ISSN 0207-4435

TARTU RIIKLIKU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

689

KVANTITATIIVLINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜS

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

Töid keelestatistika alalt

Труды по лингвостатистике

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 689 ВЫПУСК ОСНОВАНЫ В 1893.г.

KVANTITATIIVLINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜS

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

Töid keelestatistika alalt
Труды по лингвостатистике

TARTU 1984

Käesolevas kogumiku "Töid keelestatistika alalt" kümnendas väljaandes ("Kvantitatiivlingvistika ja tekstide automaatanalüüs") on avaldatud Tartu Riikliku Ülikooli rakenduslingvistika uurimisgrupi liikmete ja väliskaastöötajate artiklid.

В настоящем, десятом выпуске сборника "Труды по лингвостатистике" ("Квантитативная лингвистика и автоматический анализ текстов") опубликованы статьи сотрудников Исследовательской группы по прикладной лингвистике Тартуского государственного университета и исследователей из других городов.

This tenth issue of "Papers on Linguo-Statistics" ("Quantitative Linguistics and Automatic Text Analysis") consists of papers by members of the Research Group of Applied Linguistics at Tartu State University and guest authors.

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА. ОСНОВНЫЕ КОМПОНЕНТЫ[1]

В.М. Андрющенко

Задача создания машинного фонда русского языка пришла к нам извне и была определена "внешней постановкой"[2]: теория и практика программирования в последние годы овладели такими структурами данных и алгоритмами их обработки, которые по своей сложности сопоставимы с данными естественных языков. А практика использования ЭВМ во всех сферах жизни и труда выдвинула перед нами задачу овладеть общением с ЭВМ на естественном языке и научиться обрабатывать на машине документы в их естественной языковой форме. Когда первые такие системы вошли в эксплуатацию, стало понятно, насколько им не хватает мощности лингвистического обеспечения в виде средств для быстрой и надежной разработки и включения в систему словарей и грамматик, их пополнения и уточнения. В то же время стало понятно, что действительно глубокое, полное и всестороннее лингвистическое обеспечение систем обработки данных может быть создано лишь при условии автоматизации труда лингвистов и обеспечения их современными средствами поиска и обработки информации. В прикладных областях автоматические словари и грамматики по своей сложности не уступают традиционным, в то же время промышленность не может ждать десятилетиями, пока будут написаны соответствующие труды и подготовлены нужные данные - сроки разработки прикладных систем исчисляются месяцами. Так чисто качественный разрыв между академической и прикладной лингвистикой сегодня приобрел количественную форму и становится серьезным препятствием на пути массовой автоматизации производства и управления. Актуальность этой проблемы с каждым днем возрастает: теперь она может быть рассмотрена также в связи с обсуждением проектов ЭВМ 5-го поколения, которые должны обладать встроенными, т.е. аппаратно реа-

---

[1] В основе данной статьи лежит доклад автора "Концепция и архитектура машинного фонда русского языка", прочитанный им на конференции по проблемам создания машинного фонда данных для автоматизированной системы лексикографических исследований (Москва, 21-23 февраля 1983 г.).

[2] Впервые эта задача сформулирована А.П.Ершовым в докладе на Всесоюзной конференции Диалог 78" Моск. обл., 26-28 сентября 1978 г. (см.: Ершов, 1982).

лизованными языковыми процессорами и программно-аппаратными средствами обеспечения банков знаний.

Таким образом, накопленный опыт решения лингвистических задач на ЭВМ, научно-технические задачи сегодняшнего и в особенности завтрашнего дня позволяют поставить задачу создания системы комплексной автоматизации лингвистических исследований и разработок, состоящей из накопленных лингвистических данных, объективированных словарями, грамматиками и другими лингвистическими источниками, и программного обеспечения, предназначенного для использования этих данных и для конструирования новых лингвистических объектов - словарей, грамматик, языковых процессоров, которые в свою очередь могут войти в фонд в качестве единиц хранения и источников новых данных и средств для новых разработок. Такую развивающуюся систему лингвистических данных мы называем МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА.

Машинный фонд русского языка - это прежде всего система автоматизированных (т.е. программно-управляемых) картотек, содержащих текстовые, словарные, грамматические, программные и документальные источники данных о русском языке во всем объеме этого понятия. В понятие данных о русском языке мы включаем: данные о лексике и семантике в виде словарных статей системы словарей русского языка, сегментов текстов, содержащих употребление каждого учтенного слова, статистических данных об употребительности слов и различных комментариев к слову содержащихся в грамматиках и лингвистических трудах; данные о грамматике русского языка в виде тезауруса, адресующего категориальный состав грамматических форм и конструкций, и свода правил, характеризованных условиями применимости в определенном структурном или лексико-семантическом контексте; данные о фонетике и фонологии, о морфемике и морфонологии, включаемые в состав словарных статей и грамматической информации; диалектологические данные в виде словарей и диалектологических анкет; социо- и психолингвистические данные в виде источников, применяемых в этих дисциплинах, и т.д. Любой факт, квалифицируемый как явление русского языка, должен найти в конечном счете отражение в этих картотеках либо в виде объекта хранения, либо атрибута одного или нескольких объектов.

Программное обеспечение таких картотек должно содержать средства для выборки фактов и явлений, средства для статистической оценки их употребительности, степени их связанности и взаимообусловленности, средства поиска и группировки фактов, обладающих теми или иными заданными характеристиками. Программное обеспечение должно также включать средства ввода и редактирования данных, лингвистического анализа текстов, оперирования словарями как целыми источниками, средства литературного и технического редактирования, автоматической корректуры и автоматического набора.

Значения таких средств для лексикографии трудно переоценить: владение ими означает, что в машине каждый словарь может храниться в готовом для полиграфичес-

кого воспроизведения виде в качестве "исправленного и дополненного издания". Эти же средства нужны для подготовки лингвистических трудов. Их перевод в будущем в микрокомпьютерную форму создаст персональные средства автоматизации труда языковедов.

Со временем такой фонд языковых источников и фактов мог бы дополнить многочисленные и объемные словарные и справочные картотеки данных по русскому языку Академии наук и вузов, а также многочисленных коллективов отраслевых научно-исследовательских и информационных институтов и служб, связать их воедино, обогатиться программным обеспечением, разработанным для автоматического анализа и синтеза русского текста и общения с ЭВМ на естественном языке, и таким образом способствовать как повышению научного уровня инженерно-прикладных разработок, так и обогащению академических материалов о русском языке за счет включения в фонды Академии терминологических фондов отраслевых институтов, результатов и достижений вычислительной лингвистики.

С информационной точки зрения машинный фонд русского языка представляется как распределенный автоматический банк данных, управляемый несколькими согласованными по интерфейсу системами управления базами данных. Наиболее общими требованиями к таким СУБД являются: возможности многоаспектного поиска данных, иерархическая организация объектов хранения, возможности оперирования целыми базами, например, словарями, простота прикладного программирования, возможности телекоммуникационного доступа. Конкретные базы данных могут создаваться и существовать в различных организациях, пополняться за счет кооперативного сотрудничества с другими организациями и из центрального фонда. Пополнение, в частном случае, создание баз данных, должно быть возможным в двух аспектах: по номенклатуре хранимых объектов, например, словарных статей, и по их атрибутам, например, путем передачи из одной базы данных в другую определенных частей соответствующих словарных статей.

Базы данных, содержащие объекты одинаковых типов (например, словарные статьи, модули текстов) с содержательно совместимыми атрибутами и находящиеся под управлением одной логической СУБД, можно считать информационно однородными.

С точки зрения информационной однородности выделяются следующие максимальные компоненты машинного фонда русского языка: ГЕНЕРАЛЬНЫЙ СЛОВНИК, ТЕРМИНОЛОГИЧЕССКИЙ ФОНД, АКАДЕМИЧЕСКИЙ СЛОВАРНО-ГРАММАТИЧЕСКИЙ ФОНД, ИЛЛЮСТРАЦИОННО-ТЕКСТОВОЙ ФОНД, ЛИНГВОСТАТИСТИЧЕСКАЯ БД, ФОНД ЯЗЫКОВЫХ ПРОЦЕССОРОВ, ФОНД ЛИНГВИСТИЧЕСКИХ АЛГОРИТМОВ И ПРОГРАММ, ЛЕКСИКОГРАФИЧЕСКАЯ БАЗА ФОНДА И ИНФОРМАЦИОННО-СПРАВОЧНЫЙ ФОНД.

ГЕНЕРАЛЬНЫЙ СЛОВНИК может быть создан как семейство информационно однородных баз данных на основе Сводного словника, созданного в Словарном Секторе Института русского языка АН СССР в г. Ленинграде и других словникоподобных словарей, таких как Грамматический сло-

варь русского языка А.А. Зализняка, Орфографический словарь, выходящий скоро из печати Орфоэпический словарь, Русский семантический словарь Ю.Н. Караулова и др.

Объектом хранения в этой БД является вокабула, т.е. слово, являющееся потенциальным заголовком какой-либо словарной статьи и именем определенного лексического значения, его атрибутами - идентификаторы значений, а значениями атрибутов - имена и входы баз данных, в которых данное слово или его формы зафиксированы в качестве атрибутов каких-либо объектов - словарных статей, текстов, грамматических правил, статистических сводок, научных статей и т.д. Кроме того, в ГЕНЕРАЛЬНОМ СЛОВНИКЕ слово должно снабжаться дополнительной, наиболее общей информацией, относящейся к слову в целом, и в норме такой, как произношение его форм, формообразование, набор сем (с учетом возможных дополнений и исправлений к этой информации, содержащихся в других, адресуемых СЛОВНИКОМ базах данных). В качестве поискового индекса к СЛОВНИКУ может использоваться морфемный справочник русского языка.

ГЕНЕРАЛЬНЫЙ СЛОВНИК дает наиболее общее представление о словарном составе русского языка и его представлении в различных словарях, связывает воедино словари между собой и словари с текстами и другими источниками, позволяя получать для каждого слова всю зафиксированную информацию о нем.

Другим типом семейства словарных баз данных является ТЕРМИНОЛОГИЧЕСКИЙ ФОНД, в котором могут быть выделены информационно однородные подсемейства отраслевых информационно-поисковых тезаурусов, терминологических ГОСТов, многоязычных автоматических словарей и автоматических энциклопедических общих и отраслевых словарей и энциклопедий. Основой такого фонда может стать Автоматизированная система ведения информационно-поисковых языков в ГАНТИ и Макротезаурус этой системы, разработанные во ВНИИКИ Госстандарта СССР, а также автоматизированная система стандартизованной терминологии, функционирующая в этом же институте (Автоматизированная.., 1982).

АКАДЕМИЧЕСКИЙ СЛОВАРНО-ГРАММАТИЧЕСКИЙ ФОНД образует несколько подфондов. Одним из таких подфондов должен стать свод академических словарей - наиважнейших источников зафиксированных на сегодняшний день знаний о русском языке.

В советской лексикографии утвердилось учение о системе словарей как о такой совокупности, "которая позволяет описать лексическую систему языка (словарный состав языка) в ее полном объеме. Для русской лексикографии эту систему словарей должны составить: словарь современного русского литературного языка, исторический словарь русского языка XIX в., исторический словарь русского языка XVII в., словарь русских народных говоров, словарь древнерусского языка XI-XVII вв. Эту систему должны дополнять: словарь синонимов, фразеологический словарь, словарь антонимов, словообразовательный

словарь и другие типы специальных лексикографических изданий. В совокупности эти словари должны выполнить задачу, которую ставил перед своим словарем-тезаурусом А.А. Шахматов" (Сороколетов, 1978). Понятие системы словарей возникло и разрабатывалось в связи со стремлением сохранить гуманитарную традицию русских словарей, ставящую естественный предел допустимым объемам словарных статей и их сложности, что приводит к многообразию типов словарей, каждый со своим составом лексикона, схемами словарных статей и рубриками лексикографического описания. Современные средства автоматической лексикографии позволяют вернуться к шахматовской традиции, спрятать внутрь базы данных всю действительную сложность и объемность описания, сделав для пользователя "видимой" каждый раз ту часть тезауруса и в том представлении, которое ему необходимо и соответствует его лексикографическому восприятию. В автоматической лексикографии понятию типа словаря соответствует понятие режима обращению к нему; путем ограничения на выдачу словарных статей и их компонентов словарь может быть во внешней форме представлен в нужном объеме (большой, средний, малый), в нужном аспекте (толковый, переводной, семантический, словообразовательный и т.д., синонимов, антонимов, конверсивов, фразеологизмов и т.д.). Почти по любому лексикографическому параметру, зафиксированному в книге (Караулов, 1981), возможна выдача соответствующего словарного материала с соответствующей перестройкой словарной статьи. Современные средства управления базами данных позволяют перевести учение о лексикографических параметрах в алгоритмический план.

Естественным расширением словарного академического фонда являются историко-этимологический, диалектологический, топонимический и другие фонды, комплектуемые из соответствующих источников. Подбор атрибутов словарных статей и их увязывание в обобщенную словарную статью должны соответствовать требованиям всестороннего описания лексики, т.е. содержать грамматические,лексикологические, семасиологические, стилистические, фонетические, диалектологические,исторические, энциклопедические, библиографические и другие параметры.

Представляется целесообразным непосредственно связать собственно словарные базы с базой данных Академической грамматики, поместив в качестве адресующего индекса к текстам Грамматики ее словарный и предметный указатели, через которые нужные места Грамматики свяжутся с нужными местами словарных статей.

Машинный ИЛЛЮСТРАЦИОННО-ТЕКСТОВОЙ ФОНД явится аналогом картотек цитат. Наряду с собственно цитатами, подбираемыми исследователями по определенным правилам, в него следует включить также полные тесты образцовых с языковой точки зрения произведений, включая памятники истории языка, а также определенным образом систематически сделанные выборки из текстов разных жан-

ров, стилей, форм речи и временных срезов. Кроме того в ИЛЛЮСТРАЦИОННО-ТЕКСТОВОЙ ФОНД следует поместить также различные специальные формы: речения, пословицы, поговорки, штампы, образцы телеграфного стиля, рекламные формулы и т.п. Каждое вхождение определенного слова может быть адресовано общим индексным словоуказателем, так что ИЛЛЮСТРАЦИОННО-ТЕКСТОВОЙ ФОНД - это одновременно и конкорданс, и частотный словарь, и обратный словарь, и просто словоуказатель, в зависимости от режима обращения к нему.

На основе ИЛЛЮСТРАЦИОННО-ТЕКСТОВОГО ФОНДА может быть сформирована ЛИНГВОСТАТИСТИЧЕСКАЯ БАЗА ДАННЫХ, в которой разместятся статистические данные, позволяющие вычислять употребительность слов и других явлений, степени взаимосвязи и взаимообусловленности явлений, статистические характеристики текстов и другие оценки.

В значительной степени работа по извлечению из текстов, словарей и грамматик необходимой информации может быть автоматизирована уже накопленным к настоящему времени и созданным в рамках машинного фонда русского языка программным обеспечением. Это обеспечение может быть разделено на два класса: программные комплексы, реализующие морфологический, синтаксический анализ текстов, и программы различного назначения, свободно комбинируемые в целях формирования программных комплексов какого-либо специального назначения. Первый класс мы называем языковыми процессорами, точнее процессорами русского языка, второй - прикладными лингвистическими программами. Четкую границу между ними провести трудно, но в общем случае можно утверждать, что "настоящие" языковые процессоры наряду с программами, реализующими собственно лингвистический анализ или синтез, включают в себя достаточно полные формальные грамматики и словари с встроенными в них или эксплицитно сформулированными и записанными на определенном формальном языке лингвистическими алгоритмами. Языковые процессоры - это комплексы программ и данных, реализующие определенные модели лингвистического разбора и/или конструирования текстов или модели понимания и/или порождения текстов. В известной мере можно сказать, что языковые процессоры, в рамках своей тематической области и реализованной в них модели являются машинными фондами языка в наиболее узком, собственном смысле слова. Таких ядерных фондов сегодня может быть несколько и они могут образовать отдельный ФОНД ЯЗЫКОВЫХ ПРОЦЕССОРОВ, остальные прикладные программы образуют ФОНД ЛИНГВИСТИЧЕСКИХ АЛГОРИТМОВ И ПРОГРАММ.

Назначение ФОНДА ЯЗЫКОВЫХ ПРОЦЕССОРОВ двоякое. С одной стороны, языковые процессоры - это готовые инструменты автоматической обработки данных и человеко-машинного общения на естественном языке. Именно в получении таких инструментов заинтересованы разработчики математического обеспечения и аппаратуры современных ЭВМ. С другой стороны, языковые процессоры - это действующие модели языка, а область их разработки и экс-

плуатации - по-новому понимаемая экспериментальная лингвистика сегодняшнего дня.

В области конструирования языковых процессоров накоплен уже значительный опыт, а в ряде случаев инженерно реализованы наиболее глубокие теоретические модели анализа и синтеза слов, словосочетаний, предложений, связного текста и диалога. Работы, опубликованные в последние годы в серии "Общение с ЭВМ на естественном языке" (Нариньяни, 1979-1982), в книге Э.В. Попова под тем же названием (Попов, 1982), в сборнике (Актуальные.., 1982), показывают, насколько более высокие требования к описанию и интерпретации языковых фактов предъявляет автоматическая обработка по сравнению с обычными лексикографическими и грамматическими описаниями, ориентированными на человека. Суть отличия лингвистического обеспечения автоматизированных систем от "лингвистического обеспечения общечеловеческой деятельности" (если можно так выразиться) скорее состоит не в строго формальном построении первого и в неформальном, гуманитарно ориентированном построении второго, а в более тонкой дифференциации в системе лексических и грамматических значений, средств связи выражений на всех уровнях языковой структуры, в более тонкой согласованности параметров разных уровней. В ряде словарей, разработанных в качестве лингвистического обеспечения языковых процессоров, число параметров описания слов в каждом из разделов словарной статьи может на порядок превосходить число параметров обычных словарей. Эти разработки открывают новые перспективы и для общей лексикографии, но вряд ли могут быть реализованы в ней в условиях ручного труда лексикографов, без поддержки средств автоматизации, в традиционной картотечной или полиграфической форме словарей. В современных языковых процессорах мы видим прообраз лингвистических автоматов будущего, реализованных в микрокомпьютерной форме в виде персональных ЭВМ, незаменимых спутников слепых и глухих, помощников переводчиков, редакторов, корректоров, справочников для всех и каждого. Размещение в микрокомпьютерных базах данных лексикографических источников в виде словарей для языковых процессоров приводит к новому современному пониманию ленинских слов о словарях "для учения всех".

Следующим компонентом машинного фонда русского языка является ЛЕКСИКОГРАФИЧЕСКАЯ БАЗА ФОНДА, образуемая типовой автоматизированной лексикографической системой, т.е. системой программ и данных, предназначенных для автоматизации основных лексикографических работ - сбора и размещения в базах данных текстовых и словарных источников, отслеживания всех этапов разработки словаря, организации работы лексикографов за терминалом, производства необходимых сортировок и группировок лексики, производства операций над словарями и текстами, отслеживания выполнения требований к формированию словарных статей и др.

В настоящее время в автоматической лексикографии существуют, на наш взгляд, три главных проблемы, без

решения которых немыслимо дальнейшее существенное продвижение в направлении создания крупных автоматических словарей.

Первая проблема носит чисто технологический характер и состоит в накоплении и поддержании в актуальном состоянии больших текстовых и словарных источников, насчитывающих миллионы и десятки миллионов словоупотреблений. Такие источники не могут создаваться и переноситься на машинные носители в одном месте и в короткий промежуток времени. Обычно они создаются разными коллективами, в различных местах, в течение длительного периода. Это требует единой системы их кодифицирования, членения, разметки, отслеживания их прохождения через все этапы обработки, системы их учета и комплектации. В Московском университете разработана АЛС УНИЛЕКС (Андрющенко, 1982), в которой применяется некоторая абстрактная и формальная система членения текстовых и словарных совокупностей на пакеты,тексты, модули, сегменты, поля и элементы, снабженная своим языком разметки, основанным по польской инверсной записи формул (разграничитель выполняет роль оператора). Это позволяет связать с выделяемыми фрагментами текста или словарной статьи определенные имена, а с этими именами и разграничителями - определенные программы обработки, что придает описанию словарной статьи вид грамматики непосредственно-составляющих с приписанными им атрибутами. В известной мере - это шаг в направлении алгоритмизации проектирования словарей, организации исходных данных и работы поддерживающего программного обеспечения. Разработаны также форматы другой необходимой для учета и контроля за прохождением работ управляющей информации, сопровождающей словарные статьи и тексты, размещаемые в системе. Сообщение об этом было опубликовано в "Методике первичной обработки лингвистических данных в интерактивном режиме" (Андрющенко, 1981). Однако до полного решения проблемы еще далеко: нужна, во-первых, достаточно полная и тонкая классификация лингвистических источников для словарей и, во-вторых, классификация видов лексикографических работ.

Вторая проблема связана с разработкой процедур группового контроля за соблюдением заданного формата словарных статей и непосредственного сопоставления форматов словарных статей для лексики разных групп. Эта проблема хорошо известна и в традиционной лексикографии. В работе (Шведова, 1981) говорится: "Для того чтобы избежать разнобоя в словаре, лексикограф должен работать не с алфавитным списком слов, а с определенными их лексическими группировками внутри отдельных частей речи. Только на этом пути могут быть достигнуты единство в разграничении значений, однотипность толкований, непротиворечивые приемы подачи фразеологизмов и примеров, последовательность и единообразие стилистических характеристик. Иными словами,словарь станет не просто квалифицированно составленной коллек-

цией слов, за которой стоит лексикограф со своим знанием языка и со своим пониманием отдельного слова, а таким научным произведением, которое опирается на специально для него разработанные эталоны описания слов, так или иначе объединенных формально и семантически". В плане автоматизации и эта проблема частично решается введением формального языка для задания структуры словарной статьи. Однако для действительного решения проблемы не хватает именно той классификации слов, разработка которой заложена в цитированной статье Н.Ю. Шведовой.

Третья проблема связана с построением специальных лексикографических процессоров. В идеале хотелось бы, чтобы такой процессор, получив на входе формализованный проект словаря и доступ к базам данных, содержащим источники для формирования словаря, формировал бы новую базу данных, содержащую в качестве объектов структурно связанные фрагменты словарных статей, предусматриваемые проектом. Конечно, нельзя ожидать от такого процессора, чтобы он писал толкования значений слов или квалифицированно подбирал иллюстрации. Достаточно потребовать, чтобы он в нужных местах словарной статьи приводил возможные варианты толкований или справки для их формирования или фрагменты конкорданса. Остальная работа могла бы быть выполнена лексикографом вручную за терминалом или по распечатке с последующим вводом корректирующих данных. Разумеется, в качестве своей составной части лексикографический процессор должен иметь языковой процессор, предназначенный для анализа обрабатываемых текстов, дифференциации контекстуальных условий и реализованных в них значений. Решение этой проблемы имеет значение и в чисто прикладной области - в эксплуатации уже созданных языковых процессоров. Здесь эта проблема носит характер пополнения словаря по входным сообщениям. С точки зрения любого автоматического словаря каждое не содержащееся в нем слово - неологизм, и требуется "догадаться" о его формальном статусе и хотя бы приблизительно - о значении. Другая форма той же проблемы известна в терминологических банках данных: необходимо выдавать ответы на запросы в форме сложных слов или словосочетаний, не содержащихся непосредственно в качестве входов словарных статей. Таким образом, решение этой проблемы является условием и для решения многих прикладных задач.

В качестве лингвистического обеспечения ЛЕКСИКОГРАФИЧЕСКОЙ БАЗЫ ФОНДА необходимо создать Толковый статистико-комбинаторный словарь, аналог известного ТКС с добавлением статистической информации путем отсылки к ЛИНГВОСТАТИСТИЧЕСКОЙ БАЗЕ ФОНДА, примеров к каждой рубрике словарной статьи - путем отсылки к ИЛЛЮСТРАЦИОННО-ТЕКСТОВОМУ ФОНДУ, информации о тезаурусных отношениях - путем отсылки к ГЕНЕРАЛЬНОМУ СЛОВНИКУ и некоторой дополнительной информации. ТСКС станет новым типом академического словаря, предназначенным для использования в вычислительной среде. Он будет включать лексику последней четверти XX века.

Проект машинного фонда русского языка предусматривает также разработку группы информационно-справочных систем. Сюда относятся: библиографическая информационно-справочная система, документальная информационная система, информационная система для накопления и обработки социолингвистических и диалектологических анкет и ичформационно-обучающая система (информатор фонда). Типы этих систем хорошо известны в информатике, однако каждая из них в МАШИННОМ ФОНДЕ РУССКОГО ЯЗЫКА будет иметь свои особенности.

Особенностью библиографической системы будет наличие двух взаимодействующих ИПЯ - информационно-поискового языка лингвистики и языка примеров, т.е. русского языка; особенностью документальной системы будет наличие в ней средств создания, оформления и полиграфического воспроизведения документов и публикаций, в частности, автоматизированных средств создания аппарата изданий; информационная анкетная система будет включать в себя полный пакет статистических и графических программ, управляемых проблемно-ориентированным языком пользователя; информатор фонда - это комбинация информационной системы о составе фонда с обучающей системой, целью которой является обучение пользователей работе в фонде и помощь им в случае их ошибок и в затруднительных ситуациях.

Связь между различными базами данных фонда должна осуществляться одной или несколькими МОНИТОРНЫМИ СИСТЕМАМИ АВТОМАТИЗАЦИИ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ. Каждая из мониторных систем в зависимости от свойств операционной системы и свойств СУБД должна обладать следующими функциями:

- запуск пользовательских задач в пакетном режиме в качестве своих подзадач, задач операционной системы или задач СУБД;
- организация диалогового взаимодействия с пользователем, запуск его задач в диалоговом режиме и обеспечение диалога пользователя со своей задачей;
- интерпретация языка пользователя, формирование и передача в СУБД описаний пользовательских баз данных и запрошенных пользователем функций обработки;
- передача результатов обработки во внешние файлы или в определенные пользователем базы данных;
- анализ процесса диалога пользователя с мониторной системой и генерация недоопределенных пользователем функций обработки.

Язык взаимодействия пользователя с МОНИТОРНОЙ СИСТЕМОЙ должен быть прост и эффективен. Это - язык первого уровня, состоящий из обобщенных директив.

Пользовательский язык второго уровня, предназначенный для более тонкого программирования лингвистических работ, например, для записи формальных грамматик, словарной информации в словарях для языковых процессоров, лингвистических алгоритмов, должен разрабатываться в рамках конструирования языковых процессоров и использоваться в их диалоговых режимах. Необходим также пользовательский язык третьего уровня, позволяющий опреде-

лять в системе формальные языки, т.е. в конечном счете строить для них интерпретаторы.

Разделение пользовательского языка на три уровня позволяет выделить простейшие функции обработки и сделать их доступными самому массовому пользователю-лингвисту, создать простую систему автоматизации лексикографических работ путем применения типовых программных средств. Язык, точнее языки второго уровня, - это собственно языки вычислительной лингвистики, они предназначаются для работающих в этой области; языки третьего уровня предназначаются для разработчиков специализированных систем программирования, в частности, программных средств машинного фонда русского языка.

Изложенный проект создания машинного фонда русского языка направлен на решение следующих основных задач:

- создать возможность эффективной централизованной разработки и поставки промышленности и НИИ лингвистического обеспечения для разрабатываемых систем общения с ЭВМ на русском языке и систем обработки документов в естественной языковой форме;
- создать систему комплексной автоматизации лингвистического труда: составления словарей, поиска и обработки научной информации, анализа текстов, проведения классификационных работ, подготовки аппарата изданий и т.п.;
- заложить понимаемую в современном смысле сокровищницу данных о русском языке во всем объеме этого понятия, во всех его временных и территориальных формах.

Мы считает, что переход к новым методам сбора, хранения, анализа и сопоставления данных о языке, новые методы создания и новые формы лингвистических источников, таких как автоматические словари и грамматики, могут быть жизнеспособными и эффективными, если они опираются на общую филологическую традицию и культуру, на глубокое изучение языка и учет информации о нем во всех формах его существования. Однако соединение лингвистической традиции и новых задач практики нужно осуществлять на путях новой информационной технологии, развиваемой системами обработки данных на естественном языке в интеллектуальной среде человеко-машинного общения.

ЛИТЕРАТУРА

Автоматизированная система ведения информационных языков АСВИЯ. Информационные материалы. - М.: Изд-во ВНИИКИ Госстандарта СССР, 1982, № 4.

Андрющенко В.М. Автоматизированная лексикографическая система UNILEX. (Основные проектные решения). - В кн.: Вычислительная лингвистика. Теоретические аспекты. Вопросы автоматизации лексикографических работ/ Под ред. В.З.Демьянкова. - М.: Изд-во Моск. ун-та, 1982, с. 104-119.

Андрющенко В.М. Методика подготовки и первичной обработки лингвистических данных в интерактивном режиме. - М., 1981 (препринты № 138, 139 Ин-та русского языка АН СССР).

Актуальные вопросы практической реализации систем автоматического перевода. (Материалы первого совместного советско-французского семинара, состоявшегося в Москве в 1977 г.). Ч. 1 и П. - М.: Изд-во Моск. ун-та, 1982.

Ершов А.П. К методологии построения диалоговых систем: феномен деловой прозы. - В кн.: Вопросы кибернетики. Общение с ЭВМ на естественном языке. - М.: Наука, 1982, с. 3-20 (Научный совет по комплексной проблеме "Кибернетика" АН СССР).

Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. - М.: Наука, 1981.

Нариньяни А.С. 1979-1982: Вопросы разработки прикладных систем/ Под ред. А.С, Нариньяни.- Новосибирск: Изд-во ВЦ СО АН СССР, 1979; Синатксический и семантический компонент лингвистического обеспечения/ Под ред. А.С. Нариньяни. - Новосибирск: Изд-во ВЦ СО АН СССР, 1979; Представление знаний и моделирование процессов понимания/ Под ред. А.С.Нариньяни. - Новосибирск: Изд-во ВЦ СО АН ССС,1980; Формальное описание структуры естественного языка/ Под ред. А.С. Нариньяни. - Новосибирск:Изд-во ВЦ СО АН СССР, 1980; Разработка формальной модели естественного языка/ Под ред. А.С. Нариньяни.- Новосибирск: Изд-во ВЦ СО АН СССР, 1981; Лингвистические процессоры и представление знаний/ Под ред. А.С. Нариньяни. - Новосибирск: Изд-во ВЦ СО АН СССР, 1981; Прикладные и экспериментальные лингвистические процессоры/ Под ред. А.С. Нариньяни. - Новосибирск: Изд-во ВЦ СО АН СССР, 1982; Формальное представление лингвистической информации/ Под ред. А.С. Нариньяни. - Новосибирск: Изд-во ВЦ СО АН СССР, 1982.

Попов Э.В. Общение с ЭВМ на естественном языке. - М.: Наука, 1982.

Сороколетов Ф.П. Традиции русской советской лексикографии. - Вопросы языкознания, 1978, № 3.

Шведова Н.В. Однотомный толковый словарь (специфика жанра и некоторые перспективы дальнейшей работы). В кн.: Русский язык. Проблемы художественной речи. Лексикология и лексикография. - М.: Наука,1981, с. 171.

GRUNDKOMPONENTEN EINER SPRACHDATENBANK DES RUSSISCHEN

Wladislaw M. Andrjuschtschenko

R e s ü m e e

Im vorliegenden Aufsatz werden Grundzüge eines Projekts beschrieben, das auf die Schaffung einer Sprachdatenbank des Russischen gezielt ist. Die bezweckte Sprachdatenbank wird folgende Grundkomponenten enthalten: einen Generalwortindex des Russischen, einige Textdatenbanken für die Literatur des XI. bis XX. Jh, einige vielsprachige Terminologiebanken, einige wissenschaftliche lexikalisch-grammatikalische Sprachdatenbanken, eine linguostatistische Datenbank, einen Fond linguistischer Prozessoren des Russischen, einen Fond linguistischer Algorithmen und Programme, ein automatisiertes System für Lexikographie und einige Informationssysteme für bibliographische, dialektologische und soziolinguistische Daten. Diese Komponenten werden unter Steuerung einiger Monitorsysteme Sprachwissenschaftlern an linguistischen Institutionen als eine verteilte automatische Datenbank zur Verfügugng gestellt.

СВЯЗИ, ЕДИНИЦЫ И ЕДИНСТВА СВЕРХФРАЗОВОГО УРОВНЯ ЯЗЫКА

В.Е. Берзон, М.С. Блехман, Р.Г. Пиотровский

Введение. Разработка лингвистики связного текста (ЛСТ) может стать ключем для разрешения многих теоретических и практических вопросов коммуникации и актуализации языковых единиц в речи. Ведь выбор из парадигм языка тех или иных альтернативных единиц, в том числе и значений, часто детерминируется не столько статусом этих единиц в системе языка или их ролью в отдельно взятом предложении, сколько ситуативной пресуппозицией и коммуникативным замыслом всего текста (ср. правила выбора определенного артикля в текстах на заподноевропейских и балканских языках)*. Обладая объективными приемами для выявления этого замысла и реализующей его сверхфразовой структуры, лингвист получает возможность двигаться при анализе связного текста и его составляющих не только снизу вверх, т.е. от мелких элементов к более крупным, как это часто делается, но и сверху вниз, т.е. от семантики текста, гиперсинтаксиса, кросс-референции через семантико-синтаксическую структуру отдельных предложений, к малому синтаксису и семантике словосочетаний. Преимущество второго, "нисходящего" анализа заключается в том, что он выявляет системную значимость каждой лингвистической единицы (ЛгвЕ) (Ельмслев Л., 1960,с. 288-290; Хьюз Дж., Митчом Дж., 1980, с. 14-18) как относительно ЛгвЕ данного уровня, так и по отношению к единицам более высоких уровней.

Исследования в области ЛСТ имеют первостепенное значение для развития нейролингвистики (Лурия А.Р., 1979, с. 187-216) и инженерно-лингвистических аспектов искусственного интеллекта (Рафаэл Б., 1979, с. 351-354; Пиотровский Р.Г., 1981, с. 36-37). Включение инженерно-лингвистических вопросов в проблематику ЛСТ должно выполнять в первую очередь методологическую функцию. Дело в том, что современное языкознание, преодолевая созерцательную гносеологическую установку, направленную на построение правдоподобных схем и выведение идеальных вариантов, создает новую, конструктивную гносеологию, ориентированную на верификацию и

---

* См. Пиотровский Р.Г., 1960, с. 50 и сл.; Piotrowski R.,1965; Brainerd B., 1972; Вайнрих Х., 1978; Шкирич А.А. и др.,1979.

корректирование этих схем и инвариантов путем инженерного воспроизведения на их основе реальных лингвистических объектов-вариантов и их совокупностей (Супрун А.Е., 1978, с. 118; Зотов А., 1979).

Исходя из этой установки задачами ЛСТ следует считать;

1) выявление релевалентных для сверхфразового уровня (СУ) признаков и связей;

2) выявление с их помощью единиц указанного страта и установление их иерархий;

3) экспериментальную проверку реальности выделенных признаков ЛгвЕ и построенных на их основе теоретических схем и гипотез. Попытаемся реализовать эту программу.

1. Типы единиц, связей и единств на СУ. Будем считать минимальной единицей ЛСТ предложение, мексимальной - собственно текст. Различия этих единиц сводятся к двум пунктам.

Во-первых, предложение отличается от текста тем, что оно синтаксически структурировано, т.е. является синтаксической единицей, в то время как текст организуется в основном с помощью семантических средств (Бархударов Л.С., 1980, с. 53) и, поэтому, синтаксическим образованием признан быть не может.

Во-вторых, законченный текст является в идеале завершенным, самонасыщенным образованием, тогда как предложение, будучи изъятым из текста, теряет некоторую часть своей смысловой (семантико-прагматической) информации (Гиндин С.И., Леонтьева Н.Н.,...; Леонтьева Н.Н., 1981). Например, истинный смысл первого предложения (заглавия) - "Вишневый сад" и полное значение последней фразы чеховской пьесы - Наступает тишина и только слышно, как далеко в саду топором стучат по дереву (А.П. Чехов. Избранные произведения, т. 2. Рассказы, повести и пьесы. Л.: Лениздат, 1960, с. 825), становятся понятными только после прочтения всей пьесы. Эта антиномия предложения и текста обнаруживается и в научной прозе, не говоря уже о бытовой речи. Например, вопрос, содержащийся в первой части заглавия известной книги Э. Шредингера "Что такое жизнь" ("What Is Life?"), определенным образом прогнозирует содержание всей книги, представляющей по замыслу автора ответ на этот вопрос. Текст же насыщает предшествующее ему заглавие философским и биофизическим содержанием.* Смысловое насыщение рассмотренных предложений осуществляется без привлечения каких-либо лексико-грамматических средств.

---

* Антиномия предложения и текста устраняется лишь в тех случаях, когда связный текст сводится к одному предложению, как это имеет, например, место в некоторых афоризмах К. Пруткова или "Фрашках" ("Пустяках") Я. Кохановского (XVI в.).

Однако, наряду с имплицированным насыщением, в языке широко используется насыщение, реализующееся с помощью эксплицитных языковых средств (ср. следующую реплику Раневской из 1-го действия "Вишневого сада": Спасибо, родной, (S_0). Я привыкла к кофе (S_1). Пью его днем и ночью (S_2). Спасибо, мой старичок (S_3), там же с. 781). Предложение S_2, будучи вырванным из контекста, оказывается непонятным. Его насыщение информацией из S_1 эксплицируется с помощью синтаксического приема-репризы: ...кофе. Пью его днем и ночью.

Критерий насыщения является наиболее адекватным признаком выделения единиц, с которыми должна работать ЛСТ. Его существо заключается в том, что каждое предложение связного текста рассматривается на фоне других, окружающих его в контексте предложений $...S_f$, $S_h...S_j$, $S_k....$ (Ревзин И.И., 1978, с. 144). При этом некоторые предложения, окружающие S_i, насыщают его информацией, необходимой для того, чтобы оно воспринималось аналогично смыслу автора. Восприятие от насыщающих его предложений приводит к тому, что контекстная обусловленность этого предложения, не говоря уже о чисто формальных отношениях (ср. его-кофе), становится для непосвященного читателя абсолютно неясной. Что же касается заглавия пьесы "Вишневый сад", то оно может в принципе иметь любое насыщение - от социально-философского до плодоводческого.

Можно выделить три вида насыщения:

1) одностороннее ($S_h \rightarrow S_i$, $S_i \leftarrow S_j$), при котором предложение S_i насыщается предложениями S_h или S_j (ср. отношения предложений S_1 и S_2 в реплике Раневской);

2) взаимное ($S_f \rightleftarrows S_h$), которое характеризует отношение между заглавием книги Э. Шредингера ("Что такое жизнь") и текстом всей книги;

3) отсутствие насыщения $S_j \parallel S_k$; в качестве примера можно указать на отношение предложений S_0 и S_i, а также S_2 и S_3 в реплике Раневской.

Как уже говорилось, с точки зрения его выражения насыщение может быть эксплицитным, т.е. посредством поверхностно-синтаксической сверхфазовой (ПСС) и имплицитным, т.е. передаваемым через имеющую формальных показателей глубинно-синтаксическую сверхфазовую связь (ГСС).

Хотя только что введенные критерии еще не объясняют, каким образом происходит формирование целостного текста, с их помощью можно выделить крупные, чем предложение, сверхфазовые образования, образующие промежуточный уровень между предложением и связным текстом. В качестве таких сверхфазовых образований мы выделяем с в е р х ф р а з о в ы е е д и н и ц ы (СЕ), среди которых будем различать: п о в е р х н о с т н о - с и н т а к с и ч е с к и е, точнее - морфолого-синтаксические с в е р х ф р а з о в ы е е д и н и ц ы (ПСЕ) и г л у б и н н о - с и н т а к с и ч е с к и е, точнее - семантико-синтаксические с в е р х ф р а з о в ы е е д и н и ц ы (ГСЕ).

ПСЕ представляет собой такую цепочку, в которой одно предложение является автосемантичным, т.е. семантически насыщенным, а остальные - синсемантичными, т.е. семантически ненасыщенными, и каждое предложение эксплицитно связано посредством ПСС хотя бы с одним из предложений цепочки. Таким образом, под ПСЕ понимается синтаксически организованная сверхфразовая единица, центром которой является автосемантичное предложение. Примером ПСЕ служит следующая цепочка предложений: важнейшей теоретической задачей исследования терминов и терминосистем является исследование языковой и логико-семантической сущностей термина, которые проявляются в его синтаксическом функционировании (S_i). Эта проблема имеет и большое практическое значение (S_{i+1}). (Р.Ю. Корбин и др. Экспериментальное исследование... - НТИ, сер. 2, 1978, № 11, с. 8.

Следующее предложение этого текста ... Прикладное значение термина определяется прежде всего тем, что он является основным носителем научной и технической информации в специальных текстах (S_{i+2})... автосемантично и поэтому в данную ПСЕ не входит.

Глубинно-синтаксическая сверхфразовая единица, также как и ПСЕ, включает в качестве организующего центра одно автосемантичное предложение, однако в отличие от ПСЕ составляющие его предложения связаны между собой имплицитной ГСС. Важной особенностью этих единиц является то, что каждая ГСЕ может быть превращена в поверхностно-синтаксическую единицу посредством "восстановления" недостающих лексико-синтаксических средств (эти трансформации рассматриваются в разделе 3 статьи).

СЕ с помощью ПСС и ГСС могут группироваться в более крупные сверхфразовые объединения, содержащие несколько автосемантических предложений. Такие объединения мы будем обозначать традиционным термином "сверхфразовые единства" (СФЕ). Примером СФЕ может служить приведенная выше реплика Раневской.

Сверхфразовое единство, состоящее только из поверхностно-синтаксических единиц, назовем синтаксическим СФЕ. Частным случаем синтаксического сверхфразового единства является СФЕ, состоящее из одной ПСЕ. В реальных текстах наиболее распространенными являются СФЕ смешанного типа, представляющие собой сочетание поверхностно-синтаксических и глубинно-синтаксических единиц.

2. Выделение и анализ сверхфразовых единиц, организованных с помощью эксплицированных поверхностно-синтаксических сверхфразовых связей.

В функции ПСС могут выступать:

- морфологические и морфолого-синтаксические категории, например, категории вида и времени (Реферовская Е.А., 1983, с. 111, 177; Пиотровский Р.Г., 1956, с. 154-180; Маслов Б.А., 1975);
- различные синтаксические приемы типа эллипсиса (Откупщикова М.И., 1982, с. 71-73);

- знаки препинания, например, точка с запятой, вопросительный знак;

- лексические коннекторы (ЛК), под которыми мы будем понимать слова и словосочетания, используемые для объединения предложений в ПСЕ.

Основное внимание мы уделим именно лексическим коннекторам, во-первых потому, что они наиболее регулярно по сравнению с другими подобными средствами употребляются в текстах большинства подъязыков и функциональных стилей русского и английского языков, а во-вторых, потому что другие ПСС могут быть легко трансформированы в лексические коннекторы. Например, предпрошедшее время Past Perfect в английском языке, которое часто требует для своего насыщения информацию о предшествующем прошедшем действии из предыдущих предложений, может быть трансформировано в предложение, начинающееся с коннектора after this event (основной глагол в новом предложении будет стоять в Past Indefinite)

Таблица 1

Классификация лексических коннекторов (к.)

Классы и подклассы коннекторов	Примеры	
	английские	русские
1	2	3
1. Коннекторы (указывающие на связь между объектами)		
1.1. Анафорические к.		
1.1.1. Коннекторы сходства и различия	the, this, such, another, other	этот, такой, упомянутый, предыдущий, описанный
1.1.2. Конспекторы-заместители	he, she, one, those	он, оно, она, они и их производные
1.2. Катафорические к.	the following	следующий
2. Коннекторы (указывающие на логическую связь между целыми предложениями (суждениями)		
Анафорические к.		
2.1.1. Темпоральные к.	simultaneously, then, thereafter	одновременно, с тех пор, в то же время, после этого
2.1.2. Коннекторы-локализаторы	here, there	здесь, там

1	2	3
2.1.3. Уточняющие к.	for example, in particular	например, в частности
2.1.4. Результативные к.	finally, in total	в конце концов, в результате
2.1.5. Причинно-следственные	then, thus, therefore	следовательно, поэтому, таким образом
2.1.6. Перифрастические к.	in other words, that is	другими словами, короче говоря, то есть
2.1.7. Сравнительные к.	similarly, likewise	аналогично, соответственно
2.1.8. Дополнительные к.	also, besides, furthermore,	более того, дополнительно, также, наряду с этим
2.1.9. Противительные к.	but, yet, however	но, тем не менее, однако, наоборот
2.1.10. Подтверждающие к.	indeed, really	действительно, в самом деле
2.1.11. Соединительные	and	и
2.2. Катафорические (пояснительные) к.	namely, let	пусть, а именно
2.3. Двусторонние перечислительные к.	firstly, secondly	во-первых, во-вторых, с одной стороны, с другой стороны.

