TARTU RIIKLIKU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

549

KEELESTATISTIKA JA TEKSTI KVANTITATIIVSED SEADUSPÄRASUSED

ЛИНГВОСТАТИСТИКА И КВАНТИТАТИВНЫЕ ЗАКОНОМЕРНОСТИ ТЕКСТА

Töid keelestatistika alalt

VI

Труды по лингвостатистике

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 549 ВЫПУСК ОСНОВАНЫ В 1893.г.

KEELESTATISTIKA JA TEKSTI KVANTITATIIVSED SEADUSPÄRASUSED

ЛИНГВОСТАТИСТИКА И КВАНТИТАТИВНЫЕ ЗАКОНОМЕРНОСТИ ТЕКСТА

Töid keelestatistika alalt

VI

Труды по лингвостатистике

ТАРТУ 1980

О СЕРИЙНЫХ ДВУЯЗЫЧНЫХ УЧЕБНЫХ ЧАСТОТНЫХ СЛОВАРЯХ

П.М. Алексеев, М.Е. Копылова

За последние десять лет в учебной лексикографии стали выделяться двуязычные частотные словари-минимумы. Публикация таких словарей, хотя появляются и отдельные, "разовые издания[1], уже принимает серийный характер[2].

Ниже предлагается рассмотреть две продолжающиеся серии, которые имеют вполне широкое распространение благодаря сравнительно большим тиражам. Одна из них издается в ГДР, другая в нашей стране.

Профессор Лейпцигского университета им. Карла Маркса Л.Хоффманн, хорошо известный своими работами в областях методики преподавания иностранных языков, лексикологии, лингвостатистики, исследований речевой коммуникации, руководит составлением иноязычно-немецких частотных словарей-минимумов по различным отраслям науки и техники. Опубликованные под его редакцией словари предназначены для чтения русских, английских и французских текстов немецкими студентами, специализирующимися в математике, физике, химии, медицине, ветеринарии и животноводстве, строительной инженерии и педагогике.

Шесть из них изданы однотипными выпусками в типографском оформлении; каждый включает в себя русско-немецкий, англо-немецкий и французско-немецкий минимумы[3]. Два другие - русско-немецкий и англо-немецкий (педагогика) отпечатаны на ротапринте.

Структурно все они идентичны. Открывается очередной словарь изложением сущности статистического подхода к отбору лексики, затем следуют пояснения и рекомендации адресату. Отобранные в каждый из минимумов около 1,2 тыс. слов (сюда входят также несколько словосочетаний как отдельные единицы) представлены двумя списками: по убыванию частот - без переводов - и в алфавитном порядке - в сопровождении немецких эквивалентов.

Перед словами в частотном списке указаны ранги (порядковые номера) частот, а не самих слов, в отличие от существующей в статистической лексикографии практики[5]. Общее число слов, попавших в публикацию, читатель может узнать поэтому лишь из предисловия, а количество их с той или иной частотой

от начала списка - только проделав необходимые подсчеты вместо составителя. При словах указаны также относительные частоты (десятичные дроби, полученные делением абсолютных частот на общую длину обследованных текстов) с точностью до шести знаков после запятой. Такая скрупулезность представляется излишней, поскольку массового адресата словаря в лучшем случае могут заинтересовать наиболее общие и наглядные характеристики, а именно ранги самих слов и исходные, абсолютные частоты. Очевидно, что сведения об употребительности, например, двух слов, воспринимаются легче по их абсолютным частотам 10 и 15, чем по относительным 0,000286 и 0,000429.

Составители не сообщают объемов своих словарей за пределами попавших в минимумы единиц. Информация о суммарной длине текстов для каждого словаря, равной 35 тыс. словоупотреблений в каждом случае, в словарях отсутствует, и ее можно найти лишь в одной из статей Л.Хоффманна[6].

Однако сочетающиеся таким образом недостаточные и избыточные статистические характеристики едва ли помешают использовать словарь в тех применениях, на которые он рассчитан. Они вызывают известную неудовлетворенность у более искушенного в количественной лингвистике и статистической лексикографии читателя. Частотный список помогает преподавателю в дозировке вводимой лексики и в оценке учебных текстов и сопровождающих их лексических минимумов. Обучаемый может самостоятельно контролировать свое владение наиболее употребительной лексикой.

В алфавитно-частотной части каждого словаря иноязычное слово сопровождается указанием на ранг его частоты и на относительную частоту. Далее следуют немецкие переводы входных слов. В некоторых "параллельных" случаях обнаруживаются различия, вполне естественные, если учесть, что в составлении разных словарей участвовали и разные исполнители. Для иллюстрации приведем пример трактовки разными словарями английского слова **average**:

математика	123/0,000086	Durchschnitt, Mittel, Durchschnitts-
физика	91/0,000685	durchschnittlich, Durchschnitts-, mittlerer 0,000371; Durchschnitt, Mittelwert 0,000285; den Durchschnittwert nehmen von, mitteln; durch-

		schnittlich betragen 0,000029
химия	101/0,000143	Durchschnitt, durchschnittlich
медицина	82/0,000514	Durchschnitt 0,000285; durchschnittlich, Durchschnitts- 0,000229
животноводство и ветеринария	62/0,000942	Durchschnitt;durchschnittlich; Durchschnitts-; durchschnittlich betragen
строительное дело	97/0,000400	Durchschnitt, Mittel(wert) 0,000200; durchschnittlich, Mittel- 0,000200
педагогика	94/0,000229	Durchschnitt

Различия, как видно из примера, проявляются в различной качественной и количественной регистрации значений, в порядке представления речи правой, переводной половиной словарной статьи.

В дальнейшем составителям следует, видимо, придерживаться более унифицированных процедур анализа текстового материала и презентации данных в словарях. Это лишь повысит их информативность и расширит круг применений и "потребителей".

Некоторые изменения принципов анализа текста уже заметны на протяжении серии. Если в шести выпусках серии отсутствуют английские и французские артикли с их частотами (и это сразу бросается в глаза), то в последнем (животноводство и ветеринария) приведены английские the и a.

Русские минимумы заключаются списками наиболее употребительных суффиксов и префиксов (без частот), английские и французские - списками суффиксов.

Коллектив проф. Л.Хоффманна делает существенный вклад в учебную двуязычную лексикографию. Один из авторов настоящего обзора в течение ряда лет использует английские минимумы по математике, физике и химии, содержащиеся в этой серии, в работе с будущими учителями для школ с преподаванием некоторых предметов на английском языке, которых готовит ЛГПИ им. А.И. Герцена. Словари серии Л.Хоффманна находят определенный спрос в нашей стране. Представленные в них сведения являются важными и интересными и в лингвистическом плане. Данные группы Л.Хоффманна используются советскими лингвистами, инте-

ресующимися типологией русского, английского, французского и немецкого языков, типологией функциональных стилей и частных терминологических систем. Надо надеяться, что круг охватываемых серий подъязыков будет значительно расширен.

Оптимизацию обучения иностранным языкам в отечественной средней и высшей школе (общей и специальной) ставят перед собой в качестве главной цели составители серии двуязычных ЧС-минимумов, выпускаемой Воениздатом (Военным издательством МО СССР).

Словари этой серии отличаются один от других в несколько большей степени, чем вышеописанные, и это объясняется разными причинами.

Во-первых, составители (все они участвуют в системно-вероятностных исследованиях текста, которые осуществляются группой "Статистика речи", руководимой проф. Р.Г. Пиотровским из ЛГПИ им. А.И. Герцена) с каждым выпуском ищут наиболее удобные для адресата и издателя форму представления материала. При этом отводимый на публикацию объем в печатных листах играет не последнюю роль в выборе того или иного варианта. Во-вторых, учебная цель, являющаяся главной при подготовке словаря к публикации, входит, тем не менее, в целый комплекс задач, которые составитель имеет перед собой при отборе текстов, их анализе, идентификации лингвистических единиц, при использовании "ручных" или машинных операций, при оформлении результатов анализа текста. В-третьих, если один составитель ограничивается таким анализом на уровне словоформ (и, далее, слов-лексем), то другой включает в рассмотрение также и уровень словосочетаний. При регистрации текстовых словосочетаний используются разные подходы. В одних случаях это неформальное выделение сегментов текста, которые могут восприниматься его читателем как всякого рода клише, штампы, а не только собственно фразеологизмы в обычном понимании. В других случаях словосочетание определяется позиционно как левая или правая (в зависимости от языка) дистрибуция опорного слова (или словоформы). Облегчаемая таким путем автоматизируемая, машинная процедура членения текста приводит к довольно большому частотному перечню микросегментов, из которого в результате дальнейшей ручной обработки извлекается соответствующий ЧС словосочетаний, более отвечающий требованиям к учебному словарю[7].

Такие соображения и учитываются составителями публикуе-

мой Воениздатом серии двуязычных ЧС-минимумов.

Словари этой серии базируются на выборках объемом от 200 тыс. словоупотреблений и выше. Опубликованы 8 словарей[8], еще один[9] ожидается к выходу в 1980 г. Три из них составлены с помощью ЭВМ. Для читателя, работающего в школе, могут представить интерес прежде всего два из них и, в некоторой степени, еще четыре, если использовать их на уроках по научно-техническому переводу. Два другие рассчитаны на слушателей военных учебных заведений и на гражданских учащихся для курсов военного перевода.

В словарях этой серии представлен первый отечественный опыт массового издания однотипных ЧС-минимумов, поэтому в ходе подготовки очередной публикации и имеют место определенные изменения в оформлении словарей.

В открывающем серию англо-русском ЧС-минимуме по электронике однословные термины и терминологические сочетания образуют два алфавитночастотных списка (раздельное представление слов и словосочетаний в собственно частотных списках совершенно естественно). В дальнейших выпусках словосочетания вводятся в соответствующие статьи опорных слов, поскольку такой способ оказался более удобным для адресата словарей. Приложение к минимуму содержит обзор существовавших к моменту опубликования ЧС английского языка. В словнике этого минимума представлены 2,8 тыс. слов-терминов с частотами не менее 2 (из общего их числа 4,1 тыс., выделенного из 7,2 тыс. слов-лексем, к которым сведены 10,5 тыс. разных словоформ, зарегистрированных в выборке объемом 200 тыс. словоупотреблений). ЧС-минимум содержит также 2,4 тыс. терминосочетаний с частотами не менее 2 (из общего числа 9 тыс. сочетаний в выборке такого же, как и для слов, объема). В алфавитно-частотных списках каждая единица сопровождается указанием на частоту и русским переводом. Кроме того, при каждом однословном термине указано количество текстов (из 200), в которых он встретился, а при каждом терминосочетании - количество разделов электроники (из 5), в которых оно зарегистрировано. В специальных таблицах приводятся ранги и частоты всех единиц минимума. Тираж словаря 20 тыс. экземпляров.

Англо-русский ЧС-минимум газетной лексики более соответствует традиционному оформлению словника. Словосочетания в нем приведены в статьях гнездовых слов. В целях статистической корректности, однако, каждое сочетание приводится лишь

по одному разу: например:, a good deal of введено в статью deal и не повторяется в статье good. Помещенные в минимуме 3,8 тыс. слов с частотами не менее 4 (из общего числа разных слов, равного I2,6 тыс, к которым сведены 23,6 тыс. разных словоформ) были зарегистрированы в выборке объемом 200 тыс. словоупотреблений из I8 крупнейших газет Англии и США. Около 2 тыс. газетных статей отражают общественно-политические события, научную, культурную и спортивную жизнь, экономику, светскую и полицейскую хронику. В минимум включены I,9 тыс. словосочетаний с частотами не менее 2 (на частоту I приходятся 7,4 тыс. сочетаний), выписанных из текстов такого же объема, но уже только из одной газеты - Morning Star (20 номеров). Составители предполагали, что на уровне слов различия между газетами разных направлений в общем обнаруживаются не так сильно[I0], и поэтому такой минимум можно с оговорками рекомендовать для чтения любой газеты. Что касается учебного минимума словосочетаний, то он должен был базироваться на тех текстах, с которыми приходится иметь дело в нашей стране основной массе обучаемых иностранному языку. Анализ был ограничен одной английской газетой, отражающей тот стандартный вариант английского языка, которому учат в советской школе. Таким образом, слова с их частотами извлекались из одного корпуса газетных текстов, а словосочетания - из другого. Англо-русский алфавитно-частотный минимум содержит гнездовые английские слова, словосочетания, частоты этих единиц и русские переводы с учетом описываемых газетными текстами ситуаций. Частотный список слов упорядочен по убыванию частот с указанием самих частот слов и их рангов. Тираж словаря 34 тыс. экземпляров. Готовится второе, расширенное издание этого словаря с добавлением выборки в I00 тыс. словоупотреблений для слов и словосочетаний.

Немецко-русский ЧС-минимум газетной лексики оформлен аналогичным образом. Алфавитно-частотная часть содержит 3 тыс. слов (из общего числа 4,9 тыс. слов, объединяющих 28,4 тыс. разных словоформ) с частотами не менее 2 и 2,I тыс. именных словосочетаний (общее число разных словосочетаний не указано), также с частотами 2 и выше. При единицах минимума приводятся их частоты и русские переводы. Крайне интересны сведения частотного списка слов, где кроме их рангов и частот в выборке 200 тыс. словоупотреблений даются раздельно частоты для двух половин выборки, представленных газетами

ГДР и ФРГ. Эти данные имеют вполне определенный смысл для социолингвистических наблюдений. Аббревиатура DDR в газетах ГДР встретилась 250 раз, а в западногерманских 5 раз. Соответствующие цифры для слов Staat 137 и 48, Arbeit 103 и 3, Volk 104 и 30, Betrieb 130 и 8, Frieden 94 и 6, Genosse 95 и 3, Kampf 64 и 20, Bürger 63 и 11. Словарь выпущен тиражом 20 тыс. экземпляров.

Подобное оформление имеет немецко-русский ЧС-минимум по электронике. В него отобраны 3 тыс. однословных терминов и 137 терминосочетаний с частотами 2 и более из общего числа 6,4 тыс. слов (к которым сведены 20 тыс. словоформ) в текстах объемом 200 тыс. словоупотреблений. Количество одноразовых единиц не указано. Входные единицы алфавитно-частотного списка даются с частотами и русскими переводами. В частотном списке приведены вместе и однословные термины, и терминосочетания, и в этом заметное отличие данного ЧС-минимума от других. Тираж словаря - 15 тыс. экземпляров.

Французско-русский ЧС-минимум по электронике включает в себя только однословные термины. Их в нем 3,4 тыс. с частотами не менее 2; общее число терминов, встретившихся в текстах объемом 200 тыс. словоупотреблений, равно 4,7 тыс. Они выделены из 5,9 тыс. слов (10,3 тыс. разных словоформ). В алфавитно-частотном списке кроме частоты термина и его русского перевода сообщается также количество текстов (из 200), в которых зарегистрирован термин. Имеются статистическая таблица рангов и частот терминов и таблица распределения слов по частям речи в тексте и словаре. Словарь издан тиражом 10 тыс. экземпляров.

На выборке объемом 210 тыс. словоупотреблений основан англо-русский ЧС-минимум по судовождению. Тексты обработаны на ЭВМ. Из 9,4 тыс. разных словоформ извлечен ЧС терминов (общие количества разных слов и разных терминов не сообщаются). В минимум включены 2,7 тыс. однословных терминов с частотами не ниже 4 и 2,7 тыс. терминосочетаний с частотами 2 и более. Единицы алфавитно-частотного списка имеют при себе указание на частоты и русские переводы. В частотном списке однословных терминов приводятся их ранги и частоты. Тираж словаря - 15 тыс. экземпляров.

Англо-русский и русско-английский военные ЧС-минимумы составлены с помощью ЭВМ на материале штабных и боевых документов. Объем выборки для первого равен 200 тыс. словоупот-

реблений; из 6,4 тыс. разных словоформ, то есть 3,9 тыс.разных слов в минимум отобраны 2,9 тыс. терминов с частотами не ниже 2. Алфавитно-частотный список включает в себя также 2,4 тыс. терминологических словосочетаний. Входным единицам этого списка приписаны их частоты и русские переводы. В частотном списке однословных терминов приводятся их ранги и частоты. Словарь издан тиражом 15 тыс. экземпляров.

При составлении второго словаря этой "микросерии" также использована ЭВМ. Объем текстов равен 750 тыс. словоупотреблений. В словарь введены все 3,2 тыс. разных слов выборки, а также 1,8 тыс. терминологических словосочетаний со всеми частотами, включая 1. Словарь, таким образом, является первым из минимумов, в котором представлены полные инвентари использованных в выборке единиц. Алфавитно-частотный список содержит слова и терминологические сочетания с указанием их частот и английских переводов. В частотном словаре слов приводятся их ранги и частоты. Приложены по отдельности частотные таблицы для слов и терминосочетаний. Тираж словаря 20 тыс. экземпляров.

Публикация серийных ЧС-минимумов отражает достаточно четко определившиеся тенденции в учебной лексикографии и статистической лексикографии. Во-первых, это дифференцированный (отраслевой, ситуативный) подход к созданию учебного минимума, поскольку получение универсального, пригодного "на все случаи жизни" минимума - задача недостаточно реалистичная. Во-вторых, это частотный отбор лингвистического материала в минимум как альтернатива стихийной, интуитивной статистике наряду с другими, не более определенными критериями. В-третьих, это четкая организация и планирование работ по отбору и анализу текстов, используемых для ЧС-минимума. В-четвертых, это максимальная унификация самых элементарных, но фундаментальных приемов отбора и анализа текстов - определение состава и объема выборки, идентификация лингвистических единиц, корректная формулировка и использование доступных каждому филологу и преподавателю языка базисных статистических понятий. Наконец,в-пятых, это стандартное оформление учебных ЧС-минимумов, сообщение элементарной, но необходимой цифровой информации о текстах и словаре.

Такие соображения необходимо учитывать при подготовке серийных ЧС-минимумов[II].

Примечания

I. См. первый отечественный двуязычный ЧС-минимум, Бигагаев Р.И., Гукасянц Э.Р., Михайлова Г.Н., Нигматова Ф.Н., Соловьева И.Н., Шарипов Г.Ш. Частотный русско-узбекский словарь-минимум. - Ташкент, 1967, а также: Киссен И.А. Словарь наиболее употребительных слов современного узбекского литературного языка (высокочастотная лексика подъязыка художественной прозы). - Ташкент, 1972 (в этом словаре при узбекских словах указываются их русские переводы); Экономико-статистический словарь (частотный) (сост. В.В. Морозенко). - М., 1974; см. также нечастотный англо-русский словарь-минимум для чтения газетных текстов, который сопровождается ЧС английских словоформ и ЧС английских слов, в раб.: Учебное пособие по переводу английских политических текстов (сост. Н.Н. Маницына и др.). - Владивосток, 1971.

2. Береснев С.Д., Соловьева А.И. Зоотехнический частотный словарь немецкого языка. Свердловск, 1968 (ротапринт); Береснев С.Д., Есаулкова М.Т. Ветеринарно-зоотехнический частотный словарь немецкого языка. Свердловск, 1969 (ротапринт); Шанаурова Г.Ф. Медицинский частотный словарь немецкого языка. - Свердловск, 1969 (ротапринт). Все эти словари дают русские переводы входных немецких слов. Следует упомянуть еще один словарь, который с некоторыми допущениями мог бы использоваться как учебный: Англо-русский частотный словарь по электронике (под ред. И.И. Убина). - М., 1977 (ротапринт).

3. Fachwortschatz Medizin. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1970; Fachwortschatz Physik. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1970; Fachwortschatz Chemie. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1973; Fachwortschatz Mathematik. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1976; Fachwortschatz Bauwesen. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1976; Fachwortschatz Tierproduktion und Veterinärmedizin. Häufigkeitswörterbuch russisch, englisch, französisch. Leipzig, 1978.

4. **Häufigkeitswörterbuch russisch. Pädagogik. Potsdam,1975:, Häufigkeitswörterbuch englisch. Pädagogik. Potsdam, 1976.**

5. Это уже отмечалось в раб.: Алексеев П.М. Семантические частотные словари. - В кн.: Статистика речи и автоматический анализ текста - I972. - Л., I973, а также в рецензии П.М. Алексеева и Н.Н. Яблонской на словари серии Л.Хоффмана в **"Fremdsprachen"** I974, № 4.

6. **L. Hoffmann. Häufigkeitswörterbücher der Subsprachen von Wissenschaft und Technik (einige Bemerkungen über Prinzipien und Methoden ihrer Erarbeitung). - Fachsprachen und Sprachstatistik. Berlin, 1975, S. 39.**

7. Подробнее эти процедуры рассматриваются в раб.: Алексеев П.М. Статистическая лексикография. - Л., I975, с. 38-39; он же: К методике составления переводных словарей-минимумов на основе одноязычных частотных словарей. - В кн.: Методы анализа текстов. - Минск, I975.

8. Частотный англо-русский словарь-минимум по электронике (сост. Алексеев П.М.).-М., I97I; Частотный англо-русский военный словарь-минимум (сост. Л.Л. Нелюбин).-М., I974; Частотный англо-русский словарь-минимум газетной лексики (сост. П.М. Алексеев и Л.А. Турыгина).-М., I974; Частотный французско-русский словарь-минимум по электронике (сост. В.К. Кочеткова).-М., I975; Частотный немецко-русский словарь-минимум газетной лексики (сост. А.С. Ротарь и В.А. Чижаковский).-М., I976; Частотный немецко-русский словарь-минимум по электронике (сост. М.Г. Зореф).-М., I977; Частотный русско-английский военный словарь-минимум (сост. Л.Л. Нелюбин).-М., I978; Частотный англо-русский словарь-минимум по судовождению (сост. К.Ф. Лукьяненков и В.Н. Сергеева). - М., I978.

9. Частотный англо-русский физический словарь-минимум (сост. П.М. Алексеев, М.Е. Каширина, Е.М. Тарасова).В план Воениздата на I98I г. входит также: Частотный англо-русский словарь-минимум по квантовым генераторам (сост. Н.С. Манасян).

IO. Тем не менее, определенная разница в частотах тех или иных слов не может не иметь места, и это хорошо видно на материалах немецко-русского ЧС-минимума газетной лексики.

II. В ЛГПИ им. А.И. Герцена планируется еще одна серия частотных минимумов группы "Статистика речи".В 1980 г., по-видимому, выйдут лексико-терминологические англо-русские минимумы по психологии (около 3 тыс. единиц) и по математике (около 2 тыс. единиц).

ON SERIAL BILINGUAL FREQUENCY DICTIONARIES

P. Alekseyev, M. Kopylova

S u m m a r y

A survey of bilingual frequency dictionaries (minima) meant to serve language teaching purposes present some principal features of two "mass" series of such dictionaries - one by the group headed by Prof. L. Hoffmann (GDR) and the other by the "Language Statistics" group headed by Prof. R. Piotrowski (USSR).

Points subject to criticism are stressed. Special emphasis is made on the need for more unification in representing linguostatistical data in the dictionaries.Some information is given on other, serial and non-serial bilingual frequency dictionaries published and being published in the USSR.

ДИАЛЕКТИКА ТИПОЛОГИИ И АТРИБУЦИИ

Павел Вашак

А. Методология

Проблема, которую мы будем решать, на самом общем уровне определения касается решения взаимного отношения и внутреннего упорядочения двух объектов: элемента X и множества M. Далее дана функция (критерий) K, в рамках которой решаем отношения и упорядочение пары (X, M). Мы исходим из общего определения задачи, чтобы ее показать как общенаучную.

Взаимное отношение элемента X и множества M определяет и два взгляда на задачу: типология, атрибуция. Мы сосредоточимся на их свойствах, далее покажем диалектику их отношения: без типологии нельзя проводить атрибуцию, без атрибуции нельзя проводить типологию. В зависимости от определения можно атрибуцию понимать как типологию и наоборот. Для решения задачи применим системный подход: пусть задан критерий K, на основе которого из генерального множества M_0 получим множество M. Очевидно, что критерий K является сложной функцией и применяется постепенно. Если мы, например, исходим из множества M_0 чешских авторов (M_0 было определено тоже на основе какого-то критерия!), постепенным применением критерия "время", "жанр", "стиль", "идеология" и т.д. мы дойдем до множества M, содержащего, например, авторов категории "чешский романтизм". Ясно, что системообразующий критерий K, определяющий систему "чешский романтизм", является функцией различных переменных: время, жанр, стиль, идеология, литературное направление, лексическое богатство и т.д. Критерий K позволяет провести абстрактное сечение наблюдаемого объекта, а также его исследование на основе различных наук и методологических подходов.

Уже сейчас ясна диалектика отношений между типологией и атрибуцией. Определим сейчас атрибуцию как процесс решения: принадлежит ли или нет элемент X множеству M; типологию - как процесс решения внутреннего упорядочения множества M - все в рамках какой-то функции (критерия).

Применяя критерий K на генеральное множество M_0 с целью определить какое-то множество M_I авторов, например, для изу-

чения их лексического богатства, мы постоянно проводим атрибуцию и типологию.

На основе критерия К находим в множестве R_o всех отношений между элементами M_o множество отношений R между элементами M.

Схематически:

критерий	К	M_o	R_o
↓		↓	↓
уровень исследования		исследов. множ. M	R

Целью системного анализа является выборка (лучше сказать открытие) на множестве всех отношений таких отношений, которые существенны для изучаемой проблемы и, таким способом, тоже для изучаемой системы. Отношения релятивно стабильные, инвариантные против изменений разного рода, называются структурой. Существен прежде всего процесс создания системы, т.е. "стройка системы на исследуемом объекте". При таком определении системы как (M, R) молча предполагается, что из множества всех отношений были избраны отношения релятивно стабильные.

Отношения бывают различного характера: напр., материальные, энергетические, информационные и т.д.; в зависимости от этого имеются также различные типы систем. Для систем общественно-научных (лингвистика, искусствоведение) типичны отношения информационные, связанные с передачей и трансформацией информации.

Помимо множественного определения системы, которая для объективной реальности является гипотетическим конструктом (абстрактной системой), существует понятие реальной системы. Ее базисом является реальное состояние объекта.*

Понятие системы свидетельствует о том, что предметом исследования всех наук являются системы, так как "науки занимаются реальным миром, который является системой систем. Значит научные дисциплины являются теориями специальных систем. (Gudermuth P., Kriesel W., 1973).

Системный подход здесь очень удобен, так как позволяет показать диалектику типологии и атрибуции в лингвистике, искусствоведении и др. науках. Таким способом возможно на

* П.М. Алексеев (1977) различает дедуктивные порождающие процедуры и индуктивное моделирование текста.

абстрактном уровне прогресс в одной науке перевести в другую науку.

Б. Типология

С проблемой типологии на разных уровнях ее определения должна столкнуться каждая научная дисциплина. В результате наблюдения за определенной частью объективной реальности получается множество элементов (данных явлений, измерений и т.д.), внутреннюю функциональную структуру которых мы ищем. Этот процесс состоит, во-первых, из определения функциональной важности элементов, во-вторых, именно из определения их типологии. Таким образом, мы подходим к формулировке гипотезы (или гипотез) о реальности, за которой мы следим. Но это утверждение надо еще уточнить и оба понятия, наблюдение и гипотезу, привести к диалектическому совпадению. Наблюдение объективной реальности происходит под влиянием определенной гипотезы, которую полученные данные об объективной реальности то подтверждают, то опровергают, то изменяют и т.д. Нашей целью является определение, в какой степени исследуемое множество элементов (полученное либо "случайно", либо на основе предварительной гипотезы) создает систему; далее, какие элементы создают подсистему, каковы взаимоотношения между отдельными элементами и т.п. Неважно, называют ли эту проблему - более или менее синонимически - типологией, дискриминацией или дифференциацией.

Другое восприятие типологии исходит из множества М, определенного уже конкретной функцией. Исходя из конкретного характера множества М (тексты, стили, авторы, литературный период, языки и т.д.), мы установим элемент Х и одновременно должны решить вопрос, является или не является он составной частью М. В общем плане восприятия можно эту проблему назвать атрибуцией по отношению к множеству М.

С типологией в собственном смысле слова и с атрибуцией мы встречаемся как в лингвистике и литературоведении, так и в смежных дисциплинах, напр., в информатике (документалистике). Так, при индексировании документа процесс состоит в атрибуции документа по отношению к данному множеству (напр. к десятичной классификации). Подобно тому искусствоведы определяют комплект памятников архитектуры, которые оказались под влиянием определенного стиля, определенного приема декоративности, определенного архитектора; подобно тому поступают

в живописи, в археологии или в истории культуры при определении отдельных культур и т.д.

Математическая (квантитативная) лингвистика, а также и приложения математики в литературоведении решают проблемы типологии и атрибуции прежде всего на основе определения частоты единиц языка разных планов, причем в изучении художественного текста надо еще учитывать структуру текста (см. Tuldava J., 1971).

В рамках лингвистики речь идет о единицах лингвистических (слово, предложение, грамматические категории и т.д.), в рамках "нелингвистики", искусствоведения, о единицах неязыковых планов художественного произведения и его контекстов генезиса и общественной реализации. В теории литературы можно взять идею, тему, сюжет, тип литературного героя, жанр, и т.д. И здесь можно говорить о частоте, категории квантитативности как о регистрации количества наблюдений или оппозиции "присутствие-неприсутствие" явления. Обобщенная категория квантитативности искажает однородность проблемы.

В общем, проблему можно сформулировать следующим образом: пусть дано множество элементов A_1, A_2, ... A_m (напр., авторов, текстов, стилей, языковых и литературных периодов и течений, поколений и т.п.), и множество явлений p_1, p_2, ... p_m. Задача сводится к установлению сходств и различий A_j между собой с точки зрения исследованных явлений p_i, образуют ли они однородную систему, выполняющую определенную функцию, какова внутренняя структура этой системы и т.п.

Проблему типологии (дифференциации) двух и более элементов решает математическая лингвистика в общем плане следующими способами:

1) на основе исследования одного явления, например, длины слова или предложения, частоты слов, частей речи или форм слова, индексов и коэффициентов, выражающих разные аспекты словарного состава (индекс повторения, константа Юла, отношения Гиро, коэффициенты, предложенные Бодером, Буземанном, Фишером и др.) (см., например, Tuldava J., Villup A., 1976). Например, критерий "частота слов" в принципе разделит систему лексического состава четко на три функциональных подсистемы, конечно, неограниченных: слова формальные, слова полнозначные и слова узкотематические.

Среди приведенных явлений нет никаких "простых" объектов, на самом деле все они являются сложными системами. Например,

при исследовании длины предложения надо учитывать построение текста художественного произведения (полоса рассказчика, полоса персонажей, смежные формы между ними), надо различать распределение длины слов так. наз. формальных и полнозначных и т.д. (см. Вашак П., 1974). Тут заметны, с одной стороны, стремление к точному определению исходных понятий в соответствии с их функцией, а, с другой стороны, релятивность элемента и системы;

2) на основе исследования большого числа явлений, которые при анализе исследуемых элементов комбинируются, комплектуются или обрабатываются подходящей математической процедурой (напр., факторным анализом, вратиславской таксономией и т.д.), последнее время по программам автоматической классификации на ЭВМ (classification automatique).*

Найденная типология не является доказательством реальных отношений между изучаемыми, не только объективно сформулированной гипотезой, которую нужно проверить с помощью качественного анализа. Результаты, полученные с помощью математической процедуры, мы должны всегда интерпретировать на "языке" той части объективной реальности, которая подвергалась исследованию и о которой мы сформулировали гипотезу (гипотезы).

Так, при создании частотных словарей "языка вообще" исходят из множества (системы) S реализованных текстов; но реальная система является по отношению к изучаемому "языку вообще" абстрактной системой. Поэтому одним из основных критериев качества частотных словарей (или языка, функционального стиля, автора и т.п.) являются прежде всего качества системы S . С другой стороны, в системе S можно в зависимости от ее общей функции определить подсистемы, которые исполняют определенную, частичную функцию. В теории частотных словарей подсистемам отвечают множества реализованных текстов в зависимости от характера составляемого словаря (напр., подсистемы жанров, подсистемы устной и письменной речи, текстов отдельных авторов при создании словаря данного поколения и т.п.).

Поэтому уже определение системы является значительным шагом вперед по пути к познанию данной конкретной области, к

* См. специальный номер журнала: Information et Sciences Humaines, No. 25, Paris 1975; далее работы: Benzécri J.P., 1973; Lazarsfeld P.F., Henry N.W., 1968; Пиотровский Р.Г., 1975.

определению отдельных элементов, их взаимоотношений, диалектики и противоречий. Следовательно, типология является процессом, раскрывающим связи отдельных элементов исследуемого множества (системы) и суть его функции.

В. Атрибуция

Этим понятием обозначается в лингвистике, текстологии, искусствоведении и т.д. чаще всего установление авторства анонимного (псевдонимного) произведения. Но установление авторства - авторская атрибуция - является особым случаем общей атрибуции, не-авторской. Мы определили атрибуцию как процесс решения, принадлежит ли или нет элемент X к множеству M с точки зрения какой-то функции **F** . Множество M создано и типологически разграничено на основе различных функций (критериев), например, множество авторов, стилей, периодов времени, литературных школ и поколений, идейных направлений и т.д. В этом смысле атрибуция в первой фазе дает типологию множества M. Если мы, например, хотим атрибутировать какой-то текст X, нужно в первой фазе решить, на каком языке он написан, т.е. проводится атрибуция по отношению к множеству существующих и исторических языков. Ясно, что можно иметь в виду текст, записанный в любой системе знаков; нужно еще узнать, совпадает ли язык текста с языком произведения и идет ли речь о переводе или о псевдопереводе. В дальнейшей фазе проводится атрибуция в отношении периода языка, стиля, жанра, идейного направления и т.д. Системообразующий критерий (функция) **F** , определяющий состав множества M, решает также вопрос об уровне атрибуции - авторской или не-авторской. Множество M, по отношению к которому проводится атрибуция, называем сопоставительным множеством.

Атрибуция и типология - это два взгляда на одну и ту же проблему. При атрибуции решается:

$$X \in M; \quad \text{или} \quad X \notin M$$

по отношению к критерию (функции) **F** .

С точки зрения типологии возможно понимать атрибуцию следующим образом:

$$\text{создаем} \quad M' = X \cup M$$

и решаем, является ли множество M' однородным по отношению к **F** , Подобно этому решается и обратная задача (типология - атрибуция).

Основой любой атрибуции является принцип, который мы называем принципом параллельности. На основе различных методов получаем совокупность информаций о X и M. На основе сравнения обеих совокупностей информации, их подобия, тождества и параллельности мы решаем вопрос о принадлежности X к множеству M. Ясно, что этот принцип применяется в связи с методом получения информации, т.е. нельзя его считать "тотальным тождеством", а только тождеством в рамках метода наблюдения. Если мы, например, сосредоточимся на явлениях языковых и стилистических, выраженных статистически, то можно решить вопрос на основе принципа параллельности в смысле математической статистики (текст согласия).

Здесь не идет речь о возможных методах атрибуции (подробнее см. Vašák P., 1980). Напомним только, что возможны следующие методы:

а) метод документальный и фактический, основанный в принципе на фактах литературно-исторических;

б) метод идейно-тематический, основанный на конфронтации системы идей атрибутированного текста и мировоззрения предполагаемого автора (литературные группы, школы и т.д.);

в) метод языковой и стилистический, основанный на понятии индивидуального стиля и конфронтации лингвостилистических данных о тексте X и текстах из множества M.

Можно еще привести четвертый метод, который не является методом в собственном смысле, но более специальным случаем. Особенно в старой литературе не всегда дошел до нас оригинальный текст произведения (архетип) и произведение существует только в форме вторичных копий, списков, печатных текстов и т.д. различного происхождения и текстовой достоверности. Текстолог исходит из множества дошедших до нас текстовых источников $\{X_i\}$ с целью: 1) найти генеалогию элементов множества, т.е. показать схему текстовой зависимости; ее графическим образом является стемма. Здесь проводится типология множества $\{X_i\}$ по отношению к функции "перенос текста произведения X^{n}"*; 2) реконструировать текст произведения.

Нельзя здесь решить комплекс текстологической проблематики. Остается фактом, что понятие авторства здесь меняется,

* Подробнее см. Vašák P., 1980, а также нашу работу: Текстология - система - стемма (Prague Studies in Mathematical Linguistics, 7- в печати) и Froger J., 1968.

