным в качестве аналога позиции, открывающей ТЕ, взять массив НП. Введем на нем меру, соответствующую C_x. Пусть U_x - число предложений типа x во всем тексте, а U_x^T - число начальных предложений типа x. По аналогии с (6) определим

$$C'_x = U_x^T : U_x . \qquad (12)$$

Если между позицией справа от ТГ и массивом НП действительно существует сходство, для величин C'_x следует ожидать выполнения аналогов (2) - (5). Эти ожидания оправдываются - выполняются неравенства (табл. 3):

$$C'_б > C'_у , C'_п > C'_б , C'_н > C'_б , C'_2 > C'_3 , \qquad (13)$$

$$C'_2 > C'_1 . \qquad (14)$$

Последнее из неравенств (13) требует пояснений. Для 3-го лица характерно выражение одного и того же денотата с помощью синонимов, перифраз и т.п. Однако эти способы не включены в отношение субституции, используемое нами. В результате недооценивается способность 3-го лица к продолжению цепочек замещений, что способствует выполнению данного неравенства. Таким образом, оно выражает не только ожидаемое соотношение между импульсами 2-го и 3-го лица, но и отмеченное несовершенство модели. Заметим, что аналог (4) по-прежнему не выполняется. Представляет также интерес неравенство

$$C'_п > C'_н , \qquad (15)$$

выполняющееся для S -членения текста, - в отличие от ТЕ-членения. Можно показать, что в его основе лежат некоторые особенности поведения 3-го лица, а именно, типов ЗП и ЗН.

В целом можно считать установленным, что при S -членении начальные предложения являются удовлетворительным соответствием для позиций справа от ТГ. При этом полученные в первой части соотношения между импульсами сохраняются, несмотря на новую трактовку самого импульса.

Найдем теперь соответствие для позиций слева от ТГ. Казалось бы, в этой роли должны выступать предложения, заканчивающие сегмент. Однако обработка данных на этом массиве привела к отрицательным результатам: ни одна из ожидаемых закономерностей не оказалась устойчиво выполненной. Эмпирический поиск дал довольно неожиданный результат: искомым аналогом позиции слева от ТГ оказался массив СП.

Перейдем к изложению результатов, полученных для этого

массива. Начнем с определения аналога меры B_x. Пусть, как и раньше, U_x - число предложений типа x во всем тексте, а V_x^T - число свободных предложений типа x. По аналогии с (I) следовало бы в качестве искомой меры взять отношение $V_x^T : U_x$. Однако целесообразно несколько изменить это определение, чтобы "очистить" новую меру от влияния неравенств (I3), (I4), связанных с массивом НП. Исключим последний из текста, положив

$$B'_x = V_x^T : (U_x - U_x^T).$$

Заметим, что при ТЕ-членении такая предосторожность была излишней ввиду малочисленности предложений, расположенных справа от ТГ.

Как видно из таблицы 3, для B'_x выполняются аналоги (2), (3), (5):

$$B'_в < B'_у,\ B'_2 < B'_3,\ B'_п < B'_б,\ B'_н \le B'_б. \qquad (I6)$$

Таблица 3

	$\frac{C'_в}{C'_а}$	$\frac{C'_п}{C'_б}$	$\frac{C'_н}{C'_б}$	$\frac{C'_2}{C'_1}$	$\frac{C'_2}{C'_3}$	$\frac{B'_в}{B'_а}$	$\frac{B'_2}{B'_3}$	$\frac{B'_п}{B'_б}$	$\frac{B'_н}{B'_б}$	$\frac{B'_1}{B'_3}$
I	I,67	I,42	I,08	I,3I	I,4I	I,I2	0,59	0,63	I,04	0,7I
2	2,05	I,07	I,I0	I,70	I,6I	0,99	0,67	0,59	I,05	0,54
3	I,57	I,55	I,7I	I,I9	I,I7	0,90	0,80	0,55	0,84	0,74
4	2,00	I,32	I,I7	I,II	I,36	0,85	0,90	0,79	0,97	0,77
5	I,8I	I,36	I,2I	I,40	I,40	0,77	0,74	0,86	0,94	0,69
6	I,47	I,40	0,99	I,I9	I,II	0,86	I,06	0,75	0,89	0,92
7	I,82	I,6I	I,43	I,38	I,52	0,75	0,95	0,63	0,97	0,56
8	I,76	I,46	I,37	I,I2	I,22	0,97	0,68	0,83	I,00	0,65
9	I,44	I,I6	0,96	I,23	I,46	0,94	0,70	0,42	0,75	0,73
I0	I,95	I,54	I,37	I,74	I,68	0,97	0,93	0,8I	0,90	I,07
II	I,3I	I,I0	I,I3	I,28	I,35	0,82	0,69	0,90	I,06	0,55
I2	I,33	I,II	I,04	I,I6	I,07	0,73	0,69	0,85	0,96	0,63
PPI	2,47	I,36	I,03	I,59	I,67	0,88	I,37	0,79	0,85	I,09
PP2	0,96	I,3I	I,I2	2,25	I,84	0,7I	I,80	0,64	0,7I	I,45

Имеет место также "зеркальное" соответствие (I5): $B'_п < B'_н$. Таким образом, массив СП действительно играет роль позиции слева от ТГ при S-членении. Первое лицо и на массиве СП ведет себя аномально - выполняются неравенства

$$B'_1 < B'_2,\ B'_1 < B'_3, \qquad (I7)$$

обратные соответствующим соотношениям для B_x.

Возвращаясь к содержательной трактовке импульса, отметим, что результаты, полученные для S-членения, не противоречат, по нашему мнению, взгляду на импульс как на меру "внутренней" реакции слушателя, хотя реализуется она уже по-иному.

Остановимся также на результатах обработки разговорной речи. Как видно из таблиц I-3 (строки РРI, РР2), для нее в основном наблюдаются те же закономерности. Все отклонения от них относятся к левым позициям ЛД, ЛМ, СП. На первом месте - ЛМ-позиция, в которой из пяти рассмотренных неравенств выполняется для обеих строк РРI, РР2 только одно. Отметим также, что неравенства, связывающие в СП-позиции величину B'_3 с B'_2 и B'_1, обратны для обеих строк РРI, РР2 соответствующим неравенствам (I6), (I7). Однако малый объем выборки не дает оснований для каких-либо выводов из этих наблюдений.

Обсудим теперь в целом совокупность полученных выше неравенств. Как отмечалось, мы ограничиваемся простейшими типами предложений, характеризуемых одним параметром. Рассматриваемые в работе неравенства связаны с попарным сравнением типов, относящихся к одной грамматической категории. Это приводит к семи сравнениям: трем - в системе лица (I-2, I - 3, 2-3), трем - в системе времени (П-Н, П-Б, Н-Б) и одному - для целевой установки (У-В). Каждое допускает проверку на шести позициях: четырех - для ТЕ-членения (ЛД, ЛМ, ПД, ПМ) и двух для S-членения (НП, СП). Для каждого типа в каждой из этих шести позиций подсчитывается определенная числовая числовая характеристика (меры В, С). Для сравнений У-В, 2-3 на всех шести позициях и для сравнений П-Б, Н-Б на всех позициях, кроме ПД и ПМ, наблюдаются устойчивые неравенства между соответствующими мерами В, С. При этом если известен знак неравенства в одной из позиций, его можно однозначно получить и во всех остальных. Именно, знак неравенства для правых позиций ПД, ПМ, НП противоположен знаку для левых - ЛД, ЛМ,СП, внутри же этих подгрупп знак сохраняется.

В соответствии с принятой в работе точкой зрения мы считаем, что в основе неравенств, подчиняющихся этому правилу, лежат импульсные соотношения, то есть определенная упорядоченность импульсов сравниваемых типов. Для оппозиций "вопрос-утверждение" и (в меньшей степени) "2 лицо - 3 лицо" такая гипотеза представляется интуитивно очевидной. Для оппозиций "прошедшее-будущее" и "настоящее-будущее" это менее очевидно, однако маловероятно, чтобы соотношения, до такой степени сходные по своему поведению, имели разное происхож-

дение. Вообще, будем считать, что в основе всякого сравнения, подчиняющегося описанному правилу, лежит некоторое импульсное соотношение. Его знак совпадает со знаком соответствующего неравенства для мер С в правых позициях и противоположен знаку неравенств для мер В в левых. Тем самым, если восстанавливать знак импульсного соотношения таким способом, он не зависит от позиции, на которой производится сравнение.

Обсудим, как ведет себя с этой точки зрения I-е лицо. В ЛД-позиции для него получен самый низкий в системе лица импульс. Для позиций ЛМ, ПД, ПМ и НП соотношение между импульсами I и 2 лица сохраняется, но для I и 3 лица устойчивого соотношения обнаружить не удается. Наконец, в СП-позиции I лицо оказывается впереди и 2, и 3 лица. Ввиду этого нельзя говорить об импульсном характере неравенств, связанных с I лицом, и, возможно, даже о самом импульсе I лица.

