ISSN 0207-4435

TARTU RIIKLIKU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

628

KEELESTATISTIKA JA ARVUTUSLINGVISTIKA

ЛИНГВОСТАТИСТИКА И ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА

Töid keelestatistika alalt
Труды по лингвостатистике

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
AL'JSTATUD 1893.a. VIHIK 628 ВЫПУСК ОСНОВАНЫ В 1893.г.

KEELESTATISTIKA JA ARVUTUSLINGVISTIKA

ЛИНГВОСТАТИСТИКА И ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА

Töid keelestatistika alalt

Труды по лингвостатистике

ТАРТУ 1982

ТЕКСТ И ЯЗЫК - ЦЕЛОСТНОСТЬ И ОРГАНИЗМЕННОСТЬ

М.В. Арапов

1. В данной статье предлагается модель количественной организации языка на лексическом уровне. Здесь мы хотели бы изложить л о г и к у нашего подхода - что и с помощью каких средств предлагается объяснить, - а детали математического аппарата и анализ эмпирического материала читатель найдет в цитируемых публикациях. Как и каждая модель, данная предполагает существенное упрощение реальной картины, но должна позволять делать содержательные выводы. В данном случае - вычислять на основе одних количественных характеристик языка другие.

Основными понятиями модели будут ц е л о с т н о с т ь - свойство текста, которое, в частности, позволяет находить многочисленные параметры его лексической структуры на основе минимального набора исходных данных, - и о р г а н и з м е н н о с т ь - согласованность лексических структур всех текстов одного языка. Реальные тексты только в большей или меньшей степени приближаются к целостным, а согласование нарушается многими факторами.

2. Начнем с анализа понятия целостности, причем отправной точкой для нас будут эксперименты по так называемой свободной классификации (Звонкин, Фрумкина, 1980).

Пусть нам дана коллекция M из L предметов (обычно несколько десятков), - например, пуговиц или монет. Эти предметы должны отличаться друг от друга настолько, чтобы число возможных оснований классификации было большим и у испытуемого не возникало мысли, что экспериментатор имеет в виду какую-то определенную классификацию, а задача его сводится к тому, чтобы угадать, какую именно. С другой стороны, коллекция M должна быть достаточно однородной, т.е. не должна провоцировать испытуемого разделить ее на несколько частей и классифицировать эти части порознь.

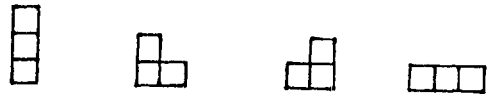
Задача испытуемого состоит в том, чтобы разложить M на отдельные "кучки" - непересекающиеся классы, причем ни число, ни размеры классов не фиксированы. Опишем, как осуществляется процесс классифицирования, сознательно при этом пре-

небрегая деталями.

Испытуемый берет предмет x (□) и кладет его в "кучку" X_I. В ячейке классификации X_I – это первый объект. Дальше испытуемый мог положить второй объект в ту же ячейку X_I или "основать" новый класс X_2. Т.е. на втором шаге его классификация могла выглядеть как (а) или как (б):

(а) (б)

На третьем шаге объект мог быть положен в классы X_I, X_2 или основан класс X_3, т.е. возникает одна из четырех диаграмм (в), (г), (д) и (е):



(в) (г) (д) (е)

Из цитируемой работы видно, что для описания результатов экспериментов по свободному классифицированию совершенно не существенна нумерация классов, а поэтому диаграммы типа (г) и (д), отличающиеся только нумерацией классов, мы можем не различать. Договоримся, что на диаграммах мы будем всегда располагать классы в порядке невозрастания, и все диаграммы, которые при таком способе изображения совпадут, будем считать одной диаграммой. Такое соглашение не только упрощает интерпретацию результатов, но в определенном смысле устраняет асимметрию классов-ячеек (X_I, X_2,...) и классов, составленных из объектов, которые попали в "свой" класс X_I, X_2 ... X_N первым, вторым, k-тыми (такие классы мы будем обозначать $У_I$, $У_2$...$У_p$(на приведенном ниже рисунке 1 два У-класса

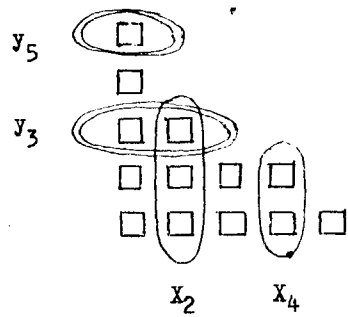


Рис.1

выделены двойной линией. Для описания результатов экспериментов нам понадобятся оба вида классов.

Заметим, что произвольная пара из одного X- и одного У-класса либо не пересекается, либо содержит в пересечении ровно один элемент. Рассмотрим наименьший У-класс, пересекающийся с данным X-классом из n элементов. Пусть объем этого У-класса m элементов. Легко видеть, что если среди X-классов нет классов одинакового объема, то m совпадает с рангом r класса X, т.е. количеством X-классов, имеющих объем $\geqslant n$. Способ обобщить понятие ранга r на случай классов, совпадающих по объему, см. Арапов, Ефимова, Шрейдер, 1975.

Число классов объемов ровно в j элементов выражается через разность объемов У-классов: $\Delta m_j = |Y_j| - |Y_{j+1}| = m_j - m_{j+1}$

Если описывать результаты экспериментов в виде монотонной диаграммы, то легко заметить следующее.

1). Независимо от того, что подвергалось классификации - пуговицы или точки на плоскости, - и кто ее осуществлял, диаграммы будут похожими.

2). Испытуемые не стремились к тому, чтобы классы X (и тем самым У) были "равнозаполненными". Но они старались, чтобы между размерами классов не было слишком больших разрывов, т.е. классы образовали н е п р е р ы в н у ю (по какой-то шкале) последовательность.

В одном из экспериментов объекты (пятикопеечные монеты) были подобраны так, чтобы наиболее очевидный способ их классифицировать приводил к появлению класса, который был много больше остальных. "После эксперимента испытуемых просили дать самоотчет. В частности, экспериментатор задавал вопрос: "Почему вы не положили все пятаки [с данным признаком - М.А.] вместе?" На это, как правило, следовал стандартный ответ: "Потому что их для этого слишком много" (Звонкин, Фрумкина, 1980, с. 4).

3). Испытуемые как бы отталкивались от двух крайностей: сложить все объекты в одну ячейку и образовать для каждого объекта свой класс (т.е. построить один У-класс). В обоих случаях им казалось, что поставленная перед ними задача не выполнена. Решая ее одни испытуемые тяготели к большей степени детализации (их диаграммы вытянуты по оси X, т.е. X-классы малы), а другие - к большей степени обобщения - у них диаграммы были вытянуты по оси У, У-классы малы.

4). Этот вывод дан в работе скорее намеком: выделяя из признаков существенные, те, по которым строились Х-классы, испытуемые руководствовались как бы "сверхпризнаком" - совершенством полученной классификации, которая проявлялась в соразмерности выделенных Х и У-классов.

3. Ситуации, которые интересуют нас в данной статье, несколько отличаются от ситуации, которая исследовалась в экспериментах по свободной классификации. В интересующих нас случаях объекты, с одной стороны, не заданы заранее: они возникали и одновременно классифицировались, а, с другой стороны, сама совокупность не была случайной коллекцией, а результатом системно организационной деятельности. Создание текста на естественном языке - очевидно, один из видов системно организационной деятельности.

Результаты классификации таких коллекций более устойчивы по сравнению с классификациями, полученными в ходе рассмотренных в п. 2 экспериментов. Устойчивость проявляется в том, что построенная диаграмма сохраняет свою форму независимо от природы коллекции и субъекта, производящего классификацию. Под формой классификационной диаграммы понимается некоторое инвариантное, не зависящее от объема коллекции M, соотношение размеров Х- и У-классов.

Достаточно простое и единообразное описание этой формы потребует предварительного, вообще говоря, нелинейного преобразования системы координат. Точный вид этого преобразования, которое зависит от природы рассматриваемых коллекций и способа классифицирования, до сих пор является предметом споров (Арапов, 1977). Не задерживаясь на этом вопросе, предположим, что соответствующее преобразование выполнено, а тем самым параметры в формулах, которые обычно используются для описания подобных диаграмм (см. ниже), оказались равными, соответственно 1 или 0.

Рассматриваемые в ретроспективе попытки описать форму диаграммы напоминает притчу о слепцах, которые с разных сторон ощупывали слона. Так, анализируя тексты на естественном языке, стенограф Эсту (Estoup, 1916), а вслед за ним Кондон (Condon, 1928) обратили внимание на соотношение объемов n_r наиболее крупных Х-классов и пришли к зависимости, которая позже (и не совсем справедливо) получила название "закона Ципфа":

$$n_r \cdot r \approx const \qquad (1)$$

На совсем другом биологическом материале к той же зависимости пришел английский ботаник Виллис (Willis, 1922).

При этом ни один из этих исследователей не обратил внимание, что за два с лишним десятка лет до них известный итальянский экономист В.Парето (Pareto, 1897) описал диаграмму той же формы. Только его интересовали как раз Х-классы с наименьшим объемом. Он нашел соотношение между числом классов Δm_k, содержащих ровно k объектов, $k = 1,2,\ldots$:

$$\Delta m_k k^2 \approx const \qquad (2)$$

В той же форме нашел данную зависимость А.Лотка (Lotka, 1926), но его коллекция состояла из научных работ, а в отдельную ячейку классификации он собирал работы, написанные одним ученым. Ту же зависимость для употребления редких слов нашел Дж.Ципф (Zipf, 1935).

Несколько лет спустя библиограф Бредфорд (Bradford, 1934), классифицируя научные работы, но уже по месту их публикации (периодическим изданиям), приходит к той же диаграмме, но описывает ее в виде

$$N_r (\log r)^{-1} \approx const, \qquad (3)$$

где под N_r имеется в виду суммарный объем первых по рангу классов $N_r = \sum_{n_i \le n_r} n_i$

Приведенный перечень способов описать возникающую при классификации "гиперболическую лестницу" (удачное выражение Л.С.Козачкова, 1973) не претендует на особенную полноту. Но он значительно бы вырос, если бы мы учитывали не только те способы описания, которые полностью эквивалентны, но и способы эквивалентные только приближенно. Более полный перечень классификаций, в которых появлялась диаграмма, описываемая зависимостями (1)-(3), приводится, например, в работах Арапов, Ефимова, Шрейдер, 1975, Haight, 1966.

История, как возникло и укрепилось убеждение, что упомянутые здесь зависимости описывают одну и ту же диаграмму, заслуживает, наверное, более подробного рассказа (частично это сделано, Арапов-Либкинд, 1982), но сейчас нам важно отметить, что, хотя реальные диаграммы удовлетворяют каждой из этих зависимостей, ни одна из этих зависимостей не содержит всей информации о диаграмме. Так зависимости (1) может удовлетворять и диаграмма, состоящая из одного Х-класса (с рангом 1),

а зависимости (2) - диаграмма из единственно У-класса, т.е. результат такой аномальной классификации, когда каждый объект попадает в собственный единичный класс.

Существенно большую информацию о диаграмме можно извлечь из зависимостей (1) и (2), если потребовать, чтобы они выполнялись одновременно. Тогда видно, что диаграмма зеркально симметрична относительно биссектрисы координатного угла, и при зеркальном отражении Х-класс ранга переходит в У-класс, содержащий объекты, которые попали в свой Х-класс тыми по порядку. Так, первому по величине Х-классу соответствует класс объектов, которые в свой Х-класс попали первыми: например: самое частое слово в тексте употребляется столько раз, сколько в тексте различных слов. Эту симметрию очень легко нарушить. Возьмем текст, которому соответствует "зеркальная" диаграмма и удвоим его. Диаграмма для удвоенного текста заведомо не будет симметричной.

Более детальное описание классификационной диаграммы мы дадим в п. 7, но закон (пока еще не сформулированный), которому удовлетворяют все такие диаграммы, мы будем называть, придерживаясь традиции, законом Ципфа.

5. Уже первым исследователям закона Ципфа стало понятно, что этот закон не похож на предельные распределения математической статистики. Само по себе увеличение размеров выборки отнюдь не гарантировало, что закон будет выполняться более точно. Сам Ципф признавал, что закон выполняется только на определенном классе выборок - не слишком малых и не чересчур больших (Фрумкина, 1964). Однако, как мы увидим дальше, дело не в размерах, а, если так можно сказать, в "качестве" выборки.

Но странным поведение закона Ципфа, в частности, его плохая воспроизводимость (Арапов, Либкинд, 1982), кажется только в рамках определенной модели текстообразования, принятой в лингвистике (Лесохин, Лукьяненков, Пиотровский, 1982, с. 186-7)*. Эта модель состоит в том, что с каждой клеткой классификации X_i связывается определенная вероятность $p(X_i)$ попадания классифицируемого объекта в эту клетку. В рамках этой модели появление классификационной диаграммы и ее устойчивость объясняется устойчивостью вероятностей $p(X_i)$ в гене-

* Более подробно о "классической" модели см.: Арапов, Ефимова, Шрейдер, 1975; Арапов, 1981.

ральной совокупности (речи). Вероятность p_i определяет объем соответствующей клетки классификации. В каждом отдельном тексте-выборке эта классификация воспроизводится со случайными отклонениями, размер которых определяется законом больших чисел. С данной точки зрения, язык в его количественном аспекте будет полностью описан, если будет собрано достаточно эмпирических данных, чтобы для каждой клетки X_i с именем i оценить p_i с требуемой точностью. В рамках данной модели не имеет особого значения то, что вероятности p_i образуют последовательность с общим членом, совпадающим с законом Ципфа.

Модель, которой мы коснулись, подразумевает определенную научную программу, причем шагом в ее выполнении является, например, составление частотного словаря.

Выполнение этой научной программы позволило собрать ценную информацию о языке, хотя сама модель приводит к логическим трудностям и ее трудно проверить (см. Арапов, Ефимова, Шрейдер, 1975; Арапов, 1981). Не останавливаясь сейчас на этих трудностях, мы покажем только, что в рамках этой модели нельзя объяснить результаты экспериментов по свободной классификации.

В ходе этих экспериментов классифицируются предметы, имеющие характерные признаки. Вероятно, год выпуска для монеты или число отверстий для пуговицы - имеют устойчивую вероятность проявления в совокупности всех монет или пуговиц. Но к результатам классификации эти признаки имеют лишь косвенное отношение: если классификация получается несоразмерной, то испытуемый готов игнорировать эти признаки, чтобы сохранить свойства целого - форму классификации. И поэтому при ее описании априорные признаки - имена ячеек классификации (вроде года выпуска для монет) - не играют роли.

Обсуждая эксперименты по свободной классификации, имеет смысл говорить о вероятности, но не о вероятности попадания объекта в определенную ячейку, а о вероятности, что классификация в целом будет иметь определенную форму.

Но как только мы переходим к поиску конкретных зависимостей, характеризующих эту форму, так сразу встает вопрос: "А что же подвергается классификации?".

6. В 70-е годы исследователями, занимающимися различными предметными областями, было высказано убеждение, что закон

Ципфа описывает количественную структуру не случайной выборки, а целостной совокупности (например, географического региона с развитыми внутренними экономическими связями - Jonson, 1980; тематически связанной совокупности научных документов - Арапов-Либкинд, 1977). По отношению к языку соответствующая гипотеза была впервые выдвинута Ю.К.Орловым (Орлов, 1970), который предположил, что в качестве такой совокупности можно рассматривать полный текст совершенного литературного произведения в отличие от отрывка из такого произведения или корпуса, составленного из таких текстов. Причем, если мы имеем дело с "ципфовским" текстом, то достаточно ничтожного числа основных характеристик, чтобы полностью описать количественную структуру, но если текст не является целостным и не подчиняется закону Ципфа, то его структура может сильно варьироваться от случая к случаю, и для нее уже не удается найти такого простого описания. (Позднее, Ю.К.Орлов предложил более сложную модель, в которой структура широкого круга "нецифовских" текстов описывается как результат регулярной "деформации" ципфовского текста; см., например, Орлов, 1978).

Сравнительно простыми средствами можно показать, что конгломерат текстов и целостный текст тех же размеров устроены по-разному. Например, учитывая закон Ципфа и все дополнительные известные сейчас данные о значении его параметров для определенных языков, жанров и даже авторов, можно попытаться экстраполировать на конгломераты текстов полученные на замкнутых текстах соотношения объема текста и словаря.

Возьмем пушкинскую "Пиковую даму", ее длина - 6606 словоупотреблений, объем словаря - 1863 слова. По данным "Словаря Пушкина" (Фрумкина, 1964) всего А.С. Пушкин в своем творчестве использовал примерно 21 тыс. различных слов. Если бы он использовал их так, как в "Пиковой даме", то написал бы около 100 000 слов текста. В действительности пушкинское наследие в 5,5 раза больше.

Дж.Джойс в своем известном романе "Улисс" использовал 29 тыс. различных слов. Объем его романа, для которого закон Ципфа выполнен довольно точно, - 260 тыс. слов. В.Шекспир, использовав словарь на несколько сот слов меньше словаря "Улисса", но написал текст в 3,4 раза более длинный - 880 тыс., однако, это объем собрания его сочинений.

На том же самом объеме, что и объем "Улисса", составлен английский частотный словарь (Howes, I966), но в нем почти в три раза меньше слов - 9 7000. Различие состоит в том, что упомянутый словарь описывает словоупотребление сразу многих лиц*.

Из приведенных примеров видно, что в целостном тексте разнообразие слов существенно больше, чем в корпусе текстов того же объема. Однако провести границу между целостным и нецелостным текстом, используя в качестве критерия только закон Ципфа, оказалось очень сложным (см. Арапов, I98I а).

Во-первых, оказалось, что уточнения требует сам закон Ципфа (это было сделано нами в ряде публикаций, см. Арапов, Ефимова, I975, Арапов, I977), причем уточнений закона может быть несколько.

Во-вторых, можно предложить не одну методику установления соответствия между непрерывными зависимостями (типа (1)-(3)) и дискретной диаграммой, полученной в результате классификации единиц конкретного текста, несколько способов выделения случайной составляющей, "маскирующей" проявление основной тенденции (см. Арапов, I977) и т.д.

В-третьих, достаточно неопределенно само понятие целостного текста. Хотя это понятие, несомненно, часть нашей культуры (целостность литературного произведения является даже объектом правовой защиты, см. Потапова, I98I), оно далеко не однозначно: немыслимо, конечно, изъять строфу из "Евгения Онегина" или дописать страницу к "Преступлению и наказанию", но уже можно опубликовать стихи или рассказ вне цикла, к которому они отнесены автором. Это два противоположные конца шкалы, и граница между целостным и нецелостным текстом лежит где-то посредине.

Оставаясь в рамках чисто феноменологического подхода, практически невозможно собрать доказательства,нужные для того, чтобы окончательно принять или отвергнуть гипотезу о связи целостности и "ципфовости" текстов. Недостаточно сравнивать реальные тексты с "ципфовским" эталоном, нужно понять природу этого закона и хотя бы примерно представить себе, в чем более глубокое различие между "ципфовскими" текстами и текстами, которые не удовлетворяют этому закону.

---

* дальнейшие примеры см. Арапов, I976, I978, I98I.

7. Вернемся снова к форме классификации - X- и У-классам с численностями n_i и m_j. Поставим вопрос: как приписать вероятность самой форме классификации, т.е. способу разложить коллекцию М по ячейкам? Действительно, в зависимости от числа классов и их объемов (при одном и том же размере коллекции L) число способов осуществить данную классификацию может быть различным. Вероятностью классификации естественно считать величину пропорциональную числу способов ее осуществления (с коэффициентом пропорциональности, выбранным так, чтобы сумма вероятностей всех способов классификации составила единицу).

Тонкость состоит в том, что подсчитать число способов осуществитель классификации может, только указав, какие из этих способов мы считаем тождественными, а какие собираемся различать (в статистической физике, где данный подход является обычным, соглашения о различении и отождествлении способов классификации называются "статистиками").

В качестве примера рассмотрим два способа ввести классификацию (две "статистики") на множестве М из трех элементов $M = \{1,2,3\}$. X-классы - предполагается, что среди них не будет пустых, - мы будем обозначать римскими цифрами. Саму классификацию удобно представить в виде таблички, где элементу $x \in M$ (верхняя строка) будет сопоставлен класс, в который он попадает. Задача подсчета числа возможных классификаций сводится при этом к подсчету числа различных табличек.

Статистика I. Очевидно, что если класс один, можно построить только одну табличку:

$$\begin{pmatrix} 1 & 2 & 3 \\ I & I & I \end{pmatrix}$$

Если классов два, то они могут состоять из одного или двух элементов, способы классификации отличаются тем, какой элемент М попал в "одиночку":

$$\begin{pmatrix} 1 & 2 & 3 \\ I & II & II \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 \\ II & I & II \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 \\ II & II & I \end{pmatrix}$$

Таким образом, при двух классах возможно три варианта классификации. Но если классов будет три, то число этих вариантов возрастет до шести:

$$\begin{pmatrix}1&2&3\\I&II&III\end{pmatrix}\begin{pmatrix}1&2&3\\I&III&II\end{pmatrix}\begin{pmatrix}1&2&3\\II&I&III\end{pmatrix}\begin{pmatrix}1&2&3\\II&III&I\end{pmatrix}\begin{pmatrix}1&2&3\\III&I&II\end{pmatrix}\begin{pmatrix}1&2&3\\III&II&I\end{pmatrix}$$

Т.е. если исходить из приведенных выше соображений, то наиболее вероятной классификацией М будет разбиение на три единичные класса.

Статистика II. В рассмотренном выше способе классификации (статистика I) большую роль играло то, элемент с каким номером попал в данный класс. Мы считали различными такие классификации, в которых объемы классов совпадали, но различались сами элементы. Можно было бы принять и другое соглашение: считать несущественным, с помощью каких именно элементов данный класс набирает определенную численность.

И здесь при одном классе будет только один вариант классификации. В случае трех классов разбиение будет только одно: шесть приведенных выше табличек различаются признаком, который мы объявили несущественным. А вот разбиений на два класса будет два. Соответствующие таблички удобнее рисовать, обозначая элементы не цифрами, а палочками (номер класса, в отличие от статистики I, где классы различаются только размерами, считается существенным признаком):

$$\begin{pmatrix}|&|&|\\I&I&II\end{pmatrix}\qquad\begin{pmatrix}|&|&|\\I&II&II\end{pmatrix}$$

Таким образом, если мы выберем второй способ классификации, то наиболее вероятным будет разбиение на два класса (вариант статистики I в физике называется статистикой Больцмана, а статистики II – статистикой Бозе-Эйнштейна).

Кроме способа классификации, форму классификации определяют дополнительные "лимитирующие" условия. Таким условием является, например, фиксированное число классифицируемых объектов $\sum$. Но для того, чтобы получить нетривиальные формы классификации, исходя, например, из статистики I, нужны более сильные лимитирующие условия. В физике эти условия появляются из законов сохранения. Такой подход использовал в частности Ю.А.Шрейдер (1967) при выводе закона Ципфа. Он ввел фактически понятие "энергии" текста и потребовал ее сохранения. Но оправдать введение энергии, имея в виду весь диапазон проявления закона, едва ли возможно. Поэтому в 1978 г. в совместной работе (Арапов-Шрейдер, 1978) мы пошли другим путем:

использовали в качестве лимитирующего условие максимальной вероятности еще одной, дополнительной классификации на том же множестве.

Дополнительная классификация – это уже известная нам классификация на У-классы. Если выбрать статистику I, то наиболее вероятной классификацией по Х-классам, как видно из разобранного примера, будет классификация М на L единичных классов (ни одно слово в тексте не повторяется), при тех же условиях наиболее вероятная классификация по У-классам предполагает, что в тексте будет только одно слово с частотой L. Но если потребовать их совместного максимума (или, эквивалентно, максимума суммы энтроний Х- и У-разбиений), то удовлетворяющая этому требованию классификационная диаграмма будет нетривиальна. Обозначим координаты вершин a и b соответствующего многоугольника n_a, m_a, n_b, m_b (см. рис. I), тогда классификационная диаграмма будет задаваться неравенством

$$|m_a n_a - m_b n_b| \leq \min(n_a, m_a, n_b, m_b) \quad (4)$$

Выражение (4) является дискретным аналогом закона Ципфа (I) – (2) и содержит всю информацию о классификационной диаграмме (см. Арапов, Шрейдер, I978).

Естественно, что убедительность описанного вывода закона Ципфа зависит от того, насколько убедительно обоснованы постулаты, лежащие в его основе. Одним из способов обосновать эти постулаты (две зависимые друг от друга классификации, принцип экстремума, статистика I и т.д.) – это показать, что те же принципы **mutatis mutandis** приводят к другим наблюдающимся в природе и языке законам распределения частот. В какой-то степени эта задача была решена в работах Арапов-Крылов, I980, Арапов, I98I.

8. Предложенный подход к описанию количественной структуры текста дает ответ на один принципиальный вопрос: почему слова в пределах текста так резко различаются по частоте употребления. Ответ состоит в том, что в указанном выше смысле ципфовское распределение является оптимальным. Однако в этом выводе текст рассматривается in toto и ничего не говорится о структуре его частей. Чтобы получить информацию о структуре частей текста, мы предлагаем взглянуть на него не как на нечто готовое, удовлетворяющее экстремальному принципу, а как на процесс. Мы попытаемся формализовать тот са-

мый процесс построения диаграммы, с которого мы начали в п.2.

Состояния этого процесса будет удобно представлять в виде диаграмм типа (а) - (б) и (в) - (е), приведенные там диаграммы соответствуют возможным состояниям процесса на втором и третьем шаге. Пусть на k-ом шагу построена диаграмма, изображенная на рис. 2. Покажем, как к этой диаграмме можно добавить еще один элемент.

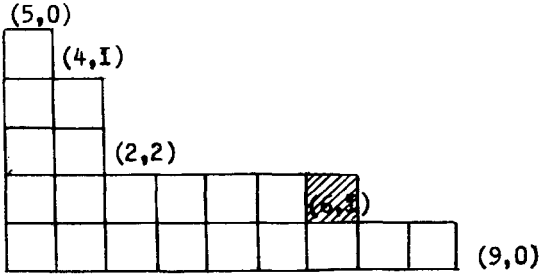


Рис.2

Поскольку диаграмма на каждом шагу должна оставаться монотонной, $k+1$-ый элемент должен попасть в "паз", где у него будет сосед слева (с координатой m) и снизу (с координатой n). На приведенной диаграмме выделен элемент, который попал в паз с координатами $n=1$ и $m=6$ (незаполненными остались пазы с координатами (5,0), (4,I), (2,2), (9,0).

Чтобы получить не просто монотонную диаграмму, а диаграмму определенной конфигурации, нужно соответствующим образом определить вероятности заполнения различных пазов, т.е. задать вероятности перехода от данного состояния к одному из возможных на следующем шаге. Тем самым будет определен вероятностный процесс.

Определим сначала "вес" w паза (n,m), положим его равным произведению $(n+1)(m+1)$. Вероятность его заполнения пусть будет обратно пропорциональна его "весу", т.е. определяется условиями $p=\frac{c}{w}$ и $\sum p=1$, где сумма берется по всем "пазам". С помощью довольно простых рассуждений можно показать, что на любом шаге наиболее вероятной будет диаграмма,

удовлетворяющая соотношению (4), т.е. закону Ципфа. Но кроме наиболее вероятной реализации, процесс будет иметь и другие реализации, далекие от ципфовской. Им также можно будет приписать определенные вероятности.

Из данной модели текстообразования можно получить ряд следствий, относящихся к внутренней структуре целостного текста: распределению по длине текста k-тых употреблений слов, слов, использованных в тексте k раз, интервалов между словами с заданным свойством и т.д. Одно из очень простых следствий состоит в том, что разнообразие лексики в н у т р и целостного текста растет от начала текста к концу по тому же закону, по которому возрастает лексическое богатство целостных текстов, упорядоченных по их длине (подразумевается, что все параметры закона Ципфа для этих текстов совпадают). Или, другими словами, "правильной" частью целостного текста, подчиняющейся тому же закону, что и весь текст, будет его любой непрерывный отрезок, содержащий н а ч а л о т е к с т а. Часть целостного текста, которая не содержит начала, может и не удовлетворять соотношению (4). Отсюда возникает гипотеза, что текст, в котором мы интуитивно выделяем несколько "начал" (текст в тексте), может не быть целостным.

Сопоставление модели с опытными данными наверное обнаружит ее приближенный характер: модель, например, не учитывает сильные ограничения на повторяемость слов в пределах фразы или сверхфразовых единств.

9. До сих пор мы пытались объяснить, как ведут себя частоты слов на одном уровне организации речи - на уровне целостного текста. Но не менее важно объяснить, почему одинаковые слова в р а з н ы х текстах (или в разных частях одного текста с несколькими "началами") употребляются согласованно (см. Арапов, Ефимова, Шрейдер, 1975 а; Арапов, Тер-Гаспарян, Херц, 1978). В рамках "классической" модели (п. 5) объяснение состоит в ссылке на устойчивые вероятности употребления этих слов, но это предположение является слишком "жестким". Реальный механизм должен быть более гибким.

Говоря о структуре текстов, мы ограничивались выводом соотношений для объемов X- и У-классов, полностью игнорируя вопрос, какие слова попадают в эти классы. Рассмотрим теперь собрание всех слов данного языка - тезаурус $\mathcal{T} = \{u_s\}$ - и укажем для каждого слова u_s из этого словаря вероятность попа-

дания в класс X_r – $P(u_s, X_r)$ Эта вероятность – мера той части всех текстов $T_{s,r} \subset T$ в которой слово u_s занимает место X_r.

Для того, чтобы задать $P(u_s, X_r)$ аналитически, перейдем от слов и классов к их измеримым свойствам. Для класса это свойство – место в классификации, т.е. ранг r , а для слова – особая величина (сложность), смысл которой мы сейчас поясним.

Слова любого естественного языка можно упорядочить по степени их "необходимости" для построения текста. У тех слов, без которых текст построить нельзя, ранг будет мал (а частота, соответственно, велика). К этим словам относятся служебные элементы, некоторые числительные, "полувспомогательные" глаголы, в меньшей мере – существительные с большим метафорическим "потенциалом" (термины родства, названия частей тела и проч.). Важно, что нет слов, которые употреблялись бы в любом тексте редко: всегда найдется текст, в котором данное слово будет употребляться сколь угодно часто. Словам с устойчивой и (в среднем) большой частотой противостоят слова с неопределенной частотой, которая сильно колеблется от текста к тексту.

Если, кроме двух упомянутых групп, рассматривать и все промежуточные случаи, можно будет говорить о непрерывно изменяющейся характеристике s – "размытости" слова по шкале рангов, эта характеристика и называется сложностью. На основании приведенных выше соображений и некоторых других формального характера, можно выбрать функцию от двух переменных $P(r, s)$ которая и определяет, как согласовано употребление слов в различных текстах.

Не останавливаясь на формализме (Арапов, и др., 1975 б, Арапов и др., 1978), с помощью которого эксплицировано понятие о р г а н и з м е н н о с т и, перечислим некоторые задачи, решаемые с помощью этого формализма.

1) Пусть у нас есть два словаря, составленные на основании двух различных текстов из множества T. Мы можем эти словари сравнить и для любой части этих словарей теоретически найти число общих слов, причем результат зависит только от одного числа, определяющего "расстояние" между сопоставляемыми текстами (словарями).

2) Если на подмножество текстов $\mathbf{T} \subset T$ наложены определенные ограничения, относящиеся к их замкнутости, объему и рас-

ε

стоянию между ними, то можно построить вероятностное пространство, элементарными событиями в котором будут употребления слов в корпусе, составленном из T'. Закон, связывающий эти вероятности для частых слов, будет близок к закону Ципфа, но для редких существенно от него отличаться.

3) При тех же ограничениях можно найти распределение слов по числу текстов из $T' \subset T$, в которых это слово употреблено. Между задачами 3) и 4) есть простая связь.

4) Характеристика сложности слова S связывается значимой при решении проблем, никак на первый взгляд не связанных с задачей описания согласованного употребления слов в текстах; она определяет темпы исторического изменения соответствующих частей словаря, определенной зависимостью связаны сложности слов, находящиеся в данных лексико-грамматических отношениях (Thorndike, 1943, Арапов, 1980) и т.д.

10. Таким образом, предложенная модель включает два механизма. Один из них - весьма универсальный - это механизм дифференциации частот в отдельных "клетках организма" - текстах. Другой - вероятно, специфический для языка - механизм согласования частот в отдельных текстах, который обеспечивает взаимопонимание членов одного языкового коллектива.

Л И Т Е Р А Т У Р А

Арапов М.В. Две модели рангового распределения. - Вопросы информационной теории и практики, вып. 4(31). - М.: ВИНИТИ, 1977, с. 3-42.

Арапов М.В. Измерение лексического богатства текстов.- В кн.: Język.Poetyka.Tekst.Warszawa:PWN,1978,s.187-210.

Арапов М.В. Регулярность семантических отношений. - Научно-техническая информация. Сер. 2, 1980, № 9, с. 1-8.

Арапов М.В. Классификация и распределения в лингвистике. - Семиотика и информатика, вып. 17. - М.: ВИНИТИ, 1981, с. 120-147.

Арапов М.В. Системный анализ лексической структуры текстов. - В кн.: Системные исследования. Ежегодник 1980. - М.: Наука, 1981 а, с. 372-403.

Арапов М.В., Ефимова Е.Н. Понятие лексической структуры текста. - Научно-техническая информация. Сер. 2, 1975, № 6, с. 3-7.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. - Научно-техническая информация. Сер.2, 1975, № 1, с. 9-20.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. Ранговые распределения в языке и тексте. - Научно-техническая информация. Сер. 2, 1975 а, № 2, с. 3-7.

Арапов М.В., Крылов Ю.К. Mathematical Models of Classification in Application to Some Problems of Statistical Linguistices - In:Computational Linguistics and Related Topics.Symposium.Tallinn,1980,pp.14-16.

Арапов М.В., Либкинд А.Н. О понятии замкнутого информационного потока. - Научно-техническая информация. Сер. 2, 1977, № 6, с. 1-15.

Арапов М.В., Либкинд А.Н. Научные документы в зеркале информатики. - Вопросы информационной теории и практики,вып. 46. - М.: ВИНИТИ, с. 47-81.

Арапов М.В., Тер-Гаспарян Л.И., Херц М.М. Сравнение частотных словарей. - Научно-техническая информация. Сер.2, 1978, № 4, с. 20-29.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссиметрии системы. - Семиотика и информатика, вып. 10. - М.: ВИНИТИ, 1978, с. 74-95.

Звонкин А.К., Фрумкина Р.М. Свободная классификация: модели поведения. - Научно-техническая информация. Сер. 2, 1980, № 6, с. 1-6.

Козачков Л.С. Системы потоков научной информации. - Киев: Наукова думка, 1973. - 196 с.

Лесохин М.М., Лукьяненков К.Ф., Пиотровский Р.Г. Введение в математическую лингвистику. - Минск: Наука и техника, 1982. - 263 с.

Орлов Ю.К. О статистической структуре сообщений, оптимальных для человеческого восприятия. - Научно-техническая информация. Сер. 2, 1970, № 8, с. 11-16.

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

Потапова И.Е. Правовое регулирование перевода произведений

науки, литературы и искусства по советскому законодательству. Автореф. дисс. канд. юрид. наук. М., 1981.

Фрумкина Р.М. Статистические методы изучения лексики. – М.: Наука, 1964.

Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текста. – В кн.: Проблемы передачи информации, вып. I. – М.: Наука, 1967, с. 57-63.

Bradford S.C. Sources of Information on Specific Subjects.– Engineering,1934,137,n 3350,p.85-86.

Condon E. Statistics of Vocabulary.-Science,1928,67,N 1733 .

Estoup J. Gamme stenographique.Paris,1916.

Haight T.A. Some Statistical Problems in Connection with Word Association Data.-Journal of Mathematical Psychology,1963,3,p.217-33.

Howes D.A. Word Count of Spoken English.- Journal of Verbal Learning and Verbal behavior,1966,5,p.572-604.

Jonson G.A. Rank-size Convexity and System Integration: View from Archeology.-Economic Geography (USA),1980,56,n 3, p 234-247.

Lotka A. The Frequency Distribution of Scientific Productivity.-Journal of the Washington Academy of Science,1926, 16,n 12,p.317-323.

Pareto V. Cours d'économie politique,v.2.Lousanne,1897.

Thorndike E. Derivation ratios.-Language,1943,19,p.27-37.

Willis J.C. Age and Area.Cambridge:Cambridge Univ.Press,1922.

Zipf G.K. Selected Studies of the Principle of Relative Frequencies in Language.Cambridge (Mass.),1932.

TEXT AND LANGUAGE - WHOLENESS AND ORGANISMLIKENESS

Mikhail Arapov

S u m m a r y

A précis of author's works in which an unorthodox approach to quantitative organization of language on its lexical level has been outlined. According to the present theory a certain probability should be attributed not to the occurence of an individual word (as in the standard theory) but to the lexical pattern of the text as a whole. A holistic concept of an individual "closed" text is proposed; its formal properties including the distribution of the word frequencies are derived from a number of simple postulates. It is demonstrated that the Zipf distribution of the frequencies is the most probable under those postulates. Some examples and semantic features of such Zipfian texts are discussed.

Language is considered as an organism that consists of many cells - "closed" texts whose lexical patterns are coordinated. A certain probabilistic mechanism of the coordination is briefly outlined.

ЧАСТОТА СЛОВА И ЕГО ЗНАЧИМОСТЬ В СИСТЕМЕ ЯЗЫКА

(некоторые уроки поисков "внутреннего оправдания" частотных словарей в лингвистике и лингводидактике)

С.И. Гиндин

Памяти Игоря Васильевича Рахманова - благодарно помнящий единственную встречу с ним

автор

0. О задачах и жанре этой статьи

В 20-м столетии лингвистика под влиянием как запросов практики, так и внутреннего стремления к осознанию собственных методов уделяла большое внимание связи между парадигматикой (свойствами языка) и синтагматикой (свойствами речевой цепи) и операциям, с помощью которых можно перейти от анализа второй к выводам о первой. Среди синтагматических свойств особое и очень важное место занимают статистические характеристики языковых единиц и их последовательностей, в частности - статистические характеристики слов, описываемые с помощью так называемых частотных словарей.

Интерес и надежды, которые частотные словари вызывали и вызывают у представителей самых разных областей теоретической и прикладной филологии, объясняются не только тем, что основные характеристики, лежащие в основе построения таких словарей, - число употреблений данной лексической единицы в некотором корпусе текстов (частота) или число текстов, в которых она употреблена (ранг, или употребительность), - носят количественный характер, а процедура их подсчета весьма проста и допускает автоматизацию. Уже первые составители таких словарей заметили, что небольшое количество наиболее частых слов покрывает львиную долю общего числа словоупотреблений в любом тексте. Постепенно наивное изумление стало уступать место поискам более глубокой и уже собственно языковой мотивированности структуры и свойств частотного словаря. В частности, у исследователей естественно возникла идея о тесной

связи между частотой слова в речи и его парадигматической значимостью*, важностью, существенностью места, занимаемого этим словом в лексико-семантической системе языка: чем чаще употребляется слово, тем более важное место в системе языка оно занимает. Коннектор "чем... тем" в подобных формулировках означает феноменологически фиксируемую связь, направление же причинной зависимости естественно полагать противоположным: системная важность слова обусловливает его повышенную употребительность.

Гипотезы, тем или иным образом уточняющие указанную идею, и эксперименты по их проверке не раз описывались в литературе. Данная статья не имеет целью изложение еще одного варианта уточнения или проверки, или - тем более - какое-либо окончательное решение проблемы. Ее задача - рассмотреть в единой перспективе известные автору** подходы к решению и попытаться осмыслить те эвристические уроки, которые можно извлечь из анализа внутренней логики каждого из них и из их сравнения. Думается, что эти уроки небесполезны для правильной постановки и планирования тех будущих исследований, которые могли бы претендовать на более или менее кардинальное освещение проблемы взаимосвязи парадигматических и статистических характеристик слова. Более специальная задача второго раздела статьи состоит в том, чтобы побудить лингвистов-теоретиков (и, в частности, "лингвостатистиков") с большим вниманием отнестись к опыту, накопленному в такой прикладной

---

* Родственные представления о связи частоты и значимости в применении уже не к языку в целом, а к корпусу произведений одного автора и к отдельному тексту возникли, соответственно, в поэтике и информатике. Развитие и структура основанных на них методов "автоматического реферирования" разобраны в Гиндин С.И. 1977. Там же в § I.4 обсуждается соотношение подобных представлений с вытекающим из статистической теории связи К. Шеннона выводом об информативности редких слов, а на с. 69 - само понятие "значимости", его отношение (в случае отдельного текста) к планам выражения и содержания и к одноименному, но не совпадающему с ним соссюровскому понятию.

** Работы, вышедшие до 1974 г. когда готовился первый вариант статьи, выявлялись в ходе специального разыскания по всем доступным автору указателям публикаций по лингвостатистике и информатике. Позднейшие работы привлекались лишь в меру "текущей" библиографической информированности автора. Сознательно были оставлены в стороне работы по частотным словарям и "поэтическим мирам" отдельных писателей, требующие ввиду своей многочисленности и богатства содержащегося в них материала отдельного рассмотрения.

дисциплине, как лингводидактика (методика обучения языкам), и создающему более добротный и надежный в некоторых отношениях фундамент для всестороннего анализа заглавной проблемы, чем иные из немедленных прорывов в квантификацию у лингвистов-теоретиков.

