АКАДЕМИЯ НАУК СССР
ИНСТИТУТ ЯЗЫКОЗНАНИЯ

Статистика речи и автоматический анализ текста

ИЗДАТЕЛЬСТВО «НАУКА»
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ
ЛЕНИНГРАД · 1971

Сборник представляет собой коллективную монографию, посвященную статистическому и теоретико-информационному описанию основных европейских языков. Книга содержит также результаты автоматической переработки текстов, написанных на этих языках.

Машинные программы, изложенные в статьях сборника, имеют вероятностный характер и предполагают предварительное информационно-статистическое описание соответствующих текстов.

Сборник рассчитан на широкие круги лингвистов, математиков и инженеров, занимающихся проблемами информатики.

**СТАТИСТИКА РЕЧИ
И АВТОМАТИЧЕСКИЙ АНАЛИЗ
ТЕКСТА**

Утверждено к печати
Институтом языкознания
АН СССР

Редактор издательства *А. А. Зырин*. Художник *М. И. Разулевич*
Технический редактор *Г. А. Смирнова*. Корректор *Ф. Я. Петрова*

Сдано в набор 25/XI 1970 г. Подписано к печати 15/VII 1971 г. Формат бумаги 60×90¹/₁₆. Печ. л. 20 + 3 вкл. (³/₄ печ. л.) = 20.75 усл. печ. л. Уч.-изд. л. 34.63. Изд. № 4264. Тип. зак. № 1106. М-11059. Тираж 2800. Бумага № 2. *Цена 2 р. 04 к.*

Ленинградское отделение издательства «Наука»
Ленинград, В-164, Менделеевская лин., д. 1

1-я тип. издательства «Наука». Ленинград, В-34, 9 линия, д. 12

7—1—1, 7—1—4
204—70 (I пол.)

Сокращения

АДД	— автореферат докторской диссертации.
АКД	— » кандидатской ».
АЛ	— сб. «Автоматизация в лингвистике», Москва—Ленинград, 1966.
АН	— Академия наук.
ВММ	— сб. «Вычислительные машины и мышление», Москва, 1967.
ВП	— Вопросы психологии, Москва.
ВФ	— Вопросы философии, Москва.
ВЯ	— Вопросы языкознания, Москва.
ГПИ	— Государственный педагогический институт.
ГПИИЯ	— » » » иностранных языков.
дв. ед.	— двоичная (-ные) единица (-цы).
ИЯ	— сб. «История языкознания XIX—XX веков в очерках и извлечениях», ч. II, Москва, 1963.
КД (рукоп.)	— кандидатская диссертация (рукопись).
л.	— лист.
ЛГУ	— Ленинградский государственный университет.
МГУ	— Московский » ».
МКЧСА	— сб. «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ», Издание Ленинградского государственного университета, 1966.
МП	— сб. «Машинный перевод», Москва.
НДВШ	— Научные доклады высшей школы, Москва.
НЛ	— сб. «Новое в лингвистике», Москва.
НС	— новая серия.
п.	— пункт.
СР	— сб. «Статистика речи», Ленинград, 1968.
ЭВМ	— электронно-вычислительная машина.
adv	— adverb.
art	— article.
attr	— attributive, Adjektiv (нем.).
aux	— auxiliary (verb).
cj	— conjunction.
dem	— demonstrative.

ger — gerund.
imp — imperative.
inf — infinitive.
interrog — interrogative.
introd — introductory.
l — link (verb).
mod — modal (verb).
n — noun.
NLC — Natural Language and Computer, New York.
not — notional (verb).
num — numeral.
pI — participle I.
pII — participle II.
part — particle.
past indef — past indefinite.
pref — prefix, Präfix (нем.).
pres indef — present indefinite.
pron — pronoun.
prp — preposition.
rel — relative.
sing — singular.
v — verb.
vbl n — verbal noun.

Часть I. СТАТИСТИЧЕСКАЯ СТРУКТУРА ТЕКСТА

Р. Г. Пиотровский и Л. А. Турыгина

АНТИНОМИЯ «ЯЗЫК — РЕЧЬ» И СТАТИСТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ НОРМЫ ЯЗЫКА

ВВЕДЕНИЕ

§ 1. Нерешенные вопросы соссюровской антиномии «язык — речь»

Существо соссюровского учения о языке, являющегося фундаментом современной лингвистики, состоит, как тонко подметил Л. Ельмслев, в различии между языком и речью.[1] Проблема антиномии «язык — речь» продолжает оставаться центральным вопросом всей послесоссюровской лингвистики.

Современная структурно-математическая лингвистика выступает в роли науки-«лоцмана» для ряда общественных и естественных дисциплин,[2] поэтому исследование антиномии «язык — речь» приобретает общеметодологическое значение. И действительно, для всякой науки, с одной стороны, оперирующей конечным множеством значимых базовых элементов, которое выступает в роли «кода» («языка»), а с другой — имеющей дело с потенциально бесконечным числом порождаемых этим множеством «сообщений» («текстов», «речевых фраз», «композиций» и т. п.), изучение антиномии «код — сообщение» («язык — речь») оказывается одной из центральных проблем. Эта ситуация возникает не только в гуманитарных дисциплинах семиотического цикла (теория семиотики,[3] структурное литературоведение и фольклор, фильмология, структурная теория музыки и театра, социальная и инженерная

[1] Л. Ельмслев. Язык и речь. Русск. пер. ИЯ, стр. 111.
[2] S. Marcus. Lingvistica, ştiintă-pilot. Studii şi cercetări lingvistice, XX, 3, 1969, pp. 235—245.
[3] Ch. S. Peirce. Collected Papers, vol. VIII. Cambridge (Mass.), 1958, § 334; Ch. Morris. Signs, Language and Behavior. New York, 1946, p. 20; R. Carnap. Introduction to Semantics. Cambridge (Mass.), 1961; H. Reichenbach. Elements of Symbolic Logic. New York, 1966, а также другие работы.

психология и др.),[4] абстрактной экономике,[5] теории организации армии и флота[6] или естественных науках, связанных, подобно биологии, с кодированием информации.[7] Она присутствует также в технических дисциплинах (машиностроение, электроника и т. д.), рассматривающих вопросы рекурсивной композиции технических устройств из конечного числа элементарных единиц типа деталей машин или же полупроводников, сопротивлений и т. п.

Основные положения соссюровского учения о языке и речи известны.

«Язык, — писал Ф. де Соссюр, — есть система знаков, выражающих идеи, а следовательно, его можно сравнивать с письмом, с азбукой для глухонемых, с символическими обрядами, с формами учтивости, с военными сигналами и т. д. и т. п.

«Язык — это. . . система, потенциально существующая в каждом мозгу или, лучше сказать, в мозгах целой совокупности индивидов, ибо язык не существует полностью ни в одном из них, он существует в полной мере лишь в массе. . .[8]

«Язык не есть функция говорящего субъекта, он — продукт, пассивно регистрируемый индивидом. . . Он есть социальный элемент речевой деятельности вообще, внешний по отношению к индивиду, который сам по себе не может ни создавать язык, ни его изменять. Язык существует только в силу своего рода договора, заключенного членами коллектива.

«. . .Образующая язык совокупность звуковых и концептуальных различий является результатом двоякого рода сближений — ассоциативных (здесь — п а р а д и г м а т и ч е с к и х, — *Р. П. и Л. Т.*) и синтагматических (здесь в значении — в а л е н т н о с т н ы х — *Р. П. и Л. Т.*); как те, так и другие в значительной мере устанавливаются языком; вот эта совокупность установившихся (узуальных) отношений и составляет язык и определяет его функционирование».[9]

Иные свойства имеет речь.

Для нее характерна «экзекутивная (выполняющая) сторона», т. е. реализация возможностей системы.[10] Одновременно «. . .речь есть индивидуальный акт воли и понимания, в котором надлежит различать: 1) комбинации, при помощи которых говорящий субъект пользуется языковым кодексом с целью выражения своей личной мысли; 2) психофизический механизм, позволяющий ему объективировать эти комбинации».[11]

«Она, — продолжает женевский лингвист, — сумма всего, что говорят люди, и включает: а) индивидуальные комбинации, зависящие от воли говорящих, б) акты,. . . необходимые для выполнения этих комбинаций. Следовательно, в речи ничего нет коллективного: проявления ее — индивидуальны и мгновенны; здесь нет ничего, кроме частных случаев. . .».[12]

Эскизный характер соссюровского определения антиномии «язык — речь» был ясен уже ближайшим ученикам и последователям де Соссюра.[13] Существует по крайней мере четыре вопроса, на которые ни «Курс», ни материалы к нему, ни разъяснения непосредственных его интерпретаторов не дают сколько-нибудь последовательного ответа.

Первый вопрос состоит в определении границы между языком и речью.