так как оно не касается авторства произведения, но любой фазы генезиса его текста, разночтения, варианта и т.д. В рамках системы "генезис текста произведения и его коммуникативного расширения" текстолог реконструирует образ исторического текстового процесса, проводит атрибуцию каждой его фазы. Мы вводим четвертую категорию (г):логику исторического текстового процесса. Между атрибуцией авторской и не-авторской нет никакой разницы; разграничение зависит только от характера функции F , которая определяет состав сопоставительного множества M.

Если мы решим авторство произведения X, это означает,что в состав множества M входят потенциальные авторы X.

Для любого метода атрибуции (а), (б), (в) нужно в текстах предполагаемого автора и во всех контекстах их генезиса найти инвариантное явление. Только после нахождения инварианта, относящегося к предполагаемому автору, можно поставить атрибуционный вопрос: принадлежит ли текст X к найденному инварианту. Этот принцип называем принципом инвариантности*.

В рамках атрибуции языково-стилистической и идейно-тематической приходим к схеме: произведение (текст) X, далее множество m текстов предполагаемого автора однородного функционально-стилистического типа. К такой совокупности текстов мы приходим применением операций не-авторской атрибуции (время, жанр, стиль и т.д.). Обрабатываем эти m текстов, причем метод обработки зависит от способа наблюдения. Пусть мы найдем качественное или количественное явление, которое является "разумно константным". Эту константность - т.е. инвариантность - в квантитативном случае обеспечим статистическим критерием согласия. Когда явление не характерно для изучаемого периода времени, у нас есть ясная характеристика автора. Это значит, что мы нашли m наблюдений какого-то авторского инварианта. В случае, когда качественное (количественное) значение явления для текста X не входит в состав инварианта предполагаемого автора, его авторство текста X опровергается. Чем больше наблюдений инварианта, т.е. чем

* Например, Дж. Юл ввел как принцип инвариантности константу K; близкие значения K для разных текстов он считал доказательством их совместного авторства. К сожалению, он не решил вопроса о зависимости функции K при переходе к другому функциональному стилю. Таким же образом при исследовании длины предложения как потенциальной характеристики авторского стиля он исходил из текстов, однородных с точки зрения построения текста.

больше текстов предполагаемого автора обрабатывалось, тем более вероятным является заключение об авторстве. Но квантитативное доказательство здесь не является единственным, этим способом мы получили тоже доказательство квалитативное. Поскольку мы обработали все тексты предполагаемого автора из периода, когда он мог написать исследуемый текст X, и во всех текстах был обнаружен инвариант, из которого X исключался, текст X данному автору с полным правом не приписывается. Возможно, что найденный инвариант является в реальности нормой или тенденцией идейной, стилистической, языковой, далее привычкой редакционной, эстетической и т.д. В таком случае нельзя говорить об авторском инварианте; но даже такое заключение возможно применить для атрибуции, а именно, для отличения индивидуального от общего. У нас есть m доказательств из всех текстов предполагаемого автора исследуемого периода времени, когда он в какой-то степени подчинялся норме, привычкам, тенденции и т.д. различного характера. В случае, когда текст X не входит в состав этого ряда, X не принадлежит к найденному инварианту. Отсюда логическое заключение - автор такого произведения не написал. Это значит: когда мы найдем для предполагаемого автора в исследуемом периоде какой-то инвариант, то проверка инварианта по отношению к "норме" не нужна в том случае, если текст не имеет такого инварианта. Добавим, что понятие нормы означает только какую-то историческую тенденцию и не является историческим законом.

Если значение исследуемого явления для атрибутированного текста входит в состав инварианта, является ли это доказательством авторства? В таком случае нужно проверить, не является ли найденный "инвариант" в данном периоде времени нормой или общей тенденцией. Если он представляет собой общую тенденцию, тогда он не выражает ничего индивидуального, и его нельзя применять как базис атрибуции. Напротив: если инвариант является действительно индивидуальной характеристикой, в состав которой входили также значения инварианта для атрибутированного текста, - тогда это является серьезным аргументом единого авторства. Всегда представляется более легким отказ от признания авторства (атетез), так как достаточно, чтобы единственный факт говорил против авторства. Напротив, доказательств авторства - т.е. принадлежности X к инварианту - никогда не достаточно. Из этого вытекает, что

нельзя построить иерархию методов атрибуции. Ясно, что процесс нахождения инвариантов идет через изучение индивидуальных стилей (см. монографию: Статистичнӗ параметри..., 1967).

Итак, атрибуция имеет следующие этапы: элемент X, сопоставительное множество M, принцип инвариантности, принцип параллельности. Такая логическая схема верна для атрибуции авторской и не-авторской. Типология и атрибуция является, таким образом, однородной проблемой, касающейся различных вариантов отношений элемента X, множества M, функции (критерия) F.

ЛИТЕРАТУРА

Алексеев П.М. Квантитативная типология текста. Авд. Л., 1977.

Вашак П. Длина слова и длина предложения в текстах одного автора. - В кн.: Вопросы статистической стилистики. Киев, 1974, с. 314-329.

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

Benzécri J. P. L'analyse des données. Paris, 1973.

Froger J. La critique des textes et son automatisation. Paris, 1968.

Gudermuth P., Kriesel W. Kybernetik und Weltanschauung. Leipzig-Jena-Berlin, 1973.

Lazarsfeld P.F., Henry N.W. Latent Structure Analysis. Houghton Mifflin, 1968.

Tuldava J. Sõnapikkus ilukirjanduslikus proosas. - Keel ja Kirjandus. Tallinn, 1971, nr. 10, lk. 583-591.

Tuldava J., Villup A. Sõnaliikide sagedusest ilukirjanduslikus proosa autorikõnes. - Töid keelestatistika alalt, I. Tartu, 1976, lk. 61-106. (TRÜ Toimetised, vihik 377).

Vašák P. Metody určování autorství. Praha, 1980.

DIALECTIQUE DE LA TYPOLOGIE ET DE L'ATTRIBUTION

Pavel Vašák

R é s u m é

L'article traite la relation dialectique entre le processus de la typologie et le processus de l'attribution du texte. Dans la première partie l'auteur part de la notion cybernétique du systeme, présentée comme l'ensemble des éléments et l'ensemble des relations entre ces éléments; on montre le processus de la création d'un système.

Par la typologie on comprend le processus de recherche de l'organisation fonctionelle de l'ensemble M dans le cadre d'une fonction F. On donne les exemples de la typologie en linguistique mathématique. On considère l'attribution comme un processus de décision lorsqu'un élément X appartient ou non à un ensemble M (appelé l'ensemble de comparaison) dans le cadre d'une fonction F. Conformément à la structure concrète de l'ensemble M on distingue l'attribution d'auteur et de non-auteur. La décision si X appartient ou non à l'ensemble M est basée sur deux principes expliqués dans l'article, c. à d. principe d'invariance et principe de similarité.

L'auteur de l'article décrit brièvement le système des méthodes possibles de l'attribution qui sont documentaire, idéologique et thématique, linguistique et stylistique. Plus loin on étudie encore l'attribution du processus textuel, décrit par un stemma. Si on fait la conjonction de l'élément X et l'ensemble M on peut considérer l'attribution comme la typologie de la conjonction; il en est de même à l'envers. Alors dans chaque processus de l'attribution on fait en même temps la typologie et vice versa.

О ЧАСТОТНОМ СЛОВАРЕ РУССКОЙ ОБИХОДНОЙ ПИСЬМЕННОЙ РЕЧИ

А.С. Григорьева

Первый опыт составления частотного словаря лексики, используемой в русской обиходной письменной речи, предпринят на материале личной корреспонденции. При комплектовании выборочного корпуса текста не удалось следовать процедурам, рекомендуемым лингвостатистической практикой. Это объясняется теми же трудностями, с которыми сталкивается исследователь устной речи: далеко не каждый информант-носитель языка согласен подвергнуться наблюдению.

Единственное ограничение на текстовой материал, которое было соблюдено, состоит в том, что авторы всех писем имеют среднее, законченное или незаконченное высшее образование (в основном это студенты филологических специальностей, их родители и друзья). Большинство корреспондентов - женщины. В словаре нетрудно заметить значительный вес форм женского рода и некоторое количество лексических единиц, описывающих реалии из жизни вуза и школы. Однако основной массив словарных единиц относится к повседневным ситуациям, в которых каждый человек может оказаться вне зависимости от его социальной принадлежности. Кроме того, лексика словаря отражает особенности именно эпистолярного функционального стиля. Эта сфера использования языка его рядовыми носителями, кстати, очень плохо, если вообще хоть как-либо описана в современной лингвистике.*

Каждый грамотный носитель языка пишет письма, но очень немногие (в среднем) участвуют в иной, кроме эпистолярной, продуктивной письменной речевой деятельности. В частности, из общего числа носителей русского языка в СССР старше 15 лет лишь 0,1% являются профессиональными литераторами и работниками печати. А это значит, что эпистолярная речь по участию в ней "рядовых" носителей языка, по количеству и общему объему произведений занимает второе место после устной

* Обзор зарубежных работ см.: П.М. Алексеев, А.С.Григорьева, М.Е.Каширина. Статистические исследования лексики писем. - В кн.: Теория языка и инженерная лингвистика. Л., 1973.

продуктивной речевой деятельности. И тем не менее основное внимание лингвистов концентрируется на изучении нормированных, отредактированных письменно-литературных текстов. Таким образом из сферы интересов науки о языке и речи выпадает важнейшая область лингвистического поведения человека, являющаяся как бы промежуточным звеном между повседневной устной и письменно-литературной речью.

Настоящий частотный словарь русской эпистолярной речи мог бы рассматриваться в качестве первого приближения к получению адекватных описаний массового продуктивного использования русского языка в его письменной форме.

Словарь составлен вручную на материале 300 писем 100 корреспондентов. В текстах общей длиной 100 тыс. словоупотреблений встретилось около 15 тыс. разных словоформ. Употребления собственных имен и географических названий вошли в объем выборочного корпуса, но не учтены в представленной здесь словарной статистике. Ниже приводятся около 1 тыс. самых частых словоформ в порядке убывания их частот, начиная от частоты 3230 и кончая частотой 10. Сведения об остальных частотах даются в конце списка. Слева от словоформ указаны их ранги, справа частоты.

1	и	3230	42	все	282
2	в	2688	43	из	280
3	не	2470	44	нет	274
4	я	2308	45	ну	273
5	что (союз,мест.)	1624	46	же	265
6	на	1570	47	хорошо (нар.)	239
7	а	1329	48	теперь	223
8	с	1323	49	да (част.)	216
9	у	949	50	ведь	212
10	ты	892	51-52	его, или	208
11	но	856	53	даже	204
12	как (союз)	839	54	ничего	203
13	меня	780	55	время	200
14	мне	765	56	когда	199
15	очень	752	57	вас	197
16	это	736	58	для	196
17	все	710	59	там	195
18	так	681	60	может	193
19	по	573	61	чтобы	190

20	тебе	544	62	знаю	I89
2I	за	499	63	ли	I88
22	то (союз,мест.)	484	64-65	пока, тоже	I87
23	тебя	479	66	день	I84
24	еще	476	67	конечно	I80
25	уже	462	68	сегодня	I77
26	мы	450	69	раз	I73
27	к	440	70	вы	I72
28	о	393	7I-73	вам, здесь, нам	I7I
29	бы	389	74	есть	I67
30	если	387	75	буду	I62
3I	от	38I	76	письма	I53
32	сейчас	379	77	быть	I5I
33-34	было, он	364	78	этом	I48
35	вот	352	79	привет	I47
36	только	35I	80-8I	был, просто	I46
37	нас	345	82-83	наверное, они	I44
38	да	333	84	была	I42
39	она	330	85	много	I39
40	письмо	327	86	нужно	I36
4I	будет	292	87-89	до́ма, надо, пото́м	I32
90	ни	I3I	I42-I43	думаю, уж	80
9I	себя	I28	I44	пожалуйста	79
92	всего	I27	I45-I46	одна, плохо	76
93	писать	I25	I47-I50	были, мама, тогда	74
94	знаешь	I23		что-то	
95	можно	I22	I5I	скоро	73
96	со	I2I	I52-I54	всех,вчера,сказать	72
97	как (мест.)	II9	I55	лучше (нар.)	7I
98-99	пиши, совсем	II8	I56	получил	70
I00-I0I	после, себе	II6	I57-I59	мой, об, папа	69
I02-I03	без, их	II5	I60-I6I	жизнь, сам	68
I04	больше (нар.)	II4	I62	такая	67
I05	вообще	II2	I63-I64	каждый, нее	66
I06	сразу	I04	I65-I66	желаю, мои	65
I07-I09	могу, хотя	I03	I67-I72	ним, почему, работу, свидания, такой, трудно	64
II0	здравствуй	I0I	I73-I74	делать, которые	63
III	получила	I00	I75-I79	деньги, жду, поэтому	62
II2	моя	99		своей, хоть	

II3-II5	где, пишу спасибо	98	I80-I86	вместе, все-таки, ней, них, обязательно, работы, сам	6I
II6-II7	дела, хочу	97			
II8	этого	96	I87-I90	большое, всем, напиши, целую	60
II9	год	95			
I20-I2I	написать	94	I9I-I92	дня, сделать	59
	потому		I93-I96	года, него, при, человек	58
I22-I24	во, кажется, немного	92	I97-200	будем, ладно, хотелось, через	57
I25-I26	почти, хочется	90	20I-204	долго, пришлось, такие, том	56
I27-I28	жизни, тобой	89			
I29-I3I	да (союз), домой, общем	88	205-207	написали, особенно, тут	55
I32	этот	87	208-2I0	уважаемый, часто, эти	54
I33-I34	времени, один	84	2II-2I2	кто, погода	53
I35-I36	два, твое	83	2I3-2I5	большой, ко, мое	52
I37-I38	опять, того	82	2I6-2I8	наш, новый, под	5I
I39-I4I	ему, правда (вв.сл.), твой	8I	2I9-222	видела, лет, несколько, тем	50
223-224	говорит, перед				49
225	свою				48
226-228	идет, им, этим				47
229-236	вроде, году, дорогая, интересно, как-то, мало, надеюсь, часов				46
237-240	завтра, который, мной, пишешь				45
24I-245	более, город, извини, милая, эту				44
246-253	жить, какой, напишу, никогда, первый, приехать, пусть, такое				43
254-259	будут, весь, насчет, наши, раньше, туда				42
260-268	вечером, главное, люблю, наших, нормально, оно, писала, работе, т.к.				4I
269-277	друг, крепко, кроме, лишь, обо, поздравляю, придется, собой, три				40
278-284	будешь, жаль, иногда, летом, месяца, пишет, правда (сущ.)				39
285-289	вечер, кино, сказала, сколько, также				38
290-303	бывает, говорят, давно, дней, думать, можешь, нельзя, понимаешь, своих, снова, совершенно, твоего, уважением				37
304-3II	адрес, месяц, нашей, пришла, рада, слова, твои, хожу				36

312-319	знать, неделю, одно, ответ, приходится, свои, свой, хорошего	35
320-326	ехать, место, получили, раза, решила, слов, тот	34
327-333	воскресенье, говорить, живу, маме, отпуск, самое чего	33
334-339	видишь, здоровья, знает, те, утром, что-нибудь	32
340-351	лето, новости, пор, приехала, работа, равно, рождения, свое, следующий, стало, таких, хотела	31
352-367	больше (прил.), ваше, второй, вышло, дни, живет, институт, институте, конце, настроение, наша, недавно, последний, прошу, твоей, этой	30
368-385	всей, возможно, возможность, ждать, живет, из-за, какие, кстати, могла, осталось, по-старому, рублей, самого, своего, случилось, ходили, хотел, числа	29
386-400	всю, где-то, дома, которая, нашего, одной, поняла, приветом, работать, случае, стоит, тетя, учиться, ходила, экзамены	28
401-416	годом, девочки, другой, здравствуйте, куда, марта, недели, некогда, передавай, писал, почему-то, прямо, рядом, субботу, хорошая, хороший	27
417-439	вами, вопросы, дали, должны, заниматься, здоровье, мог, немножко, нового, новым, письме, приехал, работает, ребята, слишком, столько, т.д., т.е., теперь, увидеть, утра, хочешь, часа	26
440-461	ваши, вечера, всем, высылаю, имею, легко, моей, написал, никак, никаких, одну, получилось, пошли, прошло, стала, страшно, счастья, таким, ужасно, успехов, черт, школе	25
462-482	вернее, видеть, дальше, две, днем, довольно, достать, между, меньше, над, нем, основном, писем, пишите, сделала, сначала, спать, сюда, тепло, улице, эти.	24
483-507	вдруг, денег, доченька, какая, которое, море, мою, никто, ночью, нравится, ответа, ответить, парень, получается, получить, приедешь, работаю, сентября, сижу, сказал, слово, уехал, учебе, ходить, хочет	23
508-531	двадцать, добрый, должен, здорово, курс, моего, мысли, пальто, понравилось, последнее, постараться, своим, сдала, смогу, смотри, снег, стал, такого, твоих, тому, холодно, читать, читаю, языка	22
532-552	быстро, говорила, июня, кем, кончаю, который, купить, отец, отношении, по-прежнему, пошел, приеду, скучно, таком, твою, учеба, фото, хуже, целый, экзамен, этому	21
553-570	будь, весна, забыла, знала, значит, института, каникулы, конференции, любовь, найти, наконец, порядке, праздником, сих, смотреть, считаю, чувствую, ясно	20

571-603 августа, боюсь, вашей, вероятно, вещи, впечатление, встречи, говорил, городе, днях, занимаюсь, зато, звонила, июля, какой-то, квартире, легче, лучше, молчание, наконец-то, одного, поговорить, понять, помню, пришел, следует, сожалению, твоя, узнать, февраля, фильм, хватает, января I9

604-628 высылать, говоря, должна, друга, живут, занятия, затем, извините, иначе, конца, которую, назад, около, приедет, работают, рад, смысле, собираюсь, уехала, хорошее, хорошие, человека, чему, чтоб, этих I8

629-662 будто, видимо, встретила, дорогой, думала, душе, ибо, имеет, именно, какие-то, мая, минут, моих, напишите, например, неужели, никуда, ними, отлично, отношения, передать, подробно, пожалуй, по-моему, посмотреть, пошла, праздник, праздники, самый, сдать, скорее, ходить, часть, чего-то I7

663-705 брать, ваша, вашего, годы, двух, думаешь, думал, жалко, живешь, заводе, зачем, идут, каким, кое-что, лежит, людей, люди, мамой, наилучшего, никого, нужна, общежитии, одним, окончательно, остается, пару, поздравление, понимаю, почту, представляешь, про, радость, ребят, родители, самом, своем, сказали, смогла, театр, точно, тяжело, час, чуть I6

706-748 апреля, будете, видел, возможности, всей, встретимся, голова, дал, далее, далеко, дать, ждала, ждем, забыл, идти, иметь, квартиру, которым, месте, начал, нашем, нечего, ноября, оба, по-видимому, правильно, предствить, прислала, руки, сдавать, собирается, сообщить, соскучилась, стоит, счет, тех, ужас, учебу, хватит, хотим, честное, число, экзаменов I5

749-790 английский, благодарю, болеет, вашу, весело, взять, во-первых, всему, выслать, гости, дети, дождь, других, зима, каждого, книги, которых, мамы, мать, могут, нами, напишешь, невозможно, обещал, обещают, октября, очевидно, передай, повезло, посылку, прелесть, просила, сессию, сидит, сильно, ту, удовольствием, узнала, учится I4

791-849 бог, большая, вашим, вопросу, вперед, впереди, вся, говорю, города, группы, глаза, дала, девчонки, действительно, делах, другие, живете, зачет, знал, зовут, изменений, каждую, как-нибудь, конец, короче, куплю, курсе, курсы, любви, мороз, обещал, однако, одни, отвечаю, открытку, отпуска, первое, песни, подумать, пожелать, поздравления, послала, представляю, прежде, преподаватель, придет, просьба, прошел, пятница, рано, самой, сессия, смешно, смотрели, старый, статей, студенты, части I3

850-911 армии, бабушка, больницу, больших, везет, видно, возьми, вспоминать, вышла, голову, дай, девочка, дороге, другое, души, замуж, знакомых, истории, каждой, каком, кого, которого, котором, купила, купили, мнение, многое, молодец, нашли, некоторое, неплохо, новая, ночь, обратно, I2

	одному, отдохнуть, первая, получать, понравился, попасть, последние, посылаю, права, пришлю, приятно, проходит, прошлом, пять, самое, свете, сделали, словом, солнце, считать, та, такую, той, трех, удалось, хотели, цедую, школу.	
912-990	береги, билет, верится, вещь, вид, вижу, вообще-то, вполне, встречать, второго, где-нибудь, декабря, делается, должно, дом, дорогие, дорогу, другом, желаем, интересует, итак, каждому, карточки, комнате, личной, маленькая, месяцев, можем, напр., нашу, необходимо, новом, новые, общежитие, одном, остался, папе, первых, подумала, поехать, поздравить, пойду, понял, положение, получу, помочь, посмотрели, приезжай, приезжла, приехали, примерно, послать, причем, провела, прошли, просить, прости, работала, работаю, рады, разговор, редко, руб., сами, сделал, сердце, серьезно, смотрели, стол, столь, странно, счастье, текстов, телеграмму, тему, ума, честно, школы, экзамена	II
991-1079	августе, адреса, благодарна, благополучно, болит, виду, во-вторых, вода, встретить, встречала, всякие, говоришь, готовиться, девушка, делаю, деле, достаточно, другим, еду, ездили, занята, знаем, имя, катер, книг, книгу, кончать, купит, лекции, лес, литература, маленький, менее, мере, места, мечтаю, надеяться, народ, настоящее, наступающим, начала, начать, нетерпением, нигде, никакого, объяснить, опишу, остались, остальное, отдыхаю, отъезда, первые, письму, пить, план, побывать, позже, попал, пора, привык, приходит, пришлю, простите, речи, самые, свободное, своим, связи, сдал, служба, случай, сообщите, список, спокойно, стали, сто, сходить, съездить, твоим, уйти, университет, успела, учусь, ходим, чаще, чем-то, четыре, чувство, язык	I0

С частотой 9 зарегистрированы 129 словоформ, с частотой 8 - 135, с частотой 7 - 210, с частотой 6 - 281, с частотой 5 - 359, с частотой 4 - 550, с частотой 3 - 1075, с частотой 2 - 2254, с частотой 1 - 9071 словоформа.

ON THE FREQUENCY DICTIONARY OF EVERYDAY WRITTEN RUSSIAN

Azissa Grigoryeva

S u m m a r y

A list of a thousand words most frequently used in everyday written Russian is given rank-ordered. The text corpus containing 300 personal letters of 100 correspondents, mainly students, their parents and friends, equivalent to some 100,000 running words was processed manually, which yielded a total of 15,000 different word-forms, proper and geographical names excluded. Data on numbers of word-forms with frequencies not covered by the list are given as well.

К ВОПРОСУ ОБ ИСПОЛЬЗОВАНИИ РАСПРЕДЕЛИТЕЛЬНОГО СЛОВАРЯ ДЛЯ РЕШЕНИЯ ЛИНГВОМЕТОДИЧЕСКИХ ЗАДАЧ

Е.А. Калинина

Математическая лингвистика рассматривает закономерности, проявляющиеся в языке, не как абстрактные суждения, а как правила высокой статистической вероятности. Статистический анализ речи с использованием математической лингвистики и теории вероятностей лежит в основе двух методов: во-первых, "статистического" (Пиотровский, 1958) или "вероятностно-статистического" (Головин, 1971), применение которого основывается на самой природе языка, представляющего собой систему, каждой единице которой присущи определенные количественные характеристики, которые в речи представлены частотой, а в языке - вероятностью и выражают объективные свойства языка в его функционировании и развитии (Головин, 1971). Статистические методы в лингвистике (в частности частотные словари) давно широко применяются в лексикографии, при составлении различных методических пособий, при написании учебников и т.д.

Другим методом математической лингвистики является метод структурно-вероятностного анализа. Поскольку "структурный анализ должен быть определен как приложение теории множеств к исследованию языка" (Андреев, 1967), то именно структурно-вероятностный метод устанавливает связь между структурными оппозициями в языке и вероятностными соотношениями в речи. Таким образом, в основе этого метода лежит количественное изучение лингвистических явлений, однако оно находится в тесной связи с качественным исследованием этих явлений.

В то же время структурно-вероятностный анализ тесно связан с социолингвистикой, "разделом языкознания, изучающим социальную дифференциацию, то есть различные его социальные диалекты" (Ахманова, 1966).

В настоящее время центр исследовательской активности в рамках социолингвистики перемещается на изучение подъязыков или малых языковых подсистем.

Подъязык - это прежде всего набор языковых элементов, отношения и пропорции между ними, заданных однородной тематикой. Подъязыки отличаются друг от друга прежде всего веро-

ятностными спектрами своих лексических наборов (Андреев, 1967).

Существование возможности достаточного формального описания отдельных подъязыков, а также наличие "определенного изоморфизма между внутренней организацией языка и внутренним строением речи" (Андреев, 1967) явились предпосылкой для использования структурно-вероятностного анализа при изучении малых языковых подсистем, что, в свою очередь, позволило увидеть общие языковые закономерности в их отдельных проявлениях при изучении текстов узкой социальной тематики. "Каждый подъязык обладает индивидуальной, только ему присущей системой структурно-вероятностных доминант, отличной, как от иных доминантных систем, принадлежащих другим подъязыкам, так и от общеязыковых. Именно эти различия между доминантными подсистемами и определяют собой подъязыковую специфику.

Структурно-вероятностные доминанты, образующие собой доминантную сетку, могут принадлежать к разным уровням языка: морфологическому, синтаксическому, лексическому и др." (Андреева, 1976).

Доминантные подсистемы и их свойства могут стать основой для создания научных грамматик, которые, с одной стороны, учитывают вероятностные характеристики языка, а с другой, мотивированы социолингвистикой.

Результаты исследования морфологического, синтаксического, лексического и др. уровней языка методом структурно-вероятностного анализа в рамках определенного подъязыка могут быть использованы в методике преподавания языков. Сейчас такая работа начата в области исследования лексического уровня языка в группе структурно-вероятностного анализа при Горьковском институте иностранных языков.

Неслучайно первые попытки применить структурно-вероятностный анализ для решения лингвометодических и учебно-методических вопросов начаты именно с лексической системы языка, т.к. подъязыки прежде всего, как указывалось выше, отличаются друг от друга вероятностным спектром своего лексического состава.

Лексический аспект в процессе обучения языку обладает своими специфическими особенностями, своими закономерностями, и их необходимо учитывать при решении учебно-методических задач.

Современная методика характеризуется системно-структур-

ным и системно-функциональным подходом в описании лексической системы языка в процессе обучения, его презентации и усвоении.

"Системно-структурный подход означает рассмотрение слова как элемента целостной лексической системы, находящейся во взаимосвязи с другими ее элементами на основе определенных закономерных отношений, а также многообразных способов связей единиц этой системы, их внутренней организации.

Системно-функциональный подход предполагает рассмотрение внутренних (системных) отношений в языке и изучение особенностей (характера) его функционирования в речи" (Ахманова, 1969).

Такой подход к изучению лексического уровня языка, принятый в современной методике, соответствует тем выводам, к которым приходят ученые, проводящие исследования в области лексики тех или иных подъязыков с применением структурно-вероятностного анализа. "В системе языка на лексическом уровне не существует стройная схема его организации, которая является результатом взаимодействия внутрилингвистических и экстралингвистических факторов, в последнем случае заданных спецификой функционирования языка" (Жигачева, 1976).

Одной из наибольших трудностей при системно-структурном изучении языка, а также в обучении лексическому аспекту языка является то, что любой язык находит конкретное проявление только в речи, в речевой деятельности в виде множества различных текстов.

Другим препятствием на пути к изучению лексики является открытый характер словарного состава, невозможность исчерпывающего исчисления множества лексических единиц. Поскольку количество единиц обучения на лексическом уровне практически бесконечно, - обучение лексике должно быть определенным образом нормированным, т.е. ограниченным и управляемым.

Количество лексических единиц, подлежащих усвоению, и их состав определяются целями и условиями обучения. Ясно, что без продуманной системы ввода словарного материала и специальной работы над ним нельзя достичь желаемого эффекта в обучении языку (Колесникова, 1978).

Как раз структурно-вероятностный метод и предлагает такой способ описания и представления лексической системы языка, который позволяет расчленить на отдельные компоненты эту сложную и необозримую структуру, частично формализовать про-

цедуру ее описания (хоть и не всей системы лексики). Такими компонентами, микросистемами в общей системе языка и являются подъязыки, "малые языковые подсистемы", о которых говорилось выше.

В структурно-вероятностном анализе лексическая система языка изучается с привлечением концепции распределительного словаря (Андреев, 1979)*.

Концепция распределительного словаря базируется на следующем положении: частота появления слова в текстах есть функция двух агрументов. Прежде всего, это количество авторов, употребляющих ту или иную лексику (здесь имеют место взаимодействия между внутриязыковыми системными отношениями и экстралингвистическими условиями функционирования языка), и затем, с какой интенсивностью отдельные авторы употребляют данную лексику (чисто субъективный фактор). Кроме того, каждая лексика имеет свою подъязыковую меру представленности, которая задана экстралингвистическими условиями функционирования языка в обществе.

Поскольку распределительный словарь является структурно-вероятностным, то дистрибуция лексем в нем представлена в том виде, в каком она представлена на узком синхронном срезе. Анализ дистрибуции выполняется при помощи сопоставления рассматриваемого подъязыка с другими подъязыками того же уровня.**

Распределительный словарь представляет собой многоуровневую систему подъязыков, где один (или каждый) из компонентов предыдущего уровня развертывается в группу подъязыков последующего уровня и где таким образом используется основной принцип построения языковой системы - ее иерархичность.

Например, рассмотрим структуру распределительного словаря подъязыка электроники. Он состоит из трех уровней.

Первый уровень - общеязыковый - включает в себя набор из

---

* Между характеристиками распределительного и частотного словарей существует отчетливо выраженная доминантная корреляция. (См. Калинина, 1975). О наличии статистической связи между вариациями показателя частоты и коэффициента распространенности говорится и в статье Ю.В. Ванникова и Л.В. Малова (1969).

** Но исследование однотипного и, притом, ограниченного определенным подъязыком лексического материала структурно-вероятностным и вероятностно-статистическим методами дает не сводимые друг к другу результаты. (См. Калинина, 1976). Поэтому эти словари следует считать словарями двух разных типов.

десяти подъязыков (электроника, математика, физика, химия, экономика, медицина, газетная хроника, драма, проза, стихи). Указанные десять подъязыков общеязыкового уровня не репрезентируют, конечно, языка в целом, но их перечисление вполне надежно отражает ту часть лексики, которая устойчиво инвариантна к произвольной тематике.

Второй уровень распределительного словаря состоит из нескольких подъязыков более низкого ранга. Например, если второй уровень - уровень электроники, то он будет охватывать основные разделы электроники. Всего восемь подъязыков: квантовая электроника, полупроводниковая электроника, электронная и ионная эмиссия, автоматика и телемеханика, вычислительная техника, электронная оптика, газовый разряд и газоразрядные приборы, электронные и ионные приборы.

Третий уровень распределительного словаря, состоящий из подъязыков специального уровня, распадается на более узкие специальности: транзисторы (или полупроводниковые триоды), полупроводниковые диоды, фотоэлектричество, термоэлементы, физические процессы в полупроводниках.

Обычно в распределительном словаре три уровня, но может быть и больше.

Каждая лексическая единица распределительного словаря характеризуется рядом численных показателей. Принадлежность слова к тому или иному подъязыку определялась его средней относительной представленностью в обследованных текстах. При этом основным инструментом структурно-вероятностного анализа является коррелятивная функция, которая является показателем отношения условной вероятности* языкового явления к его независимой вероятности** и выступает в качестве объективной оценочной меры, позволяет определить степень специфичности каждого из рассматриваемых слов.

В зависимости от набора тех или иных численных характеристик, в основе которых лежит степень специфичности слова, а именно ее численный показатель, выраженный коррелятивной функцией, все слова можно расклассифицировать по разным ти-

---

* Условная вероятность представляет собой обобщенную встречаемость лингвистических единиц, связанную с определенными условиями.

** Независимая вероятность представляет собой обобщенную относительную частоту анализируемых лингвистических единиц в произвольном тексте.

пам. Преобладание одного типа специфичности над другими на разных уровнях зависит от тематических особенностей уровней, от их рангов в иерархии подъязыков (Калинина, 1975).

В лексической системе определенного подъязыка имеется постоянная часть, инвариантная к подъязыковой тематике, полупостоянная часть, изменяющая свой состав в зависимости от уровня вообще, а также вариабельная лексика, общая исключительно для какого-нибудь одного уровня, состав которой полностью зависит от состава подъязыков этого уровня (Жигачева, 1976).

Выше уже говорилось, что в группе структурно-вероятностного анализа при Горьковском институте иностранных языков начата работа по использованию структурно-вероятностного анализа в области лексики, в целях обучения языкам.

Некоторые соображения по использованию распределительного словаря в методике преподавания иностранных языков изложены в статье М.И. Ивашкина и Т.Н. Ткаченко "Стратификация лексики для методики преподавания иностранных языков на базе распределительного словаря" (в печати) и представлены в таблице, приводимой ниже.

Состав и структура словаря	Обеспечивает в результате овладения им
I	2
I. Общеупотребительность лексики, политематичность	I. Актуализация словаря по любой изучаемой теме
2. Отражение вероятностных характеристик	2. А. Совершенствование механизма вероятностного прогнозирования
	Б. Регуляция системы "Сокращение поиска"
	В. Повышение предсказуемости следования элементов в речевой последовательности
3. Комплексность	3. Использование в каждом виде речевой деятельности
4. Общее представление о словаре	4. А. Расширение ориентирования основы внутренних умственных действий
	Б. Сокращение числа ошибочных действий
5. Укрупнение оперативной единицы (лексической синтагмы)	5. А. Ускорение операций выбора и комбинирования
	Б. Нормализация темпоральных качеств речи

Других работ по использованию структурно-вероятностного анализа в методике преподавания иностранных языков в опубликованной печати нам не встретилось.

Автор настоящей статьи попытался обобщить отдельные результаты, полученные в исследованиях, проводимых в рамках структурно-вероятностного анализа, с точки зрения возможности их применения при обучении иностранным языкам и наметить некоторые пути использования структурно-вероятностного анализа, и в первую очередь распределительного словаря, для решения лингвометодических и учебно-методических вопросов.

Структура и материал распределительного словаря дают возможность изучать системность лексики, во всем ее многообразии: устанавливать связь внутренней структуры языка с экстралингвистическими факторами: выявлять вариантность и инвариантность языковых элементов и семантические качества лексических единиц.

При изучении лексической системы определенного подъязыка или стиля метод структурно-вероятностного анализа дает также возможность исследовать такие словарные явления языка, как антонимия и синонимия, при этом семантическая характеристика слова выражается точными численными показателями.

Исследования явления синонимии методом структурно-вероятностного анализа, проведенные Ж.Ф. Коноваленко (1975, 1978), Р.Н. Артемовым (1973) и Г.Ф. Куртышко (1975), позволяют сделать вывод, что синонимия представляет собой подсистему лексической системы языка. Каждый подъязык обладает индивидуальной, только ему присущей подсистемой синонимов. Организация синонимических рядов имеет вероятностный характер, она меняется от подъязыка к подъязыку. Функция ведущего синонима может переходить от одного слова к другому при изменении сферы применения синонимов или совсем исчезать. Большинство синонимических рядов не имеет ведущих слов.

Явление синонимии синхронно; среди причин, вызывающих его, особенно очевидна связь с подъязыковой тематикой. Именно поэтому система синонимов имеет очень подвижные границы. Общеязыковых, иррелевантных к подъязыковой тематике синонимов очень мало. Они представляют собой синонимические инварианты языка и являются тем ядром синонимии, которое наиболее заметно в лексической системе языка (Коноваленко, 1978).

На материале распределительного словаря проводилось исследование антонимической лексики с тем, чтобы выявить кри-

терии антонимии и получить объективную меру антонимичности. Структурно-вероятностный анализ антонимических связей слов по частям речи позволил сделать вывод, что в глаголе антонимия представлена наиболее широко (Переверзева, 1979). Об этом же свидетельствуют данные в исследованиях А.И. Пупковой (1975, 1976) в противовес мнению В.М. Завьяловой (1973), которая, пользуясь системно-структурным методом, считает, что антонимичность среди прилагательных более репрезентативна. Можно также утверждать, что данное лексико-семантическое отношение, в основном принадлежащее внутренней структуре языка, в значительной степени имеет подъязыковую природу, характер их парной сопряженности зависит как от сферы употребления, так и от внеязыковых экстралингвистических факторов.