I. Исследования по теоретической лингвостатистике

I.I. Гипотеза Ципфа и первые опыты ее уточнения и проверки. Самое раннее из известных мне обоснований гипотезы о взаимосвязи частоты и парадигматических характеристик слова было дано американским психологом и лингвистом Джорджем Кингсли Ципфом посредством аналогии между употреблением различных слов в речи и использованием различных орудий в работе ремесленника: "... ремесленник может обнаружить, что определенное размещение (или определенный способ (pattern) расположения) совместно используемых различных близлежащих орудий позволяет легче выполнить некоторую специальную задачу, чем находящееся в большем отделении специализированное орудие... Назовем это потребностью (urge) в экономичном размещении (permutation) более удобных (easier) орудий; а эта потребность, ввиду своего постоянного характера, приводит в ходе временно́го развития к весьма любопытным результатам. Именно, она приводит к тому, что чаще используемые орудия оказываются и наиболее разнообразными по своим функциям, тогда как реже используемые будут, как правило, более специализированными. Короче говоря, между частотой использования орудий и разнообразием их применения возникает прямая зависимость. В терминах слов и их значений – считая слово переводным эквивалентом орудия, а каждое значение – эквивалентом некоторого конкретного способа применения этого орудия – ожидаемая закономерность может быть сформулирована так: существует прямая зависимость между числом различных значений слова и относительной частотой его употребления" (Zipf G.K. 1945 a, p. 144).

Если способ обоснования, примененный здесь (да и самое желание отыскать глубинные психо-биологические механизмы, стоящие за феноменологически выявленной закономерностью) и остался в науке идиосинкратически связанным с именем самого Ципфа, то идея о связи частоты и числа значений* получила

* Эту идею, равно как и предложенную им позже формулу для ее количественного выражения, не следует смешивать с гораздо более популярным в лингвостатистике "законом Ципфа",

довольно широкое распространение. Рассуждение Ципфа предопределило вместе с тем и менее бросающееся в глаза, но очень существенное свойство всех последующих лингвистических трактовок проблемы - связывать частоту прежде всего с такими парадигматическими свойствами слова, которые допускают автономную целочисленную квантификацию, т.е., попросту говоря, непосредственный подсчет*. Скажем, Ципф рассматривает не относительную важность значения (или значений) слова для характеристики внеязыковой действительности, не весомость значений в системе, а именно число значений.

Следующий шаг в исследовании проблемы был вскоре сделан тем же Ципфом (Zipf G.K. 1945 b). Изложив результаты экспериментальной проверки сформулированной в предыдущей статье гипотезы (сведения о частоте брались из частотного словаря Торндайка (Thorndike E.L. 1931), а сведения о числе значений - по составленному тем же автором на базе предварительного "частотного словаря значений" толкового словаря (Thorndike E.L. 1941), Ципф попытался придать зависимости точную количественную формулировку. Опирался он при этом на предположение (опять аппеляция к психобиологическим механизмам!), что и говорящий, и слушающий стремятся к минимизации своих речевых усилий. По мнению Ципфа, стремления говорящего и слушающего противоречат друг другу: для говорящего удобней всего использовать для выражения всех значений одно и то же слово, а для слушающего, наоборот, наиболее экономным будет вариант, при котором каждое значение выражается с помощью особого слова. Противоборство этих двух тенденций к экономии должно привести к установлению некоторого промежуточного оптимума числа значений у каждого слова. Этот оптимум и есть искомая зависимость числа значений от частоты: "Число различных значений слова будет стремиться к равенству с квадратным корнем из его относительной частоты. Исключение из

описывающим связь частоты слова с номером последнего в частотном словаре. Попытка использовать для статистического описания полисемии обе идеи Ципфа предпринята в: Тулдава Ю.А. 1979; при этом структурно аналогичной "Закону Ципфа" оказалась (см. с. 121-125) формула, описывающая распределение слов по числу значений не в тексте, а в словаре языка.

* В терминах теории измерений это есть ограничение не просто "количественными" свойствами слова, но теми из количественных, которые допускают "первичное" измерение и притом по простейшей из шкал - по "абсолютной" (см.: Суппес П., Зиннес Дж. 1967, с. 25-27, 20).

этого правила, возможно, составят несколько дюжин самых частых слов" (Zipf 1945 b, p. 255)*. Как видим, Ципф одновременно с закономерностью прозорливо указал и на вероятность важных исключений из нее - число значений не может расти столь же неограниченно, как и частота.

Аналогичные идеи развивались и в Zipf G.K. 1949. Вскоре после выхода этой книги Ципф умер (в 1950 г.), но направление дальнейших попыток уточнения и проверки гипотезы о связи частоты и парадигматических свойств слова долго еще определялось той формулировкой, которую он придал данной гипотезе**. Американский психолог Нобл, дав интерпретацию понятия значения в терминах экспериментально наблюдаемых стимулов и реакций, вывел экспериментальным путем уравнение зависимости между объективной оценкой употребительности слова - его "знакомостью" (familiarity) и число его значений (Noble C.E. 1953).

В предисловии к книге Guiraud P., 1954 ципфовские закономерности уже не столько исследовались, сколько иллюстрировались на конкретном примере как своего рода "патент на благородство" лингвостатистики. О каких-либо исключениях, наличие которых предполагал сам Ципф, у Гиро уже нет и речи. Правда, затем П. Гиро попытался вывести из ципфовской формулы и его же "закона", описывающего связь частоты слова с номером слова в частотном словаре и с числом слов, имеющих данную частоту, некоторую производную закономерность. Он утверждал, что число слов n, имеющих S значений, обратно пропорционально s^2: ns^2 = const. Но эта новая формула приводимыми им числовыми данными по 545 словам подтверждалась не в достаточной степени (Guiraud P. 1954, p. 1-3). Не смущаясь этим, П. Гиро, рассмотрев далее связь между частотой слова и числом фонем в нем, сделал далеко идущее обобщение: "Частота слова связана с множеством его звуковых, морфологических, семантических, этимологических свойств; она является их отражением и образом, и именно через ее посредство мы можем попытаться представить и проанализировать эти свойства" (там же, p. 4). Так ципфовское "число значений" уступило место

* Отнесение Э.И. Королевым данной формулы Ципфа к первой из названных его статей и к "50-м годам", равно как и его утверждение, что И.В. Рахманов разрабатывал свою методику специально "для английского языка" (Андрукович П.Ф., Королев Э.И. 1977, с. 1) представляются явным недоразумением.

** Судя по аннотации в Bailey R., Doležel L. 1968, в этом же направлении ведется анализ и в статье Baker S.J. 1950, оставшейся для меня недоступной.

некоей неопределенной совокупности свойств слова, относящихся к различным уровням языковой системы.

С точки зрения анализа общей проблемы, вынесенной в заголовок нашей статьи, переход к рассмотрению множества системных характеристик слова можно было только приветствовать. Но ведь и исходная гипотеза Ципфа, связывавшая частоту всего лишь с одной переменной, все еще оставалась, подобно прочим его лингвостатистическим идеям, скорее блестящей догадкой, вызовом будущим исследователям, чем окончательно установленным фактом*. Тем больше трудностей должно было возникнуть при обращении к целому набору переменных.

I.2. Переход к рассмотрению нескольких парадигматических свойств слова. В исследовании В.А. Московича (1969, гл. 1, раздел 3), выполненном на материале семантического поля цветообозначений в английском и украинском языках, исходная гипотеза была сформулирована так: "чем частотней слово, тем оно активнее - т.е. тем многочисленнее и сильнее по связи (sic! - С.Г.) в системе слов" (там же, с. 23). Самый термин "активность" заимствован исследователем (см. с. 46) из обсуждаемой ниже в § 2 работы И.В. Рахманова (1960), оттуда же взяты и наборы свойств, которые наряду с изучавшейся Ципфом "полисемией" рассматриваются как "аспекты активности": способность слова к словопроизводству, способность к словосложению и способность входить в состав фразеологизмов. При квантификации каждый из этих параметров слова отождествляется с числом соответствующих его коррелятов, будь то слова, производные от данного, сложные слова или фразеологизмы, в состав которых оно входит. Раз важно лишь число связей, то между собой все связи фактически уравниваются. Поэтому из двух сторон активности, отраженных в процитированной формулировке исходной гипотезы исследования, на деле измеряется и исследуется только одна - "число" связей, а не их "сила".

Выбор в качестве составляющих активности слова указанных четырех характеристик представляется мотивированным как с точки зрения организации эксперимента - именно эти 4 характеристики легко вычислить по обычным толковым словарям, - так и с точки зрения строения современной науки о слове. ох-

* Ср. вряд ли справедливое, но весьма показательное отрицание самой возможности опоры на работы Ципфа, принадлежащие, якобы, вчерашнему дню лингвостатистики, в Herdan G., 1966.

ватывающей лексикологию, словообразование и фразеологию. Конечно, если в распоряжении лингвиста будет словарь, совмещающий в себе, напр., обычные синонимический и толковый словарь, то к 4 параметрам легко будет присоединить и пятый – число синонимов. Более принципиальный вопрос заключается, однако, не в расширении числа характеристик, а в трактовке внутренних соотношений между уже введенными характеристиками. Есть ли активность (значимость) слова в системе языка простая совокупность ("вектор") выделенных характеристик, измеренных в независимых и, возможно, несопоставимых друг с другом единицах, или она может быть рассмотрена как результат некоторого взаимодействия этих характеристик и оценена с помощью каких-то дополнительных числовых операций над их значениями? Как увидим ниже, И.В. Рахманов и его соавторы, строя таблицу, в столбцах которой для каждого из слов записывались оценки по каждому из семи (считая и частоту) выделенных свойств, не вводили в рассмотрения какой-либо единой универсальной функции от этих семи свойств. Для каждого очередного слова соответствующая ему строка таблицы подвергалась индивидуальному неформальному обсуждению. В отличие от этих исследователей, В.А. Москович, выбирая подмножество из той же совокупности свойств, считает возможным рассматривать активность слова как некоторую пятую характеристику, вычисляемую на основе четырех исходных характеристик и так же, как они, выражаемую не вектором, а числом.

Уже это допущение является достаточно сильным. Но исследователь идет еще дальше, еще до эксперимента предопределяя и характер взаимодействия компонентов активности. Значения всех четырех исходных характеристик сильно разнятся между собой. По данным всех привлеченных словарей количество зафиксированных производных слов всегда регулярно превышает число зафиксированных в них же разных значений того же слова примерно в 10 раз. В связи с этим В.А. Москович производит специальное преобразование полученных величин. Именно весь промежуток изменения каждой из четырех величин, от ее максимального значения до минимального, делится на 10 разных частей и каждому слову сопоставляется уже не само найденное по словарю значение этой величины, а номер той из частей, в которую попадает это значение. После такого преобразования все четыре составляющих активности становятся не просто соизмеримыми друг с другом (что по-прежнему, хотя сам исследователь и не осознает этого, остается лишь одним из исходных

допущений), но и сопоставимыми по своим значениям. Теперь их среднюю можно использовать без боязни, что распределение этой средней будет в основном определяться распределением лишь одной или двух из составляющих. Из возможных средних В.А. Москович выбирает среднюю арифметическую. Тем самым введенный им коэффициент активности слова фактически представляет активность как некоторую равнодействующую четырех составляющих. Ни причины, ни последствия такого выбора исследователем специально не обсуждаются, очевидно, средняя арифметическая выбрана в силу своей простоты и привычности. Между тем сама применимость средней арифметической нуждалась в данном случае в доказательстве. Легко заметить, что после описанного преобразования каждая из 4-х исходных характеристик слова измеряется уже не по абсолютной шкале, а по так называемой шкале порядка (Суппес П., Зиннес Дж. 1967, с. 21). Известно (см., напр., Фрумкина Р.М. 1971, с. 30), что на такой шкале средняя арифметическая не применима. Естественно, что применимость аналогичной операции сразу к нескольким шкалам данного типа тем более может вызывать сомнения.

Проверку основной гипотезы о связи между частотностью и активностью можно вести двумя путями. Можно разбить слова на группы по их частоте и подсчитать среднюю активность слов в каждой из этих групп. Так поступал (применительно к параметру "число значений") в цитированной книге П. Гиро, использует этот путь и В.А. Москович. Получив для каждого слова индекс его активности, он вычислил средний коэффициент активности для трех групп прилагательных, выделенных по частотному принципу. Для первой группы, в которую входят наиболее частые прилагательные, имеющие частоту от 151 до 36 на выборку в 200 тыс. слов (это четыре цвета спектра green "зеленый", blue "синий, голубой", red "красный", yellow "желтый", три важнейших смешанных цвета white, black, grey "серый", а также brown) - средний индекс активности оказался равным 0,77. Во второй группе, называемой "группой средней частотности" (Москович В.А., с. 21) и которую, на наш взгляд, было бы вернее назвать группой редких (малочастотных) слов (от 12 до 3 раз на 200 тыс. слов) средний индекс активности уже только 0,3. К этой группе относятся два оставшихся цвета спектра orange и violet "фиолетовый, лиловый" и такие цвета, как rose "розовый", purple "пурпурный", pink "ярко-розовый", olive "оливковый", khaki "хаки", scarlet "алый".

Наконец, в третьей группе, слова которой названы исследователем "редкими" и которые еще вернее было бы назвать "редчайшими" - каждое из них встретилось 1 или 2 раза на 200 тыс. слов - средний коэффициент активности и вовсе равен 0. Это названия всевозможных оттенков, характерные по большей части для речи портных и женщин - lemon "лимонный", sandy "песочный", coral "коралловый", chocolate "шоколадный", creamy "кремовый" и т.д.

Другой путь проверки основной гипотезы состоит в сравнении распределений частоты и активности. Такое сравнение проводится путем вычисления коэффициента ранговой корреляции списков исследуемых слов, в одном из которых эти слова расположены по убыванию частоты, в другом - по убыванию активности. Соответствующее сравнение для английских цветообозначений было проведено 6 раз (использовались два различных источника для определения активности и три источника сведений о частоте). Коэффициент корреляции во всех случаях оказался выше 0,9, а в четырех - больше, чем 0,96. В двух экспериментах он равнялся +0,99 - значение, крайне редко встречающееся в практике лингвистических исследований и свидетельствующее о "почти-функциональной" зависимости между частотностью и активностью. Для украинского материала коэффициент составил +0,98 (там же, с. 45). Сам исследователь, по-видимому, склонен рассматривать перечисленные результаты как вполне доказательные и даже использует их для корректировки "ошибок" в традиционной "семантической классификации" поля цветообозначений (см. там же, с. 50-51). Подобный энтузиазм представляется, однако, преждевременным. Даже если отвлечься от указанных выше сомнительных моментов в методике подсчета самого коэффициента активности, и характер изложения, и сам план эксперимента оставляют без выяснения еще целый ряд существенных вопросов. Во-первых, число цветообозначений, фигурирующих в многочисленных таблицах данного раздела книги, сильно варьирует, и нигде не сказано, какой именно совокупности слов отвечает тот или иной из полученных коэффициентов. Даже для единственного из 7 экспериментов, все шаги которого документированы отдельными таблицами (активность по словарю Фанка и Вагнэллса, частоты - по выборке), остается неясным, получен ли коэффициент корреляции 0,99 для 13 слов из таблицы 21, 12 слов из таблицы 25 или какой-либо большей совокупности. Поэтому, несмотря на обилие числовой информации, результаты В.А. Московича оказываются фактически не

проверяемыми на ее основе.

Более принципиальный характер имеет второй дефект. Ввиду пионерского характера предпринятой попытки количественного описания активности было бы особенно важно оценить корреляцию не только между частотой и средним арифметическим четырех параметров, но и между частотой и отдельными параметрами и между частотой и различными комбинациями параметров по два и по три. Наличие всех этих коэффициентов позволило бы судить об удельном весе каждого из параметров. Вместе с тем их наличие дало бы и определенную информацию для оценки, хотя бы и косвенной, выявленных выше методических допущений исследователя: насколько оправдан выбор самих параметров и невзвешенной средней в качестве их суммарного представителя. Правда, данные таблиц 19 и 20 (там же, с. 36)позволяют читателю самостоятельно провести такое независимое ранжирование по каждой из 4 составляющих активности, вычисленных по словарю Уэбстера, и независимое же сопоставление полученных рядов с частотным рядом, и как будто бы показывают существенно меньшую корреляцию между частотой и отдельными составляющими активности. Использование этих данных, однако, опять же затруднено отсутствием указаний на то, сколько именно слов учитывал при подсчете своих коэффициентов сам В.А. Московичь - расхождение в числе слов наблюдается даже между этими двумя таблицами.

Наконец, работа В.А. Московича не позволяет сделать сколько-нибудь уверенных прогнозов о справедливости исходной гипотезы за пределами рассмотренного материала. Дело не только в ограниченности последнего - максимальный из приводимых списков английских прилагательных цветообозначений (там же, с. 20) насчитывает всего 26 единиц, - но и в его специфичности. Поле цветообозначений, как и поле родства, слишком "идеальный" объект для обнаружения лексикологических закономерностей, и при экстраполяции найденных в нем регулярностей на другие части лексики надо быть очень осторожным. Очевидно, понимая это, исследователь попытался в другой части своей книги (гл. Ш, раздел 3) проверить гипотезу о зависимости между частотой и активностью, применив ту же методику к качественно иному материалу - 22 словам, отобранным из частотного словаря по теме "Двигатели внутреннего сгорания". И данные о частоте, и сведения об активности определялись как по общеязыковым, так и по техническим словарям (различной степени специализации). Всего было проведено 10

попарных сравнений частоты и активности (полученные значения коэффициента ранговой корреляции представлены в таблице 56 - там же, с. 157). Как и можно было ожидать, между частотой в специальных текстах и активностью, измеренной по общеязыковым словарям, корреляция практически отсутствует ($\rho \approx 0$). Также обстоит дело при сравнении "общеязыковой" частоты с активностью, измеренной по специализированному автотракторному словарю. В то же время исследователь считает возможным утверждать, что "между частотой и активностью слов в технических текстах обнаруживается корреляция <...>. Высокая степень корреляции отмечается также и при сравнении частот слов и их активности в языке вообще <...>. Хотя абсолютные величины коэффициентов ранговой корреляции оказываются более низкими, чем в эксперименте, проведенном на материале цветообозначений, хорошо прослеживается зависимость между частотой слов и их активностью" (там же, с. 157). К сожалению, эти утверждения не согласуются с теми цифрами, на которых они должны были бы основываться. Коэффициенты для частоты и активности по техническим словарям немногим превышают +0,2, а максимальный из коэффициентов для частоты и активности, определенных по общеязыковым словарям, едва превосходит порог 0,5, который обычно считается минимальным в статистических исследованиях. К тому же обнаружилась существенная разница в коэффициентах корреляции частоты по Торндайку с активностью по уэбстеровскому и оксфордскому словарям.

Таким образом, дополнительная проверка исходной гипотезы не подтвердила наличия зависимости между активностью и частотой ни в пределах субъязыка, ни для языка в целом. Две серии экспериментов - на материале цветообозначений и на материале технических терминов - дали прямо противоположные результаты. Сам В.А. Москович этой противоположности - сознательно или бессознательно, решать не беремся - не заметил, а "более низкие величины коэффициентов корреляции" во второй серии объяснял тем, что зависимость между частотой слов и их активностью вычислена в ней "не на материале системно организованной группы слов, а на материале произвольно выбранного списка слов" (там же, с. 157-158). Такое объяснение трудно признать удовлетворительным. Во-первых, мы не знаем характера влияния системного выбора слов ни на частотность, ни на активность - очевидно, что в предположении справедливости основной гипотезы, выбрав самые частые слова из нескольких сравнимых по частоте "системно организованных групп", мы мо-

жем получить большее согласие между частотой и активностью, чем внутри каждой из этих групп. Во-вторых, если бы гипотеза о зависимости между частотой и активностью выполнялась лишь в пределах замкнутых семантических полей, она не могла бы служить основой при отборе лексики для общеязыковых словарей и утратила бы значительную долю своего практического интереса.

Предложить убедительное альтернативное объяснение указанного расхождения мы не беремся. Не исключено, что сами рассмотренные совокупности слов (26 в одном случае и 22 в другом) слишком малы для получения статистически достоверных выводов. Так или иначе, расхождения в результатах двух серий экспериментов В.А. Московича слишком разительны, чтобы хоть одну из них можно было признать сколько-нибудь доказательной.

1.3. Обнаружение многообразия конкретных форм зависимости. Тем не менее сама возможность квантификации различных аспектов значимости слова была в книге Московича продемонстрирована, и можно было ожидать, что проверка столь заманчивой гипотезы на этом не кончится. Так и случилось – следующая попытка была предпринята в рамках информатики, для которой – ввиду больших размеров словарей информационных систем и трудностей их составления – потенциальные практические выгоды в случае подтверждения гипотезы о взаимосвязи частоты и активности оказались бы особенно велики. Исследование, описанное в статье Андрукович П.Ф., Королев Э.И. 1977, выполнялось в ходе работ по изучению и совершенствованию политематической информационно-поисковой системы "Алмаз". Частоты лексических единиц определялись на массиве расширенных заголовков рефератов, которые для этой системы фактически являются при поиске заместителями реальных документов общей длиной в 700.000 словоупотреблений, а парадигматические свойства единиц определялись по их поведению в том же массиве либо самими авторами, либо экспертами-разработчиками системы. Благодаря политематичности системы в ее документах достаточно много не только узко-специальной, но и общеязыковой лексики, что делает результаты исследования интересными и для изучения языка в целом. Однако при таком общелингвистическом их использовании не следует забывать, что лексические единицы информационно-поискового языка не тождественны внешне совпадающим с ними лексемам естественного

языка, да и свойства их могут оказаться совсем другими, вряд ли мы согласимся считать, что слова время, период, момент или сигнал и команда, трактуемые исследователями (там же, с. 6) как синонимы, являются таковыми и в общенародном русском языке*. Поэтому конкретные выводы П.Ф. Андруковича и Э.И. Королева могут использоваться лингвистами лишь после тщательной перепроверки. Вместе с тем тщательность и методическая выверенность, характеризующие планирование и исполнение собственно математико-статистической части данного исследования, и его масштабы, ставшие возможными благодаря активному вовлечению ЭВМ, делают его в целом очень поучительным для лингвистов.

В число изучаемых парадигматических свойств слова П.Ф. Андрукович и Э.И. Королев включили 3 из 4-х свойств, рассматривавшихся Б.А. Московичем. Четвертый признак, "число фразеологизмов", авторы заменили на признак "число словосочетаний, в которых данное слово является аргументом некоторой лексической функции" - как известно, понятие "лексической функции" (см., напр., Апресян Ю.Д. и др. 1969), являющееся одним из достижений современной семантики, в определенном смысле может рассматриваться как обобщение понятия фразеологизма. Кроме того, в данном исследовании словам сопоставлялись еще два признака - число синонимов и число инициальных аббревиатур, образованных от словосочетаний, содержащих данное слово.

Независимой переменной считалась частота, слова группировались не по заданным интервалам частот, а по определенному числу слов в группе - по 20.

Что касается зависимых, парадигматических переменных, то в отличие от Б.А. Московича П.Ф. Андрукович и Э.И. Королев не стали объединять их в некоем едином показателе, а предпочли исследовать связь с частотой для каждого параметра в отдельности. Это позволило им получить значительно большую и более надежную информацию. Для оценки среднего значения распределения каждого из зависимых параметров использовалась не средняя арифметическая по группе слов, а медиана, менее подверженная смещениям из-за отдельных "аномальных" представителей группы.

Из всех 6 параметров ципфовская гипотеза о пропорци-

* На мой взгляд, методика определения синонимов вызывает ряд сомнений и применительно к данному информационному языку.

ональности корню квадратному из частоты слова выполняется лишь для числа лексических функций. Число синонимов оказалось связанным линейной зависимостью уже не с квадратным, а с кубическим корнем из частоты. Вид зависимости между частотой и числом аббревиатур остался неопределенным, известно лишь, что "большинство" аббревиатур "приходится на наиболее частые слова ($\geq$ 400 словоупотреблений)" (там же, с. 7). Что же касается числа значений, дериватов и сложных слов, то они также растут пропорционально корню кубическому из частоты, но лишь в некотором "среднем" промежутке не очень малых и не очень больших частот. "Со 110–130 словоупотреблений число дериватов не увеличивается, при 140–160 словоупотреблениях перестает увеличиваться число значений" (там же, с. 6), где-то, судя по графику, перед 500 словоупотреблениями существенно замедляется и рост числа дериватов. Обнаружение подобной верхней границы подчеркнуло значение ранее обычно игнорировавшейся прозорливой оговорки Ципфа "о нескольких дюжинах самых частых слов", носящей, по-видимому, не технический, а принципиальный характер. То обстоятельство, что аналогичная граница не обнаружена при анализе изменения числа синонимов, связано, возможно, с отмечавшимся выше произволом в установлении синонимических отношений, в частности, с завышением числа синонимов из-за систематического зачисления в синонимы слов, связанных родо-видовыми отношениями.

Одновременно авторы обнаружили и существование нижней границы зон пропорциональной зависимости: при низких частотах (для количества значений – до 40–50 словоупотреблений, для числа дериватов – до 15–20, для сложных слов, судя по графику, – до 20–50) значения парадигматических характеристик либо практически постоянны, либо растут значительно медленнее, чем в зоне средних частот.

Если Л.Ф. Андрукович и Э.Н. Королев выявили возможность варьирования характера взаимосвязи от одного парадигматического параметра к другому, то Ю.А. Тулдава, рассматривая, как и Дж. Ципф, лишь один параметр, показал, что характер зависимости может меняться от языка к языку. По его данным (Тулдава Ю.А., 1979, с. 126–129), в эстонском языке зависимость числа значений m от абсолютной частоты F лучше всего аппроксимируется "модифицированной экспоненциальной функцией", а в русском языке применение аналогичной формулы приводит к завышению в области малых частот. Правда, переход к другой, степенной, или "аллометрической", функции $m = \alpha F^{\gamma}$

дает хорошие результаты для обоих языков (а по предположению Ю.А. Тулдавы, она "подходит для большого числа" и других языков). Но и при таком подходе языковая специфичность зависимости не исчезает, а находит свое отражение в варьировании констант α и γ.

Построив доверительные интервалы для числа значений у слов, принадлежащих к различным зонам по частоте, Ю.А. Тулдава обнаружил, что в эстонском языке у высокочастотных слов число значений бывает не только меньше ожидаемого (что соответствует предсказанию Ципфа и результатам Андруковича и Королева), но и значительно больше ожидаемого (Тулдава Ю.А., с. 132-136).

I.4. Некоторые общие черты лингвостатистических подходов к исследуемой проблеме. Как видим, работы последних лет продемонстрировали большую пестроту изучаемого класса явлений. Необходим новый синтез, и в его преддверии представляется полезным суммировать некоторые общие черты, объединяющие рассмотренные лингвостатистические работы независимо от масштаба и методической обоснованности каждой из них. Может быть, такое суммарное представление поможет найти какие-то пути для будущих исследований, а также подскажет, на что именно лингвостатистикам имело бы смысл обратить внимание в трудах представителей других дисциплин, обращавшихся к данной проблематике.

1. Во всех рассмотренных нами работах исследуются лишь такие системные свойства слов, которые допускают непосредственное (первичное) измерение (и более того - целочисленный подсчет). В связи с этим более тонкие и, возможно, более существенные аспекты, определяющие значимость, относительную важность слов, но допускающие лишь косвенное, производное выражение через непосредственно измеряемые величины*, могут оставаться в тени.

2. Увлечение лишь непосредственно измеримыми величинами, по-видимому, сказывается и в характерном для большинства работ обращении направления исследуемой зависимости. Поскольку трудно предположить, что такие внешние свойства слова, как число его дериватов или образованных с его участием сложных

* Такое производное выражение можно получить, напр., с помощью факторного анализа непосредственно измеренных величин. Характерно, однако, что факторный анализ, столь распространенный сегодня в лингвостатистике, в исследованиях взаимосвязи значимости и частоты до сих пор не нашел применения.

слов и фразеологизмов могут детерминировать частоту употребления самого этого слова, то ищется и исследуется уже зависимость этих свойств от частоты. Пока под значимостью слова разумеются в первую очередь такие внутренние свойства, как характер значения или степень необходимости последнего для регулярного осуществления коммуникации, частота естественно (см. выше преамбулу к данной статье) выступает в качестве детерминируемой переменной, ограничение внимания лишь внешними аспектами превращает ее в детерминирующую (см., напр., явную и четкую формулировку в: Андрукович П.Ф., Королев Б.И. 1977, с. 4). Безусловно, связи "от частоты" по отношению к некоторым парадигматическим свойствам могут быть и первичны (напр., в силу стремления пишущих к избеганию навязчивых повторений рост употребительности слова должен вести к увеличению числа синонимов, а не наоборот). Но надо четко осознавать, что они не исчерпывают всего явления.

3. Все разобранные нами подходы строго синхроничны. Между тем связи значимости (важности) и частоты могут особенно рельефно приоткрываться в исторической динамике, напр., в случаях роста числа значений слова или его употребительности. Существование исторических словарей и частотных словарей для разных хронологических "срезов" делает данный подход, по крайней мере в принципе, реализуемым.

4. С возможностью "исторического" подхода связаны и желательность и целесообразность наблюдения за индивидуальным поведением отдельных слов в противовес господствовавшей до сих пор тенденции к валовому обследованию совокупностей слов.

2. Опыт составления учебных словарей-минимумов

Поскольку, как мы видели в предыдущем разделе, в теоретическом плане проблема соотношения частоты и парадигматической значимости исследована пока явно недостаточно, имеет смысл проследить, как трактовали эту проблему представители различных областей прикладной лексикографии. В первую очередь нам следует обратиться к так называемой "учебной лексикографии", имеющей наибольший опыт использования данных частотных словарей, сравнительно хорошо описанный в научной литературе.

Необходимость массового обучения иностранным языкам не могла не поставить перед преподавателями и теоретиками-методистами вопрос о том, чему, собственно, следует обучать, какова должна быть конечная цель и всего процесса обучения, и

его базисного этапа. Так возникла проблема определения того минимума языковых средств, который должен усвоить обучающийся и которому его, собственно, и следует обучать. Естественно, что одним из первых или даже первым из возможных ответов на этот вопрос явилась идея о том, что в первую очередь надо усвоить и знать то, что чаще всего употребляется в речи. В плане изучения лексики это означало, что при отработке лексического минимума следует опереться на данные частотных словарей, включая в минимум определенное число наиболее частых слов. Собственно, первые частотные словари английского языка - Дж. Ноулза и Р. Элдриджа - создавались именно в связи с задачами обучения языку: разработкой системы чтения для слепых и словаря-минимума для рабочих-иммигрантов, желающих овладеть разговорной речью (см.: Алексеев П.М. 1971, с. 161). И в дальнейшем связь между статистической и учебной лексикографией оставалась столь же тесной (ср. само название знаменитого словаря Thorndike 1931 и ряда других подобных словарей).

Существенно, что связь эта носила характер двусторонний. То, что обучение языкам долго рассматривалось как основная область применения частотных словарей и как основное оправдание их существования (см. обзор в: Фрумкина Р.М. 1964, с. 5), в общем, вполне естественно - не забудем,что и развитие теоретической лингвостатистики, и создание систем автоматической обработки информации относятся уже ко второй половине 20-го века. Но вплоть до конца сороковых годов наблюдалось и обратное: если не считать "логико-семантического" метода, выдвинутого и воплощенного сначала Ч.Огденом и А.Ричардсом в языке "Basic English", а затем М. Уэстом*, метод отбора лексики на основе частотных словарей был фактически единственным теоретическим методом, которым располагала учебная лексикография и который использовался при создании большинства учебных минимумов (см. библиографические указания в: Головачева А.К., Санаева О.П. 1973; специальная работа Bongérs М.Н. 1947 осталась для меня недоступной). Противовесом опоре на частотный словарь являлся лишь эмпиризм, опора на субъективное чувство языка, а, как писали создатели

* Сущность этого метода отбора (см. о нем: (Gougenheim G. e.a. 1964, ch. 2) состоит во включении в словарь-минимум тех слов, с помощью которых можно описать смысл (дать определение) возможно большего числа прочих слов данного языка. По существу, это прообраз теории семантических множителей в современной семантике.

знаменитого "элементарного французского языка", "совсем не учитывать частоты и создавать учебный словарь эмпирически - значит обречь себя на создание работы недолговечной (fragile) и субъективной. Методы, которыми пытались этот эмпиризм исправить, представляются нам лишь паллиативами" (Gougenheim G. e. a. 1964, p. 137).

При всей справедливости приведенного высказывания нужно отметить, что и возражения "эмпиристов" были небеспочвенны. Ведь лежащий в основе использования частотных словарей тезис: "... поскольку все слова запомнить невозможно, то надо начинать с самых "важных", а самыми "важными",очевидно, являются те, которые чаще всего нужны, чаще всего употребляются" (Фрумкина Р.М. 1964, с. 5), - использует неопределяемое понятие "важности". Ввиду его неопределенности легко возникал соблазн просто приравнять "важность" к частотности и тем самым отказаться от дальнейшего исследования этого понятия и выяснения его сущности. Возражения "эмпиристов" (укладывавшиеся, как правило, в формулы "А мне кажутся "важными" вот такие-то редкие слова" или "Вы говорите, что знания стольких-то самых частых слов достаточно для понимания 85% словоупотреблений любого текста. Но, может быть, оставшиеся 15% и есть самое важное, без чего текст в целом все равно останется непонятным?") как раз и напоминали адептам частотного метода отбора, что их основной тезис пока не доказан и остается, по сути дела, лишь одной из возможных гипотез.

Однако на стороне этой гипотезы была по крайней мере объективируемость и последовательность предлагаемой процедуры, возражения эмпиристов, препятствуя догматизации этой гипотезы, не содержали каких-либо столь же конструктивных критериев "важности" и поэтому не могли составить ей жизнеспособной альтернативы. Лишь в середине 40-х - начале 50-х годов были, наконец, предприняты серьезные попытки противопоставить критерию частотности другие критерии "важности" слова или дополнить его другими критериями, связанные с именами крупнейших советских и французских методистов. Мы имеем в виду опыт коллектива создателей так называемого "элементарного французского языка" (именуемого ныне также 1-й ступенью "основного (fondamental) французского языка") и более ранние, но, к сожалению, гораздо менее известные лингвистам работы по созданию словаря-минимума по иностранным языкам для советской средней школы в Научно-исследовательском институте методов обучения Академии педагогических

наук, проводившиеся под руководством И.В. Рахманова и при участии виднейшего теоретика лексикографии акад. Л.В. Щербы. Именно из-за подчеркнутой анти- или нестатистичности исходных установок опыт этих исследований по лингводидактике и оказывается, на мой взгляд, особенно важным для лингвостатистики.

2.2. Роль частоты в системе критериев И.В. Рахманова. Непосредственная критика статистического метода отбора лексики, содержавшаяся в Рахманов И.В., Щерба Л.В. 1947, с.6-7, серьезного значения не имеет. Порицая использование частоты, авторы в то же время одобряли (там же, с. 7) использование в работах Палмера показателя употребительности, или ранга, слов, не сознавая, видимо, что ранг также является статистической характеристикой. Зато совершенно особый интерес представляет позитивная программа авторов. Выделив ряд парадигматических свойств, обладать которыми слово может в большей или меньшей степени, авторы "Словаря-минимума" фактически противопоставили измерение значимости слова по этим параметрам подсчету его частоты как более обоснованный - и с лингвистической, и с методической точек зрения - критерий "важности" слова при обучении языку. Список свойств в последовательных вариантах словаря подвергался изменениям, менялось и понимание отдельных характеристик - напр., в 1947 г. под "семантическим принципом" отбора подразумевался отбор слов, обладающих большей сочетаемостью с другими словами, а под "семантическим индексом" - некая мера этой сочетаемости (Рахманов И.В., Щерба Л.В. 1947, с. 14-15), а в вариантах 60-х годов под "семантической ценностью" слова понималась степень употребительности и общности обозначаемых этим словом "понятий и явлений" (Рахманов И.В. 1960, с. 11; Рахманов И.В. 1967, с. 19-20). Однако общая направленность - стремление к определению "важности" слова на основе не статистических, а собственно-языковых и по преимуществу парадигматических его свойств - оставалась без изменений. Правда, в число принципов отбора включалась и частотность, но если в "Словаре-минимуме" 1947 г. введенные принципы не были никак упорядочены по своей силе или очередности применения и можно было предполагать, что этот принцип стоит в перечислении на последнем месте просто в силу своей привычности для методистов, то после введения в словарях 60-х годов деления принципов на "ведущие" и "дополнительные" частотный принцип был

открыто включен в число дополнительных, т.е. тех, "которые имеют лишь второстепенное значение" (Рахманов И.В. 1967, с. 18, 17).

Таким образом, перед нами концепция, отчетливо противопоставляющая частоту и парадигматическую значимость слова. Более того, поскольку это не просто теоретическая концепция, а методическая программа, долженствующая служить руководством к действию, то можно утверждать, что в эту концепцию входит и молчаливое допущение, что между частотой и прочими выделенными свойствами нет сколько-нибудь серьезной корреляции. Попытаемся оценить степень состоятельности этой концепции и понять уроки, которые можно извлечь из ее эволюции и судьбы.

Согласно позднейшему из вариантов данной программы, помимо частотности выдвигаются шесть критериев отбора: 1) широта сочетаемости; 2) стилистическая неограниченность (т.е. возможность употребления и в письменной, и в устной речи и отсутствие территориальных ограничений на употребление); 3) семантическая ценность (в указанном выше смысле); 4) способность к словообразованию; 5) количество значений: "чем больше различных значений имеет слово, тем важнее его знать"; 6) "способность слова нести какую-либо конструктивную функцию в языке, либо в качестве компонента фразеологического сочетания, либо в качестве служебного слова". Первые три критерия считаются основными, ведущими, а остальные три, как и частотность, - дополнительными (Рахманов И.В. 1967, с. 18-23; статья Рахманов И.В. 1969 отличается лишь заменой языковых примеров).

Если рассматривать этот перечень не как перечень принципов отбора, а как перечень свойств слова, определяющих его место и связи в системе языка, то нельзя не признать, что перед нами действительно глубоко продуманное и тщательно выполненное построение, могущее служить описанием, а в какой-то степени - и определением парадигматических свойств и парадигматической значимости слова. (Легко заметить, что рассмотренные выше параметры "активности" слова по В.А. Московичу в основном повторяют как раз "дополнительные" свойства из числа выделенных И.В. Рахмановым). Особую достоверность этому описанию придает то, что составители не ограничились какой-либо одной произвольно выбранной из языковых подсистем, скажем, словообразованием или стилистическим разбиением лексики, но постарались учесть их все, что больше отвечает

природе языка как "системы систем". Но это теоретическое достоинство может обернуться и практическим неудобством и даже неэффективностью при применении данного построения для его основной цели – отбора слов для учебного минимума. Ведь самая многочисленность критериев показывает, что они могут оказаться малосогласованными или вовсе противоречащими друг другу. И.В. Рахманов и Д.Л. Щерба (1947, с. 16-17) сами отмечали, что "на практике неукоснительное соблюдение всех вышеизложенных принципов, или даже только нескольких из них, является невозможным". В этих условиях перечень принципов отбора необходимо дополнить либо каким-то правилом оптимизации, представляющим собой некоторую функцию от набора значений, принимаемых на каждом слове всеми выделенными критериями, либо точной инструкцией, определяющей порядок применения различных принципов отбора.

Ни то, ни другое сделано не было. При подготовке "Словаря-минимума" его составители заполняли на каждое слово стереотипный формуляр (см.: Рахманов И.В., Щерба Д.Л. 1947, с.15-16), в графах которого определенным образом – числом или крестиком – фиксировались значения, которые принимает на этом слове каждый из введенных критериев, а затем относительно каждого конкретного слова приходилось принимать отдельное решение и в целом ряде случаев поступаться тем или иным принципом, если в применении к данному слову он вступал в противоречие с другими" (там же, с. 17). Поскольку никаких общих правил предпочтения, отдаваемого одним критериям перед другими, не предлагалось, то тем самым открывалась лазейка для тех же "чисто субъективных поправок, основанных на собственном педагогическом опыте" или собственном чувстве языка, несостоятельность которых в качестве коррективов к принципу частотности была отмечена самими составителями (там же, с. 7). Правда, в вариантах 60-х годов появилось, как мы видели, деление принципов на основные и дополнительные, но и здесь внутри каждой из этих групп иерархия применения отсутствовала.

Поэтому практическое применение методики отбора, разработанной под руководством И.В. Рахманова, требует непременного участия высококвалифицированных специалистов с тонким чувством языка. Весьма показательно, что, когда, напр., в Секторе лексикологии и стилистики Научно-методического центра русского языка при МГУ занялись проблемой отбора русской лексики для учебных словарей, предназначенных для иностран-

цев, то следование принципам И.В. Рахманова и его соавторов осталось, насколько можно судить по опубликованным статьям, лишь почтительной декларацией. Так, П.Н. Денисов писал: "... решено было присоединиться к принципам, разработанным в Академии педагогических наук РСФСР под руководством И.В.Рахманова", но, перечислив эти принципы, тут же добавлял: "За исходный принцип был взят показатель частотности, который уточнялся по тематическому критерию и по критерию богатства синтактико-семантических связей того или иного слова. <...> Практически работа проходила так. Мы взяли частотный список Э.А. Штейнфельдт и сверили его со словарем-минимумом того же автора" (Денисов П.Н. 1969, с. 35-36. Подчеркнуто мною - С.Г.). Как видим, фактически от принципов И.В. Рахманова не осталось и следа.

Если сотрудники НМЦ РЯ предпочли остановиться на методике, допускающей более простую и конструктивно-алгоритмизуемую реализацию без предварительного исследования сравнительной эффективности отбора по частоте и по совокупности критериев И.В. Рахманова, то их выбор свидетельствует лишь о сложности практического применения последней. Но в идеале выбор методики должен, конечно, определяться прежде всего качеством отбираемого минимума. Предпочтение более простого в реализации частотного критерия более сложной, но зато и более адекватной и обоснованной системе критериев И.В. Рахманова будет не вынужденной мерой, а методически оправданным подходом лишь в том случае, если удастся статистически показать, что список слов, отбираемых на основе этого критерия, достаточно близок к тому списку, который составляется на основе методики И.В. Рахманова, что оба списка в высокой степени согласуются друг с другом, а различия либо несущественны, либо могут быть скорректированы с помощью достаточно простых и регулярных приемов. Подобный исход статистического сравнения явился бы убедительным подтверждением гипотезы о наличии тесной зависимости между частотой и парадигматической значимостью.