В заключение сделаем один подсчет. Семь попарных сравнений типов, каждое на шести позициях, приводят к 42 комбинациям, в каждой из которых в принципе можно ожидать устойчивого неравенства для мер В, С. Такие неравенства удалось обнаружить для 30 комбинаций. Тем самым для них исследуемые соотношения оказались достаточно сильными, чтобы проявиться на фоне помех. Бо́льшая часть полученных неравенств (20 из 30) объясняется при помощи категории импульса.

Можно надеяться, что понятие импульса окажется полезным и за пределами рассмотренных здесь грамматических категорий.

ЛИТЕРАТУРА

Гаспаров Б.М. Принципы синтагматического описания уровня предложения. - Труды по русской и славянской филологии, вып. 23. Тарту, 1975.

Розенбаум Е.М. Основы обучения диалогической речи на языковом факультете педагогических вузов. М., 1975.

Сливняк Д.И. О количественной связи двух характеристик предложения. - Вестник общественных наук АН Армянской ССР, № I, 1977.

Сливняк Д.И. Об одном способе статистического анализа диалога. - Вестник общественных наук АН Армянской ССР, № 6, 1979.

Теплицкая Н.И. О структуре диалогического текста. - Сб. научн. трудов МГПИИЯ им. Тореза, Вопросы романо-герм. филологии, вып. 84. 1975, с. 314-331.

STATISTICAL CHARACTERISTICS OF INTERSEGMENTAL BOUNDARIES IN THE DIALOGUE

Dmitri Slivnyak

S u m m a r y

The subject - matter of the present investigation is the statistical analysis of dialogic texts in several languages intuitively segmented into thematic units. A kind of formal segmentation based on the substitution phenomenon has also been considered. On the general background of the whole text, close to segmental boundaries, frequency shifts of sentences classified into three categories are studied: person, communicative status (affirmation/question), tense.

A specific characteristic of the sentence - impulse is introduced, which allows us to elucidate from the unified point of view, the majority of regularities obtained.

It is observed that these regularities are relatively independent irrespective of the method of segmentation as well as the coincidence of segmental and cue boundaries.

ОПЫТ КЛАССИФИКАЦИИ ТЕКСТОВ С ПОМОЩЬЮ КЛАСТЕР-АНАЛИЗА

Ю. Тулдава

В статье рассматриваются основные принципы кластер-анализа и описывается эксперимент проведения такого анализа на материале 20 текстов художественной прозы с целью сравнения результатов отдельных опытов, проведенных на основе разных наборов квантитативно-лингвистических характеристик текстов. Эксперимент проводился в Группе прикладной лингвистики ТГУ. Использовалась ЭВМ ЕС-1022 и машинная программа, разработанная Р. Ээремаа (1978а, 1978б) для практического применения кластерного метода $B_к$.*

Основные принципы кластер-анализа. Кластер-анализ можно определить как совокупность методов, предназначенных для разбиения некоторого множества объектов на группы, или к л а с т е р ы (англ. cluster 'группа, кучка, пучок') так, чтобы в каждой группе находились в некотором смысле наиболее близкие между собой объекты. Методы кластер-анализа относятся к группе процедур, именуемых в совокупности методами распознавания образов (Елисеева И.И., Рукавишников В.О., 1977, с. 9), а в более узком смысле методы кластер-анализа можно отнести к методам классификации многомерных наблюдений (см., например, Айвазян С.А. и др., 1974). Особенностью классификации многомерных наблюдений является то, что каждый объект описывается с помощью набора (множества) зафиксированных на нем признаков, причем для построения классификации таких объектов используется данный набор признаков в их взаимосвязи. Наиболее характерными чертами кластер-анализа считаются образование единой меры, охватывающей ряд признаков, и чисто количественное решение вопроса о классификации (группировке) объектов наблюдения (Боярский А.Я., 1977, с. 8).

Существует ряд разновидностей кластер-анализа, но для них является общим наличие трех основных типа данных, используемых при проведении анализа: исходные многомерные данные, данные о близости, данные о кластерах (Крускал Дж., 1980, с. 21). Соответственно можно различать три этапа исследования: на первом, подготовительном, этапе упорядочиваются ис-

* См. также статью Р. Ээремаа в настоящем сборнике.

ходные данные, а на двух последующих этапах измеряется близость (сходство или различие) между классифицируемыми объектами и конструируется кластер-система, которая объединяет объекты при различных уровнях близости. Два последних этапа выполняются, как правило, с помощью автоматических процедур классификации на ЭВМ. Решением задачи кластер-анализа является разбиение, удовлетворяющее определенному критерию качества.

Общую ситуацию при проведении кластер-анализа можно формально описать следующим образом (ср. Дюран В., Оделл П., 1977).

Имеется исходное множество $T = \{T_1, T_2, \ldots, T_n\}$ из n объектов (например, текстов), принадлежащих некоторой популяции π_T. Рассматривается некоторое множество наблюдаемых характеристик (признаков) $C = (C_1, C_2, \ldots, C_k)'$, которыми обладает каждый объект из T. Наблюдаемые характеристики могут быть как качественными, так и количественными. В данном случае рассматриваются количественные характеристики, т.е. измерения таких характеристик. Результат измерения j-й характеристики T_i объекта обозначается символом x_{ij}, а вектор $X_i = [x_{ij}]$ размерности $k \times 1$ будет отвечать каждому ряду измерений (для i-го объекта). Сказанное можно проиллюстрировать с помощью таблицы (такой вид имеет обычно таблица исходных данных):

Объекты \ Признаки	C_1	C_2	...	C_k	
T_1	x_{11}	x_{12}	...	$x_{1k} = X_1$	
T_2	x_{21}	x_{22}	...	$x_{2k} = X_2$	X
⋮					
T_n	x_{n1}	x_{n2}	...	$x_{nk} = X_n$	

Следовательно, для множества объектов T мы располагаем множеством векторов измерений $X = \{X_1, X_2, \ldots X_n\}$, которое описывает множество T. Задача кластер-анализа заключается в том, чтобы на основе данных, содержащихся в множестве X, разбить множество объектов T на m (причем $m < n$) кластеров (подмножеств) $\pi_1, \pi_2, \ldots, \pi_m$ так, чтобы каждый объект T принадлежал одному и только одному подмножеству

разбиения (разбиение на непересекающиеся кластеры). Однако, учитывая то, что во многих областях исследования (в том числе в лингвистике) реальные системы характеризуются, как правило, "размытостью" границ, в новейших приложениях кластер-анализа предусматривается также отнесение объектов по кластерам, разрешающее пересечение, т.е. конструируются кластер-системы, где кластеры могут покрываться (Эзремаа Р., 1978а, с. 91). В данной работе мы используем как обычную процедуру разбиения объектов на непересекающиеся кластеры, так и один из алгоритмов кластеризации с пересечением.

Необходимо подчеркнуть, что кластер-анализ, как и всякий другой метод классификации, субъективен и относителен в том смысле, что результаты анализа целиком определяются теми признаками, которые положены в его основу. Классификации, основанные на большом количестве и разнообразии признаков, будут, конечно, более эффективны для определения "естественного" порядка среди объектов и явлений (если удается использовать всю доступную информацию о признаках классифицируемых объектов). В других случаях, когда исследователя интересуют только некоторые свойства объектов, или когда кластер-анализ должен служить нуждам некоторых специальных практических приложений, можно довольствоваться небольшим числом специально отобранных признаков. В настоящей работе ставится как раз такая ограниченная задача – выявить возможности классификации текстов с помощью кластер-анализа на основе некоторых известных в практике квантитативной лингвистики формальных характеристик статистической структуры текста. При этом встает вопрос о сходстве результатов различных опытов, проведенных на одном и том же материале, но на основе разных наборов признаков.

Общая задача кластеризации текстов, в том числе художественных текстов, возникает в исследованиях по изучению типологии текстов (для стилистических, педагогических и др. целей), при решении задач в области информатики, аттрибуции текстов и т.д.

Исходные данные. В данной работе подвергаются кластер-анализу 20 текстов – выборок по 5000 словоупотреблений из авторской речи 20 произведений современной эстонской художественной прозы (см. Список текстов в конце статьи). Считается, что выборки по 5000 словоупотреблений (каждая из которых в свою очередь разделена на 5 порций по 1000 словоупотреблений) достаточны для выявления некоторых существен-

ных формальных показателей интересующей нас статистической организации текстов (в сравнительном плане при одинаковых объемах текстов). На материале названных 20 текстов было проведено три опыта на основе разных наборов измерений квантитативно-лингвистических характеристик текстов. Наборы характеристик следующие:

- покрываемость текста словоформами (опыт № I);
- лексический спектр (опыт № 2);
- динамика роста словаря (опыт № 3);

Конкретные исходные данные приводятся в таблицах I - 3.