Как уже говорилось выше, весь лексический состав того или иного подъязыка подразделяется при структурно-вероятностном анализе на более мелкие составные подсистемы, а именно: на нейтральную (или общую), полунейтральную, полуспециальную и специфическую лексику. В основе принадлежности слова к определенной подсистеме лежат вероятностные дифференциальные признаки, а при определении зон семантического пространства, т.е. порога специфичности данного слова, используется коррелятивная функция.

Таким образом, место лексических единиц в иерархически организованной системе подъязыков определяется при помощи количественных показателей, причем распределение лексических единиц в этой системе, позволяя выделить общую и специальную лексику (и промежуточные группы слов), дает возможность использовать определенные статистические параметры для их дифференцирования (Артемов, 1973).

При помощи вероятностно-дистрибутивного анализа имеется возможность увидеть внутреннюю схему сложно организованной лексической системы определенного подъязыка.

Особенно важной для использования в методике преподавания языков представляется возможность дифференцировать общую и специальную лексику, отличающую один подъязык от другого. Вероятностным дифференциальным признаком общей лексики при этом является равномерная дистрибуция слов.

Общая лексика первого уровня по данным Э.Г. Жигачевой (1976) составляет 25% общего числа отмеченных слов*, на сле-

* Отмеченными называются слова, прошедшие заданный порог появления на уровне.

дующих уровнях значительно увеличивается количество слов общей лексики и составляет почти половину отмеченных слов третьего уровня.

Соотношение количества слов общей лексики уровня и суммарного числа отмеченных слов

уровень	отмеченные слова	общая лексика уровня	соотношение %
I-ый	I576	370	23
2-ой	I587	605	38
3-ий	I362	636	47

Большую часть общей лексики составляет ядерная, дифференциальным признаком которой является не только равномерная дистрибуция во всех текстовых областях, но и, как правило, высокая абсолютная представленность на уровнях. В составе ядра общей лексики основную группу составляют служебные слова: предлоги, союзы, модальные глаголы, частицы, местоимения, высокое функционирование которых в системе речи связано с отсутствием номинативной функции данной лексики, выражающей общие связи и служащей для оформления высказывания. В словарный состав языка наряду с общеупотребительной лексикой входит и специальная, т.е. лексика, обслуживающая определенные подсистемы языка. Характерной чертой специальной лексики является ее большое тяготение к некоторым подъязыкам и отрицательное к большинству других подъязыков.

Распределительный словарь можно эффективно использовать для формализованного решения вопроса об определении термина и нетермина; при этом структурно-вероятностный анализ предлагает статистические параметры для дифференцирования первого понятия от второго (Лаврова, I976). Специальная лексика делится на характерные слова и терминологическую лексику.

Характерные слова занимают особое место в лексике, имеют более разнообразное распределение, чем термины, и подразделяются на характерно-нейтральные и характерно-эмоциональные. Первые относятся к какому-то определенному подъязыку (подъязыкам), но не придают тексту эмоциональной окраски. Вторые (характерно-эмоциональные) экспрессивно окрашивают текст. Среди них наблюдаются существительные, глаголы, прилагательные, наречия, числительные.

За некоторым количеством общеупотребительных слов может закрепляться новое значение, с которым они переходят в разряд терминов. Если слово приобретает дополнительное терминологическое значение, то выяснить, является ли оно преимущественно термином или общеупотребительной лексикой, можно при наличии численных характеристик распределения семантики данного слова по подъязыкам. В преобладающем большинстве случаев терминологическая лексика тяготеет ко второму и третьему уровням. Одним из требований, предъявляемых к терминам, является их однозначность. Однако это требование не всегда реализуется в терминах.

Самую многочисленную группу в терминологической лексике представляют амбивалентные слова (термины-нетермины), и внутри этой группы существует целый ряд отношений по доминации (Винченко, 1979). Эти лексемы реализуют одно и то же значение как в области, к которой они проявляют тяготение, так и в общем употреблении, либо, соотносясь с разными означаемыми, реализуют разные значения.

Таким образом, структурно-вероятностный анализ показывает, что специфическая лексика наиболее подвержена влиянию экстралингвистических факторов: в одних сферах она используется в специальном значении, в других - в общем, в третьих - в специальном и общем в зависимости от прагматики подъязыка (Жигачева, 1979).

Применение концепции распределительного словаря к исследованию лексики одновременно двух и более языков может дать объективные данные по оценке лексических элементов с точки зрения их функционально-семантической нагрузки и принадлежности к подъязыковой, а также в определении степени двухъязычной эквивалентности. Так, например, нередко применяемые русско-немецкие эквиваленты фактически расходятся с действительным функционированием данных лексем в системе речи (Артемов, 1976).

Поэтому исследование большого текстового массива с помощью структурно-вероятностного анализа может дать дополнительные критерии в лексикографической оценке слов, может помочь в создании двухъязычных учебных словарей, которые будут наиболее адекватно отражать лексикосемантические соответствия разноязычных эквивалентов.

С помощью распределительного словаря можно объективно дифференцировать основные и дополнительные значения много-

язычных слов. Внутри каждого лексического элемента можно установить семантическую иерархию через их вероятностные параметры благодаря вариациям слов по уровням. На ряд слов человеческая деятельность во всех сферах рассматриваемой иерархии не оказывает дизъюнктивного воздействия. Так, Р.Н. Артемов (1973) выделяет слова в составе общей лексики, имеющие универсальное общее значение для тридцати восьми подъязыков в четырехуровневой иерархии (давать, время, еще, который, после и др.).

Другие лексические единицы общей лексики имеют различные показатели семантической нагрузки в своей иерархии, это позволяет специфизировать семантические качества лексем - общее и дополнительные значения, т.е. полисемию, моносемию и т.д.

В связи с тем, что специальная лексика особенно подвержена воздействию экстралингвистических факторов, часто возникает дизъюнкция в семантике этой лексики. Слова приобретают дополнительные значения, что ставит их в разряд многозначной лексики.

Воздействие подъязыковой прагматики "прежде всего сказывается на полисемических словах: как набор значений слова, так и их порядок внутри набора могут претерпевать в условиях конкретного подъязыка весьма существенные изменения" (Андреев, 1975).

Если в некоторых подъязыках слово нейтрально, а в большинстве подъязыков специфично, можно говорить о полисемии.

Такие лексемы обслуживают определенную сферу человеческой деятельности по-разному: в одних подъязыках используется специальное значение, в других - общее, а в третьих подъязыках данного слова избегают. Таким образом специальное значение слова может быть конкретизировано только для определенного уровня и подъязыка. За основное - принимается значение, специфичность которого нейтрализуется на первом уровне.

Многомерность подхода к изучению полисемичных слов на основе ступенчатого членения системы языка на семантические уровни и подъязыки позволяет проследить динамику лексических элементов в системе речи, обнаружить одноуровневые и подъязыковые инварианты у части лексики, объективно определить основное и дополнительное значение слова (Ермакова, 1976).

Поскольку численные характеристики распределения слов яв-

ляются объективными оценочными показателями, они могут быть использованы в лексикографии, в частности при составлении разного типа учебных словарей. Одним из видов таких словников, например, могут служить микрословники общей и специальной лексики для определенного подъязыка (на основе которого возможно построить общий словарь для соответствующей области знания), а также для определенного уровня.

Словники, построенные по принципу убывания величины коррелятивной функции, которая показывает степень специфичности слов, можно использовать при составлении терминологических и отраслевых словарей.

Таким образом в словник, построенный на основе распределительного словаря, будут включены достоверно функционирующие в данных подъязыках лексические единицы как общие, так и специфические.

Особенно существенно для решения лингвометодических задач перейти в процессе преподавания языка от уровня отдельных слов к уровню словосочетаний.

При помощи структурно-вероятностного анализа можно произвести отбор базисных моделей и структур для начального этапа обучения, для овладения подъязыком определенной научной дисциплины.

Подъязыки и функциональные стили противостоят друг другу не только лексически, но и относительно грамматических элементов, удельный вес той или иной лексико-синтаксической глагольной валентности в разных стилях неодинаков, так что статистическая структура одного стиля резко отличается от другого. Благодаря структурно-вероятностному анализу удалось обнаружить устойчивое смещение в пропорциях глагольных валентностей через сдвиги в категориальной мере внутри рассматриваемых подъязыков, о существовании соответственной закономерности.

В основе неравномерного распределения лексико-синтаксических глагольных валентностей по подъязыкам лежат экстралингвистические факторы, связанные с зависимостью от выполняемой подъязыками коммуникативно-общественной функции, от задач общения в соответствующей сфере. Между всеми лексико-семантическими валентностями появляется взаимодействие некоторой группы единиц лексического уровня с некоторыми единицами синтаксического уровня (Морскова, 1975).

Повышение качества преподавания, использование наиболее

эффективных и оптимальных методов в преподавании иностранных языков является одним из важных принципов современного обучения. Использование результатов исследования структурно-вероятностного анализа в решении лингвометодических задач должно способствовать решению этой важной цели обучения. И в первую очередь для учащихся нефилологических вузов.

Изучение лексической системы, синтаксических конструкций и морфологической структуры определенных подъязыков, созданные на структурно-вероятностной основе лексические минимумы, учебные лексикографические пособия, научные грамматики дадут возможность преподавания языка учащимся в нефилологических вузах "как можно раньше приобщить к профессиональной коммуникации, знакомить с ее спецификой" (Костомаров, Митрофанова, 1978).

ЛИТЕРАТУРА

Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л., 1967.

Андреев Н.Д. О распределительных словарях общей лексики языка. - Структурно-вероятностный анализ языка по данным речи. Тезисы докладов. Харьков, 1979.

Андреев Н.Д. Структурно-вероятностная типология отношений между семантикой слова и его грамматическими категориями. - В кн.: Типология грамматических категорий. Мещаниновские чтения. М., 1975.

Андреева Л.Д. Типология и универсалии в системе языковых доминант. - В кн.: Исследование по структурно-вероятностному анализу. Горький, 1976.

Артемов Р.Н. Подъязыковая системность лексики и ее отражение в распределительном словаре. АР. Л., 1973.

Артемов Р.Н. Русско-немецкие терминологические эквиваленты в подъязыке спорта. - В кн.: Исследования по структурно-вероятностному анализу. Горький, 1976.

Ахманова О.С. Словарь лингвистических терминов. - М.: Советская энциклопедия, 1966.

Ахманова О.С. Словарь лингвистических терминов. М., 1969.

Ванников Ю.В., Малов Л.В. Проблемы статистической корреляции показателей частоты и распределенности слова. - В кн.: Проблемы лингвистического анализа. М., 1969.

Виниченко Г.Г. Классификация анализа специальной лексики подъязыка агрономии. - В кн.: Структурно-вероятностный анализ языка по данным речи. Харьков, 1979.

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

Ермакова В.Н. Дистрибуция лексических единиц на двух семантических уровнях (на материале распределительного словаря подъязыка машиностроения немецкого языка). - В кн.:Исследования по структурно-вероятностному анализу. Горький, 1976.

Жигачева Э.Г. К вопросу об общей лексике в системе распределительного словаря. - В кн.: Исследования по структурно-вероятностному анализу. Горький, 1976.

Жигачева Э.Г. Соотношение общей и специальной лексики в распределительном словаре. - В кн.: Структурно-вероятностный анализ языка по данным речи. Харьков, 1979.

Завьялова В.М. Антонимические отношения в сфере однокорневых имен прилагательных современного немецкого языка. АР. М., 1973.

Калинина Е.А. Исследование лексики подъязыка электроники вероятностно-статистическим и структурно-вероятностными методами. АР. Горький, 1975.

Калинина Е.А. О соотношении между частотой слова и его распределительным типом. - В кн.: Исследования по структурно-вероятностному анализу. Горький, 1976.

Колесникова А.В. К вопросу описания лексических единиц в учебных целях. - Русский язык за рубежом, 1978, № 5.

Коноваленко Ж.Ф. Результаты изучения синонимов по данным распределительного словаря. - Проблемы структурно-вероятностного анализа языков. Материалы IУ межреспубликанского семинара. Днепропетровск, 1975.

Коноваленко Ж.Ф. Синонимия в распределительном словаре. АР. Л., 1978.

Костомаров В.Г., Митрофанова О.Д. Учебник русского языка и проблема учета специальности. - Русский язык за рубежом, 1978, № 4.

Куртышко Г.Ф. Структурно-вероятностный анализ полной и частичной терминологической синонимии. АР. Минск, 1975.

Лаврова А.П. Распределительный словарь как структурно-вероятностный словарь. - В кн.: Исследования по структурно-вероятностному анализу. Горький, 1976.

Морскова В.А. Место лексико-синтаксических валентностей в различных подъязыках. - В кн.: Вопросы теории романо-германских языков. Вып. 7. - Днепропетровск: Изд-во ДГУ, 1975.

Пиотровский Р.Г. Некоторые вопросы статистического обследования лексических групп. - В кн.: Вопросы статистики речи, 1958.

Переверзева Л.Ф. Некоторые структурно-вероятностные характеристики антонимии в лексике современного немецкого языка. - В кн.: Структурно-вероятностный анализ языка по данным речи. Харьков, 1979.

Пупкова А.И. Проявление степени антонимичности в лексической иерархии. - В кн.: Вопросы прикладной лингвистики. Вып.6. - Днепропетровск, 1975.

Пупкова А.Н. О некоторых свойствах явления антонимии. - В кн.: Проблемы структурно-вероятностного анализа языков.- Днепропетровск, 1975.

ON THE USE OF DISTRIBUTIONAL DICTIONARIES FOR SOLVING LINGUO-METHODICAL TASKS

Yelizaveta Kalinina

S u m m a r y

Structural-probalistic analysis is one of the methods of mathematical linguistics. The distributional dictionary is an instrument of structural-probabilistic analysis. The conception of the distributional dictionary lies in the fact that the multidimensional observation of the functioning of an object in speech provides us with a possibility of the objective evaluation of linguistic phenomena. The results of the studies of the lexical level of the language by structural-probabilistic analysis within the frame work of the distributional dictionary may be applied for the solution of linguo-methodological and methodological tasks, e. g. for the compilation of lexical minimum vocabularies, lexical studyaids, textbooks on scientific grammar, etc.

ВЫЯВЛЕНИЕ ОПЕРЕЖАЮЩЕЙ РОЛИ ПОЭТИЧЕСКОЙ РЕЧИ В РАЗВИТИИ ЯЗЫКА С ПОМОЩЬЮ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Л.Г. Кишинская, С.Б. Кишинский

Корреляционный анализ, которому в настоящее время в лингвистике уделяют все большее внимание[*], дает возможность получить интересные данные при изучении языковых явлений в диахронии.

Как известно, с помощью этой методики исследуют взаимосвязь двух и более явлений (точнее - тесноту связи между ними). Само по себе это представляет для лингвистики значительный интерес. Однако, изучая историю языка, анализируя развитие взаимосвязанных явлений, нас также интересуют причинно-следственные отношения между этими явлениями, когда одно языковое явление изменяет свои свойства под влиянием другого на некотором отдалении, через какой-то промежуток времени, когда весьма необходимым бывает определить даже величину этого отрезка времени. В этом случае можно воспользоваться методом автокорреляции[**]. Применение автокорреляции при изучении диахронии языка мы проиллюстрируем на примере выявления опережающей роли поэтической речи в развитии языка.

Для эксперимента был взят материал английского языка XIУ -XX вв. (поэзия и проза)[x]. Каждый из двух видов речи на отрезке примерно в одно столетие был представлен одним выдающимся английским автором. Таким образом, в поэзии было получено 6 срезов (XIУ, XУ, XУI, XУП, XУШ и XIX вв.), а в прозе 7 срезов (дополнительно был проанализирован материал XX века, что диктовалось условиями эксперимента). Каждый срез был представлен десятью выборками объемом по 8000 зн. каждая[xx]

---

[*] См., например: Головин Б.Н., 1971, с. 160-166; Пиотровская А.А., Пиотровский Р.Г., 1974; Тулдава Ю., 1976 и 1977. Из зарубежных авторов отметим работы Г. Хердана (Herdan G., 1964; 1966).

[**] Об автокорреляции см.: Политова И.Д.,1972; Громыко Г.Л., 1976, с. 163-169.

[x] Список литературы, использованной в эксперименте, указан в конце статьи.

[xx] В качестве знака учитывались буквы, знаки пунктуации и пробел между словами.

по каждому виду речи. Выборки в свою очередь были сгруппированы в две суммарные по принципу: четные - нечетные. Таким образом, каждый автор был представлен двумя выборками и, следовательно, весь обозреваемый период истории английского языка I2-ю (в поэзии) - I4-ю (в прозе) выборками.

На первом этапе с помощью собственно корреляционного анализа (см. Головин Б.Н., с. I60-I66) мы попытались на материале английского языка обнаружить взаимодействие двух видов речи - поэзии и прозы в использовании различных атрибутивных конструкций.

Выбор атрибутивных словосочетаний был сделан на том основании, что они являются как раз тем участком языковой структуры, в рамках которого идет наиболее интенсивное накопление языкового инвентаря за счет все большего количества лексем, включающихся в продуктивные модели, что приводит к нарушению равновесия системы, а, следовательно, и к структурным изменениям, способным привести к появлению нового качества (см. Ярцева В.Н., I969, с. 60-6I). В этом отношении атрибутивное словосочетание является одним из наиболее чутких инструментов, позволяющих в первую очередь обнаружить наметившиеся в языке сдвиги.

Поскольку причина предшествует следствию и опережает его на некоторый отрезок времени (будем называть это расстояние шагом*), то для того, чтобы определить, какой из двух видов речи - поэзия или проза является "родоначальником" языковых изменений, необходимо изучить, изменится ли теснота связи между ними (т.е. между поэзией и прозой) в изучаемом отношении (т.е. в отношении атрибутивных словосочетаний) при сдвиге анализируемых срезов на некоторый отрезок времени (например, в один - два века).

На первом этапе был проведен корреляционный анализ с целью определить тесноту связи между двумя видами речи в синхронии в отношении функционирования в них препозитивных атрибутивных конструкций с одним зависимым членом, выраженным: а) прилагательным *(Adj.)*; б) притяжательным местоимением *(Pn)*; в) числительным *(Num.)*; г) посессивным существительным *(Ng)*; д) существительным в общем падеже *(N)*; е) причастием I *(PI)*; ж) причастием II *(PII)*.

* В общей статистике употребляют в этом случае термин "лаг". См. Громыко Г. Л., I976, с. I64.

В качестве примера в таблице I приводится расчет коэффициента корреляции (r_i) атрибутивной конструкции с зависимым прилагательным, причем x_i обозначает выборочные частоты исследуемой конструкции в поэзии, а y_i - в прозе. Следует иметь в виду, что на доверительном уровне 90% статистически существенными при десяти выборках являются выборочные коэффициенты корреляции, превышающие величину 0,55; на уровне 95% коэффициент корреляции должен превышать критическое значение 0,66. Расчеты выборочных коэффициентов корреляции (r_i), истинных коэффициентов корреляции (r_o) и коэффициентов детерминации (r^2) производились по методике, принятой Б.Н.Головиным.

Результаты корреляционного анализа даны в таблице 2(графа I). В скобках указано количество коррелируемых рядов.

На втором этапе определялось изменение тесноты связи между поэзией и прозой при сдвиге на один век (шаге в один век). При этом за точку отсчета (фактор, причину) была взята поэтическая речь. Таким образом, в анализе участвовали следующие срезы: в поэзии - XIУ, XУ, XУI, XУП, XУШ, XIX вв; в прозе - XУ, XУI, XУП, XУШ, XIX, XX вв.

Была выдвинута следующая гипотеза. Если теснота связи (величина коэффициента корреляции) будет больше предыдущей, или корреляция из отрицательной станет положительной, то это будет означать, что прозаическая речь в данном отношении (т.е. в отношении функционирования изучаемой модели) теперь ближе к тому состоянию, в котором поэтическая речь находилась примерно век назад. И, следовательно, проза осваивает с некоторым опозданием то, что появилось в поэзии. Результаты корреляционного анализа (автокорреляции) с шагом в один век приведены в таблице 2 (графа 2).

Сравним данные обеих граф. Прежде всего заметим, что статистически существенной в первой графе является величина коэффициента корреляции только конструкции Num + Н (где Н - ядро словосочетания, обычно выраженное существительным). Корреляция поэзии и прозы в отношении конструкций Adj + Н, Pn + + Н, Num + Н, P_I + Н отрицательна, что не говорит об однонаправленности идущих языковых процессов.

Во второй графе, кроме конструкций Pn + Н и P_{II} + Н, все другие имеют статистически существенные величины коэффициентов корреляции. Необходимо заметить также, что во всех случаях корреляция стала положительной, т.е. все процессы в

Таблица I

Корреляционный анализ прозы и поэзии в синхронии на материале английского языка XIУ-XIX вв. Модель *Adj* + Н. (Х – поэзия, У – проза)

Век	X_i	a_i	a_i^2	y_i	$в_i$	$в_i^2$	ав
XIУ	180	-49,5	2450,25	200	+33	529	-1633,5
	187	-42,5	1806,25	191	+24	576	-1020,0
ХУ	110	-119,5	14280,25	155	-12	144	+1434,0
	138	-91,5	8372,25	147	-20	400	+1830,0
ХУI	205	-24,5	600,25	162	-5	25	+122,5
	151	-78,5	6162,25	170	+3	9	-235,5
ХУП	241	+11,5	132,25	93	-74	5476	-851,0
	280	+50,5	2550,25	98	-69	4761	-3484,5
ХУШ	417	+187,5	35156,25	175	+8	64	+1500,0
	400	+170,5	29070,25	177	+10	100	+1705,0
XIX	208	-21,5	462,25	220	+53	2809	-1139,5
	237	+7,5	56,25	216	+49	2401	+367,5

$\bar{x} = 229,5$ $\Sigma a_i^2 = 101099$ $\bar{y} = 167$ $\Sigma в_i^2 = 17294$ $\Sigma ав = -1405$ $r_i = -0,034$

$r_o = 0$

Таблица 2

Результаты корреляционного анализа атрибутивных конструкций с одним зависимым элементом в английском языке

Атрибутивные модели	Данные корреляционного анализа	
	I. Взаимосвязь между поэзией и прозой в синхронии	2. Взаимосвязь прозы и поэзии при сдвиге в один век
А. *Adj* + Н	r_i = -0,034 (I2) r_o = 0	r_i = +0,70 (I0) r_o = +0,28; r^2 = 0,49
Б. *Pn* + Н	r_i = -0,30 (I0) r_o = 0	r_i = +0,I4 (I0) r_o = 0
В. *Num* + Н	r_i = -0,64 (I0) r_o = -0,I8; r^2 = 0,4I	r_i = +0,69 (I0) r_o = +0,28; r^2 = 0,48
Г. *Ng* + Н	r_i = +0,08 (I0) r_o = 0	r_i = +0,77 (I0) r_o = +0,40 r^2 = 0,59
Д. *N* + Н	r_i = +0,48 (I0) r_o = 0	r_i = +0,66 (I0) r_o = +0,I8; r^2 = 0,44
Е. *P I* + Н	r_i = -0,I2 (I2) r_o = 0	r_i = +0,92 (I0) r_o = +0,76; r^2 = 0,85
Ж. *P II* + Н	r_i = +0,57 (I0) r_o = 0	r_i = +0,28 (I0) r_o = 0

обоих стилях теперь идут в одном направлении. О чем это может говорить? Вероятно о том, что в прозаической речи по прошествии века усилились те языковые процессы, которые осваивались поэзией некоторое время тому назад. Об этом свидетельствует не только перемена знака при коэффициенте корреляции (минуса на плюс), но и увеличение абсолютной его величины в том случае, когда в обоих видах речи процессы были однонаправленными и на предыдущем этапе истории английского языка (конструкции N_g + H, N + H).

Однако, видимо, не все языковые явления осваиваются прозой одинаково. Об этом говорит и тот факт, что корреляционный анализ функционирования конструкции P_n + H с шагом в два века показал значительное возрастание коэффициента корреляции до статистически существенной величины +0,82.

Полученные результаты помогают объяснить и понять такое на первый взгляд парадоксальное явление, обнаруженное лингвистами при изучении истории английского языка, как тот факт, что проза Чосера стоит ближе к прозе последующих периодов, в то время, как его поэзия сохраняет архаичные лондонские черты (см. Wyld H. C., 1956, с. 52). Все дело в том, что проза Чосера скорее "поэтична", чем "прозаична". Статистический аппарат не обнаруживает существенных расхождений между его поэзией и прозой в отношении атрибутивных словосочетаний рассматриваемого типа. Следовательно и процессы, идущие в "прозе" Чосера, как бы являются провозвестниками прозаической речи последующих периодов. Конечно, необходимо еще изучить, какие элементы прозы как формирующегося в английском языке вида речи находят свое место в языке Чосера.

Следует сказать, что опережающая роль поэзии в развитии грамматических явлений в английском языке отнюдь не столь очевидна. Так, известный исследователь истории английского языка Уайльд, касаясь употребления окончания *-s* в 3 л. ед.ч. наст. вр. глагола, говорит, что трудно поверить, чтобы форма, которая стала затем единственной во всем языке, пришла в английский язык из поэзии (см. Wyld H. C., 1956, с. 335; Ярцева В.Н., 1969, с. 140-143). И это несмотря на то, что американский вариант английского языка в семнадцатом веке также отдает ей предпочтение в поэзии. Взаимосвязь различных структурных изменений в поэтической и прозаической речи требует, безусловно, дальнейшего изучения.

Нами была проверена и обратная гипотеза: не увеличивается

ли теснота связи, если за исходный пункт (фактор, причину) взять прозаическую речь? В этом случае автокорреляция с шагом в один век дала следующие величины коэффициента корреляции. Для модели *Adj* + Н была получена величина -0,87; для модели *Pn* + Н величина -0,72; для модели *Num* + Н величина +0,25; для *Ng* + Н величина -0,42; для *N* + Н величина -0,08; для *P I* + Н величина -0,57; для *P II* + Н получена +0,08.

Среди этих коэффициентов корреляции только два первых статистически существенны. Почти все имеют знак минус, что подтверждает выдвинутую гипотезу о том, что не поэзия осваивает "достижения" прозы, а наоборот. Коэффициенты корреляции моделей *Num* + Н и *P II* + Н получены со знаком плюс. Однако, сравнивая их с данными таблицы 2, мы замечаем, что они отнюдь не опровергают опережающей роли поэзии в развитии языка - величины их значительно ниже.

Дальше было выдвинуто предположение, что подобная закономерность в развитии речи может иметь место и в других языках, например, в русском. Эта гипотеза была проверена на серии выборок. Материал охватывает отрезок времени примерно в три столетия: XVIII, XIX, XX вв., т.е. период, когда, как отмечают лингвисты, в основном уже сложился национальный литературный русский язык (список использованного материала см. в конце статьи).

Для анализа были отобраны произведения выдающихся русских и советских писателей и поэтов. Всего было сделано 72 выборки, каждая объемом в 4000 знаков. Синхронные срезы производились с шагом в 40-50 лет. Каждый срез был представлен шестью выборками по каждому виду речи (поэзии и прозе), которые были объединены затем в две суммарные по принципу: нечетные - четные. Корреляционным анализом исследовались две модели, представляющие собой атрибутивные словосочетания с одним зависимым компонентом, находящимся в препозиции в контактной связи с ядром: "прилагательное + существительное" и "местоимение + существительное".

Сначала проверялась теснота связи между обоими видами речи (прозой и поэзией) в синхронии*. Анализ проводился на десяти параллельных выборках (по две на каждый срез), на материале со второй половины XVIII в. по вторую половину XX в. включительно. Полученные данные приведены в таблице 4.

---

* В качестве примера в таблице 3 приводятся расчеты коэффициента корреляции конструкции "прилагательное + существительное".

Таблица 3

Корреляционный анализ прозы и поэзии в синхронии на материале русского языка второй пол. XVIII в. - второй пол. XX в. Модель "прилагательное + существительное" (X - поэзия, У - проза)

Век	X_i	a_i	a_i^2	$У_i$	$в_i$	$в_i^2$	ав
XVIII (2)	61	-19,5	280,25	73	+2,2	4,8	-42,9
	65	-15,5	240,25	94	+23,2	538,2	-359,6
XIX (I)	86	+5,5	30,25	83	+12,2	148,8	+67,1
	124	+43,5	1892,25	71	+0,2	0,04	+8,7
XIX (2)	34	-46,5	2162,25	49	-21,8	475,2	+1013,7
	42	-38,5	1482,25	42	-28,8	629,4	+1108,8
XX (I)	138	+57,5	3306,25	68	-2,8	7,8	-161,0
	130	+49,5	2450,25	60	-10,8	116,6	-534,6
XX (2)	46	-34,5	1190,25	92	+21,2	449,4	-731,4
	79	-1,5	2,25	76	+5,2	27,1	-7,8

$\bar{x} = 80,5$ $\Sigma a_i^2 = 13136,6$ $\bar{y} = 70,8$ $\Sigma в_i^2 = 2597,4$

$\Sigma ав = +361$ $r_i = +0,06$ $r_o = 0$

Таблица 4

Результаты корреляционного анализа атрибутивных конструкций с одним зависимым элементом в русском языке

Атрибутивные модели	Данные корреляционного анализа	
	1. Взаимосвязь между поэзией и прозой в синхронии	2. Взаимосвязь прозы и поэзии при сдвиге в один век
А. Прилагательное + существительное	$r_i = +0,06$ (10) $r_o = 0$	$r_i = +0,12$ (10) $r_o = 0$
Б. Местоимение + существительное	$r_i = +0,36$ (10) $r_o = 0$	$r_i = +0,49$ (10) $r_o = 0$

Затем с шагом в полвека на тех же конструкциях был проведен анализ зависимости развития прозы от поэзии. В этом случае были использованы следующие срезы: для поэзии - перв. пол. XVIII в., втор. пол. XVIII в., перв. пол. XIX в., втор. пол. XIX в., перв. пол. XX в.; для прозы - втор. пол. XVIII в., перв. пол. XIX в., втор. пол. XIX в., перв. пол. XX в., втор. пол. XX в. Результаты анализа приведены в таблицы 4.

Наконец, последним этапом была проверка взаимодействия между обоими изучаемыми видами речи, когда за начало отсчета (фактор, причину) принималась прозаическая речь. В этом случае для прозы были взяты срезы с перв. пол. XVIII в. по перв. пол. XX в. включительно, а для поэзии - со втор. пол. XVIII в. по втор. пол. XX в. включительно (с шагом в полвека). Коэффициент корреляции модели "прилагательное + существительное" был при этом равен -0,32, а модели "местоимение + существительное" равен -0,64. Как видим, в обоих случаях коэффициент корреляции отрицательный. Если предположить, что отрицательное значение коэффициента корреляции говорит о разнонаправленности процессов освоения той или иной модели в сравниваемых видах речи (поэзии и проза), то, оставаясь отрицательным, коэффициент корреляции должен бы возрастать по абсолютной величине от века к веку. Такую же картину тогда следовало бы ожидать и при автокорреляции. Однако, как показали результаты эксперимента, ничего подобного не происходит. Таким образом, оснований для нашего предположения нет.

Сравнивая приведенные выше результаты с данными таблицы 4, можно заметить, что в русском языке наблюдается та же тенденция, которая была обнаружена нами в английском языке - поэтическая речь оказывает заметное влияние на развитие прозаической речи и опережает ее в осваивании грамматических конструкций, в совершенствовании языка как инструмента коммуникации. Конечно, скорость освоения прозой "достижений" поэзии в английском и русском языке может быть разной.

Поскольку изученный нами материал представляет языки типологически разные, возникает дополнительно ряд вопросов, требующих изучения и связанных как непосредственно с лингвистикой, так и с процессами развития человеческой психики вообще.

Л И Т Е Р А Т У Р А

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

Громыко Г.Л. Статистика. - М.: Изд-во МГУ, 1976.

Пиотровская А.А., Пиотровский Р.Г. Математические модели диахронии и текстообразования. - В кн.: Статистика речи и автоматический анализ текста. Л., 1974, с. 361-400.

Политова И.Д. Дисперсионный и корреляционный анализ в экономике. М., 1972.

Тулдава Ю. Опыт квантитативного анализа художественного стиля. - В кн.: Studia metrica et poetica, 1. Тарту, 1976, с. 122-141. (Учен. зап. Тартуского ун-та, вып. 396).

Тулдава Ю. К проблеме сопоставления субъективных и объективных характеристик стиля. - В кн.: Studia metrica et poetica, 2. Тарту, 1977, с. 82-93. (Учен. зап. Тартуского ун-та, вып. 420).

Ярцева В.Н. Развитие национального литературного английского языка. М., 1969.

Herdan G. Quantitative Linguistics. London, 1964.

Herdan G. The Advanced Theory of Language as Choice and Chance. Berlin; New York; 1966.

Wyld H.C. A History of Modern Colloquial English. London, 1956.

МАТЕРИАЛ, ИСПОЛЬЗОВАННЫЙ ДЛЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА РУССКОЙ ПОЭЗИИ И ПРОЗЫ

1. "Гистория о российском матросе Василии Кориотском и о прекрасной королеве Ираклии Флоренской земли". - В кн.: Русская литература XVII века. Л., 1970.
2. М.В. Ломоносов. Поэтические произведения. - В кн.: Русская литература XVIII века. Л., 1970.
3. Н.М. Карамзин. Деревня. Бедная Лиза. Наталья, боярская дочь. - В кн.: Русская литература XVIII века. Л., 1970.
4. Г.Р. Державин. Водопад. Избр. стихотворения. М., 1977.
5. А.С. Пушкин. Повести покойного Ивана Петровича Белкина. Иркутск, 1976.
6. А.С. Пушкин. Евгений Онегин. Барнаул, 1973.
7. А.П. Чехов. Избранные произведения. Лениздат, 1968. (Мужики. В овраге).

8. Н.А. Некрасов. Кому на Руси жить хорошо. М., 1975.
9. М. Горький. Детство. М., 1977.
10. Сергей Есенин. Стихотворения и поэмы. Саранск, 1974.
11. К. Паустовский. Черное море. Симферополь, 1973 (Синева. Умолкнувший звук); Разливы рек. М., 1974 (Маша.Шиповник. Таинственный сундук. Сказочник).
12. А. Твардовский. За далью даль. Иркутск, 1977.

МАТЕРИАЛ, ИСПОЛЬЗОВАННЫЙ ДЛЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА АНГЛИЙСКОЙ ПОЭЗИИ И ПРОЗЫ:

1. Jeofrey Chaucer. The Canterbury Tales. The Works of Jeofrey Chaucer. London, 1928.
2. Jeofrey Chaucer. Boece. The Works of Jeofrey Chaucer. London, 1928.
3. Th. Hoccleve. Minor Poems. Hoccleve's Works. London, 1892.
4. Thomas Malory. Le Morte d'Arthur. London, 1889.
5. W. Shakespeare's Historical Plays, Poems and Sonnets. London, 1909.
6. Philip Sidney. The Countess of Pembroke's Arcadia. The complete Works of Sir Philip Sidney. Cambridge, 1926.
7. John Milton. Paradise Lost. London, New York, 1888.
8. John Bunyan. The Pilgrim's Progress. Leipzig, 1855.
9. Al.Pope. The Poems of Alexander Pope,v.III. London,1940.
10. J. Swift. Gulliver's Travells by Jonathan Swift.Moscow, Leningrad, 1935.
11. Lord Byron. Child Harold's Pilgrimage. The Poetical Works of Lord Byron. London, 1894.
12. Walter Scott. The Black Dwarf. A Legend of Montrose. London and New York, 1879.
13. Jh. Galsworthy. The Man of Property. Moscow, 1974.
14. S. Maugham. The Moon and Sixpence. Moscow, 1972.

ON THE OUTSTRIPPING ROLE OF POETICAL SPEECH IN THE DEVELOPMENT OF LANGUAGE

L. Kishinskaya, S. Kishinsky

S u m m a r y

The article describes the process and some results of the study of the history of a language by means of linguostatistical methods. The paper provides data on the frequency of syntactical structures typical of English attributive constructions. The investigation ranges roughly from the 14th to the 20th centuries. The correlation between the two types of speech, poetry and prose, is calculated and discussed.