При этом нужно выяснить, какие лингвистические единицы следует относить к сфере языка (лингвистического кода), а какие — к компетенции речи (сообщения).[14]

Второй вопрос заключается в выявлении тех механизмов, с помощью которых система языка, состоящая из конечного числа

---

[4] Ср. изучение проблемы кода и сообщения в работах: P. G u i r a u d. Linguistique et critique littéraire. Cours d'été et colloques scientifiques de langue, littérature, histoire et art du peuple roumain, Sinaia, VI—VIII, 1967; V. P r o p p. Morphology of the Folktale. International Journal of American Linguistics, vol. XXIV, 4, 1957; G. P. S p r i n g e r. Language and Music: parallels and divergencies. In: For Roman Jakobson. The Hague, 1956; N. R u w e t. Musicologie et linguistique. Revue internationale des sciences sociales, 1967, № 1; V. H. V r o o m, N. R. F. M a i e r. Industrial Social Psychology. Annual Review of Psychology, vol. XII, Palo Alto (California), 1961, pp. 413—446; M. D i n u. Structures linguistiques issues de l'étude du théâtre. Cahiers de linguistique théorique et appliquée, V, 1968.

[5] Ср.: S. C h e r u b i n o. Sille nozioni di ciclo di produzione, epoca, struttura, sviluppo. Industria, 1961, № 2, pp. 171—193.

[6] H. E. E c c l e s. Logistics: Conditio sine qua non for NATO Defence. Naval Research of Logistics Quarterly, VIII, 1961, pp. 111—116.

[7] Ср.: T. E. M o i s m a n n, M. B. S h a p i r o, C. R. M e r r i l, D. F. B r a d l e y and T. E. V i n t o n. Reconstruction of Protein and Nucleic Acid Sequences: IV. The Algebra of Free Monoids and the Fragmentation Stratagen. Bulletin in Mathematical Biophysics, vol. XXVIII, 1966, pp. 235 etc.; И. И. Ш м а л ь г а у з е н. Основы эволюционного процесса в свете кибернетики. Пробл. кибернетики, вып. 4, М., 1960, стр. 121—149.

[8] Ф. д е С о с с ю р. Курс общей лингвистики. Русск. пер. М., 1933, стр. 38—42.

[9] Там же, стр. 124.

[10] Там же, стр. 38; R. G o d e l. Les sources manuscrites du Cours de linguistique générale de F. de Saussure. Genève—Paris, 1957, pp. 148, 152.

[11] Ф. де С о с с ю р, ук. соч., стр. 38.

[12] Там же, стр. 42—43.

[13] Ср.: А. С е ш е. Три соссюровские лингвистики. Критика «Курса общей лингвистики». Русск. пер. ИЯ, стр. 60 и сл.; Л. Е л ь м с л е в, *ук. соч., стр. 113*; J. M o u n i n. *Ferdinand de Saussure ou le structuraliste sans le savoir*. Paris, 1968.

[14] Ср.: А. Г а р д и н е р. Различие между «речью» и «языком». Русск. пер. ИЯ, стр. 16 и сл.; А. И. С м и р н и ц к и й. Объективность существования языка. В сб.: Материалы к курсам языкознания. М., 1954; П. С. К у з н е ц о в. О языке и речи. Вестник МГУ, серия VII филологическ., 1961, № 4; С. Д. К а ц н е л ь с о н. Лингвистическая концепция Ф. де Соссюра. В сб.: Вопросы общего языкознания. Материалы республиканского семинара преподавателей общего языкознания, Л., 1967, стр. 33—35 и др.

единиц, порождает тексты (речь) потенциально бесконечной длины.[15]

Третий вопрос предусматривает определение форм существования языка,[16] или, иными словами, требует описания его разновидностей (стилей, подъязыков).

Четвертый вопрос состоит в измерении функционального веса (функциональной отдачи) отдельных лингвистических единиц относительно как порождающих возможностей системы языка в целом, так и систем его разновидностей.[17]

§ 2. Порождающие грамматики и механизмы создания текста

Разграничения языка и речи и особенно механизмы создания текста исследуются наиболее основательно в теории порождающих грамматик. Здесь и следует искать ответ на первые два вопроса антиномии «язык—речь».

Если, например, обратиться к концепции Н. Хомского, то языком (лингвистическим кодом) следовало бы считать порождающую грамматику, состоящую из словаря (основного и вспомогательного), который включает исходные лингвистические элементы, — например, буквы (соотв. фонемы, морфы) и их классы, затем некоторый начальный символ, — например, «предложение», правила подстановки (правила НС, трансформационные правила, морфонологические правила) и, наконец, терминальные цепочки.[18] В сообщение же (речь) попали бы окончательно образуемые цепочки морфонем, которые можно сопоставлять реально существующим фразам и текстам.

Все порождающие грамматики должны включать полное описание механизмов создания текстов. Однако выясняется, что все эти описания относятся к разряду так называемых контекстно-свободных грамматик, т. е. таких процедур, которые позволяют осуществлять преобразования над элементами вспомогательного словаря без учета контекстных ограничений. В результате порождается много цепочек, не являющихся реальными предложениями данного языка.[19] Эта неадекватность порожденного и естественного текста заметно снижает объяснительную силу порождающих грамматик с точки зрения исследования механизмов создания текста из элементов системы языка.[20]

Естественные языки в целом должны описываться контекстно-зависимыми грамматиками.[21] Создание таких грамматик предусматривает выявление всех контекстных ограничений, накладывающихся на процедуру порождения текста, что в свою очередь связано с рассмотрением двух последних вопросов антиномии «язык — речь».

Действительно, соссюровское положение о том, что язык это «коллективный образец», существующий «в коллективе в форме совокупности отпечатков, имеющихся в каждом мозгу, примерно как словарь, экземпляры которого, в полне тождественные (разрядка наша, — *Р. П. и Л. Т.*), находились бы в пользовании многих лиц»,[22] представляет собой довольно грубую аппроксимацию соотношения между языком, существующим «в мозгах целой совокупности индивидов», т. е. общества, и владением этим языком у отдельного члена общества, который воспроизводит его, как признает сам де Соссюр, «конечно, не вполне одинаково, но приблизительно».[23] От языка общества к языку индивида ведет целая лестница языковых разновидностей и подъязыков, имеющих все более и более узкие сферы обращения.[24] Различия между этими градациями как раз и состоят в особом выборе и разном использовании контекстов, чем никак нельзя пренебречь при построении контекстно-зависимых грамматик. С помощью дедуктивных приемов алгебраической лингвистики учесть особенности и ограничения выбора контекстов относительно языка в целом и его разновидностей — невозможно.

---

[15] См.: Н. Хомский. Логические основы лингвистической теории. Русск. пер. НЛ, вып. IV, 1965, стр. 478—479.

[16] Ср.: В. Г. Адмони. Язык как единство системы отношений и системы построения. НДВШ, 1969, № 3, стр. 3 и сл.; Н. Н. Коротков. Норма, система и структура как этапы анализа и описания языкового строя. В сб.: Спорные вопросы грамматики китайского языка, М., 1965, стр. 9 и сл.

[17] Ср.: Р. Г. Piotrowski. Моделирование фонологических систем и методы их сравнения. М.—Л., 1960, стр. 273.

[18] Ср.: N. Chomsky, G. A. Miller. Introduction to the Formal Analysis of Natural Languages. Handbook of Mathematical Psychology, vol. II, New York, 1963, pp. 290—322 (русский перевод наиболее важных разделов этой работы см.: «Кибернетический сб.», НС, вып. I, М., 1965). Несколько иное разграничение языка и речи находим в аппликативной грамматике Шаумяна и Соболевой, см.: С. К. Шаумян и П. А. Соболева. Основания порождающей грамматики русского языка. Введение в генотипические структуры. М., 1968.

[19] Н. Хомский, М. П. Шютценберже. Алгебраическая теория контекстно-свободных языков. Русск. пер. Кибернетический сб. НС, вып. 3, М., 1966, стр. 198—200.

[20] Проблема неадекватности порожденного и естественного текста является основным оружием критиков теории порождающих грамматик, ср.: W. Winter. Transforms without Kernels? «Language», vol. 41, n. 3, p. 1, 1965, pp. 484—489; N. L. Chafe. Idiomaticity as an Anomaly in the Chomskyan-Paradigm. «Foundations on Language», vol. 4, № 2, 1968, pp. 109—127; J. Grešnik. On N. Chomsky's Strict Subcategorization of Verbs. «Linguistica», Leto 8, Ljubljana, 1966—1968, str. 83—102.

[21] Это, разумеется, не исключает возможности истолкования некоторых ситуаций естественных языков с помощью контекстно-свободных грамматик, см.: S. Ginsburg, H. G. Rice. Two Families of Languages Related to ALGOL. In: Technical Memorandum. Systems Development Corporation, Santa Monica (California), 1961; S. G. Ginsburg, G. F. Rose. Operations which Preserve Definibility in Languages. Там же.

[22] Ф. де Соссюр. Курс общей лингвистики, стр. 42.

[23] Там же, стр. 38.

[24] А. Сеше. Три соссюровские лингвистики. . ., стр. 65.

Что касается тех эмпирических приемов введения контекстных ограничений, которые предлагаются авторами порождающих грамматик, то и эти приемы оказываются практически нереализуемыми. В трансформационных грамматиках часто рекомендуется, например, обращение к информанту. Но ведь использовать показания информанта можно лишь при условии, что имеется оценка того, как лингвистическое сознание этого информанта отражает систему языка. Однако получить эту оценку можно лишь при условии, что система языка уже описана с помощью контекстно-зависимой грамматики. Как выйти из этого порочного круга, *пока неизвестно.*

Остается, разумеется, еще одна возможность превращения контекстно-свободной грамматики в контекстно-зависимую: простое перечисление всех контекстных ограничений. Однако, как показывает опыт работы в области машинного перевода, задать полный список таких ограничений, опираясь на традиционные описания языка или на данные текста, практически невозможно.