К сожалению, описания какого-либо статистического сравнения двух списков, один из которых отобран строго по частотным соображениям, а второй взят из словарей-минимумов под редакцией И.В. Рахманова или составлен по той же методике, найти не удалось. Однако результаты подобного эксперимента в какой-то мере предсказуемы на основе чисто теоретического анализа самих критериев, предложенных И.В. Рахмановым и его

соавторами. Перефразируя известную пословицу, можно сказать: "Гони частотность в дверь, она вернется в окно". В самом деле, из трех основных критериев И.В. Рахманова два, широта сочетаемости и стилистическая неограниченность, по самому своему определению должны коррелировать с частотой употребления, естественно ожидать, напр., что в выборке, охватывающей различные "стили", слова, употребляющиеся во всех "стилях", при прочих равных условиях, будут как правило, иметь большую частоту, чем слова, употребляющиеся лишь в одном из "стилей". Третий из основных критериев, семантическая ценность (в понимании, которое придано ей в словарях 60-х годов, см. выше) вообще есть ничто иное, как частота употребления, только уже не самих слов, а тематических "классов" ("концентров"), к которым эти слова принадлежат. Наконец. среди вспомогательных признаков названа "строевая функция", т.е. способность входить в состав фразеологизмов или быть служебным словом. Разделив этот критерий на два и оставив в стороне фразеологизмы, получаем критерий, тесная корреляция которого с критерием частоты не подлежит сомнению, так как все без исключения частотные словари свидетельствуют, что служебные слова имеют высокую частоту. Таким образом, из всей системы критериев, предложенных И.В. Рахмановым, лишь про три "второстепенных" - число значений, вхождение в состав фразеологизмов и словообразовательную способность - нельзя a priori сказать, связаны ли они как-то с частотой слов. Но как раз для соответствующих признаков существование определенной, хотя и довольно сложной, зависимости было продемонстрировано в эксперименте П.Ф. Андруковича и Э.И. Королева. Поэтому естественно ожидать, что сравнение частотного списка со списком, построенным по методике И.В. Рахманова, должно в принципе показать достаточно высокую их корреляцию*.

Таким образом, разработка последовательной и теоретически обоснованной антистатистической методики отбора лексики

* По отношению к трем из критериев И.В. Рахманова - стилистической неограниченности, семантической ценности и строевой способности - к аналогичным выводам пришла М.А. Скопина: "... для начального этапа обучения решение вопроса об основной массе строевых и стилистически нейтральных слов предопределяется самим частотным списком, а принцип стилистической неограниченности слова и его строевой способности оказывается соответственно избыточным. Более внимательное рассмотрение стилистически нейтральных знаменательных слов исходного частотного списка с точки зрения отраженной в них тематики показало, что в той или иной степени они охватывают почти все темы, обычно изучаемые на начальном этапе" (Скопина М.А. 1972, с. 76).

для учебных минимумов, как это часто бывает в истории науки, привела, насколько можно судить в настоящее время, как раз к обоснованию опровергаемой методики. Раньше было неизвестно, что именно отбирать для минимумов, и статистической методикой пользовались просто как наиболее объективируемой в надежде, что то, что получится, и есть то, что нужно. Теперь же, когда благодаря работам И.В. Рахманова и его соавторов мы в общем знаем, что именно нужно отобрать, оказалось, что статистический метод отбора как раз и позволяет отобрать значительную часть этого необходимого массива лексики. Сегодня не только "заинтересованные" теоретики лингвостатистики утверждают, что, хотя частотный словарь "не может быть непосредственно использован в качестве словаря-минимума", он "является необходимой и единственно надежной основой для его составления" (Фрумкина Р.М. 1964, с. 20), но и практики-методисты все чаще свидетельствуют, что в основу словарей-минимумов кладутся именно частотные списки, которые затем подвергаются своего рода "выравниванию" на основе тематических, логических и других лингвометодических критериев (см., напр., Новиков Л.А. 1969 с. 48-49). Так, в словаре-минимуме для эстонских школ по частотному словарю отобрано 75% слов (Роовет Э., Штейнфельдт Э.А. 1965, с. 13), в словнике М.А.Скопиной - 1050 слов из 1300, т.е. более 80% (Скопина М.А. 1972, с. 76-77).

2.3. Понятие частоты и понятие disponibilité. Выявлению теоретической состоятельности и корректировке частотного принципа отбора лексики для учебных минимумов невольно содействовала и критика этих принципов создателями "элементарного французского". Эти ученые обнаружили (см.: Gougenheim G. e.a. 1956; 1964), что в существующих частотных списках очень мало конкретной лексики, особенно - конкретных существительных типа зубы, вилка, автобус, ножницы, пуговица и т.п., важность знания которых для овладения языком не подлежит сомнению. Сначала было высказано предположение, что отсутствие конкретной лексики в частотных словарях объясняется тем, что все эти словари составлялись на материале письменной речи. Тогда французские ученые создали пионерский и уникальный по своим достоинствам словарь повседневной разговорной устной речи (методику его составления и сами частотные списки также можно найти в Gougenheim G. e.a. 1956). "Однако пришлось с удивлением констатировать, что в разговорной речи, так же

как и в письменной, многие конкретные существительные, которые a priori представляются очень употребительными и даже необходимыми, не обладают частотностью, достаточной для того, чтобы обеспечить им место в кратком словаре" (Мишеа Р. 1967, с. 290). Оставалось признать, что "конкретные существительные имеют весьма слабую частотность. Более того, эта частотность непостоянна, нестабильна" (Гугенейм Г. 1967, с. 302), т.е. сильно варьирует в зависимости от темы текста. Из этого факта был сделан радикальный и не вполне корректный вывод, что "конкретное существительное само по себе не имеет частотности" (там же; в подлиннике, конечно, frequence "частота"), и в противовес критерию высокой частоты для конкретных существительных был выдвинут критерий их disponibilité, т.е. их постоянного присутствия, наличия в сознании носителя языка. Была разработана и специальная методика экспериментального выявления слов, удовлетворяющих этому новому критерию - так называемых mots disponibles, заключающаяся в том, что испытуемым задают определенную бытовую тему, например, "путешествие по железной дороге" и просят записать некоторое фиксированное количество слов, относящихся к данной теме и первыми "пришедших на ум" (Гугенейм Г., с. 304; подробное описание и анализ результатов применения - Gougenheim G. e.a., 1956). Очевидно, что при условии соответствующей математической обработки получаемых анкет данная методика позволяет не только выявить "постоянно присутствующие" слова*, но и всем встретившимся словам сопоставить оценку их "степени присутствия (наличности)", причем процедура такого сопоставления будет не менее конструктивна и объективна, нежели подсчет частот тех же слов.

В лингво-методической литературе высказывалось мнение, что критерий "постоянного присутствия" представляет собой всего лишь "своеобразную подгонку под ответ" (Морковкин В.В., 1972, с. 18; там же ссылка на "сходную точку зрения" П.М. Алексеева). Мол, заранее известно, что "слова типа ключ, вилка", не попадающие в частотные словари, нужны при обуче-

* Такой перевод представляется более адекватным, чем ранее использовавшиеся варианты: обиходные, резервные, наличные слова. Ведь mots disponibles не находятся в "резерве", а все время, хотя бы и молчаливо, но используются носителями языка, и далеко не всякое слово с "обиходным" значением оказывается disponible. Перевод наличные больше соответствует оригинальному термину, однако и он не содержит важного смыслового элемента, а именно - "постоянства", высокой степени "наличия".

нии языку, вот и вводится налепляемый на эти заранее известные слова ярлык "постоянно присутствующие". Однако декларируемый здесь логический круг - лишь кажущийся. Ведь a priori известно лишь то, что при обучении необходимо сообщить обучаемому и какое-то количество конкретной обиходной лексики. Но какие именно обиходные слова должны быть включены в словарь-минимум - будет ли это, скажем, упомянутое В.В. Морковкиным слово вилка или же не упомянутое им чапельник? В этом-то месте и начинает работать критерий постоянного присутствия, и основанная на нем процедура дает составителям минимумов надежный инструмент для отбора.

Но значение этого критерия не исчерпывается лишь его практической полезностью. С его введением расширяются и становятся более адекватными современным общелингвистическим взглядам те теоретические представления о способах существования языка и системе его функций, которые лежат в основе дидактико-методических построений. Идея составления учебных словарей-минимумов на базе одних только частотных словарей по-существу опирается на молчаливую презумпцию о том, что роль языка в человеческой деятельности сводится исключительно к построению и восприятию речевых сообщений (текстов), т.е. к обеспечению коммуникации с себе подобными. Но при всей своей важности коммуникативная функция языка все же не является единственной. Многие по внешности "молчаливые" виды нашей деятельности на самом деле в той или иной мере опосредованы языком и без него были бы или немыслимы, или бы приняли совершенно иной вид. Такова, в частности, и наша ориентация в окружающей действительности, и ее освоение в повседневном опыте, связанные с вычленением и осознанием предметов и явлений. Дополнение частотного критерия критерием постоянного присутствия как раз и позволяет отразить этот аспект существования языка.

Для изучения взаимосвязи между статистическими и парадигматическими свойствами слов понятия "постоянного присутствия" и "степени присутствия" интересны по крайне мере в трех отношениях. Прежде всего, как показал эксперимент А.П. Василевича и Р.М. Фрумкиной (Фрумкина Р.М., 1971, с. 55-65), именно на постоянно присутствующие слова приходится большая часть расхождений между частотами слов в речи и субъективными представлениями об употребительности слов. Этот факт открывает перспективу сопряжения двух ветвей лингвостатистических исследований, рассмотренных выше в § 1, - тех, где

вслед за Ципфом рассматриваются частоты слов, и тех, в которых, как в **Noble C.E.**, 1953, вместо частот берутся их психологические корреляты.

С другой стороны, не исключено, что будучи по самому своему определению характеристикой психологически релевантной, степень присутствия слова в сознании, используемая как одна из мер "значимости" слова в языке, окажется теснее связанной с характеристиками употребительности слова в речи, нежели известные нам из § I парадигматические меры, вычисляемые по описанию языка (грамматика и/или словарь), абстрагированному от сознания конкретных носителей языка. На такую гипотезу наталкивает наблюдение Б.Ю. Нормана, самостоятельно и вне всякой связи с проблемами лингводидактики введшего аналогичную степени присутствия категорию "активности понятия в мыслительной практике" (более активное понятие и, **resp.**, слово "лежит как бы ближе "под рукой" в нашей памяти" - Норман Б.Ю. 1978, с. 27). Исследователь отмечает, что хотя слова дом и камень имеют одинаковое число значений (и, добавим, по гипотезе Ципфа должны были бы иметь частоты сходного порядка), первое из них, значительно "более актирное", и частоту имеет чуть ли не в 6 раз большую.

Возможность использовать "степень присутствия" как меру значимости слов должна побудить нас повнимательнее присмотреться к внутренней структуре этой категории: не входят ли и здесь, как то оказалось с критериями, предложенными И.В.Рахмановым, статистические и, шире, стохастические соображения в самую процедуру ее определения? Мнения такого рода уже высказывались в литературе. Так, Б.Ю. Норман свой тезис о большей "активности" слова дом по сравнению со словом камень прямо обосновывал тем, что "современный человек* в своей жизни чаще пользуется понятием "дом", чем понятием "камень"" (Норман Б.Ю., с. 27; подчеркнуто мной - С.Г.). А по экспликации А.П. Василевича и Р.М. Фрумкиной постоянно присутствующие слова суть не что иное, как слова, денотаты которых часто встречаются в "нашей жизненной практике" (Фрумкина Р.М., 1971, с. 51, 57-58).

Если понимать приведенные экспликации как отождествления, то трактовку Б.Ю. Нормана придется признать более корректной. Достаточно очевидно, что категория языкового сознания, каковой по определению является степень присутствия, не

* Точнее, наверное, было бы говорить о современном взрослом горожанине. Повседневный опыт ребенка или жителя неурбанизированной деревни может оказаться несколько иным.

может совпадать с параметром эмпирического распределения объектов внешнего мира и не может быть введена без обращения к сигнификатам, а не только денотатам слов. Но ежели видеть в этих экспликациях не отождествления, а указания на главный или единственный фактор, от которого зависит степень присутствия слова, то предложение А.П. Василевича и Р.М. Фрумкиной оказывается более заманчивым и интересным: частота понятий - фактор заведомо не наблюдаемый и того же порядка сложности и абстрактности, что и сама степень присутствия, а частота денотатов - нечто более простое и в принципе допускающее регистрацию и измерение.

Тем не менее и в экспликации Василевича и Фрумкиной стохастическая природа феномена "присутствия" представлена слишком упрощенно. Не все встречающееся и даже часто встречающееся вообще замечается человеком. Скажем, поскольку зубные щетки делаются из щетины, то с денотатами соответствующих выражений человек сталкивается одинаково часто. Тем не менее вряд ли приходится сомневаться в том, что в число постоянно присутствующих слов русского языка войдет лишь зубная щетка (или зубы и щетка, если допускаются лишь изолированные лексемы), но не щетина. Предопределять меру значимости соответствующих имен частота денотатов могла бы, если б "наша практическая жизнь", повседневный обиход человека представляли собой пассивную тотальную регистрацию встречающихся предметов и явлений. Но обиход на деле есть определенным образом организованная совокупность процессов, видов деятельности и предполагает отбор, "просеивание" окружающего мира. Предметы и явления существуют для человека в его обиходе (а не просто сами для себя) тогда, когда они либо непосредственно вовлечены в эти процессы и используются в них, либо осознаются в качестве условий и обстоятельств, влияющих на обращение к тому или иному процессу и на характер его протекания.

Подобная структура обихода не может не отражаться и языковым сознанием. В частности, степень присутствия слова будет зависеть а) от того, в какой степени денотат слова вычленяется носителями языка как самостоятельный и существенный элемент каких-либо процессов и/или ситуаций их протекания, и б) от того, насколько важными представляются сами эти процессы в общей структуре обихода. Конечно, оба эти параметра (их можно назвать, соответственно "степенью выделенности элемента" и "степенью важности процесса") будут связа-

ны с частотной структурой обихода: важность процесса с частотой обращения к нему, а выделенность элемента – с числом и суммарной частотой процессов, в которые тот вовлечен. Тем не менее они вряд ли будут целиком сводиться к этим частотам, да и эти последние достаточно далеки от без разбору фиксируемой частоты попадания предметов на глаза человеку.

Указанное разделение двух факторов – степени выделенности элемента и степени важности процесса (вида деятельности) – находит соответствие в строении описанной выше экспериментальной процедуры выявления постоянно присутствующих слов. Первый, не всегда осознаваемый ее этап, состоит в подготовке перечня тем, которые целесообразно было бы предъявить испытуемым, иными словами – в выявлении видов деятельности (процессов), важных в общей структуре обихода. На втором же этапе испытуемые называют те слова, денотаты которых кажутся им существенными для реализации именно данного вида деятельности.

Подобная двуступенчатость позволяет понять, в чем собственно заключается угаданная, но не вполне адекватно интерпретированная названными выше исследователями стохастичность феномена **disponibilité**. Когда при подсчете частот слов речевую выборку стараются организовать как случайную, максимально свободную от смещений, вызванных индивидуальными особенностями того или иного текста, это делается в презумпции, что получаемые значения частот смогут служить оценками "безусловных" вероятностей слов в языке или функциональном стиле. Процедура же определения степеней присутствия слов строится так, чтобы оценить условную вероятность слова при условии, что фиксирована некоторая конкретная тема беседы или размышления. В свете подобной интерпретации категории "степени присутствия" становится ясным, что "локальная частотность, определяемая при помощи статистического анализа определенного тематически однородного массива", которую Б.В.Морковкин (1972, с. 18) противопоставлял степени присутствия как подлинно статистический критерий отбора, на деле приобретает статистический смысл именно как эмпирическая оценка **степени присутствия***.

* Вопрос о надежности такой оценки нуждается, конечно, в специальном экспериментальном изучении. Помимо лобового сличения анкетных оценок с частотами в тематически однородных массивах текстов представляется интересным и сопоставление с анкетными оценками **disponibilité** частот слов в корпусе поэтических текстов одного автора, поскольку, по распространен-

Наш анализ показал, что в модифицированном виде стохастические характеристики и принципы присутствуют во внутренней структуре тех критериев отбора слов для словаря-минимума, которые разрабатывались в противовес частотному критерию или в дополнение к нему. Такой результат может служить эвристическим свидетельством в пользу реальности взаимосвязи между частотой и парадигматической значимостью (важностью) слов.

Л И Т Е Р А Т У Р А

Алексеев П.М. Частотные словари английского языка и их практическое применение. – В кн.: Статистика речи и автоматический анализ текста. – Л.: Наука, 1971, с. 160-178.

Андрукович П.Ф., Королев Э.И. О статистических и лексико-грамматических свойствах слов. – Научно-техническая информация. Серия 2, 1977, № 4, с. 1-9.

Апресян Ю.Д., Колковский А.К., Мельчук И.А. Об одном способе изучения сочетаемости слов. – Рус. язык в нац. школе, 1969, № 6, с. 61-71.

Гиндин С.И. Методы автоматического фрагментирования текста, опирающиеся на характеристики внутреннего состава фрагментов. – Семиотика и информатика, 1977, вып. 9, с.35-82.

Головачева А.К., Санаева О.П. О работах в области создания базовых языков. – В кн.: Теория языка и инженерная лингвистика. Л., 1973, с. 82-93. (Ленингр. пед. ин-т).

Гугенейм Г. Некоторые выводы статистики словаря. – В кн.: Методика преподавания иностранных языков за рубежом. – М.: Прогресс, 1967, с. 299-305.

Денисов П.Н. Принципы отбора лексики для учебных словарей. – В кн.: Вопросы учебной лексикографии. М., 1969, с.15-38.

Левин Ю.И. О некоторых чертах плана содержания в поэтических текстах. – В кн.: Структурная типология языков. – М.: Наука, 1966, с. 199-215.

Мишеа Р. Словари основной лексики. – В кн.: Методика преподавания иностранных языков за рубежом. М., 1967, с. 286-298.

ной концепции (см., например, Левин Ю.И. 1966), такой корпус является как бы тематически сфокусированным вокруг одного или нескольких тематических полей. Однако до сих пор данные частотных словарей поэтов сравнивались, насколько известно, лишь с общеязыковыми частотами.

Морковкин В.В. Сравнительный список наиболее употребительных русских слов (на материале шести словарей). – В кн.:Лексические минимумы русского языка. – М.: Изд-во МГУ,1972, с. 16-74.

Москович В.А. Статистика и семантика. Опыт статистического анализа семантического поля. – М.: Наука, 1969.

Новиков Л.А. Учебный словарь сочетаемости слов, его лингвистические основы и структура. – В кн.: Вопросы учебной лексикографии. – М.: Изд-во МГУ, 1969, с. 39-52.

Норман Б.Ю. Синтаксис речевой деятельности. – Минск:Вышэйшая школа, 1978.

Рахманов И.В. Предисловие. – В кн.: Аракин В.Д., Любимова З.М. и др. Словарь наиболее употребительных слов английского, немецкого и французского языков. М., 1960, с.4-19.

Рахманов И.В. Предисловие. – В кн.: Любимова З.М., Рахманов И.В. Словарь наиболее употребительных слов немецкого языка. – М.: Сов. энциклопедия, 1967, с. 5-36.

Рахманов И.В. Принципы отбора лексического минимума. – В кн.: Цетлин В.С. Словарь наиболее употребительных слов французского языка. – М.: Сов. энциклопедия, 1969, с.15-45.

Рахманов И.В., Щерба Д.Л. Предисловие. – В кн.: Аракин В.Д., Монигетти А.В. и др. Словарь-минимум по английскому, французскому и немецкому языкам для средней школы. – М.; Гос. изд-во иностр. и нац. словарей, 1947 (изд.4-е.1951), с. 4-26.

Роовет Э., Штейнфельдт Э.А. Словарь-минимум русского языка для 2-8 классов эстонских школ. Таллин, 1965.

Скопина М.А. О разработке словника для начального этапа обучения. – В кн.: Лексические минимумы русского языка. – М.: Изд-во МГУ, 1972, с. 75-91.

Суппес П., Зиннес Дж. Основы теории измерений. В кн.: – Психологические измерения. – М.: Мир, 1967, с. 9-110.

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. – Учен. зап. Тартуск. ун-та, вып. 502. Тарту, 1979, с. 107-141.

Фрумкина Р.М. Статистические методы изучения лексики. – М.: Наука, 1964.

Фрумкина Р.М. Вероятность элементов текста и речевое поведение. – М.: Наука, 1971.

Baker S.J. The pattern of language. – The journal of general psychology, 1950, v. 42, half I, p. 25-66.

Bailey R., Doležel L. An annotated bibliography of statistical stylistics. - Ann Arbor: Univ. of Michigan, 1968.

Bongérs M.H. The history and principles of vocabulary control. - Woerden, 1947.

Guiraud P. Les caractères statistiques du vocabulaire. - P.: Presses universitaires de France, 1954.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A. L'élaboration du français élémentaire. - P.: Didier, 1956.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A. L'élaboration du français fondamental (I-er degré). - P.: Didier, 1964.

Herdan G. [Выступление в прениях по докладу:] Guiraud P. Le sens et l'information. - In: Statistique et l'analyse linguistique. Colloque de Strassbourg. P., 1966.

Noble C.E. Meaning - familiarity relationship. - Psychological Review, 1953, v. 60, N 2, p. 89-98.

Thorndike E.L. A Teacher's word book of 20 000 words found most frequently and widely in general reading for children and young people. - N. Y., 1931.

Thorndike E.L. The Thorndike century senior dictionary. - N. Y.: Appleton-Century, 1941.

Zipf G.K. Repetition of words, time-perspective and semantic balance. - The journal of general psychology, 1945 a, v. 32, half 1, p. 127-148.

Zipf G.K. The meaning-frequency relationship of words. - The journal of general psychology, 1945 b, v. 33, half 2, p. 251-256.

Zipf G.K. Human behaviour and the principle of least effort. - Cambridge, Mass., 1949.

FREQUENCY AND THE SIGNIFICANCE OF THE WORD IN THE SYSTEM OF LANGUAGE

(Quest for the "inner justification" of word counts in linguo-statistics and linguo-didactic)

Sergei Gindin

S u m m a r y

This paper aims to analyse the results and the inner logic of different attempts to elucidate the connection between a word's frequency and its significance (importance, activity, relative weight) in the lexico-semantic system of a given language.

The first section deals with the linguo-statistical approaches to the problem. One of the most known of them, derived by V. Moskovich, is shown to be methodologically questionable. The most striking common feature of all these approaches is that their authors treat only such systemic properties of a word that admit direct (primary) measurement on the so-called "absolute scales".

Meanwhile it does not seem at all obvious that the real significance of a word in the language may be expressed with the help of such a simple type of quantitative feature. In this connection the second chapter of this paper deals with the history of the basic vocabulary control. The authors of the most sound and valid methods of the kind (I. Rahmanov, G. Gougenheim and their colleagues) tried to elaborate some purely qualitative or at least non-frequential principles for word selection. But closer examination reveals that frequency enters implicite into the inner structure of these principles. This conclusion may serve as serious heuristic evidence in favour of the existence of the correlation between a word's frequency and its paradigmatic significance.

СТАТИСТИЧЕСКИЕ МЕТОДЫ ВЫЯВЛЕНИЯ РЕГИОЛЕКТОВ НА МАТЕРИАЛЕ ЛИНГВИСТИЧЕСКИХ АТЛАСОВ

Т.Е. Зубова, А.В. Зубов

Среди разновидностей устной некодифицированной речи различают обычно городское и сельское просторечие, интердиалекты (наддиалектные койне), полудиалекты (переходные формы от диалекта к общенародной речи), местные диалекты и говоры. Их коренным отличием является отсутствие собственной аксиологической (образцовой) нормы на современном этапе развития, в эпоху их интенсивной интеграции в рамках единой общенациональной системы, с ориентиром на литературную норму.

Однако, исходя из понятия нормы как двустороннего явления (норма образцовая, сознательно регулируемая и норма объективная, выражающаяся в сосуществовании и конкуренции вариантов) интересующие нас разновидности речи должны обладать определенными нормативными характеристиками в виде предпочтения одних вариантов другим, т.е. в количественном аспекте. Подобную норму называют "статистической сущностью" или "количественным узусом" (Guiraud, 1962, с. 124) и подчеркивают, что "критерий употребительности и широты распространения территориального или социального варианта в случаях их оценки оказывается едва ли не решающим" (Граудина, 1970, с. 334).

Географические и социальные варианты языка, обозначаемые традиционно как территориальные диалекты и социолекты, обнаруживают в наши дни теснейшую и неразрывную взаимосвязь: территориальные диалекты становятся постепенно в высокоразвитых странах монофункциональными разновидностями, свойственными исключительно речи крестьянского населения, т.е. они "социологизованы" и вне социальной характеристики, в "чистом" виде уже не выделяются. Стремление отразить эту взаимосвязь в самом названии таких речевых вариантов приводит исследователей к поискам специальных терминов, в рамках которых можно было бы объединить обе стороны объекта.

Нам кажется вполне приемлемым предложение Л.М. Скрелиной (Скрелина, 1978. с. 134–135) использовать для этих целей введенный в обиход Р.Якобсоном термин "функциональный диалект". Функциональные диалекты, по мнению Л.М.Скрелиной, имеют одну и ту же систему референций в грамматике на уровне означаемых,

а их расхождения наиболее ярко проявляются на уровне означающих, в речевой реализации, что особенно удачно можно иллюстрировать материалами лингвистических атласов.

Действительно, обращение к лингвогеографическим материалам французского языка, наиболее полно представленного в этом плане среди других европейских языков (национальный атлас Ж. Жильерона и Э.Эдмона, региональные атласы, или, иначе, микроатласы), обнаруживают наличие многочисленных и разнообразных функциональных диалектов. Микроатласы Франции охватывают всю территорию страны одновременно и потому могут считаться синхронным срезом, отражающим речевые особенности жителей сельской местности за последние 20-30 лет. Их густая опорная сеть, в отличие от атласа Ж.Жильерона и Э.Эдмона, позволяет выявлять динамический характер вариантности языковых единиц (т.е. выбор ведущих форм) и осуществлять на этой основе разнообразные типологические исследования.

Известные до сих пор методы работы с микроатласами зарубежных и советских романистов П.Гардета, Т.Таверде, Ж.Б.Мартена, С.Эскофье, М.А.Бородиной и др. (Gardette, 1970; Taverdet, 1970; Martin, 1972; Martin, 1974; Escoffier, 1974; Бородина, 1966), основываясь на установлении ареалов функционирования отдельных языковых явлений или групп явлений.

Подобная методика достаточно эффективна в тех случаях, когда исследование ограничивается одним диалектом или группой родственных говоров (в границах регионального атласа), но оказывается мало пригодной для решения типологических задач с выходом на уровень межъязыковых связей, поскольку каждый раз (для каждой подсистемы в ее географических вариантах) изменяется не только величина и конфигурация ареалов, но и само их количество в границах региона. То же можно сказать и о диалектометрическом методе Ж.Сеги, действующем эффективно для характеристики каждого говора по набору дифференциальных признаков, но непригодном для типологических задач (Séguy, 1972, 1973).

Дело в том, что обобщение информации карт требует абстрагирования от массы несущественных, частых, узко специфических признаков говора и переноса центра тяжести на самые типичные, характерные для многих диалектов черты. Тем самым нарушается основной принцип методики Ж.Сеги с его строгой ориентацией и отбором именно специфических, узко локальных признаков говора.

Ареально-типологическое исследование атласов предъявляет новые требования. Постоянные колебания не только в величине и очертаниях ареалов, но и в их количестве препятствуют "привязыванию" систематизируемого материала к определенной территории и к определенному варианту речи. За набором разнообразных показателей исчезает из виду, как бы "распыляется" сам объект исследования.

В предлагаемой нами методике сопоставительного изучения микроатласов Франции основой является выбор какого-либо показателя как постоянной величины, на фоне которой рассматриваются всевозможные переменные величины. Постоянной величиной будем считать границы каждого регионального атласа, иначе, длину его опорной сети, принимаемую за 100%, а переменными величинами будут выступать различные подсистемы функционального диалекта, зафиксированные на картах атласа и сводимые в итоге в общую систему. Единицы таких подсистем (например, система артикля, местоимений, глагольных форм и т.д.) устанавливаются на основе количественного критерия, когда один, реже два-три варианта материальной реализации формы выбираются в качестве ведущих, репрезентирующих в нашей методике весь регион и воплощающих, таким образом, региональную норму. Выход на этот уровень позволяет исследователю, во-первых, подняться над пестротой многочисленных сосуществующих на территории региона факультативных вариантов, каждый из которых, по сути дела, представляет отдельный местный говор - патуа, а, во-вторых, преодолеть свойственные региональным атласам Франции расхождения по длине опорной сети, иногда весьма значительные, поскольку на уровне ведущих вариантов критерий величины самого региона нейтрализуется.

Разновидность функционального диалекта, единицы которого являются количественно ведущими для региона вариантами реализации языковых единиц, применительно к специфике материала мы называем регионолектом (Зубова, 1980). Этот термин наилучшим образом отражает своеобразие создаваемого в процессе анализа атласов лингвистического объекта. От традиционного диалекта внутри территориальных разновидностей речи регионолект обличается двумя основными чертами: границы региональных атласов, в пределах которых и функционирует регионолект, далеко не всегда совпадают с границами старых диалектов Галлии, а иногда даже значительно расходятся; по своему содержанию он представляет собой некую переходную, полудиалектную форму,

поскольку микроатласы, даже будучи ориентированы на диалектный материал, отчетливо проявляют на своих картах свойственный современной эпохе процесс "растворения" диалектов в общенародной речи.

Опыт работы с материалами микроатласов в указанной методике выявляет большое количество случаев, когда количественная норма выявляется без специальных статистических критериев, так как сама ситуация на картах отчетливо фиксирует свойственную современному сельскому просторечию унификацию форм: например, если один вариант из вариантного ряда по удельному весу (по отношению ко всей опорной сети, принятой за 100%) занимает более 50% опорных пунктов региона, то можно с уверенностью назвать его региональным вариантом. Есть также случаи, когда два варианта примерно поровну делят между собой весь регион, что, при наличии единичных случаев употребления других вариантов, не мешает двум первым выйти на региональный уровень. Так, в атласе Франш-Конте (карта 191) (ALFC, 1975) определенный артикль мн. числа les представлен тремя вариантами lé (le) занимает 46 пунктов сети, la - 42 пункта, li - 2 пункта. Очевидно, что регионалект Франш-Конте характеризуется конкуренцией общефранцузской формы lé и диалектной по происхождению формой la.

Наряду с такими не вызывающими сомнения случаями (отметим еще раз, что их большинство, так что наш метод анализа представляет собой не некую умозрительную, навязанную сверху процедуру, а лишь выявляет объективно существующую в сельском просторечии тенденцию к унификации единиц плана выражения), встречаются ситуации, когда только статистические критерии могут указать на присутствие или отсутствие ведущего, регионального варианта. Для его выявления мы предлагаем использовать два статистических показателя, дополняющих друг друга в анализе.

Первый известен как критерий "хи - квадрат" Пирсона (Карасев, 1970):

$$\chi^2 = \sum_{i=1}^{n} \frac{(m_i - n \cdot \rho_i)^2}{n \cdot \rho_i}$$

где $m_1, m_2, \ldots m_n$ - количество употреблений каждого факультативного варианта в регионе (величина m_i практически должна быть больше или равна 5);

ρ_i - вероятность встречаемости i -го варианта;

n - длина сети атласа.

В зависимости от числа степеней свободы (при 5% уровне значимости) находим границу К критической области. В случае, если величина χ^2 оказывалась больше, чем К, это означало, что можно выделить варианты в качестве региональных, но еще не указывало на сам региональный вариант в ряду. Сделать выбор среди вариантов, т.е. назвать ведущий, помогает второй статистический критерий Z (Бектаев, Пиотровский, 1974, с. 219-222), устанавливающий существенность или несущественность расхождения двух самых употребительных вариантов:

$$Z = \frac{f_1 - f_2}{\sqrt{\frac{f_1(1-f_1)}{N_1} \quad \frac{f_2(1-f_2)}{N_2}}}$$

где $N_1 = N_2$ – общее число пунктов опорной сети атласа;

f_1 – относительная частота употребления первого из сравниваемых вариантов;

f_2 – относительная частота употребления второго варианта. Если для двух наиболее употребительных вариантов величина оказывалась больше значения нижней границы критической области (Z_p = 1,96 при 5 % уровне значимости), то в качестве регионального выбирался первый вариант. В противном случае, т.е. при Z < 1,95, на уровень регионолекта выходили оба варианта. При необходимости такая процедура повторялась со вторым и третьим по употребительности вариантом. Очевидно, что если показания по первому критерию говорят об отсутствии ведущих вариантов в регионе, т.е. χ^2 оказывается меньше величины критической области К, то нет необходимости применять второй критерий и приходится констатировать, что регионолект представлен в этом случае вариантным рядом (унификации плана выражения нет).

Приведем пример, иллюстрирующий применение обоих статистических критериев. В атласе Оверни и Лимузена (ALAL, 1975) вариантность неопределенно-личной конструкции On sue sans rien faire (карта II) по способам заполнения позиции подлежащего представлена общефранцузской структурой с on (39 пунктов), личной структурой с je (21 пункт) и рядом других, менее употребительных вариантов (соответственно 10, 2 и 1 употреблений). Обращает на себя внимание тот факт, что для сравнительно большой части региона неопределенно-личный способ не характерен, там предпочтение отдается личному способу вы-

сказывания, в чем можно видеть конкретность, тематическую ясность, тесную связь с ситуацией устной речи, в целом. Это обстоятельство чрезвычайно интересно в плане соотношения разных способов высказывания, их распределенности в речи сельских жителей, однако, это отдельная тема и здесь она подробно не раскрывается.

Для того, чтобы узнать, есть ли в регионе ведущие варианты, применяем первый критерий. Находим величину χ^2 учитывая, что n = 70 (сумма употреблений трех первых вариантов 39 + 21 + 10), а $n \cdot p_i$ = 23,3:

$$\chi^2 = \frac{(39-23,3)^2}{23,3} + \frac{(21-23,3)^2}{23,3} + \frac{(10-23,3)^2}{23,3} =$$

$$= 10,58 + 0,23 + 7,59 = 18,4.$$

Граница критической области К при числе степеней свободы = 2, будет равна 5,991. Величина χ^2 = 18,4 оказывается, как видим, значительно больше критической области. Это значит, что среди вариантов в регионе есть такие, которые могут выйти на уровень региональной нормы.

Применение второго критерия при сопоставлении двух самых употребительных вариантов региона (39 и 21) дает значение

$$Z = \frac{0,56 - 0,3}{\sqrt{\frac{0,56 \cdot 0,44}{70} + \frac{0,3 \cdot 0,7}{70}}} = \frac{0,26}{\sqrt{\frac{0,46}{70}}} = \frac{0,26}{\sqrt{0,01}} = 2,6$$

Эта величина больше порогового значения 1,96, значит, расхождение вариантов носит существенный характер, и на региональный уровень выходит только первый вариант – неопределенно-личная структура с on .

Таким образом, речь отражает субъективно существующие унифицирующие тенденции, своего рода естественный отбор способов реализации языковых единиц, а применение количественных, в том числе и статистических критериев позволяет исследователю уловить эту динамику в синхронии и повысить точность отбора. Определение ведущих вариантов представляет первый и необходимый этап работы с материалами региональных атласов. В результате региональные варианты становятся релевантными признаками особой разновидности устной речи – регионолекта, и мы получаем возможность на основе общих принципов подхода к лингвогеографическому материалу проводить разнообразные сопоставительные исследования на обширной территории страны.

ЛИТЕРАТУРА

Бектаев К.Б., Пиотровский Р.Г. Математические методы в языкознании, т. II. Алма-Ата, 1974.

Бородина М.А. Проблемы лингвистической географии. М.-Л., 1966.

Граудина Л.К. Норма и статистика. - В кн.: Актуальные проблемы культуры речи. М., 1970.

Зубова Т.Е. Типологический анализ микроатласов Франции (на материале неопределенно-личной конструкции). - В кн.: Взаимодействие лингвистических ареалов. Л., 1980, с. 230-237.

Карасев А.П. Теория вероятностей и математическая статистика. М., 1970.

Скрелина Л.М. Времена французского глагола в системе и функциональных диалектах. - В кн.: Проблемы ареальных контактов и социолингвистики. Л., 1978.

ALAL (Atlas linguistique et ethnographique de l'Auvergne et du Limousin. Paris, 1975, v. I).

ALFC (Atlas linguistique et ethnographique de la Franche-Comté. Paris, 1975, v. I.

Escoffier S. Oppositions morphologiques aux confins des trois langues gallo-romanes. - Revue de linguistique romane. Paris, 1974.

Gardette P. Rencontre de synonymes et pénétration du français dans les aires marginales. - Revue de linguistique romane. Paris, 1970, t. 34, NN 135-136.

Guiraud P. Ancien français. Paris, 1962.

Martin J.-B. L'article défini en franço-provençal central. - Travaux de linguistique et de littérature. Strasbourg, 1972, X, No 1; Le pronom personnel sujet en francoprovençal central. - Revue de linguistique romane. Paris, 1974.

Séguy J. La fonction minimale du dialecte. - Les dialectes romanes de France. Paris, 1972.

Taverdet G. Géographie linguistique de Bourgogne. Dijon, 1970.

STATISTICAL METHODS OF DEFINING REGIOLECTS FROM LINGUISTIC ATLASES

T. Zubova and A. Zubov

S u m m a r y

Areal-typological investigations on the basis of linguistic atlases with a fine base network are only possible under the condition that the divergences of atlases in territory, network length and unit variance range have been removed. Such a procedure should result in revealing dynamics in statics, i. e. certain basic trends on the realisation level. The application of statistical criteria (Pearson's criterion and the criterion of variant divergency appreciability) is put forward in the present paper, for determining normal variations for each region (atlas), which represent a certain dialectological object-regiolect.

The investigation is carried out on the material of regional atlases of France, representing three linguistic areals - French, Provençal, and French-Provençal.

ОПЫТ ОЦЕНКИ ТЕСНОТЫ ФОЛЬКЛОРНОЙ СВЯЗИ ПРИБАЛТИЙСКО-ФИНСКИХ НАРОДОВ (НА МАТЕРИАЛЕ ПОСЛОВИЦ)

А. Крикманн

I. Источники

Исходные данные для нижеследующих наблюдений взяты из рукописи сборника наиболее распространенных пословиц прибалтийско-финских народов. Это издание подготовлено группой финских и эстонских фольклористов под руководством Матти Кууси, prof. emer. Хельсинкского университета. Сборник выйдет в серии "Folklore Fellows' Communications". Непосредственными источниками пословиц для данного издания служили:

финских - фольклорный рукописный архив Общества финской литературы, а также все наиболее существенные печатные источники;

карельских - книга Л.Миеттинен и П.Лейно "Karjalaisia sananpolvia" (Miettinen L., Leino P., 1971), а также некоторые другие печатные и рукописные источники;

эстонских - рукопись полного научного издания эстонских пословиц, первый том которого вышел в 1980 г. (Eesti vanasõnad I, 1980);

водских - научное издание водских пословиц (Mälk V., 1977);

вепсских - все имеющиеся рукописные тексты, основную часть которых составляют собрания эстонских лингвистов и фольклористов (М.Иоалайд, Т.-Р.Вийтсо и др.), а также имеющиеся печатные источники;

ливских - научное издание ливских пословиц в 2-х томах (Mälk V., 1981).

Принцип включения материала в прибалтийско-финский сборник следующий.

В зависимости от общего количества записанных у того или другого народа пословиц зафиксирован некий эмпирический критерий (т.е. число вариантов), например, для финских - 120 вар., карельских - 20, эстонских - 60 и т.д. Любая пословица, которая у одного или нескольких народов превышает этот "порог популярности", включается в сборник, причем тогда приводится

касающийся ее материал (текстовые примеры, статистические данные) не только тех прибалтийско-финских народов, у которых она выступает в числе "фаворитов", но и всех остальных, у которых она представлена хотя бы одним единственным аутентичным вариантом. К каждой включенной в сборник пословице прилагаются также русское, латышское, скандинавское и немецкое соответствия (по тексту от каждого), если таковые имеются и найдены.

Из сказанного почти автоматически вытекает, что находящийся в сборнике прибалтийско-финский материал не отражает без искажений действительные соотношения ни в самой "генеральной совокупности" (т.е. фольклорной традиции), даже ни в имеющейся информации (текстов, данных) об этой совокупности. С другой стороны, лад введенной принципом выбора деформации довольно прост - по сути дела, совершен сдвиг в сторону увеличения стереотипичности материала как в разрезе отдельных народов, так и относительно сводных параметров совокупности в целом, но не тотальное разрушение всех статистических структур. Поэтому представляется допустимым применить настоящий материал - множество наиболее распространенных прибалтийско-финских пословиц - хотя бы для показа принципиальной возможности получения разумно толкуемых результатов в интересующем нас вопросе. К тому же, если отношения предпочтения/избежания окажутся показуемыми на данном материале, вопреки его высокой степени стереотипичности, то можно было бы ожидать еще четче и резче выраженной картины корреляций в случае, если будет рассмотрен весь имеющийся прибалтийско-финский паремический материал в целом.

2. Обозначения; исходные данные

Первичную информацию для наших наблюдений можно представить в виде таблицы (см. табл. I), где ряды означают отдель-

Таблица I

K \ T	T_1	T_2	T_3	T_4	T_5	T_6	...	T_{895}	T_{896}	T_{897}	T_{898}	T_{899}	T_{900}
K_1	30	200	I90	20	I20	80		I6	8	65	I6		I60
K_2	3	3	I3		4	3		4	27	25	27	2I	I
K_3	95	3	4	II5	3	I		5			I6		
K_4	4							6	I	3			
K_5	I	I				6		I			3		
K_6				2		I			I				I

ные народы (K-единицы), столбцы - отдельные пословицы (T-единицы), числа в клетках таблицы - множества записей вариантов (V-единицы). (Для обозначения некоторых параметров мы в дальнейшем пользуемся и буквами R, r в том же смысле, как и в наших предыдущих работах (Krikmann A., 1979 и 1980), хотя здесь, в отличие от этих работ, никакие вычисления не основываются на измерении r-параметров.)