At the first stage the dependent variable ("y", prose) is considered to be a function of another variable, called the independent variable ("x", poetry). A knowledge of "x" will allow one to make some prediction about the "y". Thus, if the use of the models analysed in poetry is presented by "x" and the same models in prose by "y", then the relationship between the two variables may be approximated by the correlation coefficient as:

$$r_i = \frac{\Sigma ab}{\sqrt{\Sigma a^2 \cdot \Sigma b^2}}$$, where "a" and "b" must be determined from data.

At the second stage we use prose instead of poetry as the independent variable, thus "x" is a response variable and "y" is an independent one.

The main conclusions arrived at are as follows.The increase of the value of the correlation coefficient indicates that the changes in using the models analyzed in poetry take place in prose at a distance of roughly a century. The regularity of poetry leaving prose behind in using these grammatical constructions has been established (it can be called the poetry outstripping law). The conclusions drawn from the analysis of English texts were confirmed by the results obtained from the observation of Russian texts. In English the analysis is carried out on a corpus consisting of 130 samples (8000 signs each). The Russian Text is represented by 72 samples (4000 signs each). Some results of the analysis are given in tables and a list of original sources is presented.

ЧАСТОТНОСТЬ АНГЛИЙСКИХ ГЛАГОЛЬНЫХ ФОРМ В НАУЧНО-ТЕХНИЧЕСКОЙ ЛИТЕРАТУРЕ

Вальмар Коккота

Целью настоящего исследования является определение частотности личных и неличных глагольных форм (ГФ) в англо-американской научно-технической литературе и сопоставление полученных частотностей с данными в литературе, а также с соответствующими данными об учебниках английского языка средних школ и Таллинского политехнического института (ТПИ). Исследование частотности английских ГФ проведено на кафедре языков ТПИ.

В литературе неоднократно приводились данные о частотности английских ГФ в научно-технической литературе. Например, достаточно большая выборка (5845 ГФ) представлена в исследовании А. Корсакова (Корсаков А., 1978, с. 53). Однако эти данные недостаточно репрезентативны из-за того, что они взяты только из одной монографии по электротехнике, изданной в 1939 г. Выборка 1000 ГФ, по-видимому, представляет собой нижний предел по количеству ГФ для определения их частотности. Так, частотности ГФ в двух выборках по 1000 ГФ в исследовании Х. Лийва (Лийв Х., 1974, с. 118), взятые из двух произведений художественной литературы, хорошо совпадают с частотностями в выборке 40000 ГФ у А. Корсакова (Корсаков А.,1978, с. 51).

Однако при той же величине выборки (1000 ГФ) из журнала "Machinery" частота ГФ Present Indefinite составляла 81,2%, а при той же выборке из журнала "Metal Forming" частота этой же ГФ составляла только 56,9% из всех ГФ (Елохова Г.В.,1973, с. 62-69). И это в рамках одного подъязыка техники.

При несколько больших выборках в разных подъязыках также наблюдается довольно значительный разброс в частоте личных ГФ. Так, по данным А. Безуглого (Безуглый А.И., 1970, с. 61-68), частота ГФ Present Indefinite в подъязыке авиационной техники составляла 73,2%, а по данным Ж.Зениной (Зенина Ж.М., 1971, с. 91-97) частота этой ГФ в подъязыке электроники составляла 82,5% из всех ГФ. В художественной литературе частотность этой ГФ составляет 25,55% (Корсаков А.,1978, с.51).

В настоящем исследовании использована выборка из 117

статей - 17460 личных глагольных форм (ГФ), распределенных по 6 подъязыкам техники следующим образом: энергетика и теплотехника - 10,36% из всех ГФ, электротехника, автоматика, электроника и вычислительная техника все вместе - 32,85% из всех ГФ, машиностроение и технология машиностроения - 16,85%, химия и химическая технология 18,11%, строительство (строительные материалы, технология строительных конструкций, монтаж и т.п.) - 12,08%, промышленная экономика и управление - 9,75% из всех ГФ. Частотность личных ГФ настоящего исследования в сопоставлении с данными литературы приведена в таблице 1. Данные по машиностроению Г. Елоховой (2 выборки по 1000 ГФ) усреднены (Елохова Г., 1973, с. 64-65). В табл. 1 частотность активных и пассивных конструкций объединена, чтобы вести сравнение с данными литературы.

Учебник ТПИ предусмотрен для первых трех семестров и содержит оригинальные / общетехнические и научно-популярные тексты почти для всех специальностей ТПИ. Этот учебник еще не вышел из печати в момент проведения настоящего исследования. Частотность ГФ в учебнике ТПИ весьма хорошо совпадает с частотностью ГФ генеральной выборки настоящего исследования. Частотность ГФ в учебнике 10 класса для эстонской средней школы (Ehin A., Kuljus I., 1978, с. 177) весьма хорошо совпадает с частотностью ГФ в художественной литературе по А.Корсакову, представленной в табл. 1. Частотность ГФ в учебниках 8, 9 и 10 классов для русских средних школ (Старков А., Диксон Р., 1975, с. 175) находится где-то между частотностями ГФ художественной и технической литературы.

По данным таблицы 1 можно сказать, что если в художественной литературе доминируют ГФ прошедшего времени, то в научно-технической литературе - ГФ настоящего времени и увеличена доля ГФ будущего времени. Это связано с тем, что в художественной литературе описываются события, явления, персонажи и их действия чаще всего в прошлом, а в научно-технической литературе описываются устройства, процессы, явления в момент их исследования, работы, причем исследователя, конструктора или инженера чаще всего интересует то, что происходит сейчас или что будет в будущем. Указанная тенденция превалирования ГФ настоящего времени в научно-технической литературе отмечена многими исследователями. По данным Р.Резник, ГФ настоящего времени составляют 88% от всех ГФ (Резник Р., 1979, с. 11). По данным настоящего исследования, доля ГФ на-

Таблица I

Относительная частотность личных глагольных форм английского языка в различных источниках (%)

Глагольная форма (ГФ)	Художественная литература			Научно-техническая литература				Тексты в учебниках		
	по А.Корсакову	по Х.Лийву	по Рожковой	по ТПИ	авиация по Безуглому	машиностроен. по Елховой	электроника по Зенной	ТПИ	А. Старкова	А. Эхин
I	2	3	4	5	6	7	8	9	I0	II
Количество ГФ в выборке	40000	I002	около I5000	I746I	I500	2000	I707	I620	4I20	980
I. Pres. Indef.	25,5	25,5	2I,8	63,9	73,2	69,0	82,5	7I,7	45,3	23,5
2. Past Indef.	57,7	58,0	59,5	I3,8	I0,8	9,4	6,5	I5,3	35,7	62,5
3. Future Indef.	3,0	3,0	3,3	5,2	4,8	5,4	3,5	2,7	4,8	2,9
4. Pres. Perfect	3,8	4,5	I,4	5,5	3,7	9,3	3,5	3,6	4,0	2,4
5. Past Perfect	5,2	4,5	4,4	I,I	0,7	I,0	-	0,4	I,3	4,4
6. Pres. Contin.	I,9	I,4	I,0	5,0	I,3	2,0	0,5	2,6	I,6	I,2
7. Past Contin.	2,4	2,5	2,6	0,4	-	0,5	-	0,7	I,0	2,I
8. Прочие ГФ	0,5	0,5	4,0	5,0	5,5	3,6	3,5	4,0	6,3	I,0

Таблица 2

Относительная частотность личных глагольных форм в %

Глагольная форм	Все подъязыки	Энергетика	Автом. вычисл. техника	Машиностроение	Химия	Строительство	Экономика
I	2	3	4	5	6	7	8
Количество ГФ	I746I	I8I0	5736	2942	3I63	2II0	I700
I. Active Voice: Pres. Indef.	49,57	43,48	58,77	48,03	46,76	44,3I	5I,82
2. Past Indef.	I0,6	I3,8I	7,25	9,04	8,63	II,I8	II,65
3. Future Indef.	4,II	3,87	4,I6	3,60	4,7I	I,89	7,24
4. Fut. in the Past Ind.	2,5I	3,70	2,I3	3,50	2,59	2,I3	I,I8
5. Pres. Contin.	4,00	3,70	2,75	3,60	4,55	3,55	5,76
6. Past Contin.	0,35	0,50	0,I9	0,6I	0,06	0,95	0,I2
7. Fut. Contin.	0,26	0,72	0,I2	0,I4	0,I6	0,I4	0,82
8. Fut. in the Past Contin.	0,34	0,28	0,02	0,I7	0,06	I,75	0,24

Продолжение табл. 2

I	2	3	4	5	6	7	8
9. Pres. Perf.	3,86	2,I5	2,30	5,I3	4,7I	5,02	5,76
I0. Past Perf.	0,87	0,39	0,3I	0,5I	0,22	4,45	0,65
II. Fut. Perf.	0,I6	0,44	0,07	0,3I	0,09	0,I4	0,06
I2. Pres. Perf. Cont.	0,36	0,I7	0,I9	0,7I	0,I6	0,38	0,82
I3. Past Perf. Cont.	0,04	0,06	0,07	0,00	0,03	0,04	0,06
Passive Voice:							
I4. Pres. Indef.	I4,44	I4,97	I4,I4	I4,03	I9,20	I3,03	9,82
I5. Past Indef.	3,24	4,80	2,5I	3,40	2,I2	5,97	I,88
I6. Fut. Indef.	I,08	I,44	I,I0	I,05	0,88	I,09	0,35
I7. Fut. in the Past Indef.	I,32	2,43	0,98	2,58	I,0I	0,90	0,76
I8. Pres. Contin.	0,95	0,44	I,I0	I,29	0,92	I,04	0,35
I9. Pres. Perfect	I,67	2,04	I,73	2,00	2,06	I,09	0,53
20. Past Perfect	0,27	0,60	0,I0	0,27	0,I6	0,90	0,I2

стоящего времени составляет 74,85% от всех ГФ (табл. 2). В художественной литературе эта доля равняется лишь 31,25% от всех ГФ (Корсаков А., 1978, с. 51).

Данные, приведенные в табл. 1, по разным подъязыкам значительно расходятся. Для проверки этого факта в ходе настоящей работы были рассмотрены ГФ 6 подъязыков (табл. 2). Результаты исследования показали, что в рассмотренных выборках ГФ подъязыки техники имеют довольно большие различия как в случае более частотных ГФ, так и в случае менее частотных ГФ. Больше всего приближаются к частотностям ГФ генеральной выборки частотности ГФ в таких подъязыках, как энергетика, машиностроение и химия. Поскольку исследование будет продолжено, то решено на данной стадии не проверять существенность имеющихся различий в частотности ГФ в разных подъязыках техники.

По данным табл. 2 можно заключить, что ряд временных форм глагола имеет очень низкую частотность (ниже 0,5%) и может в принципе быть исключен из дальнейшего рассмотрения, ибо их частотность вряд ли заметно изменится. Несколько неожиданным является факт довольно высокой относительной частотности ГФ Future in the Past. Сравнительно высока также доля ГФ продолженного вида, особенно настоящего времени. По данным Р. Резник, доля ГФ продолженного вида в научных текстах составляет 2,99% от всех ГФ, а в художественной литературе - 7,14%. По данным табл. 2 эта доля составляет 6,3%. По данным А. Корсакова (в табл. 1) доля ГФ продолженного вида в художественной литературе составляет 4,3% от всех ГФ.

В табл. 3 приведены данные по соотношению активных и пассивных ГФ в разных исследованиях. Судя по ним, можно сказать, что относительная частотность активных и пассивных ГФ, рассмотренных в ТПИ подъязыков техники, довольно близка к частотностям всей генеральной выборки. Выборка по экономике приближается по своим данным к выборке Р. Резника по текстам гуманитарных наук (уменьшена доля пассивных ГФ). Данные частотности генеральной выборки ТПИ практически совпадают с данными Н. Любченко по физико-техническим текстам (Любченко Н, 1972, с. 61-69). Частотность пассивных ГФ заметно выше в английских патентах, поэтому в дальнейшем следует провести полный лингвостатический анализ этого подъязыка техники. Для языка патентов характерна традиционная форма оставления на заднем плане автора изобретения, законченность и полнота

Таблица 3

Соотношение активных и пассивных ГФ (%)
в разных исследованиях

№№ п/п	Автор исследования (характер источника)	Величина выборки ГФ	Active Voice	Passive Voice
I	2	3	4	5
I.	ТПИ (все подъязыки вместе)	I746I	77,0	23,0
2.	ТПИ (энергетика)	I8I0	73,3	26,7
3.	ТПИ (автоматика, вычисл. техн.)	5736	78,3	2I,7
4.	ТПИ (машиностроение)	2942	75,4	24,6
5.	ТПИ (химия)	3I63	73,7	26,3
6.	ТПИ (строительство)	2II0	76,0	24,0
7.	ТПИ (экономика)	I700	86,2	I3,8
8.	Р.Резник (тексты точных наук)	300	68,I	3I,9
9.	Г.Елохова (журнал "Машинери")	I000	42,7	57,3
I0.	ТПИ (американские патенты)	460	56,2	43,8
II.	ТПИ (английские патенты)	5I0	32,5	67,5
I2.	Г.Елохова (журнал "Метал форминг")	I000	53,5	47,5
I3.	Р.Резник (тексты законов)	250	73,I	26,9
I4.	Р.Резник (тексты гуманит. наук)	4I5	84,4	I5,6
I5.	Н.Любченко (физ.-техн. тексты)	I09	73,5	26,5
I6.	ТПИ (тексты учебника ТПИ)	I620	79,0	2I,0
I7.	ТПИ (тексты учебника А. Старкова)	2770	9I,0	9,0
I8.	ТПИ (упраж. учебника А. Старкова)	4040	94,3	5,7
I9.	ТПИ (тексты учебника А. Эхин)	980	95,0	5,0
20.	ТПИ (упражн. учебника А. Эхин)	I2I0	90,3	9,7
2I.	Р.Резник (художеств. литература)	546	95,4	4,6

всех действий и процессов. В текстах учебника английского языка ТПИ доля пассивных ГФ совпадает с их долей в научно-технической литературе, а в учебнике английского языка для эстонских средних школ доля пассивных ГФ совпадает с таковой в художественной литературе по данным Р.Резник. Соотношение активных и пассивных ГФ в текстах и упражениях учебников средних школ несколько разное.

В ходе настоящего исследования в тех же статьях была определена относительная частотность неличных глагольных форм. По некоторым подъязыкам техники, рассмотренным в настоящем исследовании, доли личных и неличных ГФ не совпадают. Так, доли неличных ГФ составляли: по энергетике и теплотехнике - 8,54%, по электронике, автоматике, электронике и вычислительной технике - 37,I8%, по машиностроению - 22,0%, по химии - I4,6%, по строительству - 6,82% и по промышленной электронике - I0,86% от всех неличных ГФ. Отсюда видно, что в текстах по строительству доля неличных ГФ заметно меньше, чем доля личных ГФ.

По данным табл. 4 можно сказать, что во всех подъязыках техники инфинитивных ГФ больше всего. В текстах по энергетике и строительству причастий столько же, сколько инфинитивных ГФ. Наиболее стабильным является относительная частотность герундиальных ГФ. Поскольку эта частотность не очень мала, то эта стабильность может быть некоторым доказательством надежности полученных результатов. К сожалению, в момент написания статьи не было литературных данных о частотности неличных ГФ, хотя это являлось объектом диссертационного исследования Е.И. Ковалевой (Ковалева Е., I972, с. 26). Более подробные сведения об относительной частотности неличных ГФ приведены в табл. 5. По этим данным можно сказать, что неличные ГФ довольно равномерно распределены по видо-временным группам и конструкциям, поскольку почти нет неличной ГФ с частотностью ниже 0,5%. Примечательно, что причастие II имеет бблшую частотность, чем причастие I во всех подъязыках техники. Практически равную с причастием II частотность имеет конструкция инфинитива с модальным глаголом. Бросается в глаза то, что перфектные неличные ГФ имеют самую низкую частотность среди других рассмотренных неличных ГФ.

Как в случае личных ГФ, так и в случае неличных ГФ их частотность в разных подъязыках, а порой и внутри одного подъязыка, по данным Г. Елоховой, имеет значительные расхож-

Таблица 4

Относительная частотность неличных глагольных форм (%)

Пор. №	Глагольная форма	Данные ТПИ	По Н. Любченко в физ.-техн. текстах	По Г. Елоховой в машиностроении	Подъязыки техники (данные ТПИ)						Учебники			
					энергетика	автом. вычислит. техника	химия	строительство	машиностроение	экономика	ТПИ, тексты	А. Старкова, тексты	А. Старкова, упражнения	А. Эхин, тексты
I	2	3	4	5	6	7	8	9	I0	II	I2	I3	I4	I5
	Количество ГФ	6885	II62	I000	588	2560	I005	470	I5I4	748	425	I565	2253	204
I.	Герундий	20,I	8,8	30,9	I8,7	20,0	22,I	25,I	I8,03	I9,8	I5,6	I3,4	I0,4	8,3
2.	Инфинитив	47,0	54,8	20,6	40,5	43,2	53,4	37,9	53,5	48,8	57,0	46,4	68,3	57,I
3.	Причастие I	II,4	нс	I7,7	I7,9	I3,9	4,6	8,I	I2,0	8,3	8,3	2I,3	II,I	2I,6
4.	Причастие II	I5,8	нс	30,8	I9,2	I6,7	I6,I	I4,7	I4,5	I3,6	I3,2	I8,3	8,6	I2,0
5.	Прочие ГФ причастия	5,6	нс	нс	3,7	6,2	3,8	I4,3	I,9	9,5	5,9	0,6	I,6	I,0
6.	Все ГФ причастия	32,9	36,4	47,5	40,8	36,7	24,5	37,0	28,5	3I,4	27,4	40,2	2I,3	34,6

Примечание: "нс" означает - нет сведений.

Таблица 5

Относительная частотность неличных глагольных форм (в %)

Неличные глагольные формы	Все подъязыки	Подъязыки техники					
		энергетика	автом. вычисл. техн.	машиностроение	химия	строительство	экономика
I	2	3	4	5	6	7	8
Gerund:							
1. Simple	I6,33	I5,30	I6,I7	I4,40	I6,52	2I,70	I7,9I
2. Perfect	0,76	0,5I	0,50	I,06	0,60	I,06	I,20
3. Constructions	3,02	2,89	3,36	2,57	4,98	2,34	0,67
Infinitive:							
4. Indefinite	22,56	I4,96	24,69	20,54	2I,59	I8,94	28,88
5. Perfect	0,45	0,I7	0,3I	0,66	0,09	I,9I	0,27
6. Passive	3,67	4,08	2,46	4,75	4,38	4,04	4,I4
7. Subject with Inf.	2,56	3,9I	2,22	I,45	3,58	2,34	3,6I

Продолжение табл. 5

I	2	3	4	5	6	7	8
8. Object with Inf.	3,I	3,23	I,72	0,86	9,75	3,83	2,94
9. Modal v. with Inf.	I2,9I	I3,78	I0,82	20,6	I2,44	5,96	8,82
I0. Modal v. with Perf. Inf.	I,73	0,34	I,0I	4,62	I,59	0,86	0,I3
Participle:							
II. Present	II,44	I7,86	I3,87	I2,02	4,58	8,09	8,29
I2. Past	I5,84	I9,22	I6,68	I4,53	I6,I2	I4,68	I3,64
I3. Perfect	I,55	0,I7	2,I0	0,39	0,40	0,2I	5,48
I4. Passive	I,86	2,2I	I,72	0,79	I,59	4,68	2,8I
I5. Object with Participle	I,05	I,02	0,66	0,53	0,89	4,89	I,07
I6. Absolute Participle Constructions	I,I5	0,34	I,68	0,20	0,89	4,47	0,I3

дения. Эти расхождения в частотностях ГФ могут быть связаны со следующим:

1) с точностью работы, качеством классификации. Если в случае ГФ точность фиксации ГФ в текстах зависит от внимательности и компетентности исследователя, то в случае неличных ГФ могут возникнуть и спорные моменты, например, в случае классификации инговых ГФ.

2) с величиной выборки ГФ, распределением ГФ по разным авторам, журналам, монографиям и функциональным стилям.

3) с характером функционального стиля и тем, описанных в научно-технических статьях и монографиях. Например, обзор развития отрасли по частотности использованных ГФ отличается явно от описания конструкции устройства или технологического процесса. Последние наверняка также различны по частотности ГФ на единицу объема текста и по относительной частотности ГФ между собой. Отдельным объектом лингвостатистического анализа могут быть тексты инструкций, тексты описаний изобретений, реферативных материалов и т.п.

На основании вышеприведенного, можно сделать следующие выводы:

1. В научно-технической литературе относительная частотность английских личных ГФ настоящего времени примерно в три раза выше частотности личных ГФ прошедшего времени, в то время как в художественной литературе частотность личных ГФ прошедшего времени примерно в два раза выше частотности ГФ настоящего времени. Частотность ГФ будущего времени в научно-технической литературе также выше, чем в художественной литературе.

2. Между подъязыками техники как по данным настоящего исследования, так и по данным литературы имеются весьма значительные различия в относительной частотности личных и неличных ГФ, причем эти различия относятся как к частотным ГФ, так и менее частотным ГФ.

3. Относительная частотность ГФ в научно-технической литературе распределена между большим числом видо-временных форм по сравнению с художественной литературой. Однако и в научно-технической литературе имеется ряд видо-временных форм английского глагола, относительная частотность которых составляет менее 0,5% и которые могут быть исключены из дальнейшего анализа из-за малой вероятности значительного изменения их относительной частотности при изменении или

увеличении всей выборки ГФ.

4. Соотношение относительных частотностей активных и пассивных ГФ в научно-технической литературе примерно 3:I, в то время как в художественной литературе доля пассивных ГФ составляет только 5-I0% от всех ГФ. В патентной литературе указанное соотношение примерно I:I.

5. Среди неличных ГФ наиболее частотными являются инфинитивные и причастные ГФ, но относительная частотность герундиальных ГФ наиболее стабильна при переходе от одного подъязыка техники к другому. Причастие прошедшего времени имеет большую относительную частотность по сравнению с причастиями настоящего времени. Все I6 рассмотренных неличных ГФ довольно частотны, только перфектная форма инфинитива имела частотность ниже I%, у остальных неличных ГФ относительная частотность была выше I%.

6. В дальнейшем следует проверить гипотезы о зависимости частотности ГФ в разных подъязыках техники от индивидуального стиля автора статьи (монографии), от названия (характера) журнала, его принадлежности к определенной стране (США, Англия или другие страны), от разновидности функционального стиля (обзор, описание изобретения, доклад конференции, прогноз развития, описывается ли конструкция устройства или технологический процесс и т.п.).

Л И Т Е Р А Т У Р А

Безуглый А.П. О возможности ограничения языкового материала при обучении чтению. - Ученые записки I МГПИИЯ им. М.Тореза, т. 53. М., I970, с. 60-69.

Елохова Г.В. Статистический анализ видовременных форм глагола в английском подъязыке станкостроения. - В кн.: Вопросы анализа специального текста, вып. Уфа, I973, с. 62-69.

Зенина Ж.М. Частотность различных времен и временных групп действительного и страдательного залогов в английской научно-технической литературе. - В кн: Вопросы романо-германской филологии и методики преподавания иностранных языков, вып. 84. Казань, I97I, с. 9I-95.

Ковалева Е.Л. Анализ частотности английских неличных глаголов. Автореф. канд. дисс. М., I972, с. 26.

Корсаков А.К. Употребление времен в английском языке, 2-е изд. Киев, 1978.

Лийв Х. О передаче значений видо-временных форм английского глагола на эстонском языке. - Linguistica, V. Tartu, 1974, с. 106-119.

Любченко Н.В. О применении математической статистики к отбору грамматического минимума физико-технических текстов для учебных целей. - В кн.: Обучение чтению в неязыковом вузе, вып. I. Томск, 1972, с. 61-68.

Резник Р.В. О взаимосвязи количественных и качественных характеристик грамматических явлений. - Иностранные языки в школе, М., 1979, № 4, с. 8-13.

Старков А., Диксон Р. Учебник английского языка (для 8, 9 и 10 классов средней школы). М., 1975.

Ehin A., Kuljus I. English, Form 10. - Tallinn; Valgus, 1978.

RELATIVE FREQUENCIES OF ENGLISH VERB FORMS IN SCIENTIFIC LITERATURE

Valmar Kokkota

S u m m a r y

Relative frequencies of the English Active and Passive, Finite and Nonfinite verb forms have been determined on the basis of 117 papers from 6 different fields of science and technology. The verb corpus amounted to 17.460 Finite and 6885 Nonfinite verb forms. The relative frequency of the Present Tense forms was about three times higher than the relative frequency of the Past Tense forms. The Active verb forms occur about three times more frequently than the Passive verb forms. The most frequent Nonfinite verb form is the Indefinite Infinite (22.6 %) followed by the Simple Gerund (16.3 %) and the Past Participle (15.8%). This English verb frequency investigation will be continued.

ЛЕКСИКО-СЕМАНТИЧЕСКИЕ ОСОБЕННОСТИ ЖАНРА.
(Опыт статистического анализа на материале английских и шотландских народных баллад)

Э.А. Краснова

Для современной семасиологии характерно большое разнообразие подходов и к предмету анализа и к методу. Но при всем разнообразии методик продолжает оставаться слабо разработанным принцип сопоставления семантических систем, тем более сопоставление семантических систем разных этапов истории одного и того же языка. Совсем нет исследований семантических систем, сосуществующих в рамках одного языка.

При этом, если исследование семантики отдельных слов или небольших групп слов подкрепляется нашим субъективным ощущением знания этой семантики, то, по замечанию А.Я. Шайкевича, "этого нельзя сказать о системах в целом, о больших семантических классах слов, о постоянно воссоздаваемых семантических шаблонах ... Здесь мы стоим перед необходимостью семантического открытия. Это верно в отношении языков прошлого и в большей степени справедливо относительно стилистических систем в пределах живого языка" (Шайкевич А.Я.,1976, с.354).

Если при исследовании семантики отдельных слов работа семасиолога ограничена каким-либо одним словом (анализом полисемии, составом сем) или небольшой группой слов (словообразовательными синонимическими рядами, семантическими полями), то при описании лексико-семантической системы в целом или хотя бы ее крупных фрагментов, исследователь, видимо, должен обратиться к достаточно большому корпусу текстов, где исследуемые слова не существуют изолированно, а помещены в естественные семантические множества, к которым они принадлежат до начала разных процедур лингвистического и логического анализа.

Такие опыты сплошного анализа текстов (иногда с применением электронно-вычислительных машин) были сделаны для текстов разной тематики (техники, психологии, юриспруденции).

В области же фольклора и художественной литературы таких исследований практически нет. Тем более существенными оказываются результаты, полученные А.Я. Шайкевичем при анализе лексики комедий Шекспира. (Шайкевич А.Я., 1974). Здесь вы-

являются законы, которые управляют появлением определенных элементов и их связь с некоторыми семантико-стилистическими явлениями, о которых ничего не было известно ни в формальной, ни в содержательной лингвистике (также, в частности, исследование распределения относительной частоты неопределенного артикля в речи персонажей Шекспира, которое обнаруживает существенное расхождение, неоднородность персонажей относительно артикля a).

Таким образом, в лингвистике сейчас ставится вопрос о том, чтобы: I) создать процедуры семантического открытия; 2) исследовать большие срезы текстов; 3) исследовать семантические системы прошлого (Шайкевич А.Я., 1976, с.353-360).

Для решения этих задач привлекается методика дистрибутивно-статистического анализа, которая должна выявить законы, управляющие появлением элементов в тексте. Эта методика строится на одном факте - информации об элементах и их распределении в тексте, и здесь используются два компонента:

I) сведения о распределении элементов во всей совокупности текстов - так называемое текстуальное распределение, т.е. высокая и низкая частота элементов. Этот компонент можно было бы назвать парадигматикой текста (или автора).

2) сведения о распределении элементов относительно другого элемента - взаимное распределение - т.е. синтагматика текста (автора).

Эти принципы и были использованы при анализе всего собрания английских и шотландских народных баллад в издании Ф. Чайльда (F. Childe, "English and Scottish popular ballads", VVI-5 London, 1957).

Предварительным этапом работы явилось составление частотного словаря-индекса всех баллад по Ф. Чайльду, как наиболее полному собранию в нескольких вариантах. Роспись производилась по одному варианту.

В отличие от авторских словарей и различного рода конкордансов, как традиционных (не автоматических), так и составленных с помощью ЭВМ, словарей жанров, ни фольклорных, ни литературных практически нет. В истории английской лексикографии известно два: "Конкорданс к 5-ти средневековым аллитерированным поэмам", 1966, и "Конкорданс к елизаветинскому сонету", 1969.

Между тем потребность в такого рода исчерпывающих словниках становится все более ощутимой именно потому, что все

суждения об употреблении слов, изменении их значений об их синтагматике и парадигматике могут быть корректно поставлены лишь при условии полноты словаря, как авторского, так, видимо, и жанра.

Словари жанров могли бы дать те фонды, потребность которых сильно ощущается и в лингвистике, и в литературоведении. В связи с этим поучительно, в частности, замечание М.Л. Гаспарова: "Современный структурализм прав, когда подчеркивает, что прием не есть факт, а отношение факта к фону, на который он проецируется, что отсутствие приема может быть действеннее, чем его наличие. Но это значит, что для констатации приема мы должны знать фон так же хорошо, как факт. Нужно иметь исчерпывающую картину плюс-приемов предшествующей эпохи... Такой картины у нас нет, а она необходима" (Гаспаров М.Л., 1969, с. 514).

Знание лексико-семантических фонов, семантических систем жанров намного бы облегчило и уточнило наши суждения об индивидуальных, авторских семантических системах, дало бы возможность измерить расстояние между жанром и автором, между разными историческими этапами в развитии одного жанра и, наконец, между разными жанрами.

Сформулируем теперь задачи исследования:

I. Используя вариант дисперсионного анализа, попытаемся выделить группы слов высокого и низкого рассеивания по текстам 297 баллад. Это дало бы возможность:

а) найти словарные минимумы, присущие любому или почти любому балладному тексту, т.е. найти некоторый лексический инвариант, определяющий жанр;

б) найти слова высокой дисперсии, т.е. лексические группы, обильно представленные в одних текстах и отсутствующие в других, и тем самым обнаружить внутреннюю неоднородность жанра.

2. На материале нескольких наиболее частых слов из разных семантических центров можно проследить особенности синтагматических отношений. Связь 2-х элементов устанавливается на основе их совместного появления в строке баллады.

Общий словарь баллады рассматриваемой традиции составляет около 230000 словоупотреблений на приблизительно 3500 разных слов. Здесь сразу же становится очевидной весьма сильная ограниченность лексики жанра и высокая степень повторяемости сравнительно небольшого числа слов. Таким об-

разом, выявляется высокая каноничность жанра и большая стабильность его словарного состава от текста к тексту.

Анализ текстуального распределения лексики (слов высокой и низкой встречаемости) позволил отобрать группу самых частых слов (встреченных более 80 раз по всему корпусу баллад). Эта группа в 248 слов (служебных и знаменательных) составляет более половины всех словоупотреблений - I38396 раз или 60,2% всего словаря. Это соотношение еще раз подчеркивает, насколько замкнут словарь жанра. Видимо, можно считать, что этот минимум составляет семантико-грамматическое ядро этого типа текстов, куда наряду со знаменательными словами, входят и грамматические элементы (артикли, предлоги, союзы, частицы). С другой стороны, в самой группе довольно легко можно выделить несколько характерных семантических центров, которые определяют семантическое пространство жанра и образуют обобщенные компоненты смысла, связанные как с темой, так и с общей экспрессивной тональностью текстов. Они следующие (именования смыслов условные):

I. Сильно выраженный компонент диалогичности и обращенности речи, представленный самой большой по относительной частоте группой личных и притяжательных местоимений (здесь и далее относительные частоты приводятся на I38396 слов текста): более 32000 словоупотреблений. Распределение внутри местоименной группы представлено следующим образом: I-me-my-mine (6,8); he-him-his (5,8); you-ye-your (2,7); she-her (2,3); thou-thee-thy (1,84); they-them-their (1,8); it (I); we-us-our (0,8).

Преобладание группы местоимений I и 2-го лица (I2,I%) в их противопоставлении местоимениям 3-го лица есть существенный показатель сильной диалогичности текстов.

Этот компонент поддерживается и важностью группы глаголов говорения (около 3900 словоупотреблений - 2,77%) - say, tell, speak, call, quoth, bid, bespeak, swear и 2-мя существительными word, name.

II. Следующий по весу компонент - пространственных обозначений (I3600 употреблений или I2,6%), куда вместе с пространственными предлогами (они составляют около 52% всей группы предлогов): in, on, at, from, up, out, upon, under, over, before, into, through, входят и наречия места: there, down, away, here, far, yon. Специфическую подгруппу образуют существительные, локализующие развертывание балладных

событий и поступков персонажей: это home, land, bed, bower, town, country, hall, castle, house, sea, wood. Любопытно, что именования внешнего мира - природы, космоса, элементов ландшафта представлено крайне невыразительно - таковы относительно редкие слова sun, moon, world; примечательно, что слово sky относится к hapax legomena.

Так вырисовывается сильно выраженная замкнутость балладного пространства, его ограниченность ближайшим человеческим окружением, а именно жилищем. Одновременно семантика наречий места носит обобщенный, недифференцированный характер; действие происходит здесь, там, далеко (there, here, away, far, yon).

III. Большая группа глаголов со значением действия позволяет выделить третий компонент - активности (5,74%), значимость которого для баллады не раз отмечалась и в традиционных исследованиях (H.G. Gerould, 1957; L. Vargyas, 1967). В частности, Г.Джералд по этому поводу замечает: "Действие и действие, сосредоточенное на одной единственной ситуации есть первая константа, обнаруживаемая в балладах европейских народов" (H.G. Gerould, 1957, с. 6). Сама семантика глаголов этой группы выявляет общий характер этой деятельности; ср., например, очень частые глаголы do, make, get, give, take, let, bring, send, meet, bear, find, know, think, put, keep, save, fight, shoot (около 9000 употреблений). Сюда же может быть отнесена большая группа модальных глаголов, характеризующих общую возможность, невозможность, желательность, необходимость этой деятельности: can, may, would, must, should, might (9,36%).

IУ. Компонент активности дополняется глагольной группой, очень важной и специфической для семантики баллады - глаголов движения, выявляющих таким образом компонент повышенной динамичности мобильности персонажей жанра, которые находятся в постоянном движении, перемещении внутри своего замкнутого микромира (около 4900 употреблений или 3,5%: это come, go, stand, run, leave, stay, sit, turn, ride, lie, fall, win (в значении "идти, приходить"), сюда же может быть отнесено существительное way.

У. Развертывание деятельности и перемещения в ограниченном, замкнутом пространстве отнесено во времени, но времени, отмеченном либо вечностью (ever, never, aye), либо последовательностью событий одного за другим: now, again,

then, soon, till; из конкретных указаний времени важны night, в меньшей степени day и year. Регулярна атрибуция персонажей эпитетом young реже - old. Вся группа временных отнесений составляет около 5000 употреблений или 3,5%.

УI. В центре балладного пространства и времени - человек, лицо - мужчина, он, король, лорд, рыцарь, сэр, господин, юнош, шериф: he, man, king, lord, knight, sir, master, boy, sheriff (около 3600 или 2,56%); женская группа вдвое меньше - это она, дама, жена, королева, женщина, дева, невеста; she, lady, wife, queen, woman, maid, bride (около I600 или I,I4%). Описание лиц дано через внешние признаки (анатомические) - hand, knee, head, foot, eye, body, hair (0,94), и функционально-оценочные, выделяемые в особую группу. Ближайшее социальное окружение персонажей баллады - это семья: термины родства составляют весьма существенный компонент семантики жанра (I,07%; около I500 словоупотреблений). Является ли этот компонент универсальным признаком баллады, сказать трудно. Необходимо типологическое исследование хотя бы нескольких европейских традиций. Наблюдения В.Я. Проппа над различиями русских былин и баллад дают ему основания считать, что главное отличие первых от вторых в том, что баллада охватывает в основном сферу семьи и личных отношений. Эту же особенность отличает Д.М. Балашов, также на русском материале: "Баллада ставит в центр внимания индивидуальную человеческую судьбу. События общенародного значения, этические, социальные, философские проблемы получают в балладах отражение в виде конкретных судеб отдельных лиц и частно-семейных человеческих отношений. Герой баллады замкнут рамками своего "я" или мира своей семьи. Он сам по себе. Мир баллады - это мир лиц и семей разрозненных, распадающихся во враждебном или безразличном окружении (Балашов Д.М., I966, с. 9I).