В первом опыте рассматривается "покрываемость текста", т.е. относительные накопленные частоты словоформ при рангах i = I, i = I0, i = 50 и т.д. (ранги убывающих частот в частотном списке словоформ для данной выборки). Покрываемость текста фиксированным массивом наиболее частых словоформ считается одной из важнейших характеристик квантитативной типологии языков (Бектаев К.Б., 1978, с. 52). Показатели покрываемости текста могут служить также дифференциальными признаками индивидуальных стилей. На нашем материале видно, например (см. табл. I), что при i = I0 относительная накопленная частота словоформ колеблется от I0,I % (текст № I0) до I6,7 % (текст № I7), т.е. десять наиболее частотных словоформ покрывают у различных авторов различные доли в тексте. Для проведения кластер-анализа используются в первом опыте девять численных показателей в каждом ряду измерений, т.е. каждый текст описан набором из девяти количественных характеристик покрываемости.

Во втором опыте тексты характеризуются набором численных показателей т. наз. лексического спектра, т.е. долями словоформ с данной частотой в словаре рассматриваемого текста (выборки). Лексический спектр считается также важным типологическим показателем языков и текстов. В данном случае можно констатировать существенное колебание долей словоформ с частотой F = I (см. табл. 2), например, у текстов № 3 и № I0 этот показатель равняется 73,29 и 8I,74 % соответственно. В данном опыте используются двенадцать численных показателей лексического спектра для каждого текста.

В третьем опыте рассматриваются данные об объеме словаря в зависимости от объема текста. Фиксируется количество разных словоформ при объемах текста от N= I000 до N= 5000 словоупотреблений (см. табл. 3). Показатель объема словаря при данном объеме текста используется часто в лингвостатисти-

Таблица I

Опыт № I. Исходные данные: покрываемость текста словоформами (%)

№ текста	Автор	Ранги словоформ								
		i=1	10	50	100	500	1000	1500	2000	2500
1.	Э.Бээкман	2,7	11,0	21,4	28,3	50,6	62,6	72,6	82,6	92,6
2.	В.Гросс	2,6	12,0	23,3	30,4	52,4	64,2	74,2	84,1	94,1
3.	А.Хинт	2,8	13,8	26,7	33,7	57,5	70,7	80,7	90,7	100,0
4.	Х.Кийк	3,5	12,9	24,4	31,2	53,6	65,2	75,2	85,2	95,2
5.	Я.Кросс	5,7	13,8	25,0	31,3	52,2	62,7	72,7	82,7	92,8
6.	П.Куусберг	2,4	14,3	27,0	34,7	57,6	69,7	79,7	89,8	99,8
7.	Л.Промет	3,4	14,4	25,5	32,7	54,4	66,1	76,1	86,1	96,1
8.	В.Саар	3,4	15,3	27,9	35,7	58,9	71,1	81,1	91,2	100,0
9.	Х.Серго	3,1	10,7	20,3	27,2	50,6	62,1	72,2	82,3	92,4
10.	Р.Сирге	1,6	10,1	20,2	27,1	47,2	58,5	68,5	78,6	88,6
11.	М.Траат	3,3	15,4	27,0	33,9	55,5	66,8	76,7	86,6	96,5
12.	Э.Ветемаа	2,6	14,3	27,2	35,2	57,6	69,0	79,0	89,0	99,0
13.	А.Каал	3,7	14,1	27,2	34,8	56,5	68,1	78,1	88,2	98,3
14.	Т.Каллас	3,4	13,8	25,4	32,3	55,2	66,6	76,6	86,6	96,6
15.	Ю.Пээгель	2,5	13,1	25,4	32,1	54,5	66,1	76,2	86,4	96,6
16.	Ю.Туулик	2,9	13,1	25,9	33,4	57,1	70,1	80,1	90,2	100,0
17.	А.Валтон	3,6	16,7	30,2	38,0	61,1	72,5	82,5	92,5	100,0
18.	М.Унт	3,7	13,7	27,4	35,1	58,2	70,2	80,2	90,3	100,0
19.	Э.Нийт/Я.Кросс	3,5	11,5	21,4	27,6	48,9	59,9	69,7	79,4	89,2
20.	Ю.Смуул	3,9	11,9	22,6	28,8	50,0	61,4	71,4	81,4	91,4

Таблица 2

Опыт № 2. Исходные данные: лексический спектр (в словаре) - доля словоформ (%) с данной частотой F

№ текста	Частота словоформы											
	F=1	2	3	4	5	6	7	8	9	10	11-20	>20
1.	79,12	10,32	4,11	1,95	0,98	0,80	0,66	0,21	0,38	0,21	0,77	0,49
2.	78,70	10,65	4,22	1,93	0,96	0,64	0,43	0,39	0,25	0,22	1,04	0,57
3.	73,29	13,60	4,22	3,21	1,38	0,89	0,81	0,37	0,61	0,04	0,65	0,93
4.	78,74	10,45	3,94	2,01	1,24	0,91	0,37	0,33	0,29	0,18	0,99	0,55
5.	81,65	9,40	2,83	2,17	0,94	0,66	0,28	0,25	0,28	0,11	0,91	0,52
6.	76,00	12,16	4,38	1,83	1,04	0,80	0,88	0,44	0,32	0,36	1,03	0,76
7.	78,28	10,97	4,30	1,74	1,08	0,63	0,30	0,59	0,41	0,22	0,96	0,52
8.	75,20	12,46	4,39	2,05	1,23	0,94	0,53	0,53	0,33	0,33	1,31	0,70
9.	80,18	10,05	3,72	1,32	1,11	0,80	0,73	0,52	0,25	0,17	0,63	0,52
10.	81,74	10,08	3,36	1,47	0,65	0,52	0,39	0,23	0,13	0,29	0,72	0,42
11.	78,61	10,85	3,96	1,91	0,82	0,71	0,75	0,41	0,22	0,19	0,97	0,60
12.	77,70	11,25	3,96	1,53	1,10	0,74	0,47	0,59	0,59	0,19	1,25	0,63
13.	77,80	11,29	4,60	1,24	1,08	0,74	0,50	0,23	0,27	0,43	1,12	0,70
14.	78,57	9,89	4,19	2,21	1,31	0,94	0,49	0,34	0,30	0,22	0,94	0,60
15.	78,64	10,49	3,90	2,17	1,27	0,90	0,38	0,22	0,19	0,19	0,94	0,71
16.	74,07	12,81	5,18	2,09	1,25	0,96	0,80	0,52	0,32	0,12	1,08	0,80
17.	75,73	11,38	5,10	1,98	1,14	0,84	0,59	0,42	0,46	0,34	1,22	0,80
18.	76,12	11,44	4,43	2,42	1,41	0,64	0,56	0,32	0,48	0,24	1,13	0,81
19.	81,56	9,01	3,54	1,80	1,08	0,62	0,59	0,46	0,16	0,07	0,72	0,39
20.	80,60	10,33	3,14	1,77	1,19	0,68	0,34	0,38	0,14	0,24	0,68	0,51

Таблица 3

Опыт № 3. Исходные данные: динамика роста словаря
(число разных словоформ при различных объемах текста N)

№ текста	Автор	Объем текста N = 1000	2000	3000	4000	5000
1.	Э.Бэкман	731	1383	1865	2404	2869
2.	В.Гросс	677	1315	1859	2358	2791
3.	А.Хинт	649	1116	1597	2034	2463
4.	Х.Кийк	710	1351	1828	2315	2738
5.	Я.Кросс	723	1315	1914	2382	2851
6.	П.Куусберг	645	1166	1674	2075	2506
7.	Л.Промет	674	1212	1720	2207	2694
8.	В.Саар	633	1128	1700	2045	2439
9.	Х.Серго	734	1326	1885	2416	2876
10.	Р.Сирге	764	1397	2017	2572	3067
11.	М.Траат	689	1235	1722	2226	2696
12.	Э.Ветемаа	680	1208	1734	2162	2552
13.	А.Каал	651	1204	1690	2119	2586
14.	Т.Каллас	663	1226	1733	2179	2668
15.	Ю.Пээгель	690	1224	1700	2223	2669
16.	Ю.Туулик	624	1135	1560	2005	2491
17.	А.Валтон	588	1036	1468	1955	2373
18.	М.Унт	658	1176	1678	2133	2483
19.	Э.Нийт/Я.Кросс	740	1357	1923	2473	2983
20.	Ю.Смуул	732	1361	1917	2473	2929

ке в качестве квантитативной меры "богатства" словаря данного текста.

Следует отметить, что все выше описанные наборы характеристик (т.е. данные о покрываемости текста, о лексическом спектре и об объеме словаря) рассматриваются обычно как тесно связанные между собой квантитативные показатели статистической структуры текста. Иногда говорят даже о наличии "жесткой" связи между этими показателями (при некоторых идеальных условиях статистической организации текста, например, при условии точного выполнения закона Ципфа). Вопрос состоит в том, покажет ли наш эксперимент близкие результаты в трех разных опытах классификации реальных текстов, учитывая сказанное о взаимосвязи используемых наборов характеристик.

Матрица близости. Математической основой для классификации объектов с помощью кластер-анализа является вычисление функций на парах объектов, исходя из численных значений признаков. В результате получаются матрицы близости (матрицы сходства или различия) между объектами. В таких матрицах представлено множество из n подлежащих кластеризации объектов, для которых исходные (первичные) данные измерений упрощены до набора из $n(n-1)/2$ значений близости между объектами по всем парам объектов.