Поскольку неадекватность порождаемого и естественного текста не может быть устранена в русле существующей дедуктивной методики порождающих грамматик, постольку эта теория не дает ответа ни на вопрос о формах существования языка, игнорируя таким образом всю стилистическую проблематику, ни на вопрос об измерении функционального веса лингвистических единиц. Что же касается второго вопроса антиномии «язык — речь» — вопроса о механизмах создания текста, то и он, как мы видели, не получает в теории порождающих грамматик окончательного решения.

Общеизвестно, что порождающая грамматика усиленно ищет приемы интерпретации своих схем в реальном тексте, пытаясь с помощью методов алгебраической лингвистики дать алгоритмы описания контекстных ограничений.[25] Однако и при условии успеха этого подхода оценить функциональный вес отдельных лингвистических единиц не удается.

Разумеется, тот факт, что теория порождающих грамматик не дает ответа на все вопросы антиномии «язык — речь», никак не может умалить того большого значения, которое имеет эта теория в современном языкознании. Возможности каждой теории в достаточной степени ограничены в смысле решения основных вопросов данной науки. Если рассматриваемая теория дает ответ *лишь на часть интересующих нас вопросов, то мы должны* быть благодарны ей и за это. Для изучения неразрешенных вопросов следует привлечь иные теории и другие методы.

[25] Е. В. Гулыга. Словообразование и формообразование в аппликативной модели (на материале французского языка). АКД. Бельцы, 1969.

§ 3. Язык и речь или «система — норма — речь»?

Вопрос о границах языка и речи, равно как и вопрос о формах существования языка, рассматривался во многих работах по теоретическому (нематематическому) языкознанию. При этом уже первые интерпретаторы де Соссюра указывали, что при рассмотрении соотношения языка и речи целесообразно использовать еще одну категорию, занимающую промежуточное и связующее положение между языком, выступающим как обязательная для коллектива «система знаков»[26] и являющим собой устойчивое «равновесие взаимно обусловливающих себя элементов»,[27] и речью, проявления которой «индивидуальны и мгновенны», где «нет ничего, кроме суммы частных случаев».[28] Этот промежуточный компонент имплицитно присутствует в рассуждениях А. Сеше, предлагающего выделять лингвистику организованной речи и подчеркивающего, что язык является «системой произвольных знаков, совокупность которых составляет узус в какой-то определенный момент в обществе»[29] (разрядка наша, — *Р. П. и Л. Т.*). Об этом третьем промежуточном компоненте — узусе *говорит* также Л. Ельмслев. При этом *под* термином «узус» понимается не система чистых отношений, но «язык как совокупность навыков, принятых в данном социальном коллективе и определяемых фактами наблюдаемых манифестаций».[30]

Наиболее подробно вопрос о промежуточном компоненте рассмотрен в работах Э. Косериу.[31]

Существо концепции этого автора состоит в том, что, принимая в целом соссюровское противопоставление «язык — речь», он предлагает различать в первом члене этого противопоставления, языке, два аспекта: «систему» и «норму».

В результате соссюровская двухчленная схема речевой деятельности *«язык — речь» заменяется трехчленной схемой* «система — норма — речь».

Сам Э. Косериу следующим образом расшифровывает понятия системы и нормы.

«Система есть „система возможностей, координат, которые указывают открытые и закрытые пути" в речи, „понятной" данному коллективу. Норма, напротив, это „система обязательных реализаций",. . . принятых в данном обществе и данной культурой: норма соответствует не тому, что можно сказать, а тому, что уже „сказано" и что по традиции „говорится" в рассматриваемом об-

[26] Ф. де Соссюр, ук. соч., стр. 40.
[27] Там же, стр. 120.
[28] Там же, стр. 43, R. Godel. Les sources manuscrites du Cours de linguistique générale. . ., p. 155.
[29] А. Сеше, ук. соч., стр. 61—62.
[30] Л. Ельмслев. Язык и речь, стр. 113.
[31] E. Coseriu. 1) Sistema, norma y habla. Montevideo, 1952; 2) Forma y sustancia en los sonidos del lingnaje. Montevideo, 1954, pp. 25—32; 3) Синхрония, диахрония и история. Русск. пер. НЛ, вып. III, 1963, стр. 156 и сл.

ществе. Система охватывает идеальные формы реализации определенного языка, т. е. технику и эталоны для соответствующей языковой деятельности; норма же включает модели, исторически уже реализованные с помощью этой техники и по этим шаблонам».[32]

Итак, понятие нормы (соотв. узус у Л. Ельмслева) включает те ограничения, которые накладываются на систему «возможностей, координат». Используя терминологию порождающих грамматик, можно сказать, что комбинация системы и нормы создает контекстно-зависимую грамматику.

А теперь посмотрим, как соотнесены норма и речь.

«Языковые элементы, которые обнаруживаются в тексте, представляют собой, — по мнению Э. Косериу, — . . . языковые навыки» говорящих, определяемые нормой. «Для каждого говорящего, — продолжает он, — язык — это умение говорить, знание того, как говорят в определенном обществе и в соответствии с определенной традицией. На основе такого знания говорящий создает свои высказывания, которые в той мере, в какой они совпадают с высказываниями других говорящих, . . . составляют (или могут составлять) язык, засвидетельствованный в речи. В этом смысле всякий говорящий лишь в исключительных случаях создает свои собственные модели; языковые навыки он непрерывно приобретает от других говорящих».[33]

В приведенных рассуждениях Э. Косериу имплицитно присутствует еще один аспект общего понятия нормы — идея нормированности, организованности текста (ср. уже упоминавшийся призыв А. Сеше о создании лингвистики «организованной речи», а также мысль Л. В. Щербы о том, что речевая деятельность индивида вместе с обусловленным ею текстом является социальным продуктом).[34] Эта, пока еще интуитивная, «ясно-смутная», как говорил Лейбниц, идея найдет впоследствии свое оформление в понятиях «система речи» [35] или «норма речи», противопоставляемых понятию «норма языка».[36]

---

[32] Там же, стр. 174—175; ср.: E. Coseriu. Sistema norma y habla, стр. 59.

[33] Синхрония, диахрония и история, стр. 176.

[34] Л. В. Щерба. О трояком аспекте языковых явлений и об эксперименте в языкознании. ИЯ, стр. 362. Ср. сходное понимание нормы у Б. Гавранека, который рассматривает ее как «совокупность употребляемых языковых средств в отличие от конкретных языковых высказываний, где можно констатировать только то, что имеется» (Б. Гавранек. Задачи литературного языка и его культура. Русск. пер., в сб.: Пражский лингвистический кружок, М., 1967, стр. 339).

[35] Н. Д. Андреев. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л., 1967, стр. 22.

[36] В. Г. Гак. Проблема лексико-грамматической организации предложения (на материале французского языка в сопоставлении с русским). АДД. М., 1968, стр. 8.

Новое научное понятие может дать заметный прогресс в науке лишь тогда, когда применение этого понятия опирается на разработанную методику и технологию его использования. Одно лишь постулирование понятия нормы (соотв. нормы языка, нормы речи) лишь в малой степени продвигает разрешение трудных вопросов антиномии «язык — речь».

Одной из основных технологических и методических задач, связанных с применением понятия нормы, является определение меры нормированности в языке и речи. Введение такой меры становится обязательным условием при разрешении вопроса о функциональном весе отдельных лингвистических единиц.

При выборе меры нормированности нужно учитывать следующие особенности нормы.

Во-первых, существование нормы языка или его разновидности «предполагает возможность выбора» [37] определенной лингвистической единицы при условии существования иных вариативных (синонимических) возможностей.

Во-вторых, существование нормы речи предусматривает некоторую упорядоченность в той массе повторяющихся лингвистических элементов, которые составляют любой текст.

При исследовании повторяющихся массовых явлений должны применяться методы, чувствительные к категориям порядка и беспорядка. К таким методам относятся в первую очередь приемы теории вероятностей и математической статистики. Вариативность (синонимичность) средств языка предполагает случайность в выборе лингвистических единиц плана выражения; закономерности же, связанные со случайностью выбора, изучаются, как известно, также с помощью аппарата теории вероятностей и математической статистики. Наконец, употребительность отдельных лингвистических единиц, связанная с их функциональным весом, может быть измерена с помощью вероятности — основного понятия теории вероятностей и математической статистики.

Кстати, само понятие нормы имплицитно предусматривает присутствие вероятностных оценок; правда, таких оценок здесь только две: *1* — допустимость употребления данного варианта и *0* — запрещение его употребления. Однако, как хорошо показал Н. Д. Андреев, использование только двух этих оценок не может описать «реальную картину языковой действительности»; [38] чтобы получить достаточно богатое описание норм речи, необходимо использовать весь интервал вероятностной меры от нуля до единицы.

---

[37] Г. В. Степанов. О двух аспектах понятия языковой нормы (на испанском материале). В сб.: Методы сравнительно-сопоставительного изучения современных романских языков, М., 1966, стр. 226.

[38] Н. Д. Андреев, ук. соч., стр. 21.

Таким образом, сама логика разрешения антиномии «язык — речь» приводит языкознание к использованию статистического аппарата и вероятностной меры.