УП. События и персонажи получают в текстах стабильные оценочные атрибуции - штампы, где можно выделить несколько подгрупп: I) качественные: а) положительные good, fair, bonny, bold, merry, gay, true, best, noble, right и б) один очень частый отрицательный - false 2) количественные: а) группа местоименных количественных опеределителей, где особенно характерны обобщающие all, every, any, none также both, some, another б) традиционный ряд так называемых магических чисел, так или иначе присутствующих почти в каждом

балладном тексте: one, two, three, seven, first, hundred в) группа слов, выражающих степень и размер: little, more, full, long, many, fast, great, high, than, so.

УШ. Лирический компонент смысла (0,9): love, heart, marry, sweet, dear.

IX. Минорно-фатальный (0,6) - die, slay, blood, woe, sore.

X. Предметный компонент (0,57) - bow, sword, ship, gold, silver.

XI. Цветовой компонент (0,36) - особенно выделяются green и red в меньшей степени white.

XII. Растительно-животный (0,34) - horse, steed, tree.

XIII. Экзистенциальный (0,19) - life, live.

XIУ. Божественный (0,16) - god.

Выделенный перечень обобщенных компонентов смысла очерчивает, видимо, своего рода семантический инвариант жанра. Каждый из этих компонентов играет свою формообразующую роль в структуре баллады и позволяет выявить ту систему, которая скрывается за группами текстов или отдельными текстами (стилистическую семантическую систему, по определению А.Я. Шайкевича, 1976, с. 360).

Очерченная таким образом общая семантика жанра требует, очевидно, дальнейших уточнений и дифференциации. Естественно предположить, что внутри жанра возможны разбиения, группировки, т.е. известная внутренняя неоднородность.

Основным средством измерения внутренней неоднородности в статистике служит дисперсия, определяемая как средний квадрат отклонений по подвыборкам от общей арифметической средней. Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость.

Во многих случаях изучение самых общих параметров распределения уже позволяет классифицировать элементы. Здесь используется простейший вариант анализа, вычисляемый как отношение ранга частоты слова к рангу текстов, в которых оно встречается $A = \frac{Rf}{Tf}$. Для этого в отобранной группе самых частых слов (248) их располагают в порядке убывания их частот и в порядке убывания числа текстов, в которых слово употреблено. Отношение их порядков или рангов дает приближенное представление о дисперсии (характере рассеивания слова). Проведенные подсчеты позволяют выделить 2 группы слов - слова низкой дисперсии - равной и выше единицы, т.е. отно-

сительно равномерного рассеивания по всем текстам и выявляющие однородность их разброса; к слова высокой дисперсии - ниже единицы. Первую группу можно, по методике А.Я. Шайкевича, назвать "строевыми" элементами. Сюда входят, как следовало ожидать, все грамматические слова (артикли, предлоги, союзы, частицы). Наряду с ними в этой же группе особо следует отметить группу сочинительных союзов and, or, but как показателей сильно выраженного паратаксиса, в противопоставлении к подчинительным союзам for и if, не вошедшим в эту группу. Обнаруживаются и специфические строевые слова, принадлежащие собственно языку жанра. К словам низкой дисперсии оказались отнесенными из перечисленных выше групп следующие:

I) местоимения личные и притяжательные: I - me, my - mine, he - his, you - your, them - their, us - our; указательные: this, that относительные: what, which, who; неопределенно-личные: all, both, some, any, none, every, another.

2) вся группа локализующих существительных, за исключением 2-х (sea, wood), и все наречия места;

3) все глаголы, передающие значения активности, действия (за исключением fight);

4) все глаголы движения и существительное way;

5) все обозначения времени, кроме then, again;

6) из мужской группы лишь boy из женской группы: wife, woman, maid, bride;

из терминов родства - все, кроме son и brother; все названия частей тела;

7) из оценочных атрибуций: good, fair, merry, own, gay, true, best, noble, right, false;

весь ряд чисел: one, two, three, seven, first, hundred;

8) слова, относящиеся к лирическому компоненту: heart, marry, sweet, dear (кроме love);

9) слова, относящиеся к смерти и скорби: die, slay, blood, woe, sore;

I0) из предметной группы: sword, ship, gold, silver;

II) все цветовые обозначения: green, red, white;

I2) оба наименования коня: horse, steed;

I3) "жизнь", "жить" - life, live.

Общее число слов низкой дисперсии - больше половины отобранного списка - I6I слово или 64,5%. Таким образом, вырисовывается особая грамматика жанра английской баллады, грам-

матика, основанная не на регулярности морфологического выражения, а лексико-семантической регулярности, некоторой семантический шаблон, весьма стабильно повторяемый от текста к тексту. Тем самым обнаруживается и большая однородность фольклорного жанра (сравнительно, например, с авторскими текстами, исследованными А.Я. Шайкевичем).

Так же сильно отличается от авторских текстов и вторая группа слов высокой дисперсии, т.е. слов, обильно представленных в одних текстах и отсутствующих в других, что обнаруживает их неоднородность по данным признакам. Если у Шекспира выделяется большая группа слов, у которых отмечается высокая дисперсия, то в случае баллад она довольно немногочисленна - всего 34 слова или 13,7% всего избранного списка самых частых слов (при этом общая длина всех текстов шекспировских комедий приблизительно совпадает с длиной текстов баллад - около 290000 и 230000 слов соответственно).

В эту группу не вошло ни одно служебное слово; из глаголов говорения к словам неравномерного рассеивания оказываются отнесенными: say и архаичное quoth; из других глаголов fight, из местоимений - женская группа she, her, весь ряд местоимений 2-го лица ед. числа thee - thou - thy; вариант местоимения 2-го лица мн. числа ye (примечательно, что you имеет регулярный разброс); местоимения they и we и указательное these; из наречий времени - then, again; из существительных - два, указывающих на место действия: sea, wood; и одно из предметной области - bow (лук). Больше всего слов высокой дисперсии относится к группе названий лиц и семейных отношений: это man, lord, king, knight, sir, master (т.е. практически все слова мужской группы, за исключением boy); lady - queen (королевско-феодальные термины); son и brother. Тексты обнаруживают неоднородность и относительно слов love и god (также высокой дисперсии). Из прилагательных неравномерно рассеиваются little, bonny, full, bold и old.

Эта группа слов неравномерного рассеивания даст основание полагать, что могут обнаружиться группы текстов, где они обильно представлены, и тексты, где они не встречаются или встречаются весьма редко. Таким образом, может быть получена некоторая классификация внутри жанра по небольшому числу признаков. Для решения этой задачи может быть использована формальная методика статистической классификации текстов, уже применявшаяся для новоанглийского материала (Шайкевич А.Я.,

1969) и для классификации персонажей шекспировских комедий (Шайкевич А.Я., 1974). Основное средство анализа - подсчет коэффициента корреляции между парами текстов на основе небольшого списка слов, что является предметом дальнейших исследований.

Сильная ограниченность лексико-семантического ядра жанра дополняется, с другой стороны, стандартностью и устойчивостью синтагматических связей наиболее частых слов на уровне строки или четверостишия. Приведем хотя бы некоторые из них. Сравним распределение таких оценочных эпитетов как bonny, bright, true, false, dear, fair, good, gay, green, red, white, merry. Сочетаемость прилагательных весьма избирательна. Так, bonny предпочтительно определяет boy, bride, lass, may (дева) и ряд слов, начинающихся с (b), что вызвано соображениями аллитерации. Синонимичное прилагательное fair сочетается с lady и maid, но никогда с lass и may . Bright стоит в атрибуции к lady (из ряда именований женщины), а dear - только к терминам родства (mother, brother, father, daughter, sister). Good определяет существительное lord, но никогда не определяет lady и lass, в то же время как gay предпочтительно определяет lady (70% всех сочетаний).

Сочетания с названиями цветов очень устойчивы. Red регулярно определяет gold; white - bread и money. Специфично для баллады сочетание green sea.

Не менее стандартны сочетания и эмоционального ряда: merry men, true love (в отличие от false love, встреченного I раз).

Вырисовывается сильная клишированность атрибутивных сочетаний, превращение их в формулы, функционирующие в жанре как фразеологические единства.

Показательна и стилистическая дифференциация атрибуций к существительным в синонимических рядах типа man - king - lord - sir - knight; woman - wife - lady - maid - maiden - may; streed - horse. Man определяется через ряд возрастных, социальных, внешних, этических атрибуций: brave, stark, sturdy, worthy, bloody, beastly, hardy, fierce, cursed, good, old, tall, naked; well-made; если же обратиться к синонимам lord - king - sir - knight, то картина меняется: lord всегда определяется через good (77% всех сочетаний); king обязательно noble; sir всегда kind и good; knight - gentle, courteous, curtly.

Так же четко дифференцируются атрибуции и внутри женской группы: слова стилистически нейтральные woman и wife получают возрастные, нравственные и эмоциональные характеристики: old, young, sorry, grieved, greeting, woeful , но слова lady, maid, maiden, may как правило, получают характеристики, описывающие внешность: lady всегда fair, gay, bright; maid - pretty lovely, gentle; may - rare, well - faired и т.д.

Наличие словаря-индекса позволяет выделить, таким образом, специальную фразеологию жанра, набор формульных сочетаний, которые функционируют как готовые блоки или на уровне атрибутивных сочетаний AN типа red gold, или на уровне строки: when the bell was rung and the mass was sung; или даже целого двустишия: When the bell was rung and the mass was sung, and all men were bound for bed.

Рассмотренный материал и методика анализа позволяют сделать следующие предварительные выводы:

I. Относительно большая группа слов низкой дисперсии (равномерного рассеивания) обнаруживает сильно выраженную лексико-семантическую однородность данного типа текстов.

2. Сильная замкнутость тематики жанра прослеживается в явной доминации компонентов человеческого микромира и его ближайшего окружения, как правило, бытового,и подчинения остальной тематики этой доминирующей сфере.

3. Характерна тенденция к обобщенному, лишенному индивидуализации, способу описания, что подтверждается:

а) высокой частотой таких атрибуций, как good, bonny, fair, стоящих в препозиции к большому числу существительных самой разнообразной семантики; например, fair является определением к lady-maid, Scotland, England, ship, steed, и т.д.; bonny-kmaid, hall, cow, castle, cock и т.д.

Такой широкий диапазон сочетаемости этих определений приводит к известному опустошению их семантики и автоматичности их употребления. Они становятся некоторыми обязательными элементами и формируют "грамматику" жанра.

б) преобладанием в ряде семантических центров слов с общим значением: do, make, come, go, give, take; обобщенных местоименных определений: all, every, each, any, none; временных наречий: ever, never, aye, now, then; пространственных наречий here, there, yon. Они подчеркивают определенную универсальность балладного времени и пространства, известный

фатализм и неизбежность балладных событий.

4. Предложенная попытка индексации и частичного анализа лексико-семантических структур английской и шотландской народной баллады могла бы быть полезна для общей типологии текстов и для сравнительного анализа истории этой семантической системы в английской фольклористике и художественной литературе; в частности, на фоне лексико-семантического инварианта народной баллады могло быть проведено исследование английской литературной баллады Вордсворда, Кольриджа и других романтиков.

ЛИТЕРАТУРА

Балашов Д.М. История развития жанра русской баллады. Петрозаводск, 1966.

Гаспаров М.Л. Вступительная заметка к статье Б.И. Ярхо "Методология точного литературоведения (набросок плана)". - Труды по знаковым системам, IУ Тарту, 1969.

Шайкевич А.Я. Опыт статистического выделения функциональных стилей. - Вопросы языкознания, 1969, № 1.

Шайкевич А.Я. Дисперсионный анализ лексики шекспировских комедий. - Сборник научных трудов МТИИЯ им. М.Тореза, вып. 91. М., 1974.

Шайкевич А.Я. Дистрибутивно-статистический анализ в семантике. - В кн.: Принципы и методы семантических исследований. М., 1976.

Donow H.S. A Concordance to the Sonnet Sequence of Daniel, Drayton, Shakespeare, Sidney, Spenser. Carbondale, 1969.

Gerould H.G. The Ballad of Tradition. N. Y., 1957.

Kottler B., Markmann. A Concordance to Five Middle English Alliterative Poems. 1966.

Vargyas L. Researches into the Medieval History of Folk Ballads. 1967.

LEXICAL AND SEMANTIC DISTINCTIONS OF A GENRE

(a statistical analysis of English and Scottish popular ballads)

Eleonora Krasnova

S u m m a r y

On the basis of a complete concordance made up for 298 ballads of F. Childe's five-volume collection of "English and Scottish popular ballads", 1957, an attempt has been made to investigate the nature of choice and distribution of the most frequent words (both notional ones and auxiliaries) in the ballads. By applying the statistical procedure of dispersion analysis to 248 words two groups of words have been obtained: those of low dispersion and high dispersion. The first group comprises the so-called structure elements inevitably available in practically each text of the genre under consideration. This particular group reveals a highly homogeneous character of the ballad type, whereas the other group of high dispersion gives evidence of a certain differentiation within the genre and provides further grounds for formal correlation analysis and classification of the ballad texts.

АВТОМАТИЗИРОВАННАЯ СИСТЕМА ПОДБОРА ВИДА РАСПРЕДЕЛЕНИЯ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ

М. Милюна, Т. Якубайтис

Приступая к изучению какого-то явления, исследователь может ставить своей целью установление в нем определенных закономерностей, иными словами, выдвинуть определенную гипотезу. В статистической гипотезе в качестве объектов и закономерностей, подлежащих анализу, выступают параметры генеральной совокупности и форма распределения, которая представляет собой наиболее полную характеристику изучаемой случайной величины (Дружинин Н.К., 1973).

Выявление закона распределения случайной величины - один из важнейших этапов статистического анализа. Во многих областях приложения статистики, при решении многих задач форму теоретического распределения можно считать известной. Так, при изучении ряда технологических процессов исходят из предположения о нормальности распределения. В некоторых физических и биологических процессах предполагаемым теоретическим законом является распределение Пуассона (Смирнов Н.В., Дунин-Барковский И.В., 1969).

Признаки, явления и процессы, изучаемые общественными науками, исследованы недостаточно с точки зрения особенностей их распределения. Не составляет в этом смысле исключения и лингвистика.

Между тем глубокий анализ статистической структуры текста требует знания формы распределения единиц, составляющих структуру. В этом направлении сделано еще немного.

В основном, работы, в которых рассматривается закон распределения лингвистических единиц, посвящены проверке гипотезы нормальности распределения. Это вполне естественно, поскольку нормальный закон занимает особое место в математической статистике. Он наиболее изучен, т.к. ранее других привлек внимание статистиков. Этот закон является предельным для ряда других законов. Можно сказать, что наличие нормального распределения позволяет максимально широко использовать аппарат математической статистики. Поэтому во всех ее приложениях, приступая к рассмотрению эмпирических распределений, прежде всего проверяют их соответствие нормальному закону

распределения, если закон распределения заранее неизвестен.

Накопленный к настоящему времени опыт изучения распределений языковых единиц разных уровней показывает, что лексемы, как правило, не имеют нормального распределения, в то время как распределение грамматических форм и категорий в большинстве случаев достаточно хорошо описывается нормальным законом.

Следует, однако, отметить, что некоторые исследования основаны на столь малом числе наблюдений, что их доказательная сила вызывает сомнение. Кроме того, в экспериментах по проверке вида распределения языковых единиц не всегда принимается во внимание необходимость правильной организации и проведения эксперимента. Поэтому, оценивая в целом состояние обсуждаемого вопроса, некоторые авторы констатируют "полную неразработанность математических основ лингвостатистических распределений" (Городецкий Б.Ю., 1977, с. 28).

На первом этапе развития лингвостатистики, когда высказывались предварительные соображения и делались пробные вычисления, можно было удовлетвориться небольшими объемами текстов и несложными расчетами. Когда этот этап был пройден, исследователи оказались перед необходимостью перейти к серьезным экспериментам с привлечением огромных текстовых массивов и применением сложного аппарата математической статистики. Это обстоятельство в известной мере затормозило, на наш взгляд, развитие лингвостатистических исследований в целом. Любые расчеты на больших текстовых массивах трудно осуществить вручную, а лингвист далеко не всегда имеет возможность использовать вычислительную технику.

Между тем современная вычислительная техника открывает широкие перспективы для автоматизации научных исследований. Это в большой степени касается тех работ, которые используют методы математической статистики.

Как мы уже отмечали, "современные ЭВМ отличаются от предыдущих не только увеличением быстродействия и объема памяти, но и усложнением функций, оснащенностью большим набором пакетов стандартных программ и комплексом различных аппаратов и устройств. Эти особенности новых ЭВМ позволяют устранить или значительно уменьшить многие трудности, с которыми сталкиваются языковеды, обращаясь к помощи машин. Прежде всего это относится к возможности усовершенствовать методику ввода-вывода и коррекцию информации, поиска и исправления

ошибок, контроля и редактирование результатов обработки материала" (Якубайтис Т.А., 1979).

Однако расширение возможностей ввода и оперирования с большими массивами информации - это только одно из направлений процесса автоматизации научных исследований.

По мере расширения сферы приложения методов математической статистики все актуальней становится вопрос об автоматизации самого процесса статистического анализа. Дело в том, что экономист, историк или лингвист, обратившись к методам математической статистики для обработки экспериментальных данных, далеко не всегда обладает достаточной математической подготовкой, чтобы сделать надежные теоретические обоснования выбора той или иной статистической модели или корректности применения определенной методики. Поэтому остро ощущается необходимость разработки алгоритмов и программ, позволяющих автоматизировать и этот этап научного анализа.

В Институте электроники и вычислительной техники АН Латвийской ССР создана интерактивная система оптимального выбора гипотез о статистических распределениях. Эта система решает задачу автоматизированного подбора вида и параметров распределения на ЭВМ для данной экспериментальной выборки и проверки адекватности подобранной модели.

Система ориентирована на широкий круг потребителей из различных областей науки и техники, имеющих дело со статистической обработкой экспериментальных данных и статистическим моделированием на ЭВМ.

Вопрос автоматизированного подбора вида распределения решен на базе новейших достижений теории вероятностей и математической статистики, дополнен новыми исследованиями и использует такие возможности новой вычислительной техники и математического обеспечения, предоставляемых системой коллективного пользования, как режим диалога и режим реального времени, передачу данных эксперимента и результатов обработки на расстояния по линиям связи, экранные пульты, терминалы и т.д.

Для выбора типа распределения на основании обработки экспериментальных данных был разработан обладающий наибольшей наглядностью и поддающейся автоматизации графический метод подбора типа распределения.

Первоначальной основой для разработки этого метода послужили работы Л.Н. Большева (1965) и Г. Хана и С. Шапиро (1969).

Данный метод иллюстрируется рисунком I, где в плоскости распределений $\beta_2 O \beta_1$ показаны области, линии и точки для различных распределений.

Величина $\sqrt{\beta_1} = \mu_3 / \mu_2^{\frac{3}{2}}$ является нормированным показателем асимметрии, где $\mu_k = \int_{-\infty}^{\infty} (x - m_x)^k f(x) dx$ центральный k-ый момент, $\beta_2 = \mu_4 / \mu_2^2$ определяет относительный показатель эксцесса.

Если вдоль осей прямоугольной системы кординат откладывать отрезки, соответствующие величинам β_2 и β_1, то в плоскости $\beta_2 O \beta_1$ различным распределениям будут соответствовать области, линии и точки. Непрерывные распределения имеют три вида параметров, определяющих соответственно центр распределения, его масштаб и форму. Не все распределения содержат каждый из этих параметров. Например, нормальное распределение не имеет параметра формы, а у бета-распределения два таких параметра. Этим объясняется то, что нормальное распределение на рисунке представлено точкой N (3; 0), а бета-распределение - целой областью. Наряду с часто рассматриваемыми распределениями более общими методами описания эмпирических данных являются распределения семейств Джонсона и системы кривых Пирсона.

Прямая линия АВ, имеющая уравнение $\beta_2 - \beta_1 - 1 = 0$, представляет собой верхнюю границу для допустимых точек (β_2, β_1), так как не существует распределений, для которых $\beta_2 - \beta_1 - 1 < 0$.

Прямая НК, уравнение которой $8\beta_2 - 15\beta_1 - 36 = 0$, служит нижней границей точек с координатами (β_2, β_1) для кривых Пирсона.

Для практического использования графиков плоскости распределений необходимо вычислить выборочные оценки показателей асимметрии и эксцесса и нанести точку с такими координатами на рисунок. Был решен вопрос о доверительной области экспериментальной точки, т.е. было найдено уравнение эллипса (Милюна М.К., 1978), ограничивающего эту область, в прямоугольной системе координат β_2 о $\sqrt{\beta_1}$ и осуществлен переход на используемую систему координат $\beta_2 O \beta_1$. Без решения этого вопроса невозможна была бы автоматизацией всей обработки.

Предусмотрен учет соображений исследователя о предполагаемом виде распределения. При работе в диалоговом режиме - это ответы пользователя на вопросы системы, а в режиме без диалога - массив ISS (одна перфокарта), содержащий "I" для

тех видов распределений, по которым надо вести статистическую оценку модели.

Описанная система была опробована на материале биологических исследований (Miļūna, M., Buss, H., Buss, K., 1975). В настоящей работе приведены результаты использования системы для анализа вида распределения лингвистических единиц.

Исходным материалом послужили латышские тексты художественные, научно-технические, газетно-журнальные и официально-деловые общим объемом I.060.000 словоупотреблений*.

Эта текстовая совокупность разбита на I2 подсовокупностей (подъязыков). Их перечень дан в таблице I.

Таблица I

Индекс	Объем	Тип текста	Индекс	Объем	Тип текста
ША	I00 тыс.	Худ. проза	*OF*	60 тыс.	Официально-деловые докум.
ШВ	I00 тыс.	Драматургия	П*L*	60 тыс.	Газеты (литерат. и искусств.)
ШС	I00 тыс.	Поэзия	ПР	60 тыс.	Газеты (статьи по политике)
IУВ	I00 тыс.	Естеств.науки	П*D*	60 тыс.	Газеты (спорт и разн.)
IУС	I00 тыс.	Физико-мат. науки	ПТ	60 тыс.	Газеты (наука и техника)
IУ*S*	I00 тыс.	Обществ.науки	П*J*	60 тыс.	Газеты (педагогика, просвещение)

В качестве изучаемой случайной величины взято количество существительных (S) в отрезке связного текста длиной в I000 словоупотреблений. Рассматривалось распределение S в каждом подъязыке, в их комбинации и во всей текстовой совокупности в целом.

Индекс каждой комбинации образован из индексов составляющих его подъязыков. Например, индекс ШАВ означает, что комбинация состоит из подсовокупностей ША и ШВ. Индекс Ш + IУ соответствует совокупности, состоящей из подсовокупностей ША, ШВ, ШС, IУВ, IУС, IУ

* Более подробно о принципах подбора и обработки текстов см. во введении к каждому тому четырехтомного "Частотного словаря латышского языка", Рига, 1966-1976 гг.

Рассмотренные подсовокупности II и III тома вместе образуют совокупность с индексом II + III.

Предварительное рассмотрение массивов показало, что в подъязыках ШАи II имеются отрезки текстов (или элементы вариационного ряда), которые значительно отличаются по количеству от основной массы наблюдений. Причиной появления таких элементов может быть нарушение условий опыта, которое прошло мимо внимания наблюдателя. С другой стороны, возможно, что сильно отклонившееся измерение является элементом данной генеральной совокупности, вероятность появления которого весьма мала. В обоих случаях "выскакивающий" элемент независимо от его происхождения может существенно исказить статистическую оценку математического ожидания и в еще большей степени оценку среднеквадратического отклонения. Изменится также расположение экспериментальной точки и размеры ее доверительной области в плоскости β_2 о β_1. Например, наличие трех "выскакивающих" элементов в выборке ПD приводит к смещению экспериментальной точки и ее доверительной области в правый нижний угол рисунка I, причем доверительная область при заданной доверительной вероятности 0,95 имеет очень большие размеры. Для данного лингвистического материала как правило отбрасывались максимальные по величине элементы. Таким образом, выборка ША' получалась из выборки ША путем отбрасывания одного элемента, выборка D' из выборки D отбрасыванием 3-х элементов, выборка $DL'JTP$ из выборки $DLJTP$ отбрасыванием тоже 3-х элементов. Дисперсия D и среднеквадратическое отклонение S изменились следующим образом:

Таблица 2

	ША	$ША^I$	D	D^I	$DLJTP$	$DLJTP'$
D	2I38	I879	6467	2783	509I	4338
S	46,24	43,35	80,42	52,75	7I,35	65,86

Для решения вопроса о том, считать ли отклоняющиеся элементы принадлежащими данной генеральной совокупности или подлежащими исключению из выборки, использовалось приближенное равенство для вероятности появления выброса (Большев Л.Н., Смирнов Н.В., I965):

$$\rho_1 \approx N\left[0{,}5 - \Phi_0\left(\beta\sqrt{\frac{N}{N-1}}\right)\right],$$

где $\beta = \frac{x_n - m}{\sigma}$ для наибольшего элемента x_n

и $\beta = \frac{m - x_1}{\sigma}$ для наименьшего элемента x_1

(m - математическое ожидание, σ - среднеквадратическое отклонение),

Φ_0 - функция Лапласа.

Если вероятность ρ_1 ниже принятого уровня значимости ρ, то элемент исключался из рассмотрения. В рассматриваемых расчетах задавалось $\rho = 0{,}05$.

Результаты обработки показали, что экспериментальные точки всех выборок (после отбрасывания некоторых "выскакивающих" элементов) сконцентрированы в очень малой области QZQ' (рис. I) плоскости распределений $\beta_2 O \beta_1$. Это свидетельствует об однохарактерности обрабатываемого материала, о принадлежности выборок одноприродному явлению. Это было настолько характерным свойством, что попадание экспериментальной точки в другую часть рисунка наводило на мысль о том, что это "выскакивающий" элемент и такой вывод подтверждался последующим анализом.

Что же представляют собой по содержанию эти "выскакивающие тексты"? Три текста из подъязыка ПД - это объявления, где дается перечень требующихся предприятиям специалистов, так что в известной мере не вполне "естественные тексты". Текст из подъязыка ША представляет собой отрывок из повести Я. Ниедре о П. Стучке. Текст носит публицистический характер, и количество существительных в нем характерно для публицистики (466).

Интерактивная система подбора вида распределения для данных выборок отобрала гипотезы о следующих распределениях: нормальном, лог-нормальном, гамма, бета, Вейбулла. Система также произвела статистическую оценку адекватности подобранных моделей.

Оценка согласия гипотезы с данными производилась по критерию *хи* -квадрат или критерию Пирсона. Для применения этого критерия данные необходимо сгруппировать. Использовано два вида группировок - на равные и на равновероятные интервалы. В первом случае для определения интервалов группировки

отыскивают минимальное и максимальное значения элементов выборки и делят сегмент $[x_{min}, x_{max}]$ на определенное число k равных интервалов.

Количество интервалов k определяется по формуле

$$k = min([1 + 3{,}322 \cdot lg\, N], 20),$$

где $[\ ]$ означает целую часть числа.

Затем отыскивалось:

- количество наблюдений n_j, попавших в j-й интервал;
- вероятность попадания p_j в j-й интервал при условии, что изучаемая случайная величина имеет заданную, проверяемую на согласие, теоретическую функцию распределения;
- значение статистики Пирсона

$$\chi_0^2 = \sum_{j=1}^{k} \frac{(n_j - Np_j)^2}{Np_j} + HIM$$

для данного эмпирического ряда где HIM - слагаемое для модификации статистики Пирсона. Если полученное значение χ_0^2 таково, что $P(\chi^2 \geq \chi_0^2) \geq \beta$, то с доверительным уровнем β гипотеза о согласии с проверяемой теоретической функцией распределения принимается.

Из этой схемы применения критерия Пирсона ясно, что

1) значение χ_0^2 зависит от величины и положения интервалов группировки;
2) необходимо объединение некоторых интервалов на краях;
3) случайная величина χ^2 дает только "интегральную" характеристику отклонений эмпирических частот от теоретических.

Слагаемые входят в формулу для вычисления значения χ^2 симметричным образом, хотя они далеко не равнозначны с точки зрения теории вероятностей. Поэтому надо рассматривать более мелкие интервалы на тех участках, где частота меняется очень быстро. При группировке на равновероятные интервалы разбивать на интервалы нужно не область значений эмпирического ряда, как это делается при группировке на равные интервалы, а область возможных значений изучаемой случайной величины x. Интервалы группировки нужно определить в соответствии с предполагаемым законом распределения. При разбиении на равновероятные интервалы вероятность попадания случайной величины в любой интервал $\Delta_j = (y_{j-1}, y_j)$ будет одинакова и равна

$$p_j = P\{x \in \Delta_j\} = F(y_j) - F(y_{j-1}) = \frac{j}{k} - \frac{j-1}{k} = \frac{1}{k} = p.$$

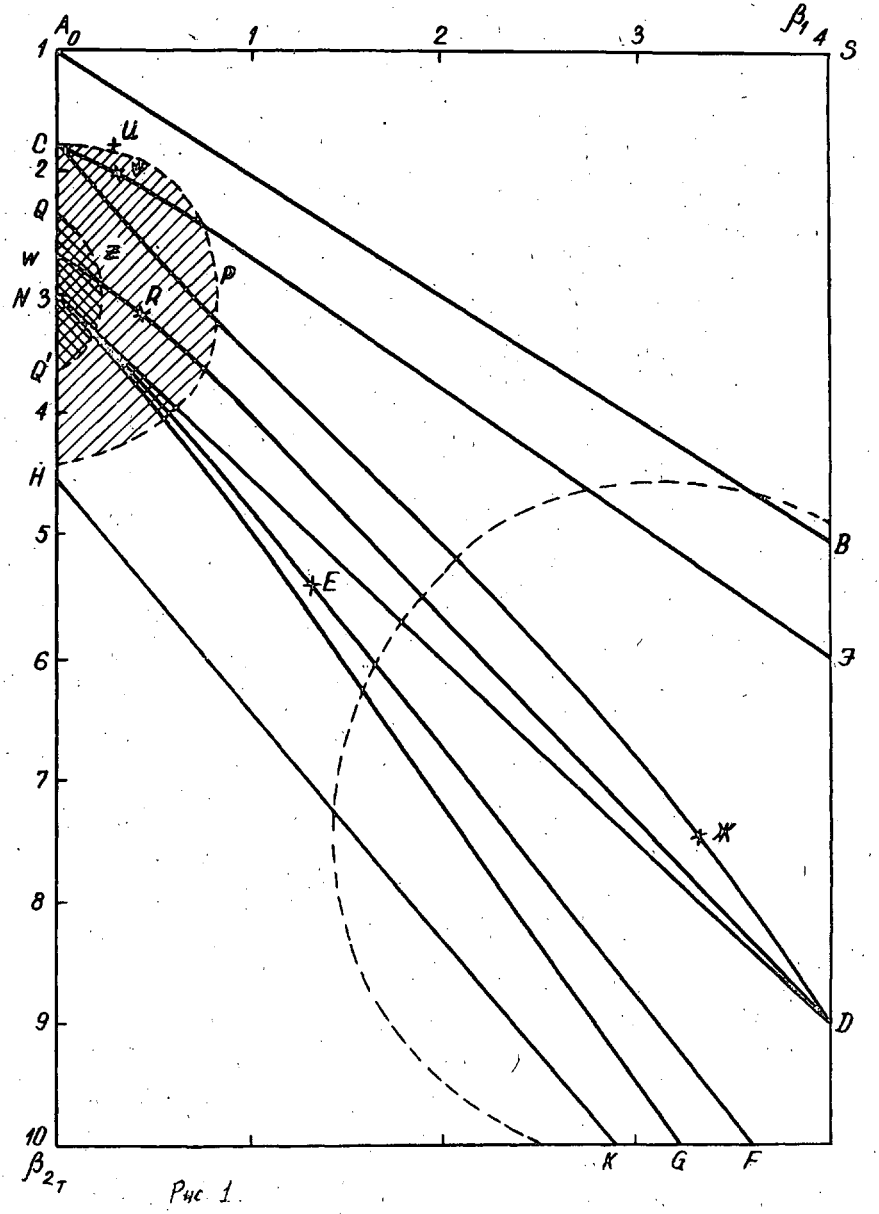

Рис 1.

Рис. I. График определения типа распределения

AB - верхняя граница всех распределений;

ABS - критическая область;

ABJC - область *U* -образного бета-распределения или кривой Пирсона I (*U*) типа;

JCD - область *J* -образного бета распределения или кривой Пирсона I (*J*) типа;

CDN - область бета-распределения или кривой Пирсона I типа;

DNG - область кривой Пирсона УI типа;

NGKH - область кривой Пирсона IУ типа;

ABFN - область семейства S_B Джонсона;

ниже *NF* - область семейства S_U Джонсона.

ND - гамма распределение или кривая Пирсона Ш типа;

NF - логарифмически-нормальное распределение или семейство распределений S_L Джонсона;

NG - кривая Пирсона У типа;

НК- нижняя граница кривых Пирсона;

АС - кривая Пирсона П (*U*) типа;

CN - кривая Пирсона П типа;

NH - кривая Пирсона УП типа;

WD - распределение Вейбулла;

NT - t - распределение или распределение Стьюдента;

Точки для распределений: С-равномерное и параболическое, *E*- экспоненциальное, *N* -нормальное, *R* -Рэлея, Е-экстремальных значений, *U* -треугольное,

V - полутреугольное;

СРН - доверительная область обработанных выборок;

QZQ' - область попадания экспериментальных точек выборок.

Таким образом, при заданном количестве интервалов k устраняется всякий произвол в выборе длины и положения интервалов, группировка эмпирических данных соответствует предполагаемой природе случайной величины.

Границы интервалов y_j находят следующим образом:

$$y_j = \psi\left(\frac{j}{k}\right), \quad j = (1, \ldots, k-1),$$

где $\psi(z)$ - функция, обратная гипотетической функции $F(x)$, выдвинутой в качестве верной.

Результаты вычислений по критерию Пирсона для S приведены в таблицах 3 и 4. Для каждой выборки по десяти распределениям вычислены вероятности согласия $I - P(x^2 > x_0^2)$ гипотезы о предполагаемом виде распределения с экспериментальными данными по критерию Пирсона.

Гипотеза согласия с данным распределением отвергается, если эта вероятность меньше 0,05.

Первое число в клетке - это значение вероятности согласия при группировке на равновероятные интервалы, второе - на равные интервалы. Когда в клетке приведено одно число, то подразумевается группировка на равные интервалы.

Из таблиц видно, что для данного лингвистического материала из рассмотрения можно исключить следующие законы распределения: экспоненциальный, Рэлея, χ^2 -квадрат, Стьюдента и Эрланга, как имеющие очень малые значения вероятностей согласия или даже чисто нулевые.

Как видим, статистическая оценка модели подтвердила правильность предварительного подбора гипотез о видах распределения. На основе таблиц 3 и 4 составлена таблица 5, где для каждой подсовокупности текстов ранжированы и обозначены римскими цифрами по степени адекватности подвергавшиеся проверке виды распределения. Общая их последовательность в убывающем порядке, исходя из вероятности согласия с экспериментальными данными, следующая: закон нормальный, логнормальный, гамма, бета, Вейбулла, Эрланга, Стьюдента.

На описанном этапе эксперимента анализировались каждый из 12 подъязыков, а также комбинированные подсовокупности, составленные из художественных и газетно-журнальных текстов, поскольку проведенные ранее расчеты показали, что художественные тексты более сходны с газетно-журнальными, чем с научно-техническими (Якубайтис Т.А., 1978).