Задачи кластер-анализа можно решать в терминах матрицы сходства или в терминах матрицы различия. Матрицы сходства обычно конструируются на основе коэффициентов подобия или коэффициентов связи (корреляции). Матрицы различия конструируются на основе показателей "расстояния" (обзор различных мер близости см. Елисеева И.И., Рукавишников В.О., 1977, с. 31 и след.). Выбор метрики для измерения расстояния определяется природой исходных признаков и целью классификации.

В данном исследовании мерой близости было выбрано обычное евклидово расстояние, исходя из следующих содержательных соображений: при данных наборах признаков и при равных объемах текстов все значения признаков (т.е. отдельные компоненты вектора) можно считать равноправными, и численные различия между отдельными значениями признаков сравниваемых текстов можно считать существенными для определения расстояния между текстами. Однако для того, чтобы избежать слишком большого веса больших численных значений отдельных признаков по сравнению с малыми значениями, необходимо выравнивать диапазоны изменения значений признаков с помощью н о р м а л и з а ц и и исходных данных (обычным способом, т.е. вычитанием

среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице, см. Дюран Б., Оделл П., 1977, с. 40). Евклидово расстояние (d) определяется формулой:

$$d(X_s, X_t) = \left[\sum_{j=1}^{k}(x_{js} - x_{jt})^2\right]^{0,5},$$

где x_{js} и x_{jt} - нормированные значения признаков, k - число измерений. Значение $d(X_s, X_t)$ для заданных векторов X_s и X_t считается эквивалентным расстоянию между самими объектами (текстами) T_s и T_t соответственно выбранному набору признаков $(C_1, C_2, \ldots, C_k)'$. Предполагается, что близость между текстами свидетельствует о близости стилей авторов в отношении некоторых скрытых для прямого наблюдения индивидуальных особенностей, выражающихся в устойчивых квантитативных (лингвостатистических) характеристиках текста.

Результаты измерения близости между объектами представляются в компактной форме в матрицах сходства или различия. В качестве примера рассмотрим матрицу различия по данным опыта № I (см. табл. 4). Из таблицы видно, что наиболее близкими текстами в отношении распределения численных данных о покрываемости текста являются тексты № 6 (П. Куусберг) и № I2 (Э. Ветемаа). Измерение евклидова расстояния между рассматриваемыми признаками этих текстов дает результат $d(X_6, X_{I2}) = 0{,}4843$. Действительно, сравнение соответствующих векторов измерений (см. табл. I) показывает большое сходство в распределении значений признаков:

X_6 : 2,4 - I4,3 - 27,0 - 34,7 - 57,6 - 69,7 - 79,7 - 89,3 - 99,8;

X_{I2}: 2,6 - I4,3 - 27,2 - 35,2 - 57,6 - 69,0 - 79,0 - 89,0 - 99,0.

С другой стороны, наиболее отдаленными друг от друга оказываются тексты № I0 (Р. Сирге) и № I7 (А. Валтон): $d(X_{I0}, X_{I7}) = I0{,}2680$. Распределение данных о покрываемости в этих текстах имеет следующий вид:

X_{I0} : I,6 - I0,I - 20,2 - 27,I - 47,2 - 58,5 - 68,5 - 78,6 - 88,6;

X_{I7} : 3,6 - I6,7 - 30,2 - 38,0 - 6I,I - 72,5 - 82, 5 - 92,5 - I00,0.

Информацию, содержащуюся в матрице различия, можно использовать для т. наз. многомерного шкалирования с помощью графа близости (Крускал Дж., I980, с. 28) или для проведе-

Таблица 4

Матрица различия (опыт № 1)

2	1.4611											
3	5.3617	3.9964										
4	2.5769	1.4538	3.2575									
5	4.3725	4.0719	5.5683	3.0506								
6	5.3637	3.9391	0.7856	3.3244	5.7253							
7	3.6244	2.3276	2.5682	1.2021	3.3984	2.5159						
8	6.3672	4.9824	1.4508	3.9755	5.6336	1.5964	3.0214					
9	8.7676	2.0062	5.8272	2.9402	4.2927	5.8930	4.0485	6.7973				
10	2.7255	3.8704	7.8028	5.1366	6.4167	7.6636	6.0375	8.8212	2.7971			
11	4.5347	3.1826	2.3168	2.1642	3.9322	2.1159	0.9883	2.3976	4.9957	6.8272		
12	5.2008	3.7588	1.0047	3.0634	5.3671	0.4043	2.2032	1.5588	5.7277	7.5047	1.7507	
13	4.9022	3.5617	1.6995	2.4035	4.0986	1.8104	1.5535	1.7477	5.3895	7.3821	1.2646	1.4529
14	3.5801	2.2673	2.2881	1.1116	3.5399	2.3343	0.5034	2.0904	3.9892	6.0862	1.2342	2.0517
15	3.1364	1.6916	2.4310	1.4587	4.3989	2.2812	1.3690	3.4102	3.7185	5.4495	1.8997	2.1047
16	4.9212	3.5952	0.5932	2.8813	5.2925	1.1242	2.3689	1.8867	5.3642	7.4077	2.3423	1.2319
17	7.8868	6.4856	2.9912	5.4478	6.6846	2.9905	4.4170	1.6369	8.3249	10.2686	3.6163	2.0520
18	5.6760	4.3496	1.2499	3.2873	4.9879	1.6678	2.5788	1.1345	6.0637	8.2148	2.2572	1.5259
19	1.9238	2.9767	6.7757	3.5936	3.9543	6.7644	4.5764	7.5386	1.6603	2.8112	5.4342	6.5156
20	1.7488	2.2807	5.7308	2.4899	2.8876	5.7746	3.4960	6.4225	1.6263	3.5746	4.3825	5.5074
№ текста	1	2	3	4	5	6	7	8	9	10	11	12

14	1.4096						
15	2.2806	1.2075					
16	1.6427	2.0152	2.1287				
17	3.1376	4.3666	4.9106	3.4742			
18	1.1355	2.3116	2.9122	1.3307	2.5723		
19	5.9388	4.6636	4.9383	6.3750	8.9358	6.8645	
20	4.7869	3.5676	3.6385	5.3288	7.8184	5.7109	1.2719
№ текста	13	14	15	16	17	18	19

ния кластер-анализа, предусматривающего дискретное комбинаторное представление, чаще всего в виде дерева (дендрограммы).

Конструирование кластер-системы. При конструировании кластер-системы, или кластеризации, исходят из данных о близости между объектами, то есть, образно говоря, в алгоритмах кластеризации матрицу близости берут в качестве входа, а разбиение на кластеры является выходом (Крускал Дж., 1980, с. 22). Методы кластеризации можно разделить на иерархические и неиерархические (обзор наиболее известных разновидностей кластер-анализа см. Айвазян С.А. и др., 1974, с. 99 и след.). Иерархические процедуры кластеризации бывают двух типов - агломеративные и дивизимные (разделительные). Принцип работы агломеративных алгоритмов состоит в последовательном объединении в кластер сначала самых близких, а затем и все более отдаленных друг от друга объектов. В разделительных иерархических процедурах, наоборот, множество объектов последовательно разбивается на группы. В данном исследовании была выбрана разновидность агломеративного иерархического метода кластеризации. Для практического решения вопроса на ЭВМ был использован метод B_k, представляющий собой усовершенствованный вариант т. наз. Кэмбриджского алгоритма (подробнее см. Ээремаа Р., 1978а, с. 61 и след.). При использовании метода B_k можно обобщенно говорить о k- кластеризации, характеризуя параметром k допустимую покрываемость кластеров до k элементов. Если имеется n объектов, подлежащих кластеризации, то параметр k может принимать целочисленные значения из отрезка $[1, n-2]$. Отметим, что при $k = 1$, т.е. при 1-кластеризации (совпадающей в данном случае с методом "одной связи", или "ближайшего соседа") получаются непересекающиеся кластеры, и их можно представить в виде дендрограммы (диаграммы-дерева). При $k > 1$ это уже невозможно. В таком случае результат анализа на ЭВМ печатается как разбиение на уровне, определенном вперед или автоматически на ЭВМ. В данной работе используются два подхода: 1-кластеризация и 2-кластеризация (разбиение при $k = 1$ и $k = 2$; последний подход используется в качестве вспомогательного). Важным фактором при проведении анализа является уровень классификации, обозначаемый символом h (подробнее см. Ээремаа Р., 1978 а, с. 57 и след.).

Как уже было сказано, при использовании агломеративного иерархического метода разбиение объектов на кластеры совер-

шается ступенчато. Процесс кластеризации начинается с того, что (на 1-м шагу) два наиболее близко расположенных объекта (в первом опыте тексты № 6 и № 12, см. табл. 4) объединяются и рассматриваются как один кластер. Это приводит к тому, что число объектов уменьшается и становится равным $n - 1$, причем один кластер будет содержать два объекта, а $n - 2$ остальных по одному. Процесс можно повторять до тех пор, пока все объекты не сгруппируются в один большой кластер. Результаты такого процесса обычно изображаются графически в виде диаграммы-дерева или дендрограммы, и с помощью отдельных таблиц с результатами кластеризации на каждом шагу (дендрограммы и таблицы выдаются в готовом виде на ЭВМ). Дендрограмма дает возможность наглядной интерпретации всего хода последовательной кластеризации (по данным наших опытов см. рис. 1 – 3). Весь процесс кластеризации в данных опытах заканчивается на 19-м шагу (при $n = 20$), где все объекты (тексты) объединяются в один кластер.