Хотя коннекторы принадлежат к разным частям речи, все они объединяются в особый функционально-грамматический класс, характеризующийся следующими особенностями:

- ЛК служит маркером насыщенности своего предложения,
- входя в состав предложения S_i, ЛК как бы "заглядывает (Адмони В.Г., 1973, с. 354) в одно из предшествующих (S_h) или последующих (S_j) предложений текста, обеспечивая связь S_i с этими предложениями,
- значение и синтаксическая функция ЛК может быть выявлена лишь на сверхфразовом уровне.

Лексические коннекторы оформляют в тексте как одностороннее, так и двустороннее насыщение предложений.

Одностороннее отношение $S_h \rightarrow S_i$ устанавливается с помощью анафорических коннекторов, например: Всего в тезаурусе представлено 100 таких дескрипторов (S_h). Для

всех их составлены дескрипторные статьи (S_i). (Г.Г.Васильева. О тезаурусе историзмов... - НТИ, сер. 2,1982, № 5, с. 20).

Отношение $S_i \leftarrow S_j$ устанавливается катафорическими ЛК, например: Пусть λ -интенсивность запросов в системе (S_i - прим. авт.) Тогда $P_j\lambda = \lambda$ определяет интенсивность запросов некоторого j-ого типа (S_j -прим. авт.) (Б.П. Николаев, Ж.Ф. Сергазин. Система...- НТИ, сер. 2, № 5, 1982, с. 10).

Наконец двустороннее отношение $S_h \leftrightarrows S_i$ устанавливается с помощью перечислительных коннекторов, например: Общественное значение определяется, во-первых, потребностью в объективных ориентировках при составлении планов (S_h - прим. авт.). Во-вторых прогноз чрезвычайно важен как защита от шока будущего (S_i - прим. авт.) (Ю.А. Шрейдер. Методологические проблемы... - НТИ,сер. 2, 1982, № 5, с. 2).

Важной особенностью ЛК является их омонимичность ложным коннекторам, т.е. таким ЛЕ, которые, совпадая по написанию с истинным ЛК, не принимают участия в связывании отдельных предложений. Проиллюстрируем соотношение истинных и ложных ЛК на примере двух употреблений в текстах союза но: Единообразная оцифровка трех взаимоортогональных осей позволяет высчитать число "кубиков", заключенных в пространстве между ними (S_h, прим. авт.). Но и ничего кроме такого подсчета сопоставление мер информации и времени не дает и дать не может (S_i прим. авт.) (В.П. Троицкий. Предвосхищение... - НТИ, сер. 2, 1982, № 6, с. 3).

Маркером ненасыщенности предложения S_i является союз но, одновременно он связывает оба предложения в составе ПСЕ. Напротив, предложение Возможно, не меньшее значение для освоения времени может иметь своеобразное концептуальное "сращивание" его не только с категорией пространства, но и с категорией информации (S_i, прим. авт.) (Там же, с. 3) насыщено (автосемантично),и союз но не является здесь ЛК.

Определение истинности потенциального ЛК без привлечения экстралингвистических сведений предполагает анализ структуры и лексического состава не только предложения, в которое он входит, но и зачастую и окружающих предложений. В частности, для решения вопроса о роли английского (французского, немецкого, румынского и т.д.) определенного артикля необходимо выяснить, имеется ли в предшествующем тексте антецедент оформленного этим артиклем существительного. При наличии антецедента артикль является ЛК, при отсутствии - нет.

Объективное выявление в тексте ЛК и их функций дает возможность построить процедуру выделения в тексте ПСЕ, а затем синтаксических СФЕ. Поскольку ПСЕ представляет собой группу предложений, объединенных формально выраженной синтаксической сверхфразовой связью, ее обрывы служат правой и левой границами ПСЕ. В качестве примера рассмотрим английский текст: ... (S_0), прим. авт.) All long, repeating chains are polymers,re-

gardless of how many monomers are used (S_i, прим. авт.) B u t when a polymer family includes copolymers, the term homopolymer is used to identify the single monomer type (S_2, прим. авт.). A n e x a m p l e is the acetal family (S_3, прим. авт.). Final properties of a copolymer depend on the percentage of monomer A to monomer B (S_4, прим. авт.). (Introduction to Polymer Chemistry.-Machine Design, 1976, v. 48, No. 5, p. 122). и его перевод:

... (S_0, прим. авт.). Все длинные, повторяющиеся цепи являются полимерами вне зависимости от количества используемых мономеров (S'_1, прим. авт.). Однако в тех случаях, когда семейство полимеров включает сополимеры, применяется термин "гомополимер", соответствующий полимеру, состоящему из мономеров одного типа (S'_2, прим. авт.). В качестве примера можно привести ацеталь (S'_3 прим. авт.). Конечные характеристики сополимера зависят от процентного отношения А к мономеру В (S'_4, прим. авт.).

Структура сверхфразовых связей текста в обеих версиях рассмотренного примера представлена на рис. 1.

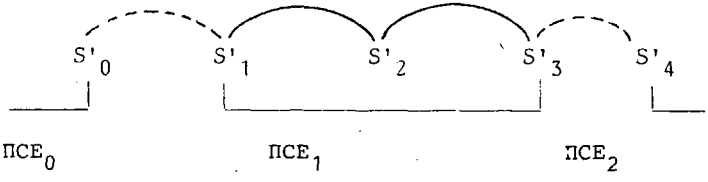


Рис. 1

Здесь сплошные линии указывают на наличие синтаксической сверхфразовой связи, а пунктирные помечают обрывы этой связи. Первый обрыв служит признаком начала $ПСЕ_1$, второй указывает на ее конец.

Как уже говорилось, элементарные ПСЕ могут группироваться с помощью ПСС в единицы более высокого уровня - в синтаксические сверхфразовые единства. Примером такого СФЕ может служить приведенный выше отрывок из "Вишневого сада". Его организация показана на рис. 2.

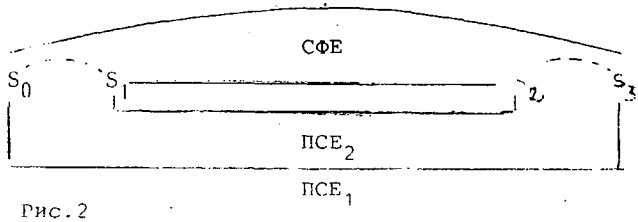


Рис.2

Рассмотренные схемы иллюстрировали "чистые" синтаксические сверхфразовые отношения. В то же время, классификация ЛК помогает определить и семантический характер выявленных в тексте ПСС. Так, семантико-синтаксическое представление сверхфразовой структуры, приведенной на рис. 1, показано на рис. 3.

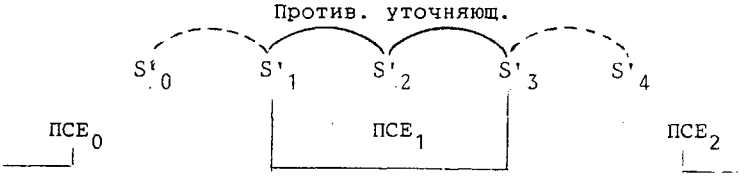
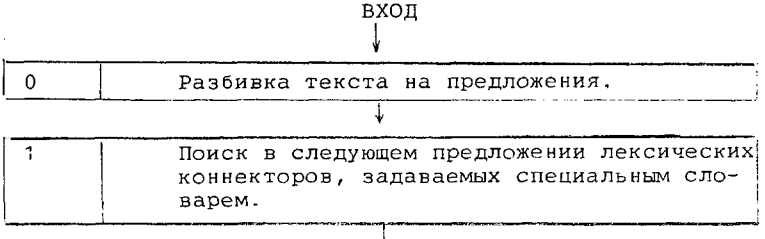
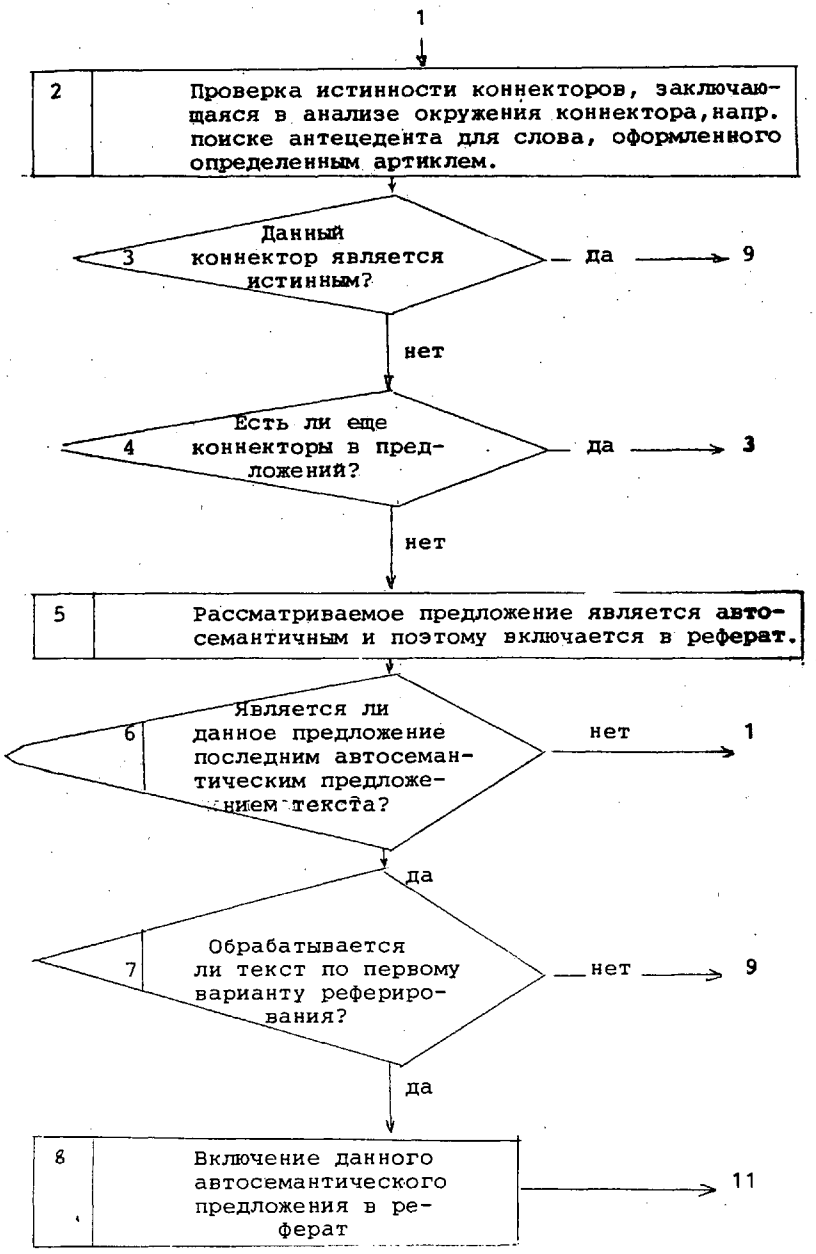


Рис. 3

Мы рассмотрели основные принципы организации синтаксических сверхфразовых связей. Однако в современной науке, использующей конструктивный подхол, построение идеальной схемы не должно венчать процесс исследования. В нашем случае его завершающим звеном, замыкающим обратную связь между идеальной моделью межфразовых отношений и ее эмпирической текстовой основой, является алгоритмическое воспроизведение этой модели в виде формальной распознающей грамматики, которая может использоваться в лингвистическом автомате для выделения и анализа ПСЕ, синтаксических СФЕ и автоматического реферирования этих текстов на основе полученных сведений. При этом коннекторы используются грамматикой в качестве "фулькрумов" (Гарвин П.Л., 1980), т.е. опорных точек при межфразовом синтаксическом анализе. Указанная распознающая грамматика реализована в виде алгоритма (рис. 4), который работает в двух режимных вариантах. При реализации первого варианта (рис. 5б) в реферат включаются автосемантичные предложения с указанием типов удаленных синсемантических предложений. Второй вариант выдает реферат, включающий только автосемантичные предложения (рис. 5в). (Программная реализация алгоритма осуществлена А.А. Захаровым).




1
2
Проверка истинности коннекторов, заключающаяся в анализе окружения коннектора, напр. поиске антецедента для слова, оформленного определенным артиклем.
3
Данный коннектор является истинным?
да
9
нет
4
Есть ли еще коннекторы в предложений?
да
3
нет
5
Рассматриваемое предложение является автосемантичным и поэтому включается в реферат.
6
Является ли данное предложение последним автосемантическим предложением текста?
нет
1
да
7
Обрабатывается ли текст по первому варианту реферирования?
нет
9
да
8
Включение данного автосемантического предложения в реферат
11

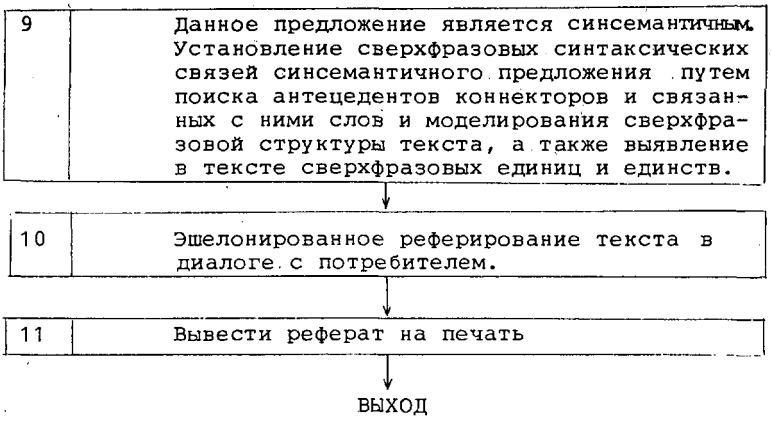


Рис. 4

В распечатке первого варианта кроме текста реферата приведена следующая информация:

а) предшествующее предложению двузначное число, которое указывает на порядковый номер данного предложения в исходящем тексте;

б) символ ¤ , указывающий на абзац в исходном тексте;

в) латинская буква в скобках при ЛК (M, L, C, B), служащая кодом класса, к которому принадлежит ЛК, и следующее за ней слово - имя класса.

Реализация алгоритма и анализ машинных результатов служат с одной стороны, подтверждением реального существования гипотетических заданных синтаксических сверхфразовых связей. С другой стороны, машинный эксперимент открывает определенные эвристические горизонты для дальнейших исследований в области лингвистики связного текста. В частности, эксперимент показал, что реферат, сформированный автоматом из автосемантичных предложений, в большинстве своем хорошо передает общее содержание целого текста. Этот факт, в свою очередь, указывает, что исследование семантико-синтаксических связей между автосемантичными предложениями текста помогло бы глубже проникнуть в механизм порождения связного текста.

3. Выявление и анализ СЕ, организованных с помощью имплицитных ГСС.

Выделение и анализ СЕ, организованных с помощью ПСС, является, как было показано выше, формально разрешимой задачей. Будет ли реальным решение аналогичной задачи относительно единиц, порожденных имплицит-

ными сверхфразовыми связями? Для выявления в связном тексте глубинно-синтаксических единиц необходимо разработать аппарат формализации ГСС, т.е. для любого предложения S_i нужно уметь определить, находится ли оно с окружающими предложениями в отношении, которое может быть трансформировано в ПСС посредством восстановления в S_i элидированного формального показателя сверхфразовой связи. Учитывая универсальный характер ЛК, можно свести сформулированную выше задачу к проблеме формального восстановления опущенных коннектовор, т.е. к устранению их эллипсиса (Берзон В.Е. и др., 1982).

Прежде всего следует принять во внимание, что эллипсис коннектора, указывающего на наличие сверхфразовой связи, должен компенсироваться в предложении какими-то иными средствами*. В соответствии с приведенной выше квалификацией входящие в одну подгруппу ЛК выражают близкие по смыслу отношения между предложениями. Тогда каждому классу ЛК должен соответствовать особый набор дополнительных показателей связи, употребляемых при эллипсисе ЛК. В частности, эллипсис противительных ЛК обычно сопровождается лексическим противопоставлением и синтаксическим параллелизмом в связанных предложениях S_i и S_h, например: При поверхностном диалоге осуществляются редактирование и отладка программ, составляемых аналитиком на языках программирования высокого уровня (S_i, прим. авт.). При глубинном диалоге аналитик вмешивается непосредственно в ход решения задачи (S_h, прим. авт.) (Г.С. Поспелов, А.М. Разин, Проблема интеграции... - НТИ, сер. 2, 1981, № 4, с. 6).

Эллипсис противительного ЛК но в предложении S_i компенсируется лексическим противопоставлением при поверхностном диалоге - при глубинном диалоге.

Таким образом алгоритм перехода от ГСС и ПСС может быть осуществлен на основе трансформационных правил (ТП), выработанных на основе сравнения синтаксических структур и лексического наполнения синонимичных предложений, одни из которых содержат определенный лексический коннектор, а в других эллипс этого ЛК компенсируется лексико-синтаксическими средствами. Анализ текста с целью формального выявления в нем с помощью ТП глубинно-синтаксической сверхфразовой связи состоит из следующих этапов.

1. Поиск в тексте синтаксически насыщенных предложений.

2. Проверка синтаксически насыщенного предложения S_i на эллипс коннекторов, с соблюдением следующих условий:

а) при проверке эллипсиса анафорических коннекторов производится последовательное сравнение синтаксических структур и лексического наполнения со структурами и наполнениями предложений, находящимися на определенном количестве шагов влево от S_i;

* Ср. задачу передачи определенного артикля средствами русского языка при переводе с западно-европейских или балканских языков.

б) при проверке эллипсиса катафорических коннекторов осуществляется аналогичное сравнение S_i с последующими предложениями.

3. В случае выявления эллипсиса коннектора устанавливается глубинно-синтаксическая связь между S_i и S_h (S_j). ТП описанного вида используются в лингвистическом автомате, производящем автоматическое индексирование научно-технических текстов на основе анализа их рефератов. Здесь в поисковый образ документа включаются лишь те ключевые слова (КС), частота употребления которых в анализируемом тексте превышает некоторую пороговую величину, зависящую от объема реферата. Однако при подсчете статистических "весов" КС часто не учитываются анафорические замещения слова или словосочетания A, входящего в предложение S_h, существительным N, входящим в предложение S_i, которое связано с S_h ГСС. Это явление сильно искажает картину соотношения весов КС и отрицательно сказывается на последующем информационном поиске в ИПС. Разработана процедура, которая обеспечивает лингвистическому автомату возможности учета подобных случаев с последующим увеличением "веса" соответствующих КС. Она реализируется в виде последовательности следующих операций:

1. Поиск в тексте потенциального заместителя, т.е. слова, зафиксированного в тезаурусе, не имеющего признака глагольности и не оформленного указательным местоимением.

2. Поиск в предшествующих предложениях антецедента (A) т.е. замещенного слова или синтаксической вершины замещенного словосочетания. Слово считается антецедентом, если значение замещающего слова шире, чем значение A, либо A совпадает с N (формообразующие флексии при сравнении отсекаются) причем множество слов, синтаксически зависящих от N, является подмножеством слов, зависящих от A.

3. Определение фактического количества вхождений в текст КС.

Для иллюстрации описанной процедуры воспользуемся следующим текстом, представляющим собой реферат научной статьи: ... исследованы свойства материалов, полученных п р е с с о в а н и е м электролитического Cu - порошка крупностью - 10 меш., распыленнного Al - порошка крупностью 100 меш. и карбонильного N_i - порошка со средним резмером частиц 3/4 мкм. П р е с с о в а н и е проводили под давлением 3,5 тс/см^2 с последующим спеканием при 950 град. С в течение 4 часов в вакууме. (РЖ "Металлургия" 1980, 3 Г 481).

Формальный анализ отрывка с помощью ЛА на выявление вхождений КС предусматривает следующие операции.

1. Во втором предложении реферата найдено слово прессование, зафиксированное в тезаурусе и не оформленное указательным местоимением. Это слово не получает признака глагольности (т.е. глагольного окончания).

2. В первом предложении для данного КС найден антецедент прессованием. У потенциального заместителя нет

ИСХОДНЫЙ ТЕКСТ

ARTIFICIAL INTELLIGENCE **
PLATH W . J . , IBM T SES AND DEVLOP . VOL 20 . NO 4 . JULY 1976 **
REQUEST : A NATURAL LANGUAGE QUESTION-ANSWERING SYSTEM **
BASIC DESIGN FEATURES **

O THE FIRST BASIC DESIGN FEATURE , THE USE OF RESTRICTED NATURAL ENGLISH , IS D
CTATED BY THE REALITIES OF THE PRESENT STATE OF THE ART OF FORMAL DESCRIPTION OF
NATURAL LANGUAGES :
SPECIFICALLY , THE FACT THAT NOTHING REMOTELY APPROACHING A COMPLETE GRAMMAR OR
SEMANTICS OF ALL OF ENGLISH (OR OF ANY OTHER NATURAL LANGUAGE) EITHER EXISTS
NOW OR APPEARS LIKELY TO MATERIALIZE IN THE NEAR FUTURE .
THE REQUEST APPROACH TO RESTRICTED ENGLISH (WHICH IS SIMILAR IN CERTAIN RESPECT
S TO THOSE ADOPTED IN THE SYSTEMS OF WOODS [6] AND WINOGRAD [5]) INVOLVES SHARP
LY LIMITING THE SEMANTIC SCOPE OF THE ENGLISH MATERIAL TO BE COVERED .
THIS IS ACCOMPLISHED BY FOCUSSING ON ONE RELATIVELY WELL DEFINED UNIVERSE OF DIS
COURSE AT A TIME , FOR EXAMPLE , THE "WORLD" OF A BUSINESS STATISTICS DATA BASE

HAVING THUS GREATLY RESTRICTED WHAT THE USER CAN "CONVERE" WITH THE COMPUTER ABO
UT , WE THEN SEEK TO PROVIDE HIM WITHIN THAT DOMAIN WITH A FLEXIBILITY O SYNTACTI
C AND LEXICAL EXPRESSION APPROACHING THAT OF NORMAL INGLISH .
O RESTRICTING A NATURAL LANGUAGE SUBSET IN THE MANNER JUST DESCRIBED HAS TWO MAJ
OR ADVANTAGES :
FIRST , IT REDUCES THE SEMANTIC UNIVERSE THAT MUST BE HANDLED TO A SIZE THAT IS
POTENTIALLY TRACTABLE FOR PURPOSES OF FORMAL ANALYSIS AND " UNDERSTANDING " BY A
COMPUTER ;
SECOND , IT LEADS DIRECTLY TO MAJOR NONARBITRARY REDUCTIONS IN THE RANGE OF VOCA
BULARY (AND , TO A LESSER EXTENT , THE RANGE OF SYNTACTIC CONSTRUCTIONS) THAT
MUST BE COVERED IN THE SUBSET , SINCE THERE IS NO NEED TO INCLUDE WORDS , CONSTR
UCTIONS , OR MEANINGS OF WORDS NOT RELATED TO THE SUBJECT MATTER THE USER WILL N
ECESSARILY BE DEALING WITH .
O IN A RESEARCH PROJECT SUCH AS THAT ON REQUEST , ONE CONCEIVABLE DRAWBACK OF A
NARROW SEMANTIC FOCUS IS THE POSSIBILITY THAT SOLUTIONS WORKED OUT FOR A SPESIFI
C DOMAIN OF DISCOURSE MAY NOT BE READILY EXTENDABLE TO OTHERS .
IN THE HOPE OF MINIMIZING SUCH DIFFICULTIES , WE HAVE CHOSEN TO WORK INIT ALLY W
ITH THE WORLD OF BUSINESS STATISTICS , BECAUSE IT APPEARS TO BE REPRESENTATIVE O
F A LARGE AND IMPORTANT FAMILY OF DATA BASES INVOLVING PERIODIC , NUMERICAL DATA

DESPITE THE EXISTENCE OF IDIOSYNCRATIC DIFFERENCES , THE VARIOUS MEMBERS OF THIS
FAMILY - SUCH AS WEATHER DATA , CENSUS DATA , AND PRICE AND WAGE STATISTICS -=
SHARE A BROAD RANGE OF SEMANTIC RELATIONSHIP , INCLUDING NOTIONS OF TIME , COMPA
RISON , AND VARIOUS HIGHER ORDER FUNCTIONS OF THE PRIMITIVE DATA (E G , SUMS ,
AVERAGES , RATIOS , RATES , MAXIMA , AND MINIMA) .
BECAUSE THESE SHARED SEMANTIC RELATIONSHIPS TEND TO BE EXPRESSED LINGUISTICALLY
IN A VERY SIMILAR MANNER FOR ALL OF THE DOMAINS IN QUESTION , THE PROSPECTS APPE
AR QUITE FAVORABLE FOR A SUBSTANTIAL CARRYOVER OF RESULTS FROM ONE CASE TO THE N
EXT .
O THE SECOND MAJOR FEATURE OF REQUEST'S APPROACH TO NATURAL LANGUAGE PROCESSING
IS THE TREATMENT OF INPUT QUERIES IN RESTRICTED ENGLISH AS HIGH-LEVEL-LANGUAGE E
XPRESSIONS THAT ARE TO BE COMPILED INTO EXECUTABLE CODE .
AS IN THE CASE OF COMPILERS FOR FORMAL LANGUAGES , THE PROCESS CONSISTS OF TWO D
ISTINCTIVE PHASES :
A PARSING PHASE , IN WHICH THE STRUCTURE OF THE INPUT LANGUAGE EXPRESSION IS DET
ERMINED , AND A TRANSLATION (O- SEMANTIC INTERPRETATION) PHASE , IN WHICH THE
RESULTING STRUCTURAL DESCRIPTION IS MAPPED INTO OBJECT LANGUAGE CODE .
IN REQUEST , THE MECHANICS OF THE LATTER PROCESS CLOSELY RESEMBLE THOSE EMPLOYED
IN CONVENTIONAL COMPILERS , IN THAT THEY ARE BASED ON A SCHEME ORIGINALLY PROPO
SED BY KNUTH [4] AS A GENERALIZATION OF STANDARD SYNTAX-DIRECTED TRANSLATION TECH
NIQUES .
A SIMILAR DEGREE OF CORRESPONDENCE DOES NOT EXIST FOR THE PARSING PHASE , HOWEVE
R , BECAUSE WE HAVE DESIGNED IT AROUND TRANSFORMATIONAL GRAMMAR , A FORM OF LING
UISTIC DESCRIPTION THAT DIFFERS MARKEDLY FROM ANYTHING USED IN COMPILERS FOR FOR
MAL LANGUAGES .
O THE THIRD BASIC DESIGN FEATURE OF REQUEST , EMPLOYMENT OF LINGUISTIU ANALYSIS
BASED ON A TRANSFORMATIONAL GRAMMAR , WAS ADOPTED IN AN ATTEMPT TO DEAL WITH THE
COMPLEXITY AND DIVERSITY THAT ARE CHARACTERISTIC OF EVEN RESTRICTED SUBSETS OF
NATURAL LANGUAGE AS WE HAVE DEFINED THEM .
THE KEY PROPERTIES OF A TRANSFORMATIONAL DESCRIPTION ARE : 1) THE DEFINITION OF
TWO DISTINCT LEVELS OF LINGUISTIC STRUCTURE - SURFACE STRUCTURE AND UNDERLYING S
TRUCTURE - AND , 2) THE SPECIFICATION OF A FORMAL MAPPING RELATING THEM .
THE USE OF SUCH GRAMMATICAL MODELS AND THEIR RELEVANCE TO THE PROBLEMS OF DEV
ELOPING USER-ORIENTED SUBSETS OF NATURAL LANGUAGE ARE NOW EXAMINED IN SOME DETA

BASIC DESIGN FEATURES

ВАРИАНТ I

```
00. □ THE FIRST BASIC DESIGN FEATURE , THE USE OF RESTRICTED NATURAL ENGLISH , IS D
    ICTATED BY THE REALITIES OF THE PRESENT STATE OF THE ART OF FORMAL DESCRIPTION
    OF NATURAL LANGUAGES :
01. (M;ПОЯСНЕНИЕ)
02. (L;ПОВТОР)
03. (L;ПОВТОР)
04. (C;СЛЕДСТВИЕ)
05. □ RESTRICTING A NATURAL LANGUAGE SUBSET IN THE MANNER JUST DESCRIBED HAS TWO MA
    JOR ADVANTAGES :
06. (M;ПОЯСНЕНИЕ)
07. (L;ПОВТОР)
08. (L;ПОВТОР)
09. (L;ПОВТОР)
10. (L;ПОВТОР)
11. (L;ПОВТОР)
12. □ THE SECOND MAJOR FEATURE OF REQUEST'S APPROACH TO NATURAL LANGUAGE PROCESSING
    IS THE TREATMENT OF INPUT QUERIES IN RESTRICTED ENGLISH AS HIGH-LEVEL-LANGUAGE
    EXPRESSIONS THAT ARE TO BE COMPILED INTO EXECUTABLE CODE .
13. AS IN THE CASE OF COMPILERS FOR FORMAL LANGUAGES , THE PROCESS CONSISTS OF TWO
    CONSECUTIVE PHASES :
14. (M;ПОЯСНЕНИЕ)
15. (L;ПОВТОР)
16. (R;ПРОТИВОПОСТАВЛЕНИЕ)
17. □ THE THIRD BASIC DESIGN FEATURE OF REQUEST , EMPLOYMENT OF LINGUISTIU ANALYSIS
    BASED ON A TRANSFORMATIONAL GRAMMAR , WAS ADOPTED IN AN ATTEMPT TO DEAL WITH T
    HE COMPLEXITY AND DIVERSITY THAT ARE CHARACTERISTIC OF EVEN RESTRICTED SUBSETS
    OF NATURAL LANGUAGE AS WE HAVE DEFINED THEM .
18. THE KEY PROPERTIES OF A TRANSFORMATIONAL DESCRIPTION ARE 1 ) THE DEFINITION OF
    TWO DISTINCT LEVELS OF LINGUISTIC STRUCTURE - SURFACE STRUCTURE AND UNDERLYING
    STRUCTURE - AND  2 ) THE SPECIFICATION OF A FORMAL MAPPING RELATING THEM .
19. (L;ПОВТОР)

                                                    15.58.59,DURATION 00.01.20
```

Рис. 5б

ВАРИАНТ 2

BASIC DESIGN FEATURES

```
00. □ THE FIRST BASIC DESIGN FEATURE , THE USE OF RESTRICTED NATURAL ENGLISH , IS D
    ICTATED BY THE REALITIES OF THE PRESENT STATE OF THE ART OF FORMAL DESCRIPTION
    OF NATURAL LANGUAGES :
05. □ RESTRICTING A NATURAL LANGUAGE SUBSET IN THE MANNER JUST DESCRIBED HAS TWO MA
    JOR ADVANTAGES :
12. □ THE SECOND MAJOR FEATURE OF REQUEST'S APPROACH TO NATURAL LANGUAGE PROCESSING
    IS THE TREATMENT OF INPUT QUERIES IN RESTRICTED ENGLISH AS HIGH-LEVEL-LANGUAGE
    EXPRESSIONS THAT ARE TO BE COMPILED INTO EXECUTABLE CODE .
13. AS IN THE CASE OF COMPILERS FOR FORMAL LANGUAGES , THE PROCESS CONSISTS OF TWO
    CONSECUTIVE PHASES :
17. □ THE THIRD BASIC DESIGN FEATURE OF REQUEST , EMPLOYMENT OF LINGUISTIU ANALYSIS
    BASED ON A TRANSFORMATIONAL GRAMMAR , WAS ADOPTED IN AN ATTEMPT TO DEAL WITH T
    HE COMPLEXITY AND DIVERSITY THAT ARE CHARACTERISTIC OF EVEN RESTRICTED SUBSETS
    OF NATURAL LANGUAGE AS WE HAVE DEFINED THEM .
18. THE KEY PROPERTIES OF A TRANSFORMATIONAL DESCRIPTION ARE 1 ) THE DEFINITION OF
    TWO DISTINCT LEVELS OF LINGUISTIC STRUCTURE - SURFACE STRUCTURE AND UNDERLYING
    STRUCTURE - AND  2 ) THE SPECIFICATION OF A FORMAL MAPPING RELATING THEM .

                                                    11.36.51,DURATION 00.01.21
```

Рис. 5в

зависимых слов, тогда как у антецедента зависимые есть. На основании полученной информации ЛА делает вывод о том, что во втором предложении опущен ЛК это (ср. Это прессование проводили...), а значит именная группа прессованием электролитического порошка замещена существительным прессование. Иначе говоря, во втором предложении имеется в виду то же прессование, что и в первом.

3. Всем замещенным терминам, зависимым от слова прессование приписывается на одно вхождение в текст больше, чем при первоначальном подсчете "веса" КС. В частности, количество вхождений термина электролитический принимается равным 2, хотя это КС в явном виде употреблено в реферате только один раз.

Следует отметить, также, что разработка указанных трансформационных правил должна позволить выявить в тексте не только СФЕ, организованные с помощью глубинно-синтаксической связи, но и СФЕ смешанного типа. В качестве иллюстрации рассмотрим следующий фрагмент научного текста: Поскольку дальнейший ход истории нам сегодня известен, мы можем вместе с автором (2) убедиться, сколь же не соответствует экстраполяционный прогноз реальному течению исторических событий (S_0 - прим. авт.). Один из рассказов Анатолия Франса построен на эффекте ложного прогноза, который делает римлянин Галлион, исполняющий должность проконсула Греции в начале правления императора Нерона (S_1 - прим. авт.). Отзвуки доходящих до него споров Павла из Тарсиса с коринфскими иудеями кажутся ему явлениями, не представляющими серьезного интереса для судеб Римской империи (S_2 - прим. авт.). Эти споры, с точки зрения римского аристократа Галлиона, относятся к миру трущоб и отбросов общества (S_3 - прим. авт.). Рассказ кончается рассуждением Галлиона о предстоящем возникновении новой имперской религии ... (S_4 - прим. авт.) (Ю.А. Шрейдер, Методологические проблемы ... - НТИ, сер. 2, 1982, № 5, с. 2).

Проанализовав сверхфразовые синтаксические связи в отрывке, построим структуру, изображенную на рис. 6.

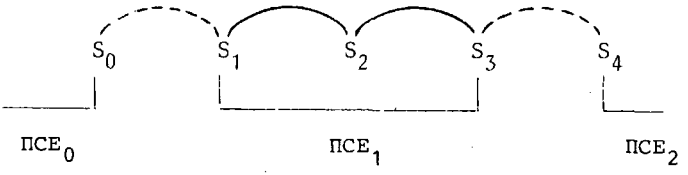


Рис. 6.

Если же принять во внимание имплицитную анафорическую связь S_4 и S_1, сопровождаемую эллипсисом ЛК этот (ср. Рассказ кончается ... - (Этот) рассказ кончается), то получим более глубокое представление того же текста (рис. 7).

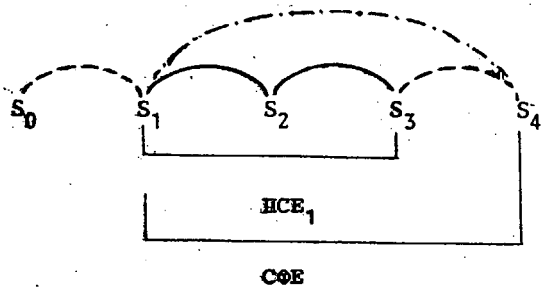


Рис. 7.

Итак, учет глубинно-синтаксической связи позволяет выделить в приведенном фрагменте сферхфразовое единство смешанного типа, сформированное из предложений S_1, S_2, S_3, S_4, которое, в свою очередь, состоит из двух сверхфразовых единиц: поверхностно-синтаксический (S_1, S_2, S_3) и глубинно-синтаксический (S_1, S_4).

4. Заключение. За двадцать лет исследований в области теории и практики ЛСТ языковедам удалось выделить основные понятия и наметить главные направления исследований. В то же время большинство работ носит здесь созерцательно-описательный характер, и поэтому выработка технологии этих исследований, точных приемов обнаружения основных единиц сверхфразового уровня и методов работы с ними находятся пока в зачаточном состоянии. Между тем, современное языкознание, вступающее в период конструктивного изучения и воспроизведения языка и речи, требует создания во всех своих областях объективной, однозначной и эффективной методики анализа конкретного материала. Это требование времени имеет прямое отношение и к ЛСТ: без конструктивных исследовательских приемов лингвистика текста останется лишь сводом более или менее остроумных предположений. Задача настоящей статьи как раз и состояло в том, чтобы, идя по пути разработки конкретной технологии, наметить методы выделения основных критерий, связей и единиц сверхфразового уровня, а также определить ту экспериментальную методику, с помощью которой можно было бы объективно проверять выдвигаемые гипотезы.

ЛИТЕРАТУРА

Адмони В.Г. Синтаксис современного немецкого языка. - Л.: Наука, ЛО, 1973.

Бархударов Л.С. Структура предложения и структура текста. - В кн.: Лингвистические проблемы текста. Сборник научных трудов МГПИИЯ им. М.Тореза, вып. 158. - М.: МГПИИЯ им. М. Тореза, 1980.

Берзон В.Е., Блехман М.С., Ефремова А.А., Захаров А.А., Иванова Т.В., Полонская О.Р. Об одном подходе к разработке ЛО и МО автоматического анализа сверхфразового уровня языка. - в кн.: Переработка текста методами инженерной лингвистики. Тезисы докладов всесоюзной конференции. - Минск: МГПИИЯ, 1982, с. 50-53.

Вайнрих Х. Текстовая функция французского артикля. Перевод с англ. - Новое в лингвистике. Выпуск УШ, Лингвистика текста. М.: Прогресс, 1978, с. 370-387.

Гарвин П.Л. Метод фулькрумов - 12 лет спустя. - Международный форум по информации и документации, 1980, Т. 5, № 2, с. 24-26.

Гиндин С.И., Леонтьева Н.Н. Проблемы анализа и синтеза целого текста в системах машинного перевода, диалоговых и информационных системах. - ВЦП. Обзорная информация. Сер. 2. Машинный перевод и автоматизация информационных процессов. М., 1978.

Ельмслев Л. Пролегомены к теории языка. Перевод с англ. - Новое в лингвистике. Выпуск 1. - М.: Изд-во иностранной литературы, 1960.

Хьюз Дж., Митчом Дж. Структурный подход к программированию. Перевод с англ. - М.: Мир, 1980.

Зотов А. Соотношение фундаментальных и прикладных исследований - актуальные аспекты. - Коммунист, 1979, № 10, с. 60-63.

Леонтьева Н.Н. Семантика связного текста и единицы информационного анализа. - НТИ, серия 2, 1981, № 1, с. 21-29.

Лурия А.Р. Язык и сознание. - М.: Изд-во МГУ, 1979.

Маслов Б.А. Проблемы лингвистического анализа связного текста /надфразовый уровень/. - Таллин: Изд-во Таллинского пединститута, 1975.

Откупщикова М.И. Синтаксис связного текста. - Л.: ЛГУ, 1982.

Попов Э.В. Общение с ЭВМ на естественном языке. - М.: Наука, Главная редакция физико-математической литературы, 1982, с. 179-180.

Пиотровский Р.Г. Очерки по грамматической стилистике французского языка. - М.: Издательство литературы на иностранных языках, 1956.

Пиотровский Р.Г. Очерки по стилистике французского языка. Л.: Учпедгиз, 1960, с. 50 и сл.

5

Пиотровский Р.Г. Лингвистические аспекты "искуственного разума". - ВЯ, 1981, № 3, с. 36-37.
Рафаэл Б. Думающий компьютер. Перевод с англ. - М.:Мир, 1979, с. 351-354.
Ревзин И.И. Структура языка как моделирующей системы.- М.: Наука, 1978.
Реферовская Е.А. Лингвистические исследования структуры текста. - Л.: Наука, ЛО, 1983.
Супрун А.Е. Лекции по языковедению. - Минск: Изд-во ЮГУ им. В.И. Ленина, 1978.
Шкирич А.А.,Берзон В.Е., Блехман М.С. Об одной функции определенного артикля в английском языке. - Романское и германское языкознание. Выпуск 1. Вопросы экспериментальной фонетики и прикладной лингвистики (Сборник научных статей). - Минск: Изд-во МГПИИЯ, 1979, с. 161-166.
Brainerd B. Article use as an Indicator of Style in English Language. - In: Linguistik und Statistik /Hrsg. von S. Jager. - Braunschweig: Vieweg, 1972, S. 11-32.
Piotrowski R. Folosirea articolului hotărît de către scritorii români. - Omagiu lui Alexandru Rosetti la 70 de ani. - Bucureşti: Editura Academiei Republicii Socialiste România, 1965, p. 693-696.

CONNECTIONS, UNITS AND UNITIES ON THE INTERSENTENCE LEVEL

V. Berzon, M. Blekhman, R. Piotrovski

Summary

The paper deals with intersentence relations in scientific texts. Distinguished are explicit and implicit connections, the former being based on the so-called lexical ties, or connectors. Russian and English ties are classified from the point of view of their logical meaning. Linguistic procedures are suggested for explicating intersentence units and unities. Natural language processing systems are outlined based on the procedures introduced; their work is illustrated. Results are discussed and admitted to be promising.

ЗАКОН ЦИПФА-МАНДЕЛЬБРОТА И ЕДИНИЦЫ РАЗЛИЧНЫХ УРОВНЕЙ ОРГАНИЗАЦИИ ТЕКСТА

М.Г. Борода, А.А. Поликарпов

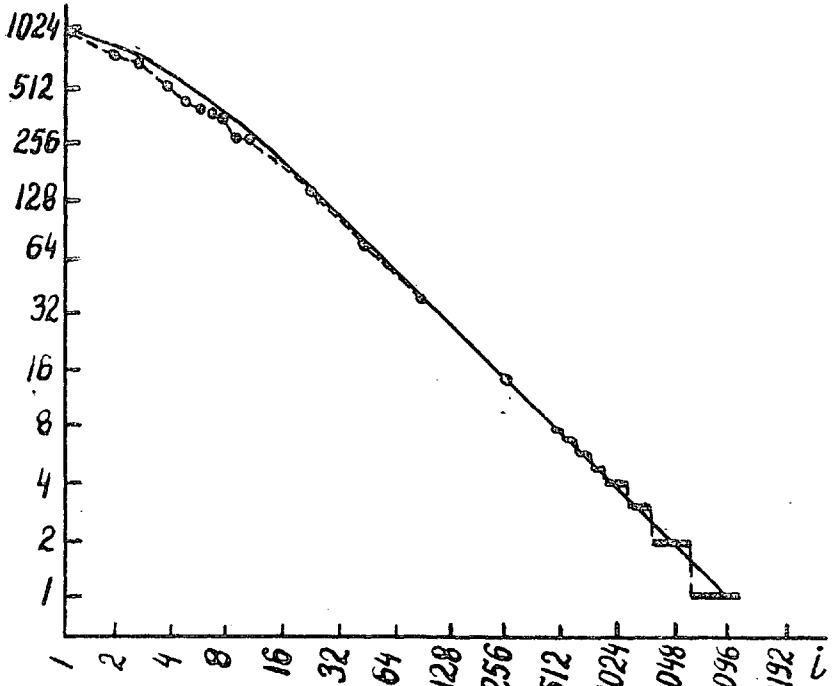
1. Исследование общих принципов организации текста количественными методами привлекает к себе возрастающее внимание специалистов различных профилей - от лингвистов, искусствоведов и психологов до математиков. Значительное место здесь занимают исследования организации повторяемости в тексте элементов того или иного рода, в особенности, работы, связывающие эту организацию с выполнением в тексте закона Ципфа-Мандельброта. В последние годы здесь был получен ряд новых результатов, выявлены важные закономерности внутреннего устройства текста (в особенности - художественного), показана связь организации повторяемости малых элементов в нем с его целостностью, законченностью, эстетической значимостью (Орлов, 1970; 1975; 1982; Борода, 1974; 1977; 1979; 1981), с типом языка, на котором написан текст (Поликарпов, 1976; 1979), построены математические модели этой организации (Орлов, 1970; 1976; Арапов, Ефимова, Шрейдер, 1975; Тулдава, 1979; 1980; и др.).