Применяются следующие обозначения:

- $\sum T$ - число T-единиц (т.е. отдельных пословиц) в таблице (= 900);
- $\sum K$ - число K-единиц (т.е. отдельных народов) в таблице (=6);
- $\sum R$ - число клеток с ненулевым показом в таблице (= 2915);

} суммарные показатели

- t_{K_i} - число пословиц, записанных у народа K_i;
- k_{T_i} - число народов, представленных в пословице T_i;
- v_{K_i} - общее число аутентичных пословичных текстов у народа K_i;

} векторные показатели

- $t_{\cap K_i K_j}$ - число пословиц, общих для народов K_i и K_j;
- $v_{\cap K_i(K_j)}$, $v_{\cap K_j(K_i)}$ числа записей, которыми народы K_i и K_j представлены в их общей части $t_{\cap K_i K_j}$;
- $\sum t_{\cap K_i}$ - сумма всех $t_\cap$-показателей народа K_i по всем его общим частям с остальными K попарно;
- $\sum v_{\cap K_i}$ - сумма всех $v_\cap$-показателей народа K_i по всем его общим частям с остальными K попарно;

} характеристики общих частей

- ФИН (=K_1) - финский;
- КАР (=K_2) - карельский;
- ЭСТ (=K_3) - эстонский;
- ВОД (=K_4) - водский;
- ВЕП (=K_5) - вепсский;
- ЛИВ (=K_6) - ливский.

} сокращения названий K

Данные о распределении суммарных показателей t_{K_i} и v_{K_i}

приведены в таблице 2, общее распределение рангов k_T - в таблице 3, картина распределения k_T-рангов по отдельным K - в таблице 4.

Таблица 2

K_i	t_{K_i}	r_{K_i}
ФИН	768	83067
КАР	700	10044
ЭСТ	686	46618
ВОД	267	1150
БЕП	265	1591
ЛИВ	229	896
Σ	2915 (ΣR)	143361 (ΣV)

Таблица 3

k_T	$p(k_T)$	$k_T \cdot p(k_T)$
1	82	82
2	202	404
3	259	777
4	185	740
5	120	600
6	52	312
Σ	900 (ΣT)	2915 (ΣR)

Таблица 4

K_i \ k_T	1	2	3	4	5	6	$\Sigma = t_{K_i}$
ФИН	12	168	244	175	117	52	768
КАР	19	130	212	170	117	52	700
ЭСТ	30	72	232	180	120	52	686
ВОД	2	6	24	85	98	52	267
БЕП	18	17	35	59	84	52	265
ЛИВ	1	11	30	71	64	52	229
$\Sigma = k_T \cdot p(k_T)$	82	404	777	740	600	312	2915 (ΣR)

Мы не будем следить здесь более подробно за характером распределений этих общих параметров, поскольку это не входит в нашу проблематику. Обращаем внимание только на два характерных обстоятельства:

1) по величине t_{K_i} материалы рассматриваемых народов распределяются на две четкие группы, т.е. на "большие" и "малые", а величины $t_{\cap K_i K_j}$ образуют соответственно три довольно четких "кластера": а) "большой с большим", б) "большой с малым" и в) "малый с малым";

2) чем больше t_{K_i}, тем ниже соответствующий ему средний

ранг k_r, т.е. тем ниже степень стереотипичности материала народа K_i.

3. Наблюдения

Ниже описываются две попытки определения оценок тесноты "паремической связи" прибалтийско-финских народов, а затем оценивается теснота связи отдельных прибалтийско-финских материалов с русским и латышским. Применяемый вычислительный аппарат прост, главными элементами в нем являются:

1) нахождение линейных трендов (линий регрессии) корреляционных полей;

2) процедуры стандартизации данных по средним и отклонениям;

3) вычисление т.н. коэффициентов коллигации (λ-коэффициентов).

Наблюдение I.

Оценка тесноты связи основывается на мощностях v-множеств интересующих нас K. Числовые данные (исходные и результирующие) приведены в таблице 5.

Вариант А.

Ход вычислений был следующий.

I. Мы положили $x_o = \Sigma v_{\cap K_i} \cdot \Sigma v_{\cap K_j}$; $y_o = v_{\cap K_i(K_j)} \cdot v_{\cap K_j(K_i)}$ и составили график корреляционного поля. Связь между x_o и y_o оказалась явно криволинейной.

2. Перешли к логарифмической шкале, положив $x = \ln(\Sigma v_{\cap K_i} \cdot \Sigma v_{\cap K_j})$; $y = \ln(v_{\cap K_i(K_j)} \cdot v_{\cap K_j(K_i)})$. По визуальной оценке, вид корреляционного поля можно было считать линейным (см. рис. I).

3. Установили трендовую норму по обычной технике, т.е. способом наименьших квадратов. Оказалось: $\mathrm{norm}(y) = \text{I.26II}x - \text{8.548I}$.

4. Вычислили отклонения $\Delta_o y = y : \mathrm{norm}(y)$ для каждой конкретной пары $K_i K_j$.

5. Поскольку отклонения действительных y от линейной нормы не проявили явной дополнительной корреляции по отношению к x, решили отказаться от их вторичного нормирования.

6. Оказалось, что оценки Δ_o еще не являются пригодными в качестве окончательных оценок, поскольку характеристики $\Sigma\Delta_{oK}$ по каждому K значительно различались друг от друга.

Таблица 5

K_i	K_j	$v_{\Pi K_i(K_j)}$	$v_{\Pi K_j(K_i)}$	Вариант А					Вариант Б		Вариант В	
				x	y	norm (y)	$\Delta_0 y$	$\Delta^* y$	$\Delta_0 y$	$\Delta^* y$	x	$\Delta^* y$
ФИН	КАР	76798	9050	22.365	20.359	19.657	1.036	1.031	0.702	0.574	20.542	1.031
ФИН	ЭСТ	62000	40691	23.965	21.649	21.675	0.999	1.009	-0.026	0.136	22.077	1.008
ФИН	ВОД	23251	1065	20.565	17.027	17.387	0.979	0.979	-0.360	-0.312	18.375	0.979
ФИН	ВЕП	19018	1214	20.739	16.955	17.607	0.963	0.984	-0.652	-0.269	18.700	0.984
ФИН	ЛИВ	17578	782	20.218	16.436	16.950	0.970	0.995	-0.514	-0.130	18.125	0.996
КАР	ЭСТ	7391	31543	21.931	19.267	19.110	1.008	0.987	0.157	-0.235	19.964	0.988
КАР	ВОД	3773	1022	18.531	15.165	14.822	1.023	0.990	0.343	-0.162	16.262	0.991
КАР	ВЕП	3510	1317	18.705	15.346	15.042	1.020	1.011	0.304	0.133	16.587	1.010
КАР	ЛИВ	2272	677	18.184	14.246	14.385	0.990	0.984	-0.139	-0.307	16.013	0.982
ЭСТ	ВОД	21359	1079	20.131	16.953	16.840	1.007	0.991	0.113	-0.100	17.797	0.990
ЭСТ	ВЕП	14569	1242	20.305	16.711	17.059	0.980	0.986	-0.348	-0.229	18.122	0.987
ЭСТ	ЛИВ	20567	875	19.784	16.706	16.402	1.019	1.027	0.304	0.428	17.548	1.029
ВОД	ВЕП	634	878	16.905	13.230	12.772	1.036	1.033	0.458	0.465	14.420	1.035
ВОД	ЛИВ	497	406	16.384	12.215	12.114	1.008	1.008	0.101	0.109	13.845	1.008
ВЕП	ЛИВ	462	316	16.558	11.891	12.334	0.964	0.984	-0.443	-0.101	14.170	0.983

K_i	$\Sigma v_{\Pi K_i}$
ФИН	198685
КАР	25996
ЭСТ	128729
ВОД	4297
ВЕП	5113
ЛИВ	3056

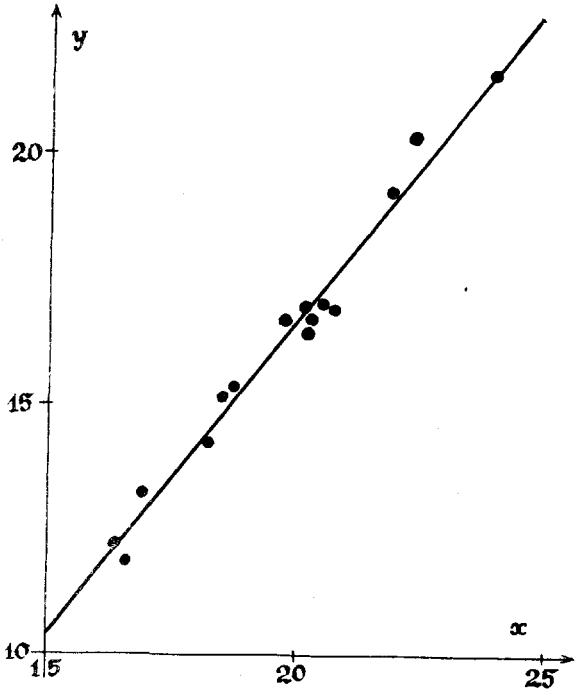


Рис. I

7. Чтобы Δ-характеристики всех K_iK_j были объективно сравнимыми между собой, провели стандартизацию данных по следующей схеме. Для каждой пары K_iK_j были вычислены оценки следующей степени

$$\Delta'_{K_iK_j} = \Delta_{oK_iK_j} - \frac{\bar{\Delta}_{oK_i} + \bar{\Delta}_{oK_j}}{2}$$

($\bar{\Delta}_{oK_i}$ и $\bar{\Delta}_{oK_j}$ означают средние по K_i и K_j), затем вновь

$$\Delta''_{K_iK_j} = \Delta'_{K_iK_j} - \frac{\bar{\Delta}'_{K_i} + \bar{\Delta}'_{K_j}}{2}$$

и т.д., до тех пор, пока $\bar{\Delta}$ всех K можно было считать равными (и тем самым - нулевыми) на желаемой степени точности.

8. Полученные оценки $\Delta^*_{K_iK_j}$ (см. табл. 5) считали окончательными.

Вариант Б.

$\Delta.y = y - norm\,(y)$, все остальное, как в варианте А. Результирующие Δ^* см. в таблице 5.

Вариант В.

$x = \ln (v_{K_i} \cdot v_{K_j})$ (значения v_K см. в табл. 2) соответственно, трендовая норма $norm\,(y) = 1.1647x - 4.0063$; все остальное, как в варианте А. Полученные Δ^*-оценки см. в таблице 5.

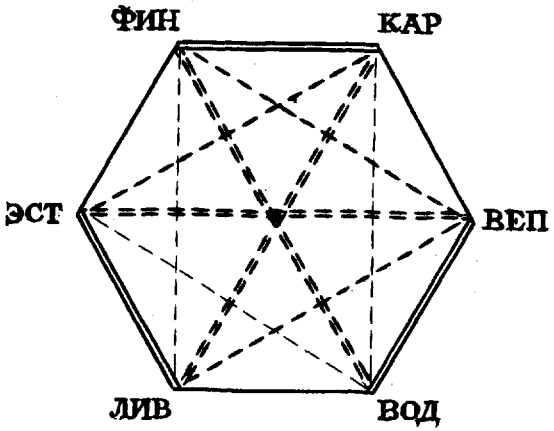
Результаты

1. Связь между K оказывается показуемой и достаточно регулярной.

2. Картина соотношений между K довольно однородна во всех вариантах наблюдения.

3. В качестве графической модели, описывающей степень предпочтения/избежания между K, удобно применить шестиугольник (рис. 2), вершины которого представляют отдельные K в порядке следования

... ФИН ⇄ КАР ⇄ ВЕП ⇄ ВОД ⇄ ЛИВ ⇄ ЭСТ ⇄ ФИН ... .



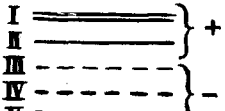


Рис. 2

В этом случае наши числовые результаты почти идеально соответствуют конфигурации, где графы наиболее сильных позитивных связей (I) и (II) соответствуют сторонам шестиугольника, графы же наиболее сильных негативных связей (У) - ребрам между диаметральными K, а графы промежуточных связей (III) и (IУ) - остальным ребрам.

4. Единственное отклонение эмпирического результата от этого идеала касается отношения ЭСТ - ВЕП, которое выделяется чрезмерно малым с точки зрения идеала негативным показателем. Этот факт, видимо, не является отражением каких-либо историко-географических обстоятельств, а скорее всего вытекает из текстологической специфики вепсского материала (очень поздние даты собраний, малое количество собирателей, сравнительно много повторов одних и тех же единиц, записанных из уст разных информантов). Следствием этих факторов можно, на наш взгляд, считать и некоторую девиацию в k_T-распределении вепсского материала (см. табл. 4).

5. После "насильственного" перемещения отношения ЭСТ - ВЕП на его "теоретическое" место картина связей будет выглядеть примерно следующим образом:

ранг тесноты	пары
I	ФИН - КАР ЭСТ - ЛИВ ВОД - ВЕП
II	ФИН - ЭСТ КАР - ВЕП ЛИВ - ВОД
III	ФИН - ЛИВ ВОД ЭСТ - КАР
IУ	ФИН ЛИВ ВЕП ЭСТ - КАР
У	ФИН - ВОД КАР - ЛИВ ЭСТ - ВЕП

Наблюдение 2.

Оценка тесноты связи основывается на мощностях t-множеств. Исходные и результирующие числовые данные приведены в таблице 6.

Таблица 6

K_i K_j	$t_{\cap K_i K_j}$	x	y	norm (y)	$\Delta_0 y$	$\Delta' y$	$\Delta^* y$	$\Delta^{**} y$
ФИН КАР	589	13.001	12.757	12.692	0.065	1.518	2.131	2.155
ФИН ЭСТ	555	12.960	12.638	12.612	0.026	0.565	0.141	0.152
ФИН ВОД	194	11.918	10.536	10.588	-0.052	-0.341	-0.849	-1.961
ФИН ВЕП	158	11.830	10.125	10.417	-0.292	-1.776	-1.266	-0.541
ФИН ЛИВ	153	11.727	10.061	10.217	-0.156	-0.872	-0.158	0.193
КАР ЭСТ	470	12.848	12.305	12.395	-0.090	-1.637	-1.918	-1.671
КАР ВОД	179	11.805	10.375	10.369	0.006	0.036	-0.329	0.140
КАР ВЕП	168	11.717	10.248	10.198	0.050	0.277	0.932	0.198
КАР ЛИВ	126	11.615	9.673	10.000	-0.327	-1.672	-0.817	-0.829
ЭСТ ВОД	197	11.765	10.566	10.291	0.275	1.585	0.185	0.046
ЭСТ ВЕП	162	11.677	10.175	10.120	0.055	0.295	-0.087	-0.828
ЭСТ ЛИВ	172	11.574	10.295	9.920	0.375	1.858	1.681	2.305
ВОД ВЕП	77	10.634	8.688	8.094	0.594	1.528	1.062	2.305
ВОД ЛИВ	54	10.532	7.978	7.896	0.082	0.198	-0.067	-0.538
ВЕП ЛИВ	35	10.443	7.111	7.723	-0.612	-1.395	-0.641	-1.132

K_i	t_{K_i}
ФИН	704
КАР	629
ЭСТ	604
ВОД	213
ВЕП	195
ЛИВ	176
Σ	2521

Вычислительная процедура в значительной степени аналогична примененной в предыдущем наблюдении.

I. Из наблюдаемого материала была удалена группа пословиц с признаком $k_T = 6$ (52 по количеству). (Ради удобства в описании настоящего наблюдения параметры (t_{K_i} - 52) и ($t_{\cap K_i K_j}$ - 52) все же обозначены как t_{K_i} и $t_{\cap K_i K_j}$.)

2. Было положено $x_o = t_{K_i} t_{K_j}$; $y_o = t^2_{\cap K_i K_j}$. Пункты соответствий $x_o(y_o)$ были нанесены на график. Как можно было ожидать, связь оказалась криволинейной.

3. С переходом к логарифмической шкале $x = \ln(t_{K_i} t_{K_j})$; $y = 2 \ln t_{\cap K_i K_j}$ корреляционное поле приобрело линейный вид (см. рис. 3).

4. Была вычислена трендовая норма для y, т.е. найдены параметры линейной регрессии. Оказалось: $\mathrm{norm}(y) = 1.9427x - 12.5649$.

5. Были вычислены отклонения $\Delta_o y = y - \mathrm{norm}(y)$ для каждой конкретной пары $K_i K_j$.

6. Визуальный осмотр корреляционного отношения $x' = x$; $y' = \Delta_o y$ (см. рис. 4) показал сильную дополнительную корреляцию y' по отношению к x', т.е. значения $\Delta_o y$, соответствующие большим и малым значениям x, не оказались объективно сравнимыми в их "сыром" виде.

7. Поэтому было проведено дополнительное нормирование Δ_o. Исходя из общего вида корреляционного поля и средних значений $\bar{x}(|\Delta_o y|)$ по кластерам "малый с малым", "малый с большим" и "большой с большим" для нормирования была применена функция типа

$$\mathrm{norm}(|\Delta_o' y|) = \frac{a}{e^{\frac{x}{b}} + b} - c$$

(a, b, c - постоянные с эмпирическими значениями $a = 101.75$, $b = 0.123$, $c = 0.110$).

8. Далее опять были вычислены новые варианты оценок

$$\Delta' y = \Delta_o y : \mathrm{norm}(|\Delta_o y|).$$

Возникшее корреляционное поле можно было считать линейным (см. рис. 5).

9. Суммы $\sum \Delta'_{K_i}$ полученных оценок (а тем самым и их средние) и здесь оказались различающимися друг от друга. Поэтому была проведена стандартизация Δ'-оценок с применением точно такой же техники, как в пункте (7) предыдущего наблюдения.

10. В результате стандартизации у нас опять получились Δ^*-оценки, при которых $\bar{\Delta}^*_{K_1} \approx \bar{\Delta}^*_{K_2} \approx \ldots \approx \bar{\Delta}^*_{K_6} \approx 0$. Однако полученная картина Δ^*-оценок не показалась нам достаточно четко интерпретируемой, поэтому процесс стандартизации продолжили следующим путем.

10.1. Δ^*-оценки мы стандартизировали и по средним их абсолютных величин $|\Delta^*_{K_i}|$, вычислив оценки следующей степени

Рис. 3

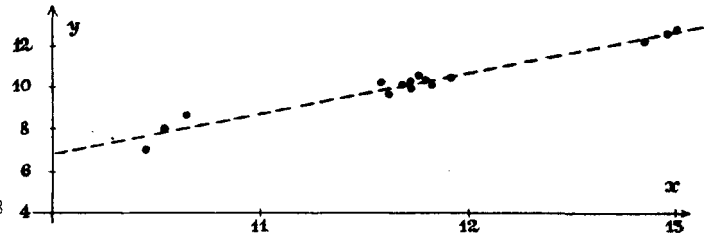


Рис. 4

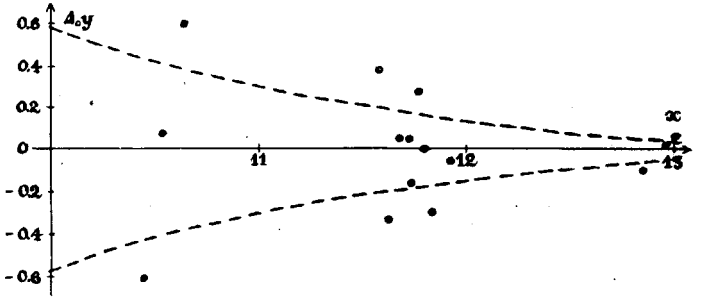


Рис. 5

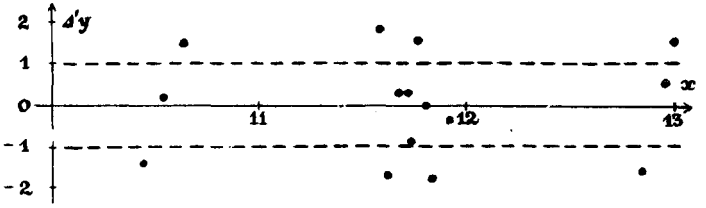
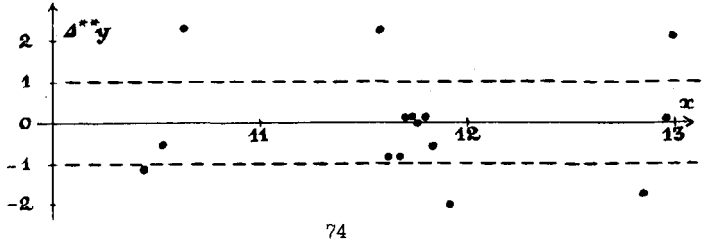


Рис. 6



$$\Delta^{*\prime}_{K_i K_j} = \Delta^{*}_{K_i K_j} : \frac{|\bar{\Delta}^{*}_{K_i}| + |\bar{\Delta}^{*}_{K_j}|}{2}$$

и продолжали итерации, пока средние значения Δ-оценок всех K не уравнились до желаемой степени точности.

10.2. Стандартизация Δ^{*} по средним абсолютных величин уничтожила равенство их средних действительных значений. Для восстановления этого равенства была вновь повторена процедура, указанная в пункте (9) выше, затем опять стандартизация по абсолютным величинам и т.д. В итоге таких поочередных итераций Δ- и $|\Delta|$-характеристики наших $K_i K_j$ стали постепенно приближаться друг к другу. В момент, когда вся система была максимально близка к идеалу

$$\begin{cases} \bar{\Delta}^{**}_{K_1} = \bar{\Delta}^{**}_{K_2} = \ldots = \bar{\Delta}^{**}_{K_6} = 0; \\ |\bar{\Delta}^{**}_{K_1}| = |\bar{\Delta}^{**}_{K_2}| = \ldots = |\bar{\Delta}_{K_6}| = 1, \end{cases}$$

процесс был прерван, и имеющиеся в тот момент Δ^{**}-оценки пар $K_i K_j$ считались окончательными. (Графический облик поля этих окончательных Δ^{**}-оценок показан на рис. 6.)

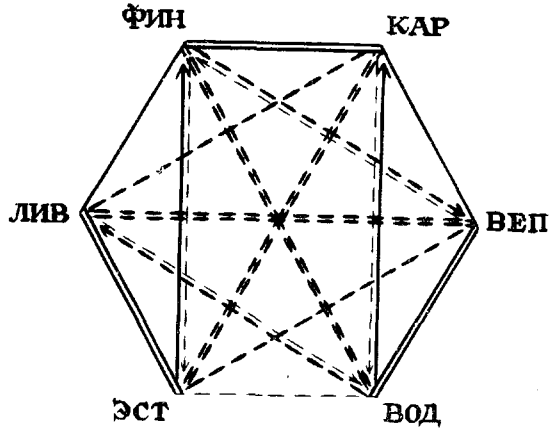
Результаты

Для более удобного сравнения результатов обоих наблюдений в качестве опорной модели мы опять применяем шестиугольную фигуру с расстановкой K-вершин по принципу "наименьшего сопротивления".

Сопоставление конечных оценок в наблюдениях (I) и (2) сразу же выявляет значительные различия.

I. Для представления результатов последнего наблюдения на шестиугольнике лучше подходит такая конфигурация K, где позиции ЭСТ и ЛИВ переменены (см. рис. 7; ср. рис. 2).

2. Связь II ранга не является регулярно двусторонней, как это было в наблюдении (I). Взаимность наблюдается только в отношениях ФИН - ЛИВ и КАР - ВЕП, зато ЭСТ и ВОД не связываются между собой, как можно было ожидать, а находят партнеров по вертикальным ребрам (ЭСТ → ФИН; ВОД → КАР). Такая картина связей как бы маркирует своеобразную "ось симметрии", проходящую между ЭСТ и ВОД. Интересно отметить, что в ряду I5 убывающих/возрастающих Δ^{**}-оценок оценка пары ЭСТ - ВОД занимает 8-е место и по значению она наиболее близка к нулю.



I
II
III
IV
V

Рис. 7

3. Связь У ранга устанавливается, как и в предыдущем наблюдении, между диаметральными K, но из прежних пар У ранга сохранена только ФИН - ВОД, остальные же (КАР - ЭСТ и ВЕП - - ЛИВ) соответствуют изменившейся расстановке K.

С другой стороны, результаты наблюдений (I) и (2) во многом сходны. Наряду с рассмотренными нами было испытано еще несколько иных методик вычисления. Во всех наблюдениях самой тесной оказалась связь между **ФИН и КАР; ЭСТ и ЛИВ; ВОД и ВЕП.**

Хорошие, как правило, оценки выпали также парам ФИН - ЭСТ, ФИН - ЛИВ, КАР - ВЕП, КАР - ВОД, а также ЛИВ - ВОД.

Наиболее сильная отрицательная связь наблюдалась между **ФИН и ВОД; (ЭСТ, ЛИВ) и (КАР, ВЕП).** Притом ФИН и ВОД с абсолютной последовательностью выступали как партнеры У ранга, их же числовая оценка связи оказалась самой низкой в обоих наших наблюдениях.

В итоге можно сказать, что полученная нами картина тесноt связей довольно хорошо соответствует традиционным лингво-

и этногеографическим представлениям.

Наблюдение 3

Как было отмечено, к изданию прибалтийско-финских пословиц прилагаются и параллели соседних неприбалтийско-финских народов - русские, латышские, скандинавские (в первую очередь шведские) и немецкие. Таким образом, издание предлагает некоторые возможности и для оценки тесноты фольклорных связей, пересекающих границы прибалтийско-финского ареала. Эти возможности, однако, нельзя считать слишком благоприятными, так как по сей день отсутствуют достоверные сводные характеристики паремических фондов большинства наших неприбалтийско-финских соседей, включая самые общие оценки мощности их как "генеральных совокупностей". А у некоторых из этих народов в силу специфики их источников и состояния архивного дела этот пробел может оказаться в принципе невосполнимым. Можно с уверенностью предсказать, что сборник пословиц-фаворитов прибалтийско-финских народов внесет заметный вклад в улучшение предпосылок для дальнейшего исследования северо-европейских фольклорных связей. Более углубленное изучение паремиологических отношений прибалтийско-финских народов с другими станет, по-видимому, возможным лишь после создания полного сопоставительного свода прибалтийско-финских пословиц, да и тогда в основном только "глядя с прибалтийско-финской стороны".

К настоящему моменту в нашем распоряжении имелись данные только о русских (сокр. РУС) и латышских (сокр. ЛАТ) соответствиях прибалтийско-финских пословиц-фаворитов. На этих данных было сделано следующее простое наблюдение с целью выяснения относительного веса русских и латышских параллелей в фонде каждого прибалтийско-финского народа.

Наблюдение было проведено в двух вариантах: 1) на основании t_n-параметров и 2) v_n-параметров. Для каждой пары KK' (K' означает РУС или ЛАТ) был вычислен т.н. коэффициент коллигации по формуле

$$\lambda_{AB} = \frac{p(AB)}{p(A) \cdot p(B)}$$

($p(AB)$ - частота совместного наступления событий A и B; $p(A)$ - частота события A; $p(B)$ - частота события B).

Результаты первого круга λ-вычисления были заново трактованы как исходные данные, по ним вычислены λ-коэффициенты

следующего круга и т.д. до того, как средние всех рядов и столбцов таблицы уравнились (и стали приблизительно равными единице) на желаемой степени точности.

Величины исходных $t_{\cap}$- и $v_{\cap}$-множеств показаны в таблице 7 (а, б); окончательные значения Λ-коэффициентов - в таблице 8 (а, б). Как видно, результаты обоих вариантов наблюдения

Таблица 7а

K \ K'	$t_{\cap}$ РУС	$t_{\cap}$ ЛАТ	Σ_K
ФИН	394	319	713
КАР	373	284	657
ЭСТ	387	343	730
ВОД	192	163	355
ВЕП	199	141	340
ЛИВ	177	200	377
$\Sigma_{K'}$	1722	1450	3172

Таблица 7б

K \ K'	$v_{\cap}$ РУС	$v_{\cap}$ ЛАТ	Σ_K
ФИН	35697	26780	62477
КАР	5768	3889	9657
ЭСТ	28743	27611	56354
ВОД	880	747	1627
ВЕП	1299	913	2212
ЛИВ	726	824	1550
$\Sigma_{K'}$	73113	60764	133877

Таблица 8а

K \ K'	Λ_t РУС	Λ_t ЛАТ
ФИН	1.023	0.977
КАР	1.054	0.946
ЭСТ	0.978	1.022
ВОД	1.000	1.000
ВЕП	1.090	0.910
ЛИВ	0.857	1.143

Таблица 8б

K \ K'	Λ_v РУС	Λ_v ЛАТ
ФИН	1.054	0.948
КАР	1.104	0.896
ЭСТ	0.927	1.073
ВОД	0.988	1.012
ВЕП	1.085	0.917
ЛИВ	0.845	1.155

весьма сходны. Учитывая суммы Λ обоих наблюдений, прибалтийско-финские материалы можно упорядочить в следующий ряд по степени их близости к русскому/латышскому:

максимум тесноты связи с РУС

ВЕП — КАР — ФИН — ВОД — ЭСТ — ЛИВ

максимум тесноты связи с ЛАТ

С историко-географической точки зрения такую последовательность можно считать вполне естественной.

ЛИТЕРАТУРА

Eesti vanasõnad I. 1 - 5000. Toimetanud A. Krikmann ja I. Sarv. Koostanud A. Hussar, A. Krikmann, E. Normann, V. Pino, I. Sarv ja R. Saukas.- Tln.: Eesti Raamat,1980.

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

Krikmann A. Towards the typology of Estonian folklore regions. Paper presented to the Fifth International Finno-Ugric Congress (Turku 1980). Preprint KKI-16. Tln., 1980.

Miettinen L., Leino P. (toim.) Karjalaisia sananpolvia. - Helsinki: SKS, 1971.

Mälk V. Vadja vanasõnad eesti, soome, karjala ja vene vastetega. - Tln.: Eesti Raamat, 1977.

Mälk V. (koost.) Liivi vanasõnad eesti, vadja ja läti vastetega I - II. - Tln.: Eesti Raamat, 1981.

AN ATTEMPT OF MEASURING THE DENSITY OF FOLKLORISTIC CORRELATION BETWEEN BALTO-FINNIC PEOPLES (ON PROVERBIAL MATERIALS)

Arvo Krikmann

S u m m a r y

In the article an attempt has been made to measure the density of folkloristic correlation between balto-finnic peoples on the ground of their proverbial materials. Initial data come from the manuscript of the edition of balto-finnic favourite proverbs, completed by a group of Finnish and Estonian paremiologists. As to calculating methods, elements of regression analysis, data standardization procedures etc. have been applied.

ОБ ОДНОЙ ПАРАДИГМЕ ЛИНГВОСТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Ю.К. Крылов

1. В настоящее время имеется большое количество экспериментальных данных, касающихся математической лингвистики, биологии, экономики и других информационных систем, которые показывают наличие определенного подобия и устойчивости структур частот классифицируемых элементов в том или ином классе эмпирических выборок. Естественно стремление понять вероятностный механизм, порождающий наблюдаемые распределения. Однако, как неоднократно отмечалось (Петров Б.Н., Яблонский А.И., 1980, с. 4 и 63; Городецкий Б.Ю., 1977, с.28), эта важная проблема до сих пор не имеет однозначного решения.

Нельзя не согласиться с Араповым М.В. и Шрейдером Ю.А. (1978, с. 75), что основным недостатком проведенных исследований является отсутствие системности рассмотрения, которая позволила бы установить место указанных закономерностей в парадигме известных распределений, например, физических.

Конечно, "термодинамический" подход, основанный на аналогиях между "сложностью" гуманитарных систем и энергией, далеко не всегда является оправданным при рассмотрении "самоорганизующихся" структур. Однако трудно представить, что статистический механизм, ответственный за возникновение распределений типа распределения Ципфа-Мандельброта, находится вообще вне парадигмы статистик теоретической физики. В связи с вышесказанным прежде всего остановимся на общепринятой трактовке этих статистик: Больцмана, Бозе - Эйнштейна, Ферми - Дирака и Линден - Белла (последней подчиняются элементарные объемы фазового пространства звездных систем).

2. Задачу классификации объектов любой природы естественно понимать как задачу о стахастическом распределении N частиц в $\mathcal{L}$ ячейках "фазового пространства", или "пространства классификации". При этом за вероятность $P(\vec{n})$ того или иного распределения принимается величина, пропорциональная числу способов $S(\vec{n})$, которыми данное распределение может

быть реализовано. Здесь вектор $\vec{n}=(n_1,\dots n_\mathcal{L})$ характеризует численности заполнения "частицами" отдельных классов эквивалентности, $\sum_\mathcal{L} n_\mathcal{L}=N$.

В статистике Больцмана частицы принято считать принципиально отличимыми друг от друга, так что перестановка двух частиц, находящихся в разных ячейках, дает новое распределение. Число частиц в одной ячейке не ограничено. При этом вероятность фиксированного распределения дается хорошо известным выражением

$$P(\vec{n})=\frac{N!}{\mathcal{L}^N\prod_\mathcal{L} n_\mathcal{L}!}=\frac{N!}{\mathcal{L}^N n_1!n_2!\dots n_\mathcal{L}!\dots n_\mathcal{L}!} \qquad (1)$$

В статистиках Ферми – Дирака и Линден – Белла полагается, что в каждой ячейке может находиться не более одной частицы. Однако в отличие от статистики Ферми – Дирака, в которой частицы считаются принципиально не различимыми, в статистике Линден – Белла делается обратное предположение (см., например, Агекян, 1974, с. 16). Тем не менее обе статистики приводят к одинаковым вероятностям распределения:

$$P=\frac{1}{C_N^\mathcal{L}}=\frac{\mathcal{L}!(N-\mathcal{L})!}{N!} \qquad (2)$$

Наконец, в статистике, предложенной Бозе и Эйнштейном, частицы также полагаются принципиально не различимыми, однако никаких ограничений на их число в ячейке не накладывается. Соответствующая вероятность распределения описывается формулой:

$$P=\frac{1}{C_{N+\mathcal{L}-1}^N}=\frac{N!(\mathcal{L}-1)!}{(N+\mathcal{L}-1)!} \qquad (3)$$

Таким образом, в традиционном рассмотрении фундаментальное значение приобретают отношения различимости и неразличимости частиц. Однако понятие различимости и неразличимости

частиц самым непосредственным образом связано с различимостью и неразличимостью классов эквивалентности, т.е. ячеек пространства классификации. Отнюдь не случайно, что вероятности распределения в статистиках Ферми – Дирака и Линден – Белла даются одним и тем же выражением (2). Этот факт лишь отражает известное равенство числа различных перестановок $S_1 = \frac{N!}{n_1!\ldots n_i!\ldots n_\mathcal{L}!}$ с повторяющимися элементами числу разбиений множества из N элементов на $\mathcal{L}$ классов эквивалентности $S_2 = \frac{N!}{n_1!\ldots n_\mathcal{L}!}$. Действительно, допустим, что в нашем распоряжении имеется N идеально гладких шаров $\mathcal{L}$ различных окрасок. Непосредственное нанесение меток на шары невозможно. Расположим шары в какой-нибудь последовательности в "фазовом" пространстве, например AABBABCA. Но ведь этим мы уже перенумеровали шары, если ячейки перенумерованы!

Из сказанного вытекает, что при рассмотрении парадигмы возможных статистик имеется две, по-существу, эквивалентные возможности. В первом случае можно ввести математическую структуру в виде множеств классов эквивалентности и классифицируемых элементов, которые находятся в определенных отношениях (понимаемых в теоретико-множественном смысле) различимости и неразличимости, описываемых заданными свойствами. Варьируя эти свойства, можно получать различные статистики. Однако спектр функционалов, описывающих вероятности тех или иных распределений, будет значительно беднее множества пересечений рассматриваемых логических возможностей. Как и в случае статистик Ферми – Дирака и Линден – Белла мы будем весьма часто приходить лишь к различным интерпретациям одинаковых распределений.

3. Однако при рассмотрении парадигмы возможных классификаций существует и другой путь. Будем исходить из постулата о том, что наличие не одного, а двух классифицируемых элементов предполагает существование по крайней мере одного признака, по которому мы можем отличить их друг от друга. (В какой-то мере этот постулат аналогичен принципу Паули, согласно которому в квантово-механической системе не может быть двух микрочастиц, характеризуемых одним и тем же набором квантовых чисел). До тех пор, пока такой признак не определен, рассуждения о наличии двух "экземпляров" тождест-

венных элементов теряют смысл. В процессе анализа качество этого признака может меняться, но обязательно на любой стадии рассмотрения наличие хотя бы одного различия элементов по отношению друг к другу не может быть утеряно.

Разобьем построение парадигмы на два этапа. На первом из них мы лишь постулируем различимость классов эквивалентности, т.е. ячеек пространства классификации. Это утверждение соответствует интуитивным представлениям о том, что классификация возможна лишь в том случае, когда известны правила, по которым она осуществляется. На этой стадии можно ограничиться лишь подсчетом разных способов размещения частиц в "фазовом" пространстве при той или иной "стратегии" его заполнения. В дальнейшем, когда ячейки для частиц будут уже фиксированы, распределениям частиц можно приписать дополнительные статистические веса в зависимости от тех или иных дополнительных условий, заданных на множестве классифицируемых частиц. Последние, однако, должны быть сформулированы в терминах, не зависящих от местоположения частиц в ячейках фазового пространства, т.е. в форме внешних по отношению к данной классификации отношений эквивалентности (см. например, п. 6).

4. До сих пор мы не конкретизировали геометрию пространства классификации. Если рассматривать его как цепочку линейно-упорядоченных ячеек, то в соответствии с введенным постулатом число этих ячеек $\mathcal{L}$, т.е. число элементов множества $\{L_\alpha\}$ классов эквивалентности $\mathcal{L} = Card\{L_\alpha\}$ должно быть не меньше числа элементов классифицируемого множества $\{N\}$, т.е. $\mathcal{L} \geq N$. При этом в нашем рассмотрении попадание двух частиц в одну и ту же ячейку фактически означает их отождествление, другими словами-классификацию одной и той же частицы, т.к. классифицируемые элементы пока нами различаются лишь в соответствии с ячейками, в которые они отображаются.

Условие $\mathcal{L} \geq N$ может быть, однако, снято, если наряду с множеством классов эквивалентности $\{L_\alpha\}$ рассматривать сопряженное или ортогональное множество классов эквивалентности $\{\bar{L}_{\bar{\alpha}}\}$. Во избежание недоразумений следует подчеркнуть, что здесь речь не идет об увеличении полноты описания за счет расширения рассматриваемой классификации введением дополнительных классов эквивалентности. Отношение эквивалентности $Q(\bar{L})$ присуще данной классификации, хотя его природа в рамках этой классификации может быть и не конкретизирована.

Возможность, более того, необходимость его рассмотрения при $\mathcal{L}<N$ связана с тем, что само определение понятия множества предполагает различимость его элементов. (Два множества $\{A, B\}$ и $\{A, A, B\}$ считаются равными, т.е. по определению являются одним и тем же множеством). Таким образом, любое разбиение произвольного множества на классы эквивалентности $\{L_1, L_2, \dots L_z \dots L_{\mathcal{L}}\}$ подразумевает, что каждое подмножество, соответствующее классу эквивалентности L_z (в свою очередь являющееся множеством), в качестве своих элементов содержит не тождественные элементы. Именно наличие сопряженного признака $\bar{Q}(\bar{L})$, $\bar{L}_{\bar{z}} \in \{\bar{L}_1, \dots \bar{L}_{\bar{\mathcal{L}}}\}$ и обуславливает их различие. Соответственно, вместо условия $\mathcal{L} \geq N$ получаем лишь требование $\mathcal{L} \cdot \bar{\mathcal{L}} \geq N$.

В связи с вышесказанным в дальнейшем под классификацией будем понимать отображение множества классифицируемых объектов на фазовое пространство, являющееся декартовым произведением $\{L \times \bar{L}\}$ множеств не одного, а двух сопряженных признаков. В таком представлении разбиение множества ячеек пространства классификации на классы эквивалентности $\{L_z\}$ или $\{\bar{L}_{\bar{z}}\}$ будет соответствовать не отождествлению элементов отдельных подмножеств, а лишь их соотнесению по сходству к разным классам. Ячейки, принадлежащие одному и тому же классу эквивалентности L_z, различаются по сопряженному признаку $\bar{Q}(\bar{L})$, тогда как ячейки, относящиеся к одному и тому же классу $\bar{L}_{\bar{z}}$, не эквивалентны друг другу по признаку $Q(L)$. Условие же сопряженности признаков $Q(L)$ и $\bar{Q}(\bar{L})$ означает, что любое пересечение $L_z \cap \bar{L}_{\bar{z}}$ содержит лишь один элемент – ячейку, которая в результате классификации может быть либо занята одним из классифицируемых объектов, либо остаться пустой. С геометрической точки зрения факторизация пространства классификации по сопряженному признаку соответствует переходу от линейно упорядоченной цепочки классов эквивалентности к матрице, содержащей $\mathcal{L}$ столбцов и $\bar{\mathcal{L}}$ строк.