Критерий качества. Процесс последовательной кластеризации может сам по себе дать ценную информацию при анализе взаимоотношений между объектами данной совокупности. Но встает вопрос, где (на каком шагу, на каком уровне) достигается "оптимальное" решение кластер-анализа. Этот вопрос связан с т. наз. критерием качества, или критерием оптимальности кластеризации. Критерий качества определяется различным путем, причем отмечается, что "выбор того или иного критерия осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему" (Айвазян С.А. и др., 1974, с. 85). С формальной точки зрения оптимальное разбиение определяется требованием наибольшей однородности внутри кластеров и возможно большего различия между кластерами. Для этого существуют особые количественные оценки ("функционалы качества разбиения"). Но в практической работе, в зависимости от конкретного материала и целей исследования, критериями качества могут служить, например, возможность содержательной интерпретации найденных кластеров или согласованность полученной классификации с теоретическими представлениями (см. Елисеева И.И., Рукавишников В.О., 1977, с. 11). Вопрос о критерии качества тесно связан и с выбором необходимого числа кластеров, которое определяется либо априорно (в зависимости от конкретных условий эксперимента), либо в процессе разбиения множества объектов на кластеры. В данном

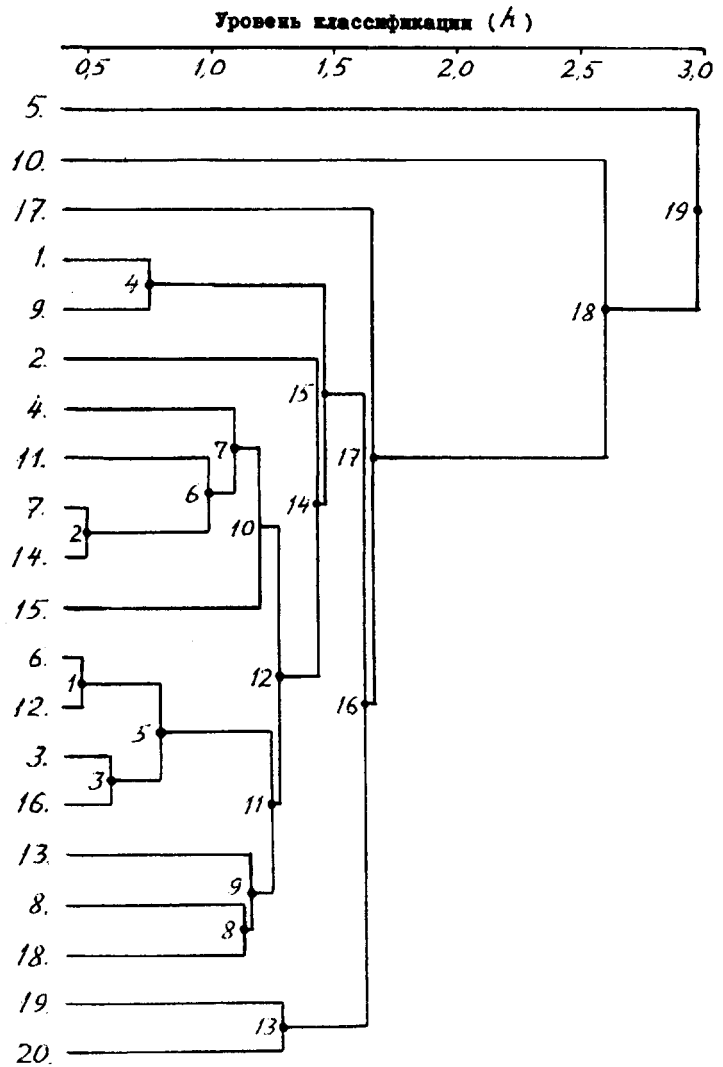


Рис. 1. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения показателей покрываемости текста словоформами (опыт № 1). Цифры слева - номера текстов. Цифры в схеме - номера шагов объединения текстов в группы.

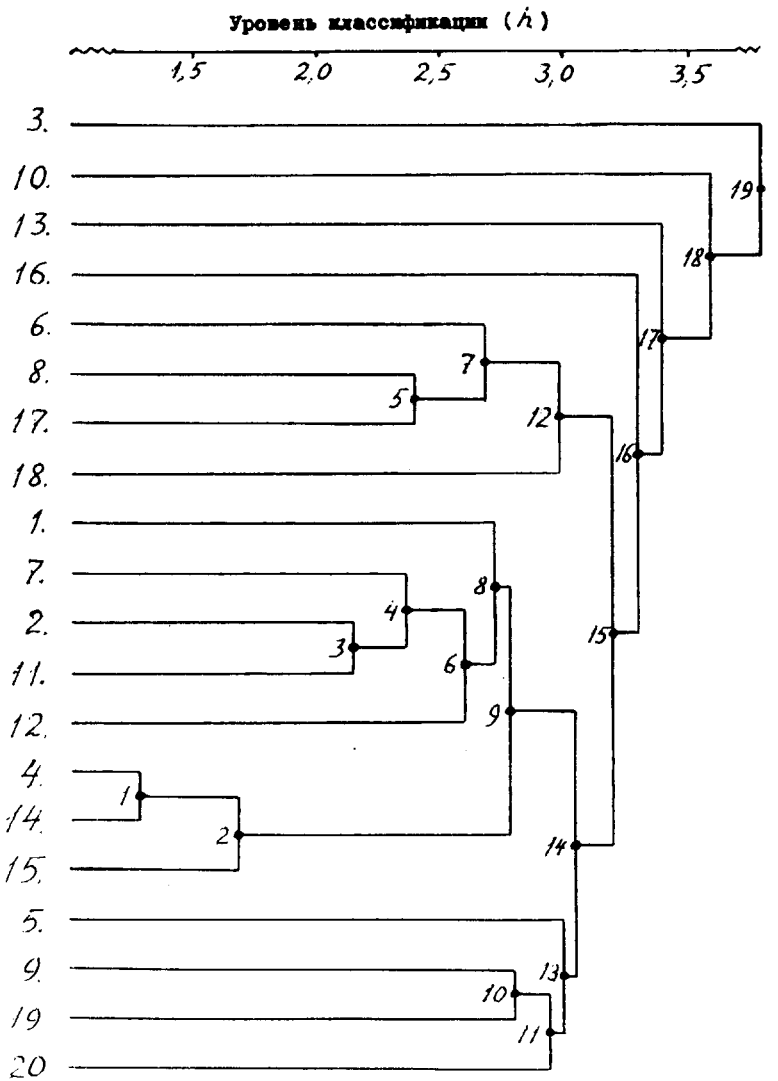


Рис. 2. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения лексических спектров на уровне словаря (опыт № 2).

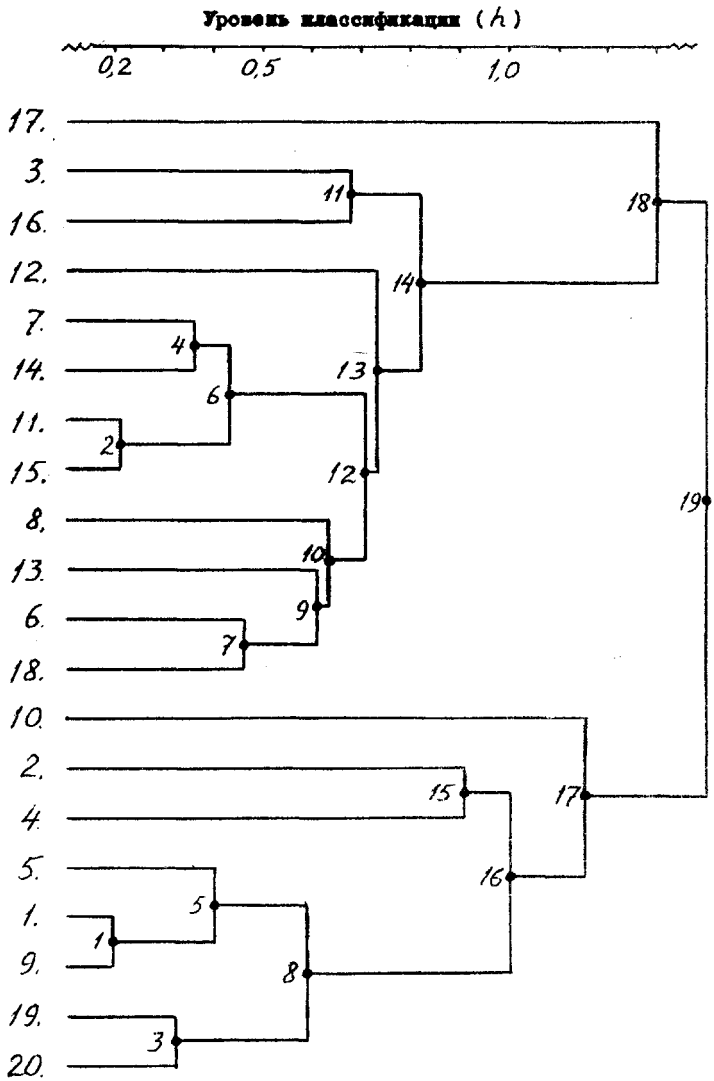


Рис. 3. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения динамики роста словаря в связи с увеличением текста (опыт № 3).