Само собой разумеется, что все наши рассуждения носят пока умозрительный и предположительный характер. О вероятностно-статистической природе нормы можно будет говорить лишь в том случае, если это свойство нормы будет выявлено с помощью научного эксперимента. Более того, если наш эксперимент обнаружит в тексте такую статистическую упорядоченность, которая отражает лингвистическую стабильность нормы,[39] мы сможем более уверенно говорить о целесообразности введения третьего компонента в антиномию «язык — речь». Если же наш эксперимент не обнаружит статистической упорядоченности и нормированности в тексте, то целесообразность самого понятия нормы будет поставлена под вопрос. Иными словами, основной технологической задачей нашей работы является постановка такого эксперимента, который недвусмысленно ответил бы на вопрос о том, которая из двух схем — новая трехчленная схема Э. Косериу или старая двухкомпонентная схема Ф. де Соссюра — является наиболее удобной для описания взаимодействия языкового кода и порождаемого им речевого сообщения.

Предположим, что 1) наш эксперимент даст положительный результат, 2) будет установлено, что норма представляет собой производящую систему языковых навыков, которые измеряются вероятностной мерой, а реализацией этой системы является текст (речь), выступающий в виде последовательности дискретных случайных событий. В таком случае возникает следующий вопрос: с помощью какой вероятностной схемы (или схем) следует описывать основные лингвистические свойства этой последовательности. Этот вопрос нельзя считать праздным: хотя и установлено, что статистическая структура текста «в основном идентична конечной марковской цепи с быстро затухающими связями вправо»,[40] на практике мы постоянно сталкиваемся с примерами того, что t-связная цепь Маркова речевого текста может быть успешно

[39] Следует иметь в виду, что далеко не всякую статистическую упорядоченность текста, написанного на естественном языке, следует считать признаком существования нормы. В частности, ни закон Эсту—Ципфа—Мандельброта (см.: B. M a n d e l b r o t. On the Theory of Word Frequencies and on Related Markovian Models of Discourse. Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics, vol. XII, Providence, Rhode Island, 1961, pp. 190—219), ни наблюдения над статистической структурой текста, осуществленные Юлом (см.: G. U. Y u l e. The Statistical Study of Literary Vocabulary. Cambridge, 1944) или Эпштейном (см.: H. J o s s e l s o n. The Russian Word Count. Detroit, 1953, pp. 23—35), еще не говорят об объективном существовании нормы.

[40] B. M a n d e l b r o t, ук. соч., стр. 191.

оценена с помощью схем «*нуль*-связных цепей», в частности с помощью схемы Пуассона.[41]

Для разрешения указанных вопросов мы воспользуемся результатами проведенного Л. А. Турыгиной статистического обследования современных английских публицистических текстов общей длиной в 200 тыс. словоупотреблений.[42]

Глава I. НОРМА И ЕЕ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ ПРИРОДА

§ 1. Устойчивость распределения вероятностей употребления словоформ как показатель нормированности текста

Изучение проблемы нормированности текста мы начнем с исследования устойчивости распределения вероятностей абсолютных частот $P(F)$ словоформ. Как показывают исследования предшественников, эмпирические распределения большого числа словоформ не подчиняются известным в математической статистике теоретическим схемам,[43] иногда отнесение эмпирического распределения к той или иной теоретической схеме зависит от условий нормировки (ср. статью К. Б. Бектаева и К. Ф. Лукьяненкова в настоящем сборнике).[44] Поэтому мы не будем заниматься сопоставлением эмпирических и теоретических распределений, а, считая, что функции $P(F)$ для всех словоформ неизвестны, сосредоточим внимание на сравнении эмпирических распределений отдельных словоформ в разных выборках.

Возьмем две выборки из английских публицистических текстов по 100 тыс. словоупотреблений каждая. В свою очередь каждая из выборок разбивается на 20 порций по 5 тыс. словоупотреблений.

Фиксируя частоты данной словоформы в каждой порции, получаем распределения этой словоформы относительно первой и вто-

[41] В. М. К а л и н и н. Развитие схемы Пуассона и ее применение для описания статистических свойств речи. АКД. Л., 1964; ср. также: Р. Г. П и о т р о в с к и й. Информационные измерения языка: Л., 1968, стр. 16; Р. А. К а з а р я н. Оценка энтропии армянского текста. Изв. АН Армянской ССР, физико-математические науки, XIV, 4, Ереван, 1961.

[42] См.: Л. А. Т у р ы г и н а. 1) Частотный словарь английских и американских газетных текстов. СР, стр. 180—184; 2) Статистическая интерпретация антиномии язык и речь. АКД. Л., 1970.

[43] Л. Е. М а ш к и н а. О статистических методах исследования лексико-грамматической дистрибуции. АКД. Минск, 1968, стр. 13—28; Т. А. М и к е р и н а. Некоторые статистические приемы лексико-морфологического описания функционального стиля (на материале английских текстов по судостроению). АКД. Л., 1967, стр. 14—15; О. А. Н е х а й. Статистика и автоматический анализ текста. АКД. Минск, 1968, стр. 10.

[44] Ср. также: К. Ф. Л у к ь я н е н к о в. Лексико-статистическое описание английского научно-технического текста с помощью электронно-вычислительной машины (подъязык судовых механизмов). АКД. Минск, 1969.

рой выборок. Получив распределения для той или иной словоформы, мы обращаемся к сравнению этих распределений. Как правило, характеристики этих распределений точно не совпадают, но отличаются одна от другой на некоторую величину.

При интерпретации таких различий мы используем метод гипотез. Применительно к задачам нашей работы существо этого метода сводится к следующему. Формулируются две гипотезы — нулевая (H_0) и альтернативная (H_1). Согласно нулевой гипотезе, расхождения между распределениями словоформ в обеих выборках являются результатом случайных (нелингвистических) колебаний. А это свидетельствует, что статистическое строение текста обусловлено определенной нормой.

Альтернативная гипотеза утверждает, что расхождения в обеих выборках не случайны, а обусловлены лингвистическими факторами; большое число словоформ не имеет устойчивого распределения $P(F)$. Иными словами, если оказывается справедливой альтернативная гипотеза, это свидетельствует об отсутствии статистической нормированности текста.

Поскольку функции $P(F)$ для исследуемых словоформ неизвестны, необходимо при сравнении распределений и при анализе гипотез выбрать такой прием, который не зависел бы от вида исследуемой функции распределения. В математической статистике в таких случаях используются так называемые порядковые критерии, в частности критерий Смирнова, критерий Вилкоксона и $\varkappa$-критерий.[45] Для массовых обследований текстов наиболее пригоден критерий Вилкоксона: он более прост в практическом употреблении и при этом в той же степени надежен, как и два других критерия.

Рассмотрим технологию этого критерия. Пусть имеются две независимые выборки. Первая разбита на n_1, а вторая на n_2 порций. Интересующая нас словоформа встретилась в i-той порции первой группы $F_i^{(\mathrm{I})}$ раз ($1 \leqslant i \leqslant n_1$), а в j-той порции второй группы — $F_j^{(\mathrm{II})}$ раз ($1 \leqslant j \leqslant n_2$). Расположим теперь все значения $F_i^{(\mathrm{I})}$ и $F_j^{(\mathrm{II})}$ в одну строчку в порядке возрастания численного значения F, не обращая внимания на верхние и нижние индексы. В результате получаем смешанный вариационный ряд. Инверсией в таком ряду называется случай, когда F_j^{II} располагается перед F_i^{I} независимо от точного положения F_i^{I} и F_j^{II} в вариационном ряду. Полная сумма u числа инверсий в вариационном ряду есть случайная величина, численное значение которой, установленное в результате опыта, и является критерием Вилкоксона.

В качестве примера рассмотрим распределение частот словоформы government 'правительство' по порциям двух независимых выборок (как уже указывалось, каждая выборка имеет длину

[45] Л. Б. ван дер Варден. *Математическая статистика*. Русск. пер. М., 1960, стр. 325—357.

в 100 тыс. словоупотреблений и охватывает 20 порций по 5 тыс. словоупотреблений каждая). Распределение частот по порциям представлено в табл. А.