5. Перейдем к интерпретации имеющихся статистик в рамках вышеописанной схемы. Отметим, что различимость частиц позволяет рассматривать статистику Больцмана в соответствии со схемой независимых испытаний Бернулли. Считая, что первая частица с вероятностью $\frac{1}{\mathcal{L}}$ попадает в любой из столбцов I строки, вторая – в любой из столбцов II – строки и т.д. для безусловной вероятности $P(n_1)$ того, что класс L_1 будет содержать ровно n_1 частиц, пол. им:

$$P(n_1) = C_N^{n_1}\left(\frac{1}{\mathcal{L}}\right)^{n_1}\left(\frac{\mathcal{L}-1}{\mathcal{L}}\right)^{N-n_1} = \frac{N!\,(\mathcal{L}-1)^{N-n_1}}{n_1!(N-n_1)!\,\mathcal{L}^N} \quad (4)$$

Оставшиеся $N-n_1$ частиц с вероятностью $\frac{1}{\mathcal{L}-1}$ могут попасть в любой из столбцов, соответствующих классам L_z $z \in \{2 \ldots \mathcal{L}\}$. Пусть в столбце, соответствующем классу L_2, оказалось n_2 заполненных ячеек. Тогда условная вероятность

$$P(n_2/n_1) = C_{N-n_1}^{n_2}\left(\frac{1}{\mathcal{L}-1}\right)^{n_2}\left(\frac{\mathcal{L}-2}{\mathcal{L}-1}\right)^{N-n_1-n_2}$$

а вероятность пересечения:

$$P(n_1)\cdot P(n_2/n_1) = C_N^{n_1}\cdot C_{N-n_1}^{n_2}\cdot\frac{(\mathcal{L}-2)^{N-n_1-n_2}}{\mathcal{L}^N}$$

и т.д.

Легко видеть, однако, что требование упорядоченности попадания частиц в соответствующие строки матрицы не является существенным. Ячейки для частиц, отображающиеся в I столбец, могут выбираться независимо от порядка их отображения. При этом выбор может быть осуществлен $C_N^{n_1}$ разными способами. В качестве строк, заполняемых частицами второго столбца, могут быть выбраны любые из $N-n_1$ оставшихся свободными. В дальнейшем при переходе к каждому последующему столбцу, выбор ячеек остается произвольным с точностью до условия, что ячейки выбираются из предварительно не занятых строк. Соответственно число разных способов выбора:

$$S(N,\mathcal{L},\vec{n}) = \prod_{z=1}^{\mathcal{L}} C_{N-\sum_{s=1}^{z-1} n_s}^{n_z} = \frac{N!}{n_1!\,n_2!\ldots n_z!\ldots n_{\mathcal{L}}!}$$

При этом в статистике Больцмана подразумевается, что имеет место равенство численностей всех классов сопряженного разбиения ($m_z = 1$), что для числа всех равновозможных случаев дает $(C_{\mathcal{L}}^1)^N = \mathcal{L}^N$ и соответственно приводит к формуле (I).

В физических приложениях статистик Ферми-Дирака и Линден-Белла фазовое пространство разбивается на слои, содержащие фиксированное количество ячеек. При подсчете числа воз-

можных распределений частиц по фазовому пространству формула (2) применяется к каждому отдельному слою. Число ячеек в слоях определяется апостериорно из физических соображений. Соответственно и мы, в рамках рассматриваемой парадигмы, будем использовать (2) применительно к отдельным столбцам матрицы классификации. Последнее равносильно замене принятого в статистике Больцмана ограничения $m_{\bar{z}} \equiv 1$ на $m_{\bar{z}} \leq \mathcal{L}$, что, естественно, снимает условие $\bar{\mathcal{L}} = N$. В результате $\bar{\mathcal{L}}$ – число классов $\bar{\mathcal{L}} = Card\{\bar{L}_{\bar{z}}\}$ сопряженного разбиения исходного классифицируемого множества (строк классификационной матрицы) выступает в роли дополнительного независимого параметра рассматриваемой классификации, аналогично $\mathcal{L}$ – числу классов эквивалентности $\mathcal{L} = Card\{L_z\}$ по основному признаку. Для каждого класса эквивалентности (столбца) выбор n_z заполненных ячеек этого столбца возможен $C_{\bar{\mathcal{L}}}^{n_z}$ независимыми способами. При этом общее число способов выбора занятых ячеек для распределения с заданным вектором $\vec{n}$ дается выражением:

$$S(N, \mathcal{L}, \bar{\mathcal{L}}, \vec{n}) = \prod_{z=1}^{\mathcal{L}} C_{\bar{\mathcal{L}}}^{n_z}$$

Так как вся матрица содержит $\mathcal{L}\cdot\bar{\mathcal{L}}$ ячеек, выбор из них N заполненных может быть осуществлен $C_{\mathcal{L}\bar{\mathcal{L}}}^{N}$ равновозможными способами. Таким образом, для вероятности классификации в случае статистик Ферми-Дирака и Линден-Белла получим

$$P(N, \mathcal{L}, \bar{\mathcal{L}}, \vec{n}) = \frac{\prod_{z=1}^{\mathcal{L}} C_{\bar{\mathcal{L}}}^{n_z}}{C_{\mathcal{L}\bar{\mathcal{L}}}^{N}} = \frac{\bar{\mathcal{L}}!^{\mathcal{L}} (\mathcal{L}\cdot\bar{\mathcal{L}} - N)!\, N!}{(\mathcal{L}\cdot\bar{\mathcal{L}})! \prod_{z=1}^{\mathcal{L}} (\mathcal{L} - n_z)!\, n_z!} \qquad (5)$$

Статистика Бозе-Эйнштейна в рамках рассматриваемой парадигмы может быть интерпретирована как такая стратегия заполнения пространства классификации, при которой выбор ячеек осуществляется не по столбцам, а по строчкам. При этом на возможные способы выбора в пределах отдельной строки накладывается дополнительное ограничение, в известной степени противоположное ограничению на выбор ячеек в статистике Больцмана. Если в последней выбор ячеек для класса L_z разрешался лишь из "еще пустых" строк, то в статистике Бозе-Эйнштейна, наоборот, при заполнении каждой последующей строки выбор

разрешен только из столбцов, заполненных на предыдущем шаге классификации. Таким образом, для первой строки имеем $C_{\mathcal{L}}^{m_1}$ возможностей. Они комбинируют с $C_{m_1}^{m_2}$ независимыми способами выбора ячеек на втором шаге и т.д. Здесь координаты вектора $\vec{m}$, характеризующего численности строк классификационной матрицы, уже считаются упорядоченными по убыванию, т.е. $m_1 \geq m_2 \geq \ldots \geq m_i \geq \ldots \geq m_{\mathcal{L}}$. В результате

$$S(N, \mathcal{L}, \vec{m}) = C_{\mathcal{L}}^{m_1} C_{m_1}^{m_2} \ldots C_{m_{N-1}}^{m_N} = \frac{\mathcal{L}!}{\prod_{i=0}^{N} (\Delta m_i)!}$$

где

$$\Delta m_0 = \mathcal{L} - m_1 \ldots \quad \Delta m_i = m_i - m_{i+1} \ldots \quad \Delta m_N = m_N$$

Как известно (см., например, М. Борн, 1965), общее число равновозможных способов в статистике Бозе-Эйнштейна равно $C_{N+\mathcal{L}-1}^{N}$. Соответственно, для вероятности классификации в этом случае будем иметь:

$$P(N, \mathcal{L}, \vec{m}) = \frac{N!\,(\mathcal{L}-1)!\,\mathcal{L}!}{(N+\mathcal{L}-1)! \prod_{i=0}^{N} (\Delta m_i)!} \qquad (5)$$

6. Здесь необходимо внести дополнительные уточнения в понятие различимости классов эквивалентности. До сих пор мы считали, что как L_i, так и L_j лишь попарно различимы. Однако различимость элементов некоторого множества по отношению друг к другу еще не означает их различимости в совокупности. В этой связи при рассмотрении той или иной классификации возможны две постановки задачи. В первом случае нас может интересовать вероятность того, что класс L_1 будет содержать ровно n_1 частиц, в классе L_2 окажется n_2 частиц и т.д. Во второй постановке мы можем искать вероятность того, что численность <u>какого-то</u> класса окажется равной n_1, <u>какого-то</u> n_2 и т.д. Когда все $\Delta n_i = n_i - n_{i+1} = 0$, количества вариантов, которыми может осуществиться каждое из этих распределений, пропорциональны друг другу. При этом коэффициент пропорциональности $\mathcal{L}!$ не зависит от численностей классов и мы приходим к одному и тому же наиболее вероятному распределению.

Ситуация, однако, изменяется, если среди численностей n_τ есть кратные. В этом случае прежде всего необходимо четко представить себе, с какой из двух задач мы фактически имеем дело. Различимость классов эквивалентности по отношению друг к другу подразумевает возможность их нумерации. Последнюю можно осуществить различными способами. Если априорно, до классификации, все эти способы эквивалентны по отношению друг к другу, мы имеем дело лишь с попарной различимостью классов эквивалентности. Для различимости в совокупности необходимо, чтобы различие классов могло быть формализовано вне рассматриваемой классификации, т.е. чтобы априорно мы могли отдать предпочтение одному из способов возможной нумерации по отношению к другим.

В связи с вышесказанным в дальнейшем будем отдельно рассматривать обе задачи классификации. Приведенный выше вывод уравнений (4), (5) относится к первому случаю, так как он инвариантен по отношению к нумерации классов эквивалентности

Для того чтобы перейти ко второй задаче выражения (4), (5) достаточно умножить на $\mathcal{L}![\prod_{\tau=0}^{\mathcal{L}}(\Delta n_\tau)!]^{-1}$. Наоборот, при рассмотрении статистики Бозе-Эйнштейна мы упорядочили классы эквивалентности L_τ по убыванию их численности и т.о. получили решение второй задачи. Разделив (6) на $\mathcal{L}!/[\prod_{\tau=0}^{\mathcal{L}}(\Delta n_\tau)!]$, получаем хорошо известное выражение (3).

7. В задаче I каждый из рассмотренных выше функционалов, взятый сам по себе, приводит к тривиальным наиболее вероятным распределениям. Для статистик Больцмана, Ферми-Дирака и Линден-Белла это равновероятное распределение, в то время как для статистики Бозе-Эйнштейна вообще любой набор численности классов обладает одинаковой вероятностью осуществления. Проведенное рассмотрение показывает, однако, что парадигма возможных статистик допускает естественное расширение.

Пусть стратегия выбора ячеек пространства классификации аналогична той, которая приводит к статистике Бозе-Эйнштейна, но в отличие от последней будем заполнять классификационную матрицу не по строчкам, а по столбцам. Упорядочим численности классов $n_1 \geq n_2 \geq \ldots \geq n_\tau \geq \ldots \geq n_{\mathcal{L}}$. Тогда заполненные ячейки в столбце наибольшей численности могут быть выбраны $C_{\mathcal{L}}^{n_1}$ различными способами. На возможные способы выбора ячеек в каждом последующем столбце наложим ограничение, заключающееся в том, что из условия $L_\tau \cap L_{\bar{\tau}} = \emptyset$ следует, что для любых $\bar{\tau} \in \{1,2,\ldots \bar{L}\}$, $\tau, k \in \{1,2,\ldots \mathcal{L}\}$ $L_k \cap L_{\bar{\tau}} = \emptyset$, если $n_k \leq n_\tau$. Тогда для выбора ячеек в следующем столбце остается $C_{n_1}^{n_2}$ воз-

можных комбинаций и т.д. В результате в задаче I вероятность распределения, соответствующего заданному вектору $\vec{n}$, будет равна:

$$P_1(N,\mathcal{L},\bar{\mathcal{L}},\vec{n}) = \frac{\mathcal{L}!}{\prod_{z=0}^{\mathcal{L}}(\Delta n_z)!\cdot\sum_{\vec{n}\in\{\vec{n}\}}\mathcal{L}!\left[\prod_{z=0}^{\mathcal{L}}(\Delta n_z)!\right]^{-1}} \quad (6)$$

а в задаче 2:

$$P_2(N,\mathcal{L},\bar{\mathcal{L}},\vec{n}) = \frac{\mathcal{L}!\,\bar{\mathcal{L}}!}{\prod_{z=0}^{\mathcal{L}}(\Delta n_z)!\prod_{\bar{z}=0}^{\bar{\mathcal{L}}}(\Delta m_{\bar{z}})!\cdot\sum_{\vec{n}\in\{\vec{n}\}}\mathcal{L}!\,\bar{\mathcal{L}}!\left[\prod_{z=0}^{\mathcal{L}}(\Delta n_z)!\prod_{\bar{z}=0}^{\bar{\mathcal{L}}}(\Delta m_{\bar{z}})!\right]^{-1}} \quad (7)$$

Здесь суммирование ведется по всем возможным численностям классов - $\{\vec{n}\}$ при фиксированных "внешних" параметрах классификации N, $\mathcal{L}$, $\bar{\mathcal{L}}$. Также обозначено: $\Delta n_0 = \mathcal{L} - n_1$, $\Delta n_z = n_z - n_{z+1}$ ... $\Delta n_{\mathcal{L}} = n_{\mathcal{L}}$, $\Delta m_0 = \bar{\mathcal{L}} - m_1$, ... $\Delta m_{\bar{z}} = m_{\bar{z}} - m_{\bar{z}+1}$ ... $\Delta m_{\bar{\mathcal{L}}} = m_{\bar{\mathcal{L}}}$. Статистики (6) и (7) можно интерпретировать как такие распределения, для которых максимальная вероятность численностей классов обеспечивается при условии, что и их "первая производная" распределена наиболее вероятно. Покажем, что при варьировании дополнительных условий, обладающих естественным конструктивным содержанием, они описывают весьма широкий спектр различных эмпирически наблюдаемых лингвостатистических распределений.

Частотное распределение букв и фонем связного текста

Хорошо известно, что эмпирические распределения частот встречаемости букв (см., например, Григорьев В.И., 1980) и фонем (Sigurd B., 1968) в первом приближении весьма близки к геометрической прогрессии. Насколько известно автору данной работы, этот факт до настоящего времени не получил достаточно убедительного теоретического обоснования. В связи с этим прежде всего отметим, что распределение, обеспечивающее максимальную вероятность функционалу (6) и учитывающее лишь условие нормировки, соответствует геометрической прогрессии для разностей численностей классов классифицируемых элементов. Еще более хорошее согласие с экспериментом получается, если искать экстремум (6) при заданной величине "энтропии" текста:

$$H=\sum_{z=1}^{\mathcal{L}} \ln n_z! \qquad (8)$$

Подчеркнем, что H в рассматриваемом случае обладает конструктивным смыслом как мера информационной способности знаковой системы того или иного языка. С другой стороны, в соответствии с известным из вариационного исчисления принципом взаимности (см. Смирнов В.И., Крылов В.И., Канторович Л.В., 1933, с. 51) минимизация (6) при условии (8) равносильна поиску экстремалей для (8) при условии (6), т.е. может интерпретироваться как отыскание наиболее вероятного распределения для классификации, подчиняющейся статистике Больцмана, при условии, что разности численностей классов также распределены наиболее вероятно, что для такой "самоорганизующейся" системы, какой является язык, представляется вполне естественным.

Максимум вероятности (6) соответствует минимуму функционала:

$$W=\sum_{z=1}^{\mathcal{L}} \ln (\Delta n_z)! \qquad (9)$$

Таким образом будем минимизировать (9) при дополнительных условиях (8) и нормировки $\sum_{z=1}^{\mathcal{L}} n_z = N$, $\sum_{z=1}^{\mathcal{L}} \Delta n_z = n_1$ – частоте класса наибольшей численности.

Для нахождения оптимального распределения воспользуемся методом неопределенных множителей Лагранжа, что с учетом известного ассимптотического представления $\ln z! \approx z(\ln z - 1)$ сводит задачу к нахождению экстремума функционала:

$$F=\sum_{z=1}^{\mathcal{L}} \left[\Delta n_z (\ln \Delta n_z - 1) + \alpha \Delta n_z + \beta n_z + \mu n_z (\ln n_z - 1)\right],$$

где α, β, μ – неопределенные множители Лагранжа. Потребовав равенства нулю частных производных $\frac{\partial F}{\partial n_z}$, получим систему рекуррентных соотношений, описывающую численности классов наиболее вероятного распределения:

$$\begin{aligned}
&\ln \Delta n_1 = \alpha + \beta + \mu \ln n_1 && \Delta n_1 = n_1 - n_2 \\
&\ln \Delta n_2 = \alpha + \beta 2 + \mu (\ln n_1 + \ln n_2) && \Delta n_2 = n_2 - n_3 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\ln \Delta n_z = \alpha + z\beta + \mu \sum_{s=1}^{z} \ln n_s && \Delta n_z = n_z - n_{z+1}
\end{aligned} \qquad (10)$$

$$\ln \Delta n_{\mathcal{L}} = \alpha + \beta\mathcal{L} + \mu \cdot \sum_{s=1}^{\mathcal{L}} \ln n_s , \qquad \Delta n_{\mathcal{L}} = n_{\mathcal{L}}$$

При $\mu = 0$ она превращается в геометрическую прогрессию для разностей численности классов. В общем случае (10) содержит $\mathcal{L} + 3$ неизвестных $n_1, n_2, \ldots n_r \ldots n_{\mathcal{L}}$, α, β, μ и может быть решена численно, если потребовать, чтобы α, β и μ удовлетворяли экспериментально наблюдаемым значениям

$$\tilde{H} = \sum_{r=1}^{\mathcal{L}} \ln \tilde{n}_r ! , \quad \tilde{N} = \sum_{r=1}^{\mathcal{L}} \tilde{n}_r \qquad \text{и} \qquad n_1 = \tilde{n}_1 .$$

Соответствующие вычисления были проделаны для букв русского алфавита (данные Григорьева В.И.) и фонем языка Kaiwa (данные Sigurd. B). Результаты представлены в табл.I и на рис.I. Легко видеть, что теоретические кривые не только проходят ближе к экспериментальным точкам, чем предложенная B. Sigurd эмпирическая аппроксимация в виде геометрической прогрессии (прямые линии на рис. I), но, что самое главное, объясняют S - образный характер распределения, наблюдаемого экспериментально.

Квантитативно-системные закономерности полисемии

В статистической физике для получения нетривиальных результатов наиболее вероятные распределения ищутся не на множестве всех логически возможных векторов $\{\vec{n}\}$, а лишь на отдельных его подмножествах, удовлетворяющих определенным законам сохранения моментов искомых распределений, которые теряют свою очевидность в случае произвольной классификации. Тем не менее в ряде задач лингвостатистики первый момент распределения обладает вполне естественным содержанием. С такой ситуацией мы имеем дело, например, при изучении распределения частот встречаемости r - буквенных слов, для которого сумма

$$\sum_{r=1}^{\mathcal{L}} r n_r = M \qquad (11)$$

равна числу букв в исследуемом тексте. Подобным образом при рассмотрении полисемии выражение (11) характеризует суммарное количество значений того или иного словаря. Если дополнительное условие (8) заменить на (11), то, проведя рассуждения, аналогичные вышеизложенным, для разностей численно-

стей классов получим выражение:

$$\Delta n_r = \exp(\alpha + \beta r + \gamma r^2) \qquad \Delta n_r = n_r - n_{r+1} \quad (12)$$

Здесь постоянные α, β, γ должны определяться из условий $\sum_{r=1}^{\mathcal{L}} \Delta n_r = \tilde{n}_1$, $\sum_{r=1}^{\mathcal{L}} n_r = \tilde{N}$, $\sum_{r=1}^{\mathcal{L}} r n_r = \tilde{M}$ где $\tilde{n}_1, \tilde{N}, \tilde{M}$ - экспериментально наблюдавшиеся значения, соответственно, числа однозначных слов $\tilde{n}_1$, объема словаря $\tilde{N}$ и полного количества значений в словаре $\tilde{M}$. При этом $\mathcal{L}$ - максимальное количество значений у слов (слова) наибольшего семантического содержания должно быть таково, чтобы выполнялось условие $(\Delta n)_{\mathcal{L}} = n_{\mathcal{L}}$.

Данные сопоставления теоретических распределений (12) с фактическими для словаря Ожегова (собственные подсчеты автора) представлены в таблице 2. Можно констатировать хорошее совпадение экспериментальных и теоретических результатов, особенно на "хвостах" распределений, что обеспечивается положительностью величины γ.

Здесь следует обратить внимание читателя, что распределение (12) является дискретным вариантом хорошо известного нормального закона. Однако существенно, что в рассматриваемой статистике ему подчиняются не сами численности классов, а их разности. Если для обычного нормального распределения коэффициент при γ обязан быть отрицательным (в противном случае имеет место расходимость при $r \to \infty$), то для (12) γ может быть как отрицательным, так и больше нуля. Последнее связано с тем, что распределение (12) принципиально конечно, так как расходимость "автоматически" снимается условием $\Delta n_{\mathcal{L}} = \mathcal{L}$.

Ранговые лингвостатистические распределения

При рассмотрении частотной структуры нас интересует вопрос о вероятности того, что частота каких-то (см. п. 6) (априорно безразлично каких) слов будет равна заданному значению n_r. Соответственно в этом случае мы уже имеем дело с задачей 2. Покажем, что распределение Ципфа-Мандельброта является одним из предельных вариантов экстремального распределения для функционала (7). Очевидно, что максимум вероятности (7) соответствует минимуму

$$W = \sum_{r=1}^{\mathcal{L}} \ln(\Delta n_r)! + \sum_{r=1}^{\mathcal{L}} \ln(\Delta m_r)! \quad (13)$$

Ограничимся рассмотрением непрерывной аппроксимации (I3), в качестве дополнительного условия учитывающей лишь условие нормировки $\sum_{z=1}^{\infty} n_z = N$. Тогда после замены сумм соответствующими интегралами задача [illegible] к поиску экстремума функционала

$$\bar{I} = \int_{0}^{x_0} [\ln \Gamma(1-y) - y' \ln \Gamma(1-\frac{1}{y'}) + \lambda y] dx = \int_{0}^{x_0} F(x, y, y') dx \qquad (I4)$$

Здесь $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$ - стандартное обозначение Гамма-функции; знак минус перед производной в ее аргументе обусловлен тем, что $y' < 0$ (как принято, слова текста ранжированы в порядке убывания их частоты). Из вариационного исчисления известно, что $y = y(x)$ обеспечивающая экстремум (I4), должна удовлетворять уравнению Эйлера:

$$\frac{d}{dx} \frac{\partial F(x, y, y')}{\partial y'} - \frac{\partial F(x, y, y')}{\partial y} = 0 \qquad (I5)$$

если обозначить $\Psi(z) = \frac{d \ln \Gamma(1+z)}{dz}$ с учетом известной аппроксимации логарифмической производной Гамма-функции $\frac{d\Psi(z)}{dz} = \frac{1}{z} - \frac{1}{2z^2}$ (I5) конкретизируется в виде

$$y'' \cdot \frac{1-y'}{2y'^2} = \lambda \qquad (I6)$$

Интегрирование (I6) приводит к параметрическому представлению экстремального распределения:

$$\begin{cases} y = \frac{1}{2\lambda} (\ln q - q) - C_2 \\ x = -\frac{1}{2\lambda} (\frac{1}{q} + \ln q) - C_1 \end{cases} \qquad (I7)$$

Если в (I7) пренебречь логарифмической зависимостью по сравнению с линейной и обратно-пропорциональной, то окончательно после исключения параметра q получим:

$$y = \frac{A}{x + C_1} - C_2 \quad \text{где} \quad A = \frac{1}{4\lambda^2} \qquad (I8)$$

Распределение (I8) принадлежит к классу гиперболических распределений. Оно лишь наличием слагаемого C_2 отличается от известной формулы Мандельброта $y = A(x+B)^{\gamma}$, в которой $\gamma = 1$, что весьма близко к экспериментально наблюдаемым значениям γ. При $C_1 = C_2 = 0$ (I8) переходит в классическое однопараметрическое распределение Ципфа $y = Ax^{-1}$.

Распределение (I8) представляется даже более естественным, чем формула Мандельброта. От формулы Ципфа оно отличается лишь возможностью сдвига в направлениях, параллельных координатным осям. Такое преобразование обеспечивается наличием постоянных C_1 и C_2. Сдвиг в направлении оси абсцисс вполне понятен. Он обуславливает инвариантность рангового распределения по отношению к определению ранга, т.е. позволяет отсчитывать ранг как от 0, так и от I или приписать рангу полуцелые значения I/2, 3/2, и т.д. Критичность ранговых распределений по отношению к малым смещениям в направлении оси ординат подробно обсуждалась в работах (Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А., I975; Арапов М.В., Ефимова Е.Н., I975; Арапов М.В., I977). Именно высокая чувствительность гиперболических распределений к малым изменениям производной на хвосте распределения побудила авторов этих работ для описания эмпирических распределений считать, что γ может несколько отличаться от I.

Постоянная C_2 в формуле (I8) в какой-то мере играет ту же роль, что и параметр γ в формуле Мандельброта. С другой стороны, наличие C_2 снимает многие трудности, связанные с расходимостью распределения Ципфа-Мандельброта при $\gamma \leq I$. Действительно, C_1 и C_2 всегда можно выбрать так, чтобы кривая (I8) пересекала каждую из координатных осей. Обозначим координаты точек пересечения $(0, y_0)$ и $(x_0, 0)$. Очевидно, что $y_0 = \frac{A}{C_1} - C_2$, $x_0 = \frac{A}{C_2} - C_1$. Тогда условие нормировки можно записать в виде:

$$N = \int_0^{x_0} \left(\frac{A}{x+C_1} - C_2 \right) dx = A\left(\ln \frac{A}{C_1 \cdot C_2} - 1 \right) + C_1 C_2 \quad (19)$$

Теперь ступенчатую функцию, соответствующую дискретному аналогу непрерывного распределения, следует определить так, чтобы площадь всех прямоугольников гистограммы точно давалась выражением (19). Это требование легко выполняется, если положить, что частота слова ранга r (при $|y'|>1$ т.е. для головы распределения) равна среднему значению $y = y(x)$ в интервале $[r-1, r]$:

$$n_r = \int_{r-1}^{r} \left(\frac{A}{x+C_1} - C_2 \right) dx = A \ln \frac{r+C_1}{r-1+C_1} - C_2 \quad (20)$$

Для "хвоста" распределения, т.е. в области $|y'|<1$, перейдем к выражению для обратной зависимости

$$x = \frac{A}{y+C_2} - C_1 \quad (21)$$

Тогда

$$\mathcal{L} = \int_0^1 x(y)\, dy = A \ln \frac{C_2+1}{C_2} - C_1 \quad (22)$$

будет давать объем $\mathcal{L}$ словаря текста. Наконец, m_κ – число κ – разовых слов в тексте определим разностью интегралов:

$$m_\kappa = \int_{\kappa-1}^{\kappa} x(y)\, dy - \int_{\kappa}^{\kappa+1} x(y)\, dy = A \ln \frac{(\kappa+C_2)^2}{(\kappa+C_2)^2 - 1} \quad (23)$$

Для вычисления постоянных A, C_1, C_2, соответствующих конкретному тексту, в формулах (19) (22), следует положить $N = \tilde{N}$, $\mathcal{L} = \tilde{\mathcal{L}}$, где $\tilde{N}$ и $\tilde{\mathcal{L}}$ – наблюдавшиеся в эксперименте длина текста $\tilde{N}$ и объем его словаря $\tilde{\mathcal{L}}$. Третье уравнение для определения A, C_1 и C_2 можно получить потребовав $n_1 = \tilde{n}_1$ – фактической частоте наиболее частого слова. Однако, учитывая, что в области малых рангов экспериментальные частоты испытывают значительные флуктуации, лучше использовать условие

$$\tilde{n}(z)=\sum_{s=1}^{z}\tilde{n}_s=\int_0^z\left(\frac{A}{x+C_1}-C_2\right)dx=A\ln\frac{z+C_1}{C_1}-C_2 z \quad (24)$$

где z равно, скажем, 5 или 10.

Система уравнений (19), (22), (24) однозначно определяет параметры распределения. Для получения ее точного численного решения можно воспользоваться итерационным методом. Сходимость итерационного процесса обеспечивается последовательностью итераций по схеме:

$$A_{s+1}=\frac{\tilde{N}-C_{2s}\cdot C_{1s}}{\ln A_s-\ln C_{2s}C_{1s}-1}$$

$$C_{2,s+1}=\left[\exp\frac{\tilde{z}+C_{1s+1}}{A_{s+1}}-1\right]^{-1}$$

$$C_{1,s+1}=z\left[\exp\frac{\tilde{n}(z)+zC_{2s}}{A_{s+1}}-1\right]^{-1}$$

В качестве нулевого приближения можно взять $C_{20}=C_{10}=1$, $A_0=\tilde{N}$.

Подробное сопоставление теоретического распределения (18) с частотными структурами, наблюдаемыми экспериментально, и с другими моделями ранговых распределений (например, М.К. Орлова) в контексте данной статьи, к сожалению, невозможно, так как привело бы к значительному увеличению ее объема. В связи с этим ограничимся описанием зоны редкочастотных слов ($1 \le K \le 15$) лексических спектров некоторых известных литературных произведений. Соответствующие данные приведены в таблице 3. Предварительный анализ показывает, что предлагаемая модель согласуется с экспериментом по крайней мере не хуже, чем другие лингвостатистические модели. Основным же ее достоинством является то, что формула (18) получена чисто вероятносто-комбинаторными рассуждениями, без каких-либо феноменологических допущений о виде функции распределения.

В заключение автор не может не высказать глубокую благодарность М.В. Арапову за внимание и интерес к работе. Именно статья М.В. Арапова и М.А. Шрейдера (1978) пробудила интерес автора к рассматриваемому кругу вопросов. В дальнейшем многие из них неоднократно обсуждались автором с М.В. Араповым

в личных беседах, что весьма способствовало работе над данной проблемой.

ЛИТЕРАТУРА

Агекян Т.А. Теория вероятностей для астрономов и физиков. – М.: Наука, 1974.

Арапов М.В. Две модели рангового распределения.– Вопросы информационной теории и практики, № 4. – М.: ВИНИТИ, 1977.

Арапов М.В., Ефимова Е.Н. Понятие лексической структуры текста. – Научно-техническая информация. Серия 2, № 6. М., 1975, с. 3-7.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. – Научно-техническая информация. Серия 2, № 1. М.: 1975, с. 9-20.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. – Семиотика и информатика, № 10. М., 1978, с. 74-95.

Борн М. Атомная физика. – М.: Мир, 1965.

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

Григорьев В.И. Frequency distribution of letters and their ranks in a running text. – Symposium: Computational linguistics and related topics. Tallinn, November 24 – 26 1980, p. 43-47.

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

Петров В.М., Яблонский А.И. Гиперболические распределения и их применения. – М.: Знание, 1980.

Смирнов В.И., Крылов Б.И., Канторович Л.В. Вариационное исчисление. – Л.: Изд-во КУБУЧ, 1933.

Sigurd B. Rank-Frequancy Distributions for Phonemes. – Phonetica 18, 1968, p. 1-15.

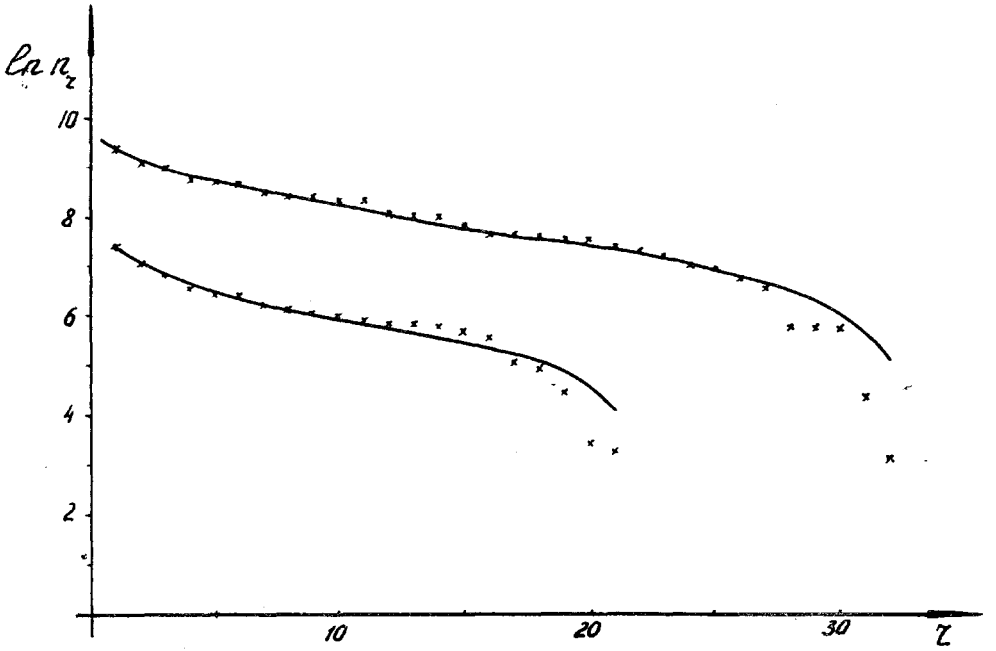


Рис. I. Экспериментальные и теоретические распределения частот букв русского языка и фонем языка Kaiwa

Таблица I

Численности букв русского языка и фонем языка

Русские буквы		Фонемы языка Kaiwa	
Теоретические частоты	Экспериментальные частоты	Теоретические частоты	Экспериментальные частоты
II4I0	II4I0	I767	I767
9726	8624	I266	II27
8385	8002	973	9I3
7302	6536	785	708
64I7	6097	659	650
5684	5926	568	633
5070	5072	499	508
455I	4674	445	469
4I07	4492	402	427
3724	4I57	365	4I5
339I	4I40	334	37I
3098	3098	306	346
2839	3095	280	342
2607	2977	257	33I
2398	2488	234	298
2209	2092	2II	258
2036	2090	I88	I54
I877	I98I	I63	I40
I729	I939	I35	87
I590	I9I2	I03	3I
I460	I6II	62	25
I336	I490	0	0
I2I7	I373		
II02	II30		
990	I0I2		
880	857		
770	685		
658	323		
545	3I0		
427	304		
30I	8I		
I62	22		
0	0		

α = 7,6722918
β = 0,6287636
μ = - 0,09329457

α = 6.853508
β = I,6854I8
μ = - 0,3I0670I

Таблица 2

Распределение многозначных слов в словаре Ожегова в зависимости от количества значений

Число значений	Весь словарь теор.	эксп.	Существительные теор.	эксп.	Глаголы теор.	эксп.
I	23456	23456	I2522	I2522	7042	7042
2	6I95	6288	2994	3033	223(	23I5
3	I836	I682	8II	735	770	69I
4	6I3	649	25I	287	288	270
5	232	232	89	85	II7	II7
6	I00	II0	37	42	52	52
7	50	55	I8	I7	26	3I
8	27	34	I0	I3	I4	I7
9	I7	23	7	7	8	I4
I0	II	I6	5	3	5	II
II	8	7	4	4	4	3
I2	6	2	0	0	3	2
I3	4	2			2	2
I4	3	0			2	0
I5	2	2			I	2
I6	I	I			I	I
I7	0	0			0	0
	α = II,23776		α = I0,75063		α = 9,73888	
	β = - I,53426		β = - I,6460I6		β = - I,298554	
	μ = 0,0526936		μ = 0,0573946		μ = 0,0372021	

Таблица 3

Лексические спектры зоны редкочастотных слов

Частота слова	"Поднятая целина" Шолохов		"Дым отечества" Паустовский		"Реалисты" Писарев		"Горе от ума" Грибоедов		"Коро... Дир" пер. Пастернака	
	теор.	эксп.	теор.	эксп.	теор.	эксп.	теор.	эксп.	теор.	эксп.
I	8233	8048	4009	3798	3090	3092	2I99	2079	2664	2578
2	3I25	3066	I27I	I277	I05I	I058	436	526	563	6I9
3	I660	I688	633	705	536	546	I94	2I0	254	269
4	I032	I097	38I	393	326	3I5	III	I28	I45	I65
5	704	699	254	264	2I0	I80	7I	86	94	85
6	5I2	559	I82	209	I58	I55	50	47	66	78
7	389	424	I37	I53	II9	I42	37	33	49	47
8	305	30I	I07	92	93	III	28	24	38	42
9	246	259	85	87	75	70	23	I3	30	29
I0	203	2I5	70	84	6I	72	I8	I5	24	22
II	I70	I67	58	58	5I	55	I5	I6	20	2I
I2	I44	I45	50	59	43	46	I3	II	I7	I3
I3	I24	I20	42	60	37	33	II	I0	I3	8
I4	I08	III	37	36	32	2I	9	I5	II	I3
			32	33	28	I8	8	4		
					25	26	7	3		
							6	II		
	$\tilde{N}$ = I93360		$\tilde{N}$ = 56666		$\tilde{N}$ = 48354		$\tilde{N}$ = I3425		$\tilde{N}$ = I9490	
	$\mathcal{L}$ = I8498		$\mathcal{L}$ = 7827		$\mathcal{L}$ = 6348		$\mathcal{L}$ = 3343		$\mathcal{L}$ = 4I6I	
	$\tilde{R}(5)$ = 23363		$\tilde{R}(5)$ = 703I		$\tilde{R}(5)$ = 6023		$\tilde{R}(5)$ = I9I0		$\tilde{R}(5)$ = 3394	
	A = 23730,6I2		A = 7775,205		A = 6944,997		A = I90I,048		A = 2533,54I	
	C_1 = 2,98I464		C_1 = 3,398867		C_1 = 3,6I9294		C_1 = 2,885828		C_1 = I,7734II	
	C_2 = 0,8469632		C_2 = 0,5754913		C_2 = 0,6685934		C_2 = 0,2077895		C_2 = 0,2397509	

A PARADIGM OF LINGUOSTATISTIC DISTRIBUTIONS

Yuri Krylov

S u m m a r y

In this article the well-known statistics of Boltzmann, Fermi-Dirac, Bose-Einstein, and Linden-Beli are examined within the framework of a common combinative-probabilistic scheme. The investigation has shown that the paradigm of these statistics allows a natural extension. The results of the theoretical analysis are compared with the findings of some experiments of examining the frequency of occurrence of letters and phonemes in coherent texts, the quantitative-systemic laws of polysemy, and the frequency distributions of lexics in completed literary works. Good fit between the theoretical and empirical data has been obtained.

ТИПОЛОГИЯ ЛИНГВОСТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Г.Я. Мартыненко

Период исключительного внимания лингвостатистиков к технике обработки экспериментальных данных, по-видимому, подходит к концу. На первый план выдвигается интерес к общим принципам, раскрывающим вероятностную природу языка. Лингвистическая мысль движется при этом в двух, дополняющих друг друга направлениях. Первое направление характеризуется интересом к вероятностно-статистической интерпретации коренных лингвистических понятий и проблем, таких, как антиномия языка и речи, соотношение плана выражения и плана содержания, взаимодействие парадигматики и синтагматики, соотношение понятий система, структура, поле и т.п. Для второго направления характерен интерес к лингвистическому осмыслению основных вероятностно-статистических категорий, таких как вероятность и частота, статистическая совокупность, статистическая закономерность, статистическая группировка, статистическое распределение и др.

Данная работа принадлежит ко второму из названных направлений. В ней предпринята попытка обобщения опыта использования статистических распределений в лингвистике. Лингвостатистическая значимость этой категории объясняется не только тем, что теория распределений занимает центральное место в общей теории статистики, не только тем, что статистические распределения являются удобным средством компактного представления и анализа экспериментальных данных, но, главным образом, тем, что естественные языки обладают рядом принципиальных свойств, присущих вероятностным системам в целом: единством иррегулярности и устойчивости в классе событий, единством автономности и зависимости событий, единством порядка и беспорядка в классе событий (Кравец А.С.,1976, 56). Вероятностные распределения являются главным выразителем и средством описания этих свойств естественного языка как вероятностной системы.

Основная задача работы – выделение типов лингвостатистических распределений. При определении степени существенности

классификационных признаков мы руководствовались в первую очередь теми сторонами лингвистической реальности, которые могут быть интерпретированы в системном плане, а лишь затем – формально-математическими соображениями.

Предлагаемая система типов распределений не претендует на полноту и строгость, это скорее перечень аспектов, в рамках которых выделяются конкретные типы распределений. Всего использовано I2 аспектов, каждый из которых строится в виде бинарных противопоставлений.

I. Противопоставление "вероятностная теория – лингвистическая реальность"

В рамках этого противопоставления можно говорить об эмпирических распределениях, отражающих результат группировки данных лингвостатистического наблюдения, и теоретических распределениях, выбранных для описания закономерности, которой подчиняется фактическое распределение.

Хотя число потенциально возможных теоретических распределений очень велико, все многообразие эмпирических рядов стараются свести к хорошо исследованным теоретическим вариантам. Многие из них занимают в статистике особое положение, либо потому, что они обладают желательными математическими свойствами, либо потому, что описываемый ими механизм вариации строго или хотя бы приблизительно соответствует вариации, характерной для конкретного фрагмента действительности. Важной чертой теоретических распределений является то, что многие из них могут рассматриваться как частные случаи более широких семейств.

2. Противопоставление "гауссовость – негауссовость"

С теоретико-вероятностной точки зрения все реальные и мыслимые варианты распределений распадаются на две группы: семейство гауссовых и семейство негауссовых распределений. К первой группе относятся теоретические распределения, описываемые тем или иным вариантом экспоненциальной функции. В роли эталона этой группы выступает нормальное распределение, к которому сходятся все варианты гауссового семейства кривых распределения (биноминальное распределение, распределение Пуассона, распределение Стьюдента и многие другие). В роли эталона второй группы выступает уравнение неравносторонней гиперболы. Это уравнение считается идеальным образцом (или

асимптотикой), к которому устремляются все варианты распределений гиперболического типа (распределение Ципфа-Парето, Эсту-Лотки, Мандельброта и др.). Отличительной чертой негауссовых распределений является бесконечность моментов любого порядка, в частности дисперсии.

История изучения негауссовых распределений сравнительно молода. Их основные свойства были впервые исследованы Полем Леви (Lévy P., 1925), который в частности установил, что при композиции двух распределений гиперболического типа получается распределение того же вида. Из этого следует, что негауссовы распределения с теоретико-вероятностной точки зрения принадлежат к классу устойчивых распределений.