исследовании, где одной из основных целей кластеризации является проверка на сходство результатов разных опытов, желаемый уровень разбиения определялся эмпирически в процессе последовательной I-кластеризации в сравнении с результатами метода 2-кластеризации и при одновременном сопоставлении результатов разных опытов. В то же время учитываются общие требования об относительной однородности внутри кластеров (по особым мерам единичной и средней "стабильности" кластеров на отдельных уровнях разбиения; такие оценки выдаются на ЭВМ).

Первый опыт. Рассмотрим процесс I-кластеризации подробнее на основе данных первого опыта (см. рис. I и табл. 5).

Таблица 5

Процесс последовательной кластеризации текстов по данным первого опыта

Шаг (№)	Уровень (h)	Объединяются
I	0,484	(6.)(I2.)
2	0,503	(7.)(I4.)
3	0,593	(3.)(I6.)
4	0,768	(I.)(9.)
5	0,789	(6.I2.)(3.I6.)
6	0,988	(7.I4.)(II.)
7	I,II2	(7.I4.II.)(4.)
8	I,I35	(8.)(I8.)
9	I,I36	(8.I8.)(I3.)
I0	I,208	(7.I4.II.4.)(I5.)
II	I,250	(6.I2.3.I6.) (8.I8.I3.)
...		
I9	2,888	(5.) (все остальные)

Из таблицы видно, что на достаточно раннем уровне (на 5-м шагу) образуется кластер (6.I2.3.I6), который остается неизменным (устойчивым) до II-го шага. На 7-м шагу образуется кластер (7.I4.II.4.), к которому на I0-м шагу присоединяется изолированный до тех пор текст I5. На 9-м шагу образуется кластер (8.I8.I3.). На основе оценок стабильности можно констатировать довольно высокую однородность этих кластеров. Поэтому остановимся для пробы на I0-м шагу (уровень классификации h = I,208). Общий результат - I0 кластеров (см. рис. I), при этом кластерами считаются и изолированные объекты (I-элементные кластеры):

(7.4.11.14.15.)
(6.12.3.16.)
(13.8.18.)
(1.9.)
(19.) (20.) (5.) (10.) (17.) (2.)

Этот результат можно сравнить с решением на 11-м шагу (при уровне классификации h = 1,250):

(7.4.11.14.15.)
(6.12.3.16.13.8.18.)
(1.9.)
(19.) (20.) (5.) (10.) (17.) (2.)

Здесь объединяются два кластера: (6.12.3.16) и (8.18. 13.). Следует проверить, не сказывается ли в данном случае т. наз. эффект сцепления (характерный для метода "одной связи"), который состоит в том, что "единственное непредставительное значение близости может вызвать на раннем уровне объединение двух несхожих кластеров" (Матула Д.В., 1980,с. 92). При объединении названных двух кластеров близкими оказываются тексты 3. и 18. (расстояние d = 1,2499, см. табл. 4), но, например, тексты 6. и 13. довольно отдаленны друг от друга (d= 1,8104). Для дополнительной проверки было решено использовать параллельный метод 2-кластеризации, где допускается пересечение двух кластеров по одному элементу (см. Ээремаа Р., 1978а и 1978б). Представим результат, полученный с помощью ЭВМ. На соответствующем уровне (h = 1,451) образуются 9 кластеров, причем один из них (13.18.) имеет по одному общему элементу с двумя другими кластерами:

(7.4.11.14.15.13.)
(6.12.3.16.8.18.)
(1.9)
(19.20.)
(13.18.)
(5.) (10.) (17.) (2.)

Таким образом, связь между кластерами оказывается более сложной, чем на 11-м шагу 1-кластеризации. Текст 13. в составе кластера (13.8.18.) тяготеет к кластеру (7.4.11.14. 15.), а не к кластеру (6.12.3.16.), хотя на более высоком уровне кластеризации текст 13. сохраняет связь с этим кластером через текст 18., с которым он образует отдельный кластер с пересечением.

С помощью подобных проверок удается определить тот оптимальный (по условиям нашего эксперимента - максимальный) уровень, на котором целесообразно остановиться при 1-кластеризации, т.е. при применении процедуры, дающей непересекающиеся кластеры. В данном конкретном случае таким уровнем оказался h = 1,208 (10-й шаг при последовательной кластеризации).

Сравнение классификаций. Напомним, что в трех опытах, проведенных на одном и том же материале 20 текстов, были использованы такие наборы признаков (покрываемость текста, лексический спектр, динамика роста словаря), которые обычно считаются взаимосвязанными и близкими показателями статистической структуры текста. Следовательно, можно было ожидать и близких результатов кластер-анализа по данным трех опытов.* Если сравнивать соответствующие дендрограммы последовательной кластеризации (см. рис. 1 - 3), то на первый взгляд удается обнаружить сходство только в отдельных точках: например, в первом и третьем опытах тексты 1. и 9., а также тексты 7. и 14. объединяются в один кластер на ранней стадии кластеризации (на 4-м и 1-м шагах и на 2-м и 4-м шагах соответственно). Но в общем приходится констатировать, что структуры дендрограмм мало похожи друг на друга. Поэтому постараемся сравнить такие этапы (стадии) кластеризации, которые на основе ранее описанных критериев могут считаться "оптимальными". На этом основании можно представить следующие наборы кластеров по данным трех опытов:

	Опыт № 1	Опыт № 2	Опыт № 3
"А":	(4.7.11.14.15.)	(1.2.4.7.11.12.14.15.)	(7.11.14.15.)
"Б":	(8.13.18.)	(6.8.17.18.)	(6.8.13.18.)
"В":	(1.9.)	(5.9.19.20.)	(1.5.9.19.20.)
"Г":	(3.6.12.16.)	-	-

Изолированные тексты (1-элементные кластеры):

(2.)(5.)(10.)	(3.)(10.)(13.)(16.)	(2.)(3.)(4.)
(17.)(19.)(20.)		(10.)(12.)(16.)
		(17.)

При сравнительном анализе обнаруживается, что существуют отдельные общие моменты как в образовании, так и в составах

* Как известно, в настоящее время отсутствуют точные методы оценки близости результатов классификаций (особенно иерархического типа). Пока можно воспользоваться лишь некоторыми приближенными методами оценки (см., например, Балашов Л.А. и др., 1973).

кластеров.

Кластер, условно названный кластером "А", выступает в близких вариантах во всех трех опытах. Устойчивым ядром кластера "А" являются 7., 11., 14., 15. (авторы: Л. Промет, М. Траат, Т. Каллас, Ю. Пээгель), к ним примыкает текст 4., который встречается в кластере "А" в первом и втором опытах (автор: Х. Кийк).

В кластере "Б" общими для всех опытов являются тексты 8. и 18. (В. Саар, М. Унт), к ним примыкают тексты 6. и 13. (П. Куусберг, А. Каал).

В кластере "В" по данным второго и третьего опытов встречаются тексты 5., 9., 19., 20. (Я. Кросс, Х. Серго, Э.Нийт/Я.Кросс, Ю. Смуул); в первом опыте тексты 5., 19. и 20. остаются на выбранном уровне классификации изолированными, т. е. составляют 1-элементные кластеры, но можно констатировать, что тексты 19. и 20. через несколько шагов объединяются в один кластер (см. рис. 1). В то же время текст 5. (Я.Кросс) остается изолированным до последнего шага кластеризации. Текст 5. отличается от всех других текстов тем, что в первом опыте (т.е. по данным покрываемости текста словоформами) он имеет чрезвычайно высокое численное значение самого частого слова (5,7 % при среднем значении 3,2 %; см. табл. 1); этот показатель нейтрализуется в лексическом спектре и динамике роста словаря, судя по тому, что во втором и третьем опытах текст 5. присоединяется к кластеру "В".

Кластер "Г" устанавливается только в первом опыте, в него входят тексты 3., 6., 12., 16. (А. Хинт, П. Куусберг, Э. Ветемаа, Ю. Туулик). Но этот кластер образуется на раннем уровне, на 5-м шагу, и остается неизменным до 11-го шага, что свидетельствует в большой устойчивости кластера.

Кроме многоэлементных кластеров представляют интерес и 1-элементные кластеры (на выбранном уровне классификации). Во всех трех опытах неизменно изолированным остается текст 10. (Р. Сирге). Тенденцию к изоляции обнаруживают также тексты 2. (В. Гросс), 3. (А. Хинт) и 17. (А. Валтон), которые составляют 1-элементные кластеры в двух случаях из трех.