Конкретно, исследование организации повторяемости слов в тексте литературного произведения, проведенное в названных работах Ю.К. Орлова, показало, что организация эта существенно связана с объемом текста (числом N_0 всех словоупотреблений в нем) и относительной частотой p_{max} самого частого слова в нем и что связь эта реализуется через закон Ципфа-Мандельброта

$$p_i = \frac{K}{(B+i)^\gamma} \qquad (1)$$

при $K = 1/\ln(Zp_{max})$, $B = Kp_{max}^{-1} - 1$, $\gamma = 1$ и Z, равном объему N_0 полного текста. В частности, упорядоченный по невозрастанию набор $\{p_i\}$ относительных частот встреч различающихся слов в тексте (объема, как правило, не менее 10 000 и не более 100 000 словоупотреблений - подробнее об этом - ниже) описывался выражением (1) при названных значениях K, B и γ (Рис. 1). Словарный запас V такого текста и число V_m различающихся слов, каждое из которых встретилось в нем m раз, описывались следствиями (1) - выражениями соответственно

$$V(Z) = \frac{Z - p_{max}^{-1}}{\ln(Zp_{max})} \qquad (2)$$



А.С. Пушкин. Капитанская дочка. Описание частотной структуры текста законом (1) при γ =1 и $Z = N_0$. По горизонтальной оси отложены номера (ранги) частот, по вертикальной - частоты соответствующих рангов. Фактические значения частот даны черными кружками, теоретическая кривая, построенная по (1) - **сплошной линией.** Рисунок заимствован из работы (Орлов, 1975).

и

$$V_m(Z) = \frac{V(Z)}{m(m+1)} \qquad (3)$$

Относительная погрешность оценки реально наблюдаемых в тексте значений V и V_m этими выражениями не превышала 20-25 процентов, а в ряде случаев была заметно меньшей (Орлов, 1976 и др.).

Аналогичные результаты были получены в исследовании организации повторяемости малых мелодических единиц в музыкальном тексте. Именно, используя выделенную в работе (Борода, 1973) однозначно определенную мелодическую единицу "Формальный мотив" (F-мотив), позволяющую членить любую мелодию с тактовой структурой строго алгоритмически и без пропусков на некоторые естественные "микросегменты" (Рис. 2), и полагая два F-мотива одинаковыми в том и только том случае, когда один из них **можно** было получить из другого сдвигом по высоте на некоторый - в частности, нулевой- интервал, удалось показать, что основные характеристики организации повторяемости F-мотивов в музыкальном тексте - соотношение объема F-мотивного словаря текста и его длины (числа N_0 всех употреблений F-мотивов в нем), связь частоты m и числа V_m различающихся F-мотивов, встретившихся в тексте ровно m раз каждый, наконец, закономерности "устройства" ранжированного по невозрастанию набора частот встреч F-мотивов в тексте - описываются соотношениями (1)-(3) с той же примерно точностью, что и для слов в случае литературного текста* (Борода, 1974; 1979; 1981, и др.) - см. Рис. 3.

---

* Отметим, что данная форма закона Ципфа-Мандельброта - ниже, в основном тексте, будем называть ее "канонической"- является частным случаем выведенного в работе (Орлов, 1970) "обобщенного закона Ципфа-Мандельброта". Другая форма этого закона возникает при γ, отличном от единицы и нуля, когда параметры K и B даются выражениями

$$K = [(1-\alpha)/(\alpha(p_{max}^{1-\alpha} - Z^{\alpha-1}))]^{1/\alpha}; \quad B = K^{\alpha} p_{max}^{-\alpha}; \quad \alpha = 1/\gamma \qquad (1a),$$

объем словаря - выражением

$$V = V_\alpha(Z) = K^{\alpha}(Z^{\alpha} - p_{max}^{-\alpha}) + 1 \qquad (2a)$$

и объемы групп m-разовых элементов в тексте - выражением

Как можно видеть из соотношений (1)-(3), выполнение закона Ципфа-Мандельброта в тексте в указанной выше форме, при данном значении параметров K, B и γ, при Z, равном объему N_0 полного текста связано с особой уравновешенностью структуры повторяемости элементов в тексте, когда, например, число m-разовых элементов убывает с увеличением m (это особенно заметно для малых значений m - см. графики на Рис. 1 и 3 со ступеньками, длина которых увеличивается с уменьшением частоты m) или когда число элементов, встречающихся в тексте по одному разу каждый - неповторяющихся в нем - составляет половину словарного словарного запаса этого текста, т.е. оказывается равным общему числу различных повторяющихся элементов в нем (см. об этом: Орлов, 1975; Борода, 1979). Закономерно, что выполнение этого закона в художественном тексте оказалось связанным с важными эстетическими его свойствами: уравновешенностью формы литературного и музыкального произведения, взаимоотношением (взаимной корректировкой) по повторам между макро- и микро-уровнями организации музыкального текста, проявлением репризности и контраста в организации музыкального цикла типа сонаты, симфонии, прелюдии и фуги, наконец, с законченностью текста. Как показали исследования, выполнение закона Ципфа-Мандельброта в "канонической форме" в литературном или музыкальном тексте существенно зависит от того, является ли этот текст законченным, полным. Например, выполняясь в исследованных циклических музыкальных произведениях, закон этот не выполнялся, как правило, на их отрывках, даже таких тематически самостоятельных и замкнутых по форме, как часть цикла. Не выполнялся он и на отрывках литературных произведений, имевших на своем полном тексте "каноническую" ципфо-мандельбротовскую структуру повторов слов. Словарный (F-мотивный) запас таких отрывков и число m-разовых слов (F-мотивов) в них плохо прогнозировались выражениями (2)-(3) - отклонения фактических значений от их прогнозов составляли 50-60 процентов и выше; ход убывания частот слов (F-мотивов) в

$$V_m = V_{m,\alpha}(Z) = V_\alpha(Z)\left(m^{-\alpha} - (m+1)^{-\alpha}\right) \qquad (3a).$$

Как отмечается в названной работе, предварительные исследования показали связь значения γ с типом структурной единицы, на уровне которой текст рассматривается. К сожалению, дальнейшие исследования в этом плане (начатые Ю.К. Орловым на уровне букв и несколько "механических" единиц типа диграмм и триграмм и, совместно с одним из авторов данной работы, на уровне мелодических интервалов и их последовательностей - также несколько механически членящих мелодию) не проводились, и полученные в них результаты остались, в целом, на уровне предварительных. Ниже мы вернемся к этому вопросу.

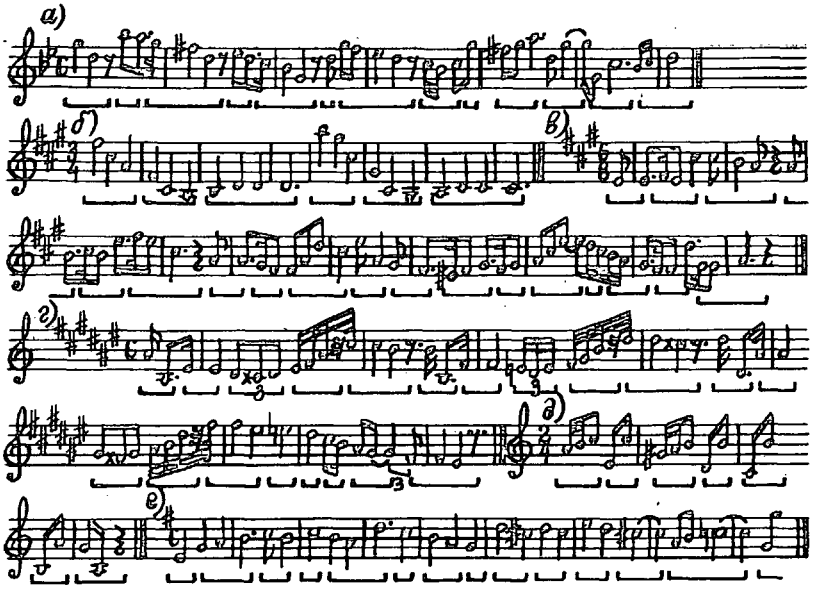


Рис. 2. Сегментация мелодических отрывков из музыкальных произведений различных стилей на F-мотивы: а) Д. Тартини. Соната для скрипки с ф-но "Покинутая Дидона", ч. Ш; б) И. Гайдн. Симфония № 45 ("Прощальная"), ч. 1; в) Ф. Шуберт. Весенний сон; г) А. Скрябин. Этюд ор.8 № 12; д) Д. Шостакович. Фуга ор. 87 № 2; е) Д. Шостакович. Фуга ор. 87 № 4. F-мотивы выделены скобками ⌒. Отметим, что при членении мелодии на F-мотивы звук, за которым следует пауза, удлиняется за счет этой паузы (ср. ниже с Рис. 4).

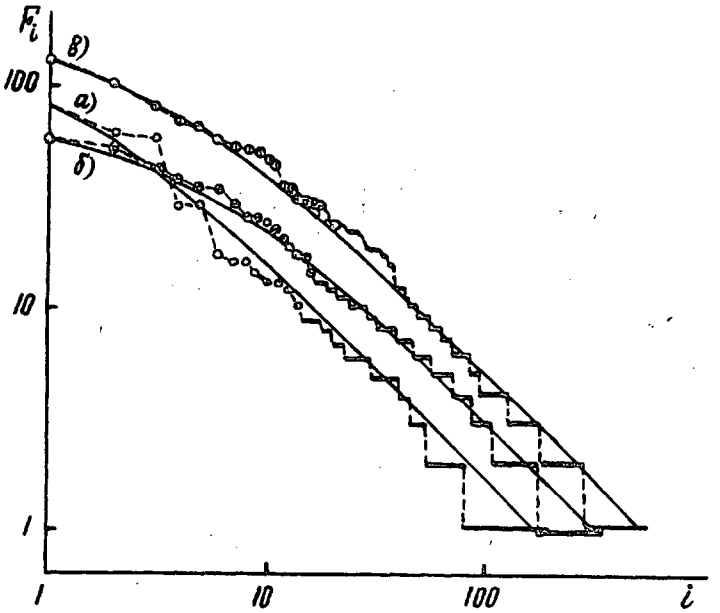


Рис. 3. Описание частотной структуры музыкальных текстов законом Ципфа-Мандельброта (1): а) Г.Ф.Гендель. Фуга № 2 для органа; б) И. Гайдн. Симфония № 45 ("Прощальная"); в) Ф. Шопен. Соната № 3. Обозначения те же, что на Рис. 1. Маштаб логарифмический.

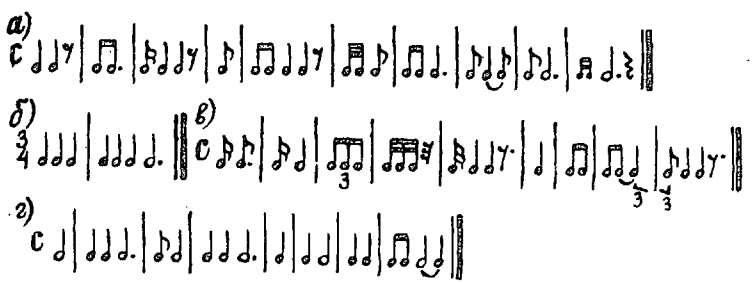


Рис. 4. Выделение ритмических F-мотивов в мелодии. На рисунке приведены словари ритмических F-мотивов для некоторых мелодических отрывков из Рис. 2: а) Д. Тартини. Соната "Покинутая Дидона", ч. Ш; б) И. Гайдн. "Прощальная симфония", ч. 1; в) А. Скрябин. Этюд ор. 8 № 12; г) Д. Шостакович. Фуга ор. 87 № 4.

отрывке, как правило, плохо соответствовал выражению (1) при Z, равном длине отрывка и названных выше K, B и γ.

Таким образом, проведенные исследования литературных текстов - в подавляющем большинстве объемом от 10^4 до 10^5 словоупотреблений, т.е. не малых, но и не очень крупных, связанных с жанром повести или небольшого романа - и музыкальных текстов относительно крупной формы /небольшая симфония, инструментальная соната, прелюдия и фуга, и т.д.) с объемами в среднем около 1000 F-мотивоупотреблений (разброс объемов исследованных музыкальных текстов был также примерно десятикратным - от 200 до 2300 F-мотивов, с "упором" на тексты объемом в 800-1200 F-мотивов) выявили важную о б щ н о с т ь внутренней организации литературных и музыкальных текстов как целостных "художественных сообщений" на уровне малых структурных единиц - слов в литературном тексте, F-мотивов - в музыкальных.

С другой стороны, проведенные Ю.К. Орловым исследования показали, что крупные литературные тексты с объемом более 10^5 словоупотреблений - связанные, как правило, с жанром романа или большой повести - нередко выполняют закон Ципфа-Мандельброта в "канонической форме" не на своем полном объеме (т.е. не при $Z=N_0$), а на отдельных частях; структура повторов слов в относительно коротких текстах объемом менее 10^4 словоупотреблений также не описывалась законом Ципфа-Мандельброта в его "канонической форме", напоминая, как отмечает Ю.К. Орлов, частотную структуру отрывков "ципфо-мандельбротского текста" (как отмечается в /Орлов, 1978/, "частотная структура коротких текстов похожа на структуру отрывков из более длинных текстов, выполняющих на своем полном объеме закон Ципфа-Мандельброта" - цит. изд., с. 65). Все это позволило Ю.К. Орлову говорить о неоднозначности прогноза частотной структуры для крупных текстов ("может быть, ципфовский объем (объем, на котором текст выполняет закон Ципфа-Мандельброта в "канонической форме" - М.Б. и А.П.) совпадает с объемом полного текста, может быть в длиной части" (там же, с. 65)), невозможности этого прогноза для относительно коротких текстов и неясности границ "промежуточной зоны" (10 000 + ?), в которой тексты следуют закону (1) в его "канонической форме" не всегда /там же/.

Сложность ситуации с выполнением закона Ципфа-Мандельброта (1) при $Z=N_0$ явилась одним из источников более общей модели частотной структуры лексики, разработанной в (Орлов, 1978). В рамках этой модели предполагается, что для любого текста T объема N_0 существует текст T' объема Z, $Z=N_0$, $Z<N_0$ или $Z>N_0$, выполняющий закон (1) в его "канонической форме" и такой, что словарный запас V текста T и число V_m m-разовых слов в T - т.е. основные характеристики организации повторов слов в T - зависят функционально от словарного запаса $V(Z)$ текста T' и отношения $x=\frac{Z}{N_0}$:

$$V = V(N_0|Z) = V(Z) \cdot \frac{\ln x}{x-1} \qquad (4),$$

$$V_m = V_m(N_0|Z) = V(Z) \cdot \sum_{i=m}^{\infty} \frac{(1-x)^{i-m}}{i(i+1)} \qquad (5).$$

Исследование реальной частотной структуры литературных текстов различных стилей и временной принадлежности в рамках этой модели (с определением значения Z по фактическому словарному запасу исследуемого текста T, так чтобы оно отличалось от значения, определенного по (4), не более, чем на одно слово) показали как ее хорошие прогнозирующие возможности для текстов различного объема (некоторые "но" в этом плане отметим ниже), так и способность ее объяснить ряд накопившихся в лингвостатистике и связанных с трактовой закона Ципфа-Мандельброта как "речевого феномена" парадоксов. С другой стороны, отношение Z к фактическому объему исследуемого текста могло даже для априори высокохудожественных и целостных литературных произведений заметно отличаться от единицы (Орлов, 1978) и связь между Z и N_0 проявилась лишь при исследовании корреляции их логарифмов (выявившем существенную положительную корреляцию $\lg Z$ и $\lg N_0$ для текстов этого круга). Кроме того, как отмечено в названной работе Ю.К. Орлова, "для всех текстов и выборок, словарный запас которых превышает 1000 слов, обнаружилось значимое з а в ы ш е н и е выражением (5) числа одноразовых слов в среднем по всем текстам на 8 %"; такое же по величине з а н и ж е н и е обнаружилось на трехразовых словах)Орлов, 1976, с. 74); такого рода систематические смещения прогноза в важнейшей для структуры повторов редкочастотной зоне словаря текста вызывают некоторую настороженность к "модели Z ". Наконец, существенной содержательной сложностью для этой модели - точнее, ее свойством, вызывающим, как кажется авторам данной статьи, наибольшие содержательные возражения - является тот факт, что в ее рамках относительно короткие законченные тексты, связанные с жанром рассказа, небольшой повести трактуются как о т р ы в к и: организация повторяемости слов в этих текстах описывается "моделью Z " при значении Z, заметно большем полной длины этого текста. Трактовка высокохудожественного текста "малого жанра" как отрывка в смысле организации повторов - важнейшей составляющей художественной формы - вызывает определенный интуитивный протест, тем более, что в случае высокохудожественных образцов этого жанра их авторами прилагаются особые усилия для достижения "гармонии формы", завершенности-

ти текста (литературоведение располагает многими свидетельствами такого рода).

Конечно, обсуждение достоинств и сложностей "Z-подхода" не является целью данной статьи. Однако, важно отметить, что само построение этого подхода в исследовании организации повторяемости слов в литературном тексте, в анализе выполнимости в нем закона Ципфа-Мандельброта существеннейшим образом связано с постулированием того, что в любом случае такая организация должна быть подчинена закону (1) и его следствиям - для $Z=N_0$ или $Z>N_0$, или $Z<N_0$ - при $\gamma=1$.

В настоящей статье показано, что содержательный смысл этого постулата и возможность отказа от "единственности значения γ" в исследовании структуры повторов композиционно значимых единиц в тексте (в частности, литературно-художественном тексте) связаны с у р о в н е м е д и н и ц ы, на котором **текст** рассматривается. Показано, что принятие гипотезы "множественности значений γ" - кстати, основанной и на соображениях, высказанных в работах Ю.К. Орлова - позволяет выявить неизвестные ранее изоморфизмы в организации литературного и музыкального текстов на их различных уровнях. Основная цель данной работы - еще раз подчеркнем, отнюдь не направленной на полемику с концепцией Z - описать первые результаты проведенных авторами исследований организации повторяемости единиц различных уровней в художественном тексте, ввести в обиход анализа новые структурные единицы и обсудить - или, по крайней мере, поставить - некоторые содержательные проблемы, возникающие в связи с многоуровневым анализом художественного текста, с исследованием того, как и при каких условиях выполняется в нем - или не выполняется - закон Ципфа-Мандельброта.

2. В первую очередь, отметим, что закон Ципфа-Мандельброта, как он исследован в художественных текстах на сегодняшний день, характеризует только один уровень иерархической организации текста - уровень слов в тексте литературном, уровень F-мотивов - в музыкальном. Выявленная значимость этого закона для литературных текстов на уровне слов априори естественна, ибо слово, по всей видимости, является центральной единицей номинации в языке - по крайней мере, в языках типа русского. Однако важную роль в любом литературном тексте играют и единицы других уровней языка - как более общие, чем слово (например, гиперлексемы - см. ниже), так и более конкретные, чем слово (лексико-семантические варианты). Единицы, более общие, чем F-мотив и генетически с ним связанные, могут быть выявлены и в музыкальных текстах - например, на ритмическом уровне их организации (см. ниже). С другой стороны, как уже было отмечено выше, прослеживается связь выполнения закона Ципфа-Мандельброта в "канонической форме" в текстах на лексическом уровне с их принадлежностью преимущественно к "средней" форме, с диапазоном текстового объема от 10^4 до 10^5 **словоупотреблений.**

В свете этих фактов возникают следующие вопросы:

а) если закон Ципфа-Мандельброта в его "канонической" форме не выполняется в литературном (музыкальном) тексте на уровне слов (F -мотивов), то значит ли это, что он не выполняется в тексте вообще, что повторяемость слов или F-мотивов в тексте регулируется не "ципфо-мандельбротовскими", а какими-то совсем другими принципами? Может быть, этот закон при γ=1 или, по крайней мере, при γ, близком к единице, выполняется на других единицах текста - также языково и композиционно значимых? С другой стороны, может быть, этот закон выполняется в тексте на уровне слов (F - мотивов) не в "канонической форме, при γ =1, а при γ, заметно отличном от единицы?

б) если **выполнение** закона Ципфа-Мандельброта на уровне слов в "канонической форме", с γ =1, связано **со "средним"** объемом произведения и, соответственно, со "средней" формой, если на "коротких" (соответственно, на "длинных") текстах наблюдаются существенные отклонения структуры повторов слов от (1)-(3) при Z, равном объему полного текста - в частности, резкое завышение объема словаря в сравнении с (2) на "коротких" текстах и его занижение на текстах "длинных" - то, может быть, в таких текстах закон (1) при γ =1, $Z=N_0$, выполняется на единицах более общих, чем слово (в коротких текстах), или более конкретных, чем слово (в длинных текстах)?

в) связаны ли как-либо между собой структуры повторов композиционно значимых единиц различных уровней в тексте? В частности, если единица данного уровня образуется объединением в один класс нескольких единиц **другого уровня /например, объединением слов в классы по признаку семантического их родства и т.п./. то как будут связаны друг с другом организации повторяемости этих единиц в тексте?**

Вопросы эти - их можно высказать и как предположения - **представляются априори содержательными. Действительно**, например тот факт, что тексты "малой формы" ведут себя таким образом, что их лексемный словарь оказывается, как правило, **завышенным** в сравнении с прогнозом по (2) при γ =1 и $Z=N_0$, свидетельствует, как нам кажется, о том, что более актуальной единицей кодирования в них является единица более общая, чем слово, имеющая на том же объеме текста меньший объем словаря вследствие своей большей обобщенности - сведения к каждой такой единице в виде ее вариантов нескольких разных слов. Наоборот, заметное "недобирание" словаря на литературном тексте "большой формы" в сравнении с прогнозом по (2) при Z, равном полной длине текста, свидетельствует о реальности ведущей оптимизации словарной (и частотной, разумеется) структуры текста на **уровне единицы, более конкретной, чем слово.**

При такой постановке проблемы встает ряд внешне

формально-методических, но на самом деле глубоко содержательных вопросов об используемых критериях различения единиц данного типа, об их естественности, органичности для данного текста. Содержательность этой проблемы - не только не тривиальной, но напротив - чрезвычайно сложной и пока еще очень далекой от окончательного решения -связана, в частности, с тем, что если выделенные единицы реально существуют, являются единицами авторского кодирования содержания, единицами оптимизации структуры текста (и, в частности, структуры повторов в нем), то критерии их различения должны быть относительно простыми, естественными, интуитивно легко воспринимаемыми. В связи с этим естественно возникает и такой вопрос: если закон Ципфа-Мандельброта выполнен в тексте на его полной длине при данном типе единицы и данном критерии различения, то выполнится ли он (а если да, то как при этом изменится γ) при существенном изменении критерия различения?

Как нам представляется, ответы на поставленные вопросы позволили бы не только прояснить ряд пока неясных проблем границ и условий реализации закона Ципфа-Мандельброта в законченном художественном тексте, но и более органично включили бы проблематику этого закона в более общую проблематику устройства языка и принципов и законов построения текста.

Ниже описаны первые результаты исследований авторов в этом плане на материале музыкальных и литературных текстов. Вначале приведены результаты, полученные для музыкальных текстов, где выбор нового критерия различения /и, в связи с этим, переход на новый уровень рассмотрения текста) диктовался особенно простыми и естественными соображениями.

3. Приведенные выше результаты исследования организации повторяемости F-мотивов в музыкальном тексте, выявившие подчиненность этой организации закону Ципфа-Мандельброта в его "канонической форме" при условии различения F-мотивов одновременно по ритмической и интервальной структуре (см. выше критерий различения F-мотивов) закономерно вызывают вопрос: какой будет эта организация, если различать F-мотивы только по ритмической структуре - т.е. если считать два F-мотива одинаковыми только в том случае, когда у них одинаково число звуков и i-й по порядку звук первого F-мотива равнодлителен i-му по (тому же) порядку звуку второго F-мотива? /далее будем говорить о таких F-мотивах, что они совпадают по ритмической структуре; в противном случае будем говорить, что они различаются по ритмической структуре).

Вопрос этот априори содержателен уже потому, что ритмика играет важнейшую, можно сказать, основополагающую "структурирующую" роль в мелодии, в ее развитии, является своеобразным стержнем этого развития. Естественно было предложить, что и в организации повторяемости ритмических F-мотивов, т.е. F-мотивов, рассматриваемых как последовательности определенных длительностей вне учета звуковысотной стороны (назовем

их далее F_R-мотивами - см. Рис. 4), также существуют определенные общие, метастилистические закономерности. Более того, можно было предположить, что и в этой организации, на важнейшем для музыкального текста ритмическом уровне будут наблюдаться - как и для "полных", ритмомелодических F-мотивов - "ципфо-мандельбротские соотношения" (1)-(3) объема словаря и объема текста, убывающих частот F-мотивов, и т.д.

Для исследования были взяты музыкальные тексты XVII-XX вв., выполняющие закон Ципфа-Мандельброта при $\gamma=1$ на своей полной длине в условиях одновременного различения F-мотивов по ритмике и интервалике. Для каждого текста определялись:

а) число F_R-мотивов в нем, различающихся по ритмической структуре - объем V^R словаря F_R-мотивов данного текста;

б) относительные частоты $\{p_i^R\}$ встреч каждого из F_R-мотивов этого словаря в тексте.

Проведенный анализ этих данных **не подтвердил** гипотезу о следовании организации повторяемости F_R-мотивов в музыкальном тексте "канонической форме" закона Ципфа-Мандельброта. Упорядоченные по невозрастанию наборы частот $\{p_i^R\}$ принципиально не укладывались на кривую (1) при $\gamma=1$, $K=1/\ln(Zp_{max})$, $B=Kp_{max}^{-1}$ и $Z=N_0$ (рис. 5). Существенно расходились с теоретическим прогнозом по (2) и фактические значения объема V^R словаря F_R-мотивов: значение V^R оказалось, как правило, значительно меньше своего теоретического прогноза. Например, прогноз по (2) словаря F_R-мотивов в сонате д.Тартини "Покинутая Дидона" ($N_0=828$, $p_{max}^R=0.2099$) равен 159.76, а фактическое значение этого словаря равно 63; в Органной фуге № 2 Генделя соответствующие значения равны 153.7 и 63, в сонате № 10 для ф-но Моцарта они составляют 233.9 и 107, в Рондо ор. 59 для ф-но Кабалевского - 116.9 и 60, в сонате № 1 для ф-но Шуберта - 197.56 и 91, в Прелюдии и фуге ор. 87 № 4 Шостаковича - 183.28 и 77, и т.д. Естественно, что и объемы групп m-разовых F_R-мотивов в музыкальном тексте прогнозировались выражением (3) плохо. В целом, подтвердилось возникшее в этих условиях предположение, что музыкальный текст выполняет закон **Ципфа-Мандельброта** на уровне F_R-мотивов не в "канонической форме", при $\gamma=1$, а при значении γ, заметно отличном от единицы. Когда такое значение γ определялось из условия совпадения (с точностью до одного F_R-мотива) теоретического прогноза F_R-мотивного словаря текста с его фактическим значением (прогноз осуществлялся выражением (2а) - см. сноску - при Z, равном полной длине текста), ход убывающих частот F_R-мотивов в тексте плохо описывался выражением (1) при соответствующих (определенных из (1а)) значениях K и B.

При этом выявились еще два примечательных факта. Во - первых, в исследованных музы - кальных текстах различной длины в числе F-мотивов и, вследствие этого, различающихся по объему V словаря

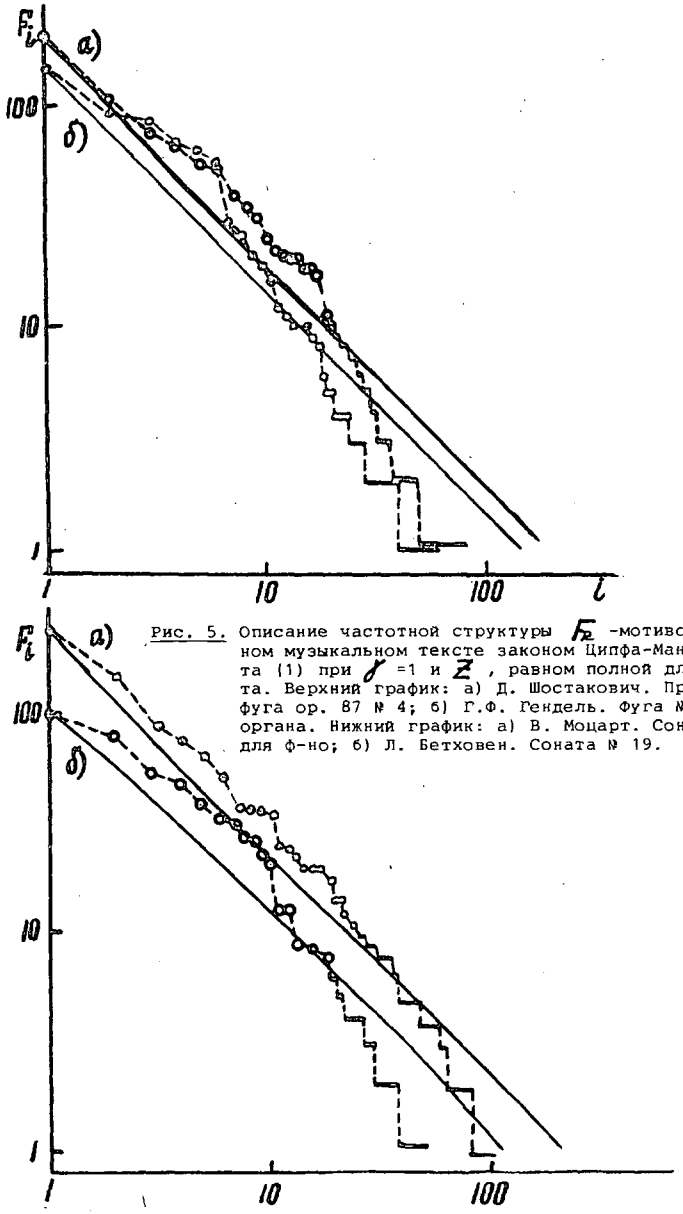


Рис. 5. Описание частотной структуры F_2 -мотивов в полном музыкальном тексте законом Ципфа-Мандельброта (1) при γ =1 и Z , равном полной длине текста. Верхний график: а) Д. Шостакович. Прелюдия и фуга ор. 87 № 4; б) Г.Ф. Гендель. Фуга № 2 для органа. Нижний график: а) В. Моцарт. Соната № 10 для ф-но; б) Л. Бетховен. Соната № 19.

"полных" F_R - мотивов, различным было и значение отношения V/V_R - в большем тексте оно оказывалось, в общем бо́льшим; при этом связь V и V_R была заметно нелинейной. Во-вторых, анализ самих словарей "полных" F - мотивов в исследованных текстах показал, что в каждом таком словаре, с одной стороны, существует много различных F_R -мотивов, имеющих в нем лишь небольшое число звуковысотных вариантов - два, три, и т.д. - или вообще встречающихся в единственном звуковысотном варианте. С другой стороны, в каждом из этих словарей наблюдалось мало различных F_R-мотивов, имеющих в нем много звуковысотных вариантов, многократно варьирующихся в данном тексте в звуковысотном плане. Все это весьма напоминало по характеру частотную структуру текста, подчиненного закону Ципфа-Мандельброта в его "канонической форме", наводя на мысль подробно исследовать организацию повторяемости (мелодической вариативности) F_R -мотивов в полном F -мотивном словаре музыкального текста, рассматривая этот словарь как квазитекст длины V и беря в качестве "частоты встреч" F_R -мотива в этом квазитексте число звуковысотных вариантов в нем данного F_R -мотива.

Проведенный анализ подтвердил сделанное предположение. Организация повторяемости (звуковысотной вариативности) F_R -мотивов в F -мотивном словаре музыкального текста, выполняющего на уровне "полных" F - мотивов закон Ципфа-Мандельброта в его "канонической форме", оказалась подчиненной этому же закону для многих исследованных текстов. И относительные расхождения объема V_R фактического словаря с его прогнозом по (2), и соотношение хода теоретической кривой, построенной по (1) при $Z=N_0$ и приведенных выше R, B и γ, с фактическим ходом убывания "частот" F_R -мотивов в F- -мотивном словаре текста, и анализ соотношений фактического числа и теоретического прогноза числа различающихся F_R-мотивов, имеющих ровно m звуковысотных вариантов в этом словаре, убеждали в таком выводе (Рис. 6). Таким образом, переход от "полных" к ритмическим F - мотивам позволили не только обнаружить на этом новом и более общем, чем "полные" F -мотивы, уровне метастилистические закономерности, но и выявить существующий в музыкальном произведении изоморфизм разных уровней F -мотивной организации, проявляющийся в аналогичности, определенном единстве принципов повторяемости малых единиц в тексте и их вариативности в словаре. Исследованные музыкальные тексты оказались в большинстве случаев по крайней мере дважды пронизанными единой закономерностью; чисто музыкальная связь "полных" и ритмических F -мотивов, как двух уровней организации текста- связь на основе ритмики как базы звуковысотных вариаций - подкреплялась единством организующих эти уровни принципов, связанных с законом Ципфа-Мандельброта.

4. Исследование организации повторов и выполнимос-

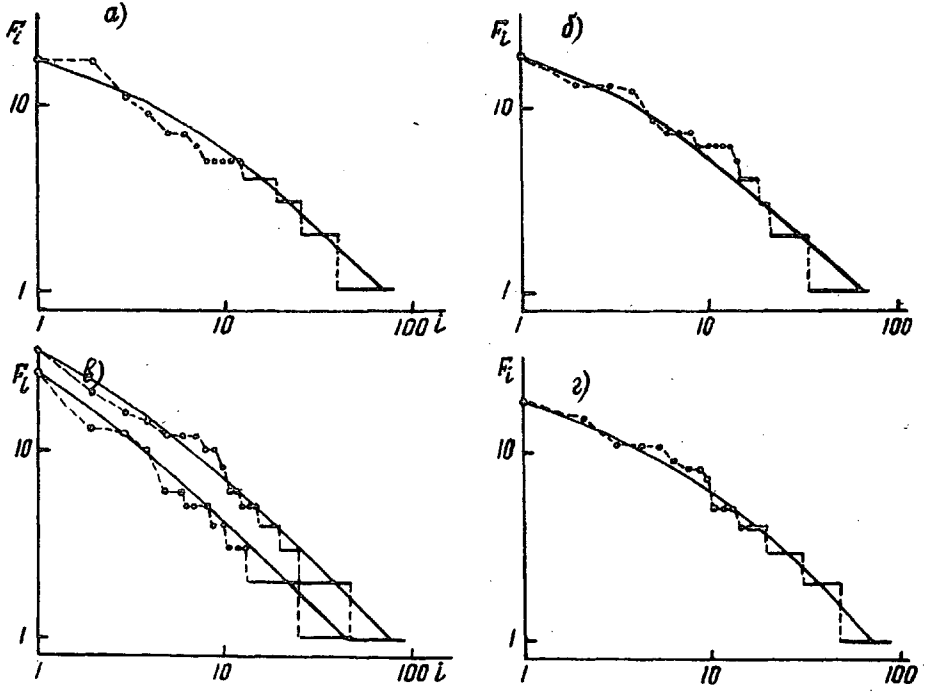


Рис. 6. Вариативность ритмических F-мотивов ("частоты их встреч") в F-мотивном словаре музыкального текста - описание законом Ципфа-Мандельброта при γ =1 и $Z \geq N_0$: а) Д. Тартини. Соната для скрипки с ф-но "Покинутая Дидона"; б) Г.Ф. Гендель. Фуга № 2 для органа; в) Л. Бетховен. Соната № 19 (нижний график), И. Гайдн. Симфония № 45 (верхний график); г) Ф. Шуберт. Соната № 1 для фортепиано. Обозначения и масштаб - те же, что и на Рис. 5.

ти закона Ципфа-Мандельброта на уровне различных малых единиц литературно-художественного текста обнаружило как существенные аналогии, так и отличия ее от ситуации в музыкальных текстах.

Прежде всего, число различных значимых уровней микроорганизации литературного текста превосходит - по крайней мере, при данном уровне изученности проблемы - число подобных уровней в тексте музыкальном. Это связано, повидимому, с тем, что естественный язык является более универсальной и, возможно, более дифференцированной семиотической системой, чем язык музыкальный. Например, для лексической подсистемы языка можно выделить, по крайней мере, три различных основных уровня единиц: уровень с л о в, уровень г и п е р л е к с е м (единиц "надлексемного" уровня, в каждой из которых отождествлено несколько однокоренных слов, принадлежащих разным частям речи, но выражающих одну и ту же лексическую "идею", осложняемую в каждом из них добавочными категориально-грамматическими компонентами процессуальности, субстанциональности и т.д./ и уровень л е к с и к о-с е м а н т и ч е с к и х в а р и а н т о в, ЛСВ /каждый из которых есть слово в одном из его значений/. Каждая из этих единиц является значимой в кодировании смыслового содержания текста, представляет собой важный этап этого кодирования, этап "разворачивания" текста. Уровень гиперлексем является в этом плане, по-видимому, одним из исходных, а уровни слова и лексико-семантического варианта оказываются последовательными этапами конкретизации, вариативности гиперлексем как неких относительно исходных единиц.

Существенно отметить, что каждая лексема объединяет в себе ряд ЛСВ, обладающих полным тождеством плана выражения и частичным тождеством плана содержания, каждая гиперлексема объединяет в себе ряд лексем на основе тождества основной части их плана выражения /корня/ и тождественности некоторой существенной части их плана содержания /лексических компонентов основных значений слов/. Это позволяет предположить, что должны существовать ещё более обобщённые лексические единицы, чем гиперлексемы. Такие единицы следующего по обобщённости уровня нами названы синонимико-антонимическими блоками гиперлексем - САБГ. Повышение степени обобщения в плане содержания заключается в игнорировании и ряда конкретно-лексических различий в близких /синонимичных/ по значению гиперлексемах, а в плане выражения - в полной нейтральности к явлению сходства или различия гиперлексем по их внешней формальной выраженности.

Следующие ступени обобщения лексических единиц должны вести к выделению лексико-семантических групп /ЛСГ/ различных уровней обобщения /Борода, Поликарпов 1984/.

Четыре названных единицы - гиперлексемы, слова, ЛСВ, синонимико-антонимические блоки гиперлексем/САБГ/

-определили собой четыре основных уровня исследования литературных текстов, описываемого ниже. Нами были выдвинуты и рассмотрены два предположения: а/тексты малой формы, не выполняющие закон Ципфа-Мандельброта в его "каноническом" виде на уровне слов, выполняют его на уровне гиперлексем или синонимико-антонимических блоков гиперлексем ; б/для каждого из четырёх исследуемых уровней существует "своё" значение γ , при котором данный текст выполняет на этом уровне обобщённый закон Ципфа-Мандельброта. Это значение определялось описанным выше методом по фактическому лексемному /соответственно, гиперлексемному, ЛСВ- и т.д./ словарю данного текста.

5.Производившаяся с литературными текстами работа заключалась в составлении частотных списков ЛСВ,слов, гиперлексем и блоков и в выявлении статистических характеристик этих списков, аналогичных описанным выше при анализе структуры повторов F- и F_k -мотивов в музыкальных текстах. Взятые для исследования тексты относились преимущественно к малым формам /будучи по объёму меньше или немного превышая 10000 словоупотреблений/, а также к средним:"Пиковая дама"А.С.Пушкина /около 6800 словоупотреблений/, его же "Капитанская дочка" /около 29000 словоупотреблений/,"Дама с собачкой" А.П.Чехова/около 5000 словоупотреблений/,"Обыкновенные атомщики" В. Ставинского (около 13 600 словоупотреблений), "В прекрасном и яростном мире" А.П. Платонова (около 4000 словоупотреблений), "Жизнь на грешной земле" А.П. Иванова (около 13 000 словоупотреблений).

Проведенный анализ с очевидностью показал справедливость предположения о "тяготении" текстов малой формы к выполнению закона Ципфа-Мандельброта в его "каноническом" виде на уровне г и п е р л е к с е м. Например, повесть "Пиковая дама" Пушкина, не выполняя этого закона при $\gamma=1$ и $Z=N_0$ на лексемном уровне, с хорошей точностью выполняла его на уровне гиперлексем - см. Рис. 7 и 8. В частности, теоретический прогноз по (2) словаря гиперлексем на полном объеме "Пиковой дамы" составил 1131.15, что лишь на 14 % отклоняется от фактического гиперлексемного словаря этого произведения; количество одноразовых гиперлексем больше своего теоретического прогноза по (3) на 16 %, для двухразовых гиперлексем (число их меньше прогноза) это отклонение равно - 8 %, для трехразовых - 27 %. Для сравнения укажем, что на уровне лексем значение словаря отклоняется от своего прогноза на 55 %, значение числа одноразовых слов - на 87 %, двухразовых - на 54 % и трехразовых - на 28 %. Вычисление точного значения γ для структуры повторов гиперлексем в "Пиковой даме" (методику см. выше) показало, что это значение очень мало отлично от единицы, будучи равным 0,93. При этом значении γ отклонения указанных эмпирических величин от их прогнозов стали еще меньшими. (Точное значение γ для частотной структуры слов в "Пиковой даме" равно 0.78).

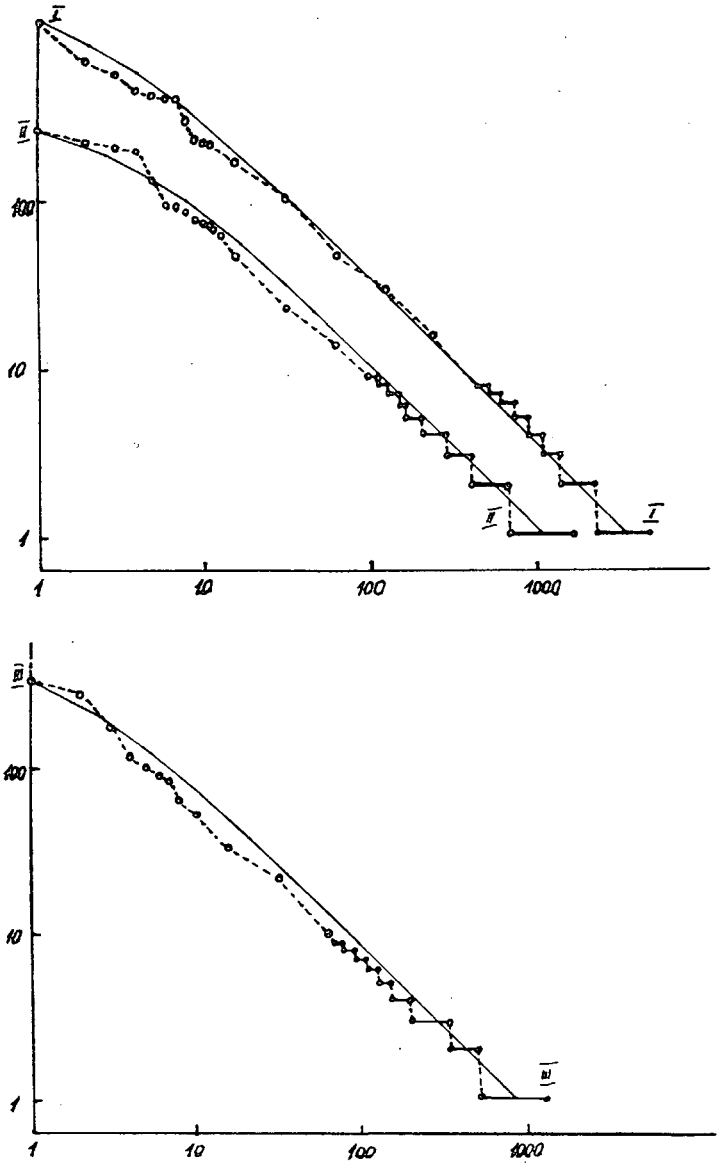


Рис. 7. Частотная структура литературного текста на уровне лексем: I – А.С. Пушкин. Капитанская дочка; II – А.С. Пушкин. Пиковая дама; III – А.П. **Чехов. Дама с собачкой.** Теоретические частоты, определенные по (1) при $\gamma = 1$ и $\overline{Z} = N_0$, даны сплошной линией, фактические частоты – точками, соединенными штриховой линией. Масштаб логарифмический.