Таблица 3

Распределения \ Выборки	ШA	ШA^{I}	ШB	ШC	ШAB	ШAC	ШBC	ШABC
Нормальное	0.9759	0.9836	0.56I8	0.4742	0.8959	0.9096	0.4I65	0.3372
	0.5672	0.9908	0.9239	0.I424	0.4I08	0.7454	0.I365	0.7I46
Лог-нормальное	0.4033	0.8794	0.5058	0.3I73	0.0608	0.7I85	0.2985	0.0006
	0.6952	0.488I	0.3085	0.3666	0.3049	0.2865	0.I977	0.0
Гамма	0.I295	0.9395	0.4560	0.0275	0.I2I4	0.82I2	0.029I	0.0094
Бета	0.I220	0.6532	0.3584	0.I042	0.5346	0.2462	0.5232	0.5586
Бейбулла	0.I380	0.4427	0.I049	0.0220	0.45I0	0.026I	0.I562	0.8922
	0.0329	0.I5I5	0.2829	0.0294	0.I029	0.0023	0.053I	0.595I
Эрланга	0.0558	0.3580	0.072I	0.0093	0.0006	0.2086	0.0	0.0
Стьюдента	0.0042	0.0I0I	0.0388	0.000I	0.0	0.0	0.0	0.0
	0.0030	0.I062	0.0200	0.0002	0.0	0.0	0.0	0.0
Хи-квадрат	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Рэлея	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Экспоненциальное	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Таблица 4

Распределения \ Выборки	ΠD	$\Pi D'$	$\Pi\chi$	ΠJ	ΠT	ΠP	$\Pi D\chi J T D$	$\Pi D\chi J T D'$	Ш + П
Нормальное	0.0407	0.7I62	0.849I	0.90I2	0.9029	0.9832	0.I966	0.8853	0.5I65
	0.0050	0.3I3I	0.92I5	0.0722	0.9996	0.9258	0.32I0	0.9409	0.I090
Лог-нормальное	0.2407	0.90I2	0.0550	0.3340	0.7596	0.4354	0.2859	0.6I34	0.0002
	0.I97I	0.4752	0.I368	0.0735	0.2968	0.8836	0.8472	0.2373	0.0
Гамма	0.00II	0.I679	0.4548	0.0I33	0.9703	0.5I76	0.3638	0.8236	0.0003
Бета	0.0004	0.2467	0.5498	0.0I03	0.8924	0.2064	0.04I4	0.4467	0.I288
Вейбулла	0.0	0.2054	0.36I8	0.6594	0.5724	0.2255	0.0	0.0454	0.4642
	0.0	0.0494	0.7654	0.0I23	0.7I46	0.404I	0.0	0.005I	0.0360
Эрланга	0.0009	0.0874	0.2075	0.0056	0.8I42	0.4095	0.0349	0.I430	0.0
Стьюдента	0.0	0.I953	0.23II	0.I468	0.I6I6	0.3600	0.0	0.0	0.0
	0.0	0.2392	0.2I35	0.0070	0.5892	0.4I25	0.0	0.0	0.0
Хи-квадрат	0.0	0.0005	0.0	0.0	0.0039	0.000I	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Рэлея	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Экспоненциальное	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Таблица 5

Распределения \ Выборки	ША	ША^{I}	ШВ	ШС	ШАВ	ШАС	ШВС	ШАВС	ПD	ПD′	Пℒ	ПЈ	ПТ	ПР	ПРℒЈТDℒЈТD′	П+Ш
Нормальное	I	I	I	I	I	I	П	П	П	П	I	I	I	П	Ш	I	I
Лог-нормальное	П	Ш	П	П	IУ	Ш	Ш	У	I	I	УП	Ш	У	I	I	Ш	У
Гамма	IУ	П	Ш	У	У	П	У	IУ	Ш	УI	IУ	У	П	Ш	П	П	IУ
Бета	У	IУ	IУ	Ш	П	IУ	I	Ш	У	Ш	Ш	УI	Ш	УП	IУ	IУ	Ш
Вейбулла	Ш	У	У	IУ	Ш	УI	IУ	I	УI	У	П	П	УI	УI	УI	УI	П
Эрланга	УI	УI	УI	УI	УI	У	УI	УI	IУ	УП	УI	УП	IУ	У	У	У	УI
Стьюдента	УП	УП	УП	УП	УП	УП	УП	УП	УП	IУ	У	IУ	УП	IУ	УП	УП	УП

На следующем этапе эксперимента были проанализированы смешанные подсовокупности, составленные из более контрастных текстов - художественных и научно-технических (Ш + IУ), а также вся выборочная совокупность текстов в целом.

На рис. 2 можем видеть, что смешанная подсовокупность Ш + IУ более ассиметрична, чем любой из подъязыков (на рис.3 показано распределение S в подъязыке ШВ). Поэтому можно заранее предположить, что как подсовокупность Ш + IУ, так и вся выборочная совокупность в целом (рис. 4) не дадут хорошего согласия с нормальной кривой.

Действительно, предварительный подбор вида распределения показывает, что в обеих совокупностях распределение может быть описано семейством S_B Джонсона, кривыми Пирсона И и П типа, а для подсовокупности Ш + IУ также бета-распределением.

Эти распределения не связаны с какой-либо определенной моделью события, но они обладают такими ценными свойствами, как возможностью апроксимировать очень разнообразные распределения вероятностей и способностью легко производить трансформацию в хорошо известные теоретические распределения.

Для алгоритмизации более удобными являются распределения Джонсона.

Плотность распределения для семейства S_B Джонсона выражается формулой

$$f(x)=\frac{\varrho}{\sqrt{2\pi}}\cdot\frac{\lambda}{(x-\varepsilon)(\lambda-x-\varepsilon)}\exp\left\{-\frac{1}{2}\left[\gamma+\varrho\ln\left(\frac{x-\varepsilon}{\lambda-x+\varepsilon}\right)\right]^2\right\},$$

где $\varrho, \gamma, \lambda, \varepsilon$ - параметры распределения;

$$\varrho>0;\quad -\infty<\gamma<\infty;\ \lambda>0;\ -\infty<\varepsilon<\infty.$$

Несмотря на внешнюю сложность приведенной формулы это распределение легко трансформируется в нормированное нормальное распределение с функцией плотности

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right).$$

Формула перехода для S_B распределения

$$z=\gamma+\varrho\ln\left(\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right).$$

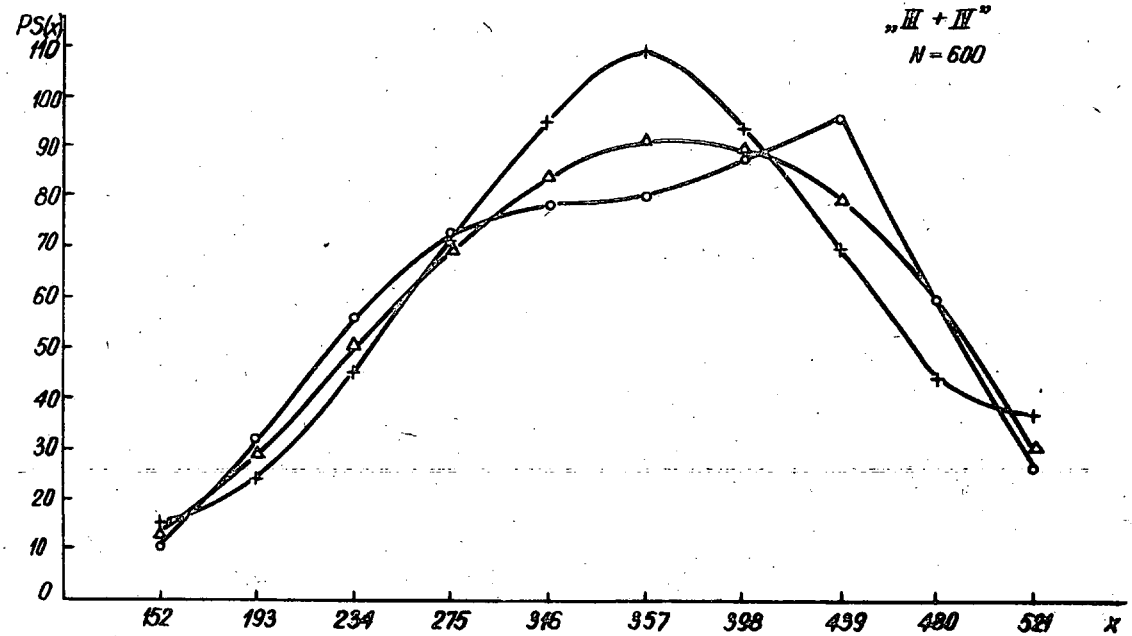

Рис. 2. Эмпирические и теоретические гистограммы

эмпирическая гистограмма;

теоретическая гистограмма нормального распределения;

$x^2 = 35,02 > x^2_{0,05} = 16,92$;

теоретическая гистограмма распределения S_B Джонсона;

$x^2 = 7,591 < x^2_{0,01} = 15,09$.

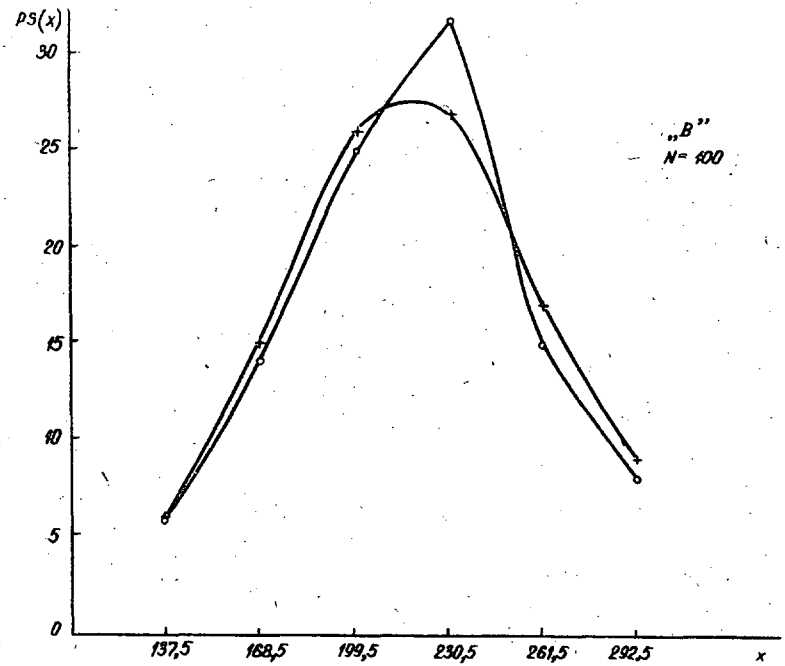

Рис. 3. Эмпирическая и теоретическая гистограммы

—o—o— эмпирическая гистограмма;

—+—+— теоретическая гистограмма нормального распределения

$\chi^2 = 1,403 < \chi^2_{0,05} = 7,815.$

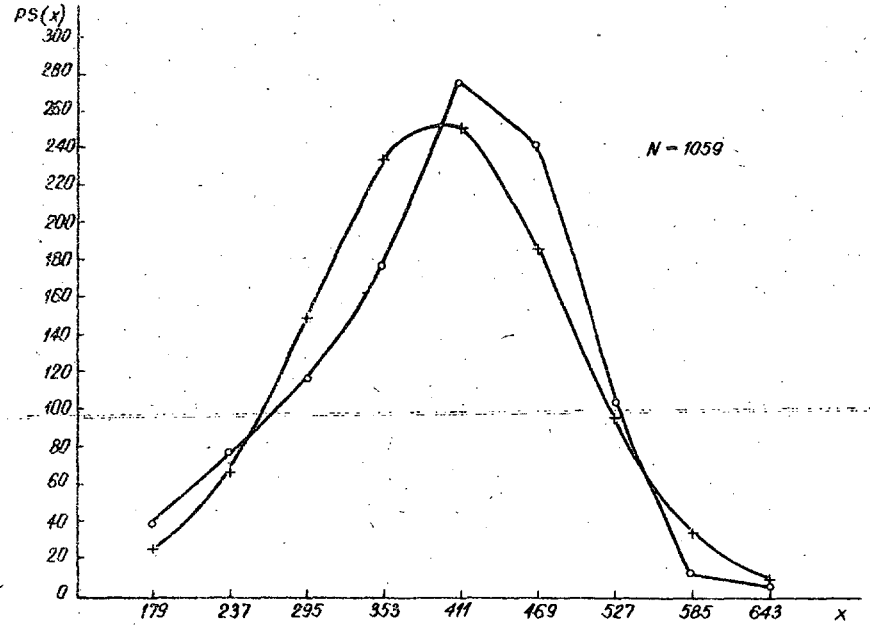

Рис. 4. Эмпирическая и теоретическая гистограммы

—o—o— Эмпирическая гистограмма;

—+—+— Теоретическая гистограмма нормального распределения;

$\chi^2 = 66{,}04 > \chi^2_{0{,}05} = 15{,}51$.

Статистическая оценка адекватности подобранной модели по критерию χ^2 -квадрат для выборки III + IУ дала вероятность согласия гипотезы о бета распределении с данными эксперимента равной 0,3659, что означает, что гипотеза не отвергается.

По семейству S_B Джонсона для этой же выборки эмпирическое значение χ^2 -квадрат равно 7,591, что меньше каждого из следующих теоретических значений χ^2 -квадрат: 9,236; II,07 и I5,09 для соответствующих доверительных вероятностей 0,90; 0,95 и 0,99.

Значит, с такими доверительными вероятностями гипотеза о согласии данных эксперимента с распределением S_B Джонсона не отвергается. При этом получены следующие значения параметров распределения S_B Джонсона:

$\lambda = 513{,}2$; $\eta = 1{,}189$; $\gamma = -0{,}211$; $\varepsilon = 81{,}90$,

где λ - ширина интервала изменения случайной величины,

η, γ - параметры формы распределения,

ε - нижняя граница случайной величины.

Проведенные исследования, разумеется, могут быть углублены и уточнены. Но настоящая работа ставила своей целью показать, что использование вычислительной техники, автоматизация процесса статистического анализа позволяют лингвисту проводить анализ на больших объемах текстов с применением достаточного широкого аппарата математической статистики, что, в свою очередь, дает возможность глубже проникнуть в природу статистических закономерностей построения текста.

Л И Т Е Р А Т У Р А

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965. - 464 с.

Городецкий Б.Ю. Лексико-статистическая инвентаризация комплекса подъязыков. - В кн.: Проблемы теоретической и экспериментальной лингвистики. - М.: Изд-во МГУ, 1977, с.21-42.

Дружинин Н.К. Логика оценки статистических гипотез. - М.: Статистика, 1973. - 345 с.

Милюна М.К. Определение доверительной области в плоскости распределений. - В кн.: Проблемы случайного поиска, 1978, т. 6, с. 325-328.

Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики. - М.: Наука, 1969. - 511 с.

Хан Г., Шапиро С. Статистические модели в инженерных задачах. - М.: Мир, 1969, - 395 с.

Якубайтис Т.А. Использование ЭВМ в лингвистических исследованиях. - Вопросы языкознания, 1979, № 3, с. 127-131.

Якубайтис Т.А. О стабильности частот и однородности текстовых совокупностей. - Известия АН Латвийской ССР, 1978, № 6, с. 69-83.

M. Miļūna, H. Buŝs, K. Buŝs. Sadalījuma funkciju analīze mežsaimnieciskos pētījumos.— Jaunākais mežsaimniecība, 17. - Rīgā: Zinātne, 1975, 85-97 lpp.

AN AUTOMATIC SYSTEM FOR SELECTING THE TYPE OF DISTRIBUTION OF LINGUISTIC UNITS

M. Miļūna, T. Yakubaitis

S u m m a r y

The use of advanced computer techniques provides good opportunities for the automation of linguistic and statistical investigations. This concerns not only the possibilities of improving the I/O methods, correcting the information, but also the automation of the analysis process itself, in particular the statistical analysis.

At the Institute of Electronics and Computer Science of the Latvian SSR Academy of Sciences an interactive system for the optimal choice of hypotheses of statistical distributions is being developed. The paper describes the operation of the system in selecting the type of distribution of parts of speech in Latvian language texts containing about one million word uses.

ПАРАБОЛИЧЕСКИЕ И ГИПЕРБОЛИЧЕСКИЕ ЗАКОНОМЕРНОСТИ ТЕКСТООБРАЗОВАНИЯ

В.Е. Остапенко

Ориентация на максимальную объективность знания языка как семиотической системы, на достижение его полной инвариантности по отношению к множеству познающих субъектов породила усиливающуюся тенденцию математизации лингвистики. Известная мысль Маркса о том, что уровень развития конкретной области науки определяется степенью математизации этой области, связана в первую очередь с тем, что языком математики можно более или менее точно описывать закономерности, открываемые в изучаемой действительности (Шрейдер, 1978).

Можно констатировать, что математизация языкознания развивается в двух направлениях. Первое из них характеризуется тем, что к исследованию одной и той же закономерности привлекается широко разветвленный математический аппарат. Например, распределение лексических единиц в тексте исследуется при помощи нормального, логнормального и распределения Пуассона (Пиотровский, 1975). Эта же проблема до известной степени решается аппроксимацией эмпирических вариационных рядов к стохастическим кривым Пирсона (Каширина, 1974). Другая закономерность текстообразования - динамика роста словаря исследуется с помощью уравнения параболы (Белоногов, Богатырев, 1973), формул В.М. Калинина и Ю.К. Орлова (Орлов, 1976).

Сущность второго направления состоит в том, что его представители пытаются найти одну формулу, удовлетворительно описывающую все основные закономерности текстообразования, т.е.

а) динамику роста словаря;

б) соотношение долей лексических единиц в тексте определенного объема;

в) соотношение объемов групп лексических единиц, имеющих частоту 1,2 ...

г) распределение лексических единиц в связном тексте.

Для первых трех пунктов эта задача с известной степенью точности решается в работе "Обобщенный закон Ципфа - Мандельброта и частотные структуры информационных единиц раз-

личных уровней" (Орлов, 1976) и поэтому было бы полезным найти более общую формулу, приемлемую для исследования всех вышеперечисленных закономерностей. Такая попытка была сделана автором данной статьи в одной из ранних работ (Остапенко, 1978) и теперь дается более полное и развернутое объяснение выдвинутой гипотезы о возможности описания единой аналитической зависимостью основных закономерностей текстообразования.

Если эмпирические вариационные ряды изобразить графически, то вышеуказанные закономерности текстообразования предстанут в виде следующих монотонно возрастающих и убывающих функций (см. рис. I-4).

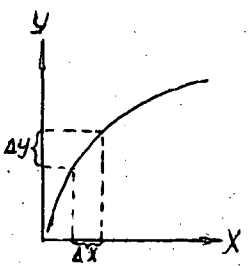

рис.I.

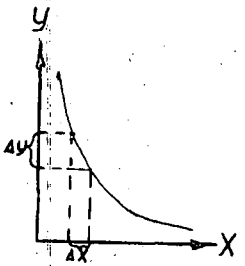

рис.2.

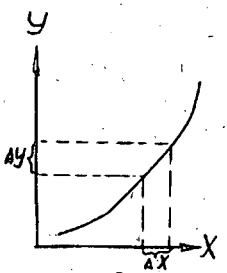

рис.3.

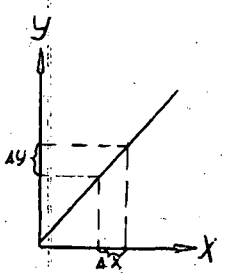

рис.4.

Для того, чтобы найти наиболее общую формулу, описывающую все интересующие нас случаи, полезно отвлечься от лингвистической сущности моделируемых закономерностей и попытаться формально определить конкретный вид аналитической зависимости. Это можно сделать при помощи обычного математического приема - исходных гипотез и дифференциальных уравнений.

I. Пусть приращение функции Δy в какой-либо момент (положительное или отрицательное - безразлично) линейно зависит

от значения этой функции в данный момент y и пропорционально приращению аргумента ΔX, т.е.

$$\Delta y = k y \Delta x \quad ; \quad \frac{\Delta y}{\Delta x} = k y .$$

Поскольку отношение $\frac{\Delta y}{\Delta x}$ при $\Delta X \longrightarrow 0$ по определению (Пискунов, 1968) есть производная функция, то, сделав соответствующие преобразования, переходим к дифференциальному уравнению, решение которого приводит к конечной формуле, соответствующей исходной гипотезе.

$$\int \frac{dy}{y} = \int k\,dx \; ; \quad \ln y = kx + C$$

$$y = C e^{kx} \qquad (I)$$

2. Пусть теперь приращение функции Δy в какой-либо момент нелинейно зависит от значения этой функции и пропорционально приращению аргумента ΔX, т.е.

$$\Delta y = k y^a \Delta x ; \quad a \neq 1 .$$

Тогда аналогично предыдущему (см. п. I) получим.

$$\int \frac{dy}{y^a} = \int k\,dx \; ; \; \frac{y^{1-a}}{1-a} = kx + C \; ; \; y = k(a-1)^{\frac{1}{1-a}} \left(x - \frac{C}{k}\right)^{\frac{1}{1-a}}$$

$$k(a-1)^{\frac{1}{1-a}} = K \; ; \; \frac{1}{1-a} = \gamma \; ; \; -\frac{C}{k} = \rho$$

$$y = K(x+\rho)^{\gamma} \qquad (2)$$

3. Пусть приращение функции Δy линейно зависит от значения аргумента и пропорционально приращению аргумента ΔX, т.е. $\Delta y = k X \Delta X$.

Отсюда получаем следующее

$$\int dy = \int kx \cdot dx \; ; \quad y = \frac{k}{2} x^2 + C \; ; \quad \frac{k}{2} = K$$

$$y = Kx^2 + C \qquad (3)$$

4. Пусть приращение функции Δy нелинейно зависит от значения аргумента X и пропорционально его приращению ΔX, т.е.

$$\Delta y = k x^b \Delta x \; ; \quad b \neq 1 .$$

Далее получаем следующее

$$\int dy = \int k x^b dx \; ; \quad y = \frac{k}{b+1} x^{b+1} + C \; ; \quad \frac{k}{b+1} = K \; ; \quad b+1 = \gamma$$

$$y = K x^{\gamma} + C \qquad (4)$$

5. Пусть теперь приращение функции Δy линейно зависит от y и нелинейно x, т.е.

$$\Delta y = k y x^b \Delta x \; , \quad b \neq 1$$

Отсюда следует

$$\int \frac{dy}{y} = \int k x^b dx \; ; \quad \ln y = \frac{k}{1+b} x^{1+b} + C \; ; \quad \frac{k}{1+b} = K \; ; \quad 1+b = \gamma$$

$$y = C e^{k x^{\gamma}} \qquad (5)$$

6. Положим, приращение Δy нелинейно зависит от y и линейно от x, т.е.

$$\Delta y = k y^a x \Delta x \; ; \quad a \neq 1$$

Тогда получим

$$\int \frac{dy}{y^a} = \int k x \, dx \; ; \quad \frac{y^{1-a}}{1-a} = k x^2 + C \; ; \quad y = \left[k(1-a) x^2 + (1-a) c \right]^{\frac{1}{1-a}}$$

$$k(1-a) = K \; ; \quad (1-a) \cdot c = p$$

$$y = (k x^2 + p)^{\gamma} \qquad (6)$$

7. Рассмотрим самый общий случай, когда приращение функции Δy нелинейно зависит от значения функции и аргумента. Причем эту гипотезу можно задать как в виде произведения, так и в виде суммы зависимостей, т.е.

$$\Delta y = k y^a x^b \Delta x \; ; \quad \Delta y = k \left(y^a + x^b \right) \Delta x \; ; \quad a \neq 1 \; ; \; b \neq 1$$

Тогда в случае произведения предполагаемых зависимостей имеем

$$\int \frac{dy}{y^a} = \int k x^b dx ; \frac{y^{1-a}}{1-a} = \frac{k}{1+b} x^{1+b} + C ; y = \left[\frac{k(1-a)}{1+b} x^{1+b} + (1-a)c\right]^{\frac{1}{1-a}}$$

$$\frac{k(1-a)}{1+b} = K ; \quad 1+b = \gamma ; \quad (1-a)\cdot c = C ; \quad \frac{1}{1-a} = n$$

$$y = (Kx^{\gamma} + C)^n \qquad (7)$$

8. Очевидно, сумма зависимостей может быть представлена в виде

$$\Delta y = k y^a \Delta x + k x^b \Delta x .$$

Тогда, опуская промежуточные преобразования, аналогичные п.п. 2, 4 получим

$$y = k (x+p)^{\gamma} + k x^{\gamma} + C .$$

Легко увидеть, что первые два члена дублируют друг друга, так как представляют собой идентичные кривые, смещенные одна относительно другой на величину p . Объединение этих членов дает конечную формулу

$$y = k (x+p)^{\gamma} + C \qquad (8)$$

Сопоставим теперь исходные гипотезы и соответствующие им аналитические зависимости (см. табл. I).

Таблица I

Аналитические зависимости, описывающие процесс текстообразования

№	Исходные гипотезы	Конечные зависимости
I	$\Delta y = k y \Delta x$	$y = C e^{kx}$
2	$\Delta y = k y^a \Delta x$	$y = k (x+p)^{\gamma}$
3	$\Delta y = k x \Delta x$	$y = k x^2 + C$
4	$\Delta y = k x^b \Delta x$	$y = k x^{\gamma} + C$
5	$\Delta y = k y x^b \Delta x$	$y = C e^{k x^{\gamma}}$
6	$\Delta y = k y^a x \Delta x$	$y = (k x^{\gamma} + p)^{\gamma}$
7	$\Delta y = k y^a x^b \Delta x$	$y = (k x^{\gamma} + C)^n$
8	$\Delta y = k (y^a + x^b) \Delta x$	$y = k (x+p)^{\gamma} + C$

Очевидно, (7) и (8) как угодно близко могут приближаться к любому другому виду (I) - (6), так как a и b принимают любые значения, вплоть до $a = I$, $a = 0$, $b = I$, $b = 0$, что автоматически обращает их в один из частных случаев, вплоть до уравнения прямой.

Таким образом, анализ исходных гипотез позволяет выбрать аналитическую зависимость, приемлемую для решения всех вышеназванных лингвистических проблем, а также тех, которые могут быть поставлены в будущем, если получаемый эмпирический ряд представится в виде монотонно возрастающего или убывающего процесса.

Формула (7) пока неизвестна в практике лингвостатистических исследований и поэтому ее возможности еще предстоит изучить экспериментально. Напротив, (8) в сокращенном виде ($C = 0$) широко известна как закон Эсту - Ципфа - Мандельброта. Очевидно, унифицированное описание закономерностей текстообразования позволит создать целостное системное представление о статистическом механизме, который представляет собой варьирование одной и той же зависимости в виде ряда парабол, гипербол и прямых линий.

Алгоритм вычисления параметров аналитической зависимости (8) представляет собой последовательность следующих операций.

$$y = k x^{\gamma}; \quad \ln y = \ln k + \gamma \ln x; \quad \ln k = c;$$

$$S = \sum_{i=1}^{n} \left\{ \left[\ln y - c - \gamma \ln x \right]^2 \right\}' \longrightarrow \min$$

$$k = \exp \frac{\left(\sum_{i=1}^{n} \ln y_i \right) \left(\sum_{i=1}^{n} \ln^2 x_i \right) - \left(\sum_{i=1}^{n} \ln y_i \ln x_i \right) \left(\sum_{i=1}^{n} \ln x_i \right)}{n \sum_{i=1}^{n} \ln^2 x_i - \left(\sum_{i=1}^{n} \ln x_i \right)^2}$$

$$\gamma = \frac{n \sum_{i=1}^{n} \ln x_i \ln y_i - \left(\sum_{i=1}^{n} \ln x_i \right) \left(\sum_{i=1}^{n} \ln y_i \right)}{n \sum_{i=1}^{n} \ln^2 x_i - \left(\sum_{i=1}^{n} \ln x_i \right)^2}$$

$$\rho = exp\ \frac{\ln y_{imin} - \ln k}{\gamma} - x_{imin}$$

$$C = k(x_{imax} + \rho)^{\gamma} - y_{imax}$$

ЛИТЕРАТУРА

Белоногов Г.Г., Богатырев В.И. Автоматизированные информационные системы. М., 1973.

Каширина М.Е. О типах распределения лексических единиц в тексте. - В кн.: Статистика речи и автоматический анализ текста. Л., 1973.

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

Остапенко В.Е. Модели лингвостатистики. - Школа-семинар по прикладной и инженерной лингвистике. Тезисы докладов. Махачкала, 1978, с. 33.

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

Пискунов Н.С. Дифференциальное и интегральное исчисления. М., 1968.

Шрейдер Ю.А. Гуманитаризация знания и управление информационной средой. - Вестник АН СССР № 9, 1978.

PARABOLIC AND HYPERBOLIC REGULARITIES OF TEXT FORMATION

Vladimir Ostapenko

S u m m a r y

The article can be considered as an attempt to compare the initial assumptions and the appropriate final formulae of showing that all text formation regularities representing a monotonously ascending and descending process may be described by two analytical dependences which at the corresponding values of the exponent of the power may turn into a parabola, a hyperbola or a straight line.

К ВОПРОСУ ОБ АНАЛИТИЧЕСКОМ ВЫРАЖЕНИИ СВЯЗИ МЕЖДУ ОБЪЕМОМ СЛОВАРЯ И ОБЪЕМОМ ТЕКСТА

Юхан Тулдава

В статье обсуждаются теоретические и практические возможности разных подходов к решению проблемы аналитического выражения зависимости между объемами словаря и текста. Рассматривается семейство формул, построенных на основе некоторых теоретических предпосылок о вероятностном процессе порождения речи и позволяющих с той или другой степенью достоверности прогнозировать объем словаря в зависимости от увеличения объема текста.

I. Постановка вопроса

Вопрос о количественных отношениях между объемом словаря и объемом текста относится к проблемам, имеющим как теоретическое, так и прикладное значение в лингвостатистике. Если удастся смоделировать процесс нарастания объема словаря в зависимости от увеличения объема текста в виде некой функции, имеющей определенный содержательный смысл, то это может не только расширить наши знания о наиболее общих квантитативных закономерностях порождения текста, но и позволит решить ряд интересных и актуальных задач прикладного характера. На основе функции, выражающей зависимость объема словаря (V) от объема текста (N), можно, например, находить неизвестное значение V по данному N, или наоборот, а также определить степень насыщения или достаточность объема выборки при проектировании автоматизированных информационных систем (см. Белоногов Г.Г., Новоселов А.П.,1971; Горькова В.И., 1972; Пиотровский Р.Г., 1975, с. 95-98). Следует отметить, что в связи с успехами компьютерной лингвистики и новыми потребностями практики становится все более актуальным изучение квантитативных закономерностей крупных текстовых массивов, в частности, в целях прогноза объема словаря далеко за пределами предварительных опытных данных. Установление формы связи между объемом словаря и объемом текста позволяет также исследовать стилостатистические особенности индивидуальных текстов или жанров /Тулдава Ю., 1974 и 1977/ и содействует

решению некоторых педагогических и психологических задач /см. например, Захарова А.В., 1967; Остапенко В.Е., 1979/.

Имеются многочисленные попытки построения эмпирических формул для выражения зависимости объема словаря от объема текста. Первые подобные формулы появились еще в 40-х и 50-х годах, например, формулы Й. Чотлоса, В. Курашкевича, П. Гиро /Chotlos J.W., 1944; Kuraszkiewicz W., 1958; Guiraud P., 1959/. П. Гиро и В. Курашкевич исходят из предположения о линейной связи между V и $\sqrt{N}$. По этому же принципу построены формулы А. Захаровой /1967/ и Т. Ковлер /1967/. Зависимость между V и $\sqrt{N}$ может действительно наблюдаться, но лишь в отдельных отрезках текста. Другие исследователи, например, Г.Хердан /Herdan G., 1966/, использовали формулу степенной функции типа $V = aN^b$ (a и b -константы). Эта формула предполагает линейную связь между $\log V$ и $\log N$, но в действительности такая зависимость наблюдается не всегда (см. ниже п. 4). Х.Сомерс /Somers H., 1959/ предлагает формулу $\log V = a(\log N)^b$, исходя из предположения о линеиной связи между $\log\log V$ и $\log\log N$. И эта формула имеет только ограниченную сферу применения /см. Тулдава Ю., 1974, с. 13/. Несколько лучших результатов добился В. Мюллер /Müller W., 1971/, применявший формулу $\sqrt{\log N/V} = a \log N$ при анализе текстов немецкого языка. Можно еще отметить попытку В. Горьковой /1972/ аппроксимировать рост ключевых слов в тексте экспоненциальной функцией, причем это удалось сделать только "скользящей" экспонентой, т.е. рядом экспонент, у которых параметры постоянно изменяются. Для решения некоторых стилостатистических задач автором этих строк была предложена дробно-линейная функция типа $V = \frac{aN}{N+b}$, которую можно использовать при исследовании тенденции роста словаря в малых выборках /Тулдава Ю., 1974 и 1977/.

Наряду с применением чисто эмпирических формул были попытки смоделировать процесс нарастания словаря, исходя из определенных теоретических предпосылок, например, основываясь на предположении о логнормальном распределении слов в тексте /Carroll J.B., 1967/ или о действии закона Ципфа /Калинин В.М., 1964; Орлов Ю.К., 1976/. Основой для выведения формул роста словаря использовались и другие известные распределения, например, распределение Вейбулла /Нешитой В.В., 1975/.

Нашей задачей является выяснение вопроса о том, можно ли установить какую-нибудь общую закономерность порождения тек-

ста, которую можно было бы использовать для выведения соответствующих формул зависимости между объемом словаря и объемом текста. Практической стороной этой задачи было бы выявление возможностей прогноза роста словаря вне диапазона наблюдений. В этой связи возникает вопрос, следует ли искать решения проблемы в "структурном", в смысле комплексном, подходе /Орлов Ю.К., 1978, Статистическое моделирование.../, т.е. когда рост словаря рассматривается в зависимости от некоторых других сторон статистической структуры текста, или целесообразнее осуществить "прямой" подход, при котором взаимосвязь между объемами словаря и текста рассматривается как исходный момент анализа на фоне некоторых внутренних факторов порождения речи.

2. Комплексный ("структурный") подход

Важным событием в квантитативной лингвистике было появление работ В.М. Калинина /1964 и 1965/, в которых была в принципе решена задача построения комплексной модели статистической структуры текста. В.М. Калинину удалось соединить в одну общую систему собственно частотную структуру (распределение Ципфа), лексический спектр (распределение Юла) и закономерность количественного роста словаря в зависимости от увеличения объема текста. Открытие имело и имеет еще в наши дни большое теоретическое значение. Однако ввиду сильной идеализации структуры реального текста (допущение о случайности и независимости появления слов в тексте и жесткий характер взаимосвязей разных квантитативных аспектов текста) эту модель редко удавалось применять на практике при лингвостатистическом анализе текста /один такой пример см. Minor J., 1971/.

Среди попыток усовершенствования комплексной модели структуры текста следует отметить работы Ю.К. Орлова (начиная с 1969 г.), создавшего оригинальную концепцию "обобщенного закона Ципфа-Мандельброта". Ю.К. Орлов предлагает более гибкий способ анализа текста в едином комплексе. Он выдвигает принцип т.н. "объема Ципфа" (который обозначается символом Z). Объем Ципфа является своего рода эталоном и точкой отсчета при анализе распределения лексики в конкретном тексте. В терминах системного анализа объем Ципфа может быть рассмотрен как некое идеальное, наиболее "благоприятное" сос-

тояние, к которому система (текст) стремится. Различные параметры статистической структуры текста определяются как бы в понятиях удаленности от состояния Z. Феномен Z Ю.К. Орлов связывает формально с условием выполнения закона Ципфа-Мандельброта, а содержательно - с понятием "целостности" текста /в отношении литературных текстов это понятие связывается с художественной полноценностью произведения, см. Орлов Ю.К., 1970/.

Свою концепцию Ю.К. Орлов и его последователи пытаются реализовать и в практических лингвостатистических исследованиях, приводя множество примеров, которые должны показать соответствие теоретических установок фактическим свойствам реальных текстов /см., например, Орлов Ю.К.,1976 и 1978; Надарейшвили И.Ш., Орлов Ю.К., 1978; Надарейшвили И.Ш.,1978/. Следует сказать, что в общих чертах концепция Ю.К. Орлова удовлетворительно согласуется с действительностью. Однако при детальном рассмотрении оказывается, что вычисления по формулам в рамках обобщенного закона Ципфа-Мандельброта не всегда дают достаточно точные и достаточно достоверные результаты, особенно когда требуется более точный стилостатистический анализ текстов при их классификации или сравнении*. Нет уверенности и в прогнозе роста словаря за пределами опытных данных. Нарастание словаря определяется в модели Ю.К. Орлова взаимодействием величин N, V и p_1 (относительная частота самого частого слова в тексте), причем требуется знать только объем словаря (V) и объем текста (N) в одной точке, чтобы прогнозировать объем словаря "вперед" или "назад". Можно привести два принципиальных возражения против такого подхода. Во-первых, любая единственная точка в потоке порождения текста вряд ли может достоверно сигнализировать о форме и тенденции роста словаря (бывают, например, тексты, в которых нарастание объема словаря идет плавно и медленно, в то время как в других текстах рост словаря может быть стремителен в начальной стадии, а потом быстро утихать). Во-вторых, частота первого слова в частотном списке (p_1) не может достоверно определить структуру текста даже во взаимосвязи с объемом

* Более подробный критический анализ возможностей практического приложения концепции Ю.К. Орлова содержится в рецензии автора "Феномен Z и связанные с ним проблемы в лингвостатистике" (депонируется в Группе прикладной лингвистики ТГУ). Некоторые примеры приводятся и в настоящей статье.