Таблица А

Распределение частот словоформы government

Порции (i) 1-й выборки	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Частоты (F_i^{I}) с-формы government в 1-й выборке	1	3	12	2	14	3	12	5	8	4	7	3	6	10	5	3	4	9	4	8	
Частоты (F_j^{II}) с-формы government во 2-й выборке	2	4	1	7	1	6	3	10	5	8	14	1	4	10	13	7	8	6	7	3	
Порции (j) 2-й выборки	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	

Расположим теперь все значения F_i^{I} и F_j^{II} в порядке возрастания численного значения F, не обращая внимания на верхние и нижние индексы. В тех случаях, когда величины $F_i^{\mathrm{I}} = F_j^{\mathrm{II}}$, вопрос об их взаимном расположении решается путем жеребьевки. Исходя из этих условий, получаем следующий вариационный ряд:

$$\begin{gathered} F_9^{\mathrm{II}} F_1^{\mathrm{II}} F_5^{\mathrm{II}} F_{12}^{\mathrm{II}} F_4^{\mathrm{II}} F_1^{\mathrm{II}} F_6^{\mathrm{II}} F_1^{\mathrm{I}} F_{16}^{\mathrm{II}} F_7^{\mathrm{II}} F_2^{\mathrm{II}} F_{12}^{\mathrm{I}} F_{20}^{\mathrm{II}} F_{10}^{\mathrm{I}} F_{19}^{\mathrm{II}} \\ F_2^{\mathrm{II}} F_{17}^{\mathrm{II}} F_{13}^{\mathrm{II}} F_8^{\mathrm{II}} F_9^{\mathrm{II}} F_{15}^{\mathrm{II}} F_{13}^{\mathrm{I}} F_6^{\mathrm{II}} F_{18}^{\mathrm{II}} F_4^{\mathrm{II}} F_{11}^{\mathrm{I}} F_{16}^{\mathrm{II}} \\ F_{10}^{\mathrm{II}} F_{20}^{\mathrm{I}} F_{10}^{\mathrm{II}} F_{17}^{\mathrm{II}} F_9^{\mathrm{I}} F_{18}^{\mathrm{II}} F_{14}^{\mathrm{I}} F_8^{\mathrm{I}} F_{14}^{\mathrm{II}} F_3^{\mathrm{II}} F_7^{\mathrm{I}} F_{15}^{\mathrm{II}} F_{11}^{\mathrm{II}} F_5^{\mathrm{I}}. \end{gathered} \quad (1)$$

Число инверсий u подсчитывается здесь следующим образом. F_1^{I} дает одну инверсию с F_3^{II}; F_4^{I} дает три инверсии с F_3^{II}, F_5^{II}, F_{12}^{II}; F_6^{I} имеет четыре инверсии с F_3^{II}, F_5^{II}, F_{12}^{II}, F_1^{II} и т. д. Всего имеем

$$\begin{gathered} 1+3+4+4+5+5+6+6+7+8+9+ \\ +9+12+14+16+16+16+18+18+ \\ +20=197. \end{gathered} \quad (2)$$

А теперь попытаемся выяснить, о чем говорит полученное нами численное значение u: требует ли оно принять или отвергнуть нашу нулевую гипотезу?

Согласно критерию Вилкоксона, нулевая гипотеза должна быть отвергнута, если количество инверсий u выходит за некоторые пределы

$$u_1(P_\beta) \text{ и } u_2(P_\beta), \quad (3)$$

где $$u_1(P_\beta) < u_2(P_\beta),$$

а P_β — уровень значимости критерия. Действительно, оказывается, что если нулевая гипотеза верна, то случайная величина имеет определенное распределение вероятностей с математическим ожиданием

$$\bar{u} = \frac{n_1 \cdot n_2}{2} \qquad (4)$$

и дисперсией $\sigma^2(u) = \frac{n_1 \cdot n_2}{12}(n_1 + n_2 + 1)$.

Если $n_1 > 3$ и $n_1 + n_2 \geqslant 20$, то это распределение с достаточной точностью может считаться нормальным.[46] Последнее обстоятельство позволяет определить пределы $u_1(P_\beta)$ и $u_2(P_\beta)$, если уровень значимости P_β задан. Введем нормированное уклонение

$$z = \frac{u - \bar{u}}{\sigma(u)}, \qquad (5)$$

и тогда

$$z_1 = \frac{u_1 - \bar{u}}{\sigma(u)} \quad \text{и} \quad z_2 = \frac{u_2 - \bar{u}}{\sigma(u)}.$$

В силу вышесказанного случайная величина z имеет нормальное распределение с параметрами: математическое ожидание 0 и дисперсия *1*.

Уровень значимости P_β *есть не что иное, как вероятность того*, что вследствие случайных колебаний величина u выйдет за пределы u_1 и u_2, а z — за пределы z_1 и z_2. Обычно выбирают $z_1 = -z_2$. Тогда связь между u_1, u_2 и P_β может быть найдена из очевидных соотношений:

$$P_\beta = 1 - \frac{2}{\sqrt{2\pi}} \int_0^{z_2} e^{-\frac{z^2}{2}} dz = 1 - 2\Phi(z_2), \quad \Phi(z_2) = \frac{1 - P_\beta}{2}, \qquad (6)$$

где $\Phi(z_2)$ — известный интеграл вероятностей. Таким образом, при заданном P_β соотношение (6) *определяет* z_2. И тогда легко находятся пределы u_1 и u_2:

$$\left.\begin{aligned} &\frac{u_1 - \bar{u}}{\sigma(u)} = -z_2, \quad u_1 = -z_2 \cdot \sigma(u) + \bar{u} = \\ &= -z_2 \sqrt{\frac{n_1 \cdot n_2}{12}(n_1 + n_2 + 1)} + \frac{n_1 \cdot n_2}{2}, \\ &\frac{u_2 - \bar{u}}{\sigma(u)} = z_2, \quad u_2 = z_2 \sqrt{\frac{n_1 \cdot n_2}{12}(n_1 + n_2 + 1)} + \frac{n_1 \cdot n_2}{2} \end{aligned}\right\}. \qquad (7)$$

Мы выбрали уровень значимости $P_\beta = 0.01$; *это означает*, что при справедливости нулевой гипотезы из 100 значений критерия Вилкоксона в среднем лишь один может выходить за пределы u_1 и u_2 (обычно P_β берется в интервале 0.01—0.05). При $P_\beta = 0.01$ соотношение (6) позволяет найти z_2, которое оказалось равным 2.6.

[46] Там же, стр. 342 и сл.

Полученное значение $z_2 = 2.6$ свидетельствует, что выбор уровня значимости $P_\beta = 0.01$ хорошо согласуется с известным в математической статистике «правилом трех сигм — 3σ», которое утверждает, что если некоторая случайная величина уклоняется *на опыте от своего математического ожидания на величину*, превышающую 3σ (при этом как раз $z \geqslant 3$), то это происходит, как правило, или за счет неслучайного воздействия на нее, или за счет изменения условий наблюдения этой величины. Последнее принципиально меняет характер распределения вероятностей случайной величины.

В нашем эксперименте $n_1 = n_2 = 20$, поэтому условия $n_1 > 3$ и $n_1 + n_2 \geqslant 20$ выполнены, u и $\sigma(u)$ оказываются при этом равными

$$\bar{u} = \frac{n_1 \cdot n_2}{2} = \frac{20 \cdot 20}{2} = 200,$$

$$\sigma(u) = \sqrt{\frac{n_1 \cdot n_2}{12}(n_1 + n_2 + 1)} = \sqrt{\frac{20 \cdot 20}{12} \cdot 41} = 37. \qquad (8)$$

В соответствии с выражением (7) *однопроцентные доверительные пределы* для критерия Вилкоксона в нашем случае будут равны:

$$u_1 = 200 - 2.6 \cdot 37 = 104, \quad u_2 = 200 + 2.6 \cdot 37 = 296. \qquad (9)$$

Нетрудно заметить, что значение u для контрольного слова government попадает в только что указанный доверительный интервал. Это дает нам право с уверенностью в 99% утверждать, что расхождения между распределениями government в обеих выборках имеют случайный, нелингвистический характер.

Результаты исследования устойчивости $P(F)$ для других словоформ из нашего частотного списка приводятся в табл. 1—7. Каждая таблица охватывает словоформы, относящиеся к определенной части речи. При этом все рассматриваемые словоформы относятся к трем группам частот общего списка: I группа: $60 \leqslant F \leqslant 15000$, II группа: $25 \leqslant F \leqslant 40$, III группа: $5 \leqslant F \leqslant 20$. Табл. 1—7 включают в общей сложности 600 словоформ. Из них только одна, began ($u = 304$), дала значение критерия Вилкоксона, *выходящее, притом очень незначительно*, за однопроцентные доверительные пределы. Между тем нулевая гипотеза оставалась бы справедливой, если бы таких случаев было шесть. На основании этого можно сделать вывод, что словоформы в английских публицистических текстах имеют устойчивые распределения $P(F)$. Если в этих текстах и встречаются лингвистические единицы с неустойчивыми распределениями $P(F)$, то вклад таких единиц не превышает здесь $\frac{1}{600} \times 100\% = 0.17\%$. Иными словами, употребление словоформ в публицистических текстах, судя по полученным величинам критерия Вилкоксона, подчиняется некоторой норме.

Таблица 1

Значения критерия Вилкоксона для именных словоформ

Словоформы	F*	u**	m***	Словоформы	F	u	m
Mr.	853	147	8	bill	87	180	12
year	298	158	10	days	86	172	12
time	282	232	8	war	86	166	10
years	246	238	11	conference	85	212	9
government	243	197	9	Ford	85	161	7
president	241	194	10	major	82	226	8
London	181	223	7	policy	82	164	10
people	180	197	10	area	81	232	6
Mrs.	179	191	9	business	78	105	11
world	179	209	8	communist	76	189	7
party	161	192	10	cup	76	240	8
week	159	246	7	members	75	215	9
cent	157	189	11	place	75	180	12
committee	156	183	8	countries	74	175	10
day	154	166	11	number	74	155	10
night	149	183	10	action	72	203	14
public	147	209	12	England	72	218	10
meeting	134	192	6	leader	72	169	10
minister	134	192	10	leaders	72	172	12
house	131	147	9	power	72	203	11
May	129	178	10	match	71	122	10
man	127	130	12	negro	71	122	12
part	123	175	10	defence	69	147	8
state	123	175	11	industry	69	158	10
work	121	200	12	association	68	186	12
West	121	200	12	interest	68	147	7
Dr.	120	161	11	miles	68	180	6
labour	120	186	15	congress	67	200	11
Sir	120	175	10	program	67	133	9
company	*119*	*155*	*14*	*final*	*66*	*141*	*9*
men	116	155	10	money	66	133	10
Britain	110	133	6	building	65	141	13
country	107	153	13	position	65	144	11
court	107	186	9	top	65	185	10
life	107	172	12	north	64	252	12
council	105	186	10	children	62	153	12
end	103	119	6	press	62	274	14
city	101	122	10	funeral	57	152	11
half	97	246	8	Aug.	40	147	9
police	94	175	10	college	40	178	11
secretary	94	206	8	death	40	246	10
chairman	91	180	8	evidence	40	147	12
board	89	215	9	island	40	153	11
office	89	200	12	opposition	40	175	8
months	88	167	10				

* — абсолютная частота.
** — значение критерия Вилкоксона.
*** — значение критерия знаков.