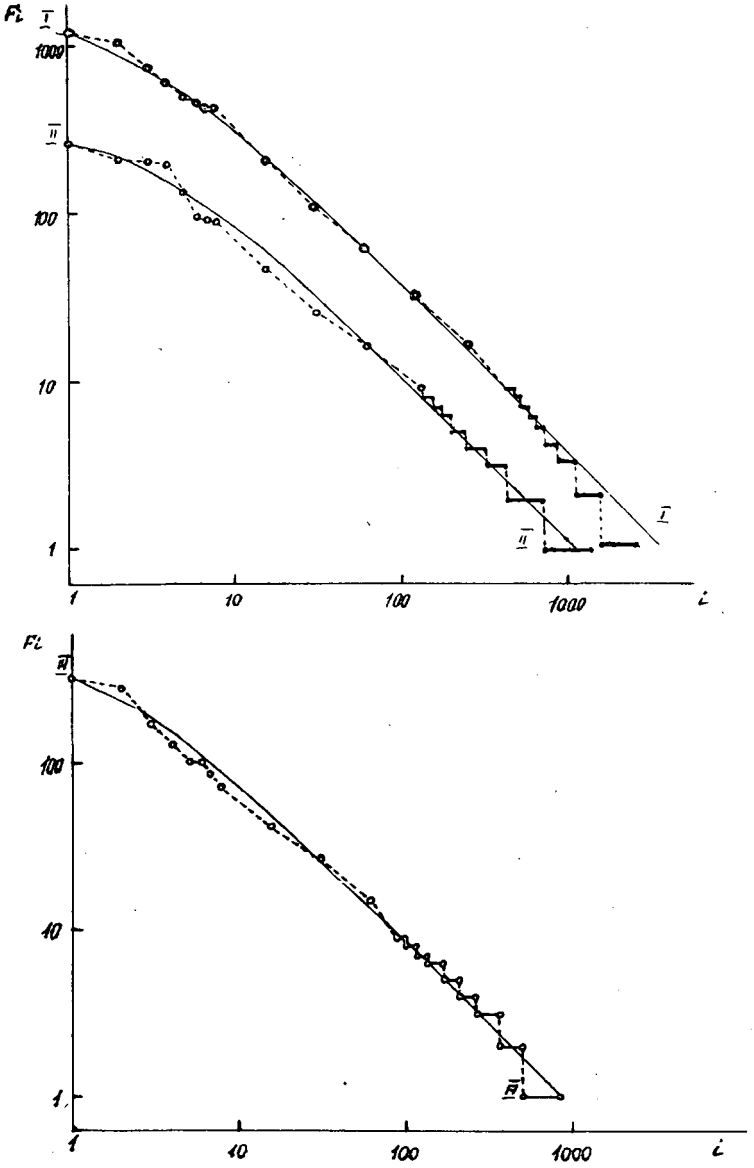


Рис. 8. Частотная структура текста на уровне гиперлексем. Тексты и обозначения - те же, что и на Рис. 7.

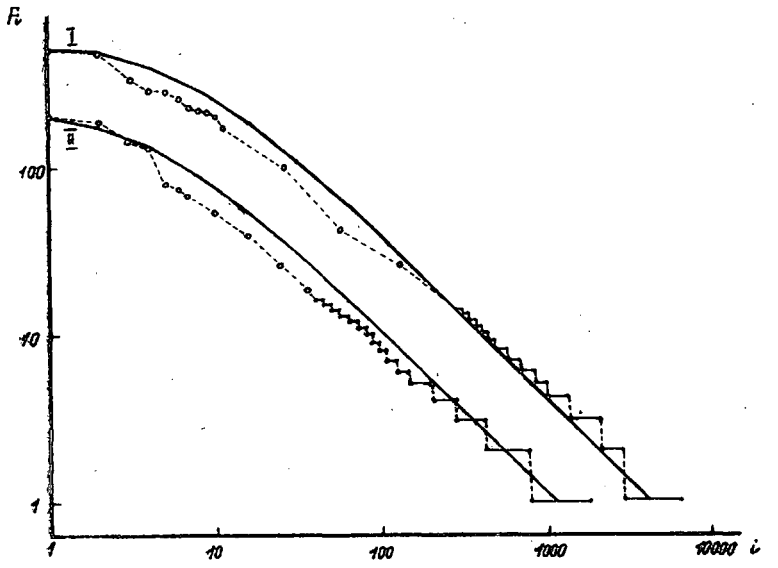


Рис. 9. Частотная структура текста на уровне лексико-семантических вариантов слов (ЛСВ). Обозначения – те же, что и на Рис. 7 и 8. 1 – А.С. Пушкин. Капитанская дочка; II – А.С. Пушкин. Пиковая дама.

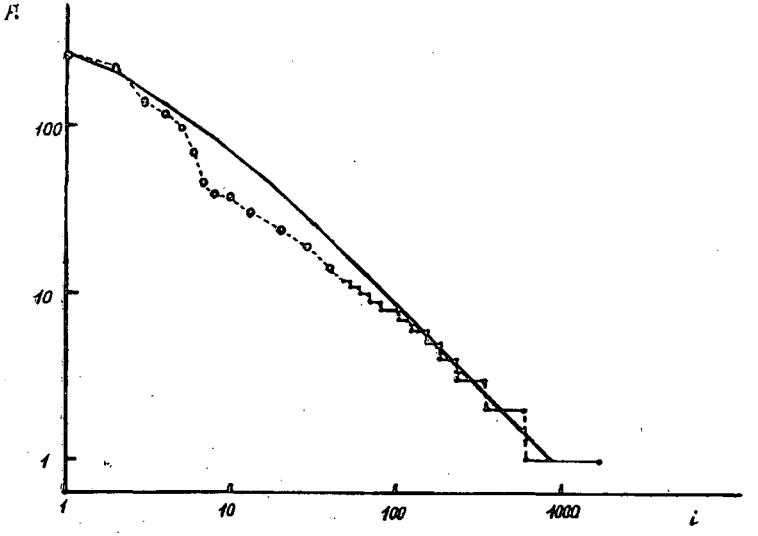


Рис. 9а. Частотная структура текста на уровне лексико-семантических **вариантов**: А.П. Чехов. Дама с собачкой. Обозначения те же, что и на рисунках 7, 8 и 9.

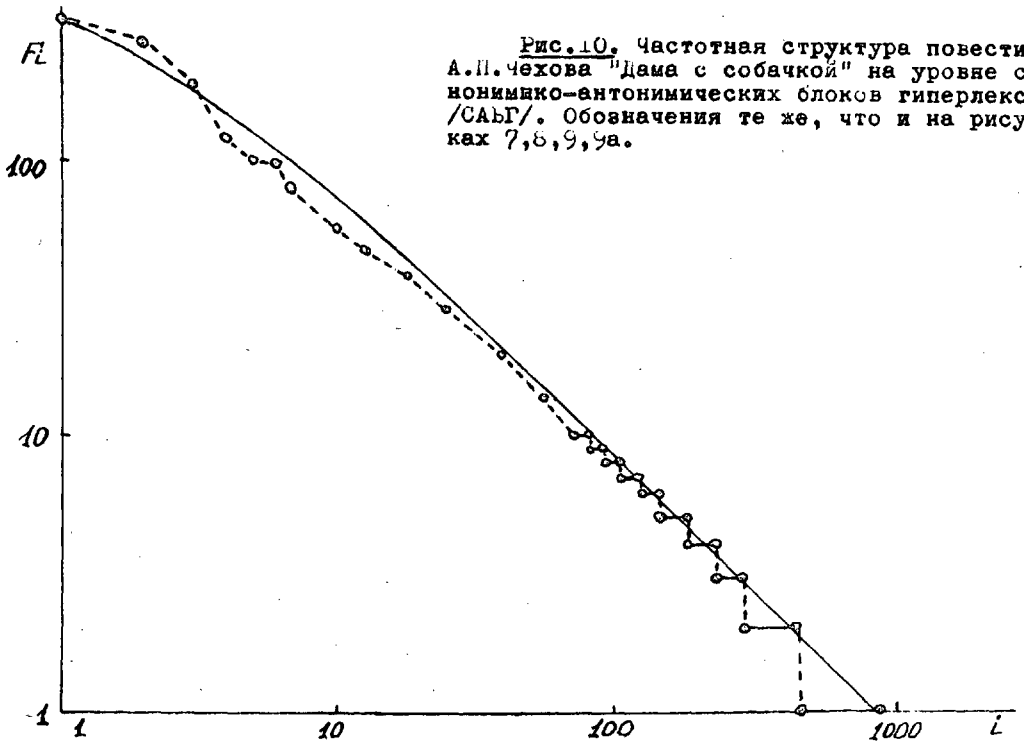


Рис.10. Частотная структура повести А.П.Чехова "Дама с собачкой" на уровне синонимико-антонимических блоков гиперлексем /САЬГ/. Обозначения те же, что и на рисунках 7,8,9,9а.

Еще большее приближение к "каноническому" виду на уровне гиперлексем обнаружила частотная структура текста "Дама с собачкой" А.П. Чехова (Рис. 8). Фактический гиперлексемный словарь отклоняется в этом тексте от своего прогноза по (2) на 0.85 % (соответствующие значения равны 832 и 839.14), небольшими оказываются и отклонения от прогноза по (3) фактических значений числа m-разовых гиперлексем в тексте. Ещё более точное соответствие хода теоретического и эмпирического частотных распределений наблюдается на уровне САБГ /см. рис. 10/. На уровне же слов фактическое значение словаря оказывается примерно в полтора раза больше своего теоретического прогноза /соответствующее относительное отклонение равно 62.74%/; естественно, значительными оказываются и относительные отклонения для m-разовых слов. Не выполняется в обоих текстах закон Ципфа-Мандельброта в "канонической форме" и на уровне частот ЛСВ (Рис. 9). В то же время закономерно, что такой текст средней формы, как "Капитанская дочка" Пушкина выполняет этот закон в "канонической форме" на лексемном уровне, но не выполняет его на уровне гиперлексем и ЛСВ (Рис. 7-8-9). Вычисление значения γ, при котором этот текст выполняет закон Ципфа-Мандельброта на уровне гиперлексем или ЛСВ, выявило заметное отличие этих значений от единицы. Аналогичные результаты были получены на других исследованных текстах: в каждом из них закон Ципфа-Мандельброта в "канонической форме" выполняется на уровне какой-либо одной единицы (притом связанной с типом, "величиной" формы текста так, как это было предположено выше) и, с другой стороны, не выполняясь в этом виде, при γ =1, на других единицах, закон этот выполнялся на них при других, отличных от 1, значениях γ.

Выявление этого факта указывает не только на особые, "неканонические" характеристики выполнения закона Ципфа-Мандельброта в литературном тексте на каждом из "нецентральных" уровней его организации (центральность, особую важность уровня единицы вполне естественно связать с выполнением на ней в данном тексте "канонического вида" закона Ципфа-Мандельброта хотя бы в силу присущей этому виду особой уравновешенности структуры повторов - см. выше), но и говорит о замечательном явлении взаимной согласованности структур повторов композиционно и языково значимых единиц на разных уровнях текста. Частотные структуры этих разных уровней оказываются взаимосвязанными (по крайней мере, в высокохудожественном тексте): если на одном из уровней лексического кодирования выполнен рассматриваемый закон (с определенным значением γ), то и на других указанных уровнях он тоже выполняется, но только с другим значением γ. Интересно, что анализ этой согласованности - по крайней мере, в свете тех фактов, которыми авторы располагают на сегодняшний день - позволяет говорить о

некотором различии музыкального и естественного языков по характеристикам структуры повторяемости в тексте их малых единиц. В музыкальных текстах повторяемость более "общих", чем F-мотив, FR-мотивов, в отличие от повторяемости аналогичных им единиц (гиперлексем) в литературном тексте, как правило, плохо описывается законом **Ципфа-Мандельброта** в его "канонической" форме, при $\gamma = 1$.

6. Полученные результаты позволяют не только уточнить условия реализации закона Ципфа-Мандельброта в художественном тексте, показать неединственность уровня его реализации и возможность "многоэтажного", согласованного его выполнения на разных уровнях, отметить "изоморфизм поведения" единиц этих разных композиционно значимых уровней текста, но и поставить ряд новых вопросов. Самым главным из них, на наш взгляд, является то, что полученные данные говорят в пользу психологической реальности нескольких уровней лексического кодирования содержания литературного текста и "микромотивного кодирования" музыкального текста - уровней, попеременно находящихся в фокусе авторского внимания. Когда какой из них попадает в этот фокус, становится "центральным", выполняющим закон Ципфа-Мандельброта в его "канонической форме" и обнаруживает связанную с этим особую уравновешенность структуры повторов в законченном тексте, гармонию, баланс повтора и неповторности? Почему в тексте малой формы таким становится уровень гиперлексем или САБГ, ещё более обобщённых единиц, а в текстах средней формы - уровень более детализированных единиц - слов? Нельзя ли также - "по инерции" - предположить, что в текстах большой формы (напрмер, в романах), с объемом более 100 000 словоупотреблений, центральным уровнем кодирования будет уровень самой конкретной, самой детализированной из рассмотренных лексических единиц кодирования - уровень лексико-семантических вариантов (ЛСВ)?

Выделение одного из уровней, на которых выполняется закон Ципфа-Мандельброта в законченном тексте на его полной длине, в качестве центрального (а именно, такого, для которого закон этот выполняется при $\gamma = 1$ со всеми вытекающими отсюда последствиями), повидимому, отнюдь не условность. Можно полагать - хотя бы из соображений связи "канонической формы" закона Ципфа-Мандельброта с неоднократно здесь отмеченной равновесностью структуры повторов - что именно на этом уровне в данном тексте (точнее, в текстах данной формы - малой, средней, большой) сосредотачивается преимущественное внимание автора - сознательное или бессознательное - по "гармонизации структуры повторов" в тексте.

Та гипотеза, которая может быть здесь высказана, заключается в том, что циклы авторской рефлексии в текстах разной по объему формы успевают развернуться в разной степени. Самые развернутые циклы этого рода могут быть реализованы в текстах большой формы - романах, крупных повестях, и т.д. - "романное время" предоставляет

здесь автору произведения наибольшие возможности. В результате этого авторское внимание может успеть учесть и сделать центральным своим объектом характеристики повторяемости единиц такого, например, "отдаленного" уровня, как уровень ЛСВ в текстах естественного языка. Менее развернутые формы, предоставляя автору меньше возможностей такого рода, закономерно оказываются связанными с фиксацией авторского внимания на единицах более общих, более приближенных к началу процесса кодирования - на словах в текстах средней формы, на гиперлексемах и САЫ - в малой. Форма, жанр, объем задают такие возможности и на них ориентируется авторская рефлексия с самого начала написания текста.

7. Подводя итоги проведенному исследованию, можно сказать, что хотя еще предстоит существенная работа по дальнейшей экспериментальной проверке некоторых выдвинутых теоретических положений, уже сейчас ясно, что они позволяют прояснить некоторые "темные места" в вопросе о выполнимости закона Ципфа-Мандельброта в художественных текстах. А именно: можно утверждать, что уровень словесного кодирования - неединственный уровень организации литературного текста, что альтернативными ему единицами, значимыми в композиционной структуре текста, являются гиперлексема и ЛСВ - единицы, более обобщенная и более конкретная, чем слово; аналогично, альтернативным F-мотиву в музыкальном тексте является FR-мотив. Можно также утверждать, что выполнение закона Ципфа-Мандельброта в "канонической" форме на одном из этих уровней зависит от жанра (формы, степени ее крупности) текста: малые литературные тексты - преимущественно до 10 тыс. словоупотреблений - тяготеют к выполнению этого закона на уровне гиперлексем и САЫ, средние - от 15-20 тыс. до 100 тыс. словоупотреблений - на уровне лексем, большие /предположительно/ - на уровне ЛСВ. Осознание реальности других единиц кодирования в художественном тексте, отличных от исследовавшихся на сегодняшний день, позволяет не только выявить некий центральный уровень кодирования в данном тексте, но и получить для других уровней, с другими значениями γ, более точный, в условиях этого закона, прогноз объема словаря и частотной структуры текста. Наконец, можно утверждать, что между тремя рассмотренными уровнями лексического кодирования в литературном тексте (соответственно, между двумя уровнями F-мотивного кодирования в тексте музыкальном) существует взаимная согласованность, изоморфизм, имеющая глубокие причины аналогичность в организации повторяемости (вариативности). Представляется весьма вероятным, что дальнейшие исследования в этом направлении - с привлечением текстов более широкого стилистического диапазона, текстов больших и меньших, чем рассмотренные, форм, как музыкальных, так и литературных, позволит не только тщательно изучить организацию повторяемости различных композиционно значимых единиц в художественном тексте, но и выявить важные принципы и закономерности его развертывания, его реальной "жизни" в процессе создания и в процессе восприятия.

ЛИТЕРАТУРА

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. - НТИ, 1975, № 1.

Борода М.Г. К вопросу о метроритмически элементарной единице в музыке. - Сообщения АН ГССР, т.71, № 3, 1973.

Борода М.Г. О частотной структуре музыкальных сообщений Сообщения АН ГССР, т. 76, № 2, 1974.

Борода М.Г. Частотные структуры музыкальных текстов. - Сб. трудов Тбилисской гос. консерватории им. З. Сараджишвили. Мецниереба, Тбилиси, 1977.

Борода М.Г. Принципы организации повторов на микроуровне музыкального текста. АКД, Тбилиси, 1979.

Борода М.Г., Поликарпов А.А. О семантических классах в литературно-художественном тексте. 1984, рукопись.

Орлов Ю.К. Обобщение закона Ципфа-Мандельброта. Сообщения АН ГССР, т. 57, № 1, 1970.

Орлов Ю.К. Частотные структуры конечных сообщений в некоторых естественных информационных системах. АКД, Тбилиси, 1975.

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

Орлов Ю.К. Модель частотной структуры лексики. - Исследования в области вычислительной лингвистики и лингвостатистики, вып. II, ч. 1 Изд-во МГУ, М., 1978.

Поликарпов А.А. Факторы и закономерности аналитизации языкового строя. АКД, М., 1976.

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

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. - Учен. зап. Тартуск. ун-та, вып. 502. Тарту, 1979.

Тулдава Ю.А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. - Труды по лингвостатистике, вып. 6. Тарту, 1980.

Boroda M.G. Häufigkeitsstrukturen musikalischer Texte (Diskussion). - Quantitative linguistics, vol. 5. Dr N. Brockmeyer's Studienverlag, Bochum, 1981.

Orlov Yu. K. Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik). - Yu. K. Orlov, M.G. Boroda, I.S. Nadarejšvili. Sprache, Text, Kunst: Quantitative Analysen. Quantitative linguistics, vol. 15. Dr. N. Brockmeyer's Studienverlag, Bochum, 1982.

THE ZIPF-MANDELBROT LAW AND UNITS OF DIFFERENT TEXT LEVELS

M.G. Boroda, A.A. Polikarpov

Summary

The paper deals with some ontological problems of the generalized form of the Zipf-Mandelbrot law (Орлов, 1970). The main point is that a well-structured text (above all a literary and musical text) can be regulated by this law simultaneously on various levels of its lexical organization with different values of the parameter γ of the law. Those are the levels of hyperlexems (units more generalized than words), words, lexico-semantic variants of words (LSV) for a literary text. It is shown that there exists only one such level ("central coding level") which follows the law in its so-called "canonical form" (with $\gamma = 1$). Moreover there has been discovered a fact of correlation between the thus defined "central coding level" of a text and its genre (its length): the hyperlexemic level is central for stories or rather short novels (usually not longer than 10^4 tokens in a text), the lexemic level is central for large stories and not very long novels (usually 10^4-10^5 tokens), the LSV level is central for long novels (usually more than 10^5 tokens). This order is correlative to successive stages of the encoding process during an author's creation of a text. There are some significant isomorphisms between literary and musical texts. In addition musical texts show a very high degree of similarity in the frequency structure of so-called F-motives and the variation structure of F_R-motives (more generalized units than F-motives).

К ПРОБЛЕМЕ ОБОБЩЕНИЯ И ИНТЕРПРЕТАЦИИ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ В СТАТИСТИЧЕСКОЙ ЛИНГВИСТИКЕ

В.Н. Бычков

В современной статистической лингвистике проявляется постоянный интерес к нахождению таких новых математических моделей и модификаций уже разработанных, которые бы максимально полно и точно описывали экспериментальные и фактические данные. Важное место занимают также попытки дать обобщенное математическое представление тем эмпирическим и теоретическим распределениям, которые получили широкое применение в лингвостатистике, и по-новому интерпретировать их с общенаучных и лингвистических представлений и данных. Это прежде всего относится к з а к о н у Ц и п ф а, относящегося к так называемым частотно-ранговым распределениям и занимающего в настоящее время одно из центральных мест в ряду лингвостатистических моделей.

Относительная простота и эффективность самого закона Ципфа и тех моделей, которые строятся на его основе, проявились при обработке самых разнообразных лингвистических массивов. Поэтому этот закон продолжает привлекать внимание тех, кто, сознавая его фундаментальность, пытается найти более точные формы его математического выражения. Теоретический анализ, проведенный рядом авторов (Хомский Н., Миллер Дж., 1967; Яблонский А.И., 1975; Mandelbrot B., 1961; и др.), позволил установить некоторые специфические особенности закона Ципфа, которые исключают возможность свести его в пределе к нормальному распределению и связывают его с асимптотикой устойчивых негауссовых распределений, разработанных современной статистической теорией, что подчеркивает универсальность этого закона в математическом отношении и снимает возможные упреки в тавтологичности его структуры.

В статистической лингвистике понятия математического закона и модели обычно понимаются в качестве равноценных. Но по отношению к понятиям закона и модели Ципфа такое представление является в общем условным, так как на основе закона Ципфа, который является одним из наиболее общих лингвостатистических утверждений, отражающих эмпирические закономерности языка и речи, уже построено несколько частотно-ранговых моделей, или конкретизированных математических описаний. Различными авторами они определяются как модификации, уточнения, приближения и обобщения исходной ципфовской модели. В настоящее время можно говорить о целой группе частотно-ранговых ципфовских моделей, которые обладают свойства-

ми системы, так как отдельные ранговые модели находятся в отношениях взаимосвязи, общего основания и дополнительности. Например, в работе П.М. Алексеева (1983) речь фактически идет о четырех таких ранговых моделях, которые определяются самим автором как распределения Ципфа в четырех приближениях, и задаются условия последовательной заменимости этих приближений. В терминах абсолютных частот четвертая модель имеет вид

$$F_i = kN \cdot i^{-\gamma - c\,lg\,i}. \qquad (1)$$

В лингвостатистике хорошо известно, что ципфовские параметры k и γ остаются постоянными лишь для ограниченного участка частотного списка, включающего 1500-2000 наиболее частотных лексических единиц. Поэтому особый интерес представляет модель (1), на основании которой делается вывод о том, что в нелинейной, логарифмической формулировке закон Ципфа лучше описывает эмпирические ряды распределений и в выборках так называемого оптимального размера, и в выборочных совокупностях, существенно отличающихся от этого "оптимума". Для обеспечения целочисленности малых частот и больших рангов предлагается ступенчатая модель (Орлов Ю.К., 1976) и "гиперболическая" лестница (Арапов М.В. и др., 1975), которые ориентированы на более точное описание хвостовой части частотно-рангового распределения, но не решают задачи интегральной аппроксимации, так как в силу своей "жесткости", которая является результатом ограниченного количества параметров в этих моделях, они не учитывают закономерного уклонения эмпирического ряда распределения при последовательном увеличении объема выборки. В принципе на основе закона Ципфа можно строить различные ранговые модели, которые могут различаться прежде всего исходными идеями и принципами, лежащими в основе всех последующих рассуждений, даже если эти принципы не сформулированы в явном виде, а вынесены, так сказать, за скобки математической модели в область возможной логической, лингвистической или теоретико-информационной интерпретации. Различаются ранговые модели уровнем аппроксимации и характером допустимых формально-математических преобразований. Общей для них является логическая связь с законом Ципфа, а также количественное и качественное единообразие параметров, "жесткость" математической структуры моделей, имплицитно ориентированных на представление о единообразии и универсальности систем различных языков и подъязыков в лексическом, грамматическом и семантическом планах.

На фоне ранговых моделей с константными параметрами в качестве принципиально новой выступает модель (1) с закономерно изменяющимся, скользящим параметром γ. Здесь мы фактически наблюдаем появление еще одного параметра, или, точнее, расщепление ранее считавшегося постоянным на всем или значительном участке распределения

параметра γ на две составляющие - собственно константу γ_0 и переменную скользящую γ_i, так как степенный коэффициент рангов в модели (1) можно представить как

$$\gamma_i = \gamma_0 + c \lg i, \qquad \gamma_i - \gamma_0 = c \lg i. \tag{2}$$

Тогда частотно-ранговую модель (1) можно переписать в виде

$$F_i = k N \cdot i^{-\gamma_i}. \tag{3}$$

Очевидно, что выражение (3) можно считать общим для других ранговых моделей, в которых γ - параметр понимается как некоторая закономерно изменяющая величина, причем характер ее закономерного изменения понимается и математически задается иначе. Более того, если рассматривать γ-константу в качестве частного случая γ_i - скользящего параметра, то выражение (3) будет общим и для других частотно-ранговых моделей.

Тот факт, что один из параметров моделей (1) и (3) находится в функциональной связи с одним или несколькими другими параметрами, переводит эти модели из разряда стационарных в разряд динамических. Для них существенным является выявление возможностей взаимно увязать переменные величины и определить пути и способы их вывода из некоторого общего основания. Для этого обратимся еще раз к графическому представлению закона Ципфа. Когда в лингвостатистике говорят о частотно-ранговом распределении, то обычно имеют в виду соотношение логарифмов рангов и частот. Если отвлечься от некоторых конкретных особенностей распределения по закону Ципфа, то в общем графически в билогарифмических координатах частотно-ранговое соотношение можно представить в следующем виде (рис. 1-2).

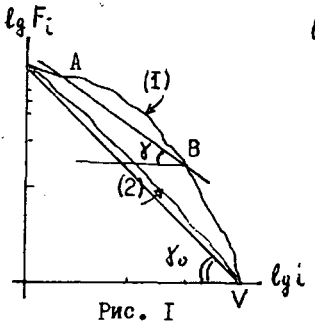


Рис. I

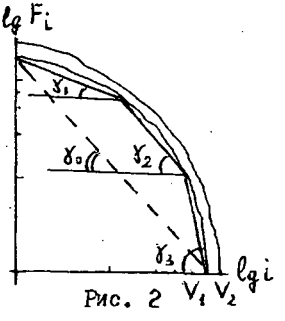


Рис. 2

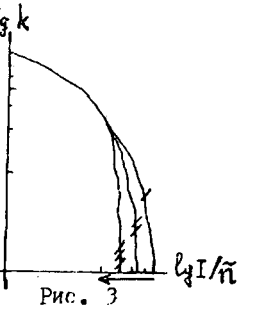


Рис. 3

На рис. 1 и 2 F и i, то есть частоты и ранги слов или словоформ, выступают в качестве переменных величин, F_{max} - абсолютная частота наиболее частого слова в заданном объеме выборки является единственной явной координатой графиков ципфовских моделей, γ -коэффициент частотно-рангового соотношения, который, как видно из графика 1, сохраняет постоянное численное значение лишь на ограниченном участке А - В частотного ряда. На графиках имеется, но обычно не работает другая необходимая координатная точка модели - i_{max}, численно равный объему словаря выборочной совокупности V. Если бы γ была постоянной на свем протяжении частотного ряда ("оптимальный объем" - кривая 2), то график частотно-ранговой зависимости представлял собой прямую или почти прямую с координатами $F_{max} = \kappa N$ (где κ - относительная частота, или вероятность наиболее высокочастотного слова выборки длиной N с/у), а так же V. Параметр γ выражался бы через соотношение логарифмов F_{max} и V.

В то время как величина k является более или менее постоянной в конкретных подъязыках, корректное определение объема словаря V в заданном объеме выборки представляет собой отдельную лингвостатистическую задачу, которая на практике решается независимо от закона Ципфа. К настоящему времени в статистической лингвистике разработан целый ряд N/V -моделей "текст-словарь", в частности, модели Г. Хердана, П.Гиро, В.М. Калинина, Ю.А. Тулдава, В.В. Нешитого и др. Сравнительный анализ этих моделей проводится в статье Ю.К. Орлова (1978). Из них можно выделить семейство параболических N/V -зависимостей, которые в общем виде предстают как

$$N = V^{\alpha} \text{ или } V = N^{\beta},$$

параметры которых у разных авторов существенно различаются, так как эти зависимости получены на основе различных исходных допущений. Но для того, чтобы стать параметром или составной частью ципфовской модели, N/V - функция должна обладать свойствами полной совместимости с законом Ципфа в лингвистическом и формально-математическом планах. Последнее требование легко удовлетворить, если вслед за Г. Херданом (Herdan G., 1964, с. 28-37) исходить из того, что первый момент (или средняя частота ципфовского распределения) для достаточно большого словаря равняется

$$\tilde{n} = \alpha V. \qquad (4)$$

Так как $\tilde{n} \cdot V = N$, то

$$N = \alpha V^2, \qquad (5)$$

то есть получаем еще одну параболическую зависимость "текст-словарь", родственную закону Ципфа: ее поэтому можно включать в семейство ципфовских моделей. Введение в выборку "инородного" фрагмента, то есть текста, относящегося к другому подъязыку или такого фрагмета из одного и того же подъязыка, но лексика и терминология которого не обладает свойством непосредственной семантической взаимосвязи и логико-понятийнсй деривации, проявляется в увеличении параметра α. Поэтому в целом постоянство величины этого параметра является показателем и некоторой мерой однородности подъязыка, признаком лексико-терминологической близости фрагментов выборки. Существенность такого параметра для ципфовских моделей не требует специальных доказательств. Подставив выражение (5) в "каноническую" модель Ципфа, получим более развернутую форму ее выражения

$$F_i = k \alpha V^2 \cdot i^{-\gamma} \qquad (6)$$

Поскольку динамическое соотношение "текст-словарь" носит параболический характер, а количественно значение γ-параметра зависит от величин максимального ранга и объема словаря выборки ($i_{max} = V$), то при увеличении объема этой выборки соответствующее изменение γ частотно-рангового ряда также с неизбежностью должно отражать характер соотношения (5). Из (6) в целом становится очевидной многомерность γ-параметра, а тезис о нелинейном, параболическом характере закона Ципфа получает еще одно логическое подтверждение.

В действительности частотно-ранговое распределение существенно отличается от "оптимального" графика линейной зависимости. И это отличие тем значительнее, чем больше объем соответствующего частотного словаря. Это проявляется особенно заметно, если в целях сохранения единого масштаба и сопоставимости графики представить не в абсолютных, а в относительных частотах. (См. рис. 3, где билогарифмическими координатными точками выступают максимальная относительная частота κ и $V/N = I/\tilde{n}$ -обратная величина средней частоты выборки.) В целом кривая распределения делается все более выпуклой из-за левого смещения нижней координаты в последовательно увеличивающейся выборки (эффект "левого смещения" как результат уменьшения обратной величины средней частоты выборки).

Сохранив в качестве исходного понимание γ_0 как соотношения логарифмов F_{max} и V, аппроксимируем эмпирическую кривую несколькими прямолинейными отрезками (двумя, тремя и т.д.). Тогда даже по графику видно, что на различных участках угловой коэффициент γ не одинаков и изменяется, возрастая при периоде от $i = 1$ к $i_{max} = V$, а при увеличении объ-

ема словаря V параметр γ меняет свое численное выражение также и в тесной связи с изменением величины V (эффект "левого смешения" - см. рис. 3). В целом же характеристикой увеличения γ может служить коэффициент d его прироста за некоторый период "времени", или за интервал перехода от $i = 1$ до $i = n$. Этот коэффициент представляет собой отношение прироста $\Delta\gamma$ за указанный интервал к общему значению γ_0. Путем последовательных преобразований через предел Эйлера получаем величину γ к концу перехода к $i_{max} = V$.

$$\gamma = \gamma_0 \cdot e^{dV}, \text{ или } \gamma = \gamma_0 \exp dV.$$

В логарифмическом представлении на графике функциональная зависимость γ от i носит линейный характер. Здесь, естественно, речь идет о натуральных логарифмах. Тогда численно коэффициент d будет определяться как отношение $\ln\gamma_0$ к $\ln V$, так как величина объема словаря равна величине максимального ранга элемента выборки. Соответственно, γ_i, или значение γ-параметра в конкретной точке частотно-рангового ряда, определяется как

$$\gamma_i = \gamma_0 \exp di.$$

Очевидно, что в такой записи γ-параметр предстает как закономерно изменяющаяся, "скользящая" величина. Тогда закон Ципфа получает следующее модельное представление

$$F_i = k d V^2 \cdot i^{-\gamma \exp di}. \qquad (7)$$

Данная модель в общем достаточно удовлетворительно описывает закономерности эмпирического ряда распределения в выборках различного объема, но и она, как и большинство других ципфовских моделей, опирается на принцип универсальности и изоморфизма структур разных языков и подъязыков. Существование некоторых общеязыковых универсальных характеристик не может служить запретом для наличия специфических особенностей в лексико-грамматической структуре конкретных языков и подъязыков, которые прежде всего проявляются на высокочастотной лексике, и, соответственно, не может быть априорного запрета на специфическое уклонение эмпирической кривой распределения в ту или иную сторону от теоретически предсказуемой. Если считать, что грамматика языка по отношению к лексике в общем выступает в качестве некоторого базиса, его инфраструктуры, тогда высокочастотный, поправочный коэффициент Мандельброта p естественно считать инфраструктурным параметром, и его необ-

ходимо сохранить в ранговой модели. В силу специфичности конкретных языков и подъязыков инфраструктурный поправочный коэффициент выступает в качестве отдельной эмпирической координаты и параметра модели. Его можно вычислить путем приведения соответствующих эмпирических и теоретических частот к равенству, например, 10 и 20-го по рангу слова частотного словаря или любых других, которые считаются наиболее специфичными в заданном отношении. Тогда частотно-ранговая модель (7) принимает вид

$$F_i = k\alpha V^2 \cdot (i+p)^{-\gamma \exp di} \qquad (8)$$

Из вывода модели (8) следует, что параметры α и γ находятся в определенной функциональной связи, которая проявляется в том, что при последовательном многократном увеличении объема словаря V параметр α количественно будет иметь тенденцию к незначительному, но постепенному уменьшению, и наоборот, при экстраполяции "назад" α-параметр увеличивается. Зависимость γ-параметра от объема словаря V была показана выше. Если требуется повышенная точность максимальных экстраполяционных расчетов, то можно показать, как и для случая γ-параметра, что в более точной репрезентации α-параметр должен принять вид

$$\alpha_0 \exp h V,$$

где h - также эмпирический параметр, смысл которого раскрывается ниже. Тогда ранговая модель (8) принимает конечный вид

$$F_i = k\alpha_0 \exp h V \cdot (i+p)^{-\gamma \exp di} \qquad (9)$$

Иными словами, в качестве "платы за точность" исходная ципфовская модель обрастет дополнительными параметрами, которые обладают свойством последовательного вычисления, свертки и укрупления в соответствии с задачами конкретного исследования, критерия точности и интерпретации. Усложненность модели (9) только кажущаяся, так как операционально она представляет собой последовательность элементарных математических вычислений. Определенная устойчивость и одновременно динамичность α-параметра имеет статистический и лингвистический смысл. Абсолютное постоянство величины α означало бы, что малая, большая и сверхбольшая выборки строятся из абсолютно однородных составляющих с одинаковой дисперсией частот, что в условиях лингвостатистического эксперимента и практики осуществить принципиально невозможно. Статистически более достоверной и, следователь-

но, относительно более однородной является бóльшая выборка. Лингвостатистически подъязык как лексико-терминологическая система предстает как функциональная структура, характеризующаяся не только некоторой мерой внутренних связей, которые в существенной степени определяют характер распределения частот, но и внешними связями с некоторыми другими подъязыками. Если считать, что внутренние связи лингвостатистической структуры относительно разнообразнее и сильнее внешних, которые обуславливают появление в выборке "инородных" элементов, то следует ожидать того, что качественно и количественно эти внешние связи будут исчерпываться относительно быстрее внутренних в последовательно увеличивающейся выборке. Тогда параметр k можно интерпретировать как меру относительного повышения однородности подъязыка в последовательно возрастающем информационном потоке, который данный подъязык обеспечивает.

В целом в модели (9) проявляется соотновительный характер всех ципфовских параметров, когда увеличение глубины представления одного из них ведет к раскрытию и расширению описания его связей с другими лингвостатистическими параметрами, к включению их в область представления данного параметра. Важным является и то, что выявленная взаимосвязь параметров модели и возможность их логической развертки открывает дополнительные возможности для лингвистического осмысления модели в целом, ее отдельных фрагментов и их взаимосвязей. Тот факт, что в модели (9) параметры носят выводимый, конструктивный характер, позволяет вычислять их единообразно и соотносить с соответствующими параметрами модели в приложении к разным подъязыкам и языкам, представленными неравными по объему выборками. Отдельный параметр модели численно выступает как функция другого или нескольких других. По-видимому, скользящий, динамический характер γ-параметра связан с тем обстоятельством, что лингвостатистические частоты и ранги языковых элементов находятся между собой не в парной, как это обычно считается, а многосторонней, опосредованной корреляционной связи. Например, в качестве третьего и четвертого взаимосвязанных, но существенно независимых элементов частотно-ранговой структуры выступают объем частотного словаря и величина их средней частоты, что соответствует лингвистическому представлению о многомерности и динамичности структуры языковых элементов и их связей. Модель (9) и ее свернутые аналоги представляют собой логическое развитие модели (1), а также тезиса о нелинейности частотно-ранговых распределений в языке, которые положены в основу этой модели, хотя и отличаются от последних пониманием специфики зависимости γ-параметра от величины ранга i, и, следовательно, возможностями вывода, вычислений и лингвистической интерпретации.

Необходимо особо подчеркнуть, что обе нелинейные частотно-ранговые модели не противоречат и ни в коей мере не отрицают закон Ципфа как таковой. Если "кано-

ническая", линейная по своей математической сути модель Ципфа-Мандельброта обусловлена условиями ее формального вывода из минимального числа наиболее общих и абстрактных лингвостатистических условий, что и обеспечивает ей статус теоретического закона, то нелинейные модели (1) и (9) представляют собой привязку этого абстактного и общего к конкретной языковой эмпирии, которая характеризуется широким диапозоном вариативности ее элементов и многомерностью их связей. Методологически лингвостатистический обобщающий процесс внутренне противоречив. Собственно статистико-лингвистическое абстрагирование как формальный процесс отвлечения от языкового многообразия выражается в максимально упрощенных по математической структуре выражениях. Лингвистико-статистическое обобщение представляет собой обратный процесс - включение в область рассуждений все более полного набора лингвистически существенных составляющих, которые получают репрезентацию в дополнительных параметрах исходной математической модели. Очевидно, что выбор конкретного обобщающего выражения диктуется в основном прагматическими соображениями и поэтому не обладает правом абсолютной истины.

ЛИТЕРАТУРА

Алексеев П.М. Методика квантитативной типологии текста. Л., 1983.

Арапов М.В., Ефимов Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. - НТИ, сер. 2. М., 1975, № 1, с. 9-20.

Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней.- В кн.: Вычислительная лингвистика. М.: Наука, 1976, с. 179-202.

Орлов Ю.К. Статистическое моделирование речевых потоков. - В кн.: Вопросы кибернетики. Вып. 45. М., Л., 1978, с. 66-99.

Хомский Н., Миллер Дж. Конечные модели использования языка. - Кибернетический сборник. Новая серия. М., 1967, № 4.

Яблонский А.И. Стохастические модели научной деятельности. - В кн.: Системные исследования. Ежегодник 1975. М.: Наука, 1975, с. 5-42.

Herdan G. Quantitative Linguistics. London, 1964.

Mandelbrot B. Final Note on a Class of Scew Distribution Functions: Analysis and Critique of a Model due to H.A. Simon. - Information and Control, 1961, vol. 4, p. 198-216.

ON THE PROBLEM OF GENERALIZATION AND INTERPRETATION OF RANK DISTRIBUTIONS IN STATISTICAL LINGUISTICS

Valery Bychkov

Summary

The article deals with a set of linguostatistical rank-frequency models in the aspect of their formal generalizations and linguistic interpretations. The author points out the insufficiency of the existing linear monoparametrical models in solving practical linguostatistical problems where greater exactness and reliability are needed. The rank-frequency relation and its dynamics in an ever-increasing sample is considered to be a result of interaction of several linguistic and extralinguistic factors formalized by means of additional parameters. The paper describes an operation of including such essential parameters into the initial Zipfian model which converts it into non-linear ones and secures a better fit between theoretical and empirical data.

ОДНОРОДНОСТЬ ТЕКСТОВ ОТНОСИТЕЛЬНО ЧАСТОТ ВСЕГО РЯДА НЕМЕЦКИХ ГРАФЕМ

Б.Н. Гвоздович

Проблема однородности текстов относительно частот лингвистических объектов или, что то же самое, проблема стабильности лингвистических частот занимает центральное место в современной лингвистической статистике. Ее важность определяется тем, что постулат об обязательной стабильности лингвистических частот, на котором построено, по сути дела, все здание лингвостатистики, в ряде случаев не находит подтверждения (см., например, Савицкий Н.П., 1966). Это обстоятельство выдвигает на передний план задачу экспериментальной проверки лингвистических частот на стабильность, т.к. "без установления на самом различном материале и при помощи самых различных статистических процедур того, насколько стабильны или нестабильны частоты лингвистических элементов, дальнейшее развитие лингвистической статистики как лингвистической науки невозможно." (Сегал Д.М., 1972, с. 83).

В настоящей работе описывается эксперимент по проверке однородности текстов относительно частот немецких графем. Задачей эксперимента было выяснить, являются ли тексты однородными по отношению к частотам всего ряда немецких графем, не касаясь при этом вопроса о стабильности частот отдельных графем.

Выбор графемы как объекта исследования определило то обстоятельство, что факты нестабильности частот зафиксированы преимущественно для таких единиц языка, употребление которых подвержено определенному влиянию со стороны говорящего, а именно, для слов и классов слов (см., в частности, Лесскис Г.А., 1964). В то же время частоты других единиц языка, например, фонем, выбор которых значительно меньше зависит от воли говорящего, оказываются сравнительно стабильными - во всяком случае частоты некоторых из них стабильны безусловно (см. Сегал Д.М., 1972, с. 136 и след.). Напрашивается предположение, что шансы на стабильность лингвистических частот врзрастают по мере того, как уменьшается зависимость употребления языковых единиц от воли человека.

Графемы, как известно, служат для обозначения фонем, но отношения между фонемами и графемами носят далеко не однозначный характер. Фонемы могут обозначаться на письме по-разному: только одной определенной графемой, например, нем. /l/, двумя разными графемами, например нем. /f/, сочетаниями графем, например, нем. /ʃ/, а также еще более сложным способом - и графемами и их сочетаниями, например, нем. /i:/.

Не менее разнообразны и отношения графем к фонемам. Графемы могут обозначать только одну определенную фонему, например, нем. M, две разных фонемы, например, нем. V, сочетание двух разных фонем, например, нем. X, дифференциальные признаки фонем, например, русск. Ь, а в ряде случаев - и фонемы, и их дифференциальные признаки одновременно, например, нем. E.

Поэтому графемы, очевидно, более независимы от смысла высказывания, характера текста, индивидуальных особенностей стиля его автора и т.д., чем фонемы. Исследование стабильности их частот позволит, таким образом, не только дать соответствующую характеристику еще одной группе языковых единиц, но и проверить высказанное выше предположение.

Исследование проводилось на выборке из произведений современных немецких писателей, газет и научных журналов общим объемом в 50.000 графем /500 отрывков по 100 графем/. При этом из каждого произведения* извлекалось 10 отрывков, которые представляли разные части произведения - его начало, середину и конец - и отделялись друг от друга разными интервалами : от 100 до 10.000 графем. Длина интервала увеличивалась по направлению от начала или конца произведения к его середине. В выборке равными долями по 100 отрывков /10 произведений/ были представлены три основных литературно-художественных жанра, а именно, проза, поэзия и драма, а также оба функциональных стиля речи, тесно связанные с письменной формой языка - научный и газетный стили.

Такое построение выборки позволило одновременно с вопросом о стабильности частот всего ряда графем вообще проверить, меняются ли шансы на стабильность:

- внутри одного и того же произведения по мере увеличения интервала между сравниваемыми отрывками текста;

- внутри одного и того же произведения в зависимости от того, какую часть произведения они представляют: начало, середину или конец;

- между разными произведениями, относящимися к одному и тому же литературно-художественному жанру или функциональному стилю.

В результате подсчетов было получено 500 частотных рядов, репрезентирующих статистическое поведение немецких графем в разных частях указанных произведений. Задачи эксперимента решались на основе сопоставления этих рядов.

Основным инструментом исследования служил критерий согласия "хи-квадрат", с помощью которого проверялась нулевая гипотеза, состоящая в предположении,

* Словом "произведение" здесь и далее ради единообразия наряду с законченными произведениями, например, пьесами, научными статьями, называются также главы романов и большие отрывки газетного текста, например, полосы.

что различия, наблюдаемые между сопоставляемыми рядами, являются несущественными, и их можно отнести за счет различий между эмпирической и генеральной совокупностями. Принятие такой нулевой гипотезы означает, что сравниваемые частотные ряды взяты из одной и той же генеральной совокупности, а, следовательно, изучаемые частоты носят стабильный характер. Ее отвержение дает основание говорить о нестабильности этих частот.