3. Противопоставление "типические - нетипические ряды"

Делению теоретических распределений на гауссовы и негауссовы в традиционной статистике соответствует деление эмпирических рядов на типические и нетипические. Типическими рядами считаются те, в которых механизм взаимодействия постоянных и случайных причин строго или хотя бы приблизительно соответствует моделям массовых процессов, обобщенных в первоначальных теоремах закона больших чисел. Все остальные ряды рассматриваются как нетипические (Кауфман А.А., 1909, с.353).

4. Противопоставление "ранг-частота"

Каждое теоретическое распределение имеет вид функции, устанавливающий связь между возможными значениями случайной величины и соответствующими им вероятностями. Существует несколько способов задания этой связи. Важнейшие из них представлены в таблице I, а в таблице 2 приведены варианты двух распределений простейшего типа: распределения Ципфа-Парето и показательного распределения. Обе таблицы построены на основании сведений, систематизированных Н. Хастингсом и Дж. Пикоком (Хастингс Н., Пикок Дж., 1980).

Из таблицы I видно, что разные способы задания распределения находятся в отношениях взаимно-однозначных отображений: интегральная форма - дифференциальная форма, прямая функция - обратная функция, вероятность события - вероятность противоположного события. В лингвостатистических исследованиях связь между различными формами задания одного и того же распределения часто не осознается, в частности не учитывается связь между двумя формами рангового распределе-

Таблица I

Способы задания функциональной связи между случайной величиной и вероятностью

Теоретико-вероятностный термин	Способы	Математико-статистический термин	Лингвостатистический термин
I	2	3	4
I. Функция распределения $F(x)$	$F(x)$ есть вероятность того, что случайная величина принимает значение, меньшее или равное x $F(x) = Prob[X < x] = \alpha$ $F(x) = \int_{-\infty}^{x} f(u)du$	Кумулятивный ряд	
2. Плотность вероятности	Функция, интеграл от которой по промежутку от x_L до x_u равен вероятности того, что случайная величина принимает значение из этого промежутка $\int_{x_L}^{x_u} f(x)dx = Prob[x_L < X < x_u]$ $f(x) = \frac{d[F(x)]}{dx}$	Вариационный ряд	Спектровое распределение
3. Обратная функция распределения (функция квантилей) $G(d)$	$G(d)$ – такое число, что случайная величина примет значение, не превосходящее α, с вероятностью $G(\alpha)$ $x = G(d) = G[F(x)]$ $Prob[X < G(\alpha)] = \alpha$	Возрастающий ранжированный ряд	

Продолжение табл. I

I	2	3	4
4. Функция выживания $S(x)$	$S(x)$ – это вероятность того, что случайная величина примет значение большее, чем x $$S(x) = Prob[X > x] = 1 - F(x)$$		
5. Обратная функция выживания (обратная функция квантилей) $Z(\alpha)$	$Z(\alpha)$ – это такое значение, которое случайная величина превосходит с вероятностью α $$Prob[X > Z(\alpha)] = \alpha$$ $$x = Z(\alpha) = Z(S(x)),$$ где $S(x)$ – функция выживания, $$Z(\alpha) = G(1-\alpha)$$ где G – обратная функция распределения.	Убывающий ранжированный ряд	Ранговое распределение

ния (возрастающим и убывающим), а также между функцией распределения и ранговым распределением. В таблице 2 показано, что между распределением Юла (спектровое распределение) и распределением Ципфа (убывающее ранговое распределение) нет никакой сущностной разницы, это всего лишь разные формы задания одного и того же распределения.

5. Противопоставление "разнообразие - ограничение разнообразия"

Порождение речи протекает на фоне двух тенденций: стремления к расширению состава разноименных единиц (увеличения разнообразия) и стремления к его сокращению (ограничение разнообразия) (Тулдава Ю., 1980). Этим противоборствующим тенденциям соответствуёт два типа количественной вариативности лингвистических единиц: вариативность состава (номенклатурная колеблемость знакотипов) и рекуррентная вариативность (колеблемость числа знакоупотреблений каждого знакотипа). Распределения, устанавливающие связь между частотой знакотипа и числом знакотипов с данной частотой называют распределением знакотипов, а распределения, в которых частоте знакотипа ставится в соответствие число знакоупотреблений - называют распределением знакоупотреблений. Распределения, связывающие накопляемое количество знакотипов с накопляемым числом знакоупотреблений, принято называть распределениями знакотип-знакоупотребление (Carrol J.B., 1967).В статистической лексикографии этот тип распределения используется при вычислении индексов лексического богатства и лексической концентрации текстов.

6. Противопоставление "строение - поведение"

Объект познания - то, на что направлена познавательная деятельность, представляет единство двух противоположных сторон: поведения и его материального носителя - строения (Смирнов С.Н., 1978). Из этого следует, что при построении статистических распределений могут быть использованы две категории варьирующих признаков: 1) признаки, отражающие внутреннюю структуру единиц совокупности, 2) признаки, указывающие на статус (функциональный вес) единиц, образующих совокупность. В соответствии с этими категориями признаков можно говорить о двух типах распределений: статусных и структурных. Текстовые словоформы могут быть упорядочены, например,

Таблица 2

Формы функционального задания показательного распределения и распределения Ципфа-Парето

Формы задания распределения	Распределение Ципфа-Парето	Показательное распределение	Обозначения и примечания
I. Кумулятивный ряд	$K-Kx^{-c}$	$K-Ke^{-\lambda x}$	K - объем совокупности
2. Вариационный ряд	$Kcx^{-(c+1)}$	$K\lambda e^{-\lambda x}$	x - величина варьирующего признака
3. Возрастающий ранжированный ряд	$\left(\frac{K}{K-r}\right)^{\frac{1}{c}}$	$\frac{1}{\lambda}\ln\frac{K}{K-r}$	r - ранг
4. Антикумулятивный ряд	Kx^{-c}	$K\lambda e^{-\lambda x}$	
5. Убывающий ранжированный ряд	$\left(\frac{K}{r}\right)^{\frac{1}{c}}$	$\frac{1}{\lambda}\ln\frac{K}{r}$	Термин "антикумулятивный ряд" предложен нами

по их размеру (числу букв, фонем, морфем), т.е. по признаку, отражающему внутреннюю структуру словоформ. В этом случае "коллектив" словоформ образует структурное распределение. Но та же совокупность словоформ может быть упорядочена по функциональному признаку, например, по их встречаемости в тексте. В этом случае образуется статусное распределение. Между двумя типами распределений нет абсолютно четкой границы: функции конкретного объекта определяются тем, как он устроен, и, наоборот, внутреннее строение объекта в значительной мере определяется его функциями. Связь между статусными и структурными распределениями характерна для организованных коллективов любой природы: от неорганической до знаково-информационной. Например, весовое содержание конкретного химического элемента в различных средах зависит от его атомного веса, численность популяции в биоценозе - от размера особи, а частота слова в тексте - от его размера, степени полисемии, объема понятия и т.п.

7. Противопоставление "элемент - совокупность элементов"

Статистическое описание совокупности объектов занимает промежуточное положение между индивидуальным описанием каждого из объектов совокупности в отдельности и описанием совокупности по ее общим свойствам, совсем не требующим ее расчленения на отдельные объекты (Колмогоров А.Н., 1954, с. 485). Статистический интерес может смещаться или в сторону изучения свойств отдельного объекта, или в сторону изучения совокупности объектов. В первом случае лингвиста интересует поведение одного и того же объекта в различных ситуациях (однопредметное наблюдение). Во втором случае исследуется поведение различных, но родственных в каком-либо отношении единиц в одной и той же или группе однородных ситуаций (многопредметное наблюдение). Различие в целях исследования ведет к построению двух контрастных типов распределений. Распределения первого типа назовем однопредметными, а распределения второго типа - многопредметными. Принципиальная непохожесть этих распределений была впервые осознана А. Кетле, который заметил резкую грань, отделяющую средние величины, полученные при многократном измерении признака одного и того же объекта (например, высоты одного и того же дома) и однократном измерении величины разнокачественных, но родственных

в каком-либо отношении объектов (например, высоты разных домов на одной и той же улице).

Различие между однопредметными и многопредметными распределениями с особой откровенностью проявляется тогда, когда результаты эксперимента представлены в виде двухмерной таблицы. Примером такой таблицы могут служить результаты анализа цветописи Гоголя, осуществленного Андреем Белым (Белый А., 1934, с. 121).

Таблица 3

Цветовой спектр Гоголя (по А. Белому)

Цвета	Группа произведений*				Статистическая средняя в %
	I	II	III	IУ	
Красное	26,6	12,5	10,3	6,4	17,4
Белое	9,5	9,0	22,0	17,0	14,0
Черное	11,0	14,1	11,8	4,8	12,0
Зеленое	8,6	7,7	9,6	21,6	9,4
Золотое	11,6	8,9	2,8	12,8	9,2
Синее	10,7	6,1	4,9	6,4	8,7
Желтое	3,5	8,5	10,3	12,8	7,0
Серое	2,6	8,9	10,5	6,4	5,8
Голубое	4,4	5,7	7,0	1,6	4,8
Серебряное	7,1	3,2	2,8	4,8	4,8
Коричневое	0,9	6,5	8,4	1,6	4,0
Розовое	3,8	0,8	2,1	3,2	2,3
Оранжевое	0,3	1,6	2,8	0,0	1,2
Лиловое	1,7	1,2	0,0	1,6	0,9

Данные, приведенные в строках табл. 3, могут рассматриваться как однопредметные распределения (вариация конкретного цвета по группам произведений), а данные столбцов – как многопредметные распределения (вариация цветовой гаммы каждой группы произведений). Эта пространственная ориентация нашла отражение в терминологии, уже получившей некоторое распространение в статистической лексикографии: однопредметные распределения называют горизонтальными, а многопредметные – вертикальными распределениями (Алексеев П.М., 1977, с. 17).

* I – "Вечера на хуторе близ Диканьки", "Вий" и "Тарас Бульба", II – повести и комедии, III – первый том "Мертвых душ", IУ – второй том "Мертвых душ".

В заключение отметим, что в роли "предметов" конкретного распределения могут выступать объекты разной степени общности: от самых элементарных, например, тех, которые используются в обычных частотных словарях, до предельно обобщенных, например, тех, которые сконструировал Андрей Белый. А это означает, что результаты статистического эксперимента находятся под сильным влиянием внестатистической концепции, положенной в основу формирования обобщенных объектов, а также интуиции, изобретательности и даже смелости исследователя.

8. Противопоставление "виртуальность - актуальность"

В лингвостатистике различают два типа совокупностей в зависимости от того, к какому уровню познавательной деятельности они относятся: к уровню непосредственного наблюдения или к сфере объективизации. В первом случае исследуется вариативность лингвистических единиц в тексте, во втором - в продуктах классификационной деятельности лингвиста: терминологических словарях, семантических сетях, информационно-поисковых тезаурусах и т.п. Распределения, строящиеся на материале объективизированных (виртуальных) единиц, называют распределениями в сфере фиксации, а распределения, строящиеся на материале текстовых (актуальных единиц) - распределениями в сфере функционирования (Марусенко М.А., 1980).

9. Противопоставление "однородность-неоднородность"

Основным условием, определяющим группу в статистическом смысле, является однородность единиц совокупности. Только однородная в качественном отношении совокупность может быть описана с помощью обобщающих показателей различного типа. В большинстве случаев вопрос формирования однородной совокупности решается на стадии планирования эксперимента путем качественного анализа с последующей стратификацией неоднородной совокупности на качественно однородные группы. В результате такого анализа строится несколько распределений, состоящих из однородных в качественном отношении единиц. Однако в ряде случаев качественная специфика лингвистических объектов такова, что прямое распознавание принадлежности конкретного объекта к данной подсовокупности осуществить не удается, т.е. неоднородность априорно не устранима. Это дает основание говорить о распределениях с априорно устранимой и

распределениях с априорно не устранимой неоднородностью. Последние возникают в том случае, когда между подмножествами лингвистических единиц нет четкой границы, она может быть установлена лишь условно с помощью формальных процедур, разрабатываемых в теории статистических группировок.

10. Противопоставление "устойчивость-неустойчивость"

Проблема устойчивости статистических чисел является центром, вокруг которого уже не одно десятилетие вращается статистическая мысль. Однако длительность общения с этой проблемой не привела к серьезным сдвигам в ее решении. Даже сегодня некоторые ученые рискуют утверждать, что "вопрос о причинах устойчивости частот можно считать только поставленным. Его решение еще не предвидится, хотя некоторые усилия в этом направлении предпринимаются" (Сачков Ю.В., 1971, 32).

Трудность решения этой проблемы заключается в том, что в явлениях реальной действительности нет той абсолютной устойчивости, которая характерна для схем, конструируемых теорией вероятностей. В статистике, включая лингвистическую, принято говорить об относительной устойчивости статистических рядов, сравнительно медленно изменяющихся во времени и пространстве. Это означает, что при интерпретации эмпирической реальности в вероятностных терминах нужно соблюдать известную осторожность. Ведь считается чуть ли не аксиомой, что аппарат теории вероятностей может быть применен лишь к таким явлениям, где частоты обладают устойчивостью (Бентцель Е.С., 1952, с. 31). И все же эта аксиома не останавливает статистиков-практиков в их усилиях познать закономерности реальных массовых процессов именно на почве теории вероятностей. Эти усилия не лишены оснований.

Во-первых, роль внешнего арбитра, "примиряющего" теорию с практикой, выполняет закон больших чисел, содержание которого не исчерпывается содержанием математическим. Это закон объективной действительности, внешне проявляющийся в устойчивости статистических чисел. "Устойчивость статистических чисел, их свойство колебаться из года в году, лишь в известных, ограниченных пределах - писал А.А. Чупров, - представляет собой эмпирически установленный факт, который, сам по себе, независимо от тех или иных теоретических истолкований, имеет громадную научную и жизненную важность. Это один из коренных, хотя и мало заметных, устоев современной культуры" (Чупров А.А., 1909, с. 287).

Во-вторых, хотя все модели, конструируемые в теории вероятностей, начиная с теоремы Бернулли, ориентированы на испытания с регламентированной, целесообразно организованной случайностью (Смолуховский М., 1927), условия, при которых предоставляется роль случаю, в них постоянно расширяются и усложняются, все более приближаясь к условиям взаимодействия случайности и необходимости в явлениях реальной действительности.

В-третьих, любой разумно организованный статистический эксперимент может обладать практической (и не только практической) ценностью и без обращения к теоретико-вероятностным схемам, но "лишь на почве теории вероятностей удается осмыслить наблюдаемые в статистических числах колебания, лишь теория вероятностей открывает возможность прагматического их истолкования в каждом отдельном случае" (Чупров А.А., 1909, с. 280).

Перечисленные факторы в совокупности создают основу для разыскания критериев устойчивости эмпирических распределений. Первый шаг в этом направлении был сделан немецким статистиком В. Лексисом, предложившим для измерения устойчивости статистических рядов особый показатель Q^2 - коэффициент дивергенции. Смысл этого показателя заключается в том, что эмпирическая дисперсия сравнивается с теоретической величиной, условно определяемой таким образом, что частость, взятая для всего ряда данных, приравнивается вероятности:

$$Q^2 = \frac{2\sum(P_i - P)^2}{m-1} : \frac{2P(1-P)}{n}$$

где m - число серий по n испытаний,

P_i - частости в сериях,

P - общая частость, приравненная вероятности.

По Лексису небольшое отличие Q^2 от единицы свидетельствует о нормальном рассеянии, если же Q^2 меньше единицы, то устойчивость ряда сверхнормальна, а если больше единицы, то устоичивость ряда поднормальна (Lexis W., 1903, 190).

В длинной серии статистических рядов, исследованных Лексисом, ни в одном случае не получилось сверхнормальной устойчивости, иначе говоря, та степень устойчивости, которая характерна для закона ошибок Гаусса, имеет по Лексису значение максимальной. В подавляющем большинстве случаев устойчивость эмпирических рядов оказалась ниже нормы, зачастую ус-

тупая последней в 5, 10 и даже в 20 раз.

Экспериментальные результаты, полученные Лексисом и его многочисленными последователями, были в дальнейшем пересмотрены А.А. Чупровым, которому удалось показать, что в явлениях действительной жизни ряды со сверхнормальной устойчивостью встречаются достаточно часто. Примечательно, что этот факт был им обнаружен на лингвистическом материале – распределении букв в сериях выборок из произведений Гете и распределении дактилей в гекзаметрах Овидия и Вергилия. Если объединить точки зрения Чупрова и Лексиса, то статистические ряды, построенные на материале реальных явлений, можно условно разделить на четыре группы: 1) ряды с наднормальной устойчивостью ($Q^2<1$), 2) ряды с нормальной устойчивостью ($Q^2\approx1$), 3) ряды с поднормальной устойчивостью ($Q^2\leq 2$),4) ряды со сверхподнормальной устойчивостью ($Q^2>2$).

11. Противопоставление "симметрия – асимметрия"

Встречи с симметричным распределением в лингвистике сравнительно редки. В подавляющем большинстве случаев исследователь имеет дело с умеренно и крайне асимметричными распределениями. Та или иная степень асимметрии может возникать вследствие 1) соединения в одной совокупности качественно неоднородных элементов или групп элементов, 2) наличия связей между единицами совокупности, 3) ограниченности начальных или крайних значений варьирующего признака каким-либо числом, 4) принадлежности лингвистической единицы к конкретному лингвистическому уровню или лексикограмматическому классу, а также по другим причинам, с трудом поддающимся систематическому учету.

12. Противопоставление "одновершинность – многовершинность"

Наряду с одновершинными (одномодальными) в лингвостатистических исследованиях часто возникают многовершинные (полимодальные) вариационные ряды. Появление многовершинности ряда почти всегда указывает на неоднородность состава того коллектива, который дал материал для построения ряда. При этом распределению некоторых явлений свойственна полимодальность по существу особенностей их строения. Впервые на одно из таких лингвистических явлений обратил внимание известный русский статистик П.А. Бихляев. На материале произведений

Тацита, Цезаря, Данте и Леопарда он показал, что распределения словоупотреблений по размеру всегда двухвершинны. Причина этого явления, по мнению Бихляева, "кроется в самом свойстве языка, который слагается из коротких частей - предлогов, союзов - и более длинных - прилагательных и существительных (Бихляев П.А., 1928, с. 163).

Полимодальность лингвостатистических рядов иногда имеет скрытый характер. Это обычно наблюдается в распределениях с большим размахом вариации, производящих впечатление крайне асимметричных. Некоторые не слишком откровенные нерегулярности в поведении таких распределений обычно списываются на счет ограниченности объема выборки.

Подведем некоторые итоги. Каждое эмпирическое распределение может быть "заиндексировано" с помощью наименований типов, выделенных в совокупности бинарных противопоставлений. Попытаемся это сделать на материале четырех распределений: 1) распределения словоупотреблений по частоте, 2) распределения словоупотреблений по размеру, 3) распределения союза "и" в сериях выборок, 4) распределения размеров терминоупотреблений. Из числа "индексирующих" исключим те аспекты, которые являются общими для перечисленных распределений, а также те, которые связаны со способами представления экспериментальных данных.

Результаты индексирования, показанные в табл. 4, нуждаются в некоторых пояснениях.

Распределения, аналогичные по смыслу тому, "статистический образ" которого приведен в первых двух столбцах табл.4, широко обсуждаются в длинном перечне общественных и естественных дисциплин: например, биологи строят распределения биологических видов по численности популяций, геохимики - распределения химических элементов по их содержанию в различных средах, науковеды - распределения ученых по научной продуктивности и т.п.

Среди статистиков нет единодушия в отношении распределений этого типа. Согласно наиболее распространенной концепции все они являются крайне асимметричными распределениями с "патологически" высокой вариацией признака, свойственной негауссовым распределениям гиперболического типа (см., например, Яблонский А.И., 1975). Вторая, менее распространенная точка зрения, которой придерживается и автор настоящей работы (Мартыненко Г.Я., 1978), заключается в том, что совокупности, дающие материал для построения многопредметных ста-

Таблица 4

Результаты индексирования эмпирических распределений

Противопоставления	Распределение частот словоупотреблений		Распределение размеров словоупотреблений	Распределение размеров терминоупотреблений	Распределение союза "и" в серии выборок
	концепция 1	концепция 2			
1	2	3	4	5	6
Гауссовость-негауссовость	Негауссово	Сумма двух гауссовых распределений	Сумма двух гауссовых распределений	Гауссово	Гауссово
Строение-поведение	Статусное	Статусное	Структурное	Структурное	Статусное
Однопредметность-многопредметность	Многопредметное	Многопредметное	Многопредметное	Многопредметное	Однопредметное
Устойчивость-неустойчивость	Сверхподнормальная устойчивость	Со сверхподнормальной устойчивостью	С подпормальной устойчивостью	С подпормальной устойчивостью	Однородное
Однородность-неоднородность	Однородное	Неоднородное, с априорно неустранимой неоднородностью	Неоднородное, с априорно не устранимой неоднородностью	Однородное	С нормальной устойчивостью
Симметрия - асимметрия	Крайне асимметричное	-	-	Умеренно асимметричное	Симметричное
Одновершинность-многовершинность	Одновершинное	Двухвершинное	Двухвершинное	-	Одновершинное

тусных распределений, качественно неоднородны. Такие совокупности могут рассматриваться как системы, состоящие из пластичной, подверженной изменениям части и относительно жесткой части, обеспечивающей сохраняемость целостности системы. Естественно полагать, что каждая из этих частей должна подчиняться своему закону распределения. На эту точку зрения частично наводят данные колонки 3 табл. 4, хорошо коррелирующие с данными колонки 2.

В заключение отметим, что список предложенных нами бинарных противопоставлений и соответствующих им типов распределений не является исчерпывающим. Вне поля нашего рассмотрения остались, например, такие аспекты, как "дискретность - непрерывность", "частота - распространенность", "редкость - частость", "статика - динамика, "одномерность - многомерность" и некоторые другие противопоставления.

Избранный нами способ систематизации не устраняет частичной "синонимии" типов распределений, обусловленной в основном причинно-следственными отношениями между бинарными противопоставлениями. Так, асимметрия распределения и тем более его многовершинность могут рассматриваться как внешнее (геометрическое) проявление того или иного варианта качественной неоднородности, а степень устойчивости - как количественное выражение неоднородности.

Из сказанного следует, что при совершенствовании предложенной схемы необходимо расширить список бинарных противопоставлений, отобрать из них наиболее существенные и изучить характер причинно-следственных отношений между ними.

Л И Т Е Р А Т У Р А

Алексеев П.М. Квантитативная типология текста. Автореферат. докт. диссертации. - Л., 1977.

Белый А. Мастерство Гоголя. - М.-Л.: ОГИЗ, 1934.

Вентцель Е.С. Теория вероятностей. - М.: Физматгиз, 1962.

Вихляев П.А. Очерки теоретической статистики. - М.: Новый агроном, 1928.

Кауфман А.А. Теория статистики. - М.: Типография Т-ва И.Д. Сытина, 1909.

Колмогоров А.Н. Математическая статистика. БСЭ, т. 26, М., 1954.

Кравец А.С. Природа вероятности (философские аспекты). – М.: Мысль, 1976.

Мартыненко Г.Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах. – В кн.: Структурная и прикладная лингвистика. Вып. I. – Л.: Изд-во ЛГУ, 1978, с. 63-79.

Марусенко М.А. Об однородности эмпирически выделяемых подъязыков и закона распределения сложных научно-технических терминов. – Научно-техническая информация. Серия 2. М., 1980, № 8, с. I-6.

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

Смирнов С.Н. Элементы философского понятия "система" как ступени развития познания и общественной практики. – В кн.: Системный анализ и научное знание. – М.: Наука, 1978, с. 60-82.

Смолуховский М. О понятии случайности и о происхождении законов вероятностей в физике. – Успехи физических наук. М., 1927, т. УП, вып. 5, с. 331-332.

Тулдава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. – В кн.: Лингвостатистика и квантитативные закономерности текста. Труды по лингвостатистике. Вып. 6. Тарту, 1980, с. 113-144.

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. – М.: Статистика , 1980.

Чупров А.А. Очерки по теории статистики. – С.Петербург: Издание М. и С. Сабашниковых, 1910.

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

Lexis W. Abhandlungen. Zur Theorie der Bevölkerungs- und Moralstatistik. Jena, 1903.

Carrol J.B. On Sampling from a Lognormal Model of Word-Frequency Distribution. – In: H. Kučera, W.N.Francis. Computational Analysis of Present-Day American English. Providence, 1967.

Lévy P. Calcul de probabilités. Part. 2, Chap. 6. Gauthier-Villars, 1925.

TYPOLOGY OF STATISTICAL DISTRIBUTIONS IN LINGUISTICS

Grigory Martynenko

S u m m a r y

The paper deals with a set of linguo-statistical distributions. The set presents the classification of the main aspects. Each aspect is composed as a binary opposition, e. g. "structure-behaviour", 'stability-instability", "virtuality-actuality", "homogeneity-heterogeneity", "variety-restriction of variety" etc.

The specific forms of distributions are identified due to the tendency of real distributions to one of the external terms of a binary opposition within each aspect. Thus, the status distributions and structural distributions are determined within "structure-behaviour", those of both a priori eliminable and ineliminable heterogeneity are obtained within "homogeneity-heterogeneity" and so on.

ЭМПИРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ЧАСТОТНОСТИ ФОНЕМ В КАЗЫМСКОМ ДИАЛЕКТЕ ХАНТЫЙСКОГО ЯЗЫКА

Ю.А. Тамбовцев

В последнее время в финно-угроведении помимо других методов лингвистического анализа все больше и больше начинают применяться методы фоностатистического анализа. Они затрагивают уже не только языки, имеющие большое количество говорящих, но и языки малых финно-угорских народов. Одним из таких языков является язык казымских ханты. Целью данной статьи как раз и является анализ эмпирического распределения частотности фонем казымского диалекта хантыйского языка. Частотное распределение фонем в речи выявляет специфические особенности каждого языка и помогает определить его близость к другим генетически или территориально соседним языкам. В работе рассматриваются частоты фонем в потоке речи, которые названы абсолютной частотностью фонем, а также частотность фонем в начале и в конце казымского слова. Сравниваются аналогичные фоностатистические характеристики, полученные на материале генетически и территориально близкого казымскому северном диалекте мансийского языка.

В качестве материала послужили тексты К. Редеи, включающие бытовые описания, рассказы, песни и игры казымских ханты (Rédei, 1968). Обработка материала происходила на ЭВМ ЕС-1033 ВЦ НГУ. Объем введенного текста составил 74 762 фонемы. Необходимо отметить, что этот же материал, но меньшего объема (24 000) уже подвергался фоностатистическому обследованию Г.Г. Куркиной, с данными которой будут сопоставляться полученные нами частотности гласных (Куркина, 1980). При этом в соответствии со статистическими законами эталонными значениями следует признать в случае расхождений данные, полученные на большей выборке (Гнеденко, 1969). В данном случае эталонной выборкой будет считаться выборка в 74 762 фонемы, так как она более, чем в три раза больше выборки Г.Г. Куркиной. Следует отметить, что при небольших выборках частоты фонем носят необъективный характер и не являются постоянными (Тамбовцев, 1980). Эту закономерность колебаний легко уви-

деть, постепенно увеличивая объем выборки и при этом отмечая величины фонемных частот (Тамбовцев, Утев, 1981). Только при значительном объеме выборки они стабилизируются (Вентцель, 1964). Иными словами, в случайных языковых явлениях имеет место некоторое объективно существующее свойство, которое имеет тенденцию оставаться постоянным и проявляется все яснее при увеличении объема исследуемого материала (Пиотровский, Бектаев, Пиотровская, 1977). Известный фонолог Д.Н.Сегал, например, прямо замечает, что в тех случаях, когда статистика основывается на единичной (притом малой) выборке, ее с трудом можно признать значимой (Сегал, 1972). При выборе объема исходного материала лингвист должен иметь в виду, что даже применение точнейших методов математической статистики не делает возможным применение малой выборки там, где необходима большая (Боярский, 1961). В языке казымских ханты *ми*, также как и Карой Редеи, выделяем 9 гласных фонем. /*ă*, *ē*, *a*, *i*, *ɔ*, *ō*, *o*, *u*, *ə*/. Что касается согласных фонем, то в отличие от К. Редеи у нас их меньше, так как вслед за Л. Верте (Верте, 1979) мы считаем фонемы /*ć*/ и /*č*/ только оттенками и не включаем их в список самостоятельных фонем. Таким образом выделяется 18 согласных фонем: /*p*, *t*, *t'*, *k*, *x*, *m*, *n*, *ń*, *ŋ*, *s*, *ś*, *š*, *w*, *j*, *r*, *l*, *ł*, *l'*/. Следовательно, в казымском диалекте хантыйского языка всего 27 фонем, столько же, сколько и в северном диалекте мансийского языка (Тамбовцев, 1977, 1979).

Данные подсчета частотностей фонем сведены в таблицы, в которых частота дается в процентах к различным величинам. Например, в таб. 1 частота дается в процентах ко всем фонемам в потоке речи, а в таб. 2 только по отношению к гласным, в то время как в таб. 3 – только по отношению к согласным и т.д. На это нужно обращать внимание в дальнейшем при сравнении различных величин в процентах, чтобы сравнивать только сопоставимые величины (Пфанцагль, 1976).

Изучая частотное распределение фонем в таб. 1, можно заметить большое сходство этого распределения с аналогичным распределением в мансийском языке (Тамбовцев, 1977, 1979, 1981). В обоих языках самой частотной согласной является фонема /*t*/, а среди гласных это фонемы /*a*/ и /*i*/ в мансийском и /*ə*/, /*ă*/ и /*i*/ в хантыйском языках. Интересно заметить, что в некоторых других финно-угорских и самодийских языках в упорядоченных рядах /*a*/ всегда стоит перед /*i*/. Это можно наблюдать во всех языках, по которым получена ла-

дежная фонологическая статистика: финскому (Setälä, 1972, 9), эстонскому (Tuldava, 1980, 83), тихвинскому диалекту карельского языка (по нашим неопубликованным данным), венгерскому (Jékel, Papp, 1974), коми-зырянскому (по нашим неопубликованным данным), эрзя- и мокша-мордовскому (Veenker, 1981), по горно- и луговому-марийскому (Veenker, 1980), а также по селькупскому (Морев, 1973) и ненецкому (Попова, 1978). Кроме того, по данным В. Феэнкера (Феэнкер, 1981) фоностатистическая обработка средне-обского текста (средне-обской диалект хантыйского языка) тоже дает порядок /а/ - /и/. Вероятно, такое распределение, когда /и/ следует за /а/, а не наоборот, в какой-то мере присуще финно-угорским языкам. Более того, тот факт, что в казымском диалекте хантыйского языка частотность /а/ больше частотности /и/, подтверждается и историческим развитием финно-угорских и самодийских языков. Это отмечается известным финно-угроведом Б.А. Серебренниковым, который пишет, что на огромном протяжении от Норвегии до нижнего течения Оби в целом ряде языков наблюдается тенденция к превращению гласных звуков разного качества в гласный /а/. Такое явление он наблюдает в саамском, ненецком, северных диалектах мансийского языка, а также в хантыйском (Серебренников, 1965). В то же время по данным Г.Г. Куркиной эти гласные следуют в упорядоченном ряду в обратном порядке, т.е. /и/ предшествует /а/. По-видимому, в эти данные вкралась ошибка, так как выше описано, что для финно-угорских языков характерен обратный порядок, т.е. /а/ - /и/. Может быть также, что в данном случае негативно проявил себя недостаточный объем выборки, что в итоге привело к неверным результатам. Общий порядок следования фонем в упорядоченном ряду и их значения также не совпадают. Расхождения существуют и в сумме долгих и кратких гласных. По нашим данным суммы долгих и кратких гласных соотносятся как 40,06:59,04. Огубленные гласные в казымском диалекте хантыйского языка среди всех гласных составляют 23,77%, а в мансийском языке этот показатель равен 25,11%. Для того, чтобы понять много это или мало и на каком расстоянии отстоят друг от друга эти два языка, исходя из такой важной фонологической характеристики как лабиализация (Зиндер, 1979), заметим, что, например, коми-зырянский язык имеет эту характеристику, равную 17,32%, а людиковский диалект карельского языка (Баранцев, 1975) - 31,25%. Или, например, русский язык имеет в сумме 16,19% лабиализованных гласных (Елкина, Юдина, 1964), а родственный

ему польский - 34,74% (Сегал, 1972). В свете этих данных становится совершенно очевидным, что по сумме употребления лабиализованных гласных в речи северный диалект мансийского языка и казымский диалект хантыйского языка очень близки.

Теперь рассмотрим мелодичность языка казымских ханты. Под мелодичностью, также как и в мансийском языке (Тамбовцев, 1977), понимается суммарная частотность в речи гласных и сонорных. В казымском диалекте хантыйского языка она составила 71,33%, в мансийском - 73,70%. Оба эти языка следует признать мелодичными, так как другие языки имеют несколько меньшую сумму гласных и сонорных, например, в русском она - 61,25%, несколько меньше в болгарском - 60,45% (Маринова, Маринов, 1964), и еще меньше в чешском - 53,40% (Ludvíková, Königova, 1967).

В своей работе Г.Г. Куркина пишет, что язык казымских ханты имеет относительно низкую вокалическую насыщенность, приводя следующие цифры: 40,6% гласных и 59,4% согласных, или 1000 гласных на 1464 согласных. По нашим данным насыщенность казымской речи согласными несколько выше: 40,37% гласных и 59,63% согласных, или на 1000 гласных 1477 согласных. Тем не менее, в отличие от Г.Г. Куркиной мы полагаем, что даже при этом условии вокалическую насыщенность языка казымских ханты нельзя считать низкой. Известно, что многие языки мира имеют на 1000 гласных 1700 и более согласных (Никонов, 1966, 1972). Именно такие языки в нашем понимании скорее следует относить к языкам с довольно низкой вокалической насыщенностью. Так, например, латышский язык с соотношением 1702 согласных на 1000 гласных нужно отнести к языкам с низкой вокалической насыщенностью. Языки же с соотношением согласных и гласных такого порядка как хантыйский или мансийский, по нашему мнению, следует относить к языкам со средней вокалической насыщенностью, см., например, данные по 56 языкам (Чистяков, 1960).

При анализе частотного функционирования фонем в языке и речи следует особое внимание обращать на то, какие фонемы чаще употребляются в маргинальных позициях, особо выделяя фонемы, характерные для абсолютного начала слова и дифференцируя их от характерных фонем абсолютного конца слова. Можно предположить, что в прошлом, так же как и сейчас (и, вероятно, так же как в будущем) носители языка подсознательно действовали и действуют таким образом, чтобы в эти ключевые, т.е. маргинальные, позиции ставились такие фонемы, распозна-

вание которых в процессе коммуникации незатруднено. Причем в позиции начала могут попадать не такие фонемы, как в позиции конца, таким образом, носителям языка будет легче распознавать начала и концы слов, и, следовательно, процесс речевой коммуникации будет протекать легче. Эту тенденцию направления развития языка в сторону улучшения процесса коммуникации понимают многие лингвисты, так, например, еще в начале века немецкий ученый Б. Хорн обращал внимание на это, говоря, что язык служит во многом цели достижения взаимопонимания, и в его развитии существует тенденция возможно лучше достигать этой цели (Horn, 1923).

Выявление закономерных явлений начала и конца слова в финно-угорских языках представляет особый интерес еще и потому, что они относятся к языкам агглютинативного типа, следовательно, начало и конец слова несут различные грамматические функции. Важность изучения характерного употребления фонем в абсолютном начале и конце в финно-угорских языках подчеркивали многие лингвисты (Лыткин, 1968; Серебренников, 1974; Хакулинен, 1953; Itkonen, 1965; Molnar, 1974 и другие).

Перейдем теперь непосредственно к рассмотрению фонем абсолютного начала и конца слова в языке казымских ханты. Начнем анализ с рассмотрения частотности фонем в абсолютном начале слова (табл. 4). Сразу же бросается в глаза, что наиболее характерными фонемами начала являются не гласные, а согласные. Первая гласная, которой является /i/, стоит в упорядоченном ряду фонем абсолютного начала только на II-м месте. В сумме согласные этого ряда составляют 84,38%, а гласные более, чем в 4 раза меньше, только лишь 15,62%. Это означает, что в потоке казымской речи более, чем в 4 раза вероятнее встретить слово, начинающееся с согласной, чем с гласной.

Сравнивая полученный нами упорядоченный ряд с аналогичным рядом гласных, построенным по данным Г.И. Куркиной, обнаруживаем, что и здесь имеются расхождения. Во-первых, порядок фонем несколько иной и, во-вторых, числовые значения частотности фонем, даже стоящих на одних и тех же местах в этих упорядоченных рядах, различны. Забегая несколько вперед, отметим, что почти такие же расхождения существуют между нашим упорядоченным рядом фонем абсолютного конца и аналогичным рядом Г.Г. Куркиной. Вероятно, что небольшой объем выборки, взятый упомянутым автором, не дает надежных результатов и при анализе маргинальных фонем.

Сопоставив десять самых частотных казымских с десятью самыми частотными северномансийскими фонемами абсолютного начала слова, приходим к выводу, что это одни и те же фонемы, т.е. налицо взаимнооднозначное соответствие девяти из десяти элементов в этих двух рядах, кроме одной хантыйской фонемы /ś/, которая соответствует в мансийском ряду фонеме /s/. Все же остальные фонемы легко идентифицируются: (отождествляются) /x, w, p, t, j, m, k, ɲ, ŋ/. Так же, как в хантыйском, в мансийском упорядоченном ряду среди первых I2 фонем почти нет гласных, а на I3 месте стоит одна и та же фонема – /ɑ/. Очень близки по величине суммы согласных и гласных в хантыйском и мансийском рядах: 84,38% и I5,62% против 79,83% и 20,I7%.

Рассуждая о близости языков, известный немецкий лингвист Г. Дерфер замечает, что совпадения некоторых языковых характеристик могут говорить о том, что а) языки генетически связаны; б) большое влияние оказали многочисленные заимствования; в) оба языка имели один и тот же субстрат (Doerfer, I98I). На наш взгляд, в данном случае схожесть фонемного функционирования в потоке речи как минимум является следствием первых двух пунктов Г. Дерфера, и есть вероятность, что схожие фоностатистические характеристики этих двух языков могут быть следствием всех трех указанных пунктов. Последующий анализ абсолютного исхода слов языка казымских ханты и языка северных манси только подтверждает этот вывод. В данной статье мы не будем подробно останавливаться на сопоставлении характерных особенностей частотного функционирования фонем в абсолютном исходе хантыйского и мансийского слова в связи с тем, что это уже сделано в другой работе (Тамбовцев, в печати), но постараемся более детально рассмотреть закономерности и характерные особенности употребления фонем в начале и конце слова в казымском диалекте хантыйского языка (таб. I0).

Совершенно закономерно бы было предположить, что фонемы имеют разные частотности в начале и конце слова. Для того, чтобы проверить эту гипотезу построим новую таблицу (таб.I0), в которой все частотности будут даны по отношению ко всем фонемам звуковой последовательности с тем, чтобы все частотности было бы можно сравнивать не нарушая закона соизмеримости.

В связи с тем, что фонемы /ɑ/ и /o/ не встречаются в конце слов, а фонема /ŋ/ не встречается в начале слов, эти

фонемы выпадают из нашего анализа, так как с точки зрения коммуникации совершенно ясно, что носитель казымского диалекта хантыйского языка, услышав фонемы /а/ и /о/ никогда не станет предполагать конец слова, а услышав фонему /ŋ/ сразу же идентифицирует конец слова. Что касается фонем, встречающихся и в начале и в конце слова, то в этом случае, вероятно, носитель языка полагается на интуитивно чувствуемую им фонемную частотность начала и конца слова. Так, например, только две фонемы /ē/ и /t'/ употребляются в начале и конце слова с одинаковой вероятностью, все остальные фонемы совершенно очевидно дифференцируются по вероятности употребления в конце против употребления в начале или наоборот. Расположив все фонемы в упорядоченный ряд по величине вероятности их употребления в совершенно определенной маргинальной позиции, заметим, что более половины фонем этого ряда употребляются в одной из позиций в 4 раза чаще, а три четверти их встречаются в определенной позиции чаще более, чем в три раза (таолица II).

Выводы

1. При фоностатических исследованиях речи огромную роль играет правильный выбор объема вводимого в ЭВМ материала. На небольших или недостаточно больших выборках результаты фоностатистического исследования будут или неверны или недостаточно объективны.

2. Не все лингвисты, применяющие фоностатистические методы, учитывают объем вводимого материала.

3. По частоте встречаемости в речи огубленных гласных казымский диалект хантыйского языка ближе северному диалекту мансийского языка, чем комизырянскому или карельскому языкам.

4. По сумме гласных и сонорных согласных казымский диалект хантыйского языка следует признать мелодичным.

5. Язык казымских ханты следует отнести к языкам со средней вокалической насыщенностью.

6. Наиболее характерными фонемами начала слова языка казымских ханты являются согласные.

7. Частотное распределение фонем начала казымского и мансийского слова очень похоже. Это, так же как и сходство частотного распределения фонем в казымской и мансийской речи, может говорить о близости двух данных финно-угорских

языков. В данном случае эта близость выводится с помощью фоностатистики.

8. Фонемы языка казымских ханты имеют различные частотные характеристики в начале и конце слова, что облегчает процесс коммуникации.