Таблица 1 (*продолжение*)

Словоформы	F	u	m	Словоформы	F	u	m
act	39	155	10	accounts	10	215	10
plans	39	119	7	admission	10	158	11
prince	39	169	8	anniversary	10	200	9
education	38	194	11	author	10	197	12
Europe	38	141	6	Asia	10	232	12
forces	38	147	8	baseball	10	203	10
Moscow	38	150	10	belief	10	271	9
period	38	138	7	benefit	10	203	10
steel	38	179	10	branch	10	226	6
arms	37	150	10	Bristol	10	229	11
boy	37	138	9	bureau	10	200	12
face	37	144	8	buyers	10	189	8
army	35	130	10	check	10	175	12
campaign	35	127	6	circle	10	175	12
football	35	127	9	waters	10	200	13
nation	35	138	10	wood	10	172	10
water	35	194	9	agents	9	220	14
budget	34	144	7	bankers	9	243	10
mother	34	119	7	Brooklyn	9	175	13
movement	34	116	10	counties	9	200	9
works	34	178	5	wheel	9	232	10
dividend	33	192	10	woods	9	138	7
information	33	113	10	article	8	178	10
Monday	33	150	8	cargo	8	212	9
April	32	183	11	center	8	203	10
areas	32	138	7	faculty	8	197	8
assurance	32	136	9	hats	8	206	10
balance	31	161	8	newsmen	8	229	9
Brown	31	158	10	acceptance	7	223	8
captain	31	158	9	writer	8	229	5
crowd	31	147	14	actress	7	144	10
advantage	*30*	*127*	*10*	boat	*7*	*200*	*10*
authority	30	158	8	bodies	7	200	8
capital	30	127	12	critics	7	223	7
cents	30	153	14	cubs	7	215	10
addition	29	172	12	dream	7	203	6
average	29	136	10	investments	7	197	10
body	29	153	11	makers	7	191	12
convention	28	218	8	composers	6	203	11
economy	28	122	10	Corp.	6	229	10
ground	28	136	12	bottom	6	200	8
banks	27	127	9	centres	6	215	9
brother	27	164	10	writers	6	200	8
choice	27	257	9	celebration	5	229	11
competition	27	178	10	animals	5	200	9
affairs	26	144	11	duties	5	229	7
amount	26	133	10	weapon	5	215	7
boys	26	127	15	villa	5	200	10
republic	26	238	10				

Таблица 2

Значения критерия Вилкоксона для адъективных словоформ

Словоформы	F	u	m	Словоформы	F	u	m
new	482	185	13	close	45	197	10
united	255	185	12	royal	43	212	10
British	245	188	9	late	42	238	8
American	185	191	10	open	42	191	11
general	165	212	8	private	40	182	6
national	164	217	11	red	40	209	12
good	142	191	10	hard	39	232	10
next	118	169	15	necessary	39	226	12
great	116	220	12	real	39	105	10
long	116	155	8	social	39	164	7
high	108	188	10	certain	37	209	12
own	105	172	11	Chinese	37	182	9
political	105	185	10	heavy	37	179	11
far	98	197	6	civil	36	152	8
nuclear	94	249	9	clear	35	164	9
Soviet	94	212	8	financial	35	147	12
foreign	90	152	13	latest	35	209	8
international	84	178	11	medical	34	223	11
little	84	194	14	net	34	200	10
military	83	200	13	previous	34	209	6
white	82	317	9	Russian	33	182	10
better	81	167	10	greater	31	277	8
local	80	260	14	various	31	194	11
old	76	215	11	democratic	30	226	9
big	74	175	14	gold	30	152	14
total	71	243	10	modern	30	229	12
later	70	191	9	ready	30	235	14
industrial	67	191	12	bad	29	206	9
best	66	194	10	due	28	191	10
possible	66	243	12	human	28	232	6
small	66	242	14	northern	27	194	10
right	65	169	6	straight	27	215	7
federal	64	194	9	commercial	25	191	8
large	63	212	10	difficult	25	185	10
main	61	164	10	direct	25	221	9
able	60	158	8	agricultural	10	158	5
full	60	226	10	broad	10	246	11
recent	59	238	9	broadcast	10	175	10
central	58	194	12	collective	10	229	11
future	57	268	10	competitive	10	229	12
higher	57	152	12	constitutional	10	172	10
several	57	152	11	dangerous	10	263	8
special	57	223	12	attractive	9	189	9
economic	56	158	10	deep	9	200	10
free	55	203	10	confident	8	172	8
French	54	220	11	dramatic	7	232	7
official	53	235	12	dry	7	200	7
round	51	226	8	busy	6	169	39
annual	48	188	8	acceptable	5	200	8
strong	48	206	12	deeper	5	175	10
short	47	185	9	cheap	5	170	10

Таблица 3

Значения критерия Вилкоксона для глагольных словоформ

Словоформы	F	u	m	Словоформы	F	u	m
is	1927	172	10	brought	68	155	7
was	1811	147	8	lost	66	271	12
be	1345	169	11	does	65	178	9
will	825	161	6	reported	65	175	12
have	802	182	11	says	63	178	7
has	798	195	10	think	63	158	10
said	718	212	7	announced	60	164	10
had	692	185	15	making	60	212	12
were	683	206	5	become	58	169	8
been	632	203	9	find	58	178	10
are	595	152	13	gave	58	212	12
would	512	206	7	expected	57	144	8
last	467	209	10	help	54	194	11
made	281	149	12	leading	54	178	10
can	223	122	16	know	53	188	12
could	214	170	10	began	52	304	9
may	191	176	9	cost	52	191	11
being	165	152	11	increased	52	141	10
do	164	169	12	including	52	155	8
like	145	130	8	provide	40	187	10
make	145	203	11	showed	40	200	9
go	132	191	15	done	39	209	8
must	131	200	10	need	39	215	13
told	121	263	12	played	39	155	6
did	120	175	10	reached	39	226	14
play	119	141	11	run	39	250	9
put	114	158	12	appeared	38	268	11
take	112	226	8	decided	38	167	10
get	107	150	10	refused	37	200	17
called	102	185	12	became	37	220	10
came	97	144	9	don't	37	147	12
won	96	212	12	rise	37	175	10
taken	94	161	8	return	37	200	13
went	94	141	11	believe	36	207	9
left	92	155	10	let	36	147	12
found	91	257	10	live	36	191	11
took	91	191	9	call	35	185	10
increase	88	147	11	continue	35	282	11
given	87	251	10	include	35	209	12
show	86	164	10	died	34	206	8
see	84	220	11	heard	34	209	10
held	82	229	13	met	34	220	6
might	82	182	10	offer	34	229	10
used	78	229	9	exchange	33	185	10
asked	75	260	10	offered	33	158	11
added	72	182	11	runs	33	212	10
come	72	197	10	spent	33	176	12
going	72	167	10	taking	33	238	9
say	71	194	14	thought	33	166	13
give	69	169	8	am	32	208	11
got	69	250	10	caused	32	176	8

Таблица 3 (*продолжение*)

Словоформы	F	u	m	Словоформы	F	u	m
change	32	194	13	achieved	10	144	9
comes	32	194	10	agree	10	200	10
continued	32	174	9	alleged	10	172	6
provided	31	208	8	applied	10	186	8
published	31	231	10	arranged	10	186	10
carried	31	203	10	buying	10	200	12
charge	30	220	8	defeated	10	215	8
considered	30	206	10	demonstrated	10	200	14
scored	30	191	14	driving	10	144	10
turned	30	176	6	breaking	9	186	6
written	30	191	13	demanded	9	186	7
beat	29	266	10	drawing	9	197	10
broke	29	212	12	acquired	8	186	8
coming	29	150	7	alter	8	186	11
worked	28	191	10	associated	8	200	8
believed	27	229	8	avoid	8	153	10
felt	27	174	9	convinced	8	215	11
looking	27	203	13	directed	8	186	14
carry	26	235	8	assured	7	186	10
changes	26	200	10	concluded	7	200	7
looked	26	229	12	drink	7	200	12
makes	26	218	10	assist	6	200	10
returned	26	212	12	deciding	6	200	12
turn	26	223	8	attributed	5	200	10