Заключение о принятии или отвержении нулевой гипотезы делалось каждый раз после сравнения найденного значения χ^2 с критическим при соответствующем числе степеней свободы (см. Урбах В.Ю., 1964, с. 395): нулевая гипотеза принималась, если значение χ^2 не превышало критического при 1%-ном уровне значимости, и отвергалась, если оно превышало критическое при этом уровне значимости. При таких условиях вероятность того, что будет отвергнута справедливая нулевая гипотеза, составляет всего один процент (Урбах В.Ю., 1964, с. 395).

Эксперимент проводился в несколько этапов. Вначале исследовался вопрос о стабильности частот всего ряда немецких графем внутри одного произведения. С этой целью сопоставлялись частотные ряды, извлеченные из одного и того же произведения, а именно, первый ряд /начало произведения/ последовательно сравнивался с девятью остальными, расположенными на все большем удалении от него и представляющими другие части произведения - середину и конец. Значение критерия "хи-квадрат" вычислялось при этом по формуле:

$$\chi^2 = \sum \frac{n_1 - n_2}{n_1 + n_2} \qquad (1)$$

где n_1 - частота графемы в 1-м ряду, а n_2 - ее частота во втором сравниваемом ряду.

Всего было проведено 450 таких сопоставлений - по 9 на каждое из 50 произведений. Во всех случаях найденное значение χ^2 позволило принять нулевую гипотезу. В табл. 1 в качестве примера приведены значения χ^2, найденные при сравнении частей некоторых произведений* /числитель/ и число степеней свободы этого критерия /знаменатель/.

Такой результат достаточно однозначно говорит

* W o l f Ch. Der geteilte Himmel. Mitteldeutscher Verlag Halle (Saale), 1963; S. 294 ff; S o l o m o n H. Loorbass. - In : Neue Stücke. Henschelverlag Berlin, 1971, S. 391 ff.; W e i n e r t E. Gedichte. Verlag Philipp Reclam jun. Leipzig, S. 68 ff.; N e u e s D e u t s c h l a n d. 31.1.1974, S. 1; S t a a t u n d R e c h t. 1972, No. 2, S. 211 ff.

в пользу стабильности частот всего ряда немецких графем в рамках одного произведения. Но чтобы быть в этом окончательно уверенным, необходимо сопоставить не только первый отрывок со всеми остальными, но и остальные отрывки между собой. Эту проверку, однако, которая крайне трудоемка, можно заменить одновременной проверкой на стабильность всех 10 частотных рядов, представляющих одно и то же произведение. Такая однократная совместная проверка применяется в статистике в тех случаях, когда имеется несколько эмпирических совокупностей, о которых предполагается, что все они являются выборками из одной и той же генеральной совокупности (см. Урбах В.Ю., 1964, с. 234). В ходе проверки подвергается испытанию нулевая гипотеза об отсутствии существенных расхождений между всеми сравниваемыми эмпирическими совокупностями.

При однократной совместной проверке на стабильность всех частотных рядов, извлеченных из одного и того же произведения, значение x^2 вычислялось отдельно для каждого ряда по формуле:

$$x^2 = \sum \frac{(n_i - n_i^o)^2}{n_i^o} \qquad (2)$$

где n_i -эмпирическая частота i-ой графемы, а n_i^o - ее же теоретическая /средняя/ частота в данном произведении. Полученные значения x^2 суммировались. В связи с большим числом степеней свободы критерия "хи-квадрат" при такой проверке* заключение о принятии или отвержении нулевой гипотезы делалось с помощью отношения Романовского:

$$\frac{/x^2 - K/}{2K} \qquad (3)$$

где К - число степеней свободы (см. Романовский В.И., 1939, с. 96). Нулевая гипотеза принималась, если это отношение было меньше трех, и отвергалось, если оно равнялось или превышало три.

Результаты проверки /некоторые из них для примера приведены в табл. 2/ подтвердили первоначальный вывод о стабильности исследованных частот, т.к. нулевая гипотеза в ходе этой проверки не была отвергнута ни разу. Таким образом можно уверенно говорить о том, что любые два отрывка одного и того же произведения являются однородными относительно частот всего ряда немецких графем вне зависимости от длины интервала, отделяющего эти отрывки друг от друга, и от места, которое отрывок занимает в рамках данного произведения.

---

* Оно получалось как результат перемножения числа сравниваемых рядов, уменьшенного на единицу, и уменьшенного на единицу числа графем в столбце, итоговом для данного произведения.

Таблица 1

ЗНАЧЕНИЯ КРИТЕРИЯ "ХИ-КВАДРАТ", НАЙДЕННЫЕ ПРИ СРАВНЕНИИ ЧАСТЕЙ ОДНОГО ПРОИЗВЕДЕНИЯ, И ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ

Автор и произведение	Сравниваемые отрывки								
	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10
К. Вольф. Разделенное небо	19.74 22	19.91 22	24.38 22	24.29 21	28.12 21	25.44 21	14.92 21	11.61 22	21.31 22
Г. Соломон. Шелопай	17.40 20	14.40 21	12.37 20	37.08 21	25.87 20	15.79 20	23.58 21	21.50 20	24.24 19
Э. Вайнерт. Стихи	22.41 21	15.35 21	10.52 21	15.24 21	13.70 21	16.31 22	8.91 22	23.35 22	19.46 21
Нойес Дойчланд	34.67 21	21.07 21	25.48 20	22.49 22	18.65 20	25.17 20	25.21 19	24.39 21	33.84 20
Государство и право	22.23 22	15.62 22	18.38 22	18.04 22	15.84 20	20.00 21	20.04 21	20.65 21	19.12 20

Таблица 2

ДАННЫЕ, ХАРАКТЕРИЗУЮЩИЕ СТАБИЛЬНОСТЬ ЧАСТОТ ВСЕГО РЯДА НЕМЕЦКИХ ГРАФЕМ ПРИ ОДНОВРЕМЕННОМ СРАВНЕНИИ ВСЕХ ЧАСТЕЙ ОДНОГО ПРОИЗВЕДЕНИЯ

Автор и произведение	Значение χ^2	Число степеней свободы	Отнош. Романовского
К. Вольф. Разделенное небо	170.34	198	1.39
Г. Соломон. Шелопай	211.37	207	0.21
Э. Вайнерт. Стихи	209.84	207	0.13
Нойес Дойчланд	213.22	207	0.30
Государство и право	159.61	216	2.07

На втором этапе эксперимента исследовался вопрос о стабильности частот всего ряда немецких графем при переходе от одного произведения к другому в пределах одного и того же литературно-художественного жанра или функционального стиля. Для этого частотные ряды, представляющие различные части одного произведения, были объединены, а полученные совокупности, относящиеся к одному и тому же жанру или стилю, попарно сравнивались друг с другом с помощью критерия "хи-квадрат", который вычислялся по формуле /1/. Результаты сравнения отражены в табл. 3.

Таблица 3

РЕЗУЛЬТАТЫ СРАВНЕНИЯ ДРУГ С ДРУГОМ ПРОИЗВЕДЕНИЙ ОДНОГО ЖАНРА/СТИЛЯ/

Жанр /Стиль/	Число сравнений	Количество случаев, когда нулевая гиротеза	
		принималась	отвергалась
Проза	45	43	2
Драма	45	37	8
Поэзия	45	43	2
Газета	45	45	-
Наука	45	45	-
Всего	225	213	12

Большое число случаев отвержения нулевой гипотезы /12 из 225/ в целом свидетельствует о нестабильности частот всего ряда немецких графем при переходе от од-

ного произведения к другому в рамках одного и того же литературно-художественного жанра или функционального стиля. В то же время обращает на себя внимание тот факт, что все случаи непринятия нулевой гипотезы приходятся на сферу художественной литературы, а при сравнении между собой "нехудожественных" произведений частоты оказались стабильными.

В качестве одного из объяснений этого факта можно предположить, что существует какая-то зависимость между частотами всего ряда графем и самобытностью текста: там, где тексты подвергаются определенному выравниванию, в первую очередь лексическому, частоты при переходе от одного произведения к другому оказываются стабильными; там же, где такого выравнивания не бывает, они нестабильны.

ЛИТЕРАТУРА

Лесскис Г.А. О зависимости между размером предложения и его структурой. - Вопросы языкознания. 1964, № 3.

Романовский В.И. Элементарный курс математической статистики. М.; Л., 1939.

Савицкий Н.П. Об устойчивости относительных частот лингвистических элементов. - Československá russistika, 1966, № 4.

Сегал Д.М. Основы фонологической статистики. М., 1972.

Урбах В.Ю. Биометрические методы. М., 1964.

THE HOMOGENEITY OF TEXTS AS REGARDS THE WHOLE SERIES OF GERMAN GRAPHEMES

Boris Gvozdovitch

Summary

The article deals with the results of the investigation of the homogeneity of texts with regard to frequency of the system of all German graphemes. The author establishes that two texts within the limits of one and the same work are always homogeneous as regards these frequencies. Texts representing different works are not homogeneous as a rule, this being true primarily in the case of extracts from fiction. In this connection the author supposes that there exists a dependence between text homogeneity and text originality.

К ВОПРОСУ О ПРИМЕНЕНИИ ТЕОРИИ СЛУЧАЙНЫХ ФУНКЦИЙ ПРИ ИЗУЧЕНИИ КВАНТИТАТИВНЫХ ОСОБЕННОСТЕЙ ЛИНГВИСТИЧЕСКИХ СИСТЕМ

(на примере терминологической системы английского подъязыка физики)

Н.С. Манасян

Теорию случайных функций и случайных процессов, которая за последние десятилетия получила интенсивное развитие и уже имеет многочисленные приложения в технике, как кажется, можно широко использовать при прогнозировании событий в разного рода вероятностно-лингвистических, инженерно-лингвистических и информационных задачах.

Случайную лингвистическую последовательность, как и любую случайную последовательность, можно определить как функцию от неслучайного аргумента t, значение которой представляет собой случайную величину; или, "случайной функцией называется функция, которая в результате опыта может принять тот или иной конкретный вид, неизвестно заранее, какой именно" (см. Вентцель, 1969, с. 370).

Если из различных подъязыков некоторого более общего подъязыка, который принимаем за генеральную совокупность, случайным образом отобраны n выборок, являющиеся подмножествами этой генеральной совокупности, и по этим выборкам составлены частотные словари (ЧС) по определенному лингвистическому признаку, как, например, в нашем случае для однословных терминов (ОТ), то получим n реализаций случайной последовательности

$$x_1(t), x_2(t), \ldots, x_n(t). \qquad (1)$$

При этом мы делаем следующие предположения.

1. Отбор терминов в ЧС является случайным, т.е. до осуществления отбора каждая терминологическая единица обладает определенной, заранее заданной вероятностью быть включенной в выборку.

2. Отбор вышеназванных лингвистических единиц осуществляется по определенному принципу (см. Шварц, с. 9). Такой отбор называется направленным отбором.

3. Будем пренебрегать объективными ошибками составителей ЧС, а также субъективным подходом составителей ЧС к отбору терминов в словник словаря.

Отметим еще, что составление ЧС или получение соответствующего вариационного ряда является независи-

мым испытанием, так как вероятность того или иного исхода каждого из опытов не зависит от того, какие исходы имели другие опыты.

Перейдем к вычислению основных, простейших характеристик случайных функций: математического ожидания, среднеквадратического отклонения и корреляционной функции. Подробное описание методики вычисления этих величин дано, например, в (Вентцель, с. 377 и сл.; Смирнов, Дудин-Барковский, с. 169-172).

В настоящее время в распоряжении лингвостатистики имеется четыре ЧС английских физических текстов, представление статистических данных в которых делает возможным их изучение с точки зрения статистических распределений и теории случайных функций. Это ЧС подъязыков электроники (Э) (Алексеев); физики твердого тела (ФТТ) (Алексеев, Каширина, Тарасова); физики элементарных частиц (ФЭЧ) (Алексеев, Каширина, Тарасова); квантовой электроники (КЭ) (Манасян). Эти словари, составленные по единой методике, принятой в группе "Статистика речи", имеют одинаковую величину выборки, на основе которой они составлены, - 200 тыс. словоупотреблений. Учитывая тот факт, что словари представляют один общий подъязык, физику, а также единую методику составления и одинаковую величину выборки, можно утверждать, что эти четыре словаря являются четырьмя реализациями случайной последовательности $X(t)$ и сопоставимы друг с другом и с генеральной совокупностью, подъязыком физики. Следует отметить, что вариационные ряды этих ЧС имеют аналогичный характер (см. Прил. табл. 1, где приводятся фрагменты выборочных данных до частоты 120, и рис. 1-4 Прил.).

По данным табл. 1 Прил. и по формулам

$$\tilde{m}_x(t_z) = \frac{\sum_{i=1}^{n} x_i(t_z)}{n}, \qquad (2)$$

$$\tilde{\mathcal{D}}_x(t_z) = \frac{n}{n-1}\left\{\frac{\sum_{i=1}^{n}[x_i(t_z)}{n} - [\tilde{m}_x(t_z)^2\right\}, \qquad (3)$$

$$\tilde{\sigma}_x(t_z) = \sqrt{\tilde{\mathcal{D}}_x(t_z)}. \qquad (4)$$

вычислим оценки математического ожидания $m_x(t_z)$, дисперсии $\mathcal{D}_x(t_z)$ и среднеквадратического отклонения. Значения $\tilde{m}_x(t_z)$ и $\tilde{\sigma}_x(t_z)$ занесены в табл. 2 Прил. При вычислениях учитывается частота 1. Диапазон частот берется от 1 до 30 по причинам, изложенным в (Манасян,

1981, с. 67). Дальнейшая наша задача заключается в том, чтобы математическое ожидание и среднеквадратическое отклонение "наилучшим" образом аппроксимировать непрерывной функцией.

Как видно из данных табл. 2 Прил., оценки математического ожидания $\tilde{m}_x(t)$ и среднеквадратического отклонения $\tilde{\sigma}_x(t)$ являются убывающими функциями от неслучайного аргумента t. Это наглядно видно из Рис. 1, где по данным табл. 2 Прил. изображены последовательности $\tilde{m}_x(t_x)$ и $\tilde{\sigma}_x(t_x)$. Ломаная 1 на рисунке изображает оценку математического ожидания случайной функции, а ломаная 2 - оценку среднеквадратического отклонения. Обе последовательности с начала спектра частот резко убывают, а потом более или менее плавно асимптотически приближаются к оси абсцисс, оставаясь все время выше нее. Обе функции положительны - математическое ожидание по своему лингвистическому смыслу, о среднеквадратическое отклонение - по определение. Это наводит на мысль, что эту функцию можно аппроксимировать некоторой непрерывной монотонно убывающей функцией так, чтобы эта аппроксимация была наилучшей в некотором смысле. Необходимость такой аппроксимации вытекает из того, что во всяком статистическом, как и лингвостатистическом, распределении, а еще точнее, в распределении терминов, неизбежно присутствуют элементы случайности, связанные со многими причинами, напрмер, с тем, что число наблюдений ограничено; взяты эти выборки, а не те, давшие результаты определенного характера. Все закономерности для случайных величин проявляются при достаточном числе наблюдений. В лингвистических исследованиях мы почти никогда не имеем дела с таким большим числом наблюдений, и в нашем случае мы вынуждены считаться с элементом случайности. Поэтому для обработки нашего эмпирического материала нужно подобрать такую теоретическую кривую, которая выражала бы лишь его существенные черты, но не случайности, связанные с недостаточным объемом экспериментальных данных. Для этой цели была предпринята аппроксимация данных при помощи семейства кривых следующих видов

$$ae^{-a_1 t},\ ae^{\frac{a_1}{t}},\ ae^{\frac{a_1}{t^2}},\ at^{a_1}. \qquad (5)$$

Проверка, проведенная при помощи метода наименьших квадратов показала, что наилучшим образом аппроксимирует наши данные функция вида at^{a_1}. Итак, функцию аппроксимирующую математическое ожидание, будем искать в виде

$$x(t) = a t^{a_1} (a>0, a_1<0), \qquad (6)$$

где a и a_1 - параметры, значение которых будут определяться из наших лингвостатистических наблюдений; т.е. $m_x(t) = ae^{a_1}$. Логарифмируя обе части (6), получим

$$\ln x(t) = \ln a + a_1 \ln t.$$

Введя обозначения

$$\ln x(t) = \eta, \ln(t) = \xi, \ln a = a_0, \qquad (7)$$

получим

$$\eta = a_0 + a_1 \xi \qquad (8)$$

А это значит, что на логарифмической плоскости зависимость между $x(t_k)$ и t_k должна быть линейной, т.е. точки $x(\xi_k, \eta_k)$ должны располагаться по прямой. Посредством некоторых преобразований (ср. Вентцель, с. 357 и сл.; Манасян, 1981; с. 64-65) получим

$$\begin{cases} a_0 + S_1 a_1 = V_0 \\ S_1 a_0 + S_1 a_1 = V_1. \end{cases} \qquad (9)$$

Из системы нормальных уравнений (9) легко найти оценки параметров функции (8)

$$\tilde{a}_0 = \frac{V_0 S_2 - V_1 S_1}{S_2 - S_1^2}, \qquad (10)$$

$$\tilde{a}_1 = \frac{V_1 - S_1 V_0}{S_2 - S_1^2}. \qquad (11)$$

Найдя $\tilde{a}_0$ и $\tilde{a}_1$ и замечая, что параметр $a = e^{\tilde{a}}$ (см. (7)), получим оценку аппроксимирующей функции

$$\tilde{m}_x(t) = \tilde{a} t^{a_1}. \qquad (12)$$

Вычисления располагаем в расчетный бланк (см. Прил. табл. 3), вычислительная схема которого описана в работах, цитированных выше. По данным этой таблицы строим точки (ξ_k, η_k) (см. Рис. 2). Как видно из этого рисунка, точки довольно близко располагаются вокруг некоторой прямой, что дает уверенность в том, что при-

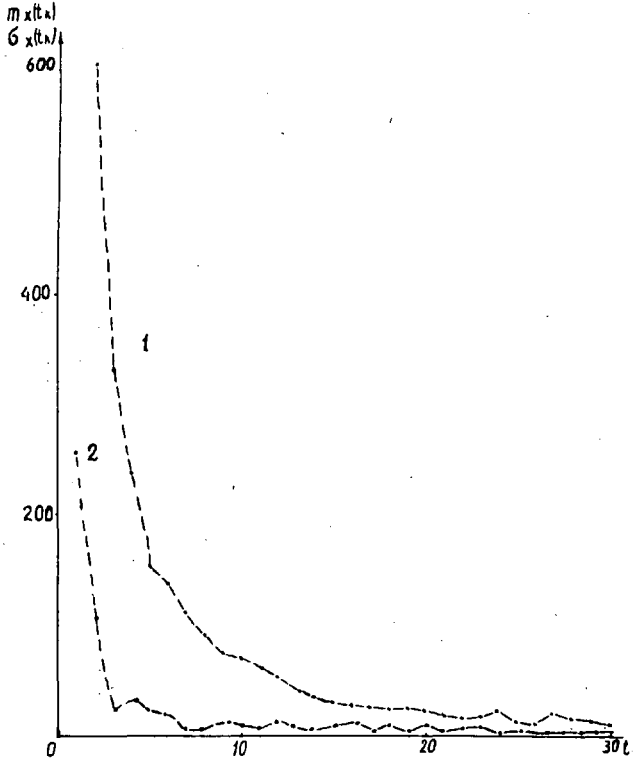


Рис.I. Распределение математического ожидания /I/ и среднеквадратического отклонения /2/ ОТ английского подъязыка физики.

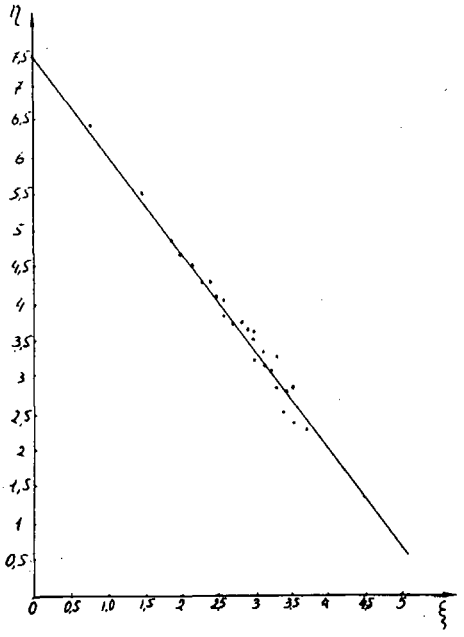


Рис.2. Выпрямленный график математического ожидания случайной последовательности ОТ.

менение метода наименьших квадратов приведет к хорошим результатам.

По способу, изложенному в (Пулькин, с. 181), графически определяем оценки параметров аппроксимирующей функции. Из Рис. 2 имеем, что

$$a_0 \approx 7.35; \quad \tilde{a}_1 = e^{\tilde{a}_0} = 1556.$$

Беря точки M_1 (0,69; 6,41) и M_2 (2,20; 4,32), соответствующие частотам 2 и 19, найдем

$$\tilde{a}_1 \approx -\frac{2.08}{1.50} = -1.39.$$

Тогда оценка математического ожидания приближенно выражается формулой

$$\tilde{m}_x(t) = 1556 \cdot t^{-1.39}. \qquad (13)$$

Полученные графическим способом данные являются ориентировочными и в какой-то мере могут служить для контролья последующих вычислений. Приступим к вычислению коэффициентом нормальной системы (9). Для этого, используя данные табл. 3 Прил., найдем суммы чисел в строках 3-6

$$\sum \xi_\tau = 74{,}66; \qquad \sum \xi_\tau^2 = 206{,}77;$$
$$\sum \eta_\tau = 117{,}79; \qquad \sum \xi_\tau \eta_\tau = 8{,}81.$$

Далее находим

$$S_1 = \frac{1}{30}\sum \xi_\tau = 2{,}49; \qquad \sum \xi_\tau^2 = 6{,}89;$$
$$V^0 = \frac{1}{30}\sum \eta_\tau = 3{,}93; \qquad \sum \xi_\tau \eta_\tau = 8{,}81.$$

Подставляя эти значения в формулы (10) и (11), получаем $\tilde{a}_0 = 7{,}32$; $a_1 = -1{,}37$. По формуле (12) находим

$$\tilde{a} = e^{7.32} \approx 1516$$

Подставляя полученные значения в (12), получим оценку математического ожидания случайной функции $X(t)$

$$\tilde{m}_x(t) = 1516 \cdot t^{-1.37} \qquad (14)$$

Теоретические частоты, вычисленные по формуле (14), представлены в Прил. табл. 4.

Сравнивая (12) и (13), замечаем, что параметры, полученные методом наименьших квадратов, хорошо согласуются с данными, полученными графически. Таким образом, оценка математического ожидания выражается степенной функцией вида at^{a_1} где $a > 0$, $a_1 < 0$, т.е. математическое ожидание выражается функцией гиперболического типа. Этого и следовало ожидать, так как все оценки реализаций случайной последовательности выражаются функцией гиперболического типа (см. Манасян, 1982, с. 8-9). Естественно, что математическое ожидание, выражаемое формулой (12), тоже описывается кривой такого же типа.

Параметр $\tilde{a}$ = 1516 имеет определенный лингвистический смысл: при t = 1 $\tilde{m}_x(t)$ = 1516, т.е. значение параметра есть среднеожидаемое число ОТ с частотой, равной единице.

Перейдем теперь к сглаживанию среднеквадратического отклонения случайной последовательности $X(t)$ тем же методом, которым сглаживалось математическое ожидание. Результаты этих вычислений представлены в Прил. табл. 4. Для оценки среднеквадратического отклонения получаем следующую формулу

$$\tilde{\sigma}_x(t) = 141 \cdot t^{-1.14}. \qquad (15)$$

По данным табл. 4 Прил. видно, что в ряде σ_x наблюдаются значительные флуктуации частот. Это можно объяснить, во-первых, тем, что имеющихся ЧС английского подъязыка физики недостаточно, чтобы сгладить флуктуации в оценке $\tilde{\sigma}_x$, а во-вторых, они объясняются а) индивидуальным подходом к отбору терминов в ЧС каждым отдельным составителем словаря и б) особенностями каждого из четырех подъязыков. Очень возможно, что второе обстоятельство, включающее пункты "а" и "б", нивелировалось бы при условии наличия большего числа ЧС.

Перечисленными выше причинами объясняется то, что имеющие место флуктуации значений оценки $\tilde{\sigma}_x$ влияют в свою очередь на значения параметров a_1, a_0 и a, которые не всегда хорошо аппроксимируют значения $\tilde{\sigma}_x(t)$.

В Прил. табл. 5 представлены значения оценок среднеквадратического отклонения случайной функции $X(t)$, вычисленные по формуле (15). Сравнивая 2 и 3 строки табл. 5 Прил., замечаем, что здесь не всегда наблюдается хорошее согласие теоретических и эмпирических данных. Это особенно заметно для частот 1, 2, 3, 7, 8, 18, 23. А для некоторых частот (их более 50 % всего диа-

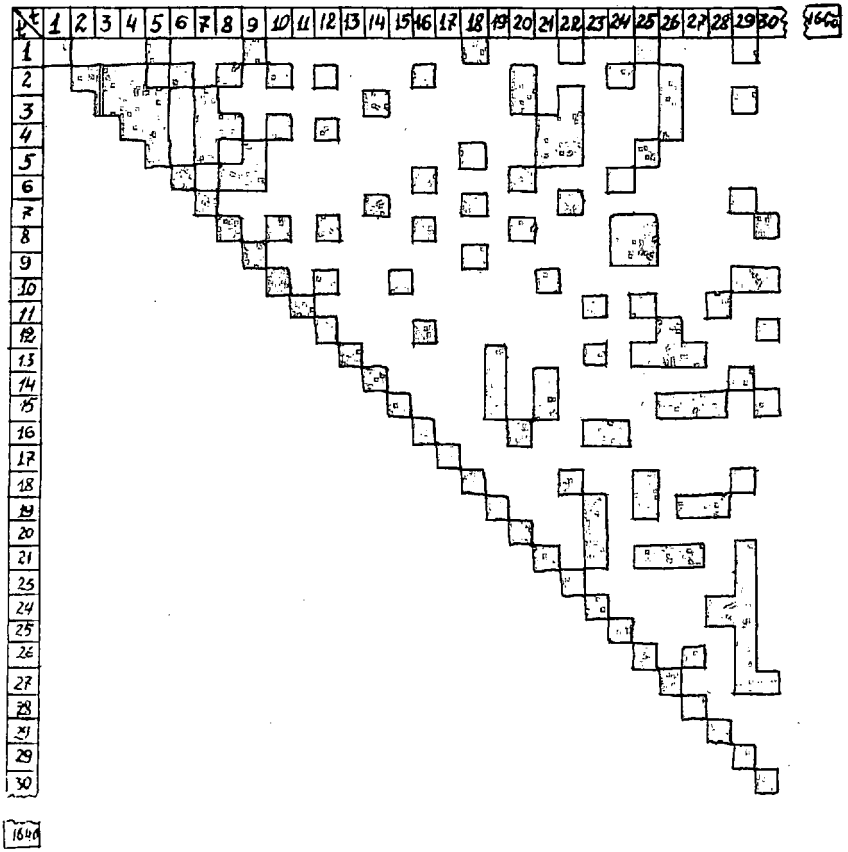


Рис. 3. Схема оценки элементов корреляционной матрицы /фрагмент/.

пазона) мы имеем поразительное совпадение эмпирических и теоретических данных.

Значительное расхождение в начале спектра объясняется причинами, изложенными выше. Кроме того, в пользу того, что распределение среднеквадратического отклонения имеет гиперболический характер, свидетельствует форма кривой эмпирического распределения $\sigma(t)$ (см. Рис. 1) и преобладание соответствий теоретических и эмпирических данных. Что касается построения доверительного интервала, то его можно построить аналогично интервалу математического ожидания (см. Смирнов, Дунин-Барковский, с. 238).

Теперь перейдем к вычислению элементов корреляционной матрицы, без которой нельзя сделать какие-нибудь качественные выводы. На Рис. 3 в схематической форме представлены оценки корреляционной матрицы (фрагмент), в которой вычислены значения $\rho_x(t_k, t_e)$, коэффициентов корреляции случайных величин (см. там же, с. 171) $X(t_k)$ и $X(t_e)$. Из-за экономии места на рисунке представлена часть значений этих величин. Черные квадратики обозначают, что данные случайные величины коррелированы. (Считается, что случайные величины коррелированы при $\rho_x(t_k, t_e) > 0{,}71$). Из схемы на рисунке видно, что некоторые элементы матрицы близки к единице, что указывает на существование функциональной зависимости случайных величин соответствующих сечений.

Наличие функциональной зависимости, по-видимому, свидетельствует о существовании семантических связей между ОТ с частотами t и t'. При этом сочетаемостью обладает не абсолютное большинство единиц с данными частотами, а преобладающее их количество; это могло бы подтверждаться тем, что величина $\rho_{tt'}$ почти никогда не бывает равна точно единице. Однако в действительности механизм сочетаемости ОТ настолько сложен, что корреляционную матрицу можно принимать лишь как грубую схему сочетаемости ОТ терминологической системы подъязыка. Проверка элементов корреляционной матрицы, начиная с частот 30-40, показывает, что начиная с этих частот, рассматриваемую последовательность можно считать приблизительно стационарной, т.е. если перейти к нормированной случайной величине с математическим ожиданием, равным нулю, и дисперсией, равной единице, то элементы корреляционной матрицы приблизительно удовлетворяют соотношению

$$K_{ij}(t + \tau) = K_{ij}(t),$$

где τ - произвольное целое число. Однако точная проверка этого требует огромного объема вычислительной работы, невозможной без привлечения электронно-вычислительной техники. Это может служить предметом дальнейших исследований статистических закономерностей в рассматриваемых лингвистических последовательностях.

На основании формулы доверительного интервала и при помощи дзета-функции Римана можно вычислить ожидаемое количество ОТ для каждого из физических подъязыков, исследуемых в настоящей работе, и всего подъязыка физики. Подробное описание этого будет дано в последующей работе автора.

ЛИТЕРАТУРА

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

Манасян Н.С. Частотный англо-русский словарь-минимум по квантовым генераторам. М., 1983.

Манасян Н.С. Статистические модели употребления терминов в специальном научно-техническом тексте. КД. Л., 1981.

Манасян Н.С. Статистические модели употребления терминов в специальном научно-техническом тексте. АКД. Л., 1982.

Пулькин С.П. Вычислительная математика. М., 1974.

Смирнов Н.В., Дудин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. М., 1965.

Шварц Г. Выборочный метод. М., 1978.

ПРИЛОЖЕНИЕ

Таблица 1

Вариационные ряды частот ОТ
в английском подъязыке
Э, ФТТ, КЭ, ФЭЧ и Φ^1 (до t = 120)

№№ пп			1	2	3	4	5	6	7	8
1	Э	x_1	1269	485	306	218	150	120	119	95
2	ФТТ	x_2	1387	562	339	220	159	127	112	96
3	КЭ	x_3	1567	718	371	287	183	149	109	86
4	ФЭЧ	x_4	948	664	331	229	129	166	117	89

	9	10	11	12	13	14	15	16	17	18
1	72	77	68	62	52	33	33	48	30	21
2	87	82	59	68	41	45	53	40	43	33
3	83	59	66	45	48	45	30	28	35	34
4	53	74	59	49	33	37	42	24	31	12

	19	20	21	22	23	24	25	26	27	28
1	32	39	22	24	39	27	14	14	18	14
2	27	27	20	18	17	28	15	15	24	19
3	31	21	27	17	20	25	13	8	17	15
4	12	21	20	24	16	23	18	13	23	18

	29	30	31	32	33	34	35	36	37	38
1	18	17	14	13	12	22	11	12	12	17
2	16	14	19	16	21	10	17	10	15	12
3	10	9	14	12	11	13	7	11	8	10
4	20	9	14	10	8	11	5	12	9	11

	39	40	41	42	43	44	45	46	47	48
1	11	7	8	13	10	11	11	10	8	8
2	9	8	2	7	2	4	4	6	5	6
3	6	10	12	8	5	7	4	7	6	5
4	3	11	4	7	10	6	5	6	2	5

	49	50	51	52	53	54	55	56	57	58
1	11	12	10	3	5	4	2	5	10	7
2	5	9	6	5	7	6	9	5	6	3
3	6	7	3	5	7	7	9	4	3	3
4	9	14	5	0	8	3	1	11	3	2

	59	60	61	62	63	64	65	66	67	68
1	4	6	7	5	3	6	2	4	6	2
2	5	5	4	1	5	2	1	7	3	7
3	5	3	3	5	1	3	6	1	2	4
4	2	2	1	5	2	3	5	6	2	5

	69	70	71	72	73	74	75	76	77	78
1	2	7	1	3	0	2	1	6	5	2
2	2	2	2	7	5	4	3	3	3	1
3	5	4	1	4	2	2	3	1	5	2
4	6	3	3	2	1	3	5	0	6	1
.	.	.	.	.	.	.	.	.	.	.

Прочерк при частоте 1 обозначает отсутствие данных в ЧС, послужившем источником.

Таблица 2

Вычисление математического ожидания $\tilde{m}_x$, дисперсии $\tilde{D}_x$ и среднеквадратического отклонения $\tilde{\sigma}_x$ случайной последовательности ОТ

	t	1	2	3	4	5	6
2	$\tilde{m}_x(t)$	1292.75	607.25	336.75	238.50	155.3	140.5
3	$\tilde{D}_x(t)$	67837.60	10825.9	721.59	1068.3	500.3	441.7
4	$\tilde{\sigma}_x(t)$	130.23	52.02	13.43	16.34	11.18	10.5

1	7	8	9	10	11	12	13	14
2	114.25	91.5	75.55	73	63	57.75	43.5	40
3	20.92	23	235.67	98	22	156.92	69.67	36
4	2.29	2.40	7.68	4.95	2.35	6.26	4.17	3

1	15	16	17	18	19	20	21	22
2	39.5	35	34.75	25	29.25	27	22.25	20.75
3	107	121.33	34.92	110	6.92	72	10.92	14.25
4	5.17	5.51	2.95	5.24	1.32	4.24	1.65	1.89

1	23	24	25	26	27	28	29	30
2	23	25.75	15	12.5	20.5	16.5	16	10.75
3	116.67	4.92	4.67	9.67	12.33	5.67	18.67	5.58
4	5.40	1.11	1.08	1.55	1.76	1.19	2.16	1.83

Таблица 3

Данные для расчета функции,
аппроксимирующей математическое ожидание
случайной последовательности ОТ
(метод наименьших квадратов)

1	t_τ	1	2	3	4	5	6
2	$\tilde{m}_x(t_\tau)$	1292.75	607.25	336.75	238.5	155.25	140.5
3	$\xi_\tau = \ln t_\tau$	0	0.69	1.10	1.39	1.61	1.79
4	$\xi 2_\tau = \ln^2 t_\tau$	0	0.48	1.21	1.92	2.59	3.21
5	$\eta_\tau = \ln \tilde{m}_x(t_\tau)$	7.16	6.41	5.82	5.48	5.04	4.95
6	$\eta_\tau \xi_\tau$	0	4.44	6.39	7.59	8.12	8.86

1	7	8	9	10	11	12	13	14
2	114.25	91.5	75.5	73	63	63	57.75	43.5
3	1.95	2.08	2.20	2.30	2.40	2.48	2.56	2.64
4	3.79	4.72	4.83	5.30	5.75	5.75	6.17	6.58
5	4.74	4.52	4.32	4.29	4.14	4.29	4.29	4.14
6	9.22	9.39	9.50	9.88	9.93	9.93	9.68	9.74

1	15	16	17	18	19	20	21	22
2	40	35	34.75	25	29.25	27	22.25	20.75
3	2.71	2.77	2.83	2.89	2.94	3.00	3.04	3.09
4	6.96	7.69	8.03	8.35	8.67	8.97	9.27	9.55
5	4.06	3.56	3.55	3.22	3.38	3.30	3.10	3.03
6	9.96	9.86	10.05	9.30	9.94	9.87	9.45	9.37

1	23	24	25	26	27	28	29	30
2	23	25.75	15	12.5	20.5	16.5	16	10.75
3	3.14	3.18	3.22	3.26	3.30	3.33	3.38	3.40
4	9.83	10.10	10.36	10.6	10.86	11.10	11.34	11.57
5	3.14	3.25	2.71	2.53	3.02	2.80	2.77	2.37
6	9.83	10.32	8.72	8.23	9.95	9.34	9.34	8.08

1	
2	
3	$\sum \xi_\tau = 74.65$ $\quad S_1 = 2.49$
4	$\sum \xi_\tau^2 = 206.77$ $\quad S_2 = 6.89$
5	$\sum \eta_\tau = 117.78$ $\quad V_0 = 3.93$
6	$\sum \eta_\tau \xi_\tau = 264.44$ $\quad V_1 = 8.81$

Таблица 4

Эмпирические $\tilde{m}$, $\tilde{\sigma}_{\tilde{m}}$ и теоретические $\tilde{m}_\tau$, $\sigma_{\tilde{m}}$ значения математического ожидания и среднеквадратического отклонения для случайной последовательности ОТ

t	1	2	3	4	5	6	7	8	9	10
$\tilde{m}$	1293	607	336	239	155	141	114	92	76	73
$\tilde{m}_\tau$	1516	588	338	229	168	131	106	89	75	65
$\tilde{\sigma}_{\tilde{m}}$	130	52	13	16	11	10	2	2	8	5
σ_m	110	48	29	21	16	13	11	9	8	7

t	11	12	13	14	15	16	17	18	19	20
$\tilde{m}$	63	58	44	40	40	35	35	25	30	27
m_τ	57	51	46	41	38	34	32	29	27	25
$\tilde{\sigma}_{\tilde{m}}$	2	6	4	3	5	6	3	5	1	4
$\sigma_{\tilde{m}}$	6	6	5	5	4	4	4	3	3	3

t	21	22	23	24	25	26	27	28	29	30
m	22	21	23	26	15	13	21	17	16	11
$\tilde{m}_\tau$	24	22	21	20	19	18	17	16	15	15
$\tilde{\sigma}_{\tilde{m}}$	2	2	5	1	1	2	2	1	2	1
$\sigma_{\tilde{m}}$	3	3	2	2	2	2	2	2	2	2

ON THE APPLICATION OF RANDOM FUNCTION THEORY WHEN STUDYING THE QUANTITATIVE PECULIARITIES OF LINGUISTIC SYSTEMS

Narynay Manasyan

Summary

A small amount of available terminological frequency dictionaries allow to use random functions theory in linguistical problems of different kind. The compilation of a frequency dictionary is presumed to be a realisation of a random process.

The article presents numerical calculations of random sequence characteristics: the estimates of expectation and standard deviation. It is possible to predict the frequency structure of sublanguage with the help of these characteristics. The calculation of correlation matrix elements enables to produce both quantitative and qualitative conclusions on functional and stochastic relations of random variable. The correlation matrix may be regarded as a diagram which represents semantical connections between one-word terms with corresponding frequencies.

ОПЫТ ДИАЛЕКТНОГО РАЙОНИРОВАНИЯ НА ОСНОВЕ АВТОМАТИЧЕСКОГО АТЛАСА ЛЕКСИКИ ГОВОРОВ

С. Муруметс

Диалектное районирование считается основным вопросом в эстонской диалектологии (Kask, 1965,с. 98). Уже в первых дошедших до наших дней фрагментах из книги напечатанной на эстонском языке в 1535-ом году, упоминается о том, что эстонский язык звучит по-разному по меньшей мере в пяти разных частях территории его распространения (Saareste, 1930, с. 80; Saareste, 1932, с. 22). Затем ценные наблюдения по этому вопросу были приведены в словарях А.В. Хупеля (Hupel, 1780, с. 5; Hupel, 1818, с.4) и в работах Ф.Й. Вийдеманна (Wiedemann, 1864, с. 1 - 3; Wiedemann, 1873, Wiedemann, 1875, с. 49 - 80).

В начале 1930-х годов в эстонской диалектологии впервые начинают применять количественный подход (Saareste, 1931). Первая попытка районирования эстонской территории по количеству совпадающих и различающихся черт диалектной речи (Saareste, 1932a) стала возможна после десятилетних систематических полевых работ (Saareste, 1932b). Обобщения основывались главным образом на морфофонетических признаках. Результаты А. Сааресте считались приемлемыми в общих чертах и через несколько десятилетий (Kask, 1965, с. 97). Первый количественный анализ чисто лексических границ в Эстонии, опубликованный в 1952-м году (Saareste, 1952), был основан на 94 лексических особенностях и был посвящен в основном самой главной границе между наречиями.

Ввод в ЭВМ первого тома "Краткого диалектного словаря" (VMS) открыл новые перспективы для лексического районирования эстонской территории. На магнитной ленте сохраняются 27 906 слов вместе с данными об их распространении по 115 территориальным единицам (в большинстве случаев совпадающим с бывшими приходами и соответствующим говору). Эту базу данных мы называем первой частью автоматического атласа эстонских говоров.

Любой метод анализа должен быть выбран в соответствии с материалом. Поскольку автоматический атлас эстонских говоров не основывается на результатах опроса по единому вопроснику, а на материале весьма различных источников (VMS, с. 5; Must, 1971), то таксономический подход, опирающийся непосредственно на анализе наличия или отсутствия разных диалектных черт, который применяется, например, Н.Н. Пшеничновой при классификации русских говоров (Пшеничнова, 1977), исключается. Более подходящим кажется методика, использованная А. Крикманном при фолк-

лорном районировании Эстонии на материале пословиц (Krikmann, 1979; Krikmann, 1980). По его примеру, связь между любыми двумя приходами у нас определяется по степени близости материала (словников). Степень близости каждой пары словников устанавливается по качеству пересечения, т.е. общих слов двух словников. Качество это, в свою очередь, вычисляется по шкале средней стереотипности (распространенности)/уникальности (редкости) слов, встречающихся в пересечении (в региональных единицах, т.е. в приходах). Затруднения возникают главным образом из-за того обстоятельства, что средняя стереотипность пересечения зависит от количества слов в нем, а количество слов в пересечении двух словников очень сильно зависит от количества слов в самих словниках (которые в использованном нами материале могут содержать от 94 до 4816 слов). Чтобы избавиться от нежелательного влияния количества на качество, была предпринята следующая процедура. Для каждого прихода через поле корреляций между абсолютными объемами и средними показателями стереотипности пересечений словника этого прихода со словниками всех остальных приходов была вычислена линеарная регрессия (по методу наименьших квадратов). Связь между любыми двумя приходами была определена как среднее значение из соответствующих двух отклонении от нормирующей регрессионной линии (по графикам обоих приходов).*

Затем приходы группируются по силе вычисленных связей: на карте по одному соединяют приходы друг с другом, пока вся территория не будет покрыта графами (присоединение прихода уже относящегося к одному графу, не допускается). Чтобы избавиться от "шума", которого не удалось "вычистить" нормированием, А. Крикманн соединяет только территориально соседние приходы. Первоначально группировка мной проводилась так же (Murumets, 1982; Murumets, 1984), но при моем материале более целесообразным оказался несколько иной путь (Murumets, 1983). Пришлось допускать соединение несоседних приходов (т.е. приходов без общей границы), если связи между ними оказывались сильнее, чем их связь с соседними. В то же время были исключены некоторые приходы (6 из 115) с мало репрезентативными словниками, у которых слишком большое количество связей с "несоседями" было сильнее, чем с "соседями". В результате была получена карта (рис. 1) с 19 группами приходов (и соответствующих говоров). Если считать силу связи между соседними говорами обратно пропорциональной "коммуникативному барьеру" между соответствующими приходами, то можно анализировать границы между группами говоров по отдельным отрезкам. Оказывается,

---

* В случае значительной корреляции между этой величиной и количеством общих слов в двух словниках процедуру следует повторить, заменив начальную стереотипность значением s*.

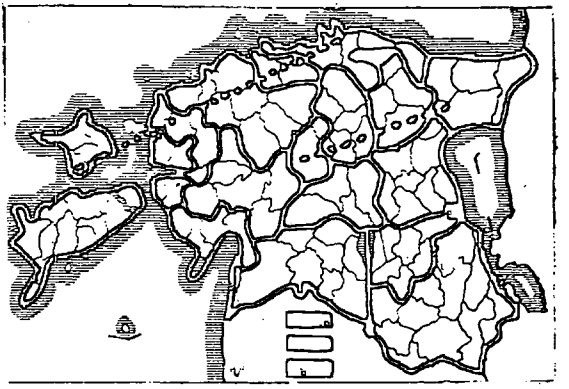


Рис. 1. Группы говоров.

что самые длинные из относительно удобных прямых "коридоров движения" диалектного слова расположены по линиям, обозначенным пунктиром на рис. 1.

Рисунок 2 получен в результате новой группировки групп на предыдущей карте, но дополнительно учтены данные как об отрезках границ, так и лексических дистанциях между всеми возможными парами изучаемых групп.