словаря и объемом текста. По модели Ю.К. Орлова получается, например, что при сравнении текстов с равными N и V прогнозируется более богатая лексика у того текста, у которого p_1 больше. В реальных текстах p_1 (и вообще начальная часть частотного списка) или вообще не коррелирует с богатством лексики /как, например, по данным эстонского языка, см. Тулдава Ю., 1977, с. 171/, или, наоборот, во многих случаях может быть признаком относительной бедности лексики (ввиду концентрации лексики в начале частотного списка). Отсюда следует, что техника вычисления роста словаря в рассматриваемой модели поставлена неудовлетворительно. Преимущество "единственной точки" и тем самым экономность подхода иллюзорны, а привлечение величины p_1 - излишняя работа (требуется составление частотного списка), особенно в тех случаях, когда целью анализа является лишь выяснение тенденции роста словаря в данном тексте*.

Ю.К. Орлов называет свой подход "структурным", так как при выведении формул роста словаря учитываются и другие аспекты статистической структуры текста, и утверждает, что только структурные (комплексные) формулы могут дать надежные результаты при прогнозе и при стилостатистическом анализе текстов /см. Орлов Ю.К., 1978, Статистическое моделирование.../. Известно, что в математических моделях обычно стремятся отразить как можно больше процессов в их связях и взаимозависимостях. Но один из специалистов по методологии математики пишет: "Однако еще не удалось построить модель, гармонично отражающую совокупность совместно протекающих процессов. Почти всегда выделяется для подробного отражения лишь один из процессов, об остальных используют лишь упрощенные данные" /Рыбников К.А., 1979, с. 110/. Не следует, по-видимому, преувеличивать роль комплексных математических моделей (особенно таких, которые построены по схеме жесткой детерминации) при решении практических лингвостатистических задач. Значимость таких моделей в другом: они дают общую картину взаимных внутренних связей в статистической организации текста и указывают на некоторые предельные состояния системы, которые могут служить отправной точкой и эталоном

* Относительная частота самого частого слова обычно стабильна в смешанных текстах данного подъязыка, но индивидуальные отклонения могут тем не менее быть большими. Поэтому, нельзя использовать среднее значение p_1 данного жанра при анализе индивидуальных текстов.

при исследовании текстов. В этом смысле концепции В.М. Калинина и Ю.К. Орлова следует считать ценным вкладом в теорию лингвостатистики.

3. "Прямой" метод

Мы исходим из предположения, что рост словаря в зависимости от объема текста можно исследовать и с некоторых других позиций, а именно с точки зрения имманентных свойств данного явления. Связь словаря и текста предстает в таком случае как исходный, первичный аспект статистической организации текста. Квантитативную зависимость между словарем и текстом следует при таком подходе изучать в связи с некоторыми наиболее общими факторами порождения текста, а не как производственный аспект в сравнении с такими сторонами статистической структуры текста, как лексический спектр или распределение слов по порядку убывания частотности. Такой конкретный подход представляется нам и практически многообещающим, так как позволяет экономнее и, по-видимому, точнее определить параметры текста при изучении конкретной проблемы зависимости объема словаря от объема текста. Мы основываемся при этом на предпосылках теории вероятностных систем, согласно которым исследуемый объект характеризуется не только случайностью параметров (низший уровень организации), но и определенной устойчивостью и регулярностью в массе случайных событий (высший уровень организации). Упорядоченность, структура во множестве событий выражается в вероятностных системах через распределения (функции), которые "характеризуют знаки низшего кода обобщенным, интегральным образом" /Сачков Ю.В., 1971, с. 133/. Следовательно, в вероятностных системах не исключается детерминированность связей, но она переносится на высший, обобщенный уровень организации. Упорядоченность и организованность в системе обусловлены некоторыми скрытыми от внешнего наблюдения глубинными причинами, которые и следует при возможности выявить.

В данном конкретном случае нарастание объема словаря в ходе порождения текста можно представить себе как вероятностный процесс, при котором на каждом шагу по ходу порождения текста осуществляется выбор между "новым", ранее не появившимся словом, и "старым", уже употребленным в данном тексте словом. Эмпирически установлено, что с увеличением текста

вероятность выбора "нового" слова постоянно уменьшается. Отсюда можно заключить, что рассматриваемый вероятностный процесс подчиняется какой-то определенной глубинной закономерности порождения текста. Содержательно процесс нарастания словаря определяется сложным взаимодействием двух противоположных тенденций: стремлением к расширению и стремлением к ограничению словарного состава данного текста. С одной стороны, в процессе говорения (писания) говорящим (пишущим) овладевает желание развертывать и расширять выбранную тему (в силу ассоциативных свойств мышления). С другой стороны, свободное развертывание речи ограничивается свойствами человеческой памяти, а также, например, необходимостью оставаться в рамках определенной тематики, что ведет к повторению уже использованных знаменательных слов; к этому прибавляется постоянное повторение структурных (служебных) слов, обусловленное системно-языковыми причинами. Эти две тенденции определяют вероятностный процесс нарастания словаря в целом, однако в рамках нашего анализа следует более четко сформулировать общий структурно-функциональный принцип, который сохранял бы ясный лингвистический смысл и в то же время мог бы быть подвергнут математическому анализу. Регулирующим рычагом и движущей силой при порождении речи можно считать принцип о г р а н и ч е н и я р а з н о о б р а з и я л е к с и к и в тексте, имеющий, по-видимому, своей более глубокой причиной некоторые филогенетические факторы (аналогом этого принципа можно рассматривать принцип уменьшения энтропии в самоорганизующихся системах).

Разнообразие лексики в тексте - несколько абстрактное понятие, но оно становится четким и наглядным, если его выразить с помощью известных в лингвостатистике мер - "коэффициента разнообразия", т.е. отношения объема словаря к объему текста (V/N), или обратного отношения (N/V), которое выражает среднюю частоту слова в данном тексте. Известно, что степень разнообразия лексики изменяется вместе с увеличением объема текста: коэффициент разнообразия V/N (именуемый также "индексом TTR " от англ. type-token ratio) монотонно уменьшается, а средняя частота (повторяемость) слова соответственно увеличивается. Приводим для примера данные об изменении степени разнообразия лексики на материале текстов из авторской речи эстонской художественной прозы (средние данные 20 выборок; объем словаря определяется по числу словоформ):

N	V	V/N	N/V
1000	686	0,686	1,46
2000	1241	0,621	1,61
3000	1762	0,587	1.70
4000	2238	0,560	1,79
5000	2692	0,538	1,86

Степень разнообразия лексики, измеряемая отношением между V и N, удобна для выявления некоторых существенных квантитативных свойств текста и в том смысле, что она (степень разнообразия) тесно коррелирует с вероятностным процессом выбора "нового" или "старого" слова на каждом шагу порождения текста. Отношение V/N отражает в определенном смысле вероятность появления (прибавления к словарю) нового слова. Действительно, если на достигнутом объеме текста в N словоупотреблений накоплен словарь из V различных слов, то вероятность прибавления нового слова к словарю приблизительно пропорциональна отношению V/N. Этим оправдывается оперирование величиной V/N вместо самих вероятностей, которые труднее поддаются выявлению и измерению.

Для выяснения некоторых дополнительных структурных свойств изучаемого явления можно представить процесс нарастания словаря как постоянное изменение "доли" использованного словаря в каком-то пространстве максимальных возможностей. Если, например, изобразить ход роста словаря графически (рис. I), то можно наблюдать постоянно замедляющееся на-

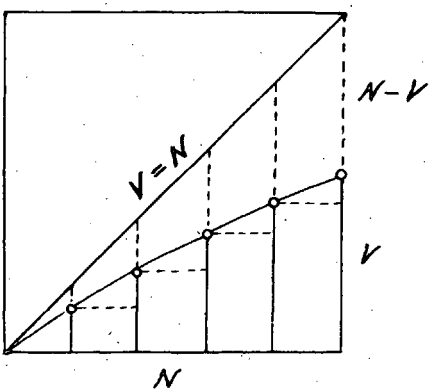

Рис. I. Нарастание объема словаря (V) в зависимости от увеличения объема текста (N).

растание объема словаря по сравнению с "базисной прямой", которая указывает на тот идеальный объем словаря, при котором слова в тексте совсем не повторялись бы ($V=N$). Такой случай встречается в действительности в самом начале порождения текста, а последующие этапы отражают неуклонный уход от начального состояния, где степень разнообразия имеет свое максимальное значение ($V/N = N/V = 1$). Как уже отмечалось, фактором, регулирующим этот уход, является обусловленный языковыми причинами процесс ограничения разнообразия лексики в реальном тексте.

В этой связи введем понятие д а в л е н и я р е к у р р е н т н о с т и (давления повторяемости слов), которое вместе с увеличением текста постепенно усиливается (отношение $(N-V)/N$, см. рис. 1), ограничивая степень разнообразия (отношение V/N). Если обозначить $V/N = p$, то давление рекуррентности (аналог избыточности) выражается через $q = (N-V)/N = 1 - V/N$; $p+q = 1$. Учитывая, что объем словаря V можно представить как некоторую долю "максимального" словаря, то исходя из вышесказанного, вычисление объема словаря можно осуществить на основе общих моделей:

$$V = Np \quad \text{или} \quad V = N(1-q).$$

4. Конструирование и проверка формул

Для наглядного представления возможностей различных решений проблемы аналитического выражения связи между объемом словаря и объемом текста применим индуктивный метод последовательной проверки гипотез и постепенного приближения к искомой форме зависимости на основе общей модели $V = Np$ (или $V = N(1-q)$). Преимущество такого подхода в том, что мы не теряем связь с нашими исходными посылками о роли фактора разнообразия в процессе порождения текста и можем конструировать соответствующие формулы на единой системной основе.

В первом приближении можно исходить из предположения, что изменение степени разнообразия лексики - это непрерывный процесс, при котором движение происходит равномерно и прямолинейно. Эмпирическая проверка показывает, однако, что зависимость между $p = V/N$ и N можно аппроксимировать линейной связью лишь на отдельных отрезках текста. Тем не менее можно на этой основе конструировать формулы (например, типа $V = N(a + bN)$ или $V = aN/(N+b)$, где a и b - констан-

ты), которые могут пригодиться для решения некоторых стилостатистических задач при исследования малых выборок /Тулдава Ю., 1974 и 1977/.

Гипотеза об экспоненциальной или логарифмической зависимости, т.е. когда предполагается наличие линейной связи между $\ln (V/N)$ и N или между V/N и $\ln N$, оправдывается лишь частично. Здесь не помогает модификация формул введением новых параметров (см. исследование В. Горьковой /1972/, в котором используется функция типа $V = V_0(1 - e^{aN})$, где V_0 - предполагаемый предел. словаря); и в этом случае функция более или менее удовлетворительно описывает ход нарастания словаря лишь на коротких отрезках текста.

Более обнадеживающим оказывается предположение о с т е п е н н о й с в я з и между V/N и N, т.е. когда предполагается, что скорости относительного изменения V/N и N остаются постоянными. В таком случае наблюдается линейная связь между $\ln (V/N)$ и $\ln N$, и зависимость между степенью разнообразия и объемом текста выражается функцией типа

$$V/N = aN^{B}, \qquad (I)$$

где a и B - константы (в данном случае, когда V/N уменьшается с увеличением N, константа $B < 0$). Эта функция по своей форме аналогична закону Ципфа /Zipf G.K., 1949/. Из этой зависимости образуем формулу зависимости между объемом словаря и объемом текста:

$$V = N(aN^{B}) = aN^{b}, \qquad (Ia)$$

где $b = B + 1$. Степенная функция такого типа называется "аллометрической" функцией, которая широко используется в разных отраслях науки при описании роста частей в отношении некоторого целого. Известно, что как ципфовская, так и аллометрическая разновидности степенной функции выражают какие-то существенные стороны функционирования целого класса сложных самоорганизующихся систем, связанных с деятельностью общения людей /ср. Яблонский А.И., 1976; Тулдава Ю., 1979, с. 130/.

Как известно, Г. Хердан /Herdan G., 1966/ считал функцию типа $V = aN^{b}$ универсальной для выражения связи между объемом словаря и объемом текста, но опытные данные по многим

языкам /см., например, Weitzmann M., 1971/ показывают, что эта функция способна описывать связь между V и N лишь локально, обычно лишь в начальной стадии порождения текста (с начала текста до ок. 4-5 тысяч словоупотреблений, т.е. примерно в рамках одного короткого рассказа). В дальнейшем темп нарастания объема словаря в реальных текстах замедляется, и прогнозы по аллометрической функции дают завышение оценки. Для примера приводим данные по эстонскому языку (см.табл.I), где использованы средние величины V (в словоформах) в точках N = 1000÷5000, и делается прогноз на N = 100.000. По формуле /Ia/ прогнозируется V = 34625, но это существенно больше, чем действительный объем сводного словаря 20 текстов при N = 100.000 (V = 30700)*, причем надо учесть, что сводный словарь всегда больше, чем словарь отдельного автора при таком же объеме текста (а мы прогнозируем словарь в данном случае среднего индивидуального текста).

Учитывая тот факт, что с увеличением текста темп нарастания объема словаря постепенно замедляется по сравнению с аллометрическим законом роста, можно попробовать прологарифмировать V и/или N в исходной формуле (Ia). Но оказывается, что такой подход не дает хороших результатов (см., например, замечания по поводу формулы Сомерса: /Тулдава Ю., 1974, с. 13/. Целесообразнее будет возвратиться к нашему исходному принципу и определить закономерности нарастания объема словаря через "степень разнообразия лексики". Заменяя в формуле (I) выражение V/N его логарифмом $\ln(V/N)$, получим

$$\ln (V/N) = -aN^{\beta}, \quad ** \qquad (2)$$

отсюда $V/N = e^{-aN^{\beta}}$ и

$$V = Ne^{-aN^{\beta}}, \qquad (2a)$$

---

* Данные авторской речи эстонской художественной прозы по 20 текстам /полные данные по отдельным текстам см. Тулдава Ю., 1977, с. 175, и по сводному частотному словарю эстонского языка см. Tuldava J., 1977, с. 164/. Здесь и далее константы формул вычисляются на основе линеаризации с помощью метода наименьших квадратов.

** Формула (2), а также ее варианты (2а), (2б) и (2в)и формула (3) на стр. 126 вместе с ее вариантом (3а) не являются следствием формулы (I). Формулы (2) и (3) имеют самостоятельное значение.

где a и β - константы, e - основание натуральных логарифмов. Формулу можно вывести и на основе фактора рекуррентности $q = 1 - V/N$, тогда приходим к выражению $\ln(1 - V/N) = -aN^{\beta}$, и формула связи между объемом словаря и объемом текста принимает вид

$$V = N(1 - e^{-aN^{\beta}}). \qquad (2б)$$

При конструировании всех вышерассмотренных формул подразумевалось, что нарастание объема словаря не имеет границ, точнее, максимальный объем словаря ограничивался "базисной прямой" (рис. 1), т.е. когда $V = N$, а объем словаря на каждом этапе порождения текста вычислялся как "доля" этого максимального объема (по схеме $V = Np$ или $V = N(1 - q)$).

Однако можно представить себе, что при порождении текста наступает момент, когда новых слов уже не поступает и словарный запас говорящего (пишущего) иссякает или становится ничтожно малым. Исчерпывание словарного запаса можно тогда рассматривать как некое предельное состояние "устойчивости", к которому система стремится. В этом смысле можно говорить, например, о "потенциальном" словарном запасе автора или авторов определенного класса текстов (в синхроническом плане)*. Предел словаря в таком случае определяется аналитическим способом, исходя из тенденции роста словаря и используя соответствующую формулу.

Обозначая гипотетический предел (максимальный объем) словаря через V_0, определяем отношение V/V_0 как долю уже использованных слов и выражение $1 - V/V_0$ как долю еще не использованных слов из данного конечного словаря V_0. Проверка показывает, что лучшие результаты дает применение последнего выражения $(1 - V/V_0)$, и формула связи между V и N принимает тогда следующий вид:

$$V = V_0(1 - e^{-aN^{\beta}}). \qquad (2в)$$

Параметр V_0 определяется методом итераций (выбирается то значение параметра, которое позволяет наилучшим образом ап-

* Если рассматривать повседневную речевую деятельность человека как продолжение какого-то длинного текста (на родном языке), то можно предполагать, что взрослый говорящий постоянно находится в состоянии "асимптотической устойчивости" в отношении своего словарного запаса.

проксимировать экспериментальные данные теоретической функцией).

Если проверить эффективность формул (2а), (2б) и (2в) на практическом материале, то оказывается, что все они достаточно хорошо описывают рост словаря в пределах опытных данных (см. табл. I-4). Но прогнозирующая сила этих формул, проверенная на материале текстов разного типа и разных языков, оказывается различной. По нашим экспериментальным данным формула (2а) несколько уменьшает оценки объема словаря при прогнозе "вперед", в то время как при экстраполяции в сторону меньших величин (прогноз "назад") формула дает хорошее приближение. По формуле (2б) получаются несколько завышенные оценки как в прогнозе вперед, так и назад.

Что касается формулы (2в), то можно сказать следующее. Оценки потенциального словарного запаса (V_o), полученные по данной формуле, не имеют стабильного характера и меняются в зависимости от выбора точек наблюдения. В таком случае никакого реального смысла нельзя вкладывать в значение V_o как "предела объема словаря" (если предположение о существовании такого предела принимается). Далее, формула дает систематически заниженные оценки при прогнозе вперед на более дальние расстояния и, как правило, завышенные оценки при прогнозе назад. Таким образом, функцию (2в) нельзя рассматривать как выражение действительной тенденции роста словаря, хотя ее можно с успехом применять при интерполяции (при правильном выборе V_o можно достичь очень хорошего приближения к опытным данным; см., например, табл. I). Формулу можно использовать и при стилостатистическом анализе текстов одинакового объема и при одинаковых условиях эксперимента; величину V_o можно в таком случае рассматривать как стилидифференцирующий фактор и как относительную меру богатства лексики данного текста.

Функции (2а), (2б) и (2в) можно рассматривать как разновидности одной и той же формы зависимости, при которой в правой части формулы имеется сложная экспонента типа e^{-aN^b} или $1 - e^{-aN^b}$. Последний случай характеризует распределение Вейбулла /Weibull W., I939/*. Будем условно называть выше-

* Как известно, первым показал применимость распределения Вейбулла к языковым данным Г.Г. Белоногов /I962/. Это распределение является исходным и для В. В. Нешитого /I975/ при конструировании формулы связи между объемом словаря и объемом текста.

указанные функции распределениями вейбулловского типа. Как показывают наши опыты, распределения такого типа описывают тенденцию роста словаря в зависимости от увеличения текста вполне удовлетворительно лишь в том случае, если ограничиться диапазоном наблюдаемых данных (особенно, если привлечь еще третий параметр V_0).

В поисках более адекватного соответствия общей тенденции роста словаря, при которой и экстраполяция давала бы более или менее достоверные результаты, будем продолжать процедуру последовательного логарифмирования компонентов исходной формулы связи между степенью разнообразия и объемом текста ($V/N = aN^{\beta}$). Если формула (2) (и ее варианты) были получены в результате логарифмирования переменной V/N в левой части формулы, то следующим этапом будет замена в правой части равенства (I) величины N на величину $\ell n\, N$, в результате чего получаем

$$\ell n\,(V/N) = a\,(\ell n\, N)^{\beta}. \qquad (3)$$

Это дает нам $V/N = e^{-a(\ell n N)^{\beta}}$ и формулу роста словаря:

$$V = Ne^{-a(\ell n N)^{\beta}} \quad * \qquad (3a)$$

Мы видим, что в отличие от вейбулловской зависимости экспонента в этой формуле содержит элемент $\ell n\, N$ вместо N (приращение функции зависит нелинейно от логарифма аргумента). Проверка показывает, что такая зависимость наиболее достоверно отражает тенденцию роста словаря в реальных текстах в широком диапазоне от самого начала текста до размеров порядка $N = 10^6$ (см. табл. 4), причем оказывается, что в принципе одна и та же формула зависимости наблюдается как в отношении отдельных, так и сводных текстов, а также при из-

---

* Формулу можно представить и в виде $V = N^{1-a(\ell n N)^{\beta}}$, где $\beta = B - 1$. В таком виде формула была описана автором в статье /Tuldava J., 1977/. В этом случае за основу берется модель $V = N^{1-d}$, где d - фактор ограничения роста словаря, причем d изменяется нелинейно в зависимости от объема текста N . Процесс нарастания словаря представляется и здесь как неуклонный уход реального текста от исходного состояния ($V = N^1$) под воздействием давления рекуррентности.

мерении объема словаря в словоформах или лексемах (см. табл. I-3 и 6).

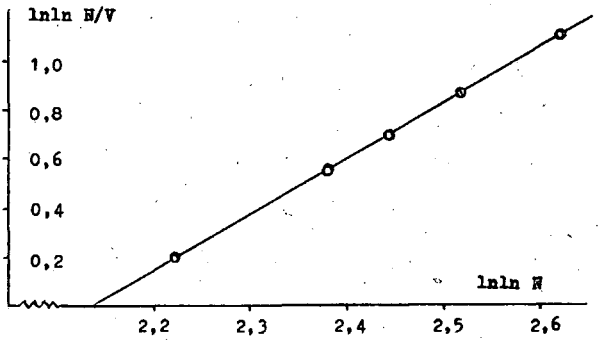

Рис. 2. Линейная связь между $\ell n \ell n (N/V)$ и $\ell n \ell n N$ (по данным ЧС английского языка Х.Кучеры и В. Фрэнсиса)

При практическом использовании можно вычислить константы в формуле (3а) на основе линеаризации: $\ell n \ell n (N/V) = A + B \ell n \ell n N$, где $A = \ell n |a|$. Линейная связь между $\ell n \ell n (N/V)$ и $\ell n \ell n N$ оказывается хорошей (см. рис. 2). Для вычисления констант A и B достаточно знать две точки по ходу порождения текста (хотя для большей достоверности рекомендуется выбрать большее число точек наблюдения)*. Например, вычисление по данным выборки из отдельного текста (табл. 2)и на основе данных сводного текста (табл. 4) дает возможность достаточно точно прогнозировать объемы полных словарей.

Для сравнения приводятся вычисления по другим рассмотренным формулам, в том числе на основе формул по параметру Z (согласно методу Ю.К. Орлова). Прогнозируемый объем словаря по параметру Z вычисляется в двух этапах: сначала определяется на основе данных о N, V и p_1 (методом итераций) "объем Ципфа" Z, т.е. такой объем текста, при котором частотная структура предположительно соответствует обобщенному

* Вычисление констант по двум точкам (N_1, V_1) и (N_2, V_2): $B = (y_1 - y_2)/(x_1 - x_2)$; $A = [(y_1 + y_2) - B(x_1 + x_2)]/2$; где $x_1 = \ell n \ell n N_1$, $x_2 = \ell n \ell n N_2$, $y_1 = \ell n \ell n (N_1/V_1)$, $y_2 = \ell n \ell n (N_2/V_2)$.

закону Ципфа-Мандельброта, и соответствующий объем словаря $V(Z)$; затем вычисляется объем словаря при данном объеме текста по формуле $V(N,Z) = V(Z)\frac{\ln X}{X-1}$, где $X = Z/N$ /подробнее см. Орлов Ю.К., 1977/. Мы уже указывали, что определение роста словаря по этому методу не дает вполне достоверных результатов (прогноз сильно варьирует в зависимости от точек наблюдения; прогноз вперед дает, как правило, систематически заниженные оценки). В этом можно убедиться и на основе приведенных примеров, хотя следует сказать, что в отдельных случаях прогноз по Z может дать и хорошие результаты (см., например, табл. I). Ю.К. Орлов объясняет различную прогнозирующую силу своего метода тем, что достоверные результаты можно получить при исследовании отдельных законченных текстов, в то время как сводные (смешанные) тексты не так хорошо поддаются анализу по формулам на основе Z. Однако наши экспериментальные данные не подтверждают упомянутого утверждения в отношении отдельных текстов (см., например, табл. 2 и 4).

Если даже результаты вычислений по Z приблизительно соответствуют действительности, то в практической лингвостатистической работе, особенно при более точном стилостатистическом анализе (сравнении, атрибуции) текстов, они не могут служить достоверными показателями, в частности при выявлении действительной тенденции роста словаря. Кроме того, приходится констатировать, что производить достаточно трудоемкие вычисления по Z просто неэкономно, когда такие же приблизительные данные можно гораздо легче получить с помощью совсем простых методов, например, с помощью формулы Курашкевича-Гиро ($V = a\sqrt{N}$).*

5. О возможностях экстраполяции

Хорошая описательная сила какой-нибудь функции в пределах опытных данных и при многократном прогнозе как вперед, так и назад наводит на мысль, что функцию можно использовать

* Пример: вычисления по формуле Курашкевича-Гиро по данным ЧС английского языка (см. табл. 4) в точке $N = 253538$ позволяет прогнозировать объем всего словаря $V = 47333$ (при $N = 1014232$). Прогноз по Z на основе той же точки дает $V = 42620$. В действительности объем всего словаря $V = 50406$, т.е. прогноз по Курашкевичу-Гиро оказывается даже лучше, чем прогноз по формуле Z.

также для экстраполяции на более дальние участки текста. Безусловно, это несколько рискованное предприятие, особенно если учесть, что мы в настоящее время не располагаем возможностями контроля таких прогнозов. Но тем не менее практика автоматической обработки текстов может предъявлять требование хотя бы приблизительно прогнозировать объем словарей крупных текстовых массивов. В таком случае правомерно использовать именно такую формулу, которая показала себя на практике наиболее стабильной и достоверной.

Такой формулой можно на основании наших опытов считать формулу (3). Она является модификацией схемы исходной степенной связи между степенью разнообразия (V/N) и объемом текста (V), причем окончательный вид формула получила логарифмированием обоих своих компонентов. Оказалось, что степень разнообразия убывает в связи с увеличением текста по сложной экспоненте, которая содержит логарифм от агрумента ($V/N = e^{-a(\ln N)^b}$). В содержательной интерпретации это означает, что разнообразие убывает в связи с увеличением текста не так стремительно, как это можно было бы предположить, и задержка убывания разнообразия чувствуется особенно на более дальних отрезках текста. В соответствии с этим фактом замедление роста словаря идет в более низком темпе.

Прогноз, который основывается на предположении о "сохранении общей тенденции развития явления во времени" и о "сохранении в основных чертах взаимосвязей прогнозируемого явления с другими явлениями" /Четыркин Е.М., 1975, с. 57/, может быть проведен только в условиях однородности явления. Под однородностью следует в данном случае понимать одинаковость состава, например, одинаковую тематику текстов и соответственный словарный запас*. При анализе смешанных (сводных) текстов требование однородности означает сохранение состава определенных исходных пропорций отдельных жанров или подъязыков и соблюдение прочих общих условий эксперимента.

С помощью формулы (3а) мы предприняли попытку прогнозирования объема словаря для разных текстов из разных языков

* В литературе по лингвостатистике иногда трактуют однородность как "независимость появления языковых единиц в тексте", как "идеальное перемешивание" единиц текста и т.п. /см., например, Орлов Ю.К., 1978/. Представляется, что в подобных случаях целесообразнее использовать термины "связность" и "несвязность" единиц текста /ср. Лукьяненков К.Ф., Нешитой В.В., 1975/.

вплоть до объема текста $N = 10^7$ (см. табл. 4 и 5). При сравнении данных следует учесть различие между жанрами (подъязыками) и различие в подсчете объема текста (в словоформах или лексемах). Например, по данным научно-технических текстов английского языка (табл. 5, е, ж) при $N = 10^7$ прогнозируется объем текста 36000-38000 словоформ, в то время как в смешанных английских текстах ожидаемый объем словаря при такой же длине текста - 148.000 словоформ (табл. 4). При сравнении данных, взятых из разных языков надо учесть и разницу в степени аналитичности языков; так, например, в текстах по электронике в английском языке прогнозируется 38.000 словоформ при $N = 10^7$, в то время как в русских текстах по электронике при такой же длине текста ожидаемый объем словаря - 94.000 словоформ. В агглютинативном казахском языке (в газетных текстах) ожидается объем словаря - 230.000 словоформ при $N = 10^7$.

Для эстонского языка мы вычислили данные по сводным текстам авторской речи художественной прозы (при условии использования текстовых выборок по 5000 словоупотреблений, разделенных на 5 порций по 1000 словоупотреблений). Ожидаемый объем словаря вычислялся отдельно для словоформ (V_c) и для лексем ($V_л$). Можно обратить внимание на изменение индекса синтетичности, т.е. отношения $V_c / V_л$, обозначающего по существу среднее число разных флективных форм на одну лексему, в связи с увеличением текста (звездочкой отмечены наблюдаемые данные, на которых был произведен расчет):

N	V_c	$V_л$	$V_c / V_л$
5000*	3000	2200	1,36
10000	5100	3600	1,42
20000	8700	5700	1,53
50000	17200	10000	1,72
100000*	30000	14600	2,05
200000	45000	21000	2,14
500000	80000	31000	2,58
10^6	120000	40000	3,00
10^7	360000	70000	5,14

(Ожидаемые объемы словарей вычислены по формулам : $V_c = Ne^{-0,0012(\ln N)^{2,85}}$ и $V_л = Ne^{-0,0019(\ln N)^{2,83}}$).

Наконец, следует упомянуть еще об одной особенности рассматриваемой функции, которая в некотором смысле является препятствием для прогнозирования на очень далекие участки

текста. Оказывается, что при применении функции $V = Ne^{-a(\ln N)^b}$ достигается предел, т.е. функция имеет максимум (в точке $N_{max} = e^{1/\sqrt{ab}}$, где $b = B - 1$). Например, по данным ЧС английского языка (табл. 4) $N_{max} \approx 10^{11}$, когда максимальное значение объема словаря $V \approx 800.000$. Заманчиво было бы считать эту точку таким уровнем порождения текста, где прирост новых слов прекращается*. Однако никаких реальных оснований для предположения о совпадении максимума функции и предела словаря не имеется, и приходится пока довольствоваться тем, что функция в общем достаточно хорошо описывает динамику роста словаря на большом протяжении текста**.

5. Заключение

Подытоживая сказанное, можно констатировать следующее.

I. Наряду с чисто эмпирическим подходом к проблеме аналитического выражения связи между объемом словаря (V) и объемом текста (N) в современной квантитативной лингвистике намечаются и некоторые направления теоретического анализа изучаемой взаимосвязи по существу. При комплексном ("структурном") подходе стараются связать разные стороны статистической структуры текста воедино, причем связь между V и N обычно выводится как нечто вторичное из других структурных характеристик текста, таких как распределение Ципфа-Мандельброта или распределение Юла. Неудовлетворенность практическими результатами такого подхода вызывает к жизни другое направление, при котором связь между V и N рассматривается

---

* После максимума функция начинает убывать, что́ не имеет лингвистического смысла (хотя чисто теоретически можно представить себе наступление точки насыщения, после чего начинается "забывание" слов). Если исследовать последовательность всех возможных значений функции, то образуется кривая, которая близка к логнормальному распределению. Лингвистическому смыслу удовлетворяет лишь левая сторона кривой ("усеченное логнормальное распределение"), т.е. та часть, которая предшествует максимуму. Но так как этот максимум, как правило, лежит достаточно далеко (для сводных текстов N_{max} обычно порядка 10^9 или 10^{10}), то он в сущности не мешает применению функции в исследовательской работе, учитывая потребности практики сегодняшнего дня.

** Интересно отметить, что формула (3а) в принципе подходит и для выражения зависимости между количеством одноразовых слов и объемом текста /см. Tuldava J., 1977/. Можно еще упомянуть о возможности использования параметров функции (B и A/B, где $A = \ln |a|$) в качестве содержательно интерпретируемых стиледифференцирующих показателей.

как исходный, первичный момент, который подлежит особому исследованию на фоне некоторых глубинных факторов порождения речи. В статье обсуждается вариант такого "прямого" подхода, когда порождение текста рассматривается как вероятностный процесс, при котором нарастание объема словаря характеризуется устойчивым распределением. Устойчивость системы достигается взаимодействием различных языковых и внеязыковых явлений, которые можно свести к интегральному понятию "ограничения разнообразия лексики", подлежащему математическому анализу. Понятие разнообразия лексики квантифицируется и определяется при таком подходе как отношение объема словаря к объему текста ("степень разнообразия" V/N), которое в свою очередь является функцией от объема текста и в количественном выражении неуклонно (но нелинейно) уменьшается вместе с увеличением текста под воздействием "давления рекуррентности (повторяемости) слов". Этот процесс имеет свой аналог в уменьшении энтропии в самоорганизующихся системах.

2. Исходя из названных предпосылок, строятся формулы связи между степенью разнообразия (V/N) и объемом текста (N), из которых прямо выводятся искомые формулы зависимости объема словаря от объема текста. За исходное берется степенная связь (с отрицательным показателем) между V/N и N, которая по своей общей форме аналогична закону Ципфа. Такая связь выполняется в начальной стадии порождения текста (в пределах короткого сообщения, содержащего ок. 4000-5000 словоупотреблений). Однако в связи с тем, что темп уменьшения степени разнообразия при дальнейшем увеличении текста не отвечает **простой степенной связи между** V/N **и** N, требуется модификация (преобразование) исходной формулы. По содержательным соображениям представляется естественным модифицировать исходную формулу путем последовательного логарифмирования переменных. Таким образом получаются три основных типа связи (a и B - константы):

$V/N = a N^{\beta}$ (линейная связь между $\ln \frac{V}{N}$ и $\ln N$);

$\ln (V/N) = a N^{\beta}$ (линейная связь между $\ln|\ln \frac{V}{N}|$ и $\ln N$);

$\ln (V/N) = a (\ln N)^{\beta}$ (линейная связь между $\ln|\ln \frac{V}{N}|$ и $\ln \ln N$).

Из этих основных типов связи выводятся соответствующие формулы зависимости между V и N, построенные по общей схеме: $V = Np$ (или $V = N(1-q)$, где p - степень разнообразия ($q = 1-p$). Экспериментальная проверка показывает,

что наиболее адекватное соответствие действительности имеет третий тип связи, из которого выводится функция

$$V = Ne^{-a(\ln N)^{\beta}}$$

(или $V = N^{1-a(\ln N)^{\delta}}$, где $\delta = \beta - 1$). Эта функция хорошо описывает зависимость между V и N от самого начала текста до объема текста порядка $N = 10^6 \div 10^7$ (предположительно еще дальше). Второй тип связи, отражающий распределения вейбулловского типа, действителен в пределах опытных данных умеренно длинного текста, а первый тип связи, приводящий к известной формуле Г. Хердана /Herdan G., 1966/, отражает взаимосвязь между V и N в начальной стадии порождения текста, являясь, таким образом частным случаем более общего третьего типа связи.

3. Эффективность предлагаемой формулы третьего типа проверялась на материале текстов из девяти языков, относящихся к различным языковым семьям и группам (эстонский; казахский; латышский; русский, польский, чешский; французский, румынский; английский). Формула одинаково хорошо описывает интересующую нас взаимосвязь как в отдельных, так и сводных (смешанных) однородных текстах (см. табл. 1-6). Это говорит в пользу того, что, несмотря на некоторые имеющиеся различия, всякие тексты, в том числе "суммарные", очевидно подчиняются какой-то всеобщей (интегральной) квантитативной закономерности порождения речи. Мы не можем, однако, утверждать, что именно предлагаемая конкретная функция представляет собой квантитативно-лингвистический закон в прямом смысле слова, но она в какой-то степени должна отражать этот закон. Пока не будет достигнуто лучшее понимание рассматриваемой функции, ее можно рассматривать как удобное и простое феноменологическое описание реально наблюдаемых фактов, которое позволяет получить достаточно надежные результаты при решении практических задач. В то же время примененный в настоящем исследовании теоретический подход раскрывает некоторые новые аспекты в изучении проблемы динамики роста словаря при порождении текста.