Выводы. С помощью параллельных опытов кластеризации 20 текстов на основе разных наборов формальных характеристик т. наз. статистической структуры текста (покрываемость текста словоформами, лексический спектр, динамика роста словаря) удалось выделить некоторые достаточно устойчивые непересекающиеся кластеры, которые в определенной степени представляют

характерные для данного языка (или подъязыка) квантитативно-лингвистические т и п ы текстов. Однако примененный метод не позволяет охватывать типизацией все тексты: в среднем 30 % текстов не попадают в устойчивые много-элементные или 1-элеметные кластеры. Отчасти это может быть объяснено недостатками примененного метода, в частности, "эффектом сцепления" кластеров, для преодоления которого приходится останавливать процесс кластеризации на довольно раннем уровне. Некоторую роль может играть и то обстоятельство, что по примененной программе часть информации матриц близости теряется (т.к. не учитываются коэффициенты различия пар объектов, наиболее отдаленных друг от друга). Но основной причиной неполного разбиения текстов на непересекающиеся кластеры следует все же считать то, что в принципе "большинство реальных классов размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен" (Заде Л.А., 1980, с. 208). Таким образом, будет целесообразно основывать алгоритмы кластер-анализа на представлении о кластере (классе, типе) как о размытом, нечетком множестве. В настоящем эксперименте был использован один из подобных алгоритмов, но только в качестве вспомогательного метода (для проверки однородности кластеров и для определения оптимального уровня классификации). В данном случае целью исследования было сравнение результатов классификации по обычному способу, т.е. по способу разбиения объектов на непересекающиеся кластеры.

При сравнительном анализе результатов трех параллельных опытов можно было констатировать значительное различие в иерархических структурах кластер-систем (см. соответствующие дендрограммы). Это различие обусловлено в большой степени тем, что наборы признаков, считающиеся близкими и тесно взаимосвязанными, в действительности не обнаруживают такого соответствия, которое необходимо для более точных расчетов. В реальных текстах нет жесткой связи между разными характеристиками статистической структуры текста. Из этого следует, что кластер-анализ на основе одного какого-нибудь набора признаков, характеризующих статистическую структуру текста, не предопределяет результаты анализа на основе другого набора аналогичных (родственных, близких) признаков, хотя некоторое (не предсказуемое) сходство между результатами анализов имеется. При этом только сходные или совпадающие результаты параллельных опытов можно считать достаточно достоверными.

СПИСОК ТЕКСТОВ

1. Э. Бээкман : A. Beekman, Kartulikuljused. Tln., 1968.
2. В. Гросс : V. Gross, Pinginaabrid. Tln., 1965.
3. А. Хинт : A. Hint, Tuuline rand IV. Tln., 1966.
4. Х. Кийк : H. Kiik, Tondiöömaja. Tln., 1970.
5. Я. Кросс : J. Kross, Kolme katku vahel I. Tln., 1970.
6. П. Куусберг : P. Kuusberg, Südasuvel. Tln., 1966.
7. Л. Промет : L. Promet, Primavera. Tln., 1971.
8. В. Саар : V. Saar, Ukuaru. Tln., 1969.
9. Х. Серго : H. Sergo, Põgenike laev. Tln., 1966.
10. Р. Сирге : R. Sirge, Kolmekesi lauas. Tln., 1970.
11. М. Траат : M. Traat, Tants aurukatla ümber. Tln., 1971.
12. Э. Ветемаа : E. Vetemaa, Väike romaaniraamat. Tln.,1968.
13. А. Каал : A. Kaal, Saaremaa laastud II. Tln., 1970.
14. Т. Каллас : T. Kallas, Puiesteede kummaline valgus.Tln., 1968.
15. Ю. Пээгель : J. Peegel, Lühikesed lood. Tln., 1970.
16. Ю. Туулик : J. Tuulik, Vana loss, Abruka lood.Tln.,1972.
17. А. Валтон : A. Valton, Luikede soo. Karussell.Tln.,1971.
18. М. Унт : M. Unt, Kuu nagu kustuv päike. Tln., 1968.
19. Э. Нийт, Я. Кросс : E. Niit, J. Kross, Muld ja marmor. Tln., 1968.
20. Ю. Смуул : J. Smuul, Jaapani meri, detsember. Tln.,1969.

ЛИТЕРАТУРА

Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. М.: Статистика, 1974.

Балашов Л.А., Гуськов А.А., Махотенко Ю.А., Смолянов О.Г. Некоторые критерии оценки классификационных систем. – Научно-техническая информация. Серия 2. М., 1973, № 7, с. 3-7.

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

Боярский А.Я. О методологических принципах и многомерном анализе. Предисл. в кн.: Дюран Б., Оделл П., Кластерный анализ. М.: Статистика, 1977, с. 5-12.

Дюран Б., Оделл П. Кластерный анализ. Перев. с англ. М.: Статистика, 1977.

Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. М.: Статистика, 1977.

Заде Л.А. Размытые множества и их применение в разпознавании образов и кластер-анализе. - В кн.: Классификация и кластер. М.: Мир, 1980, с. 208-247.

Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. Перев. с англ. - В кн.: Классификация и кластер. М.: Мир, 1980, с. 20-41.

Матула Д.В. Методы теории графов в алгоритмах кластер-анализа. Перев. с англ. - В кн.: Классификация и кластер. М.: Мир, 1980, с. 83-111.

Ээремаа Р. Общая теория конструирования кластер-систем и алгоритмы для нахождения их численных представлений. - В кн.: Статистическая обработка данных. Труды ВЦ. Вып. 42. Тарту, 1978 (а), с. 53-77.

Ээремаа Р. Алгоритм опознания кластеров к-дендрограммы. - В кн.: Статистическая обработка данных. Труды ВЦ. Вып. 42. Тарту, 1978 (б), с. 78-93.

AN ATTEMPT OF CLASSIFYING TEXTS WITH THE HELP OF CLUSTER-ANALYSIS

Juhan Tuldava

S u m m a r y

The main principles of cluster-analysis are examined and three parallel experiments of classifying 20 Estonian literary texts with the help of B_k-method are described. In these experiments three various sets of quantitative-linguistic characteristics of texts were used (accumulated relative frequencies of words in the frequency vocabularies of the texts, the distribution of word frequencies, the dynamics of vocabulary growth; see Tables 1 - 3). The results of the hierarchical agglomerative cluster-analysis are presented in the form of dendrograms (Fig. 1 - 3) and thoroughly analyzed.

ОБ ОДНОЙ ВОЗМОЖНОСТИ ПРОВЕДЕНИЯ КЛАСТЕР-АНАЛИЗА

Р.В. Ээремаа

При проведении кластер-анализа множества из n подлежащих кластеризации объектов можно выделить два этапа: первый - конструирование n(n-I)/2 - элементной матрицы близости, т.е. либо матрицы различия, либо матрицы сходства над всеми парами объектов; второй - конструирование кластер-системы, т.е. конструирование расслоения подмножеств (кластеров) множества объектов по уровням, соответствующим различным значениям близости.

В рамках системы статистической обработки данных STP - SSP - TRY Вычислительного центра Тартуского государственного университета реализовано несколько алгоритмов для проведения кластер-анализа. В данной статье опишем одну возможность кластеризации объектов, применяемую в НЦ ТГУ (см. Ээремаа Р., I978а, I978б), основой которой являются идеи B_k-метода k-кластеризации Джардайна и Сибсона (Jardine N., Sibson R., I967, I968а, I968в, I97I). При этом параметром k характеризуется допустимая покрываемость кластеров до k элементов. Если имеется n объектов, подлежащих кластеризации, то параметр k может принимать целочисленные значения из отрезка [I, n-2].

Для конкретности исходной матрицей мы возьмём матрицу различия, и в дальнейшем о ней будем говорить как об исходном коэффициенте различия (КР) d над всеми парами объектов. Конструирование кластеров B_k-методом на некотором уровне h можно проиллюстрировать следующим образом. Вырисовывается граф, вершины которого представляют исследуемые объекты , а ребрами соединяются объекты с различием не выше h . Такие объекты, которые соединены ребром, можно называть связанными между собой объектами на уровне h, также можно говорить о связи между этими объектами.

Рассмотрим одну конкретную задачу кластеризации 20 объектов. Эти объекты соответствуют данным, приведенным в опыте № I в статье Ю. Тулдава в настоящем сборнике. В таблице I

представлены первые 22 элемента возрастающей последовательности межобъектных различий и также указаны соответствующие пары объектов. На рис. I и 3 даются граф-представления межобъектных связей соответственно на уровнях 0,786 и I,45I.

Кластерам соответствуют максимальные подмножества вершин, имеющих все возможные ребра. Если такие подмножества имеют по меньшей мере k общих вершин, то соединение двух таких множеств вершин дополняется до полного подмножества прибавлением отсутствующих ребер. При том возникают искаженные связи между объектами. Теперь проверяется снова - имеют ли какие-то два подмножества по меньшей мере k общих вершин и т.д. Процесс окончен, если невозможно дополнить ни одно подмножество вершин. Получаемые максимальные подмножества вершин являются кластерами на уровне h. Граф, который таким образом получается, имеет ребра, соединяющие те пары объектов, при которых $B_{\text{д}}(d) \leq h$. Такое конструирование графа возможно при всех значениях h, которые КР d имеет на множестве исследуемых объектов.