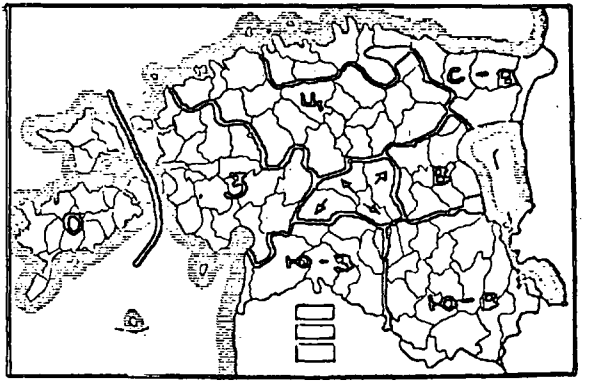


Рис. 2. Диалекты.

Карта иллюстрирует вывод о том, что эстонская территория разделяется на 8 частей, 7 из которых (острова,запад, юго-запад, юго-восток, восток, северо-восток и центральная часть северной Эстонии) характеризуются специфической диалектной лексикой. Но центральная часть территории Эстонии является слишком гетерогенной, чтобы называться самостоятельной диалектной зоной. Стрелки на карте указывают направления "коммуникативного центра тяжести" каждого их четырех приходов.*

До сих пор мы группировали приходы, исходя главным образом из одного параметра - попарной средней стереотипности s*. Но каждый приход представлен 114-элементным рядом из таких попарных средних и каждый из 115 таких рядов имеет свои среднее и экстремальные значения, стандартное отклонение, ассиметрию и эксцесс, по которым тоже можно провести классификацию, видимо, и районирование. Интересные результаты получены австрийским диалектологом Г. Гэблем, который сравнивает показатели асимметрии распределений связей, вычисленных по коэффициенту Фишера

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n \cdot s^3}$$

где x_i - значение отдельной связи, $\bar{x}$ - среднее значение всех связей одной территориальной единицы, n - число связей в ряду и s - стандартное отклонение.** Используя материал итальянского атласа (AIS), Г. Гэбль сделал вывод, что наибольший отрицательный показатель асимметрии характеризует распределения связей тех регионов, которые находятся в переходных диалектных зонах, а положительная асимметрия является признаком более самобытных и самостоятельных ареалов (Goebl, 1980, с. 77; Goebl, 1981, с. 401; Goebl, 1982, с. 46). В нашем материале все без исключения распределения имеют отрицательную асимметрию, что будто бы подтверждает выводы Г.Гэбля. Но так как у нас сила связей возрастает в порядке уменьшения их абсолютных значений, картина получается совершенно противоположной. На рис. 3 изображено территориальное расположение приходов с такими распределениями s*, коэффициент асимметрии которых ни-

* Здесь кроме качественного параметра учитывается и количество общих слов (см. карты 6-9 в Murumets, 1984). Метод был впервые использован в Murumets, 1981.

** В литературе этот коэффициент называется по-разному, напр. As в Лакин, 1980, с. 118, r_3 в Jalasto, 1978, с. 77.

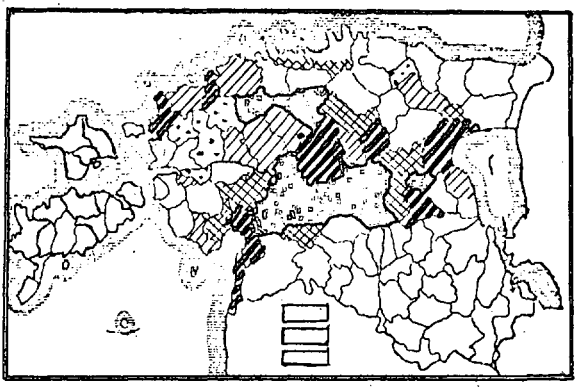


Рис. 3. Говоры с отрицательной асимметрией распределений связей.

же среднего. По карте видно, что значение коэффициента тем меньше, чем более центральным является географическое положение соответствующего прихода. Вокруг обширной переходной зоны можно выделить три больших территории - западная (ареал диалектов островов и часть ареала западного диалекта), северо-восточная (ареал прибрежного диалекта и примыкающая часть центрального), и южная (ареалы вырусского, тартусского и мульгиского диалектов).

Районы подразделяются на регионы при помощи анализа частотных полигонов отдельных распределений (рис.4). Для этого ряд связей каждого прихода был сегментирован таким образом, что интервалы измерялись стандартным отклонением от среднего. Интервалы были обозначены следующим образом:

Интервал	З н а ч е н и я
0	$s^* < \bar{s}^* - 4\sigma$
1	$\bar{s}^* - 4\sigma < s^* < \bar{s}^* - 3\sigma$
2	$\bar{s}^* - 3\sigma < s^* < \bar{s}^* - 2\sigma$
3	$\bar{s}^* - 2\sigma < s^* < \bar{s}^* - \sigma$
4	$\bar{s}^* - \sigma < s^* < \bar{s}^*$
5	$\bar{s}^* < s^* < \bar{s}^* + \sigma$
6	$\bar{s}^* + \sigma < s^* < \bar{s}^* + 2\sigma$
7	$\bar{s}^* + 2\sigma < s^* < \bar{s}^* + 3\sigma$
8	$\bar{s}^* + 3\sigma < s^* < \bar{s}^* + 4\sigma$
9	$\bar{s}^* + 4\sigma < s^*$

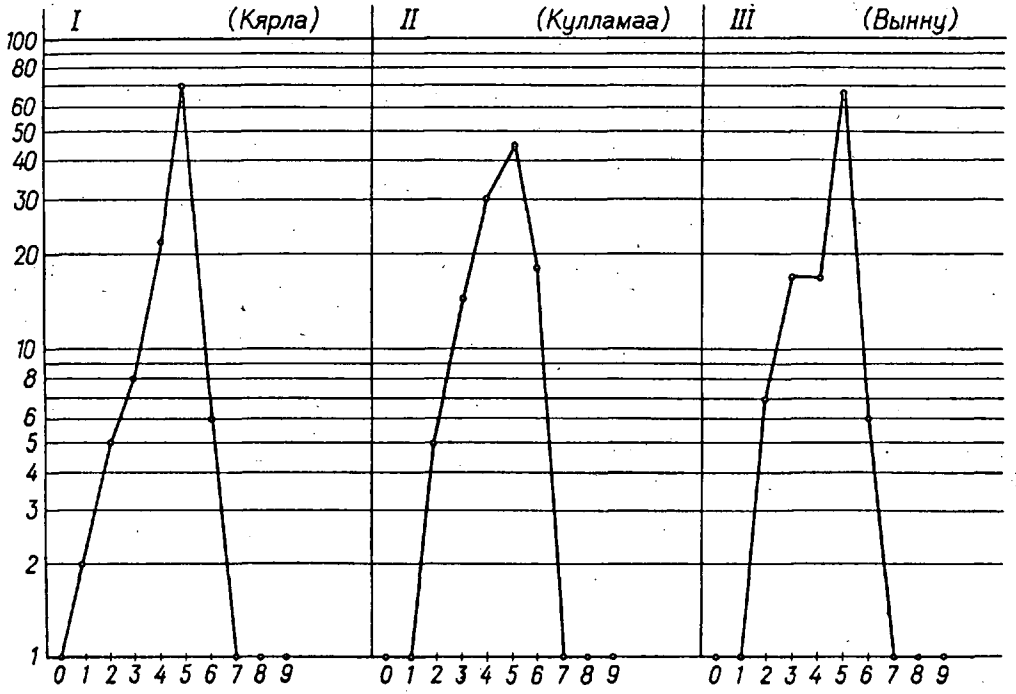


Рис. 4. Самые типичные частотные полигоны распределений связей.

При сравнении 115 полигонов оказалось возможным выделить некоторые районы по типичным значениям отдельных интервалов, а также по соотношениям значений в некоторых интервалах. Во-первых, бросается в глаза, что разница значений в интервалах 5 (мода) и 4 больше на периферии (44 в северо-восточном прибрежном районе и на островах, 47 в южной Эстонии) и меньше в центральной части территории (15). Но у южных приходов обнаруживается меньше связей в интервалах 1 и 2 вместе взятых (в среднем 0,55) и больше в интервале 3 (15), что оближает их с центральными приходами (на юго-востоке и на островах соответствующие показатели - с незначительной дисперсией - равны 2,5 и 7,5). Это отражено на карте рис. 5, где по форме распределений связей приходов выделены три основных региона. Треугольником в каж-

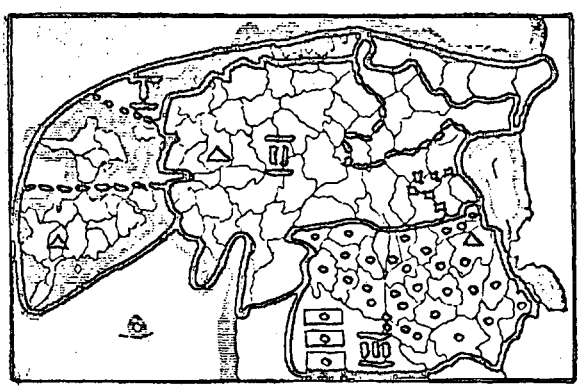


Рис. 5. Районы выделенные по особенностям полигонов распределении связей говоров.

дом регионе отмечается тот приход, рапределение связей которого является самым близким к графику средних значений интервалов по всему региону (рис. 4). В регионе I, в свою очередь, выделяется остров Сааремаа, у приходов которого вообще нет связей в интервале 0. Единственная разница между островом Хийумаа и прибрежным районом состоит в том, что у первого среднее значение в интервале 2 немного меньше, а в интервале 3 немного больше, чем у второго. В центральном регионе (II) выделяется группа северо-восточных приходов, у которых разница интервалов 5 и 4 сближает их с прибрежными приходами, но по остальным сегментам полигонов они принадлежат к цент-

ральным. У трех самых восточных приходов мода совпадает не с 5-м, а с 4-м интервалом (т.е. разница интервалов 5 и 4 отрицательна). В регионе III четких границ провести нельзя. Точкой отмечены приходы, где подобно центральному региону значение интервала 6 больше, чем в остальных (сходных с регионом I). Кружком отмечены те приходы, у которых разница интервалов 5 и 4 выше среднего значения этого региона. В абсолютном большинстве случаев эти приходы географически периферийны. Следовательно, форма кривых распределения является довольно четким параметром определяющим положение отдельных региональных единиц в общетерриториальной коммуникативной ситуации, хотя при его помощи нельзя устанавливать лексическую близость.

ЛИТЕРАТУРА

Лакин Г.Ф. Биометрия. М., 1980.

Пшеничнова Н.Н. Применение таксономического анализа к классификации говоров. - В кн.: Диалектологические исследования по русскому языку. М., 1977, с. 3-14.

AIS (Sprach- and Sachatlas Italiens und der Südschweiz/ Ed. by K. Jaberg, J. Jud, vol. I-VIII, Zofingen, 1928-1940).

Goebl H. Dialektgeographie + Numerische Taxonomie - Dialektometrie. - Ladinia IV, S. 31-95.

Goebl H. Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Wien, 1982.

Goebl H. Éléments d'analyse dialectomêtrique (avec application à l'AIS). - Revue de linguistique Romane. Strasbourg, 1981, t. 45, NN 179-180, p. 349-420.

Hupel H.W. Ehstnische Sprachlehre für beide Hauptdialekte den revalschen und den dörptschen; nebst einem vollständigen Wörterbuch. Riga und Leipzig, 1780.

Hupel H.W. Ehstnische Sprachlehre für die beyden Hauptdialekte, den revalschen und dörptschen, nebst einem vollständigen Ehstnischen Wörterbuch. Mitau, 1818.

Jalasto H. Elementaarstatistika käsairaamat. Tallinn, 1978.

Kask A. Saareste Eesti murrete uurijana. - Kodumurre 7, 1965, lk. 93-101.

Krikmann A. Some aspects of proverb distribution. - Symposium: Mathematical Processing of Cartographic Data (Tallinn, December 18-19, 1979). Summaries. Tallinn, 1979, p. 28-44.

Krikmann A. Towards the Typology of Estonian Folklore Regions. Tallinn, 1980, (Preprint KKI-16).
Murumets S. Eesti keeleala murdelisest liigendusest "Väikese murdesõnastiku põhjal I. - Keel ja Kirjandus, 1982, nr. 1, lk. 11-17.
Murumets S. Eesti keeleala murdelisest liigendusest "Väikese murdesõnastiku" põhjal II. - Keel ja Kirjandus, 1983, nr. 11, lk. 615-623.
Murumets S. On measuring interregional linguistic communication. - Symposium: Processing of Dialectological Data (Tallinn, November 23-25, 1981), Summaries. Tallinn, 1981, p. 43-77.
Murumets S. Regional vocabulary and lexical regions in Estonia. - Dialectology /Ed. by H. Goebl. Quantitative linguistics, Vol. 21, Bochum, 1984, p. 224-253.
Must M. Eesti murdearhiiv teise poolsajandi künnisel. - Emakeele Seltsi Aastaraamat XVII. Tallinn, 1971, lk. 9-31.
Saareste A. Eestikeeleala murdelisest liigendusest. - Eesti Keel, 1932a, nr. 1/2, lk. 17-40.
Saareste A. Kumb eesti murdeist soomele on lähem:põhjaeesti või lõunaeesti? - Eesti Kirjandus, 1931, nr. 1 lk. 29-44.
Saareste A. Kümme aastat eesti murrete süstemaatset kogumist. - Eesti Keel, 1932b, nr. 6, lk. 164-174.
Saareste A. Põhja-Eesti ja Lõuna-Eesti murde vahepiir. - Virittäjä, 1952, nr. 5 lk. 292-307.
Saareste A. Wanradt-Koelli katekismuse keelest. - Eesti Keel, 1930, nr. 4/5, lk. 73-96.
VMS (Väike murdesõnastik. Tallinn, 1982).
Wiedemann F.J. Ehstnische Dialekte und ehstnische Schriftsprache. - Verhandlungen der gelehrten Estnischen Gesellschaft zu Dorpat, VII. Dorpat, 1873, S. 57-80.
Wiedemann F.J. Grammatik der ehstnischen Sprache zunächst wie sie in Mittelehstland gesprochen wird, mit Berücksichtigung der anderen Dialekte. St. Pêtersbourg, 1875.
Wiedemann F.J. Versuch ueber den werroehstnischen Dialekt. - Mêmoires de l'Académie des Sciences de St.-Pêtersbourg. St. Petersburg, 1864. Sêr. 7, t. 7, No. 8.

DIALECT REGIONS FROM AN AUTOMATIC ATLAS OF DIALECT WORDS

Sirje Murumets

Summary

In the 16th-19th centuries valuable intuitive observations on the dialectal division of Estonia were made by dictionary-makers. The 20-ties of our century saw the beginning of systematic fieldwork. The conclusions made 10 years later were mainly based on morphophonetic features.

Since the 80-ties a computer version of the regional index of the Concise Estonian Dialect Dictionary has been available for analysis as an automatic word atlas. This has served as source material for the present study. According to a coefficient of lexical distance computed from the average stereotypedness of the intersection of the word lists of every two parishes, the Estonian territory was first divided into 19 groups of patois which in their turn yielded 7 dialects and one intermediate zone.

An analysis of the asymmetry of parish distributions of the coefficients yields a different division that can be interpreted in terms of the communicative status of the patois.

ОПРЕДЕЛЕНИЕ НАДЕЖНОСТИ ДАННЫХ ЧАСТОТНОГО СЛОВАРЯ

И.В. Перебейнос

Сейчас, когда составлено и опубликовано сотни частотных словарей, накоплен опыт их использования и разработаны приемы их анализа, целесообразно рассмотреть некоторые важные вопросы теории и практики статистической лексикографии. Один из таких вопросов - методы определения надежности данных частотного словаря.

Еще в 1959 г. была предложена методика определения надежности статистических данных в ЧС путем определения относительной ошибки установления частоты и задания порога относительной ошибки, ниже которого частоты считаются недостаточно достоверными. Была предложена формула вычисления относительной ошибки δ :

$$\delta = \frac{Z_p}{\sqrt{N_p}},$$

где Z_p - константа для уровня значимости 0,95, которая принималась автором (Фрумкина, 1959) равной 2 (точнее - 1,96), N - величина выборки в количестве словоупотреблений, p - относительная частота слова, для которого вычисляется относительная ошибка.

Нетрудно увидеть, что выражение под корнем равно абсолютной частоте анализируемой единицы. Таким образом, δ обратно пропорциональна корню квадратному из абсолютной частоты и ни от чего иного не зависит. Если считать приемлемой для лингвистики относительную ошибку, равную 33 %, то нижний порог статистически достоверных частот равен 35 употреблениям слова или словоформы.

Но абсолютная частота 35 имеет разный вес в выборках различной величины, как это видно из таб. 1. Если в малых выборках слово с такой частотой принадлежит к высокочастотным,то в больших оно оказывается низкочастотным. Между тем, предложенная формула этого не учитывает.

Позднее была предложена другая методика определения надежности статистических данных ЧС, базирующаяся, как и предыдущая, на формуле определения разности между наблюденной частотой единицы и ее частотой в полной генеральной совокупности (Ван дер Варден, 1960, с.45):

$$\left| k - p \right| \leq g\sqrt{\frac{pq}{n}},$$

Таблица 1

Абсолютная частота 35 в разных выборках

N	$\bar{x}$	p	N	$\bar{x}$	p
50 тыс.	0,7	0,0007	400 тыс.	0,0875	0,0000875
100 тыс.	0,35	0,00035	500 тыс.	0,07	0,00007
200 тыс.	0,175	0,000175	1000 тыс.	0,035	0,000035

Таблица 2

Статистические характеристики
доверительных границ неизвестной вероятности
при абсолютной частоте 35

N	p_1	Np_1	ε_1 %	p_2	Np_2	ε_2 %
50 тыс.	0,0005	25	40	0,00097	49	28
100 тыс.	0,00025	25	40	0,00049	49	29
200 тыс.	0,000126	25	39	0,000243	49	28
400 тыс.	0,000063	25	39	0,000122	49	28
500 тыс.	0,00005	25	40	0,000097	49	28
1000 тыс.	0,000025	25	40	0,000049	49	29

Таблица 3

Статистические характеристики $F = 35$
в разных выборках в предположении
равномерности распределения частоты

N	$\bar{x}$	$\sigma_{\bar{x}}$	ε %
50 тыс.	0,7	0,065	18,2
100 тыс.	0,35	0,048	26,6
200 тыс.	0,175	0,027	30,2
400 тыс.	0,0875	0,014	31,4
500 тыс.	0,07	0,0114	31,9
1000 тыс.	0,035	0,0058	32,5

где k - наблюденная частота, p - вероятность, т.е. частота в полной генеральной совокупности, g - константа уровня значимости 0,95, $q = 1 - p$, n - величина выборки. Отсюда вытекают значения доверительных границ, между которыми расположено истинное значение p:

$$p_1 = \frac{kn + 1/2 g^2 - g\sqrt{k(1-k)n + 1/4 g^2}}{n + g^2};$$

$$p_2 = \frac{kn + 1/2 g^2 + g\sqrt{k(1-k)n + 1/4 g^2}}{n + g^2}.$$

Предложено упрощение этих формул: изымается $1-k$, так как относительная частота даже самого частого слова намного меньше 1; kn заменено абсолютной частотой, которой оно и равно; из знаменателя изымается g^2, так как оно ничтожно мало по сравнению с величиной выборки (Алексеев, 1975, с. 47). В результате формулы приобретают следующий вид:

$$p_1 = \frac{F + 1/2 g^2 - g\sqrt{F + 1/4 g^2}}{n};$$

$$p_2 = \frac{F + 1/2 g^2 + g\sqrt{F + 1/4 g^2}}{n}.$$

Если по этим формулам определять доверительные границы p для некоторого слова в разных по величине выборки ЧС, то они зависят от абсолютной частоты этого слова и от величины выборки, на которой эта частота определена.

Определим доверительные границы вероятности для частоты 35 в разных выборках. Так как F остается постоянным, то величина в числителе одинакова для всех выборок, т.е. верхнее и нижнее значение вероятности зависит только от величины выборки. При этом оказывается (см. табл. 2), что относительная разность между этими вероятностями для всех выборок составляет 48...49%,т.е. от величины выборки не зависит. Не зависят от нее и граничные значения абсолютных частот, соответствующих вычисленным вероятностям, и границы относительной ошибки определения доверительных интервалов, вычисляемые по формулам, предложенным П.М. Алексеевым (Алексеев, 1975, с. 47):

$$\varepsilon_1 = \frac{k - p_1}{p_1};$$

$$\varepsilon_2 = \frac{p_2 - k}{p_2}.$$

Получается, что, несмотря на изменяющиеся вероятности, статистические характеристики их не зависят от величины выборки. Между тем, в статистике существует формула вычисления относительной ошибки средней частоты (Урбах, 1964), указывающая на обратную зависимость ε от величины выборки, измеряемой числом подвыборок:

$$\varepsilon = \frac{1{,}96\,\sigma}{\bar{x}\sqrt{n}} = \frac{1{,}96 \cdot \sigma_{\bar{x}}}{\bar{x}}.$$

Для вычисления этой относительной ошибки недостаточно иметь лишь одну абсолютную частоту, как в большинстве ЧС, а надо организовать выборку по всем правилам статистики, разделив ее на минимальные выборки (подвыборки), вычислив среднюю частоту и среднее квадратичное отклонение или меру колебания средней частоты.

Вычислим ε по этой формуле в предположении, что каждая выборка разделена на минимальные выборки по 1000 словоупотреблений, а абсолютная частота распределена равномерно, т.е. в минимальной выборке представлена одним употреблением. Так как нас интересует $F = 35$, то и количество минимальных выборок, в которых встретилась предполагаемая единица, должно быть таким же (см. табл.3).

Как видим, при таком подходе к вычислению относительной ошибки величина ее меняется, возрастая вместе с ростом выборки: рост выборки приводит к уменьшению $\bar{x}$, с которым ε связана обратной зависимостью. Но есть и другая причина увеличения ε: предположение о равномерности распределения не всегда правомерно. Анализ частотного словаря современной украинской художественной прозы (Частотный словарь, 1981) показывает, что встречаться только по одному разу в минимальной выборке могут слова и словоформы, частота которых по крайней мере на порядок меньше объема выборки, измеряемого количеством минимальных выборок: лишь для $F = 48$ (при $N = 500$) в одном случае количество минимальных выборок $K = F$. То же наблюдаем для $F = 43, 39, 37, 33, 32$. Начиная с $F = 26$, количество совпадений F с K увеличивается, но вплоть до $F = 10$ они не составляют и половины числа единиц с данной частотой. Этот факт интересен сам по себе: исследование распределения слова по минимальным выборкам может раскрыть такие закономерности функционирования лексики в тексте, каких нельзя вскрыть иными способами.

Таким образом, мало вероятно, чтобы при выборке в 50 тыс. словоупотреблений, т.е. всего 50 подвыборок, $F = 35 = K$. Скорее тут можно ожидать, что слово с такой

частотой будет встречаться по несколько раз в минимальной выборке, как это наблюдается для высокочастотных слов. В выборке же 200 тыс. и выше случаи $F = K$ более вероятны, что отражается и в относительном постоянстве $\mathcal{E}$ в этих выборках. Но нигде $\mathcal{E}$ не достигает 33 %, как при вычислении δ. Во всех случаях $\mathcal{E} < \delta$.

Но в случае неравномерного распределения частоты по подвыборкам, что и наблюдается в действительности, $\mathcal{E}$ должна существенно расти с увеличением степени этой неравномерности.

Интересующая нас $F = 35$ в частотном словаре современной украинской художественной прозы распределяется по подвыборкам весьма неодинаково для разных единиц: $34 \geqslant K \geqslant 10$, соответственно изменяется и $\mathcal{E}$, как это видно из таблицы 4.

Таблица 4

Соотношение K и $\mathcal{E}$ % для $F = 35$ при $N = 500$ тыс.

K	слов	словоформ	всего	$\mathcal{E}$ %
34	3	6	9	32,8
33	7	5	12	33,6...34,7
32	8	5	13	34,7...35,6
31	6	4	10	35,6...36,4
30	3	5	8	36,4...37,2
29	10	-	10	37,2...48,2
28	4	3	7	38,9...41,4
27	3	-	3	39,8
26	2	-	2	44,2...44,8
25	1	2	3	42,8...48,2
24	1	1	2	46,2...49,0
23	-	1	1	47,6
22	1	-	1	46,2
21	1	-	1	52,4
20	1	-	1	56,6
19	1	-	1	50,1
18	1	-	1	60,2
15	1	-	1	62,4
13	-	1	1	65,0
12	-	1	1	73,4
10	-	1	1	74,2
Всего	54	35	89	

Обращает на себя внимание, что для ряда K устанавливается не одно значение $\mathcal{E}$, а интервал ее значений. Получается, что количество подвыборок - не окончательный критерий для суждения о степени равномерности распределения единицы по подвыборкам. Действительно, при одном и том же K распределение может быть разным по степени равномерности. К примеру, для $K = 28$ возможны три варианта распределения:

1)		2)		3)	
x_i 0	n_i	x_i	n_i	x_i	n_i
0	572	0	572	0	572
1	22	1	23	1	24
2	5	2	4	2	2
3	1	3	-	3	1
4	-	4	1	4	1
$\sigma_{\bar{x}} = 0{,}0139$		$\sigma_{\bar{x}} = 0{,}0144$		$\sigma_{\bar{x}} = 0{,}0148$	
$\mathcal{E} = 38{,}9\ \%$		$\mathcal{E} = 40{,}3\ \%$		$\mathcal{E} = 41{,}4\ \%$	

Еще больше различия между разными распределениями при $K = 25$:

0	575	0	575	0	575
1	18	1	19	1	20
2	4	2	4	2	3
3	3	3	1	3	1
		5	1	6	1

Для этих распределений $\sigma_{\bar{x}}$ соответственно равна 0,0153, 0,0163 и 0,0172, а $\mathcal{E}$ % - 42,8, 45,6 и 48,2.

Таким образом, если принять за достоверные те частоты, для которых $\mathcal{E} \leq 33\ \%$, то следует признать, что в Частотном словаре современной украинской художественной прозы $F = 35$ может считаться достоверной лишь для 9 лексем из 89. При $F = 36$ достоверными могут считаться данные лишь для четырех слов, для которых $K \geq 35$.

Можно вычислить, какое максимальное значение $\sigma_{\bar{x}}$ обеспечивает $\mathcal{E} = 33\ \%$ для разных абсолютных частот, преобразовав формулу вычисления $\mathcal{E}$:

$$\sigma_{\bar{x}} = \frac{\mathcal{E}\,\bar{x}}{1{,}96} = \frac{0{,}33\,\bar{x}}{1{,}96}.$$

Вычисление этих граничных значений $\sigma_{\bar{x}}$ для слов и словоформ с $35 \leq F \leq 80$ и определение количества лексических единиц, имеющих достоверные данные, показывает, что вплоть до $F = 40$ они составляют не более 13,6 % от количества единиц с данной F. Далее их вес постепенно повышается:

$40 \leq F \leq 46$	в пределах	30...50 %
$47 \leq F \leq 56$	-"-	51...70 %
$57 \leq F \leq 67$	-"-	65...90 %
$68 \leq F \leq 80$	-"-	75...100 %

С ростом частоты роль неравномерности распределения ее по минимальным выборкам снижается. Так, в последней группе даже $k = 46$ может обеспечить $\varepsilon \leq 33$ %. Таким образом, если не учитывать неравномерности распределения частот, а определять надежность данных ЧС лишь по абсолютной частоте, то для $N = 500\,000$ надежными можно считать данные для $F > 68$. Вполне возможно, что для иных по величине выборок этот порог изменится, как изменятся и пороговые значения $\sigma_{\bar{x}}$.

Как представляется, определение надежности данных ЧС должно основываться на особенностях функционирования каждой лексической единицы в ЧС, а не на априорных рассуждениях, какими бы правдоподобными они ни казались.

Для того, чтобы дать в руки исследователям и пользователям сведения для вычисления надежности данных ЧС, необходимо пересмотреть схему их построения. Почему-то сложилась традиция, согласно с которой основная статистическая характеристика единицы в ЧС - абсолютная частота ее во всей выборке, на которой составлен словарь. Ни одно статистическое исследование не может базироваться на единственном показателе частоты, это хорошо известно. Вероятно, традиция эта сложилась тогда, когда первые составители ЧС не были достаточно знакомы со статистикой. Теперь обстановка изменилась. Применение методов статистики в лингвистических исследованиях получило широкий размах. И настало время строить ЧС в соответствии с требованиями статистики. Тогда не придется придумывать методы определения равномерности распределения лексики и частоты каждого слова, так как равномерность распределения определяется широко применяющимися в статистике формулами, а также законами распределения частот.

На наш взгляд, в ЧС, кроме абсолютных данных (абсолютная частота, количество текстов, количество минимальных выборок) необходимы также относительные данные: средняя частота и мера ее колебания. Они нужны как для анализа материала данного ЧС, так и для сопоставления его данных с данными других ЧС, они обеспечат и сведения для вычисления степени надежности приведенных в ЧС статистических характеристик слов и словоформ.

ЛИТЕРАТУРА

Алексеев П.М. Статистическая лексикография. - Л.: ЛГПИ им. А.И. Герцена, 1975. - 120 с.

Ван дер Варден Б.Л. Математическая статистика. - М.: Изд-во иностранной литературы, 1960. - 434 с.

Урбах В.Ю. Биометрические методы. - М.: Наука, 1964. - 416 с.

Фрумкина Р.М. Некоторые вопросы методики составления частотных словарей. - В кн.: Машинный перевод и прикладная лингвистика. - М.: МГПИИЯ им. М. Тореза, 1959, № 2 (9), с. 23-37.

Частотний словник сучасної української художньої прози в двох томах. - К.: Наукова думка, 1981. - Т. 1 - 864 с.; т. П - 856 с.

MEASURING THE RELIABILITY OF FREQUENCY DICTIONARY DATA

Valentina Perebeynoss

Summary

The reliability of word frequencies in a frequency dictionary is often measured by means of relative error δ, taking into consideration only the absolute frequency (F) of a word, but not the size of the text (N) on which it was obtained. The absolute frequency of 35 is considered reliable because for it δ = 33 %. Another method is based on calculating the reliable limits of an unknown probability. Using it we obtain different relative frequencies (P) for different N, but the reliable interval of absolute frequencies remains the same (25...49).

In statistics the reliability of a mean frequency is measured by a relative error ε calculated as the ratio of the standard error of the mean to the mean times 1.96 (the coefficient of 95 % level of reliability). ε takes into consideration not only $\bar{X}$ and N but also uniformity/non-uniformity of a word distribution in samples. The analysis of the data in the Frequency Dictionary of the present Ukrainian works of fiction (based on 500 samples of 1000 running words each) shows that each absolute frequency is characterized by a number of different values of the standard error of the mean, this influences the reliability of the frequency. For example, only for 10 % of the words having F = 35 ($\bar{X}$ = 0,07) the data may be considered sufficiently reliable ($\varepsilon \leq$ 33 %). This means that the basic statistical data in word counts must be not only absolute frequency but also $\bar{X}$ and $\sigma_{\bar{x}}$.

КВАНТИФИКАЦИЯ СВЯЗЕЙ ВНУТРИ ПРЕДЛОЖЕНИЯ КАК ИНСТРУМЕНТ ТИПОЛОГИИ

М.С. Полинская

1. Введение

В нашей работе (Полинская, 1983) был предложен метод количественной оценки связей между элементами языковой структуры, который в общем виде сводится к следующему.

Совместная встречаемость языковых единиц (элементов), парадигматически принадлежащих к разным классам, в заданном линейном интервале является определенным образом детерминированной*. Появляясь в некотором интервале I, элемент x ($x \in P$), детерминирует появление в этом же интервале элемента y ($y \in T$). Ставится задача оценить количественно, насколько появление одного элемента (=языковой единицы некоторого уровня) в пределах данного интервала детерминирует совместное с ним появление другого определенного элемента (=языковой единицы того же уровня). Подобные отношения детерминации предлагается измерять при помощи коэффициента степени жесткости (СЖ).

Пусть:

X - любой класс языковых единиц, принятый в анализе;

$i \in X_2$; $j \in X_I$;

α_{ji} - число случаев совместной встречаемости элемента j с данным элементом i (в пределах заданного интервала I, на данной выборке);

m - число элементов в множестве X_I;

n - число элементов в множестве X_2, -

тогда мера степени жесткости (сж), с которой появление в данном интервале I элемента j требует появления в I элементов множества X_2, определяется по формуле:

$$\text{сж} = \left(\frac{\sqrt{\sum_{i=1}^{m} \alpha_{ji}^2}}{\sum_{i=1}^{m} \alpha_{ji}} - \frac{1}{\sqrt{m}} \right) \cdot \frac{\sqrt{m}}{\sqrt{m} - 1} \qquad (I)$$

* Метод не различает положительной и отрицательной детерминации, т.е. случаи, когда элемент требует и когда он запрещает появление другого элемента, рассматриваются вместе. Отсутствие отношений детерминации подпадает под предельный случай - нулевую детерминацию совместной встречаемости элементов.

Для определения того, как жестко все элементы множества X_I задают совместное с ними появление в I элементов множества X_2, для каждого элемента вычисляется средне-взвешенный коэффициент; затем производится усреднение этих коэффициентов. Итак, конечная мера СЖ находится по формуле:

$$СЖ_{X_I} = \frac{\sum_{j=1}^{n} СЖ_j \cdot g_j}{n} \qquad (2),$$

где

$$g_j = \frac{\sum_{i=1}^{m} \alpha_{ji}}{\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ji}}$$

В анализе, результаты которого приводятся здесь, ставилась цель применить описанный метод для синтактико-типологически ориентированного сравнения языков.

За интервал было взято простое предложение (индексировались только связные тексты, о преимуществах таких текстов при лингвистическом анализе см. (Heath, 1979)).

Выделялись следующие классы (множества) - класс предикатов (Р) и классы актантов (гипер) ролей: Агенс (Аг), Пациенс (Пац), Экзистант (Экз), Адресат (Адр), Инструмент (Инстр), Локатив (Лок)*. Мы полагаем, что семантические классы удобнее, нежели формальные, если речь идет о сравнении разносистемных языков: кодировочные средства могут иметь совершенно разный вес в разных языках, и значит, при сопоставлении языков на формальной основе необходимо вводить предварительные процедуры ранжирования средств выражения. Опора на семантические классы создает бо́льшие возможности, хотя, как мы постараемся показать ниже, и эти классы трудно считать универсальными.

Итак, на интервале простого предложения определялось, насколько жестко появление актантов перечисленных ролей и предиката детерминирует появление их "партнеров". Результаты исследования обсуждаются ниже в виде отдельных задач.

---

* Подробно об определении и обосновании перечисленных здесь классов см. (Полинская, 1983).

П. Результаты и их обсуждение

Задача 1. Два стиля одного языка.

Сравнивались две выборки русского языка, представляющие драматическую прозу (А) и научную - научно-популярную прозу (Б). Результаты и списки выборок приведены в табл. 1.

Все таблицы строятся следующим образом. В левом столбце записываются попарно названия классов, принятых в анализе; например, Аг/Пац. В правых столбцах стоят полученные показатели, левее знака дроби - показатель СЖ, с которой левый член пары задает правый, правее - показатель СЖ, с которой правый член пары задает "партнера". Таким образом, показатели с самого начала сводятся в парные индексы и рассматриваются не как абсолютные, а как соотносительные. (В дальнейшем интересные факты можно получать, определяя для каждой пары коэффициент внутреннего соотношения, то есть $СЖ_1 : СЖ_2$ и т.п. Мы опускаем здесь этот анализ, так как он более существен при внутриязыковой формальной параметризации.)

Положим, пара Аг/Пац имеет индекс 0.21/0.06. Это означает, что класс Аг требует//запрещает класс* Пац в рамках данного интервала со степенью жесткости 0.21, а класс Пац детерминирует появление//отсутствие элементов класса Аг с СЖ = 0.06, то есть втрое меньшей. Как интерпретировать эти и другие показатели? Мы видим, что Пац относительно безразличен к Аг, как безразличен он и к ряду других классов (это видно из его низких индексов СЖ). Гораздо более "действенны" актанты ролей Аг, Адр. Обращает на себя внимание, что в парах роль/роль индексы выше, чем в парах предикать/роль, особенно для более субъектных ролей.

Это представляется закономерным. Дело в том, что исходя из семантических отношений между составляющими предложениями, мы, по нулевой гипотезе, должны получить показатели чистых семантических связей. Однако получаемые индексы, позволяющие соотносить содержательную структуру с поверхностным представлением, являются показателями комплексной связи: на них влияют все структурные компоненты, в том или ином виде доходящие до поверхностного представления.

Говоря о связях составляющих внутри предложения, мы имеем в виду, таким образом, множественные зависимости; в отношениях двух конкретных составляющих может играть преобладающую роль та или иная связь (семантическая, синтаксическая, референционная). Появление низких индексов в ряде пар может быть обусловлено тем, что семантическая связь "перекрыта" какой-либо другой связью. В паре Р/Пац семантическая связь перекрывается

---

* Ясно, что такое представление условно: друг друга в реальном предложении задают не классы, а их элементы.

Таблица 1. Два стиля русского языка

пара	выборка /А/	/Б/	пара	выборка /А/	/Б/
Р/Аг	.07/.17	.08/.17	Пац/Экз	.15/.25	.17/.26
Р/Экз	.10/.11	.09/.11	Пац/Адр	.15/.13	.165/.13
Р/Пац	.07/.07	.09/.07	Пац/Лок	.15/.09	.13/.07
Р/Адр	.10/.09	.11/.08	Пац/Инстр	.22/.11	.26/.17
Р/Лок	.12/.075	.10/.07	Адр/Экз	.15/.25	.12/.21
Р/Инстр	.08/.06	.11/.10	Адр/Лок	.17/.28	.17/.30
Аг/экз	.22/.11	.23/.12	Адр/Инстр	.16/.20	.21/.29
Аг/Пац	.18/.06	.21/.06	Лок/экз	.26/.20	.27/.185
Аг/Адр	.24/.06	.25/.07	Лок/Инстр	.21/.19	.18/.18
Аг/Лок	.20/.14	.22/.125	Инстр/Экз	.18/.16	.21/.20
Аг/Инстр	.23/.10	.23/.11			

Выборка /А/ - драматическая проза:

Л. Толстой, Власть тьмы, д. 5, сц. 1
А.П. Чехов, Иванов, первые 10 стр.
А. Вампилов, Прощание в июне, 15 стр.
Ф. Дюренматт, Ангел приходит в Вавилон /пер.Н.Оттена/, первые 12 стр.

Выборка /Б/ - научная и научно-популярная проза:

Сердечно-сосудистая патология, М., 1977, ст. 78-90
Курс квантовой механики, М., 1980, стр. 54-62
Высокомолекулярные соединения, А 23, 2, 15 стр.
А. Суперанская, А. Суслова, Современные русские фамилии, М., 1981, 15 стр.

прежде всего синтаксической связью, которая настолько сильна, что создает блок Р - Пац со слабым внутренним семантическим отношением. Та же синтаксичность связей заметна и из других индексов Р, причем с нарастанием периферийности роли "партнера" синтаксичность соответственно падает.

Там, где пара индексов резко несимметрична (особенно это ощутимо в парах с Аг, чьи индексы намного выше индексов "партнера"), можно предполагать отношения "хозяина - слуги". Подобное отношение связывает и Пац/Инстр: там, где в предложении есть Пац, он может детерминировать появление Инстр, но не наоборот.

Отношения "хозяина - слуги" обнаруживаются и для Аг/Пац: Аг гораздо жестче задает свой Пац. Возможно, это отражает языковую тенденцию строже иерархизовать данные роли, в принципе достаточно конкурирующие. На это косвенно указывает и некоторое повышение индекса Аг в данной паре в научном тексте (образец (Б)). Известные структурные икомпозиционные особенности научного текста вынуждают Аг укрепить свой статус. Кроме того, здесь действует и фактор длины предложения. Бо́льшая, чем обычно, длина предложения заставляет Аг жестче задавать свой удаленный Пац; такое повышение СЖ как бы скрепляет кон-

струкцию. Возможно, та же причина лежит в основе роста СЖ Адресата в (Б).

В целом, как видно из табл. 1, две выборки дают очень сходные результаты. Различия - в основном в области классов Адр и Инстр, чьи индексы в (Б) несколько выше. Мы полагаем, что это связано с содержательными и коммуникативными особенностями научного текста.

В остальном данные близки, расхождения - порядка сотых долей, причем они коррелируют с периферийностью ролей. Можно наблюдать одновременное нарастание или снижение показателей в рамках одной пары (по сравнению с другим образцом), что согласуется с идеей относительного коэффициента СЖ. Кроме того, в этом можно найти подтверждение двунаправленного, хотя и несимметричного, отношения между двумя классами: при изменении одной из мер происходит изменение "контрмеры". Однако пока неясно, всегда ли меры меняются согласованно, как в данном случае.

Материал задачи показывает, что метод может быть использован для описания языка, так как стилистические различия не являются возмущающим фактором.

Задача 2. Группа индоевропейских языков.

Определялись индексы для общих выборок английского, итальянского, латинского, немецкого, русского и французского языков. Материал был обработан по методике, описанной выше (данные в табл. 3), и по "узкой" методике (данные в табл. 2), согласно которой измерялись только СЖ предиката и трех классов семантических ролей - Аг, Пац и Адр.

Обнаруживается существенное сходство индексов по всем рассмотренным языкам. Сходство растет при "сужении" классификации, из чего следует, что укрупнение таксонов не может быть бесконечным: на каком-то уровне мы начинаем не приобретать, а терять информацию.

Из данных табл. 3 видно, что при существенном сходстве даже небольшая группа из шести языков может быть организована в соответствии с полученными мерами, как некоторый континуум. Континуальное представление при типологической классификации кажется более удовлетворительным, нежели дискретное, точечное. Такое представление в принципе хорошо согласуется с качественным наблюдением о том, что "чистых" типов не существует: всякий язык - это сложный комплекс разнообразных типологических характеристик, и даже с точки зрения частной типологии (в нашем случае - синтаксической) он будет гетерогенным. Мы полагаем, что использование набора индексов позволяет учесть пучок характеристик языка; если каждую отдельную характеристику рассматривать как координатную ось, то логично ожидать, что мы получим сложную многомерную пространственную модель типологического разбиения языков. Итак, "типологическое пространство" - не просто красивая метафора, а вполне адекватный образ реального положения вещей.

Данные табл. 3 во многом сходны с данными табл. 1.

Таблица 2. Индоевропейские языки, "узкая" методика

пара	язык					
	русск.	нем.	англ.	франц.	лат.	итал.
Р/Аг	.07/.16	.07/.21	.06/.18	.06/.17	.07/.20	.06/.18
Р/Пац	.07/.06	.10/.08	.09/.08	.06/.06	.09/.06	.07/.065
Р/Адр	.18/.12	.18/.14	.19/.11	.16/.12	.18/.13	.17/.11
Аг/Пац	.18/.15	.15/.13	.14/.15	.15/.12	.17/.15	.15/.14
Аг/Адр	.25/.21	.24/.15	.22/.19	.22/.17	.21/.18	.23/.195
Пац/Адр	.15/.13	.12/.08	.12/.07	.15/.09	.15/.11	.14/.10

Русск. - выборки /А/, /Б/, /табл. 1/; Народные русские сказки в зап. А.Н. Афанасьева, 10 стр.; М.Ю. Лермонтов, Соч. в 4 тт., т. 1, стр. 40 - 46; "Правда", 19.11.1980, 1 стр.;"Филиппины",сб., М., 1971, стр. 19-26; Ю. Нагибин, Итальянские тетради. 10 стр.; А.П. Чехов, Палата № 6; Р. дель Валье-Инклан, Страх /пер. А. Косс/.

Нем. - W. Hauff, Der Affe als Mensch; H. Fallada,Der Trinker, 15 S.; Burgmann A., Syntaktische Probleme in Polynesischen... (ZES, Bd. 32, 20 S.); Neues Deutschland, 8.3.1981, 2 S.

Англ. - A. Trollope, Barchester Towers, v. 1,20 pp.; E. Hemingway, To Have and to Have Not,25 pp.; Medical text, 10 pp.; Morning Star,23.10.1980, 2 pp.

Франц.- G. Bernanos, Nouvelle histoire de Mouchette, 25 pp.; Le livre magique, 15 pp,; Paris,ville enchantée, P., 1966, 10 pp.; L'Humanité,27.8. 1980, 2 pp.

Лат. - Titus Livius, Ab urbe condita, Liber XXI, 37 pp.; Nicolaus Cusanus, De Deo abscondito;Vulgata, MATt., 18-28.