Таблица I

Абсолютная частотность гласных и согласных фонем казымского диалекта хантыйского языка

I.	/t/	8429	II,27%
2.	/ə/	7453	9,97
3.	/ā/	53I5	7,II
4.	/i/	4798	6,42
5.	/ʌ/	4I2I	5,5I
6.	/n/	4053	5,42
7.	/m/	3830	5,I2
8.	/j/	3376	4,52
9.	/ē/	3264	4,37
I0.	/x/	3I58	4,22
II.	/s/	3I29	4,I9
I2.	/w/	2973	3,98
I3.	/p/	2602	3,48
I4.	/a/	2I78	2,9I
I5.	/r/	I978	2,65
I6.	/o/	I96I	2,62
I7.	/ō/	I8I7	2,43
I8.	/u/	I702	2,28
I9.	/k/	I700	2,27
20.	/ɔ/	I694	2,27
2I.	/ŋ/	I58I	2,I2
22.	/ś/	I353	I,8I
23.	/š/	94I	I,26
24.	/ń/	563	0,75
25.	/l/	399	0,53
26.	/ʌ'/	272	0,36
27.	/t'/	I22	0,I6
		74 762	I00,00%

Таблица 2

Абсолютная частотность гласных по отношению только к гласным

I.	/ə/	7453	24,69%
2.	/ā/	53I5	I7,6I
3.	/i/	4798	I5,90
4.	/ē/	3264	I0,8I
5.	/a/	2I78	7,22
6.	/o/	I96I	6,50
7.	/ō/	I8I7	6,02
8.	/u/	I702	5,64
9.	/ɔ/	I694	5,6I
		30 I82	I00,00%

Таблица 3

Абсолютная частотность согласных по отношению только к согласным

I.	/t/	8429	I8,9I%
2.	/ʌ/	4I2I	9,24
3.	/n/	4053	9,09
4.	/m/	3830	8,59
5.	/j/	3376	7,57
6.	/x/	3I58	7,08
7.	/s/	3I29	7,02
8.	/w/	2973	6,67
9.	/p/	2602	5,82
I0.	/r/	I978	4,47
II.	/k/	I700	3,8I
I2.	/ŋ/	I58I	3,55
I3.	/ś/	I353	3,03
I4.	/š/	94I	2,II
I5.	/ń/	563	I,26
I6.	/l/	399	0,90
I7.	/ʌ'/	272	0,6I
I8.	/t'/	I22	0,27
		44 580	I00 00%

Таблица 4

Частотность гласных и согласных фонем начала казымского слова

I.	/x/	I7I5	II,I3%
2.	/w/	I59I	I0,33
3.	/p/	I544	I0,02
4.	/t/	I5I8	9,85
5.	/j/	I478	9,59
6.	/m/	I324	8,59
7.	/k/	694	4,50
8.	/ś/	570	3,70
9.	/л/	557	3,62
I0.	/n/	556	3,6I
II.	/i/	555	3,60
I2.	/s/	532	3,45
I3.	/ā/	5I3	3,33
I4.	/ē/	358	2,32
I5.	/a/	348	2,26
I6.	/š/	327	2,I2
I7.	/ō/	278	I,80
I8.	/ń/	257	I,67
I9.	/ę/	20I	I,30
20.	/o/	I53	0,99
2I.	/u/	I48	0,96
22.	/r/	I04	0,68
23.	/ǭ/	54	0,35
24.	/л'/	I8	0,I2
25.	/t'/	I6	0,I0
		I5 409	99,99%

Таблица 5

Гласные фонемы начала слова по отношению только к гласным начала

I.	/i/	555	23,06%
2.	/ā/	5I3	2I,3I
3.	/ē/	358	I4,87
4.	/a/	348	I4,46
5.	/ō/	278	II,55
6.	/o/	I53	6,38
7.	/u/	I48	6,I5
8.	/ǭ/	54	2,24
		2407	I00,00%

Таблица 6

Согласные фонемы начала слова по отношению только к согласным начала

I.	/x/	I7I5	I3,I9%
2.	/w/	I59I	I2,24
3.	/p/	I544	II,86
4.	/t/	I5I8	II,68
5.	/j/	I478	II,37
6.	/m/	I324	I0,I8
7.	/k/	694	5,34
8.	/ś/	570	4,38
9.	/л/	557	4,28
I0.	/n/	556	4,28
II.	/s/	532	4,09
I2.	/š/	327	2,52
I3.	/ń/	257	I,98
I4.	/ę/	20I	I,55
I5.	/r/	I04	0,80
I6.	/л'/	I8	0,I4
I7.	/t'/	I6	0,I2
		I3 002	I00,00%

Таблица 7

Частотность гласных и согласных конца слова каз. диал. хант. языка

I.	/t/	2378	I5,34%
2.	/a/	2347	I5,23
3.	/n/	2I50	I3,95
4.	/i/	2054	I3,33
5.	/ʌ/	I052	6,83
6.	/m/	963	6,25
7.	/s/	958	6,22
8.	/ŋ/	530	3,44
9.	/w/	5II	3,32
I0.	/j/	467	3,03
II.	/r/	427	2,77
I2.	/ē/	372	2,4I
I3.	/x/	249	I,62
I4.	/p/	223	I,45
I5.	/ś/	I73	I,I2
I6.	/k/	I72	I,I2
I7.	/q̄/	I67	I,08
I8.	/š/	III	0,72
I9.	/ń/	34	0,22
20.	/ʌ'/	25	0,I6
2I.	/u/	I8	0,I2
22.	/t'/	I5	0,I0
23.	/ō/	9	0,06
24.	/e/	4	0,03
		I5 409	I00,00%

Таблица 8

Гласные конца слова по отношению только к гласным конца слова

I.	/ā/	2347	47,25%
2.	/i/	2054	4I,35
3.	/ē/	372	7,49
4.	/q̄/	I67	3,36
5.	/u/	I8	0,36
6.	/ō/	9	0,I8
		4967	99,99%

Таблица 9

Согласные конца слова по отношению только к согласным конца слова

I.	/t/	2378	22,77%
2.	/n/	2I50	20,59
3.	/ʌ/	I052	I0,07
4.	/m/	963	9,22
5.	/s/	958	9,I7
6.	/ŋ/	530	5,08
7.	/w/	5II	4,89
8.	/j/	467	4,47
9.	/r/	427	4,09
I0.	/x/	249	2,38
II.	/p/	223	2,I4
I2.	/ś/	I73	I,66
I3.	/k/	I72	I,65
I4.	/š/	III	I,06
I5.	/ń/	34	0,33
I6.	/ʌ'/	25	0,24
I7.	/t'/	I5	0,I4
I8.	/e/	4	0,04
		I0 442	99,99%

Таблица I0

Некоторые характеристики фонем начала и конца казымского слова с точки зрения частот их встречаемости в речи

I	2	3	4	5	6	7	8	9
I.	x	2,29	I3	0,33	-I2	+I,96	6,9	6,9
2.	w	2,I3	9	0,68	- 7	+I,45	3,I	3,I
3.	p	2,06	I4	0,30	-II	+I,76	6,8	6,8
4.	t	2,03	I	3,I8	+ 3	-I,I5	0,6	I,6
5.	j	I,98	I0	0,62	- 5	+I,36	3,2	3,2
6.	m	I,77	6	I,29	0	+0,48	I,4	I,4
7.	к	0,93	I6	0,23	- 9	+0,70	4,0	4,0
8.	ś	0,76	I5	0,23	- 7	+0,53	3,3	3,3
9.	л	0,7[illegible]	5	I,4I	+ 4	-0,67	0,5	I,9
I0.	n	0,74	3	2,88	+ 7	-2,I4	0,3	3,9
II.	ĺ	0,74	4	2,75	+ 7	-2,0I	0,3	3,7
I2.	s	0,7I	7	I,28	+ 5	-0,57	0,6	I,8
I3.	ă	0,69	2	3,I4	+II	-2,45	0,2	4,6
I4.	ĕ	0,48	I2	0,50	+ 2	-0,02	I,0	I,0
I5.	ś	0,44	I8	0,I5	- 2	+0,29	2,9	2,9
I6.	ə	0,37	23	0,0I	- 6	+0,36	37,0	37,0
I7.	ń	0,34	I9	0,05	- I	+0,29	6,8	6,8
I8.	ĺ	0,27	24	0,005	- 5	+0,265	54,0	54,0
I9.	ш	0,20	2I	0,02	0	+0,I8	I0,0	I0,0
20.	ч	0,I4	II	0,57	+II	-0,43	0,2	4,I
2I.	ӈ	0,07	I7	0,22	+ 7	-0,I5	0,3	3,I
22.	л'	0,02	20	0,03	+ 5	-0,0I	0,7	I,5
23.	t'	0,02	22	0,02	+ 4	0,00	I,0	I,0
24.	а	0,47	-	-	-	-	-	-
25.	о	0,20	-	-	-	-	-	-
26.	ŋ	-	8	0,7I	-	-	-	-

I. Номер фонемы в упорядоченном ряду начала
2. Символ фонемы
3. Величина частотности фонем начала
4. Номер в упорядоченном ряду конца
5. Величина частотности фонем конца
6. Разница в номере места между номером в упорядоченном ряду начала и в упорядоченном ряду конца

7. Разница между величиной частотности фонемы в начале и конце слова
8. Отношение величины частотности начала к величине частотности конца
9. Во сколько раз фонема употребляется чаще в какой-либо маргинальной позиции.

Таблица II

Вероятности появления определенной фонемы в определенной маргинальной позиции в казымском слове

I.	e	-54,0	I0.	k	- 4,0	I9.	t	- I,6
2.	$ə$	-37,0	II.	n	- 3,9	20.	n'	- I,5
3.	u	-I0,0	I2.	i	- 3,7	2I.	m	- I,4
4.	$š$	- 9,9	I3.	$ś$	- 3,3	22.	$ē$	- I,0
5.	x	- 6,9	I4.	$ʃ$	- 3,2	23.	t'	- I,0
6.	$ń$	- 6,8	I5.	w	- 3,I	24.	o	- ∞
7.	p	- 6,8	I6.	$ā$	- 3,I	25.	a	- ∞
8.	$ā$	- 4,6	I7.	l	- I,9	26.	$ŋ$	- ∞
9.	r	- 4,I	I8.	s	- I,8			

Л И Т Е Р А Т У Р А

Баранцев А.П. Фонологические средства людиковской речи. – Л.: Наука, I975, с. I78-I79.

Боярский А.Я. Математика для экономистов. – М.: Госстатизд, I96I.

Вентцель Е.С. Теория вероятностей. – М.: Наука, I964.

Верте Л.А. Некоторые дистрибутивные характеристики согласных фонем казымского диалекта хантыйского языка. – В кн.: Исследование звуковых систем сибирских языков. Новосибирск, I979.

Гнеденко Б.В. Курс теории вероятностей. – М.: Наука, I969.

Зиндер Л.Р. Общая фонетика. – М.: Высшая школа, I979.

Куркина Г.Г. Относительная частотность гласных в языке казымских ханты.–В кн.: Звуковой строй сибирских языков. Новосибирск, I980, с. 66-7I.

Лыткин В.И. К вопросу о конечных гласных финно-угорского праязыка. – СФУ, IУ, I968, с. 233-238.

Маринова М., Маринов А. Статистически изследования на фонемите в българския книжовен език.– В кн.: Български език, I964, 2-3, с. I73-I79.

Морев Ю.А. Звуковой строй среднеобского (ласкинского) говора селькупского языка. АКД. Томск, 1973.

Никонов В.А. Глоттогенез Сибири и Дальнего Востока в свете фоностатистики. – В кн.: Происхождение аборигенов Сибири и их языков. Томск, 1976, с. 41–42.

Фоностатистическое измерение межъязыковых расстояний.–В кн.: Исследования по фонологии. М., 1966, с. 285–286.

Консонантный коэффициент. – Lingua Posnaniensis, 1972, с. 44–48.

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

Попова Я.Н. Фонетические особенности лесного наречия ненецкого языка. – М.: Наука, 1978, с. 128–163.

Шфанцагль И. Теория измерений. – М.: Мир, 1976.

Сегал Д.М. Основы фонологической статистики. – М.: Наука, 1972, с. 137–138.

Серебренников Б.А. О некоторых закономерных явлениях начала и конца слова в уральских языках. – СФУ, X, 1974, с.151–157.

Тамбовцев Ю.А. Некоторые характеристики распределения фонем мансийского языка. – СФУ, XШ, 1977, с. 195–198.

Тамбовцев Ю.А. Распределение гласных фонем в мансийской поэзии. – СФУ, XУ, 1979, с. 164–167.

Тамбовцев Ю.А. Частотные характеристики гласных первого слога мансийского языка. – В кн.: Звуковой строй сибирских языков. Новосибирск, 1980, с. 72–75.

Тамбовцев Ю.А. Закономерности частотного функционирования долгих и кратких гласных в ударных и неударных слогах мансийского слога. – СФУ, XУП, 1981.

Тамбовцев Ю.А., Утев С.А. Зависимость величины частот мансийских гласных 1-го слога от величины объема выборки. – В кн.: Теоретические вопросы фонетики и грамматики языков народов СССР. Новосибирск, 1981, с. 97–105.

Феэнкер В. Проблемы фонологической статистики хантыйского языка. – В кн.: Теоретические вопросы фонетики и грамматики языков народов СССР. Новосибирск, 1981, с. 84–88, 95.

Хакулинен Л. Развитие и структура финского языка. I. М..1953.

Чистяков В.Ф. Частотность гласных и согласных в 50 языках разного грамматического строя. – Lingua Posnaniensis, УШ, 1960.

Dörfer G. The Conditions for Proving the Genetic Relationship of Languages. - In: The Bulletin of the International Institute for Linguistic Sciences. Kyoto Sangyo University. Vol. II, No. 4, 1981, p. 39-58.

Horn W. Sprachkörper und Sprachfunction. Leipzig, 1923.

Itkonen T. Proto-Finnic Final Consonants. Helsinki, 1965.

Jékel P., Papp F. Ady Endre összes költői művéinek fonémastatisztikája, Budapest, 1974.

Ludvíková M., Kônigová M. Quantitative Research of Graphemes and Phonemes in Czech. - In: The Prague Bulletin of Mathematical Linguistics. Praha, 1967, 7, p. 15-29.

Molnar F.A. On the History of Word-Final Vowels in the Permian Languages. Szeged, 1974.

Redei K. Nord-ostjakische Texte (Kazym-Dialekt) mit Skizze der Grammatik. Göttingen, 1968.

Setälä V. Suomen kielen dynamiikka. Helsinki. 1972.

Tuldava, J. Eesti keele sõnavara foneetilis-grafeemilised mõõted. - В кн.: Лингвостатистические исследования по финно-угорским языкам. Труды по лингвистике 5. (Учен. зап. Тартуск. гос. ун-та, вып. 518). Тарту, 1980.

Veenker W. Zur phonologischen Statistik der čeremissischen (marischen) Schriftsprachen. - СФУ, XVI, 1980.

Veenker W. Zur phonologischen Statistik der mordvinischen Schriftsprachen. - In: Ural-Altaische Jahrbücher, Band I. Wiesbaden, 1981.

EMPIRICAL DISTRIBUTION OF PHONEMIC FREQUENCY IN THE KAZYM DIALECT OF HANTY

Yuri Tambovtsev

S u m m a r y

The article deals with the phonostatistical research of the Kazym dialect of Hanty texts collected by K. Redei. The volume of the sample comprises 74 762 phonemes. The article stresses the importance of the greater volumes, taken for phonostatistical investigations, and argues that some results of such studies are not reliable because of the insufficient volumes.

The data obtained after computing the Kazym dialect texts are compared to the data of Mansi. The article discusses the frequency of occurrence of the Kazym phonemes in speech: a) in any position of the word (this is called the absolute frequency); b) at the beginning of the word; c) at the end of the word. The article underlines the fact that the frequency values of the same phonemes at the beginning and the end of the word are different. This difference may help to distinguish the word beginning from the word ending in the stream of fluent Kazym speech during the process of communication.

The comparison of the Hanty and Mansi phonostatistical data allows us to conclude that these two languages are in a close genetic relation to each other. Eleven tables at the end of the article illustrate the discourse and the proofs.

КВАНТИТАТИВНОЕ ИССЛЕДОВАНИЕ ГЕНЕТИЧЕСКОГО СОСТАВА ЛЕКСИКИ ЭСТОНСКОГО ЯЗЫКА

Ю. Тулдава

В статье рассматриваются разные по своему происхождению пласты эстонской лексики как части словаря и текста. Анализ проводится на основе базового частотного словаря, который представляет собой своего рода квантитативную модель текста, отражающую употребление наиболее важных, актуальных слов в живой речи. Предварительно дается краткий исторический обзор развития эстонского языка, отдельно рассматриваются правила идентификации древних слов, проводится анализ их употребления по частям речи и лексическим группам и, наконец, особо рассматривается вопрос о системной связи между возрастом и частотностью слов.

Вводные замечания. Важное место в лексикологических исследованиях занимает вопрос о генетическом составе лексики данного языка вместе с соответствующими квантитативными характеристиками. Подсчет различных слоев лексики: исконного наследия, заимствований из различных языков и собственных образований позволяет сделать обоснованные выводы о происхождении и развитии языка, о культурных связях с другими народами в разные исторические периоды, об удельном весе различных генетических групп слов в современном языке. Особый интерес представляют древнейшие элементы лексики, которые передавались в течение столетий от одного поколения к другому и дожили до наших дней. Эти элементы принадлежат, как правило, к т.н. основному словарному фонду, на котором базируется вся лексико-семантическая система языка. В рамках квантитативно-системного исследования лексики возникает вопрос о закономерностях распределения древних слов в словаре современного языка, особенно в той его части, которая охватывает наиболее частотные лексические единицы и в определенной степени совпадает с основным словарным фондом языка. Совокупность наиболее частотных слов в представительном словаре именуется "базовым словарем". В данной работе в качестве базового словаря был взят фрагмент частотного словаря лексем авторской речи художественной прозы современного эстонского языка (Каа-

sik U. и др., 1977). Фрагмент состоит из 1200 слов-лексем, имеющих частоту не менее 10 в тексте объемом 100 тыс. словоупотреблений (весь объем словаря - 14,7 тыс. лексем). В базовом словаре не учитываются собственные имена. Поскольку авторская речь в художественной прозе занимает в некотором смысле центральное место в ряду жанровых подсистем языка (многие квантитативные характеристики очень близки к средним показателям совокупности подъязыков; см., например, Tuldava J., 1980, с. 59), то можно в первом приближении считать выбранный в данной работе базовый словарь представительным для всего языка. Важно отметить также, что выбор ограниченного базового словаря одного подъязыка дает возможность соблюсти принцип однородности исследуемого материала.

Краткие исторические сведения. Как известно, эстонский язык принадлежит к прибалтийской подгруппе финно-угорских языков, которые вместе с некоторыми более дальними родственными языками составляли когда-то единую семью уральских языков*. Считается, что уральское языковое единство распалось еще в IУ тысячелетии до н.э., а первое разделение финно-угорского языка-основы, состоявшего из близких племенных диалектов, произошло в III тысячелетии до н.э. (или еще раньше, см. Serebrennikov B., 1957, с. 157). По данным археологии, прибалтийско-финские племена появились на берегах Финского залива в том же III тысячелетии до н.э. В начале первого тысячелетия н.э. на территории теперешней Эстонии произошла консолидация одной части прибалтийско-финских племен, которые по своей материальной культуре заметно отличались от соседних восточных прибалтийско-финских и балтийских племен (Moora H., 1956, с. 91). Эти племена можно уже считать эстонскими племенами, хотя между ними существовали значительные диалектные различия. В основном выделяются две группы племенных диалектов - северо- и южноэстонские. На их основе в дальнейшем образовались по существу два языка: северо- и южноэстонский,

---

* В соответствии с генеалогической классификацией, принятой в советском финно-угроведении, уральские языки распадаются на две основные группы - финно-угорские и самодийские языки. К финно-угорским языкам относятся подгруппы прибалтийско-финских, саамского, волжских, пермских и обско-угорских языков. К прибалтийско-финским языкам относятся эстонский, ливский, водский (южная группа), ижорский, карельский, вепсский (восточная группа) и финский (северная группа). Обзоры см. Майтинская К.Е., 1966; Аристэ П., Вяари Э.Э., 1966; Laanes [illegible] 1975.

которые долгое время вели самостоятельное существование, но постепенно сближались друг с другом. Только в прошлом столетии сложился окончательно единый эстонский национальный язык, который образовался главным образом на базе северо-эстонского языка, но перенял и многие южно-эстонские черты.

Если говорить о слое древних слов в лексике современного эстонского языка, то приходится учитывать сложную историю становления языка. Все же можно утверждать, что лексические элементы, встречавшиеся, например, в разных диалектах эстонских племен в первом тысячелетии н.э., составляют общее достояние эстонского языка. Это подтверждается также сведениями о том, что несмотря на племенную раздробленность и диалектные различия эстонские племена еще в конце первого и в начале второго тысячелетия н.э. сознавали себя единым народом (см. Каск А.Х., 1966, с. 35). Несмотря на отсутствие письменных памятников на эстонском языке до XУI века (начиная с XIII века появились в местных латинских хрониках лишь отрывочные записи на эстонском языке – отдельные слова и выражения, географические названия, имена людей и др.), мы можем, хотя бы в пределах базового словаря, более или менее достоверно различать древнюю лексику от более поздних инноваций или заимствований. Границей можно считать год 1200, т.е. начало XIII века – период появления чужеземных колонизаторов в Прибалтике, после чего, безусловно, произошел заметный сдвиг в развитии лексики языка в связи с изменением внешних условий (феодализация, христианизация, новые иноязычные влияния). В дальнейшем мы будем называть лексические элементы, дошедшие до нас от времен до 1200 г. н.э., словами древнего происхождения, или древней лексикой (эст. iidne sõnavara), хотя и этот пласт в свою очередь подлежит дифференциации.*

Генетическая классификация лексики. Слова уральского, финно-угорского, прибалтийско-финского происхождения и собственно эстонские образования составляют исконную, или "собственную" лексику (эст. omasõnavara) эстонского языка. Кроме

---

* Автор выражает искреннюю благодарность проф. Х.Рятсепу и с.н.с. Т.-Р.Вийтсо за ценные советы и помощь при этимологизации слов. В этой работе были использованы также следующие пособия и исследования: Alvre P., 1981a; 1981b; Ariste P., 1956; 1981; Kask A., 1970; 1972; Laanest A., 1975; Raun A., 1982; Rätsep H., 1978; 1982; Saareste A., 1952.

того, заметное место в эстонской лексике занимают заимствования. По данным исследований Х.Рятсепа (Rätsep H., 1978; 1982), в эстонском литературном языке насчитывается ок. 5800 основ, из них исконных основ 3400 (т.е. около 60 %) и заимствованных основ 2400 (около 40 %). Следует отметить, что доля исконных элементов в эстонской лексике достаточно высока по сравнению с некоторыми другими языками. Например, среди 6000 корневых слов, зарегистрированных в современном венгерском языке, исконных слов оказалось лишь 18 % (в том числе финно-угорского происхождения – 10 % и венгерские новообразования – 8 %), а все остальные слова – заимствования (см. Папп Ф., 1980, с. 21).

Основы слов эстонского языка распределяются, по данным Х.Рятсепа, следующим образом (числа приблизительные, включая некоторые неясные и сомнительные случаи):

I. Исконные основы:

уральские	180	(период до IУ тыс. до н.э.)
финно-угорские*	600	(IУ – III тыс. до н.э.)
прибалтийско-финские**	1200	(III – I тыс. до н.э.)
эстонские инновации	1400	(I – II тыс. н.э.)
	~ 3400	

II. Заимствованные основы:***

древние индоевропейские	40	(УI – III тыс. до н.э.)
–"– балтийские	150	(III – I тыс. до н.э.)
–"– германские	330	(II – I тыс. до н.э.)
–"– славянские	50	(I тыс. н.э.)
–"– латышские	40	(I тыс. н.э.)
новые заимствования (немецкие, шведские, русские, финские и др.)	1800	(после 1200 г. н.э.)
	~ 2400	

---

* Финно-угорские основы подразделяются на "праязыковые" (ок. 260), пермские (180) и волжские (160).

** Включая ок. 100 основ "протоевропейского" субстрата (Ariste P., 1956; 1981).

*** Среди заимствованных основ (эст. laentüved) не учитываются основы иностранных слов. По данным Словаря иностранных слов (Kleis R. et al., 1981), в современном эстонском языке насчитывается более 20 000 иностранных слов (эст. võõrsõnad).

О всем объеме словаря современного эстонского литературного языка (не включая диалектных и узкоспециальных терминов)

В приведенной схеме учитываются только различные основы слов, причем основа слова (эст. sõnatüvi) определяется как та часть словоформы, которая остается, если отнять от нее формообразующий суффикс (tunnus) и окончание (см. Rätsep H., 1977, с. 3). Ясно, что количество о с н о в не совпадает с количеством с л о в, так как при подсчете слов учитываются не только простые (корневые), но и производные и сложные слова. Это касается также подсчета древних слов в современном языке, где число древних слов оказывается больше, чем число древних основ, так как до наших дней дошло и некоторое количество производных и сложных слов древнего происхождения. Однако не все древние основы встречаются в ежедневной речи: многие из них сохранились только в словах периферийного значения или в диалектных словах. Кроме того, при подсчете слов в базовом словаре оказалось, что многие древние основы встречаются в новообразованиях, т.е. в производных или сложных словах более позднего происхождения, например, põhjus 'причина' (põhi 'основа' в сочетании с суффиксом -us), salapärane 'таинственный' (sala-'тайный' +-pärane 'подобный'), ajalugu 'история' (генит. от aeg 'время' +lugu 'сказание'). В некоторых словах древняя основа сочетается с заимствованным суффиксом или словом более позднего периода и т.д. Таким образом, встает вопрос об идентификации древних с л о в в современном языке. Для этого потребуется сформулировать точные правила идентификации.

Идентификация древних слов. Выше было упомянуто, что по определению к "древним" словам эстонского языка относятся слова, которые по данным сравнительно-исторических исследований существовали в древне-эстонских диалектах в периоде до 1200 г. н.э. и которые сохранились в языке до наших дней. Это значит, что "возраст" таких слов должен быть не менее 800 лет. Таким образом, к древним словам следует отнести, в первую очередь, и с к о н н ы е с л о в а древнейшего происхождения, т.е. слова уральского, финно-угорского, прибалтийско-финского происхождения и собственно эстонские образования периода до 1200 г. н.э. Кроме того, к древним словам относятся д р е в н и е з а и м с т в о в а н и я*.

можно судить по представительному Ортологическому словарю (õigekeelsussõnaraamat, 1978), охватывающему 115 000 слов.

* В переводе на основы (см. схему на с.139) можно считать, что общее число исконных основ древнейшего происхождения (до 1200 г.н.э.) в современном эстонском языке достигает 3000 (из них ок. 1000 эстонских инноваций древнего периода), а число древних заимствований - 600.

Правила идентификации древних слов в современном языке должны также охватить семантические, морфологические и фонетические критерии. В настоящей работе постулируется принадлежность к древней лексике следующих слов:

1) непроизводные слова с установленной древней основой и со значением, или с одним из значений, предположительно равным или близким к старому значению или к одному из старых значений; например, lugema (-ma - суффикс инфинитива) 'читать' < *luke- 'читать числа, считать'; linn 'город' < *liina 'городище; крепость'; aed 'сад; забор, ограда' - раньше, по-видимому, только в значении 'ограда';

2) производные или сложные слова, состоящие из древних основ и аффиксов, причем существование этих образований с данным (близким) значением в древнем языке должно быть засвидетельствовано прямо или косвенно, например, по данным древних хроник или фольклора, или на основе научных историко-лингвистических соображений. Сюда относятся некоторые производные слова с суффиксами -lane, -ne, -mine, -us, -u, -ja и др.: sugulane 'родственник', vaenlane 'враг', kuldne 'золотой', tegemine 'дело, делание', küsimus 'вопрос', vastus 'ответ', nägu 'лицо', võitleja 'борец' (ср. антропоним Votteli-(ja), засвидетельствованный в XIII веке; ср. Saareste A., 1952, с. 78) и др. К древним сложным словам можно отнести главным образом топонимы, напр., Otepää, Alutaguse. Сомнительным является отнесение к древним сложным словам образований типа vanamees 'старик' (досл. 'старый' + 'мужчина'), igaüks 'каждый' ('каждый' + 'один'), tookord 'тот раз' и т.п., которые в древнем языке могли выступать как окказиональные словосочетания. В данной работе такие сомнительные случаи не были учтены. В общей сложности в базовом словаре встретилось 67 производных слов, из них 21 слово было признано древним, а из 33 сложных слов было причислено к древней лексике только одно слово - maailm 'мир, свет' (maa 'земля' + ilm 'мир').

При идентификации древних слов учитываются регулярные фонетические изменения, происходившие в языке в ходе своего исторического развития. В эстонском языке к таким явлениям относятся апокопа (например, laps 'ребенок' < *lapsi; silm 'глаз' < *silmä); синкопа (tahtma 'хотеть' < *tahtamahen); контракция (lükkama 'толкать' < *lükkäämähen) и многие другие комбинаторные изменения и фонетические сокращения (например, teadma 'знать' < *tētämähen; ainult 'только' < *ainaolta). Особый тип сокращений представляют собой т.н. "изно-

шенные" формы (эст. kuluvormid, см. Alvre P., 1982), которые с точки зрения обычных фонетических законов являются нерегулярными изменениями. Сюда относятся такие случаи, как vaatama 'смотреть' < valatama, valvatama; ei 'не, нет, ни' – партикль-наречие, образованное из древнего глагола отрицания, по-видимому, из формы 3-го числа презенса *epi (подробнее см. Künnap A., 1982), и многие бывшие производные или сложные слова, которые в силу фонетических сокращений выступают в современном языке как простые слова, например, ning 'и, а также' < *nĭn-kä; juba 'уже' < *jo-pa (-ka/-kä и -pa/-pä – древние партикли); aasta 'год' < *aiɣast(a)aiɣan – досл. 'от времени (срока) время'. В некоторых случаях допускается изменение аффиксальной части слова при сохранении изначального значения, например, kuidas 'как, каким образом' (вопросительный суффикс -s добавлен позднее, см. Kask A., 1972, с. 25), а также слова kes 'кто', mis 'что', miski 'что-нибудь' (в номинативе этих слов добавлен -s, но исконная основа сохранена в формах партитива: keda, mida, midagi). Однако к древним словам мы не причислили те случаи, где семантическая связь сокращенных форм с исходными формами уже не ощущается, например, aga 'а, но' < *aika 'время'; väga 'очень' < *väɣen kansak 'с мощью' (vägi 'мощь; войско'); ikka 'всегда' – иллатив от iga '(человеческий) век'.

В качестве примера приводится фрагмент базового словаря, в котором древние слова отмечены звездочкой (см. табл. I)*. Автор отдает себе отчет в том, что в ряде случаев нет полной уверенности при отнесении слова к древней или новой лексике. Однако характерная для лингвистических явлений "размытость границ" не помешает установлению общих тенденций и статистических закономерностей, связанных с распределением и употреблением древних слов в современном языке.

Анализ по частям речи. В базовом словаре, т.е. среди 1200 наиболее частотных слов было установлено 652 слова древнего происхождения (54,3 %). Наиболее обширную группу составляют существительные – 228 слов (35,0 % из общего числа древних слов; см. табл. 2), далее следуют глаголы – 177 (27,1 %) и

* В табл. I приведены "основные" формы лексем: у склоняемых слов номинатив, у глагола инфинитив (с суффиксом -ma); в русском переводе дается только форма мужского рода (он, весь, большой), причем надо учесть, что эстонский язык не знает грамматического рода (например, ta можно перевести как 'он', 'она', 'оно').

Табл. 1

Фрагмент эстонского частотного словаря: первые 100 наиболее частотных слов (лексем) - звездочкой отмечены древние слова

4237	*olema	'быть'	218	*tundma	'чувствовать'
3493	*ja	'и, да'	209	kas	'ли, разве'
2598	*ta	'он' (кратк.)	207	*vastu	'против,навстречу'
1981	*see	'этот'	206	*ära	'прочь; вы-,от-'
1395	*ei	'не,ни; нет'	200	*hea	'хороший'
1300	*et	'что(бы)'	198	*välja	'наружу; вы-,ма-'
1047	*kui	'когда,если'	195	*tahtma	'хотеть'
879	*mis	'что' (мест.)	194	*mitte	'не,ни'
845	*ma	'я' (кратк.)	192	*küll	'довольно, уж'
827	*tema	'он'	190	pärast	'потом,после'
724	aga	'а, но'	189	*mõni	'некоторый'
634	*oma	'свой'	185	*istuma	'сидеть'
613	*mina	'я'	185	*naine	'жен(щин)а'
581	*ise	'сам'	181	*andma	'давать'
568	*nagu	'как (будто)'	180	rääkima	'говорить'
499	*minema	'идти'	179	sest	'так как, ибо'
496	*tulema	'прийти'	176	*vana	'старый'
493	*siis	'тогда'	175	*panema	'класть'
465	*saama	'получать;мочь'	174	*kus	'где'
448	*kes	'кто'	174	*mõtlema	'думать'
436	*nii	'так'	171	*käima	'ходить'
434	ka	'тоже, также'	167	*ega	'ни; и не'
428	*ning	'и, а также'	166	*enam	'больше'
382	*kõik	'весь'	165	*päev	'день'
373	*mees	'муж(чин)а'	164	*seisma	'стоять'
366	*või	'или'	164	*siin	'здесь, тут'
350	*üks	'один'	163	*asi	'вещь; дело'
339	*tegema	'делать'	163	*pea	'голова'
327	*veel	'еще'	159	*nägu	'лицо'
309	kuid	'но, однако'	153	*seal	'там'
304	*hakkama	'начинать;стать'	151	*aasta	'год'
304	*jääma	'оставаться'	150	*sa	'ты' (кратк.)
304	*teine	'второй; другой'	149	*kuidas	'как(им образом)'
297	*vaatama	'смотреть'	148	*peale	'на; на-,над-'
285	*aeg	'время'	146	*tagasi	'обратно, назад'
285	*teadma	'знать'	144	*läbi	'(на)сквозь;через'
285	*võima	'мочь'	139	*kord	'порядок; раз'
272	*miski	'что-нибудь'	138	*kaks	'два'
272	*suur	'большой'	136	*all	'внизу; под'
268	*nägema	'видеть'	135	*keegi	'кто-нибудь'
264	*pidama	'быть должным'	133	*jõudma	'успевать; мочь'
261	*võtma	'брать'	133	*kõige	'всего,наи-;самый'
260	*silm	'глаз'	131	*iga	'каждый,всякий'
254	*inimene	'человек'	131	*vaid	'а, но; лишь'
248	*juba	'уже'	130	isegi	'даже'
238	*nüüd	'теперь'	130	*maa	'земля; страна'
237	*käsi	'рука'	126	ikka	'всегда'
234	*ainult	'только, лишь'	125	*jalg	'нога'
230	*üle	'свыше; через'	125	*ju	'ведь, же'
223	*ütlema	'сказать'	124	*sõna	'слово'

Табл. 2

Распределение древних слов по частям речи в базовом словаре (I200 наиболее частотных лексем) и в соответствующем тексте

Часть речи	Словарь число (С)	%	Текст число (Т)	%	Т/С
Существит.	228	35,0	9648	I6,4	42,3
Глагол	I77	27,I	I5I96	25,8	85,9
Прилагат.	82	I2,6	30I6	5,I	36,8
Местоим.	28	4,3	II4I5	I9,4	407,7
Числит.	I5	2,3	II93	2,0	79,5
Наречие	73	II,2	7872	I3,4	I07,8
Релят.	40	6,I	3092	5,3	77,3
Союз	9	I,4	7425	I2,6	825,0
Всего	652	I00,0	58857	I00,0	90,3

Табл. 3

Распределение частей речи в базовом словаре

Часть речи	Всего слов данной ч.р.	Из них древних слов число	%
Существит.	422	228	54,0
Глагол	294	I77	60,2
Прилагат.	I39	82	59,0
Местоим.	4I	28	68,3
Числит.	I6	I5	93,8
Наречие	20I	73	36,3
Релят.	67	40	59,7
Союз	I7	9	52,9
Междом.	3	0	0
Всего	I200	652	54,3

прилагательные – 82 (I2,6 %). Среди частей речи объединяются пред- и послелоги (из которых большинство может выступать также в роли наречия) под названием "релятивные слова" – таких слов древнего происхождения оказалось 40 (6,I %), а "чистых" наречий было 73 (II,2 %).

Учитывая частотность древних слов в т е к с т е (соответствующем базовому словарю; см. табл. 2), наибольшую употребительность обнаружили глаголы (25,8 % из всех древних слов в тексте), т.е. каждое четвертое древнее слово в тексте – это глагол. На втором месте местоимения (I9,4 %), и только затем следуют существительные (I6,4 %). Можно отметить, что функциональные нагрузки различных частей речи (слов древнего происхождения) заметно отличаются друг от друга. Функциональную нагрузку можно измерить отношением Т/С, где Т – число словоупотреблений в тексте, С – число слов в словаре. Как видно из таблицы (см. табл. 2), наибольшей функциональной нагрузкой обладают союзы и местоимения, а наименьшей – прилагательные и существительные. Остальные части речи составляют "среднюю" группу. Функциональная нагрузка существительных (Т/С = 42,3) указывает на то, что каждое древнее существительное употребляется в данном тексте в среднем 42,3 раза, в то время как каждый глагол (при Т/С = 85,9) встречается в среднем два раза больше, чем существительное. Основные полнозначные слова (существительные, глаголы, прилагательные, наречия) вместе взятые составляют 85,9 % словаря, но в соответствующем тексте они покрывают лишь 70,7 %.

Распределение древней лексики по частям речи можно рассматривать ещё с другой точки зрения, а именно, смотря по тому, какую долю из общего числа слов данной части речи составляют древние слова (см. табл. 3). Оказывается, что наибольший процент древних слов в базовом словаре имеют числительные (93,8 %) и местоимения (68,3 %), в то время как у наречий процент древних слов совсем низкий (36,3 %). У остальных частей речи процент древних слов колеблется между 50 и 60.

Если рассматривать распределение древних слов по частотным зонам базового словаря, разделяя словарь на I2 зон по I00 слов в каждой (по убывающей частоте), то можно констатировать большие различия между отдельными частями речи (см. табл. 4). Например, если в первой сотне наиболее частотных слов базового словаря среди 9I древнего слова имеется I5 существительных, т.е. I6,8 % из общего числа древних слов в

Табл. 4

Распределение древних слов по частям речи в базовом словаре (С) и в соответствующем тексте (Т)

Частотная зона (i)	Существит.		Глагол		Прилагат.		Местоим.		Числит.		Наречие		Релят.		Союз		Всего	
	С	Т	С	Т	С	Т	С	Т	С	Т	С	Т	С	Т	С	Т	С	Т
I (I-I00)	I5	2913	23	I0I23	3	648	I5	I0665	3	792	I8	5098	7	I269	7	7369	9I	38877
2 (I0I-200)	23	2063	I8	I565	6	538	4	376	I	I07	II	I430	I4	I338	0	0	77	74I7
3 (20I-300)	23	I286	I6	906	II	576	5	289	2	II7	II	550	2	96	0	0	70	3820
4 (30I-400)	23	826	23	874	8	285	I	33	0	0	7	279	3	I20	I	4I	66	2458
5 (40I-500)	24	668	I4	402	I0	283	I	28	2	58	6	I74	3	90	0	0	60	I703
6 (50I-600)	22	498	I2	275	I0	234	0	0	2	44	5	II5	3	7I	0	0	54	I237
7 (60I-700)	I9	364	I6	309	3	57	0	0	3	55	4	77	I	2I	0	0	46	883
8 (70I-800)	I8	296	I4	232	7	II6	0	0	0	0	3	49	2	33	0	0	44	726
9 (80I-900)	I9	267	I5	2II	4	57	0	0	0	0	3	4I	I	I3	I	I5	43	604
I0 (90I-I000)	I6	I98	I2	I52	6	73	I	I3	0	0	2	25	0	0	0	0	37	46I
II (I00I-II00)	I0	I09	7	77	8	89	I	II	0	0	3	34	I	II	0	0	30	33I
I2 (II0I-I200)	I6	I60	7	70	6	60	0	0	2	20	0	0	3	30	0	0	34	340
Всего	228	9648	I77	I5I96	82	30I6	28	II4I5	I5	II93	73	7872	40	3092	9	7425	652	58857

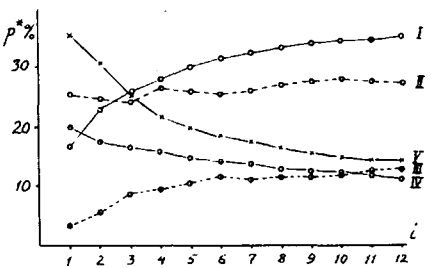


Рис. 1. Распределение древних слов в базовом словаре: связь между вложенными группами слов частотных зон (i) и накопленной частотой p*(%) различных частей речи (I - существит., II - глагол, III - прилагат., IV - наречие, V - остальные вместе взятые)

данной зоне, то в 6-ой сотне среди 54 древних слов – 22 существительных (40,7 %), а в 12-ой сотне среди 34 древних слов – 16 существительных (47,1 %). Особенно наглядно проявляется динамика изменения удельного веса частей речи при рассмотрении н а к о п л е н н ы х (кумулятивных) частот. Сделав необходимые перерасчеты (на основе данных табл. 4), мы можем проиллюстрировать это на графике (см. рис. 1), откуда видно, что удельный вес вложенных групп существительных среди древних слов постоянно увеличивается по мере приближения к зонам меньшей частотности. Из графика видно также, что доля глаголов среди древних слов остается более или менее стабильной, в то время как удельный вес прилагательных сначала увеличивается, а потом стабилизируется. Доля наречий и всех других частей речи монотонно уменьшается.

На основе данных употребительности древних слов в т е к с т е (см. колонку Т в табл. 4) можно сделать вывод, что и в тексте доля древних существительных постоянно увеличивается по мере приближения к низкочастотным зонам, в то время как глаголы и прилагательные сохраняют стабильность, а удельный вес других частей речи уменьшается. С помощью экстраполяции данных можно приблизительно прогнозировать, что доля существительных во всем словаре древних слов данного подъязыка (т.е. выходя за пределы базового словаря) составляет ок. 55 %, доля глаголов – ок. 25 % и доля прилагательных – ок. 10 %. В тексте сохраняется доминантное положение глаголов, а доли существительных и местоимений выравниваются. Большой удельный вес древних глаголов в тексте объясняется наличием в языке большого количества высокочастотных глаголов древнего происхождения.