Для примера рассмотрим этап I-кластеризации (т.е. k=1) на уровне h = 0,786. На рис. I видно, что максимальными подмножествами вершин, которые имеют все возможные ребра, являются следующие: {6,I2}, {7,I4}, {3,I6}, {I,9} и {3,6}. Среди них подмножества {6,I2} и {3,6} имеют один общий элемент и их можно соединить в подмножество {3,6,I2}. Так как теперь подмножества {3,6,I2} и {3,I6} имеют один общий элемент, то они соединяются. Таким образом, при применении B_I-метода получаются на уровне 0,786 следующие неодноэлементные кластеры {3,6,I2,I6}, {7,I4} и {I,9}. При том искаженными являются связи между объектами (3,I2), (6,I6) и (I2, I6), так как соответствующие исходные различия I,005, I,I24 и I,232 заменяются значением 0,786.

Отметим, что при I-кластеризации (B_I-методом) получается кластер-система непересекающихся кластеров, которую можно представить в виде диаграммы-дерева.

Диаграмма-дерево (I-дендрограмма), получаемая в результате I-кластеризации нашего 20-элементного множества, представлена в статье Ю. Тухдава в настоящем сборнике.

Рассмотрим теперь этап 2-кластеризации тех же объектов на уровне h = I,45I. На рис. 3 исходные межобъектные связи (таких 22) представлены сплошными линиями. Полными подмножествами, которые имеют по меньшей мере 2 общих элемента, являются {3,6,I2,I6} и {3,I6,I8}. Их соединение {3,6,I2,I6, I8} имеет 2 общих элемента с множеством {3,8,I8}. Продолжая аналогичным образом, получаются на уровне I,45I следующие неодноэлементные кластеры: {3,6,8,I2,I6,I8}, {I3,I8}, {4,7, II,I3,I4,I5}, {I,9}и {I9,20}. При этом объекты I3 и I8 при-

Таблица I

№ пп.	Различие	Объекты
I.	0,484	(6,I2)
2.	0,503	(7,I4)
3.	0,593	(3,I6)
4.	0,768	(I,9)
5.	0,786	(3,6)
6.	0,988	(7,II)
7.	I,005	(3,I2)
8.	I,II2	(4,I4)
9.	I,I24	(6,I6)
I0.	I,I35	(8,I8)
II.	I,I36	(I3,I8)
I2.	I,202	(4,7)
I3.	I,208	(I4,I5)
I4.	I,232	(I2,I6)
I5.	I,234	(II,I4)
I6.	I,250	(3,I8)
I7.	I,265	(II,I3)
I8.	I,272	(I9,20)
I9.	I,33I	(I6,I8)
20.	I,370	(7,I5)
2I.	I,4I0	(I3,I4)
22.	I,45I	(3,8)

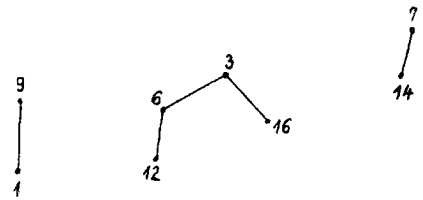


Рис. I. Граф-представление межобъектных связей на уровне h=0,786

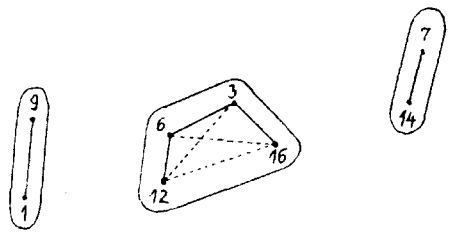


Рис. 2. Образование кластеров на уровне h= 0,786 при I-кластеризации

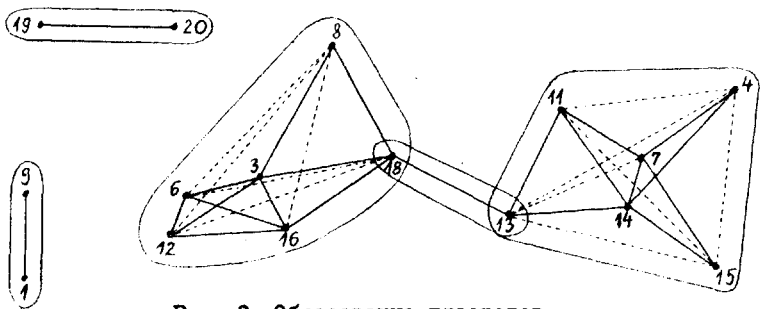


Рис. 3. Образование кластеров на уровне h= I,45I при 2-кластеризации

надлежат нескольким кластерам одновременно. На рис. 3 представлены искаженные межобъектные связи (таких II) пунктирными линиями.

Если $k > 1$, то трудно изобразить результат k-кластеризации деревом (как в случае 1-кластеризации), из-за существования покрываемых элементов в кластерах. Поэтому алгоритм k-кластеризации при $k>1$ работает в таком режиме, что сначала выпечатывается разбиение на одном уровне, определенном заранее или выбранном ЭВМ. Затем исследователь оценивает получаемое разбиение, исходя из своих профессионально-теоретических соображений и из численных показателей, характеризующих разбиение. Он либо удовлетворен полученным разбиением, либо желает представления разбиения на некотором другом уровне. Исследователь может задать то конкретное значение КР, которому соответствующий уровень его интересует, или определить уровень целочисленным значением требуемых межобъектных связей, которые надо учитывать при конструировании разбиения на уровне.

Если в нашем примере определить уровень требуемыми исходными связями 22, то выпечатается уровень 1,451.

При автоматическом определении уровня выбирается в ЭВМ наибольший уровень среди тех, которым соответствует относительно мало искаженных связей. Отметим, что уровнем g, которому соответствует относительно мало искаженных связей, считается тот, который удовлетворяет условиям.

$$O(g) \geqslant n \cdot k/2 \quad \text{и} \quad O'(g) \leqslant 3 \cdot O(g)/2,$$

где $O(g)$ означает число связей по исходным КР на уровне g, $O'(g)$ - число связей на уровне g после k-кластеризации.

При 2-кластеризации автоматически определяется в ЭВМ уровень $h = 1,451$, так как этот уровень наибольший среди тех, при которых удовлетворены вышеприведенные условия.

Отметим ещё, что разбиение всегда печатается вместе с разными численными характеристиками, характеризующими разбиения (в том числе, например коэффициенты стабильности и сплоченности каждого кластера).

Как уже было указано, при k-кластеризации возникает искажение исходного КР. Это искажение можно измерить. Применимая мера искажения Δ такова, что последовательность искажений $\Delta(d, B_k(d))$ монотонно убывает с увеличением k и достигает нуля при $k = n-1$. В реализованном алгоритме k-

кластеризации значение параметра k может быть задано заранее или же оно определяется автоматически в ЭВМ. Для определения автоматически значения параметра k, исследуют его те значения, при которых последовательность искажений $\Delta(d, B_2(d))$, $k = 1, 2, \ldots, n-2$, изменяется скачкообразно и делается выбор по определенным критериям.

Под конец отметим, что вышеизложенная реализация k-кластеризации позволяет кластеризировать при $k > 1$ до 360 объектов, при $k = 1$ число объектов неограничено.

ЛИТЕРАТУРА

Ээремаа Р. Общая теория конструирования кластер-систем и алгоритмы для нахождения их численных представлений. - В кн.: Статистическая обработка данных. Труды ВЦ. Вып. 42. Тарту, 1978(а), с. 53-77.

Ээремаа Р. Алгоритм опознания кластеров к-дендрограммы. - Труды ВЦ. Вып. 42. Тарту, 1978(б), с. 78-93.

Jardine C.J., Jardine N., Sibson R. The structure and construction of taxonomic hierarchies. - Math. Biosciences 1967, vol. 1, pp. 173-179.

Jardine N., Sibson R. A model for taxonomy. - Math. Biosciences, 1968(a), vol. 2, pp. 465-482.

Jardine N., Sibson R. The construction of hierarchic and non-hierarchic classifications. - The Computer Journal, 1968(b), vol. 11, pp. 177-184.

Jardine N., Sibson R. Mathematical Taxonomy. London: Wiley, 1971.

ON A POSSIBILITY OF THE REALIZATION OF CLUSTER-ANALYSIS

Ruth Ääremaa

S u m m a r y

In this article a graph-theoretic description of the cluster methods used in the paper of J. Tuldava is presented. These cluster methods use data in the form of dissimilarity (or similarity) coefficients on a set of objects in the process of the construction of a cluster system. The cluster system, or the k-dendrogram may be described as a hierarchy with numerical levels. Clusters - the sets of objects which are grouped at some level in the k-dendrogram - may overlap to the extent of k - 1 elements.

The clustering program has been written for the EC-1022 computer in the Computer Centre of the Tartu State University.

СОДЕРЖАНИЕ

SUMMARIES - RESÜMEES