Итал. - Italo Calvino, Gli avanguardisti a Mentone; L'Unità, 10.9.1981, 3 pp.

Таблица 3. Индоевропейские языки.

пара	языки					
	англ.	франц.	нем.	русск.	лат.	итал.
Р/Аг	.07/.20	.08/.I8	.08/.I9	.06/.I7	.07/.I8	.065/.I7
Р/Экз	.08/.II	.09/.08	.09/.I0	.09/.I0	.08/.09	.09/.095
Р/Пац	.07/.05	.09/.06	.06/.06	.07/.06	.08/.07	.07/.06
Р/Адр	.09/.I2	.09/.08	.I0/.I0	.I0/.06	.I0/.09	.09/.07
Р/Лок	.I0/.05	.I0/.06	.II/.06	.II/.07	.I2/.06	.II/.08
Р/Инс	.II/.07	.I2/.08	.I0/.075	.09/.07	.I0/.08	.095/.08
Аг/Экз	.20/.II	.2I/.I0	.23/.I4	.2I/.I2	.I8/.II	.22/.I3
Аг/Пац	.I7/.06	.I9/.05	.I6/.05	.I9/.09	.I7/.07	.I8/.06
Аг/Адр	.22/.08	.2I/.09	.24/.I0	.25/.09	.22/.II	.2I/.08
Аг/Лок	.2I/.I2	.I9/.I0	.20/.I3	.2I/.I5	.23/.I4	.22/.I45
Аг/Инс	.23/.I35	.20/.II	.22/.I2	.24/.I0	.22/.II	.2I/.I0
Пац/Эк	.I8/.28	.I6/.I9	.I9/.22	.I7/.29	.I6/.2I	.I8/.25
Пац/Ад	.I6/.I2	.I8/.I6	.I8/.I5	.I9/.I4	.20/.I6	.I8/.I55
Пац/Ло	.I4/.I0	.I6/.09	.I5/.09	.I4/.09	.I7/.II	.I5/.I0
Пац/Ин	.I9/.II	.25/.I3	.22/.I4	.24/.I5	.22/.I2	.24/.I3
Адр/Эк	.I7/.26	.I6/.I9	.I6/.22	.I3/.24	.I5/.2I	.I7/.20
Адр/Ло	.I6/.22	.I7/.29	.I5/.23	.I8/.32	.I6/.30	.I7/.27
Адр/Ин	.I8/.2I	.I5/.20	.I7/.22	.I7/.23	.I5/.24	.I6/.23
Лок/Эк	.28/.24	.26/.20	.22/.20	.26/.2I	.27/.22	.25/.23
Лок/Ин	.20/.2I	.22/.I9	.2I/.22	.20/.20	.25/.22	.2I/.2I
Ин/Экз	.I9/.I6	.20/.I5	.I8/.I6	.20/.I65	.2I/.I4	.I9/.I4

Списки выборок приведены в таблице 2.

Все те особенности, о которых шла речь в задаче 1,прослеживаются и здесь. Отметим относительно менее "влиятельное" поведение Пац, семантическую нейтральность Р и особенно высокие СЖ классов Аг и Адр. Более подробное обсуждение пар классов см. ниже (задача 8).

Задача 3. Два периода в истории одного языка - тонганский язык.

Тонганский - австронезийский язык Полинезии.Традиционно он считается эргативным, однако современный тонганский, по-видимому, во многом утратил свою эргативность и существенно аккузативизировался (Hohepa,1969), (Tchekhoff, 1979).

Мы сравнивали показатели, полученные на материале тонганских мифов, записанных в конце XIX - нач. XX вв. ($ТОН_1$), и тонганских сказок, пересказанных или сочиненных носительницей языка в сер. 70-х гг. нашего века ($ТОН_2$). Результаты приведены в табл. 4. Данные таблицы действительно позволяют говорить о существенных изменениях $ТОН_2$ по сравнению с $ТОН_1$. Видно, что изменения эти наиболее заметны в области "центральных" классов (предиката и ролей,чаще всего соотносимых с ингерентно субъектными актантами). По сравнению с динамикой латинского языка (данные приведены в (Полинская, 1983)),тонганский демонстрирует несимметричность, или несбалансированность, изменений: в отличие от латыни, меры в паре (справа и слева) меняются несогласованно,т.е.рост меры необязательно ведет к росту или падению "контрмеры".

Таблица 4. $Тонганский_1$ - $Тонганский_2$

пара	$тон_1$	$тон_2$	пара	$тон_1$	$тон_2$
Р/Аг	.13/.13	.07/.16	Пац/Экз	.17/.10	.16/.08
Р/Экз	.16/.19	.16/.09	Пац/Адр	.16/.18	.16/.20
Р/Пац	.06/.07	.09/.06	Пац/Лок	.16/.10	.14/.09
Р/Адр	.12/.15	.12/.06	Пац/Инс	.21/.12	.19/.13
Р/Лок	.10/.12	.11/.11	Адр/Экз	.18/.18	.12/.14
Р/Инстр	.15/.08	.12/.06	Адр/Лок	.18/.10	.14/.15
Аг/Экз	.16/.10	.15/.06	Адр/Инс	.19/.21	.20/.21
Аг/Пац	.12/.08	.10/.07	Лок/Экз	.14/.29	.16/.18
Аг/Адр	.11/.05	.16/.05	Лок/Инс	.18/.19	.19/.17
Аг/Лок	.18/.09	.12/.07	Инс/Экз	.18/.24	.22/.21
Аг/Инстр.	.20/.09	.18/.09			

$Тон_1$ - Gifford E., Tongan Myths and Tales, Honolulu, 1924 (30 pp.); Reiter P., Traditions tongiennes - "Anthropos", 2, 1907 (10 pp.)

$Тон_2$ - Tupou Posesi Fanua, Po Fananga, San Diego, 1975 (40 pp., random choice).

Индексы пары Р/Аг в $ТОН_2$ напоминают соответствующие индексы индоевропейских языков, что позволяет предполагать некоторую аккузативизацию языка. В $ТОН_1$ эти

индексы резко отличаются от соответствующих индексов всех других языков, рассмотренных выше.

Таким образом, статистика подтверждает типологическое изменение языка, предполагаемое на основании качественного анализа. Выясняется, что при достаточно резком изменении языка меняются показатели "верхушки", т.е. центральных классов - предиката и ролей, чаще всего ассоциируемых с грамматическими отношениями подлежащее и прямое дополнение.

Задача 4. Эргативные языки.

В табл. 5 приведены индексы, полученные для арчинского (АРЧ), баскского (БАС), самоанского (САМ) и тонганского (ТОН, расширенная выборка $ТОН_1$ предыдущей задачи).

Круг проблем расширяется, так как языки, сравниваемые здесь, различаются и генетически, и ареально, и, что немаловажно, с точки зрения формального типа (в АРЧ, БАС - развитая морфология: в САМ и ТОН словоизменение практически отсутствует).

Бросается в глаза существенное сходство между языками и их отличие от языков, описанных выше. Особенно существенным является отличие предикатных пар от соответствующих пар в других языках. В этой четверке языков Р гораздо жестче детерминирует свои актанты; особенно это заметно в паре Р/Аг, где индексы повышаются почти вдвое. Таким образом, наблюдается семантизация предикатных связей. Неясно, вызвано ли это слабой залоговой дифференциацией (тем более, что в БАС существует пассив); скорее, причины семантизации связей разнообразны.

Кроме того, обнаруживается существенное сходство АРЧ - БАС, САМ - ТОН попарно. В типологической организации всех четырех языков так же, как и в организации языков задачи 2, актуально представление в виде пространства.

Статистика подтверждает особую семантичность АРЧ* (ср. (Кибрик, 1979)) и достаточную, хотя и меньшую, чем в АРЧ, семантичность связей в БАС. Видно, что при сходстве картины в полинезийских эргативных языках и при совпадении коэффициентов пар, показатели в САМ и ТОН ниже, чем в АРЧ и БАС. О факторах, понижающих СЖ в полинезийских языках, идет речь в следующей задаче.

В заключение отметим, что рост семантичности связей наблюдается при движении вниз, то есть к периферийным ролям (ср. также оппозицию: синтаксические категории, или термы, - семантические роли, или не-термы). Отсюда следует и то, что различия между двумя типами языка - эргативным и аккузативным - наиболее ощутимы в "верхушке" и несколько стираются к периферии.

Интересное сходство с эргативными языками прояв-

* В тагальском (квалифицируемом как семантически аккузативный язык в (Кибрик, 1979, 10)) семантичность отношений статистикой не подтверждается, см. задачу 6.

Таблица 5. Эргативные языки

пара	язык			
	арч.	бас.	тон.	сам.
Р/Аг	.14/.27	.15/.23	.13/.15	.13/.19
Р/Экз	.15/.21	.13/.20	.16/.19	.15/.20
Р/Пац	.08/.09	.08/.08	.06/.07	.10/.08
Р/Адр	.07/.11	.08/.10	.12/.15	.12/.10
Р/Лок	.11/.10	.10/.10	.10/.12	.15/.11
Р/Инстр	.15/.11	.13/.10	.15/.08	.14/.05
Аг/Экз	.20/.25	.21/.22	.16/.11	.16/.09
Аг/Пац	.17/.12	.16/.10	.12/.08	.17/.11
Аг/Адр	.14/.14	.12/.13	.11/.06	.14/.12
Аг/Лок	.20/.11	.21/.09	.18/.09	.15/.06
Аг/Инстр	.30/.13	.25/.10	.20/.08	.19/.095
Пац/Экз	.20/.30	.18/.24	.17/.10	.18/.16
Пац/Адр	.14/.20	.12/.16	.16/.18	.13/.19
Пац/Лок	.21/.13	.19/.105	.16/.10	.15/.11
Пац/Инстр	.31/.20	.25/.18	.21/.12	.19/.14
Адр/Экз	.19/.20	.18/.21	.18/.18	.21/.19
Адр/Лок	.20/.11	.21/.10	.18/.10	.19/.12
Адр/Инстр	.24/.21	.22/.18	.19/.21	.18/.21
Лок/Экз	.13/.33	.15/.27	.14/.29	.16/.22
Лок/Инстр	.26/.30	.25/.255	.18/.19	.16/.21
Инстр/Экз	.19/.22	.18/.19	.18/.23	.21/.24

Арч. - Арчинский язык. Тексты и словари, М., МГУ, 1977 /40 стр./

Бас. - Estornes Lasa B., Mundua euskal-eriaren gogoan, San Sebastian, 1977 (35 pp.); N'Diaye G.,Structure du dial. basque de Maya, The Hague,1970 - texte (3 pp.).

Тон. - Выборка Тон$_1$ /см. табл. 5/, а также:Caillot A., Mythes, lêgendes et traditions des Polynêsiens,P., 1914 (12 pp.)

Сам. - Steubel O., Myths and Legends of Samoa,Wellington, 1976 (35 pp.); Krämer A., Die Samoa-Inseln, Bd. 1-2, Stuttgart, 1902 (15 pp.).

ляет русский - в области СЖ Аг (ср. особенно пару Аг/Экз), однако здесь существенно, что это сходство абсолютных, а не относительных показателей: "партнеры" Аг в русском ведут себя неэргативно.

Задача 5. Генетически родственные и типологически разные языки.

Задача отчасти дублирует предыдущие.Рассматриваются предположительно разные по типу языки внутри одной генетической и ареальной группы - полинезийские (далее - ПЯ). ПЯ принято подразделять на восточные и западные; это деление типологически релевантно: языки первой подгруппы аккузативны, второй - эргативны.

Тонганский ($ТОН_1$) и самоанский, описанные выше, относятся к западной подгруппе. К ним в этой задаче прибавляется еще один западный ПЯ - ниуэанский (НИУ) и три восточных: маркезанский, или маркизский (МКЗ), маорийский (МАО) и гавайский (ГАВ). Данные приведены в табл. 6.

Мы получаем статистическое подтверждение типологической дивергенции ПЯ, но картина расхождений отличается от ожидаемой.

Как можно было ожидать (ср. задачу 3), $ТОН_2$ гораздо ближе к аккузативным восточным ПЯ. Вполне возможно, что будь в нашем распоряжении данные о современном САМ, картина его индексов тоже была бы иной: некоторые наблюдения о деэргативизации САМ имеются в (Milner,1976), (Chung, 1978).

Таблица 6. Полинезийские языки.

пара	я з ы к			
	ниу	мкз	мао	гав
Р/Аг	.11/.11	.06/.08	.07/.10	.07/.09
Р/Экз	.17/.09	.17/.09	.17/.08	.18/.09
Р/Пац	.09/.08	.10/.06	.08/.11	.08/.08
Р/Адр	.15/.13	.14/.06	.14/.07	.12/.07
Р/Лок	.14/.11	.13/.07	.10/.11	.10/.12
Р/Инстр	.15/.10	.19/.08	.15/.04	.15/.09
Аг/Экз	.15/.09	.14/.085	.16/.09	.15/.08
Аг/Пац	.09/.10	.09/.08	.07/.07	.09/.06
Аг/Адр	.13/.08	.14/.06	.14/.08	.15/.04
Аг/Лок	.15/.085	.12/.08	.11/.07	.115/.08
Аг/Инстр	.18/.12	.19/.10	.18/.07	.19/.07
Пац/Экз	.16/.11	.17/.13	.17/.10	.16/.12
Пац/Адр	.14/.19	.26/.11	.22/.16	.21/.16
Пац/Лок	.16/.10	.16/.13	.15/.10	.16/.115
Пац/Инстр	.21/.12	.20/.13	.24/.14	.23/.11
Адр/Экз	.16/.17	.13/.15	.14/.14	.12/.14
Адр/Лок	.155/.14	.14/.20	.15/.17	.15/.155
Адр/Инстр	.20/.21	.19/.22	.25/.23	.24/.21
Лок/Экз	.16/.23	.16/.14	.17/.12	.15/.15
Лок/Инстр	.20/.19	.22/.19	.17/.16	.20/.18
Инстр/Экз	.22/.20	.18/.15	.19/.18	.20/.16

см. данные по $Тон_1$, $Тон_2$ и Сам в табл. 5, 6.

Выборка:

Ниуэ - Loeb E.M., History and Traditions of Niue,Honolulu, 1926; Hill W.M., Tongakilo, Koe Tau Tala kia Maui, New Zeal. Dept. Educ., 1965 - ca. 50 pp.

Марк.- Handy E.W., Marquesan Legends, Honolulu, 1930 (45 pp.)

Маори - Orbell M., Maori Folktales in Maori and English, Auckland,1968; Grey G.,Nga Mahi a Nga Tupuna, Wellington, 1971; Biggs B. et al. (eds.),Selected Readings in Maori, Wellington, 1969 - ca. 80 pp..

Гав. - Fornander's Selections from Hawaiian Antiquities and Folklore, ed.S.Elbert, Honolulu, 1959 - 50 pp.

Наиболее интересным кажется тот факт, что по количественным показателям НИУ располагается между восточными и западными ПЯ, причем скорее ближе к восточным (особенно по показателям "верхушки"). Это противоречит его качественной оценке как "самого эргативного" ПЯ ((Hohepa, 1969, 326). Наши наблюдения показывают, что эргативность НИУ носит формальный характер (морфологическая эргативность), семантически и синтаксически НИУ акузативен и нейтрален (о нейтральности синтаксиса см. (Кибрик, 1979); см. также (Seiter, 1980), (Chung-Seiter, 1980)). Семантическая эргативность в НИУ также не наблюдается.

Свидетельств в пользу эргативности НИУ мало (в основном это индексы класса Инстр), и это показывает,что предлагаемый метод позволяет оценивать не само поверхностное представление языка, а его связи с планом содержания. Полученные данные хорошо иллюстрируют известную консервативность морфологии и ее возможное отставание от структурного типа языка. (Это, однако, не означает, что расхождения такого рода существуют всегда или что они безграничны.)

Итак, эмпирические данные указывают на меньшую, чем можно было бы ожидать, дивергенцию внутри группы ПЯ.

Восточные ПЯ явно неэргативны и сходны с индоевропейскими языками. Однако их аккузативность носит несколько иной характер. Можно видеть, что меры,полученные для центральных классов, в ПЯ ниже, чем соответствующие меры индоевропейских языков. В связи с этим можно высказать следующее предположение.

В ПЯ происходит некоторое подавление, "перекрывание" семантических связей составляющих предложения.Однако если в индоевропейских языках можно предполагать, что семантические связи перекрыты только на небольшом фрагменте рассматриваемой системы, то в ПЯ это явление заметнее и шире.

Возможно, что отношения семантических ролей (некоторые в большей, некоторые - в меньшей степени) не являются доминирующими в структуре высказывания ПЯ и сбалансированы иными отношениями.

В наши задачи не входит качественный анализ синтаксиса ПЯ, однако имеющиеся данные позволяют предположить, что эти отношения - прежде всего референционные (хотя, например, синтаксичность связи Р/Пац напоминает универсалию, см. задачу 8). Применение к ПЯ целого ряда синтаксических текстов, и в первую очередь, правил сдвига, показывает, что они контролируются в этих языках проекцией темы (топиком), а не синтаксическими подлежащими, дополнениями и проч. Мы не останавливаемся подробно на том, как именно референционные отношения "опережают" другие, об актуальности этих отношений в австронезийских языках уже говорилось в лингвистической литературе (см., например, (Ferrell-Stanley, 1980)).

Вероятно, в связи с этим, что вопрос об универ-

сальности семантических ролей и их соотношении с планом выражения не может быть решен однозначно. Семантические роли можно рассматривать и как абсолютные,и как относительные сущности. Они абсолютны, видимо, в силу самого факта своего существования (или "метасуществования", что в данном случае несущественно), а оттого и могут считаться универсальными. Однако они неуниверсальны по своему участию в механизмах конкретного языка, по своему статусу в его структуре и по влиянию на организацию конкретного высказывания. Подобная неуниверсальность ролей важна, поскольку она сказывается на оценке разнородных связей в различных языках или на разных участках одной языковой системы.

Поскольку зашла речь о существенности и возможной доминации референционных отношений, рассмотрим еще один язык, интересный с этой точки зрения.

Задача 6. Тагальский язык: тема или роль?

Тагальский, как и ПЯ, - австронезийский язык, относящийся к западной группировке названной семьи.ПЯотносятся к восточной, или океанийской, группировке.

Ведущая роль референционных отношений в тагальском подчеркивается в целом ряде описаний (например, (Schachter-Otanes, 1972),(Schachter, 1976)). В(Кибрик, 1979) тагальский описан как семантически аккузативный язык. Несомненно, отношения данного - нового, темы - ремы, и т.п. связаны с семантикой, однако это несколько иная семантика, чем семантика ролей. Итак,различает ли тагальский "две семантики",и что существеннее для него - тема или роль?

Если семантическая аккузативность действительно существует (аналогично семантической эргативности, подобной той, что в арчинском) и если она представлена в тагальском, то можно ожидать в этом языке высоких индексов СЖ, пусть иных по распределению, чем в арчинском. Реальные показатели (табл. 7) скорее подтверждают некоторое безразличие тагальского к семантическим отношениям. Тагальский демонстрирует хорошее сходство с аккузативными референционными языками - восточно-полинезийскими. Интересно, что меры, полученные для тагальского, даже несколько ниже соответствующих индексов в ПЯ.

Итак, количественные показатели позволяют сделать вывод о том, что в соотношении темы и роли приоритет принадлежит теме, а значит, тагальский следует относить не к семантичным, а к референционным языкам.

Задача 7. Выбор альтернативной гипотезы: дирбал.

Дирбал - язык, ставший предметом многочисленных лингвистических споров. После первого восторженного признания синтаксической эргативности и некоторой исключительности дирбала наступил период сомнений,появились работы, отрицающие эргативность дирбала (Jake, 1978);(Rood, 1977);(Heath, 1979). Все оценки дирбала,в том числе и данная - количественная - грешат одним: они

основываются на ограниченном материале (Dixon, 1972), который недостаточен и рискует быть непредставительным.

Как бы то ни было, статистика, полученная на этом скудном материале (данные в табл. 8), свидетельствует против какой-либо эргативности дирбала. Мы вправе ожидать, что для дирбала будут получены индексы, иные, чем в арчинском или баскском, чья эргативность не носит синтаксического характера, и даже иные, чем в самоанском, имеющем некоторые черты синтаксической эргативности/активности. Однако индексы дирбала не похожи на индексы табл. 5.

Дирбал проявляет большое сходство с ПЯ в индексах центральных классов, "верхушки". Сходство с эргативными языками обнаруживается только в индексах Инстр* (но не "партнеров" Инстр, ср. в связи с этим эргативное поведение Аг в русском). Таким, образом, дирбал неэргативен и похож на аккузативные языки референционного типа - восточные ПЯ (об актуальности референционных отношений в дирбале можно судить по материалу в (Dixon, 1972).

Интересной чертой, наблюдаемой только в двух исследованных языках - дирбале и маорийском - является некоторое повышение СЖ Пац в паре Р/Пац. Во всех рассмотренных языках, независимо от типа, этот индекс очень низок (0.06 - 0.09), а в маорийском и дирбале он несколько повышен (соответственно 0.11 и 0.115). Это может быть связано с чертами патетивности, обнаруживаемыми в обоих языках. Однако прочие СЖ Пациенса достаточно тривиальны; не наблюдается никакого существенного отличия в индексах пары Аг/Пац. Возможно, патетивность - черта ряда языков, но не проявление индивидуального, особого типа. Поэтому предполагается, что в пределах аккузативного пространства можно выделять подсистему языков с чертами патетивности.

Итак, статистика позволяет нам высказаться в пользу гиротезы о неэргативном характере дирбала. Скорее это аккузативный язык с чертами патетивности.

Задача 8. Анализ пар классов.

Теперь рассмотрим сами пары классов, на основании которых сравнивались языки. Попытаемся выяснить, какие из использованных пар ведут себя одинаково во всех языках и похожи на универсальные, какие специфичны и могут считаться диагностирующими, наконец, в каких лингвистических целях могут быть использованы те или иные парные индексы.

Поскольку существенным является соотношение индексов в парах и поскольку материал позволяет предположить взаимодействие мер, представляется более разумным анализировать именно пары, а не индивидуальные меры или обобщенные средние**.

---

* Данные по Инстр были обработаны, несмотря на их малую статистическую надежность, и необходима осторожность в их трактовке.

** Данные средних опущены в таблицах за недостатком места; читатель легко может получить их и убедиться, что они сглаживают межъязыковые различия.

Таблица 7. Тагальский язык

пара		пара		пара	
Р/Аг	.07/.10	Аг/Пац	.08/.07	Пац/Инстр	.16/.11
Р/Экз	.14/.07	Аг/Адр	.12/.08	Адр/Экз	.14/.09
Р/Пац	.07/.06	Аг/Лок	.13/.085	Адр/Лок	.16/.17
Р/Адр	.15/.09	Аг/Инстр	.16/.11	Адр/Инстр	.18/.15
Р/Лок	.13/.08	Пац/Экз	.15/.12	Лок/Экз	.17/.155
Р/Инстр	.16/.10	Пац/Адр	.21/.15	Лок/Инстр	.18/.165
Аг/Экз	.12/.07	Пац/Лок	.14/.105	Инстр/Экз	.20/.18

Выборка: Bloomfield L., Tagalog Texts with Grammatical is, pt. I, Urbana, 1917; Intermediate Readings in Tagalog, ed. J.D. Boven, Berkeley - LA, 1968 (tt. 47-52) - cf. 45 pp.

Таблица 8. Дирбал

пара		пара		пара	
Р/Аг	.06/.10	Аг/Пац	.12/.08	Пац/Инстр	.22/.12
Р/Экз	.13/.12	Аг/Адр	.13/.02	Адр/Экз	.19/.21
Р/Пац	.07/.115	Аг/Лок	.10/.02	Адр/Лок	.11/.11
Р/Адр	.12/.08	Аг/Инстр	.20/.06	Адр/Инстр	.21/.25
Р/Лок	.10/.06	Пац/Экз	.18/.08	Лок/Экз	.12/.15
Р/Инстр	.13/.10	Пац/Адр	.21/.09	Лок/Инстр	.17/.21
Аг/Экз	.14/.09	Пац/Лок	.16/.06	Инстр/Экз	.28/.25

Выборка - тексты в (Dixon, 1972).

Может возникнуть вопрос, не стоит ли оставить пары ради более сложных объединений классов, например, ради троек. Тройки центральных классов, видимо, могут быть привлечены для внутреннего анализа (анализа в пределах языка, группы родственных языков), но сведение всех классов в тройки едва ли целесообразно.

Обратимся к анализу конкретных пар.

1. Р - Пац. Это очень интересная пара, с удивительным постоянством получающая низкие индексы в самых различных языках. В связи с этим можно предположить, что существует некоторое весьма универсальное отношение, связывающее Р и Пац в рамках предложения, и приоритет в этом отношении принадлежит синтаксической связи.

Возможно, такая устойчивая связь ответственна и за распространенное в языках (универсальное?) присоединение имени Пац глаголом или отглагольной формой (так называемая объектная инкорпорация, имеющая место и в словообразовании, и в синтаксисе): Р и Пац легко стягиваются в один блок.

Как уже было сказано, мера СЖ Пац повышена в мао-

рийском и дирбале, что связано с чертами патетивности. Существенно, что языки (стили), где патетивность не выражена, но где по каким-то иным причинам повышено число пассивов (стиль (Б) в русском, латынь Николая Кузанского), не демонстрируют изменений СЖ Пац; напротив, в стиле (Б) русского языка в этой паре наблюдается некоторый прирост СЖ предиката.

2. Р - Инстр. Типологически различительная пара. Обе меры выше во всех рассмотренных эргативных языках и в дирбале. Видимо, более жесткая связь соотносится некоторым образом с омонимией эргатива - инструменталиса; семантичность связи становится значимее, поскольку формально эргативный Аг и Инстр могут быть неразличимы.

3. Аг - Экз. Диагностирующая пара, хорошо различающая аккузативность/эргативность, а кроме того, способная указывать на существенность ролевого или референционного компонента (ср. понижение индексов этой пары в австронезийских языках и в дирбале). Если референционные отношения подчинены ролевым, индексы в паре достаточно высоки, так как при одноместном предикате роли Аг и Экз являются взаимоисключающими. Если референционные отношения не подавлены или даже "перекрывают" ролевые, возможно линейное (синтагматическое) сочетание темы-роли и роли, не являющейся темой; индексы, естественно, снижаются (ср. самые низкие индексы в тагальском, 0.12/0.07).

Понижение индексов в австронезийских языках может быть также связано с актуальной для этих языков категорией неотчуждаемой (неконтролируемой, или пациентной) и отчуждаемой (контролируемой, непациентной, или агентивной) принадлежности. Эта категория проецируется в синтаксис, и при таком последовательном категориальном делении Экзистанты более равномерно распределены по отношению к Аг, к Адр и к Пац, что снижает каждую индивидуальную меру.

4. Пац - Экз. По сравнению с предыдущей, пара менее значима типологически. В рассмотренных эргативных языках наблюдалось незначительное повышение индексов с обеих сторон, по сравнению с аккузативными показателями. В целом, Пац и Экз достаточно жестко детерминируют друг друга, что связано прежде всего с отрицательной детерминацией: одна роль препятствует появлению другой в рамках данного интервала (при одноместных предикатах). Все те факторы, о которых шла речь в п. 3, ответственны за некоторое понижение обеих мер в этой паре в полинезийских языках.

5. Аг - Пац. Пара обладает хорошей различительной способностью. В эргативных языках обе меры достаточно высоки и ожиданным образом достигают максимума в арчинском. Для индоевропейских языков характерны отношения "хозяина" (Аг) - "слуги" (Пац), см. задачу 1. В референционных языках меры взаимно низки, однако, существенно, что в раннем тонганском ($ТОН_1$) и в самоан-

ском эргативное отношение в паре пересиливает референционное, и индексы отличаются от мер в других референционных языках - в сторону повышения.

6. Типологически неинтересны и, вероятно, универсальны связи в парах Аг - Адр, Аг - Лок и Аг - Инстр, где повсеместно реализуется отношение "хозяина - слуги". Однако, судя по данным стилей русского языка, меры могут несколько различаться в зависимости от стиля: при этом они сбалансированно меняются и справа и слева, не изменяя общего отношения.

7. Достаточно постоянные отношения, позволяющие предполагать универсалию, обнаруживаются для пар Адр - Пац и Адр - Экз во всех рассмотренных языках. Для первой пары мы могли бы ожидать специфические индексы в языках с регулярной аффективной конструкцией. Однако пара оказывается чувствительной не к поверхностной реализации, а к регулярной и, по-видимому, универсальной падежной рамке глаголов восприятия - аффекта. Устойчивость индексов, надо полагать, обеспечивается не только этими глаголами, но и различными конверсивами: ситуации с "дать /кто, что, кому/" очень распространены, а следовательно, статистически значимы.

Пара - Адр - Экз обязана своим универсальным характером отношению принадлежности, выражаемому - пусть по-разному - во всех языках.

8. Пара Экз - Лок является различительной парой:

- в эргативных языках очевидна тенденция Экз жестко задавать Лок, но не наоборот (отношения "хозяина - слуги"),

- в аккузативных языках отношения достаточно жесткие (СЖ относительно высоки), но практически равные, с некоторым перевесом в пользу Лок.

Первая из названных тенденций указывает на существенность Фактитива (в частности представленного Экз) в отношениях эргативности. Отметим, однако, что в других парах мы не получили каких-либо явных указаний на фактитивные отношения, что может свидетельствовать об актуальности фактитивных отношений не на всей системе языка, а на некоторых ее участках. Интересно,что в языках с предполагаемой эргативностью (ниуэанский, дирбал, тонганский?) отмечается очень незначительный перевес в паре в пользу Экз.

Относительно некоторого перевеса СЖ в пользу Лок в аккузативных языках трудно высказаться с определенностью. Возможно ad hoc связать этот перевес с тем,что мы задаем роль Экз "сверху" и и определяем ее прежде всего как роль актанта при бытийном предикате, а сама идея нахождения/бытия некоторым образом связана с актуальностью в-Локатива.

9. Диагностирующими являются пары центральных классов Р - Аг и Р - Экз: они позволяют четко разграничивать аккузативность и эргативность, а внутри каждого типа - проводить дальнейшее деление, например, с точки зрения семантического типа.

10. Наконец, во всех исследованных языках наблюдаются достаточно жесткие отношения между классами периферийных ролей. Подобные жесткие меры связаны с отношениями отрицательной (запретительной) детерминации: они объясняются отношениями конкуренции между периферийными ролями. Поскольку в одном высказывании невозможно совместить все аспекты ситуации, приходится жертвовать какими-то, отдавая предпочтение другим. Таким образом, из нескольких примерно равных по абсолютному статусу аспектов избирается какой-то один. Так, приходится отдавать предпочтение какому-то одному актанту, например, Лок, пренебрегая другими (той же роли); возможен выбор между Кауз и Инстр. Адр и Лок и т.д. Чем периферийнее составляющие, тем строже ограничения на их совместное появление в рамках данного интервала (здесь существенно и то, что интервалом является простое предложение).

В то же время периферийные актанты очевидным образом подчинены "верхушке": они не могут задавать центральные классы жестко, а последние, напротив, нередко детерминируют появление периферийных. Таким образом, "верхушка" и периферия вступают в отношения "хозяина - слуги".

Небезынтересные факты были прлучены при сравнении уже приведенных индексов для разных выборок русского языка (см. табл. 1 и 3) с некоторыми индексами, полученными для текста русского перевода (1876 г.) Ветхого Завета*. Индексы "верхушки" не отличаются от индексов, приведенных в табл. 1 и 3. Наиболее показательны следующие пары: Адр/Лок, .14/.20; Лок/Инстр, .18/.17; Адр/Инстр, .16/.175; Кауз/Инстр, .16/.07 /показатели по общей выборке - .18/.11/; $Лок_1/Лок_2$**, .15/.12. Можно заметить, что индексы, полученные на материале библейского текста, ниже соответствующих индексов, определенных для других выборок. Это, по-видимому, объясняется известным нагромождением ситуативных актантов*** в русском переводе 1876, служившем нам материалом.

11. В таблицы не включались данные по Кауз. Для тех языков, где статистика по Кауз была набрана, выяснилось, что независимо от типа языка Кауз достаточно нейтрально ведет себя в отношении Лок и Адр (примерно равные индексы справа и слева в парах) и демонстрирует высокие СЖ в парах с Аг и Инстр (вероятно, во многом это вызвано отрицательной детерминацией - запретом на совместное появление в рамках данного интервала). Универсально жесткое соотношение наблюдается в паре Кауз - Пац, что семантически хорошо мотивировано. Возможно,

---

* Выборка - Быт., 12 - 26; Иов, 3 - 12.

** $Лок_1$ - первый по линейному порядку актант, $Лок_2$ - второй; в других выборках пара не учитывалась как несущественная.

*** В целом ряде случаев разграничение актантов и сирконстантов становится проблематичным - см. об этом интересные замечания в (Vater, 1978).

различительную способность удастся установить для пары Кауз - Экз (высокие индексы в эргативных языках, особенно СЖ Экз; более низкие - в аккузативных), однако недостаток материала не позволяет пока утверждать это.

Существенно, что пара Р - Кауз оказывается нечувствительной к способам поверхностного оформления каузативных конструкций в том или ином языке. Далее, выясняется, что в каждом отдельно взятом языке индексы в паре Р - Кауз несколько выше индексов в паре Р - Аг (особенно заметно, что СЖ Кауз > СЖ Аг); это свидетельствует о большей семантичности связей в первой паре.

Ш. ЗАКЛЮЧЕНИЕ

Как можно видеть из задач, предлагаемый метод работает и позволяет выявлять и проверять некоторые языковые закономерности и тенденции. Метод может быть применен при типологическом сравнении языков; в таком случае наиболее существенными являются показатели в парах предиката и наиболее субъектных ролей. Далее, метод может быть использован в стилистическом анализе, и тогда становятся существенными индексы периферии. Кроме того, задав какие-то иные способы классификации, мы можем использовать метод для внутриязыкового анализа, в частности, для анализа в прикладных целях (оценка актантно-предикатных и межактантных связей при автоматической обработке текста). Метод позволяет судить о предпочтительности того или иного лингвистического описания (ср. задачи 6 и 7).

Метод, естественно, может быть уточнен и дополнен. Как уже было сказано, он может быть применен не только к семантическим классам (ролям), но и к формальным; кажется перспективным объединение двух подходов.

Далее, можно предсказать два уточнения метода. Первое - это разделение СЖ на СЖ разрешения (Е) и СЖ запрета ($\bar{E}$), то есть СЖ отрицательной детерминации. Второе - эксплицитный учет и введение в таблицы показателя К = $СЖ_1 : СЖ_2$.

Перейдем к содержательным выводам. По-видимому, при построении типологических классификаций трудно получить строго дискретные группы и "линейные" иерархии. Вводимое здесь представление в виде пространства - не попытка метафоры, а некоторая модель приближения к реальности.

Рассматривая всякий язык как систему, состоящую из подсистем разного рода и их взаимоотношений, мы не можем считать эту систему ни однородной, ни неподвижной, ни тем более строго упорядоченной. При типологическом анализе необходим учет многообразных тенденций, имеющих место в языке. Ясно, что неоднородность и разнонаправленность тенденций может сказываться и на разных подсистемах, и в пределах какой-то одной. Вероятно, это относится и к распределению универсального и специфического: весьма показательно, что наиболее существенные расхождения между рассмотренными языками обнаруживают-

ся в области "верхушки", центральных классов, а при переходе к более периферийным классам ролей различия типа несколько стираются. При интуитивной оценке типа всегда есть риск, что мы неверно определили соотношение тенденций; статистика призвана верифицировать качественные гипотезы.

При выявлении сетки вероятностных связей - от плана содержания к плану выражения - обнаруживается, что связи между классами составляющих сложны, зависят от конкретных таксонов и складываются из набора компонентов. Таким образом, тип языка, в частности, характеризуется и тем, какие виды связей - синтаксическая, семантическая, референционная, - преобладают на тех или других участках системы.

Отношения семантических ролей имеют различный статус в разных языках и неодинаково значимы для разных классов составляющих. Таким образом, возможно,что семантические роли функционируют не по абсолютным и универсальным, а по специфическим правилам, отчасти меняющимся от языка к языку.

Помимо связей актантов с предикатом, интерес представляют и межактантные связи (в нашем анализе - отношения индивидуальных ролей друг с другом). Для них можно выделить по крайней мере три типа отношений: безразличное (нейтральное); конкурентное;иерархичное (отношение "хозяина - слуги"). В связи с этим встает вопрос о том, что далеко не все соотношения актантов могут быть адекватно описаны в терминах имеющихся иерархий субъектных свойств.

* * * *

Автор выражает глубокую признательность А.Я.Шайкевичу за обсуждение работы на всех ее этапах и благодарит Г.Э. Чудинова за обсуждение § 11 и помощь в расчетах.

ЛИТЕРАТУРА

Кибрик А.Е. Материалы к типологии, эргативности,0.-1., ИРЯ, предв. публ., вып. 126. М., 1979.

Полинская М.С. Метод квантификации связей между элементами языковой структуры - Уч. зап. ТГУ. Труды по лингвостатистике, вып.1X. Тарту, 1983.

Chung S. Case Marking and Grammatical Relations in Polynesian. Austin, 1978.

Chung S., Seiter W. An overview of Raising and Relativization in Polynesian. - Language, vol. 56, No. 3, 1980.

Dixon R.M.W. The Dyirbal Language of Northern Queensland. Cambridge, 1972.

Ferrel R., Stanley P. Austronesian versus Indo-European: the case against case. - In: Austronesian Studies/Ed. by P.B. Naylor. Ann Arbor, 1980.

Heath J. Is Dyirbal ergative? - Linguistics, vol. 17, No. 5/6 (219/220), 1979.
Hohepa P. The accusative-to-ergative drift in Polynesian languages. - J. Polynesian Society, vol. 78, No. 3, 1969.
Jake J. Why Dyirbal isn't ergative at all? - Chicago Liguistic Society, Regional Meeting 14,Papers. Chicago, 1978.
Milner G.B. Ergative and Passive in Basque and Samoan. - Oceanic Linguistics, vol. 15, No. 1, 1976.
Rood D. Against artificial tree branches. - Intern. J. American Linguistics, vol. 43, No. 3, 1977.
Schachter P. The subject in Philippine languages: topic, actor, actor-topic or none of the above.- In: Subject and Topic/Ed. by C.N. Li, Austin, 1976 (рус. перевод в: "Новое в зарубежной лингвистике", вып. XI, М. 1982).
Seiter W. Studies in Niuean Syntax. N.Y., 1980.
Tchekhoff C. From ergative to accusative in Tongan. - In: Ergativity/Ed. by F. Plank. L., N.Y., 1979.
Vater H. On the possibility of distinguishing between complements and adjuncts. - In: Valence, Semantic Case and Grammatical Relations/Ed. by W. Abraham. Amsterdam, 1978.

SYNTACTIC TYPOLOGY: AN ATTEMPT AT QUANTIFICATION

Maria Polinskaya

Summary

Languages of different syntactic types and different linguistic families are compared on the basis of the suggested quantitative method generally aimed at evaluating the determination relationships between the linguistic units of a given level. Applied to different functional styles the method proves to be insensitive to the stylistic differentiation within the language. Typologically similar languages yield similar quantitative characteristics irrespective of their genetic relations. On the contrary, a typologically heterogeneous group of closely related languages (Polynesian) manifests quantitative variation in accordance with the distinction "accusative - ergative". The method can be helpful in the verification of alternative hypotheses: as applied to the Dyirbal (Australia) data it evidently favors the accusative hypothesis.

The obtained indices reflect the complex relationships established between the NP - NP's and NP - VP's in the sentence, semantic, syntactic and pragmatic components being equally relevant here. Certain parameters tend to be quasi-universal and non-varying across languages, whereas others, being language-specific, prove efficient in the typological comparison. The multi-parametric typological evaluation may result in constructing the typological hierarchy ("spatial") model, which seems more adequate than that of discrete characterization.

Possible applications of the method are briefly discussed in the paper, the method being promising in automatic translation.

ОПЫТ ОПРЕДЕЛЕНИЯ МЕЛОДИЧЕСКИХ ТИПОВ И ИХ ВЗАИМОСВЯЗЕЙ С ИНЫМИ ПРИЗНАКАМИ (НА МАТЕРИАЛЕ ЭСТОНСКИХ РУНИЧЕСКИХ НАПЕВОВ)

И. Рюйтел

1. Задачи и метод классификации

В статье представлен опыт классификации народных напевов, базирующейся на моделировании мелодии на основе мелодического контекста. Задача метода классификации - создать содержательную музыкальную типологию, позволяющую дифференцировать родственные мелодии, определить мелодические семейства и их подгруппы, а также выяснить их возможные взаимосвязи и пересечения. Предварительные анализы варьирования эстонских рунических напевов показали, что изменяться могут почти все элементы напевов: ритм, звуковой и интонационный состав, амбитус, а нередко даже финальный звук, основной устой и иные ладовые опоры, вследствие чего мелодия может приобрести почти неузнаваемое звучание, и границы мелодического семейства оказываются довольно неопределенными (Rüütel, 1980; Rüütel, 1981). Где кончаются напевы одного и начинаются напевы второго мелодического семейства, порой нелегко определить. Нередки случаи, когда один мелодический тип перерастает в тип совершенно иного качества. Выяснить такие интересные с теоретической точки зрения процессы возможно лишь на основе типологического исследования. Для решения такой задачи, как музыкальная типология, когда мы кроме сложности самой проблемы имеем дело и с большими массивами материала, необходимо обратиться к современной точной науке и разработать автоматизированный метод классификации. Анализ эстонских (а также близких северо-эстонским водских) рунических мелодий показывает, что наиболее стабильным признаком мелодического типа является система опорных звуков (опорная система) мелодии, являющаяся инвариантом и неизменно сохраняющаяся во всех (или в основной части) напевах одного мелодического типа (Рюйтел, 1977; 1979; 1980). Опорная система содержит наиболее значительные элементы, основной костяк мелодии, выражая ее основное содержание. В принципе она может служить хорошей основой для составления типологии, однако ее трудно определить заранее, до разграничения самой типологической группы. Она ясно выявляется лишь при анализе последней. Поэтому в основу нашей типологии легла основная форма типологической группы напевов, представляющая собой нормативную модель данного мелодического типа. Такие модели определяются итогом автоматического статистического анализа распределения

связей отдельных элементов мелодии, т.е. мелодического контекста. Они выражают статистическую норму, которой в определенной степени соответствуют все напевы одного мелодического типа, хотя в них могут встречаться и некоторые отклонения.

Найденные нормативные модели отдельных мелодических типов на следующем этапе анализа сравниваются со всеми исследуемыми напевами. Выясняется, какой модели они больше всего соответствуют, и классифицируются в соответствующую типологическую группу. Если напев отличается от двух (или трёх) моделей равным количеством элементов, он оказывается пограничным явлением.

Основа и метод классификации соответствуют взаимоотношению понятий "тип" и "вариант" в теории фольклора, согласно которому все варианты одного фольклорного произведения имеют общую основную форму, от которой отдельные варианты могут отличаться в деталях, имея, однако, общий инвариант. В данном случае основной форме мелодического типа соответствует нормативная модель, которая и берется в основу классификации, а общий инвариант выражается в опорной системе, составляющей наиболее существенную и стабильную часть нормативной модели, сохраняемую, как правило, в отдельных вариантах мелодического типа.

Наконец, создаются структурные модели найденных мелодических групп на основе их полного мелодического контекста. Такие модели функционируют в качестве генеративных грамматик и позволяют вывести все мелодические варианты, соответствующие структурным правилам данного мелодического семейства, в том числе отсутствующие в исходном материале. Цель описанных манипуляций - проникнуть в суть музыкального мышления представителей определенной фольклорной традиции, приблизиться к познаванию внутренних процессов их музыкальной компетентности и моделировать их (ср. Pelinski, 1981).

Принципы и первые итоги применения описанного выше метода классификации и строения генеративной грамматики (на относительно небольшом материале) представлены в Рюйтел, 1979. Позже в секторе музыкального фольклора Института языка и литературы АН ЭССР было создано специальное программное обеспечение для ЭВМ ЕС 1010 (см. Ruutel, 1981, с. 4-5). В настоящей работе представляется усовершенствованная методика анализа.

2. Опыт классификации 1-строчных рунических напевов

2.1. Материал исследования

Ниже приводятся итоги применения описанного выше метода классификации для определения типологических групп северо-эстонских 1-строчных напевов с трихордовой ладовой основой. В принципе нет необходимости ограничивать исходный материал по отдельным музыкальным или иным признакам, за исключением масштабного единства (1-строчные напевы, хотя бы на первом этапе исследования, целесообразно сопоставить с 1-строчными и т.д.).