Таблица 4

Значения критерия Вилкоксона для местоименных словоформ

Словоформы	F	u	m	Словоформы	F	u	m
that	1962	116	10	both	127	221	14
he	1377	226	12	us	119	155	10
it	1292	194	13	those	113	161	12
his	950	212	12	another	108	175	14
this	825	147	13	me	107	188	12
which	688	209	6	my	102	179	10
they	677	200	8	each	83	194	12
who	580	218	13	your	57	191	13
their	503	209	10	whose	50	119	9
I	495	223	12	it's	44	243	11
no	358	185	11	themselves	43	206	12
its	345	203	8	whom	27	144	10
we	330	169	14	anything	20	182	9
her	242	215	10	I'm	20	185	11
she	234	150	9	that's	20	185	9
them	226	181	14	he's	15	243	10
him	223	182	10	none	13	194	12
what	211	212	14	I've	10	188	10
our	189	223	12	herself	8	188	12
these	175	229	14	myself	6	200	13
you	164	141	7	they're	6	200	10
such	152	130	8				

Таблица 5

Значения критерия Вилкоксона для адвербиальных словоформ

Словоформы	F	u	m	Словоформы	F	u	m
as	1002	164	10	already	71	209	12
not	734	197	15	away	70	212	8
there	439	178	7	once	57	226	11
when	436	194	14	nearly	54	220	11
up	374	206	8	yet	53	277	10
out	336	194	9	probably	52	235	12
so	270	220	10	near	51	158	8
now	246	191	6	why	48	133	14
today	236	237	12	outside	44	108	12
also	212	240	10	perhaps	43	164	7
well	200	169	12	always	40	152	14
off	157	164	10	around	37	197	10
then	146	188	11	often	35	229	8
even	143	127	10	ahead	34	167	7
where	143	238	12	along	33	221	6
much	142	229	6	above	30	235	10
just	137	209	10	across	29	169	8
down	130	169	9	frequently	10	200	8
very	130	235	10	nevertheless	9	197	8
how	98	217	8	absolutely	8	172	10
however	91	188	11	exactly	8	200	10
ago	85	246	10	anywhere	7	191	7
never	85	249	11	otherwise	7	200	7
again	80	240	9	directly	6	215	13
almost	73	191	10				

Таблица 6

Значения критерия Вилкоксона для числительных

Словоформы	F	u	m	Словоформы	F	u	m
one	526	172	10	nine	29	229	10
first	375	158	11	ten	27	200	14
two	362	200	10	billion	22	197	12
four	145	217	12	fifth	22	188	8
second	138	200	7	thousand	16	200	14
million	121	203	12	sixth	14	212	9
five	120	191	10	eighth	13	191	14
third	63	175	13	twelve	7	200	10
seven	44	194	5	eleven	7	215	10
fourth	29	212	12	fifty	5	209	10

Таблица 7

Значения критерия Вилкоксона для служебных слов

Словоформы	F	u	m	Словоформы	F	u	m
the	15004	172	10	before	226	226	11
of	7030	169	11	against	195	191	8
to	4834	150	10	between	190	232	8
and	4661	155	7	under	182	243	12
in	4591	116	11	per	176	188	10
a	4302	206	12	because	153	191	11
for	2028	155	6	since	143	206	8
on	1616	197	12	during	124	209	10
at	1424	176	8	through	119	125	8
by	1330	138	16	while	113	147	14
with	1283	246	12	until	106	232	9
from	939	254	8	without	91	188	10
an	918	141	12	among	80	138	11
but	767	238	6	although	70	212	13
or	439	150	9	whether	56	176	7
after	318	133	8	behind	52	175	10
if	277	188	13	about	192	223	8
into	274	209	10				

§ 2. Постоянство вероятности и норма

Тот факт, что основная масса словоформ имеет устойчивые распределения вероятностей $P(F)$, говорит о том, что частоты F имеют свойства, аналогичные свойствам классической случайной величины, описывающей тот или иной статистический объект. Величина F описывает некоторую лингвистическую единицу — в нашем случае словоформу. Отсюда следует, что словоформы английских публицистических текстов можно рассматривать как статистические объекты.

Первым вопросом, возникающим при исследовании статистических объектов, является вопрос о постоянстве (устойчивости) их вероятностей при переходе от одной порции к другой. В нашем случае речь будет идти о том, сохраняет или не сохраняет словоформа свою вероятность употребления при переходе от одного усредненного, по многим авторам, участка текста к другому. От ответа на этот вопрос зависит не только решение проблемы нормированности текста, но и правильный выбор «вероятностной схемы», а значит и правильное применение аппарата теории вероятностей к речи.

Известно, что вероятность p употребления некоторого явления (в нашем случае — появления словоформы) оценивается его относительной частотой $f=F/N$. Известно также, что полученные из опыта относительные частоты употребления данной словоформы в выборках N_1 и N_2 обычно не равны одна другой, т. е. $f_1 \neq f_2$.

Это несовпадение может быть объяснено как следствие или случайных колебаний f при постоянной вероятности p (нулевая гипотеза), или непостоянства значения вероятности употребления словоформы (альтернативная гипотеза). Справедливость нулевой гипотезы будет свидетельствовать об объективности существования нормы. Напротив, правильность альтернативной гипотезы укажет на отсутствие нормированности текста.

Чтобы выяснить, какая из двух гипотез верна, необходимо определить, значима или нет, с точки зрения статистики, разница между f_1 и f_2.

Анализ зависимости расхождений между f_1 и f_2 снова осложняется тем, что неизвестен вид распределения вероятностей $P(F)$. В силу этого необходимо применить такой критерий значимости, который не зависит от вида $P(F)$. Таким критерием является, в частности, критерий знаков.

Для применения критерия знаков воспользуемся проведенной выше разбивкой двух равных выборок английского публицистического текста ($N_1=N_2=100\,000$ словоупотреблений) на n порций. Для каждой выборки $n=20$ (см. выше, стр. 15). Подсчитав относительные частоты исследуемой словоформы в каждой порции, получаем два ряда независимых относительных частот:

$$\begin{aligned} &f_1^{(1)},\ f_2^{(1)},\ \ldots\ f_i^{(1)},\ \ldots\ f_n^{(1)} \\ &f_1^{(2)},\ f_2^{(2)},\ \ldots\ f_i^{(2)},\ \ldots\ f_n^{(2)}, \end{aligned} \tag{10}$$

где верхний индекс указывает на принадлежность частоты к первой или второй выборке, нижний — показывает номер порции.

Составим разности:

$$\begin{aligned} &z_1 = f_1^{(1)} - f_1^{(2)};\quad z_2 = f_2^{(1)} - f_2^{(2)} \ldots \\ &z_i = f_i^{(1)} - f_i^{(2)} \ldots\ z_n = f_n^{(1)} - f_n^{(2)}. \end{aligned} \tag{11}$$

Сосчитав число положительных значений z_i (число плюсов), получаем численное значение критерия знаков (обозначим его буквой m).

Чтобы решить, какую из двух гипотез следует принять, необходимо опытное значение критерия знаков сопоставить с границами доверительного интервала для числа плюсов. Эти границы, определяемые уровнем значимости P_β, находятся исходя из следующих соображений. Если нулевая гипотеза справедлива для каждого i, то вероятности событий $z_i > 0$ и $z_i < 0$ равны одна другой:

$$P(z_i > 0) = P(z_i < 0). \tag{12}$$

Если вероятность события z_i равна нулю, то из (12) вытекает:

$$P(z_i > 0) = P(z_i < 0) = \frac{1}{2} = \text{const}. \tag{13}$$

Из (13) следует: вероятность того, что из n знаков мы будем иметь m плюсов, описывается биноминальным распределением вида

$$P(m, n) = \frac{n!}{m!\,(n-m)!}\left(\frac{1}{2}\right)^m \cdot \left(1-\frac{1}{2}\right)^{n-m} = \frac{n!}{m!\,(n-m)!}\left(\frac{1}{2}\right)^n. \quad (14)$$

Поэтому нижняя граница открытого доверительного интервала m_1 и его верхняя граница m_2 должны определяться из условий

$$\sum_0^{m_1-1} P(m, n) = \sum_0^{m_1-1} \frac{n!}{(m_1-1)!\,[N-(m_1-1)]!}\left(\frac{1}{2}\right)^n = \frac{P_\beta}{2}$$

$$\sum_{m_2+1}^{n} P(m, n) = \sum_{m_2+1}^{n} \frac{n!}{(m_2+1)!\,[N-(m_2+1)]!}\left(\frac{1}{2}\right)^n = \frac{P_\beta}{2}. \quad (15)$$

Возьмем снова $P_\beta = 0.01$. Воспользовавшись специальной таблицей, находим, что при $n=20$ $m_1=4$, а $m_2=16$. Численные значения m для обследованных словоформ приведены в четвертой колонке табл. 1—7. Из 600 словоформ только четыре — busy ($m=39$), can ($m=16$), by ($m=16$), и refused ($m=17$) — дают для критерия знаков такие значения, которые выходят за пределы однопроцентного доверительного интервала. Иными словами, только 0.7% обследованных словоформ обнаруживают неустойчивость своих вероятностей: остальные 99.3% словоформ имеют постоянные вероятности. Это позволяет нам принять нулевую гипотезу и утверждать, что вероятность употребления словоформ в английских публицистических текстах подчиняется норме.[47]

Глава II. ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ АСПЕКТЫ ФУНКЦИОНИРОВАНИЯ НОРМЫ

§ 1. Оценка нормативности словоформ с помощью дисперсионного критерия

Выше было показано, что словоформы являются статистическими объектами, обладающими устойчивыми распределениями $P(F)$ и $P(f)$, имеющими к тому же постоянные вероятности p. Эти статистические характеристики словоформ определяются нормой языка или его разновидности — нормой, заложенной в лингвистическом сознании говорящего. Опираясь на эти результаты, перейдем к дальнейшему исследованию нормированности текста.