Лексические группы древней лексики. Среди существительных обширную лексико-тематическую группу образуют слова, относящиеся к п р и р о д е (предметы и явления природы, растительный мир). Из общего числа 54 слов данной группы базового словаря 46 слов можно признать древними. Наиболее древний слой представляют слова уральского происхождения, т.е. слова, возраст которых превышает 6000 лет. К таким словам относятся (в скобках указана частота употребления в тексте объемом 100 тыс. словоупотреблений): maa (130) 'земля', vesi (68) 'вода', kivi (45) 'камень', lumi (43) 'снег', jõgi (26) 'река'; kuu (32) 'луна, месяц', täht (28) 'звезда'; puu (49) 'дерево') и др. Среди древнейших слов имеется и некоторое

число заимствований, например, taevas (60) 'небо' (по-видимому, индоиранского происхождения, см. Ariste P., 1981, с. 23). Можно констатировать, что наиболее древние слова этой группы являются в среднем и наиболее частотными.

Названия животных представлены в базовом словаре 11 словами, из них 9 древних, например, слова древнейшего происхождения kala (26) 'рыба', koer (18) 'собака' и hiir (16) 'мышь'. Древним германским заимствованием является lammas (10) 'овца'.

По понятным причинам отсутствуют или относятся к низкочастотным зонам рассматриваемого словаря (авторской речи современной художественной прозы) многие древние слова, обозначающие растения и лесных животных, а также слова, связанные с охотой, рыболовством и т.п. Зато хорошо представлена лексика, относящаяся к человеку, к его внешности и названиям, а также его свойствам, чувствам и деятельности.

Стопроцентно древними оказываются существительные, обозначающие части (или органы) тела. Их общее число в базовом словаре - 33. Наиболее древними и, в среднем, наиболее частотными являются: silm (260) 'глаз', käsi (237) 'рука', pea (163) 'голова', süda (58) 'сердце', suu (52) 'рот', keel (48) 'язык', õlg (47) 'плечо' и др. Эти слова относятся к уральскому периоду, в их числе можно выделить слова (käsi, keel), предположительно восходящие к "ностратической" лексике (по гипотезе В.М.Иллича-Свитыча), т.е. имеющие соответствия в древнейших пластах лексики индо-европейских, алтайских и др. языковых семей (см. Rätsep H., 1978, с. 6-7).

В группе названий людей (или групп людей) насчитывается в базовом словаре 28 слов, из них 16 древних. К словам древнейшего, уральского периода относятся термины родства isa (101) 'отец' и ema (93) 'мать', а также laps (74) 'ребенок, дитя' и poeg (63) 'сын'. Слово tütar (13), по-видимому, древнее балтийское заимствование. К финно-угорской лексике причисляются высокочастотные inimene (254) 'человек' и naine (185) 'женщина'. Более позднего происхождения слова mees (373) 'мужчина', vend (53) 'брат', õde (48) 'сестра', pere (21) 'семейство, семья'. Древним заимствованием является rahvas (38) 'народ' (см. Viitso T.-R., 1982).

Ряд употребительных древних существительных выражает человеческие чувства: rõõm (37) 'радость', mure (33) 'забота; горе', õnn (19) 'счастье', viha (14) 'гнев'; человеческую

речь: sõna (124) 'слово', jutt (93) 'речь; рассказ', hääl (69) 'голос'; абстрактные свойства: jõud (45) 'сила', õigus (32) 'право; правда'; временные понятия: aeg (285) 'время', päev (165) 'день', aasta (151) 'год', õhtu (94) 'вечер', öö (52) 'ночь' и др.

Из глаголов наиболее древними (уральскими) и наиболее частотными являются такие, которые обозначают типичные действия человека, например, minema (499) 'идти', tulema (496) 'прийти', tegema (339) 'делать', nägema (268) 'видеть', võtma (261) 'брать', а также saama (465) 'получать', tundma (218) 'чувствовать'. К финно-угорской лексике причисляются высокочастотные olema (4237) 'быть', teadma (285) 'знать', pidama (264) 'долженствовать', tahtma (195) 'хотеть', andma (181) 'давать' и др. Древним германским заимствованием является käima (171) 'ходить'.

К древним исконным прилагательным относятся высокочастотные слова, выражающие меру, например, pikk (97) 'длинный; kõrge (40) 'высокий'; свойства: hea (200) 'хороший', uus (105) 'новый', ilus (64) 'красивый', puhas (44)'чистый', tugev (35) 'сильный', halb (32) 'плохой', а также vana (176) 'старый', noor (82) 'молодой' и др.; названия цветов: valge (58) 'белый', must (51) 'черный', punane (47) 'красный' и др. Слова suur (272) 'большой' и lai (30) 'широкий' считаются древними германскими заимствованиями, а hall(71)'серый; седой' является балтийским заимствованием.

Среди наречий древнего происхождения наиболее частотными являются простейшие наречия времени: siis (492) 'тогда',nüüd (238) 'теперь', hilja (51) 'поздно'; täna (46) 'сегодня';наречия места и пространства, например, siin (163) 'здесь', seal (153) 'там'; наречия меры и количества: ainult (218) 'только', enam (166) 'больше', palju (107); наречия образа действия: nii (368) 'так', kuidas (149) 'как', äkki (44) 'вдруг', hästi (43) 'хорошо' и др. К древним заимствованиям относятся ju (125) 'ведь, же' (герм.) и veel (308) 'еще' (балт.).

Большинство местоимений относится к древней исконной лексике. Среди числительных встречаются древние заимствования sada (23) 'сто' (индоиранск.) и tuhat (30) 'тысяча' (балт.).

Какие же частотные слова базового словаря не являются древними? Среди существительных это в основном эстонские новообразования – производные и сложные слова, например, tunne (57) 'чувство', ülesanne (16) 'задание', olukord (16) 'поло-

жение', а также заимствования, относящиеся главным образом к культуре и быту людей: tuba (85) 'комната', pilt (34) 'картина', kool (25) 'школа', trepp (24) 'лестница' tool (23) 'стул', masin (22) 'машина', korter (I7) 'квартира', müts (I7) 'шапка', paber (I7) 'бумага' и др. Среди глаголов новыми являются производные собственные слова, например, juhtuma (7I) 'случаться', tunduma (54) 'казаться', ilmuma (37) 'появляться'; некоторые глаголы ономатопоэтического происхождения (их возраст трудно точно определить): rääkima (I80) 'говорить'; kõlama (I4) 'звучать'; ряд заимствований: märkama (75) 'замечать', proovima (I9) 'пробовать', teenima (I6) 'служить', pressima (II) 'нажимать; прессовать' и др.

Возраст и частотность слов. В предыдущей части статьи был отмечен тот факт, что в пределах отдельных лексико-тематических групп наиболее древние слова современного языка оказались в среднем наиболее частотными. Можно предположить, что существует какая-то общая связь между возрастом и частотностью слова, о чем свидетельствуют также результаты исследований по другим языкам (из работ последнего времени см., например, Арапов М.В., Херц М.М., I974; Вейлерт А.А., I978). Неверно было бы утверждать, что связь носит здесь непосредственно причинно-следственный характер, что, например, слово употребляется часто потому, что оно древнего происхождения, или наоборот. Во-первых, не все древние слова входят в зону частых слов, и часть из них вообще исчезла из языка вследствие изменения общественных условий или вследствие того, что древние слова были заменены новообразованиями. Во-вторых, частотность слов зависит в большой степени от сферы коммуникации, так, например, некоторые древние слова, редко или вообще не употребляемые в литературном языке, могут встречаться в диалектах или появляться в языке представителей какой-нибудь узкой сферы человеческой деятельности (охота, рыболовство и др.). Все это надо учитывать при изучении взаимосвязи между возрастом и частотностью слов. Однако нельзя пройти мимо того факта, что большая часть сохранившихся в языке древних слов относится к словам общеязыкового употребления. Причину тому следует частично искать в том, что в языке имеются определенные строевые слова, часто употребляемые в речи из-за требований языковой структуры (союзы, релятивные слова, местоимения, вспомогательные глаголы). Эти слова обычно сохраняются в языке в течение долгого периода именно из-за час-

того употребления в ежедневной речи. В языке сохраняется и определенное число древних полнозначных слов. Это в основном слова, выражающие наиболее привычные и простейшие предметы и явления человеческой жизни. Они употреблялись часто и в древности, и сегодня. Из-за своей "абсолютной" частотности они сохраняются в языке долго. Кроме того, именно сохранившиеся полнозначные слова в большинстве своем являются многозначными. Это слова, которые в процессе частого употребления обрастались новыми значениями вследствие того, что они появлялись в живой речи в разных окружениях (контекстах). Это привело к размытости границ семантического объема слов и, в конце концов, к многозначности. Многозначность, в свою очередь, закрепляет частотность употребления данного слова из-за его применения в разнообразных ситуациях. Далее, известно, что существует связь между частотностью и длиной слова, т.е. частые слова имеют тенденцию к укорочению. То же самое можно констатировать в отношении сохранившихся древних слов (в качестве примера можно указать на так называемые "изношенные слова" в эстонском языке, о которых речь шла выше). Краткость слова, в свою очередь, в определенной степени содействует его частому употреблению (момент экономии речи) и т.д. Из всего сказанного ясно, что если говорить о связи между возрастом и частотностью слов, то приходится, по сути дела, констатировать наличие очень сложной сети взаимосвязей и взаимообусловленностей не только непосредственно между упомянутыми свойствами, но и между ними и особенностями общественного и языкового развития, а также особенностями языковой структуры, в том числе связь с такими моментами, как многозначность и длина слова. Следовательно, связь между возрастом и частотностью слов предстает как проявление сложного взаимодействия компонентов системы и общности их поведения, причем обнаруживаемые связи в системе подчиняются законам развития и функционирования языка.

При квантитативно-системном подходе к исследованию связи между возрастом и частотностью слов целесообразно прибегать к методу моделирования с помощью р а с п р е д е л е н и й. Требуется выявить такие системные свойства исследуемых объектов, которые могут служить основанием для построения распределений и их математической экспликации с помощью функций. Опыт исследований по квантитативной лингвистике говорит о том, что такая процедура позволяет не только наглядно и эко-

номно описать те или иные явления, но и обнаружить в них специфические квантитативно-системные закономерности.

В данном случае мы воспользуемся методикой М.В.Арапова и М.М.Херц (1974), которые рассматривают связь возраста и частотности слова как основу для построения математической модели эволюции словаря. По этой методике экспериментальные данные – слова в представительном частотном словаре – объединяются в частотные зоны (группы) по 100 слов в каждой зоне. Ранжированным зонам приписываются номера, или ранги (i). В каждой зоне выявляется абсолютная частота древних слов $F(i)$ и соответствующая относительная частота $p(i) = \frac{F(i)}{n}$ (n=100).

По данным рассматриваемого базового словаря эстонского языка выявляется (см. табл. 5), что в первой зоне (i = 1), т.е. среди 100 наиболее частотных слов словаря встречается 91 древнее слово, что составляет 0,91 или 91 % всех слов зоны; во второй зоне соответствующие показатели 77 и 0,71 (71 %) и т.д. В 12-й зоне (i = 12) древних слов оказалось только 34 из 100. Можно констатировать монотонное убывание доли древних слов по частотным зонам в связи с увеличением ранга зоны (i), и, следовательно, с уменьшением частотности слов в среднем.

Эти данные можно представить и в виде и н т е г р а л ь н о г о распределения, т.е. на основе накопленных частот. Например, относительная частота древних слов на месте i = 12 (т.е. в пределах 1200 слов) равняется 0,54 (см. табл. 5).

Как обычное (дифференциальное), так и интегральное распределения могут быть представлены также для т е к с т а, соответствующего данному словарю (см. табл. 6). Мы видим, например, что если весь базовый словарь покрывает ок. 73 % текста (72853 словоупотреблений из 100 тыс.), то в пределах текста, соответствующего базовому словарю, древние слова составляют 0,808, или 80,8 % (58857 из 72853). Забегая вперед, можно отметить, что при экстраполяции с помощью соответствующей формулы прогнозируется ок. 60-процентное покрытие всего текста (авторской речи художественной прозы) древними словами. По данным другого, более раннего исследования (Saareste A., 1952, с. 64), в котором изучалось употребление различных пластов лексики в эстонском повествовательном тексте объемом ок. 1000 словоупотреблений, исконная лексика занимает 80 %, а древние заимствования покрывают в тексте ок. 10 %. Прямое сопоставление с нашими данными здесь невозможно, так как в упомянутом исследовании исконная лексика включает и эстон-

Табл. 5

Распределение слов древнего происхождения по частотным зонам эстонского словаря: $F(i)$ - абсолютная частота и $p(i)$ - относительная частота древних слов в данной частотной зоне (i); $F^*(i)$ и $p^*(i)$ - соответствующие накопленные частоты; n - объем зоны (n=100), n^* - накопленный объем зон

i	$F(i)$	$F^*(i)$	$p(i)=\frac{F(i)}{n}$	$p^*(i)=\frac{F^*(i)}{n^*}$
I	9I	9I	0,9I	0,9I
2	77	I68	0,77	0,84
3	70	238	0,70	0,79
4	66	304	0,66	0,76
5	60	364	0,60	0,73
6	54	4I8	0,54	0,69
7	46	464	0,46	0,66
8	44	508	0,44	0,64
9	43	55I	0,43	0,6I
I0	37	588	0,37	0,59
II	30	6I8	0,30	0,56
I2	34	652	0,34	0,54
Всего	652	-	-	-

Табл. 6

Древняя лексика в эстонском тексте: распределение слов соответственно частотным зонам (i) словаря.
$N(i)$ - покрываемость текста всеми словами данной частотной зоны, $F_t(i)$ - покрываемость текста древними словами, $N^*(i)$ и $F_t^*(i)$ - соответствующие накопленные частоты

i	$N(i)$	$N^*(i)$	$F_t(i)$	$F_t^*(i)$	$F^*(i)/N^*(i)$
I	4I358	4I358	38877	38877	0,940
2	89II	50269	74I7	46294	0,92I
3	5429	55698	3820	50II4	0,900
4	37I9	594I7	2458	52572	0,885
5	2839	62256	I703	54275	0,872
6	229I	64547	I237	555I2	0,860
7	I9I4	6646I	883	56395	0,849
8	I644	68I05	728	57I2I	0,839
9	I408	695I3	604	57725	0,830
I0	I249	70762	46I	58I86	0,822
II	I09I	7I853	33I	565I7	0,8I4
I2	I000	72853	340	58857	0,808
Всего	72853	-	58857	-	-

Табл. 7

Распределение слов древнего происхождения по частотным зонам (i) в шести языках (данные по пяти языкам - см. Арапов М.В., Херц М.М., 1974).

Ранг (зона) i	Языки					
	Эст.	Франц.	Нем.	Англ.	Рус.	Чеш.
	Датировка (год н.э.)					
	I200	I200	II00	II00	600	600
I (I-I00)	9I	9I	94	92	84	75
2 (I0I-200)	77	84	88	70	57	63
3 (20I-300)	70	84	83	53	52	50
4 (30I-400)	66	7I	73	40	5I	43
5 (40I-500)	60	73	63	47	42	37
6 (50I-600)	54	52	55	32	32	36
7 (60I-700)	46	57	55	29	42	45
8 (70I-800)	44	55	59	36	35	42
9 (80I-900)	43	52	52	3I	33	32
I0 (90I-I000)	37	6I	53	3I	35	32
Всего слов древн. происх.	588	680	675	46I	463	455

Табл. 8

Эмпирическое и теоретическое распределение частот древних слов в трех языках; параметр α и коэффициент убывания $e^{-\alpha}$ по экспоненциальному закону $F(i) = 100e^{-\alpha i}$

Ранг i (зона)	Эстонский		Французский		Немецкий	
	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.
I (I-I00)	9I	9I	9I	94	94	93
2 (I0I-200)	77	82	84	89	88	87
3 (20I-300)	70	74	84	84	83	8I
4 (30I-400)	66	67	7I	79	73	76
5 (40I-500)	60	6I	73	74	63	70
6 (50I-600)	54	55	52	70	55	66
7 (60I-700)	46	50	57	66	55	6I
8 (70I-800)	44	45	55	62	59	57
9 (80I-900)	43	4I	52	58	52	53
I0 (90I-I000)	37	37	6I	55	53	50
Параметры: α	0,I		0,06		0,07	
$e^{-\alpha}$	0,905		0,942		0,932	

ские инновации нового периода.

Для сравнения на уровне словаря приводятся данные о встречаемости древних слов по частотным зонам в ряде других языков (см. табл. 7)*. В пределах первых 1000 наиболее частотных слов наибольшее число древних слов обнаруживается во французском и немецком словарях (680 и 675, соответственно), в эстонском словаре их меньше (588), а в английском словаре значительно меньше (461). Данные этих четырех языков основываются на одинаковой датировке слов (1100–1200г. н.э.), причем правила идентификации древних слов приблизительно совпадают (наши правила несколько строже лишь в отношении семантических критериев; ср. Арапов М.В., Херц М.М., 1974,с. 58). Для русского и чешского языков датировка древних слов относится к более раннему периоду, а именно ко времени распада праславянского единства (ок. 600 н.э.). Число древних слов в обоих словарях примерно одинаковое (463 и 455).

В отношении всех рассматриваемых языков можно констатировать связь возраста и частотности слов в том смысле, что доля древних слов монотонно убывает в связи с увеличением ранга частотной зоны, т.е. с уменьшением частотности в среднем. Встает вопрос о форме математической зависимости между количеством древних слов $F(i)$ или долей $p(i)$ и рангом частотной зоны i. Эту зависимость приходится рассматривать в вероятностном плане, но формально в виде функции.

В первом приближении можно считать, что убывание количества (или доли) древних слов в связи с увеличением ранга подчиняется экспоненциальному (показательному) закону, при котором средний темп убывания остается постоянным. Распределение древних слов по частотным зонам можно в таком случае представить в виде пологой экспоненты, асимптотически приближающейся к оси абсцисс (см. рис. 2). Аналитическое выражение такой кривой имеет вид:

$$p(i) = e^{-ai}, \qquad (I)$$

где $p(i)$ – вероятность появления древних слов в зоне i,
e – основание натуральных логарифмов,
a – параметр (константа).

* Данные заимствованы из работы М.В.Арапова и М.М.Херц (1974) и основываются на материалах частотных словарей: по франц. яз. (автор словаря Г.Гуженейм), нем. яз. (Ф.Кединг), англ. яз. (А.Г.Дьюи), русский яз. (Э.А.Штейнфельдт), чешский яз. (А.Елинек, И.Бечка, М.Тешителова).

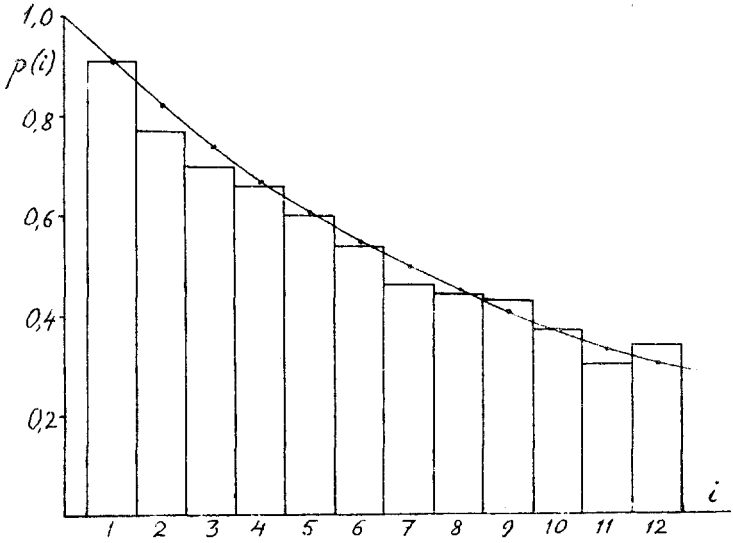


Рис. 2. Распределение древней лексики в частотном словаре эстонского языка. Связь между долей древних слов $p(i)$ и рангом частотной зоны (i). График функции $p(i) = e^{-0,1i}$

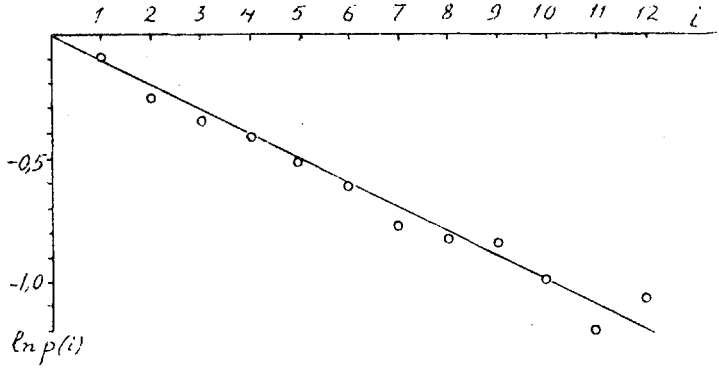


Рис. 3. Линейная связь между логарифмом доли $\ln p(i)$ и рангом i .

Как известно, такую же простую зависимость постулировал М. Сводеш (1960) в своей теории глоттохронологии в отношении вероятности сохраняемости древних слов за определенные промежутки времени в истории языка*.

Функция (I) достаточно хорошо описывает динамику убывания числа древних слов в пределах 10–12 частотных зон в эстонском, французском и немецком языках (см. табл. 8).** Особенно хорошо эта связь проявляется в эстонском языке, где за основу был взят базовый словарь одного подъязыка. На графике (рис. 3) видно, что соответствующее экспоненциальному закону условие линейной зависимости между $\ln p(i)$ и i приблизительно выполняется. Это может быть косвенным подтверждением предположения о действии экспоненциального закона в эволюции словаря согласно теории глоттохронологии (о прямой связи проблемы возраста и частотности с проблемой эволюции словаря см. Арапов М.В., Херц М.М., 1974). Однако функция (I) не подходит для всех языков и словарей, в частности для русского, чешского и английского языков (по данным рассматриваемых словарей). М.В.Арапов и М.М.Херц (1974) предлагают обобщающую формулу типа:

$$p(i) = e^{-a\sqrt{i}} . \qquad (2)$$

Хотя эта формула и подходит, например, для русского языка, она все же дает слишком приблизительные оценки по материалам других рассматриваемых словарей. Приходится искать другую обобщающую формулу. Нетрудно показать, что формулы (I) и (2) являются частными случаями более общей исходной формулы с тремя параметрами c, a и b:

$$p(i) = c e^{-a i^{b}} \qquad (3)$$

В формулах (I) и (2) параметр c заранее зафиксирован: $c = 1$, т.е. он указывает на максимум вероятности древних

---

* Формула (I) соответствует зависимости, применяемой в археологии при радиоуглеродном датировании. Это явилось прямым стимулом для создания теории глоттохронологии (см. Сводеш М., 1960, с. 24, 28). Можно еще отметить, что в формуле (I) компонента e^{-a} содержательно означает средний темп убывания, например, при $e^{-a} = 0{,}905$ доля древних слов составляет 0,905, или 90,5 % предыдущего уровня.

** В данном случае вычисляются значения абсолютных частот $F(i)$ на основе соотношения $p(i) = \frac{F(i)}{n}$, где n – объем зоны ($n = 100$).

слов. Это будет видно также при переписании формулы (3) для абсолютных частот:

$$F(i) = n e^{-a i^{b}}, \qquad (4)$$

где $F(i)$ – абсолютная частота древних слов,

n – объем группы (частотной зоны), т.е. максимум или точка отсчета, на основе которого определяется вероятность $p(i) = \frac{F(i)}{n}$. В данном случае n = 100.

Разница между формулами (1) и (2) в том, что в первом случае жестко зафиксирован параметр b = 1, а во втором случае b = 0,5 (т.к. $\sqrt{i} = i^{0,5}$). Представляется, что в данном случае можно допустить свободное варьирование параметра b, который предстает как квантитативно-лингвистический показатель, дифференцирующий языки и словари. Содержательно параметр b выражает темп убывания вероятности появления древних слов с уменьшением частотности слов. Параметр a имеет также содержательный смысл: он является показателем концентрации древних слов в начальной части частотного словаря (меньшее значение a отражает бóльшую концентрацию). Например, уточняя значения параметров по словарю эстонского языка с помощью формулы (3), мы получаем: a = 0,11* и b = 0,96. Для русского языка (в пределах i = 1 ÷ 25) параметры a = 0,32 и b = 0,51.

К вопросу аналитического описания распределения древних слов в словаре можно подойти и иначе, а именно, исходя из и н т е г р а л ь н о г о распределения слов. В таком случае надо рассматривать накопленные частоты древних слов по "вложенным" частотным зонам (в первой зоне, в первой и второй зонах вместе взятых и т.д.). При выборе соответствующей функции необходимо учесть, что рост числа древних слов имеет предел и что рост замедляется по мере приближения к пределу. Следуя соображениям, приведенным в работе Г.Шраубера (1967; цит. по: Добров Г.М., 1969, с. 155–158), можно вывести функцию типа:

$$p^{*}(i) = 1 - e^{-a i^{b}}. \qquad (5)$$

Здесь $p^{*}(i)$ – вероятность, соответствующая отношению $F^{*}(i)/F_{n}$ где $F^{*}(i)$ – накопленная частота древних слов,

---

* Параметр a равняется логарифму вероятности появления древних слов в первой зоне (при i = 1). Для выявления численного значения этой вероятности требуется найти антилогарифм. В данном случае $e^{-a} = e^{-0,11} = 0,90$ (фактически доля древних слов в первой зоне составляла 0,91; см. табл. 5).

F_n – предел числа древних слов в данной совокупности; a и b – параметры*. Формула (5) совпадает с известным в квантитативной лингвистике интегральным законом распределения В е й б у л л а (Weibull W., 1939; см. также Бектаев К.Б., Пиотровский Р.Г., 1973, с. 136-138).

В таблице 9 приводятся результаты вычислений по формуле (5) на материале шести языков при одинаковых условиях эксперимента (в пределах $i = 1 \div 10$ для всех языков). Можно констатировать хорошее соответствие между эмпирическими и теоретическими данными. В эстонском языке прогнозируется предел количества древних слов $F_n \approx 1000$ для данного словаря авторской речи художественной прозы, т.е. около 7 % всего словаря (объемом 14,7 тыс. слов). Надо учесть, что F_n прогнозирует число древних слов во всем словаре при условии, что темп роста остается неизменным и за пределами экспериментальных данных (см. рис. 4). Естественно, что для увеличения достоверности прогноза придется увеличить объем экспериментального материала. Но для сравнительного типологического анализа, по-видимому, достаточно охватывать экспериментом лишь первую тысячу наиболее частотных слов (см. также Арапов М.В., Херц М.М., 1974, с. 59). Показатель F_n можно в таком случае рассматривать как относительную оценку "архаичности" данного словаря.

Другим параметрам распределения (5) можно также придать содержательный смысл. Параметр a выражает (по отношению к F_n) степень концентрации древних слов в начале словаря (большее абсолютное значение a означает относительно бóльшую концентрацию).** Параметр b отражает темп роста, причем отношение

---

* Параметры a и b могут быть найдены методом наименьших квадратов на основе линеаризации: $\ln \ln \frac{1}{1-p^*(i)} = \ln a + b \ln i$. Здесь $\ln a$ – начальная ордината, b – угловой коэффициент. F_n определяется итеративным способом (подбирается такое значение F_n, при котором соответствие между экспериментальными и теоретическими данными наилучшее; на первой стадии это удобно делать на графике на основе линеаризации).

** Точнее, параметр a равняется логарифму вероятности $1-p^*(i)$ для первой зоны ($i = 1$), т.е. он указывает на долю еще не появившихся древних слов из общей совокупности F_n. Например, для эстонского языка $e^{-a} = e^{-0,095} = 0,909$. Следовательно, доля непоявившихся древних слов 0,909 или 90,9 %; доля уже появившихся древних слов: $1 - 0,909 = 0,091$ или 9,1 %, в данном случае 91 из 1000 (при $F_n = 1000$).

Табл. 9

Интегральное распределение частот древних слов в шести языках: эмпирические и теоретические данные; параметры распределения Вейбулла

Ранг (зона)	Эстонский		Французский		Немецкий		Английский		Русский		Чешский	
	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.
I (I-I00)	9I	9I	9I	93	94	96	92	94	84	84	75	76
2 (I0I-200)	I68	I70	I75	I77	I82	I80	I62	I58	I4I	I42	I38	I33
3 (20I-300)	238	24I	259	255	265	257	2I5	2I2	I93	I93	I88	I84
4 (30I-400)	304	305	330	327	338	329	255	258	244	242	23I	230
5 (40I-500)	364	364	403	395	40I	395	302	300	286	282	268	273
6 (50I-600)	4I8	4I7	455	459	456	458	334	337	3I8	32I	304	3I3
7 (60I-700)	464	466	5I2	5I9	5II	5I6	363	37I	360	359	349	35I
8 (70I-800)	508	5I0	567	576	570	570	399	408	395	394	39I	387
9 (80I-900)	55I	55I	6I9	630	622	622	430	43I	428	428	423	42I
I0 (90I-I000)	588	588	680	680	675	670	46I	458	463	460	455	453
Параметры:												
F_n	I000		I600		I500		900		I900		I500	
α	0,095		0,060		0,066		0,II0		0,045		0,052	
β	0,97I		0,965		0,953		0,807		0,792		0,836	
i_e	II,3		I8,5		I7,3		I5,3		50,4		34,5	

a/b указывает на степень насыщения (степень близости развивающегося процесса к пределу). На основе этого отношения можно вычислить, например, индекс $i_e = e^{-(\ln a)/b}$, который конкретно указывает на то значение i, при котором достигается 63 %-ое насыщение словаря древними словами $(0{,}63\,F_n)$.* Например, в эстонском языке (при F_n = 1000) такая точка достигается при i_e = 11,3, т.е. в 11-ой частотной зоне (по 100 слов) с начала словаря. В русском языке (при F_n = 1900) такая степень насыщения наступает лишь в 50-ой зоне (данные по шести языкам см. табл. 9). Разница в значениях показателей отражает различие в генетических структурах языков, но в данном случае, по-видимому, определенную роль играет и разный состав словарей, например, эстонский словарь составлен на основе лишь одного подъязыка, а русский словарь (частотный словарь Э.А.Штейнфельдта) представляет собой сводный словарь многих подъязыков. Следовательно, при сравнении генетических структур словарей разных языков приходится в определенной степени учитывать и жанровые особенности словарей.

Таким образом, анализ функции распределения Вейбулла в качестве модели распределения древних слов в словаре показывает, что параметры этого теоретического распределения могут в прямом или косвенном смысле служить квантитативно-лингвистическими характеристиками и стиледифференцирующими факторами при анализе лексики: параметр F_n – как общий показатель архаичности лексики (предел количества древних слов); параметр a – как показатель концентрации древних слов в начале словаря; индекс i_e (вычисленный на основе отношения параметров a и b) – как своеобразный показатель скорости насыщения словаря древними словами. С помощью функции распределения возможны также экстраполяция и интерполяция данных в пределах словаря. В общей сложности параметры распределения Вейбулла отражают как интегральные свойства, так и внутрисистемные взаимосвязи между элементами системы лексики, и вместе с тем указывают на системную связь между возрастом и частотностью слов.

---

* Данная величина получается из выражения $F^*(i)/F_n = 1 - \frac{1}{e} = 0{,}63$. Отношение $-(\ln a)/b$ выражает при линеаризации (см. выше) то значение $\ln i$ при котором $\ln\ln \frac{1}{1-p^*(i)} = 0$; отсюда $\frac{1}{1-p^*(i)} = e$ и $p^*(i) = F^*(i)/F_n = 1 - \frac{1}{e}$ ($e = 2{,}72$ – основание натуральных логарифмов).

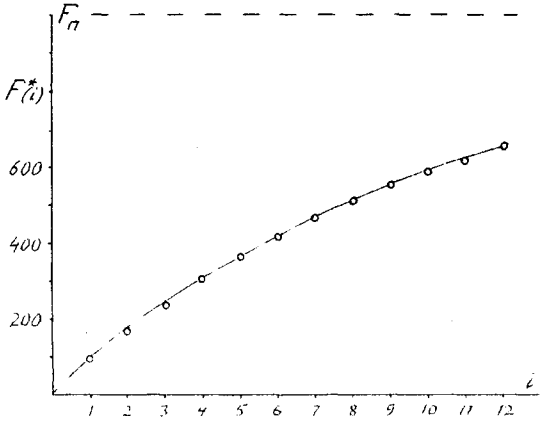


Рис. 4. Связь между накопленной частотой древних слов $F^*(i)$ и рангом частотной зоны i. $F_п$ – предел количества древних слов.

Табл. 10

Эмпирическое и теоретическое интегральное распределение древних слов в эстонском тексте по функции

$$F_t^*(i)' = 61000\,(1 - e^{-\sqrt{i}}).$$

Зона ЧС (i)	Эмпир. $F_t^*(i)$	Теоретич. $F_t^*(i)'$
I (I-I00)	38877	38559
2 (I-200)	46294	46I70
3 (I-300)	50II4	50208
4 (I-400)	52572	52745
5 (I-500)	54275	54480
6 (I-600)	555I2	55733
7 (I-700)	56395	56672
8 (I-800)	57I2I	57395
9 (I-900)	57725	57963
I0 (I-I000)	58I86	584I8
II (I-II00)	585I7	58787
I2 (I-I200)	58857	5909I

Интегральной функции (5) соответствует дифференциальная вида

$$\rho(i) = ab i^{b-1} e^{-ai^b} . \qquad (6)$$

Эта формула родственная формуле (3); последнюю можно причислить к формулам (распределениям) "вейбулловского типа", причем отличительным свойством таких формул является наличие в их правой части компонента e^{-ai^b}. Интересно отметить, что при $b = 1$ формула (6) превращается в экспоненциальное распределение классического типа, которое является, по существу, частным случаем более общего распределения Вейбулла. Выше уже говорилось о близости распределений древних слов в эстонском, французском и немецком языках к экспоненциальному распределению. Это подтверждается нашим последним экспериментом: значение параметра b по данным названных языков оказывается близким к 1 (см. табл. 9).*

Наконец, можно показать, что распределение Вейбулла подходит и для описания распределения древних слов в т е к с т е. В интегральной форме можно использовать формулу (5) с той разницей, что вместо вероятности $\rho^*(i)$ (по словарю) за основу берется вероятность $\rho_t^*(i)$ (по тексту):

$$\rho_t^*(i) = 1 - e^{-ai^b} . \qquad (7)$$

Здесь $\rho_t^*(i) = F_t^*(i) / F_{tn}$, где $F_t^*(i)$ – накопленное число древних слов (словоупотреблений) в тексте, F_{tn} – предел числа древних слов в данном тексте. Для рассматриваемого эстонского текста параметры $a = 1$; $b = 0{,}5$; $F_{tn} \approx 61000$, т.е. прогнозируется наличие в тексте 61000 древних словоупотреблений, или 61 % текста (объемом 100 тыс. словоупотреблений). Соответствие между эмпирическими и теоретическими данными хорошее (см. табл. 10).

На основе проведенного исследования можно сделать вывод, что распределение Вейбулла, как "обобщенный закон прогрессивного роста" (по Г.Шрауберу), является вполне адекватным средством описания и объяснения распределений древних слов в словаре и тексте разных языков. Параметры распределения содержательно интерпретируемы и имеют определенный лингвистический смысл. Распределение Вейбулла в качестве обобщающей модели может служить также квантитативно-системной экспликацией связи между возрастом и частотностью слов.

---

* Судя по эмпирическим данным, параметр b колеблется между 0,5 и 1. Только теоретически возможен случай, когда $b > 1$. В таком случае кривая распределения (5) превращается в логисту с точкой перегиба.

ЛИТЕРАТУРА

Арапов М.В., Херц М.М. Математические методы в исторической лингвистике. – М.: Наука, 1974.

Аристэ П.А., Вяари Э.Э. Прибалтийско-финские языки. Введение. – В кн.: Языки народов СССР. Т. Ш. – М.: Наука, 1966, с. 26-34.

Бектаев К.Б., Пиотровский Р.Г. Математические методы в языкознании. Ч. I. Алма-Ата, 1973.

Вейдерт А.А. О некоторых факторах, определяющих частоту слова в тексте. – Вопросы языкознания, 1978, № 2, с.99-104.

Добров Г.М. Прогнозирование науки и техники. – М.: Наука, 1969.

Каск А.Х. Эстонский язык. – В кн.: Языки народов СССР. Т. Ш. – М.: Наука, 1966, с. 35-60.

Майтинская К.Е. Введение. – В кн.: Языки народов СССР. Т. Ш. Финно-угорские и самодийские языки. – М.: Наука, 1966, с. 9-25.

Пап Ф. Лингвостатистика и венгерский язык. – Учен. зап. Тартуского ун-та, вып. 518. Труды по лингвостатистике У. Тарту, 1980, с. 15-37.

Сводеш М. Лексикостатистическое датирование доисторических этнических контактов. Перев. с англ. – В кн.: Новое в лингвистике. Вып. I. М., 1960, с. 23-52.

Шраубер Г. О росте научно-технических параметров с точки зрения прогнозирования. – В кн.: Материалы совещания специалистов стран – членов СЭВ и СФРЮ по обмену опытом при составлении научно-технических прогнозов. Прага, 1967, с. 41-65.

Alvre P. Uurali keelte ajaloolise foneetika harjutusülesanded ja materjalid (üksikkonsonandid). 2. tr. – Tartu: TRÜ, 1981a.

Alvre P. Uurali keelte ajaloolise foneetika harjutusülesanded ja materjalid (konsonantühendid). 2. tr. – Tartu: TRÜ, 1981b.

Alvre P. Omapäraseid kuluvorme. – Keel ja Kirjandus 1982, nr. 6, lk. 310-314.

Ariste P. Läänemere keelte kujunemine ja vanem arenemisjärk. Rmt.: Eesti rahva etnilisest ajaloost. – Tallinn: ERK, 1956, lk. 5-23.

Ariste P. Keelekontaktid. Eesti keele kontakte teiste keeltega. - Tallinn: Valgus, 1981.

Kaasik Ü., Tuldava J., Villup A., Ääremaa K. Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik. - TRÜ Toimetised, vihik 413. Töid keelestatistika alalt II. Tartu, 1977, lk. 5-140.

Kask A. Eesti kirjakeele ajaloost I. - Tartu: TRÜ, 1970.

Kask A. Eesti keele ajalooline grammatika. Häälikulugu. 2. tr. Tartu: TRÜ, 1972.

Kleis R., Silvet J., Väärі E. Võõrsõnade leksikon. 4.tr. - Tallinn: Valgus, 1981.

Künnap A. Eesti eitussõnade ei, ep ja es tausta. - TRÜ Toimetised, vihik 611. Fenno-Ugristica 9. Tartu, 1982, lk. 61-66.

Laanest A. Sissejuhatus läänemeresoome keeltesse. - Tallinn: TA KKI, 1975.

Moora H. Eesti rahva ja naaberrahvaste kujunemisest arheoloogia andmeil. - Rmt.: Eesti rahva etnilisest ajaloost. - Tallinn: ERK, 1956, lk. 41-119.

Raun A. Eesti keele etünoloogiline teatmik. - Toronto: Maarjamaa, 1982.

Rätsep H. Eesti keele ajalooline morfoloogia I. - Tartu: TRÜ, 1977.

Rätsep H. Eesti keele sõnavara ajaloo põhiprobleeme. - TRÜ Toimetised, vihik 460. Töid eesti filoloogia alalt V. Tartu, 1978, lk. 3-14.

Rätsep H. Millest koosneb eesti kirjakeele sõnavara. - Ettekanne Eesti TV-s 5. apr. 1982.

Saareste A. Kaunis emakeel. - Lund: Eesti Kirjanike Kooperatiiv, 1952.

Serebrennikov B. Mõningaid kaudseid andmeid kaasaegsete soome-ugri keelte üksikute rühmade kujunemise ajast. - Emakeele Seltsi aastaraamat III. Tallinn, 1957, lk. 153-158.

Tuldava J. Eesti keele sõnavara foneetilis-grafeemilised mõõted. - TRÜ Toimetised, vihik 518. Töid keelestatistika alalt V. Tartu, 1980, lk. 51-100.

Viitso T.-R. On Early Loanwords in Finnic. - Symposiumi 82. Suomalais-neuvostoliittolainen itämerensuomalaisen filologian symposiumi 30.8. - 2.9.1982. Esitelmien referaatit. Jyväskylä, 1982, pp. 78-79.

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

Õigekeelsussõnaraamat. 2. tr. Toimetanud R. Kull ja E. Raiet. - Tallinn: Valgus, 1978.

A QUANTITATIVE INVESTIGATION OF THE GENETIC STRUCTURE OF THE ESTONIAN VOCABULARY

Juhan Tuldava

S u m m a r y

In this article the various "strata" of different origin in the Estonian vocabulary are examined as constituent parts of the lexicon and the text. The analysis is carried out on the basis of the core of a frequency dictionary (1,200 most frequent words in the modern language) which is considered as the quantitative model of the text reflecting the use of the most important and topical words in the living language. To begin with, a short historical survey of the development of the Estonian language is given and the principles of the identification of the ancient words in modern language are stated. Then follows the quantitative analysis of ancient words (presumably inherited from the period before the 13th century of our era) according to their division into parts of speech and lexical groups. Finally, the problem of the relation between the age of the words and their frequency is discussed. The Weibull distribution is found to be suitable to serve as the quantitative-probabilistic model of the age-frequency relation. This has been demonstrated on the material of six languages (Estonian, Russian, Czech, English, German, and French).

Содержание

SUMMARIES

Ученые записки Тартуского государственного университета.
Выпуск 628.
ЛИНГВОСТАТИСТИКА И ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА.
Труды по лингвостатистике VIII.
На русском языке.
Резюме на английском языке.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, I8.
Ответственный редактор Я. Соонтак.
Подписано к печати I0.I2.I982.
МВ I29I0.
Формат 60x90/I6.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 9,76.
Печатных листов I0,5.
Тираж 450.
Заказ № I306.
Цена I руб. 50 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Пялсона, I4.