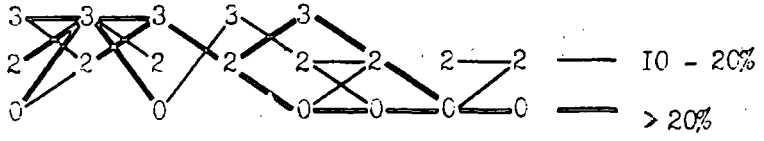
Наш метод расчитан также на равное количество счетных метрических единиц в мелодической строке (в руничес-ких напевах их, как правило, насчитывается 8). Отбор наиболее архаичного пласта эстонских рунических мелодий, т.е. 1-строчных напевов с трихордовой ладовой основой казался нам целесообразным для первичного испытания методики. А напевы определенного культурно-географического ареала (в данном случае - североэстонского) выбраны на превом этапе, во-первых, исходя от предположения, что в них мелодическое родство скорее всего выражает и генетическое (что важно при дальнейшей интерпретации результатов), а во-вторых, потому, что подавляющее большинство - 84 % имеющихся в отделениий фольклора Литературного музея АН ЭССР архивных записей с вышеназванными признаками принадлежит именно к североэстонскому ареалу, в то время как 26 % разбросано по остальной части территории Эстонии. Последние позже сравнивались с обнаруженными мелодическими типами Северной Эстонии (см. Rüütel, 1981, с. 28).

Ичходя из целей и особенностей типологического исследования (т.е. макро-уровня анализа) исследование опирается на основную форму напева, не учитывая отклонения от нее в течение песни. Такая основная форма в (звуко) записях всей песни определяется путем специального статистического анализа (см. Рюйтел, 1977), а в старых слуховых записях, составляющих основной материал данного исследования, чаще всего представлен лишь один, основной вариант.

2.2. Спределение нормативных моделей мелодических типов и разграничение мелодических семейств

На основе статистики связей соседних звуков мелодии сконструирована следущая модель исходного материала (связки, встречающиеся более чем в 25 % материала, отмечены жирной, а встречающиеся в 10-25 % - тонкой чертой; связки, встречающиеся менее чем в 10 % материала, в данной модели не учтены).

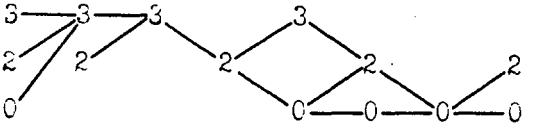
Таблица 1*



* Т.к. в руничских напевах основной звук (т.е. тонику) не всегда легко определить, то при кодировании нулем обозна-

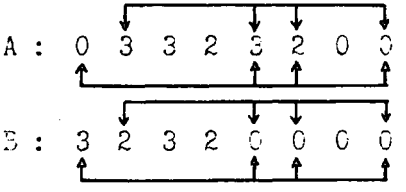
Из общей модели выделяется следущая, характеризующая основную, доминирующую подргуппу напевов:

Таблица 2



Модель содержит несколько параллельных элементов (позиции 1, 2, 5, 6 и 8), взаимоотношения которых не выявляются на данном этапе анализа. Поэтому далее разъясняются статистические отношения более отдаленных элементов мелодии. Для этой цели создается структурная модель на основе полного мелодического контекста исследуемого материала. По такой контекстуальной модели можно в данном случае определить доминирующие связи параллельных элементов модели, вследствие чего она делится на две отличающиеся в четырех позициях модели:

Таблица 3



В таблице 1 выделяется еще комбинация связок позиций 2-5: 3-0-3-2, по всей вероятности, относящаяся к нормативной модели отдельной ргуппы мелодии. Далее ис-

---

чается нижний опорный звук. Цифрами 1, 2, 3 ..., -1, -2 ... отмечается, на сколько полутонов остальные звуки звукоряда выше или ниже нуля. В дальнейшем анализе для выявления нормативных моделей мелодических типов полутоны заменяются (при помощи специальной программы ЭВМ) обозначениями ступеней звукоряда, т.к. высота той же ступенеи может варьироваться в вариантах того же мелодического типа и даже в течение одной песни (напр. малая и большая терция).

следуются отношения ее проминентных элементов (т.е. отличавшихся от моделей типов А и Б) с остальными позициями моделей, ступени которых пока остаются неизвестными. Выявляются следующие доминирующиеся связи:

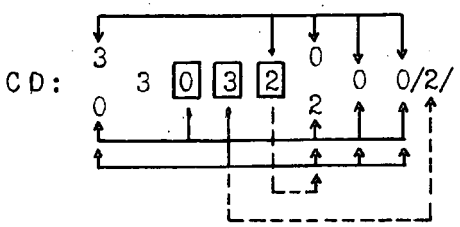
3-3-0-3-2-2-0-2
0 0 0

В найденной третьей модели также обнаруживаются некоторые параллельные элементы (позиции 1, 6 и 8), соотношения которых исследуются отдельно. При выявлении их доминирующих связей между собой и с остальными элементами (см. табл. 4) модель можно разделить на следующие подвиды:

C: 3
0 3 0 3 2 0 0 0

D: 3 2 2
0 3 0 3 0 2 0 2

Таблица 4



Последние отличаются четко лишь в двух позициях (6 и 8) и трактуются подвидами одного мелодического типа CD.

На основе найденных трех моделей А, В, и CD весь исходный материал делится на три группы. Каждый напев сравнивается со всеми моделями, причем определяется наиболее близкая к ней модель и классифицируется в соответствующую группу. Часть мелодий остается на границе двух (или трех) групп, т.к. отличается от них равным количеством элементов. В таком случае в дальнейшем анализе учитываются качественные показатели, т.е. предпочитается та модель, к опорной системе которой напев стоит ближе. Однако для выявления последнего необходима предварительная классификация основной части материала

и определение стабильности и мобильности отдельных элементов напевов соотвествующих групп (см.Rüütel,1981,12-13).

Часть мелодий не относится к обнаруженным группам, или же представляет собой принадлежащие к ним исключения (отличающиеся более, чем в трех позициях). Найденные группы мелодий в дальнейшем можно разделить на подгруппы (см. напр. Rüütel, 1981, 18-21). В данном случае ограничимся разграничением и анализом основных групп.

3. Составление контекстуальных моделей мелодических семейств и генерирование соответствующих им вариантов

Далее составляется полная контекстуальная модель для каждой группы мелодий. По ней можно определить все возможные (и не возможные) ступени звукоряда в каждой мелодической позиции при определенном контексте. Такая модель функционирует в качестве генеративной грамматики, позволяющей вывести все мелодические варианты, соответствующие правилам композиции данного мелодического типа. Подобная модель представлена в виде таблицы (см. таблицу 9), где в столбцах (1-8) указаны соответствующие мелодические позиции, а в строках - ступени звукоряда (см. обозначения в первом столбце). Цифры в столбцах 1-8 указывают количество вариантов, в которых встречается определенная ступень звукоряда при определенной ступени первой позиции (строки 1-16), затем при соответствующей ступени во второй позиции (строки 17-32) и т.д. Варианты генерируются автоматически по данным всей таблицы.

Такие таблицы в принципе не отличаются от полной контекстуальной модели, составленной в начале исследования по данным всего исходного материала и послужившей основой для определения нормативных моделей отдельных мелодических групп (семейств). По ней можно также вывести все мелодические варианты, соответствующие синтаксическим правилам, охватывающим возможности строения мелодий, содержащиеся во всем исходном материале. Однако наш опыт показывает, что более достоверных результатов можно достичь при более ограниченном и гомогенном материале. При соединении структурных правил разных мелодических семейств (а также при учете исключительных признаков и более отдаленных версий) можно наряду с достоверными реконструкциями получить и такие, которые вряд ли возможны в фольклорной практике. С другой стороны, при таких ограничениях, разумеется, отпадают и возможные контаминации разных мелодических типов, вполне возможные в традиции народной песни. Все же подобные утраты кажутся нам менее значительными, чем неизбежные нереальные конструкции, полученные при учете более гетерогенного по характеру (и более отдаленного по арелау распространения) материала. Иначе говоря, избирая "замкнутый" путь, можно лишаться некоторых реконструкций, соответствующих уровней музыкальной компетентности, и ограничить "твор-

ческие возможности", а избирая "открытый" путь, откроем тем самым дорогу и для теоретических спекуляций, не соответствующих музыкальному мышлению носителей исследуемой фольклорной традиции. Итоги последнего рода манипуляции приводились (правда, не менее обширном и гетерогенном материале) в первом варианте данной работы (Рюйтел, 1979). Ниже представлены итоги "замкнутого" пути на основе мелодий типа А, причем не учтены исключительные (лишь в одном варианте) встречающиеся признаки и более отдаленная (и в музыкальном отношении, и по региону) версия, встречаемая в песнях исполнителя из Ярвамаа (основным ариалом распространения типа А является Харьюмаа, как будет установлено ниже).

Мелодии типа А

1. Варианты, встречающиеся в исходном материале

1. 3332 3200	9. 0332 2200
2. 3302 3300	10. 0232 3200
3. 2332 3300	11. 0230 3200
4. 2332 3200	12. 3330 2200
5. 2332 2200	13. 3233 3200
6. 2302 3200	14. 0332 32-0
7. 0332 3300	15. 0323 3200
8. 0332 3200	16. 0302 320-

2. Варианты, отсутствующие в исходном материале

1. 3333 3200	6. 0333 3200
2. 3332 3300	7. 0302 3300
3. 3332 2200	8. 0302 3200
4. 3302 3200	9. 0233 3200
5. 2302 3300	

4. Определение взаимосвязей мелодического типа с регионом, несенными жанрами и ритмоформулой

Значимость статистической зависимости между рассматриваемыми признаками проверялась при помощи критерия Хи-квадрат:

$$x^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{i.} n_{.j} - n n_{ij})^2}{n_{i.} \cdot n_{.j}},$$

причем:

n_{ij} - число наблюдений в клетке таблицы i-той строки и j-го столбца

$n_{i.}$ - число наблюдений в i-той строке

$n.j$ - число наблюдений в j-том столбце

k - число строк

l - число столбцов

n - общее число наблюдений

Значение вычисленной статистики Хи-квадрат сравнивалось с табличным значением Хи-квадрат, с числом степеней свободы $(k - 1)(l - 1)$ при уровне значимости 0,05.

Теснота связи измерялась посредством коэффициента Чупрова.

$$T = \sqrt{\frac{x^2}{n\sqrt{(k-1)(l-1)}}} \quad .$$

Обоими методами было доказано, что между мелодическим типом и вышеназванными признаками существуют закономерные связи, причем наиболее существенной оказалось связь между мелодическим типом и регионом ($\chi^2 = 173{,}0$; $T = 0{,}57$), наиболее слабой - между мелодическим типом и жанром ($T = 48{,}6$; $\chi^2 = 0{,}25$).*

Далее такая зависимость исследовалась подробнее. Для этого вычислялся коэффициент, характеризующий различие эмпирических данных от теоретических (где теоретические вычислены при предположении независимости признаков). Теоретическая частота при независимости признаков вычисляется при помощи формулы

$$p_{ik} = \frac{n_i \cdot n_k}{n}, \text{ где}$$

n_i - сумма частот i-той строки

n_k - сумма частот k-того столбца

n - общая сумма наблюдений,

интересующий нас коэффициент равняется отношению $\frac{n_{ik}}{p_{ik}}$.

---

* Обработка данных произведена в вычислительном центре ТГУ при помощи стандартного программного обеспечения математической статистики.

Если эмпирическая встречаемость (n_{ik}) равна теоретической (p_{ik}), то получаем значение коэффициента 1. Если коэффициент >1, то напрашивается вывод, что между данными признаками существует в определенном смысле положительная корреляционная зависимость. Если же коэффициент <1, то мы имеем дело с отрицательной корреляцией. Положительная корреляция тем сильнее, чем больше коэффициент, отрицательная тем сильнее, чем меньше коэффициент.

При анализе корреляций между мелодическим типом и регионом сравнение абсолютных частот встречаемости и корреляционных коэффициентов особых сюрпризов не вызывает (см. табл. 5). В типе А в абсолютных цифрах преобладает материал Харьюмаа, который дает и единственную, притом сильную положительную корреляцию. Тип В имеет логично сильную положительную корреляцию с ареалом Вирумаа. Что касается типа СД, где в абсолютных цифрах и в процентах преобладает материал Ярвамаа, то здесь кроме ареала Ярвамаа обнаруживается сильная положительная корреляция с ареалом Сев. Тартумаа, который в абсолютных цифрах представлен наименьшим числом вариантов.*

При анализе связи мелодического типа и ритмоформулы (табл. 6) выясняется, что тип А кроме доминирующей в абсолютных цифрах 3-ей ритмоформулы имеет сильную положительную связь со второй и положительную связь с первой ритмоформулой, а тип СД кроме доминирующей 4-ой ритмоформулы проявляет положительную корреляцию и с первой. Здесь необходимо уточнить, что на самом деле все примеры с 1-ой ритмоформулой относятся к подгруппе С, а в подгруппе Д, которая проявляет наиболее яркий характер качельного напева, встречается только 4-ая ритмоформула. Наиболее смутную картину по абсолютным цифрам представляет жанровая таблица. Здесь коэффициенты вычислялись сначала по всему исследуемому материалу, а затем отдельно в каждом регионе. В таблице 7 представлены абсолютные цифры и коэффициенты по всему материалу. Тип А имеет наиболее сильную положительную корреляцию со свадебными песнями, с календарными и лирическими, В - с игровыми, лироэпическими, лирическими и заклинаниями (все 3 варианта заклинаний, встречаемые в материале, относятся к типу В), и СД - с качельными, игровыми, колыбельными и танцевальными, т.е. с песнями, сопровождающими движение. Еще более наглядной становится картина по таблтце 8, где указаны лишь положительные корреляционные связи мелодических типов и жанров, причем сначала по всем данным (Σ), а затем в отдельных регионах. Заметные различия в корреляциях по общим данным и по отдельным регионам наблюдаются в типе В. Если по общему материалу тип В не проявляет положительной корреляции со свадебными и календарными пес-

---

* Высокий коэффициент корреляции возникает здесь из-за того, что из имеющихся песен подавляющее большинство (3/4) относится к типу СД.

нями, то в регионе Вирумаа, который является основным ареалом распространения типа В, он весьма характерен и для свадебных и календарных песен.

Вообще таблица 8 наглядно показывает, что типы А и В представляют собой т.н. "общие напевы" для разных песенных жанров, при исполнении которых господствует речитативное начало. Если к ним и присоединяется какое-нибудь движение, напр., переступание с ноги на ногу и раскачивание, то оно, очевидно, секундарно и не оказывает столь сильного воздействия на мелодию. Тип СД зато везде проявляет предпочтение тех жанров, при исполнении которых господствует кинетическое начало, т.е. где движение, сопровождающее песню, оказывает сильное влияние на строение напева. Впрочем, и в качельных и колыбельных песнях мы имеем дело с движущимся реквизитом (если так можно выразиться), к которому песня приспосабливается, а в танцах всегда шаг четко соответствует ритму мелодии.

Лирические и лиро-эпические песни Северной Эстонии носят двоякий характер: они могут исполняться и на качелях, и без движения, связываясь соответственно то с качельными напевами, то с напевами речитативного характера. В некотором смысле "переходный" характер носят и игровые песни (на это обращал внимание уже Тампере): они по содержанию близки к лиро-эпическим, но их исполнение сопровождается каким-либо движением, правда не всегда четко сочетаемым с ритмом напева, как это может иметь место, напр., и в свадебных песнях.

Можно сделать заключение, что выявленные мелодические типы проявляют не столько жанровую, сколько именно синкретическую специфику, в первом случае - заметную связь с речевой интонацией (типы А и В), а во втором - явный кинетический характер (тип СД, особенно Д) (подробнее см.: Rüütel, 1981).

Первобытный синкретизм напева и речи, напева и единого поэтического строя явно объясняет известную жанровую расплывчатость напевов рунической песни, особенно тех видов, которые имеют речитативную основу. С другой стороны, именно на основе синкретизма формируются и некоторые жанровые особенности мелодий, особенно там, где мы имеем дело с синкретизмом напева и сопровождающего песню движения, напр. в качельных и колыбельных песнях.

Разъяснение таких важных с теоретической точки зрения проблем народного творчества, как взаимоотношения категорий песенного жанра, мелодического типа, ритма, явлений синкретизма и т.д. возможно лишь на основе содержательной музыкальной типологии и путем конкретного, объективного анализа взаимосвязей мелодического типа с иными категориями.

Таблица 5

Регион \ Мел. тип	А	В	СД	А В СД и др.	Всего
Харьюмаа	67-2,10	0-0	14-0,62	18	99
Вирумаа	1-0,03	66-2,34	17-0,77	13	97
Ярвамаа	14-0,79	7-0,44	24-1,92	10	48
Сев. Тартумаа	0-0	1-0,86	3-3,29	0	4
	82	74	58	41	255

Таблица 6

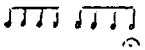
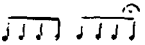
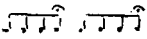
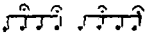
Ритма формула \ Мел. тип	А	В	СД	А В СД и др.	Всего
1. 	15-1,20	6-0,53	12-1,35	6	39
2. 	7-2,18	0-0	0-0	3	10
3. 	47-2,66	0-0	2-0,37	6	55
4. 	13-0,27	68-1,55	44-1,28	26	151
	82	74	58	41	255

Таблица 7

Жанр \ Мел. тип	А	В	СД	А В СД и др.	Всего
1. Трудовые	2-0,68	2-0,86	1	3	8
2. Календарные	5-1,11	4-0,98	3-0,94	2	14
3. Качельные	20-0,59	27-0,89	30-1,26	28	105
4. Игровые	2-0,48	5-1,32	5-1,69	1	13
5. Колыбельные	1	0	3-3,30	0	4
6. Свадебные	29-2,58	5-0,49	0	1	35
7. Лиро-эпические	5-0,86	6-1,15	4-0,98	3	18
8. Лирические	17-1,10	19-1,36	9-0,82	3	48
9. Заклинания	0	3 3,44	0	0	3
10. Танцевальные	0	0	2-4,40	0	2
11. Иные	1	3	1	0	5
	82	74	58	41	255

Таблица 8

	Песни речитативного характера	Песни с двояким характером	Песни с напевами кинетического характера
А Σ	свад. кал.	лир.	
Ха	свад. кал.	лир.	
Яр	свад.	л-эп. лир.	
В Σ закл.		л-эп. лир. игр.	
Ви закл.	свад. кал.	л-эп. лир. инр.	
СД Σ		игр.	кач. кол. танц.
Ха		лир.	кач. кол.
Ви		лир.	кач.
Яр		л-эп. игр.	кач.

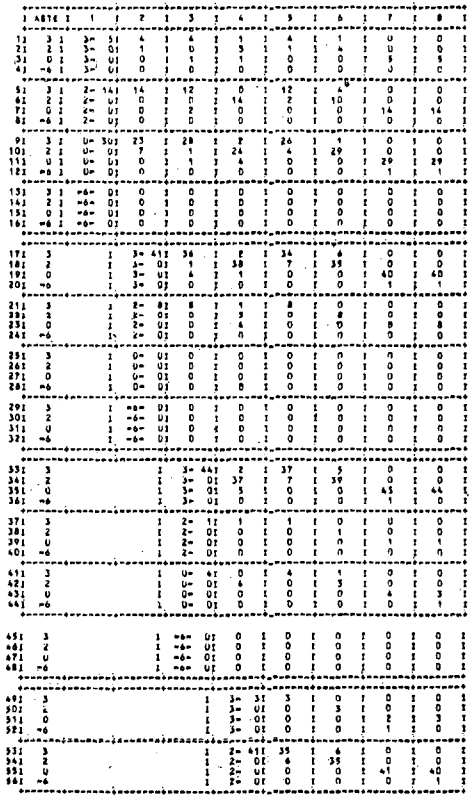
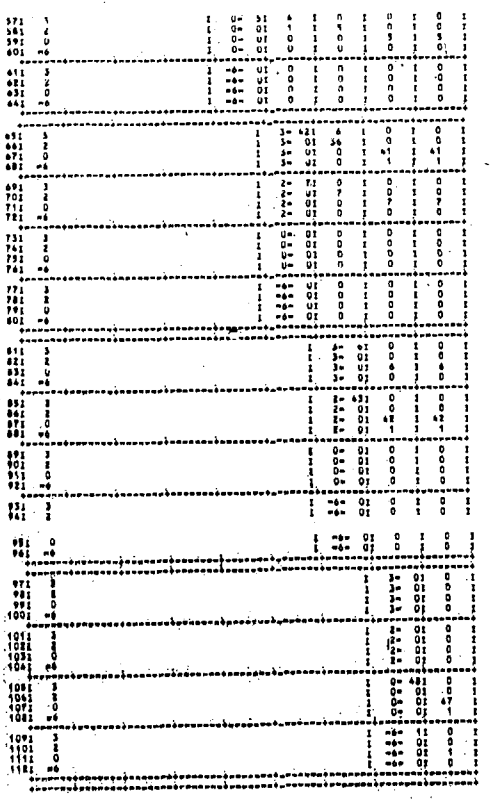


Таблица 9



ЛИТЕРАТУРА

Рюйтел И. Опорная система и закономерности варьирования мобильных элементов как структурные признаки типологии народных напевов. - В кн.: Проблемы таксономи эстонских народных мелодий. Таллин, 1977.

Рюйтел И. Опыт структурно-типологического исследования однострочных рунических напевов. Preprint KKI - 10. Таллин, 1979.

Pelinski R. La musique des Inuit du Caribou. Montréal, 1981.

Rüütel I. Mustjala regiviiside tüpoloogia. - Ars Musicae Popularis. Tallinn, 1980.

Rüütel I. Typology of Estonian runo-tunes: experiment and some results. Preprint KKI - 18. Tallinn, 1981.

ESTABLISHING TUNE TYPES AND THEIR RELATIONS WITH OTHER FEATURES (ON THE MATERIAL OF ESTONIAN RUNO-TUNES)

Ingrid Rüütel

Summary

In this article a computerized experiment in folk tune typology based upon modelling melody on the ground of melodic context is presented. The total contextual model of the investigated tune material (the North-Estonian 1-line runo-tunes) was found from which models of separate tune types were determined. On the basis of the latter the tune material was divided into separate groups. Finally, the structural models for each group were constructed which served as the grammars of the tune-families and enabled us to derive all the melody variants corresponding to the models (including those not presented in the material analysed).

The tune groups found appeared to be in a certain correlation with the topographic region, song genres and rhythmic patterns. The correlations between the items considered were examined by means of the X^2-criterion, Chuprov coefficient and the λ-coefficient, characterizing the difference between the empirical and the theoretical data.

К ВОПРОСУ О МОДЕЛИРОВАНИИ РОСТА СЛОВАРЯ

Ю.А. Тулдава

В статье рассматриваются возможности квантитативного изучения и моделирования роста лексики литературного языка на основе данных об объемах "представительных" словарей эстонского языка XVII - XX вв. Обсуждаются вероятностные модели, основанные на экспоненциальном и логистическом законах развития.

Данные о словарях. Известно, что словарный состав языка представляет собой "открытую" систему и его объем не поддается точному учету из-за словообразовательной, семантической и стилистической подвижности языка. Тем более трудно изучать рост объема лексики в течение исторического развития конкретного языка. Все же некоторое приблизительное представление о росте лексики языка может дать сравнение объемов представительных (наиболее полных и нормативных) для своего времени словарей разных периодов развития литературного языка.* В данной работе взяты за основу следующие двуязычные и одноязычные словари:

1. Гезекен (Göseken H., 1660) - 10 000 слов
2. Хупель 1 (Hupel A.W., 1780) - 14 000 "
3. Хупель 2 (Hupel A.W., 1818) - 21 000 "
4. Видеманн 1 (Wiedemann F.J., 1869) - 50 000 "
5. Видеманн 2 (Wiedemann F.J., 1893) - 60 000 "
6. Эст. ортологич. словарь (ЭОС) (Eesti õigekeelsuse sõnaraamat, 1925-1937) - 120 000 "
7. Орт. словарь (ОС-60) (Õigekeelsuse sõnaraamat, 1960) - 105 000 "
8. Орт. словарь (ОС-76) (Õigekeelsussõnaraamat, 1976) - 115 000 "

---

* Эстонский литературный язык донационального периода берет свое начало в XVI веке и проходит стадии "становления, формирования и стабилизации". В середине XIX в. формируется общенациональный литературный язык. Более подробный анализ составов эстонских словарей разных периодов см. Тулдава Ю., 1984.

Встает вопрос, можно ли на основе приведенных данных сделать некоторые заключения о закономерностях роста лексики в периоды становления, формирования и стабилизации литературного языка. Хотя прямое сравнение объемов рассматриваемых словарей затрудняется тем, что принципы составления словарей не являются одинаковыми, все же составы этих словарей можно рассматривать как отражение конкретных этапов развития языковой ситуации. В этом смысле количественные данные выражают какие-то общие тенденции в развитии конкретного языка. Измеряя рост лексики по данным словарей, мы косвенно регистрируем и лингвистические потребности носителей языка в разные периоды развития общества.

Экспоненциальный закон роста. Общеизвестным является тезис о том, что "в результате постоянного расширения сферы деятельности человека лексика каждого языка, особенно его терминологический словарь, несмотря на выпадение некоторого количества слов, неуклонно растет" (Пиотровский Р.Г. и др., 1977, с. 56). С математической точки зрения такому неуклонному нарастанию объема словаря соответствует экспоненциальный закон роста по следующей формуле (Пиотровский Р.Г. и др., 1977, с. 57):

$$L_T = L_0 e^{kT}, \qquad (1)$$

где T - промежуток времени (например, столетие), L_T - объем словаря к концу периода T, L_0 - начальный объем словаря, k - коэффициент прироста, e - основание натуральных логарифмов. По экспоненциальному закону скорость роста словаря имеет "лавинообразный" характер (скорость роста словаря пропорциональна достигнутому уровню), который может быть описан следующим дифференциальным уравнением:

$$\frac{dL_T}{dT} = kL_T \qquad (k>0),$$

где k - константа. Из уравнения вытекает, что скорость роста dL_T/dT линейно зависит от достигнутого уровня L_T, а относительная скорость роста (темп прироста) $(dL_T/dT)/L_T$ остается постоянной величиной. Решая это дифференциальное уравнение, мы и получаем уравнение экспоненты (1).

Проверка показывает, что экспоненциальному закону вполне соответствует рост лексики по данным рассматриваемых словарей в промежутке времени 1780...1937 г.,т.е. начиная с первого издания словаря Хупеля (14 000 слов) и кончая ЭОС-ом (120 000 слов). Формула (1) принимает вид:

$$L'_T = 1000e^{1,45T}$$

Соответствие эмпирических данных теоретическим хорошее (кривая I на рис. 1; см. также табл. 1; числа округлены до целых тысяч).* По условиям экспоненциального закона роста период удвоения составляет в данном случае 48 лет**, т.е. примерно за каждое полустолетие объем словаря должен удваиваться. При таком темпе роста можно прогнозировать, например, что в 2000 году объем эстонского словаря будет равен 330 000 словам, а в 2100 году - полутора миллионам слов.

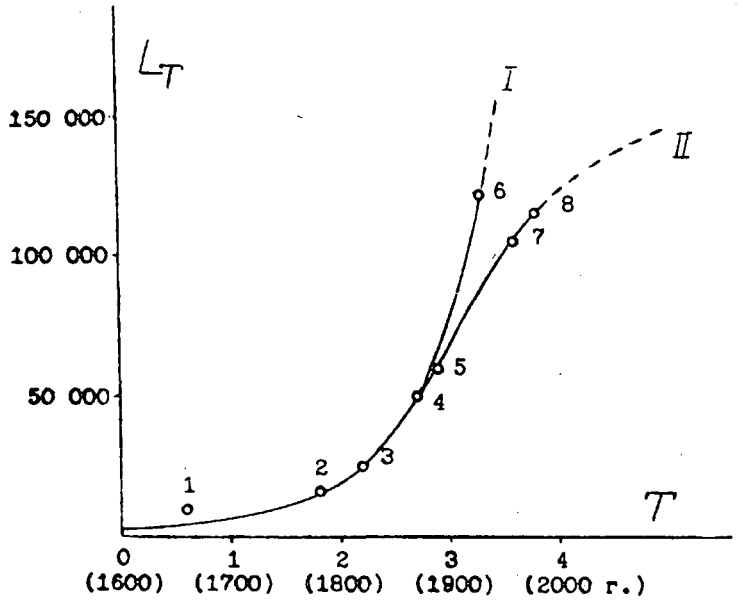


Рис. 1. Рост лексики эстонского литературного языка по данным словарей XVII - XX вв. Выравнивание и прогноз по экспоненциальной (I) и логистической (II) функциям. Цифры на схеме указывают на словари (см. табл. 1).

---

* За точку отсчета берется год 1600 ($T = 0$). Далее учитывается отдаленность в столетиях от данной точки отсчета, например, при 1780 г. (год появления первого издания словари Хупеля) $T = 1{,}8$.

** Период удвоения вычисляется как $T_y = \frac{\ln 2}{k}$. В данном случае $k = 1{,}45$, следовательно, $T_y = \frac{\ln 2}{1{,}45} = 0{,}48$, т.е. 48 лет (T_y выражает столетие).

Таблица 1

Эмпирические и теоретические данные о росте лексики эстонского литературного языка.

(L'_T - экспоненциальное, L''_T - логистическое распределение)

№ пп.	Год	T	L_T(эмпир.)	L'_T(теорет.)	L''_T(теорет.)
1.	1660	0,6	10 000	2 000	2 000
2.	1780	1,8	14 000	14 000	13 000
3.	1818	2,2	21 000	24 000	24 000
4.	1869	2,7	50 000	50 000	50 000
5.	1893	2,9	60 000	67 000	60 000
6.	1930	3,3	120 000	120 000	86 000
7.	1960	3,6	105 000	185 000	105 000
8.	1976	3,8	115 000	247 000	115 000
-	2000	4,0	прогноз:	330 000	124 000

Логистический закон роста. Представляется, что рост общеупотребительной лексики и рост объема словарей литературного языка за разные промежутки времени может характеризоваться экспоненциальным законом только в отдельные периоды развития языка. В действительности процесс роста лексики начинается медленно (период становления литературного языка), затем ускоряется и принимает "лавинообразный" характер (период формирования литературного языка), но в какой-то момент процесс роста обязательно замедляется (период стабилизации). Такой схеме развития отвечает математическая модель, выражаемая т. наз. логистической функцией:

$$L_T = \frac{L_\Pi}{1 + ae^{-kT}}, \qquad (2)$$

где L_T - объем словаря к концу периода T, L_Π - теоретический предел роста словаря (асимптота), a и k - параметры функции. Графически эта модель представляется S-образной кривой, которая выражает сначала рост с возрастающей скоростью, затем скорость уменьшается и почти прекращается по мере асимптотического приближения к некоторому пределу (кривая П на рис. 1). Формулы (1) и (2) близки друг другу в том смысле, что при $L_\Pi >> L_T$ (на начальных стадиях роста) кривая по фор-

муле (2) практически совпадает с кривой экспоненты по формуле (1); это хорошо видно также на графике. Таким образом, рассматриваемая кривая по формуле (2) по существу включает экспоненциальный тренд как первую стадию роста. Логистическая кривая имеет "точку перегиба", т.е. пункт перехода, где начинается непрерывное замедление скорости роста. За точкой перегиба кривая напоминает изображение логарифмической функции, отражающей "закон адаптационного торможения" (Налимов В.В., Мульченко З.М., 1969, с. 41 и след.). Анализ показывает, что кривая логарифмической функции по формуле $L_T = a + b \ln T$ (в данном конкретном случае $a = -160\,000$, $b = 206\,000$) практически совпадает с кривой логистической функции по формуле (2) на начальных этапах замедления роста лексики (в пределах $T = 3 \div 4$, т.е. между 1900 и 2000 гг.).

Можно констатировать, что закон логистического роста действительно имеет силу в отношении роста лексики эстонского литературного языка, если исключить из рассмотрения словарь Гезекена 1660 года и Эстонский ортологический словарь 1925-1937 гг. Вычисление параметров функции (2) на основе линеаризации* дает результат: $L_n = 150\,000$, $a = 280$ и $k = 1{,}8$, т.е. теоретические значения функции можно вычислить по формуле:

$$L''_T = \frac{150\,000}{1 + 280e^{-1{,}8T}} .$$

Соответствие между эмпирическими и теоретическими данными хорошее (кривая П на рис. 1; см. также табл. 1; числа округлены до целых тысяч). Однако здесь, так же как и при исследовании экспоненциальной функции по формуле (1), обнаруживается, что теоретические данные не совпадают с эмпирическими в отношении первого рассматриваемого словаря - словаря Гезекена 1660 года. Это

---

* Логарифмируя (2), получим: $\ln\left(\frac{L_n}{L_T} - 1\right) = \ln a - kT$,

т.е. линейную связь между $\ln\left(\frac{L_n}{L_T} - 1\right) = y$ и $T = x$.

С помощью графики на плоскости (X, Y) устанавливается то значение L_n , которое дает наилучшее приближение к линейной зависимости. Предварительный анализ на графике полезен тем, что отклонение от общей тенденции отдельных эмпирических точек (в данном случае пунктов 1 и 6) становится явным, и эти точки можно при вычислении значений параметров не учитывать. Значения параметров можно выявить обычным способом на графике или вычислить методом наименьших квадратов.

объясняется, по-видимому, тем, что наиболее ранние периоды становления литературного языка характеризуются иной тенденцией развития. Сравнение словарей XVII и XVIII вв. показывает, что литературная лексика в этот период растет не экспоненциально, а более медленными темпами. Только начиная с конца XVIII в. рост лексики принимает экспоненциальный характер, который продолжается до XX в.

Логистической функции соответствует дифференциальное уравнение

$$\frac{dL_T}{dT} = k' L_T (L_\Pi - L_T),$$

которое раскрывает "механизм" роста по этой функции (см. Налимов В.В., Мульченко З.М., 1969, с. 21)*. Здесь рост ограничен, т.к. L_Π (теоретически максимальное значение L_T) представляет собой предел роста словаря. Относительная скорость роста (темп прироста) уже не остается постоянной, она оказывается линейной функцией L_T :

$$\frac{dL_T}{dT} \cdot \frac{1}{L_T} = k'(L_\Pi - L_T).$$

Чем выше становится достигнутый уровень, тем ниже оказывается скорость роста.

Выше было упомянуто, что логистическая кривая имеет точку перегиба, за которой начинается замедление скорости роста. В данном случае этот пункт находится приблизительно при $T = 3{,}1$, т.е. в 1910 году.** Примерно в это время начинается период "форсированного развития" эстонского литературного языка (Saari H., 1979), но в то же время начинается и регулирующая (организационная) деятельность языковедов. Эти две тенденции прослеживаются и в развитии эстонской лексикографии. С одной стороны, увлечение сбором слов и включение в общеупотребительный словарь (ЭОС 1925-1937 гг.) многих устарелых слов, словесных дублетов и т.д. привело к чрезмерному наращиванию темпов роста словаря (см. кривую I на рис. 1). С другой стороны, регулирующая деятельность языковедов, которая в довоенные годы, по-видимому, не могла полностью утвердиться, начинает оказывать свое действительное влияние в послевоенные годы, завершаясь выпуском ортологических словарей 1960 и 1976 годов. Эти словари уже вписываются в

---

* Коэффициенту k' соответствует k/L_Π по формуле (2).

** Точку перегиба вычисляют по формуле $T_\Pi = \frac{\ln a}{k}$. Здесь a = = 280, k = 1,8; следовательно $T_\Pi = \frac{\ln 280}{1{,}8} \approx 3{,}1$.

общую картину развития эстонской лексикографии как естественное продолжение стабилизации литературного языка, начатой в скрытой форме еще в 1910-1920 гг. (см. кривую II на рис. 1). Конечно, и период стабилизации литературного языка имеет свои "приливы и отливы", но мы их в данном случае игнорируем и говорим только о наиболее общих тенденциях развития лексики в историческом плане.

Далее, анализируя модель роста эстонского словаря по логистической функции (2), выясняется, что для данной тенденции развития соответствует максимальный объем словаря (L_n), равный 150 000 словам. Хотя цифры в наших моделях могут иметь только приблизительное, ориентировочное значение, можно указать на то, что согласно модели формально ожидается предел $L_n \approx 150\ 000$ при условии продолжения актуальной тенденции развития лексики в периоде зрелости литературного языка (при этом не учитывается рост узкоспециальной терминологической лексики). Возможна также интерпретация предела L_n как "уровня равновесия" или "оптимального объема" словаря при данных условиях развития.* Однако в данном случае задачей исследования был не прогноз будущего. Нельзя предвидеть все потребности и технические возможности лексикографии в будущем. Поэтому установление предела по данной модели носит чисто теоретический характер и конкретное значение L_n не рассматривается как некий реальный или идеальный предел роста словаря. Предлагаемая математическая модель роста словаря носит в данной работе лишь обобщающий характер и указывает на о б щ и е т е н д е н ц и и роста лексики эстонского литературного языка по данным представительных словарей за 300-летний период.

---

* Для сравнения можно привести некоторые общеизвестные данные об объемах словарей других современных языков: Толковый словарь русского языка Д.Н. Ушакова (1935-1940) - около 85 000 слов; 17-томный Словарь совр. русского литерат. языка (1948-1965) - 120 000 слов; 6-томный Украинско-русский словарь (1953-1963) - 122 000 украинских слов; 8-томный Толковый словарь грузинского языка - 113 000 слов; Словарь совр. финского языка (Nykysuomen sanakirja, 1951-1961) - 200 000 слов; Акад. словарь шведского языка (Svenska Akademiens Ord Lista) - 160 000 слов; Акад. словарь итальянского языка -115 000 слов; Словарь совр. нем. языка (Wörterbuch der deutschen Gegenwartsprache; hrsg. von R. Klappenbach und W. Steinitz) - 100 000 слов. Особую группу среди представительных словарей составляют словари кумулятивного типа и словари, которые содержат много узкоспециальной лексики (например, Большой Оксфордский словарь английского языка, охватывающий около 450 000 единиц). Во всех случаях данные о словарях надо рассматривать в контексте исторического развития конкретного языка, учитывая также традиции в лексикографии данной страны.

Заключение. Исходя из теоретических предпосылок и в соответствии с эмпирическими данными, можно высказать предположение, что рост лексики эстонского языка за время прохождения этапов становления, формирования и стабилизации литературного языка подчиняется закону логистического развития. Такой вывод можно сделать на основе анализа роста лексики по данным наиболее полных и нормативных для своего времени словарей XVIII-XX вв. В идеале имеется в виду не кумулятивный (накопленный) рост лексики, а "чистый" прирост, исключая вышедшие из живого употребления устарелые слова. Кроме того, не подлежат учету узкоспециальные термины, а учитывается только такая общеупотребительная и специальная лексика, которая приводится в словарях типа ортологических и толковых. Логистический закон характеризует процессы роста, которые начинаются медленно с постоянным ускорением ("лавинообразный" характер роста), а затем переходят в стадию замедленного роста, приближаясь асимптотически к некоторому теоретическому пределу. Темп роста лексики вначале растет также медленно (этап становления литературного языка), затем заметно ускоряет рост почти экспоненциально (этап формирования литературного языка) и, наконец, переходит в стадию замедленного роста (этап стабилизации литературного языка). Вопрос о пределе роста словаря на данном этапе не имеет однозначного решения. Не исключается возможность прогнозирования роста лексики в периоде стабилизации литературного языка с помощью какой-нибудь другой функции, например, логарифмической функции, не имеющей предела.

По всей вероятности, закон логистического развития (возможно с упомянутой модификацией) имеет всеобщее социально-лингвистическое значение и характеризует рост лексики большинства литературных языков, хотя конкретную форму этот закон принимает в зависимости от условий исторического развития данного народа - носителя языка. Безусловно возможны и отклонения от общей тенденции развития под воздействием различного рода случайных факторов. В этом смысле процесс роста лексики языка можно рассматривать как динамическую вероятностную систему, характеризующуюся чертами вариативности и устойчивости.

Можно добавить, что логистический закон роста в различных своих конкретных проявлениях (имеется ряд вариантных формул логистического роста; см., например, Янч Э., 1974; Wingert F., 1971) считается одним из основных законов развития самоорганизующихся сложных систем, если рассматривать их развитие при достаточно больших временных интервалах. Соответствующие этому закону S-образные кривые находят себе в наши дни широкое применение не только в технике и биологии, но и в общественных науках - в экономике, демографии, а также при решении задач моделирования развития самой науки (например, Прайс Д., 1966).

ЛИТЕРАТУРА

Налимов В.В., Мульченко З.М. Наукометрия. - М.: Наука, 1969.

Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. - М.: Высшая школа, 1977.

Прайс Д. Малая наука, большая наука./Перевод с англ. - В кн.: Наука о науке. - М.: Прогресс, 1966, с. 281-384.

Тулдава Ю. Развитие лексики эстонского языка по данным словарей XVII - XX вв. - В кн.: Вопросы общего и сопоставительного языкознания. Тарту, 1984 (Учен. зап. Тартуск. ун-та, вып. 684).

Янч Э. Прогнозирование научно-технического прогресса./ Перев. с англ. Изд. 2-е. - М.: Прогресс, 1974.

Eesti õigekeelsuse sõnaraamat. - Tartu: Eesti Kirjanduse Selts, I kd. - 1925; II kd. - 1930; III kd. - 1937.

Göseken H. Manuductio ad Linguam Oesthonicam. Anführung zur Öhstnishcen Sprache. Reval, 1660.

Hupel A.W. Ehstnische Sprachlehre für beide Hauptdialekte den revalschen und den dörptschen; nebst einem vollständigen Wörterbuch. Riga und Leipzig, 1780.

Hupel A.W. Ehstnische Sprachlehre für die beyden Hauptdialekte, den revalschen and dörptschen, nebst einem vollständigen ehstnischen Wörterbuch. Zweyte durchgängig verbesserte und vermehrte Auflage. Mitau, 1818.

Saari H. Kirjakeele saatus I. Vaade sajandile. - Keel ja Kirjandus, 1979, nr. 11, lk. 661-670; nr. 12, lk. 712-723.

Wiedemann F.J. Ehstnisch-deutsches Wörterbuch. St. Petersburg, 1869.

Wiedemann F. Ehstnisch-deutsches Wörterbuch. Zweite vermehrte Auflage./Redigirt von Dr. Jacob Hurt. St. Petersburg, 1893.

Wingert F. Eine Verallgemeinerung der logistischen Wachstumsfunktion. - Biometrische Zeitschrift, Bd. 13, 1971, S. 34-72.

Õigekeelsuse sõnaraamat./Toimetanud E. Nurm, E. Raiet ja M. Kindlam. - Tallinn: Eesti Riiklik Kirjastus, 1960.

Õigekeelsussõnaraamat./Toimetanud R. Kull ja E. Raiet. - Tallinn: Valgus, 1976.

ON MODELLING VOCABULARY GROWTH

Juhan Tuldava

S u m m a r y

On the basis of the representative (the most complete and normative for the time) vocabularies of the Estonian literary language for the period of the 17th-20th centuries the growth of the Estonian vocabulary has been observed (Table 1). It has been stated that the general trends of vocabulary growth may be expressed by the logistic growth function:

$$L_T = \frac{L_n}{1 + ae^{-kT}},$$

where L_T - the volume of the vocabulary at the end of a period T (measuring a century), L_n - the theoretical limit of the growth of vocabulary (asymptote), a and k - the parameters of the function, e - the basis of the natural logarithms. This means a slow start with steady acceleration and "avalanche" growth with subsequent slowing down of the rate of growth (see Figure 1, curve II), corresponding to the initial stage and the periods of formation and stabilization of the literary language. In this case the growth of the vocabulary has been treated not cumulatively but taking into account only the "pure" increase of the vocabulary, i.e. excluding obsolete words. The highly specialized terminology has been excluded and only such terms are counted which occur in the normative vocabularies of common use (orthological and explanatory dictionaries, etc.). As an alternative, the logarithmic function (having no limit of growth) may be used when estimating the vocabulary growth at the stage of deceleration.

СОДЕРЖАНИЕ

SUMMARIES - RESÜMEES

Ученые записки Тартуского государственного университета.
Выпуск 689.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ.
Труды по лингвостатистике.
На русском языке.
Резюме на английском и немецком языках.
Тартуский государственный университет.
ЭССР, г.Тарту, ул.Юликооли, I8.
Ответственный редактор Ю. Тулдава.
Подписано к печати 24.IO.I984.
МВ IO454.
Формат 60x90/I6.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов II,28.
Печатных листов IO,0.
Тираж 550.
Заказ № 967.
Цена I руб. 70 коп.
Типография ТГУ. ЭССР, 202400, г.Тарту, ул.Пялсона, I4.