Таблица I

Наблюдаемый и ожидаемый рост словаря по данным
авторской речи современной эстонской художественной прозы (словоформы)

N	V (набл.)	V (ожид.) по формулам: /1а/	/2а/	/2б/	/2в/	/3а/	Не Z
1000	686	688 (+2)	684 (-2)	686 (0)	685 (-1)	685 (-1)	684 (-2)
2000	1241	1240 (-1)	1249 (+8)	1245 (+4)	1244 (+3)	1246 (+5)	1241 (0)
3000	1762	1752 (-10)	1760 (-2)	1756 (-6)	1756 (-6)	1757 (-5)	1759 (-9)
4000	2238	2237 (-1)	2234 (-4)	2237 (-1)	2237 (-1)	2237 (-1)	2233 (-5)
5000	2692	2705 (+13)	2680 (-12)	2697 (+5)	2693 (+1)	2693 (+1)	2692*
		Прогноз:					
10000	–	4880	4623	4791	4715	4753	4761
50000	–	19196	14115	17690	14229	16867	16804
100000	–	34625	20912	30746	19398	28432	27904
Параметры:							
a		1,9248	-0,0456	-6,4732	$-6,728 \cdot 10^{-5}$	-0,0040	–
B		0,8510	0,3071	0,2492	0,8780	2,3535	–
V_o		–	–	–	24000	–	–
Z (p_1=0,0322; V(Z)=164121)		–	–	–	–	–	$1,8 \cdot 10^6$

Таблица 2

Наблюдаемый и ожидаемый рост словаря по данным романа "Правда и право" (т.I) А.Х.Таммсааре (эстонский язык; авторская речь, лексемы); звездочкой отмечены точки наблюдения (по данным: Villup, 1978)

N	V(набл.)	V (ожид.) по формулам:				
		/2а/	/2б/	/3а/	по z'	по z''
10000	2114*	*	*	*	*	2149
20000	3124*	*	*	*	3065	*
Полный объем:			Прогноз:			
114124	7348	6259	8052	7207	6550	6735
Параметры:						
a		-0,1461	-20,4413	-0,006714	-	-
B		0,2567	-0,4837	2,4521	-	-
z' (p_1=0,00417; V(z)=3201)		-	-	-	21800	-
z'' (p_2=0,00417; V(z)=3362)		-	-	-	-	23100

Таблица 3

Наблюдаемый и ожидаемый рост словаря по данным "Пиковой дамы" А.С.Пушкина (по данным: Орлов, 1978) - лексемы

N	V(набл.)	V (ожид.) по формулам:				
		/2а/	/2б/	/2в/	/3а/	по z
1000	462*	*	*	*	*	493
2000	787*	*	*	*	*	809
3000	1067*	*	*	*	*	*
				Прогноз:		
4000	1348	1308	1325	1326	1321	1288
5000	1541	1527	1564	1567	1556	1487
6000	1752	1727	1790	1794	1776	1666
Полный объем:						
6861	1928	1887	1976	1981	1957	1809
Параметры:						
a		-0,1128	-5,376	-0,0001094	-0,01694	-
B		0,2663	-0,3126	0,7766	1,976	-
V_0		-	-	20000	-	-
z(p_1=0,038; V(z)=3833)		-	-	-	-	26500

Таблица 4

Наблюдаемый и ожидаемый рост словаря по данным ЧС английского языка Х.Кучеры и В.Фрэнсиса (Kučera, Francis, 1967) – словоформы

N	V(набл.)	(ожид.) по формулам: /2а/	/2б/	/2в/	/3/	по z′	по z″
101566	13706*	*	*	*	*	*	14716
253538	23655*	*	*	*	*	21692	*
		Прогноз:					
1014232	50406	47500	53105	49076	50617	38230	42620
10^7	–	93290	199430	95332	148000	73500	84210
50721	8749	8737	8961	8891	8853	9236	9808
10051	3009	2729	3241	3138	2968	3203	3327
2000	700–1000	760	1082	1095	902	956	978
1000	–	427	649	694	525	550	559
500	–	437	378	438	300	311	315
100	–	57	94	151	77	79	80
10	–	7	10	33	9	10	10
Параметры:							
a		-0,237	-20,287	$-7{,}21\cdot10^{-5}$	-0,00891	–	–
B		0,185	-0,429	0,6612	2,215	–	–
V_0		–	–	100000	–	–	–
z′ (p_1=0,06933; V(z)=17609)		–	–	–	–	164500	–
z″ (p_1=0,06847; V(z)=21130)		–	–	–	–	–	201400

Таблица 5

Наблюдаемый объем (V) и ожидаемый объем (V′) словаря в зависимости от длины текста (N) по данным разных языков по формуле: $V' = Ne^{-a(\ln N)^B}$

а) Латышский язык – газеты, лексемы (Latviešu val., 1969)			б) Чешский язык – технич. тексты, словоформы (Bečka, 1972)			в) Казахский язык – газеты, словоформы (Ахабаев, 1971)		
N	V	V′	N	V	V′	N	V	V′
50000	7065	7025	25000	4829	4827	25000	9088	9161
100000	9834	9919	75000	9603	9626	50000	15047	14875
200000	13389	13510	125000	13056	13050	100000	23895	23523
300000	16103	15912	175000	15858	15853	150000	29785	30378
500000	–	19200	500000	–	28200	500000	–	61000
10^6	–	24000	10^6	–	40000	10^6	–	87000
10^7	–	37000	10^7	–	114000	10^7	–	230000
Параметры:								
a	0,003736			0,01123			0,001372	
B	2,6304			2,1539			2,8488	

г) Польский язык – А.Мицкевич, словоформы, (Sambor, 1970)			д) Украинский язык – О.Довженко, словоформы (Дарчук, 1975)			е) Английский язык – суд.механизмы, словоформы (Лукьяненков, Ланитой, 1975)		
N	V	V′	N	V	V′	N	V	V′
12172	3434	3458	5000	1629	1629	50495	4871	4849
29787	6146	6044	10000	2637	2646	100970	6858	6882
48255	8026	7998	15000	3504	3482	201966	9470	9520
64510	9250	9398	20000	4195	4214	302156	11314	11360
100000	–	11800	100000	–	11500	403966	12975	12832
500000	–	25000	500000	–	28000	500000	–	14000
10^6	–	33000	10^6	–	40000	10^6	–	18000
10^7	–	60000	10^7	–	110000	10^7	–	36000
Параметры:								
a	0,00364			0,01055			0,01235	
B	2,6081			2,1783			2,2019	

ж) Английский язык – электроника, словоформы (Алексеев, I968)

N	V	V′
50000	5399	5437
100000	7853	7728
150000	9361	9371
200000	10582	10682
500000	–	15600
10^6	–	20000
10^7	–	38000

Параметры:

a 0,009152
B 2,3057

з) Румынский язык– электроника, словоформы (Еван, I966)

N	V	V′
50000	6785	6841
100000	10281	10070
150000	12477	12479
200000	14292	14454
500000	–	22400
10^6	–	30000
10^7	–	68000

0,008148
2,3086

и) Русский язык – электроника, словоформы (Калинина, I968)

N	V	V′
50000	9464	9388
100000	14062	14168
150000	17263	17803
200000	21468	20818
500000	–	33000
10^6	–	45000
10^7	–	94000

0,004284
2,5058

ЛИТЕРАТУРА

Алексеев П.М. Лексическая и морфологическая статистика английского подъязыка электроники. - В кн.:Статистика речи. - Л.: Наука, 1968, с. 120-131.

Ахабаев А. Статистический анализ лексико-морфологический структуры языка казахской публицистики. АКД. Алма-Ата, 1971.

Белоногов Г.Г. О некоторых статистических закономерностях в русской письменной речи. - Вопросы языкознания, 1962, № 1, с. 100-101.

Белоногов Г.Г., Новоселов А.П. Некоторые количественные закономерности в автоматизированных информационных системах. - Автоматическая переработка текста методами прикладной лингвистики. Материалы конференции. Кишинев, 1971, с.219- - 220.

Горькова В.И. Статистические оценки параметров совокупностей документальных информационных потоков. - Научно-техническая информация. Серия 2. М., 1972, № 12. с. 14-20.

Дарчук Н.П. Індивідуальне загальне в лексичній системі авторського стилю (на матеріалі сучасної української художньої прози). Канд.дисс. Київ, 1975.

Ешан Л.И. Опыт статистического описания научно-технического стиля румынского языка. АКД. Л., 1966.

Захарова А.В. Опыт статистического исследования устной речи ребенка. - В кн.: Исследования по языку и фольклору. Вып. 2. Новосибирск, 1967, с. 16-38.

Калинин В.М. Некоторые статистические законы математической лингвистики. - Проблемы кибернетики. Вып.11. М., 1964, с. 245-255.

Калинин В.М. Функционалы, связанные с распределением Пуассона, и статистическая структура текста. - Труды Математического института им. Стеклова. Том 29. М., 1965, с. 182-197.

Калинина Е.А. Частотный словарь русского подъязыка электроники. - В кн.: Статистика речи. - Л.: Наука, 1968, с.114-150.

Ковлер Т.Д. Исследования лексической структуры текста с количественной стороны. АКД. М., 1967.

Лукьяненков К.Ф., Нешитой В.В. Оценка степени связности слов в научно-техническом тексте. - В кн.: Вопросы лингвистики и методики преподавания иностранных языков. - Минск: Наука и техника, 1975, с. 52-62.

Надарейшвили И.Ш. Сравнительный статистический анализ лексики как метод изучения творчества писателя. - В кн.:Структурная и математическая лингвистика. Вып. 6.- Киев: Вища школа, 1978, с. 45-52.

Надарейшвили И.Ш., Орлов Ю.К. Метод полной фиксации текста при лингвостатическом анализе. - В кн.: Проблемы общей и прикладной лингвистики. *LINGUISTICA* X. Тарту, 1978,с.65-84.

Нешитой В.В. Длина текста и объем словаря. Показатели лексического богатства текста. - В кн.: Методы изучения лексики. - Минск: Изд-во БГУ, 1975, с. 110-118.

Орлов Ю.К. О статистической структуре сообщений, оптимальных для человеческого восприятия (к постановке вопроса).- Научно-техническая информация. Серия 2. М., 1970, № 8,с.11-16.

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. Модель частотной структуры лексики. - В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. - М.: Изд-во МГУ, 1978, с. 59-118.

Орлов Ю.К. Статистическое моделирование речевых потоков. - В кн.: Статистика речи и автоматический анализ текста. Вопросы кибернетики, вып. 41. - М.-Л.: Наука, 1978.

Остапенко В.Е. Динамика роста словаря как метод оптимизации преподавания иностранных языков. - В кн.:Школа-семинар по оптимизации преподавания иностранных языков с помощью технических средств. Тезисы докладов и сообщений.Кишинев, 1979, с. 51.

Пиотровский Р.Г. Текст, машина, человек. - Л.: Наука, 1975.

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

Рыбников К.А. Введение в методологию математики. - М.:Изд-во МГУ, 1979.

Сачков Ю.В. Введение в вероятностный мир. Вопросы методологии. - М.: Наука, 1971.

Тулдава Ю. О статистической структуре текста. - Советская педагогика и школа. Вып. 9. Тарту, 1974, с. 5-33.

Тулдава Ю. О квантитативных характеристиках богатства лексического состава художественных текстов. - LINGUISTICA IX. Тарту, 1977, с. 159-175. (Учен. зап. Тартуского ун-та, вып. 437).

Четыркин Е.М. Статистические методы прогнозирования. - М.: Статистика, 1975.

Яблонский А.И. Стохастические модели научной деятельности. - В кн.: Системные исследования. Ежегодник 1975. - М.: Наука, 1976, с. 5-42.

Bečka J.V. La structure lexicale des textes techniques en tchèque. - Philologica Pragensia, 1972, vol. 15, No. 1.

Carroll J.B. On Sampling from a Lognormal Model of Word-frequency Distribution. - In: Computational Analysis of Present-day American English by H.Kučera and W.N.Francis. Providence, R. I., 1967, pp. 406-424.

Chotlos J.W. Studies in Language Behaviour, IV. Statistical and Comparative Analysis of Individual Written Language Samples. - Psychological Monographs, 1944, vol. 56, pp. 75-111.

Guiraud P. Problèmes et méthodes de la statistique linguistique. Dordrecht, 1959.

Herdan G. The Advanced Theory of Language as Choice and Chance. - Berlin-Heidelberg-New York: Springer-Verlag, 1966.

Kuraszkiewicz W. Statystyczne badanie słownictwa polskich tekstów XVI wieku. - Polskie Studia Sławistyczne. Prace Językoznawcze i Etnogenetyczne. Warszawa, 1958, str. 241-257.

Kučera H., Francis W.N. Computational Analysis of Present-day American English. Providence, R. I., 1967.

Latviešu valodas biežuma vārdnīca. II sēj. Laikraksti un žurnāli. 1. daļa. / Atb. red. T. Jakubaite. - Rīga: Zinātne, 1969.

Minor J. Structure statistique d'un texte selon Kalinin. - Stud. ling. appl., 1971, No. 1, pp. 6-19.

Miller W. Wortschatzumfang und Textlänge. Eine kleine Studie zu einem vielbehandelten Problem. - Muttersprache, 81. Jg., Nr. 4. Mannheim-Zürich, 1971, S. 266-276.

Sambor J. Analiza stosunku "type-token", czyli objętości słownictwa (W) i długości tekstu (N). - Prace filologiczne. Tom XX. Warszawa, 1970, str. 65-70.

Somers H.H. Analyse mathématique du langage. Lois générales et mesures statistiques. - Louvain: Nauwelaerts, 1959.

Tuldava J. Sagedussõnastik leksikostatistilise uurimise objektina. - Rmt.: Töid keelestatistika alalt, II. Tartu, 1977, lk. 141-171. (Учен. зап. Тартуского ун-та, вып.413).

Villup A. A.H. Tammsaare romaani "Tõde ja õigus" I köite autori- ja tegelaskõne sagedussõnastik. - Eesti keele sõnavarastatistika probleeme. Töid keelestatistika alalt, III. Tartu, 1978, lk. 5-106. (Учен. зап. Тартуского ун-та, вып.446).

Weibull W. A Statistical Theory of Materials. - Proceedings of the Royal Academy of Engineering Sciences, 1939, No.15.

Weitzmann M. How Useful is the Logarithmic Type-Token Ratio? - Journal of Linguistics, No. 7. London, 1971, pp. 237-244.

Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge, Mass., 1949.

ON THE ANALYTICAL EXPRESSION OF THE RELATION BETWEEN SIZE OF VOCABULARY AND SIZE OF TEXT

Juhan Tuldava

S u m m a r y

In this article various empiricial and theoretical approaches to the problem of the relation between the size of vocabulary (V) and the size of text (N) are discussed. The author points out the insufficiency of the existing "complex" models in solving practical linguostatistical problems where greater exactness and reliability are needed (stylo-statistical analysis, text attribution, extrapolation beyond the limits of observable data, etc.). Instead of the complex models a direct method is proposed where the relation between V and N is regarded as the primary component with its own immanent properties in the statistical organization of text. The relation between V and N has to be analysed on the background of some essential inner factors of text generation. The dynamics of vocabulary growth is considered as the result of the interaction of several linguistic and extralinguistic factors which in an integral way are governed by the principle of "the restriction of variety" of lexics (an analogue of the principle of the decrease of entropy in self-regulating systems). The

concept of the variety of lexics is defined as the relation between the size of vocabulary and the size of text in the form of V/N (type-token ratio, or coefficient of variety) or N/V (average frequency of word occurrences). The coefficient of variety is supposed to be correlated with the probabilistic process of choosing "new" (unused) and "old" (already used in the text) words at each stage of text generation. The steady decrease of the degree of variety $V/N = p$ is attended by the increase of its counterpart: $(N - V)/N = 1 - V/N = q$ $(p + q = 1)$, which can be interpreted as "the pressure of recurrency" of words in real texts (analogous to the concept of redundancy in the theory of information). The formulae of the relation between V and N are constructed from the basic models $V = Np$ or $V = N(1 - q)$. For this purpose the quantitative changes of $V/N = p$ depending on the size of text are analysed.

According to the initial hypothesis the relation between V/N and N is approximated by the power function of the type: $V/N = aN^B$ (a and B are constants: $B < 0$), which leads to the well-known formula of G. Herdan: $V = aN^b$ (where $b = B + 1$). A verification shows good agreement with empirical data in the initial stages of text formation (in the limits of about 4.000 - 5.000 tokens which correspond to a short communication). Later on the rate of the diminishing of the degree of variety gradually changes. Accordingly the initial formula has to be modified and this can be done by logarithmization of the variables. The first attempt gives us $\ln (V/N) = aN^B$, which leads to some variants of the Weibull distribution. This kind of distribution shows good agreement with the empirical data within the boundaries of a text of medium length but it is not good for extrapolation. Only after balancing the initial formula by the logarithmization of both variables we obtain $\ln (N/V) = a(\ln N)^B$ and the corresponding formula for expressing the relation between V and N:

$$V = Ne^{-a(\ln N)^B},$$

or $V = N^{1 - a(\ln N)^b}$ (where $b = B - 1$), which turns out to be the most adequate formula for solving our problems. (All formulae are to be considered as independent, i.e. not derived from each other.)

The good descriptive power of the given function and the possibilities of extrapolation in both directions (from the beginning up to a text of about $N = 10^7$) are demonstrated on the basis of experimental material taken from 9 languages belonging to different typological groups (Estonian, Kazakh, Latvian, Russian, Polish, Czech, French, Rumanian, English: see Tables 1 - 6).

It could be added that the function may be applied to the analysis of individual texts as well as composite homogeneous (similar) texts and the size of vocabulary may be determined by counting either word forms or lexemes. This seems to corroborate the assumption about the existence of a universal law (presumably of phylogenetic origin) which governs the process of text formation on the quantitative level.

ЧАСТОТНЫЕ СЛОВАРИ ПОДЪЯЗЫКОВ НАУКИ И ТЕХНИКИ - КЛЮЧ К ПОНИМАНИЮ СПЕЦИАЛЬНЫХ ТЕКСТОВ

Лотар Хоффманн

Начиная с 1970 года издательство "Энциклопедия" в Лейпциге выпустило серию частотных словарей на трех языках - русском, английском, французском - под общим названием "Fachwortschatz", т.е. "Специальная лексика" . До сих пор были изданы словари по медицине, по физике, по химии, по математике, по строительному делу и по животноводству/ветеринарии (Hoffmann, L., Hrsg., 1970-1978). Другие, напр., по общественным наукам, должны появиться в ближайшем будущем. В состав авторского коллектива входят, прежде всего, научные сотрудники и преподаватели секции иностранных языков Лейпцигского университета имени Карла Маркса.

Словари составлены по единому принципу. Они содержат в себе наиболее часто употребляемых 1100-1200 лексических единиц русских, английских и французских текстов по данной специальности. Включенные в списки слова были выделены на ЭВМ на основе достоверной выборочной процедуры: качественной - удельный вес отраслей соответствующих специальностей, различных жанров научно-технической литературы и т.п. и количественной - объем.

Лексический материал каждого из трех языков располагается в двух списках по принципу частотности и по алфавиту, причем во втором списке приводятся и немецкие эквиваленты. Обзор самых продуктивных словообразовательных аффиксов дополняет лексические данные словарей.

В частотных списках слова располагаются по принципу убывающей частоты. Роль классификационного критерия играют здесь и ранг и относительная частота или вероятность лексических единиц. Приведение относительных частот облегчает их кумулирование и, тем самым, установление покрываемости текста отдельными лексическими единицами или определенным их количеством. Из частотных списков видна как вся совокупность 1100-1200 слов, так и позиция каждого отдельного слова, его роль в статистической структуре спецтекстов, богатство и разнообразие словарного фонда подъязыков науки и техники. Они обеспечивают прочный фундамент для сопоставительного анализа

терминологических систем и структур текста. Все эти аспекты играют чрезвычайную роль в работе лингвистов, авторов учебников и преподавателей иностранных языков, в упорядочении терминологии, в информатике и т.д.

Содержащийся в словарях материал предстает перед пользователем во второй раз уже в алфавитном порядке. Он используется, в первую очередь, тогда, когда интерес направлен на отдельное слово, его ранг и относительную частоту, на эквивалент в родном языке, т.е. он служит руководством не только исследователю и преподавателю, но одновременно и учащемуся, желающему знать и усвоить необходимую ему минимальную специальную лексику.

Опубликованием рангового списка самых продуктивных в словообразовании трех языков аффиксов преследуется цель повысить эффективность частотного словаря-минимума в 1100-1200 слов или, другими словами, покрываемость текста его лексическими единицами.

В словари вошли не все части речи: отсутствуют имена числительные, частицы, междометия и имена собственные из-за лингвистических и лингводидактических соображений. Пропущены и все представители других частей речи, встречавшиеся в анализированном корпусе текстов с частотностью $f < 0.000114$ и в ограниченном числе выборок.

Исходя из классификации П.М. Алексеева (1968), можно сказать, что здесь речь идет о смешанном типе "малого", отраслевого, неполного частотного словаря исходных форм лексем, полученного путем выборочного анализа, с указанием ранга и относительной частоты.

Весьма ограниченное количество зарегистрированных в словарях вышеупомянутой серии лексем имеет огромное текстообразовательное, конституирующее значение. Кумулирование их относительных частот приводит к средней 86-процентной покрываемости текста. Практическое испытание эффективности наиболее часто встречающихся в спецтекстах слов дает, как правило, лучшие результаты, в частности, тогда, когда учитываются и раскрываемые по моделям словообразования (деривация, словосложение, словосочетание; интернационализмы, искусственные символы, логико-математические выражения, собственные имена и др.) лингвистические единицы, составляющие 5-7 процентов знакового состава публикаций научно-технической литературы.

Именно здесь проявляется полезность статистического ана-

лиза лексики и опубликования частотных словарей. Они представляют научно обоснованный выбор коммуникативно существенной части лексики из общего словарного фонда, который во многих языках, включая научную терминологию и номенклатуру, составляет сотни тысяч лексических единиц. Всем словарным составом родного языка не владеет даже самый образованный человек, всему словарному фонду иностранного языка не может обучить ни школа, ни университет. Да это и никому не нужно. Большая часть элементов лексической системы любого языка никогда никем не используется.

Актуализация лексических средств языковой системы в речевой коммуникации определяется, в первую очередь, внелингвистическими факторами. Она зависит, прежде всего, от цели и содержания высказывания, которые обусловливаются общественно-продуктивной деятельностью человека. Из этого внеязыкового ограничения и вытекает и сильное сужение коммуникативных потребностей, влияющее, в конечном счете, и на употребление лексических средств. Любой (ограниченной) сфере деятельности, таким образом, соответствует (ограниченная) сфера коммуникации.

Запас слов, в котором нуждается отдельное лицо или целый коллектив в определенной сфере коммуникации, может быть составлен с помощью двух методов, коренным образом отличающихся друг от друга: первый основывается на систематическом анализе предметов и понятий, которые затем соотносятся с адекватными наименованиями; второй исходит из статистической оценки, т.е. из встречаемости лингвистических элементов в речевом потоке.

Данные, полученные от двух исследовательских процедур, конечно, разные. Из систематического анализа возникают тематические и терминологические списки, в которых приведены, прежде всего, имена существительные и другие полнозначные слова, отличающиеся высоким содержанием (узко) специальной информации. Здесь нередко заложены зачатки отраслевых словарей и тезаурусов.

В частотные списки, с другой стороны, входят, в первую очередь, неполнозначные, служебные слова, занимающие в качестве структурных элементов важные, ключевые позиции в тексте, и, кроме того, многие другие слова из самых разных сфер коммуникации, т.е. довольно общая лексика с низкой информационной нагрузкой. Количество информации в верхней зоне час-

тотного словаря повышается только тогда, когда границы специальности и, вследствие этого, подъязыка очень четкие и узкие, чего в принципе и достигла обсуждаемая здесь серия.

Из этого, однако, не следует, что, если ограничиться определенной специальностью, можно обойтись одной только тематически важной, терминологической лексикой. Ее высокая информативность является, прежде всего, имплицитным признаком отдельной лексемы, а не эксплицитным фактором связной организации слов и высказываний в тексте. Если выше было установлено, что наиболее часто употребляемые лексические единицы м о г у т выступать во многих различных сферах коммуникации, то здесь следует уточнить, что они д о л ж н ы там употребляться, чтобы обеспечить формальную и смысловую связь между основными носителями информации. Без них функционирование речевого общения оказалось/бы невозможным или, по крайней мере, несовершенным.

Итак, оба вида лексических единиц вместе создают необходимые предпосылки и для производства, и для понимания текстов научно-технической прозы как на родном, так и на иностранном языках. Оба метода анализа лексики следует поэтому признать равноправными и комплементарными.

Но это еще не все. Роль наиболее часто употребляемой лексики в овладении иностранным языком совсем иная, чем в использовании родного языка. На родном языке каждый знает эту лексику или, точнее говоря, несознательно свободно пользуется ей. Таким образом, главное внимание говорящего/пишущего сосредоточивается на носителях информации, на терминах, искусственных знаках-символах, формулах и т.п. Изучающему иностранный язык на начальном этапе не известны ни часто, ни редко употребляющиеся слова.

Часто встречающиеся слова имеют, однако,преимущество по сравнению с редко употребляющимися, заключающееся в том, что их количество совсем ограничено, между тем как их повторяемость, их доля, их конституирующая сила в тексте велики. Они образуют, так сказать, основной словарный фонд или базовый язык, который можно усвоить в самое короткое время и текст за текстом небольшими порциями обогащать его носителями информации.

Носители информации в своем общем количестве так многочисленны, даже если учесть только терминологию одной специальности, как, напр., медицина, химия, электротехника, маши-

ностроение и т.д., что их усвоение остается бесконечным, "вечным" процессом, впрочем и потому, что открытые терминологические системы постоянно обновляются. Следовательно, элементы частотного словаря предназначены для запоминания в памяти словаря вместе с ограниченным количеством тематически важных слов на ранних этапах преподавания иностранных языков; специальному словарю, напротив, суждено служить незаменимым помощником в самостоятельной работе долгие годы, если не всю жизнь. Иными словами, частотный словарь является абсолютно необходимым минимумом; специальный словарь представляет собой практически неисчерпываемый максимум лексики.

Растущее значение науки и техники во всех областях общественной жизни заставляет лингвистов более основательно рассматривать их подъязыки, стиль научной речи (Hoffmann, L., 1976а). Кроме того, и обучение иностранным языкам не может не учитывать обусловленной научно-техническим прогрессом дифференциации в использовании языковых средств, если оно намерено построить прочный фундамент международного сотрудничества.

Вследствие этого частотные словари, составленные на материале художественных произведений и печати (Josselson,H.H., 1953; Eaton, H.S.,1961; Juilland, A. et al., 1970; Штейнфельдт Э.А., 1963; и др.), могут, в известной мере, служить основой для выбора лексического минимума общеобразовательной школы, но ни в коем случае для преподавания подъязыков науки и техники в университетах, вузах и техникумах. Сопоставление с частотными словарями разных подъязыков науки и техники обнаруживает максимальное совпадение на 25-30 процентов в верхней зоне первых 1100-1200 единиц; на более низких рангах оно постоянно уменьшается.

Из этого сопоставления явствует и такой факт, что абсолютного порога между часто и редко употребляемыми словами по отношению к языку в целом не существует. За исключением нескольких лексических единиц, встречающихся в любом тексте и занимающих поэтому одно из первых мест во всех частотных списках, употребляемость других элементов словника зависит от тематики, от предмета текста. Чем специальней тематика, тем важнее роль терминологии, тем чаще на первый план выступает редко употребляемая лексика.

Общеязыковый частотный словарь (если такой действительно существует), по этим причинам, не проникает в глубь кон-

кретного текста, покрывает лишь незначительную часть его. То же самое относится и к частотным словарям, базирующимся на беллетристике и публицистике в отношении спецлитературы.

Напрашивается вывод, что точно ограниченная сфера коммуникации является наилучшей предпосылкой для успешного подбора выборок, причем подъязыки науки и техники лучше всех отвечают этому требованию. В таких условиях в состав частотного словаря входят только те лексические единицы, которые заслуживают внимания представителей данной специальности. Внутренняя структура частотного словаря тем лучше отражает статистическую структуру текста, чем правильнее распределяются выборки по всем представительным текстам специальности, т.е. чем яснее они воспроизводят специальность со всеми ее отраслями и существующими между ними пропорциями.

Именно эту цель преследует серия "Fachwortschatz". Успешность проложенного пути подчеркивается высокой покрываемостью текстов и большой долей узкоспециальной лексики. Частотные словари исследованных до сих пор подъязыков науки и техники удовлетворяют требованиям лингвистов, ожидающих от них вероятностного прогноза употребления языковых единиц в речи и организации разных видов текста. Они в то же время оказываются бесценным подспорьем для авторов учебников и преподавателей иностранных языков в подборе элементов для учебных минимумов, в составлении учебных программ, в разработке алгоритмов введения и повторения нового материала, в установлении специфики и трудностей спецтекстов, в их адаптации и комментировании. Они становятся, как это сформулировано в заглавии статьи, ключом к специальным текстам.

Составители частотных словарей придерживаются различных мнений, когда речь идет о принципах расположения материала, о достоверности абсолютных и относительных частот в связи с количеством и объемом выборок. Принципы, по которым разработана и построена обсуждаемая серия, изложены в ряде монографий и статей (Hoffmann, L., 1971 и 1975; Hoffmann, L., Piotrowski, R.G., 1979). Не будем касаться этих принципов, ибо это выходит за рамки данной работы.

Характеризуя частотные словари, изданные в Лейпциге, в более общих чертах, можно сказать, что они представляют собой пособия, фиксирующие современный уровень и тенденции развития избранных подъязыков науки и техники, способствующие их совершенствованию и овладению ими.

ЛИТЕРАТУРА

Алексеев П.М. Частотные словари и приемы их составления. - В кн.: Статистика речи. Л., 1968.

Штейнфельдт Э.А. Частотный словарь современного русского литературного языка. Таллин, 1963.

Eaton, H.S. An English, French, German, Spanish Word Frequency Dictionary. New York, 1961.

Hoffmann, L. (Hrsg.) Fachwortschatz Medizin. Leipzig, 1970. (4. Aufl., 1978.)

Hoffmann, L. (Hrsg.) Fachwortschatz Physik. Leipzig, 1970. (3. Aufl., 1976)

Hoffmann, L. (Hrsg.) Fachwortschatz Bauwesen. Leipzig, 1976.

Hoffmann, L. (Hrsg.) Fachwortschatz Tierproduktion/Veterinärmedizin. Leipzig, 1978.

Hoffmann, L. Häufigkeitswörterbücher der Subsprachen von Wissenschaft und Technik. - In: Fachsprachen und Sprachstatistik. Berlin, 1975.

Hoffmann, L. Zur maschinellen Bearbeitung sprachlicher Daten bei linguostatistischen Untersuchungen. - Deutsch als Fremdsprache, 1971, H. 2.

Hoffmann, L. Kommunikationsmittel Fachsprache.Berlin,1976(a).

Hoffmann, L., Piotrowski R.G. Beiträge zur Sprachstatistik. Leipzig, 1979.

Josselson, H.H. The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian. Detroit, 1953.

Juilland, A., Brodin, D., Davidovitch, C. Frequency Dictionary of French Words. The Hague; Paris, 1970.

HÄUFIGKEITSWÖRTERBÜCHER DER SUBSPRACHEN VON WISSENSCHAFT UND TECHNIK - EIN SCHLÜSSEL ZUM FACHTEXT

Lothar Hoffmann

R e s ü m e e

Im vorliegenden Aufsatz wird die während der Jahre 1970 bis 1978 von einem Forschungskollektiv der Sektion Fremdsprachen der Karl-Marx-Universität Leipzig veröffentlichte Reihe von Häufigkeitswörterbüchern für die Fachrichtungen Medizin, Physik, Chemie, Mathematik, Bauwesen und Tierproduktion/Veterinärmedizin (Russisch, Englisch, Französisch) vorgestellt, deren Weiterführung beabsichtigt ist.

Die Wörterbücher sind nach einem einheitlichen Muster aufgebaut. Sie enthalten die 1100 bis 1200 häufigsten lexikalischen Einheiten der genannten Fachdisziplinen in Häufigkeitslisten und alphabetischen Verzeichnissen mit den deutschen Äquivalenten, ergänzt durch Übersichten über die produktivsten Affixe der fachsprachlichen Wortbildung. Sowohl die Kumulierung der relativen Häufigkeiten als auch praktische Tests haben eine günstige Textdeckung ergeben.

Wendet man die Klassifizierung von P. M. Alexejev an, dann handelt es sich um den kombinierten Typ eines kleinen Spezialwörterbuches, das keinen Wert auf Vollständigkeit legt, sein Material dem Stichprobenverfahren verdankt und neben den Stichwörtern in ihrer Grundform Angaben zu deren relativer Häufigkeit enthält.

Erörtert werden hier in erster Linie das Verhältnis von häufiger und "seltener" (thematisch relevanter) Lexik bei der Konstituierung des Textes, die unterschiedlichen Funktionen von Häufigkeits- und Fachwörterbüchern sowie die Methoden zu deren Zusammenstellung. Ein besonderer Aspekt ist ihre Nutzung bei der Erlernung und Verwendung der Subsprachen von Wissenschaft und Technik als Fremdsprachen unter den Bedingungen des stürmischen wissenschaftlich-technischen Fortschritts und der wachsenden internationalen Zusammenarbeit in unserer Zeit.

СОДЕРЖАНИЕ

Алексеев П.М., Копылова М.Е. О серийных двуязычных учебных частотных словарях 3
Вашак П. Диалектика типологии и атрибуции 14
Григорьева А.С. О частотном словаре русской обиходной письменной речи 25
Калинина Е.А. К вопросу об использовании распределительного словаря для решения лингвометодических задач 32
Кишинская Л.Г., Кишинский С.Б. Выявление опережающей роли поэтической речи в развитии языка с помощью корреляционного анализа 47
Коккота В. Частотность английских глагольных форм в научно-технической литературе 59
Краснова Э.А. Лексико-семантические особенности жанра 73
Милюна М., Якубайтис Т. Автоматизированная система подбора вида распределения лингвистических единиц 86
Остапенко В.Е. Параболические и гиперболические зависимости текстообразования 106
Тулдава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста 113
Хоффманн Л. Частотные словари подъязыков науки и техники - ключ к пониманию специальных текстов ... 145

SISUKORD - INHALTSVERZEICHNIS - CONTENTS

Alekseyev, P., Kopylova M. On Serial Bilingual Freguency Dictionaries. - Summary in English . . . 13
Vašák, P. Dialectique de la typologie et de l'attribution. - Résumé en francais 24
Grigoryeva, A. On the Frequency Dictionary of Everyday Written Russian. - Summary in English . . . 31
Kalinina, E. On the Use of Distributional Dictionaries for Solving Linguo-methodical Tasks. - Summary in English 46
Kishinskaya, L., Kishinsky, S. On the Outstripping Role of Poetical Speech in the Development of Language (Statistical Analysis). - Summary in English 58
Kokkota, V. Relative Frequencies of English Verb Forms in Scientific Literature. - Summary in English 72
Krasnova, E. Lexical and Semantic Distinctions of a Genre. - Summary in English 85
Milūnà, M., Jakubaite, T. An Automatic System for Selecting the Type of Distribution of Linguistic Units. - Summary in English 105
Ostapenko, V. Parabolic and Hyperbolic Regularities of Text Formation. - Summary in English 112
Tuldava, J. On the Analytical Expression of the Relation between Size of Vocabulary and Size of Text. - Summary in English 143
Hoffmann, L. Häufigkeitswörterbücher der Subsprachen von Wissenschaft und Technik - ein Schlüssel zum Fachtext. - Resümee in Deutsch 152

ЛИНГВОСТАТИСТИКА И КВАНТИТАТИВНЫЕ ЗАКОНОМЕРНОСТИ ТЕКСТА.
Труды по лингвостатистике VI.
На русском языке.
Резюме на английском, французском и немецком языках.
Тартуский государственный университет.
ЭССР, г. Тарту, ул. Юликооли, I8.
Ответственный редактор Ю.Тулдава.
Корректор Н.Чикалова.
Подписано к печати 06.11.1980.
МВ-09772.
Формат 30х45/4.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 9,I3.
Печатных 9,75.
Тираж 500.
Заказ № II87.
Цена I руб. 40 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Пялсона, I4.