[47] Заметим, что нулевая гипотеза оставалась бы справедливой и тогда, если бы только 99.0% словоформ показало постоянство вероятностей, а оставшиеся шесть словоформ (1.0%) дали бы значения m, выходящие за пределы доверительного интервала.

Первый вопрос, на котором следует остановиться, состоит в оценке нормированности связей между отдельными словоупотреблениями текста.

Действительно, наша лингвистическая интуиция говорит о том, что между словоупотреблениями текста существуют лексико-грамматические связи, заданные валентностными нормами языка или его разновидности. Однако остается неясным, распространяются ли эти валентностные нормы на все словоформы, или часть этих последних употребляется независимо. Если существуют словоформы, имеющие независимое, со статистической точки зрения, употребление в тексте, то необходимо выяснить, какова их доля и к каким лексико-грамматическим классам они принадлежат.

Введем численные характеристики употребления зависимых и независимых лингвистических единиц, используя при этом следующие рассуждения.

Если бы словоформы употреблялись не только с постоянной вероятностью, но и независимо одна от другой, то тогда ситуация была бы тождественна схеме Бернулли. При этом колебания абсолютной частоты словоформы описываются биномиальным распределением

$$P(F, k) = \frac{k!}{F!\,(k-F)!}\,P^F(1-p)^{k-F}, \quad (16)$$

где $k = \text{const}$ — число словоупотреблений в одной порции, где $p = \bar{F}/k$ есть вероятность употребления словоформы, а $F = M(F) = \sum_{i=0}^{k} F_i P(F, k) = k \cdot p$ — математическое ожидание абсолютней частоты.

Будем интересоваться случаями, когда $\bar{F} \gg 1$, а $\bar{F}/N \ll 1$. В этих условиях распределение (16), как известно, может быть с достаточно высокой точностью заменено нормальным распределением вида

$$P(F) = \frac{1}{\sqrt{2\pi\bar{F}}}\,e^{-\frac{(F-\bar{F})^2}{2\bar{F}}}, \quad (17)$$

которое характеризуется математическим ожиданием и дисперсией:

$$M(F) = \bar{F} = k \cdot p, \quad \sigma^2(F) = \bar{F} = k \cdot p. \quad (18)$$

Таким образом, если распределение $P(F)$ устойчиво, вероятность употребления в тексте некоторой лингвистической единицы постоянна ($p=\text{const}$) и все единицы употребляются независимо одна от другой, то вид распределения $P(E)$ и его параметры определяются теоретически строго и даются выражениями (16, 17).

Теперь рассмотрим ситуацию, когда $P(F)$ устойчиво, $p=\text{const}$, но употребление словоформы в тексте не является независимым.

Остановимся на простейшем случае зависимого употребления: предположим, что вероятность появления рассматриваемой словоформы зависит только от того, какая лингвистическая единица стоит в тексте перед этой словоформой. Такая ситуация, как известно, описывается простой цепью Маркова. Математическое ожидание и дисперсия абсолютной частоты имеют в этом случае вид

$$M(F) = k \cdot p + (p_1 - p)\frac{1-\delta^k}{1-\delta},$$

$$D(F) = kpq\frac{1+\delta}{1-\delta} + a_n, \qquad (19)$$

где p — так называемая «предельная» вероятность, к которой, с увеличением числа испытаний, очень быстро стремится вероятность употребления слова, p_1 — вероятность употребления интересующего нас слова в первом испытании (p не зависит от p_1); $\delta = \beta - \gamma$, причем β — вероятность употребления интересующего нас слова два раза подряд, γ — вероятность употребления интересующего нас слова с другими словами $|\delta| < 1$; $q = 1 - p$; a_n — некоторая ограниченная величина.

Сопоставляя выражения (18) и (19), нетрудно заметить, что применение к тексту, вместо схемы Бернулли, простой цепи Маркова не приведет к заметному изменению $M(F)$: второе слагаемое будет малым, из-за того что вероятности P в речи малы (обычно $P_{max} < 0.05$). Вместе с тем это заметно приводит к существенному изменению дисперсии $D(F)$. Последняя по своей величине может измениться в несколько раз, причем изменение будет тем сильнее, чем больше $\delta = \alpha - \beta$ отличается от нуля.

Из всего сказанного следует, что различия в значениях дисперсии могут служить критерием степени взаимозависимости употребления словоформ текста.

Расчет обеих дисперсий производится следующим образом. Сначала для каждой словоформы определяется опытное значение математического ожидания (среднее арифметическое частот), которое равно

$$\bar{F}_{ap} = \frac{1}{n}\sum_{i=1}^{n} F_i. \qquad (20)$$

Затем вычисляется опытное значение дисперсии, которое равно

$$S^2(F) = \frac{1}{n-1}\sum_{i=1}^{n}(F_i - \bar{F}_{ap})^2. \qquad (21)$$

Согласно (18), величина $\bar{F}_{ap}$ служит оценкой дисперсии при условии справедливости схемы Бернулли. Поэтому, если $\bar{F}_{ap}$ и $S^2(F)$ близки между собой, справедлива схема Бернулли;

если же разница между ними велика, то действует схема цепей Маркова.

Различия между величинами $\bar{F}_{ap}$ и $S^2(F)$ оцениваются также с помощью метода гипотез. Согласно нулевой гипотезе, верна схема Бернулли, а $\bar{F}_{ap}$ и $S^2(F)$ отличаются одно от другого вследствие естественного разброса случайных величин. Согласно альтернативной гипотезе, различия $\bar{F}_{ap}$ и $S^2(F)$ существенны, и поэтому к данной словоформе применима не схема Бернулли, а схема цепей Маркова. Если верна нулевая гипотеза, вероятность появления словоформы не зависит от предшествующей единицы текста; если же справедлива альтернативная гипотеза, то словоформа имеет условные вероятности.

Чтобы осуществить статистический анализ обеих гипотез, сконструируем из $\bar{F}_{ap}$ и $S^2(F)$ такую случайную величину, распределение которой было бы хорошо известно. В теории вероятностей для анализа значимости отличий экспериментальной величины дисперсии $S^2(x)$ от теоретического значения $\sigma^2(x)$ случайной величины x, которая имеет нормальное распределение, применяется случайная величина

$$\chi^2 = (n-1)\frac{S^2(x)}{\sigma^2(x)}. \qquad (22)$$

При этом доказывается, что указанная величина имеет χ^2-распределение с числом степеней свободы $(n-1)$. Заметим, что n — это число измерений, по которым найдена величина $S^2(x)$; в нашем случае оно соответствует числу порций в выборке.

Согласно нулевой гипотезе, величина F также имеет нормальное распределение, у которого $\sigma^2(F) = \bar{F}$. Мы не знаем значения $\bar{F}$, однако имеем его экспериментальную оценку $\bar{F}_{ap}$, среднеквадратичное уклонение которой

$$S(\bar{F}_{ap}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(F_i - \bar{F}_{ap})^2} \qquad (23)$$

мало́ (оно уменьшается с увеличением числа измерений). В силу этого мы имеем право вместо $\bar{F}$ использовать значение $\bar{F}_{ap}$ и считать последнее квазипостоянным.

В итоге мы приходим к нужной нам случайной величине

$$\chi^2 = (n-1)\frac{S^2(F)}{\bar{F}_{ap}}, \qquad (24)$$

которая имеет χ^2-распределение с $n-1$ степенями свободы. Если задать уровень значимости P_β, то можно получить граничные значения χ^2-критерия, которые можно использовать при

статистическом анализе наших гипотез. Эти граничные значения отвечают условиям

$$\frac{P_\beta}{2} = \int_0^{\chi_1^2} P_{n-1}(\chi^2)\, d\chi^2,$$
$$\frac{P_\beta}{2} = \int_{\chi_2^2}^{\infty} P_{n-1}(\chi^2)\, d\chi^2. \qquad (25)$$

Таким образом, если для исследуемой словоформы найденное из опыта значение критерия χ^2 удовлетворит условию $\chi_1^2 \leqslant \leqslant \chi^2 \leqslant \chi_2^2$, то нет причин отказываться от нулевой гипотезы и считать, что употребление данной словоформы зависит от лингвистической единицы слева. Наоборот, если $\chi^2 > \chi_2^2$ или $\chi^2 < \chi_1^2$, то нулевая гипотеза должна быть отвергнута. А это будет означать, что употребление словоформ в тексте не является независимым.

По критерию χ^2 были обследованы 167 самых частых словоформ, извлеченных из уже упоминавшегося английского публицистического текста длиной в 200 тыс. словоупотреблений. Текст был разбит на 40 порций.

Мы снова приняли $P_\beta = 0.01$, что при $n = 40$ и согласно (25) дает граничные значения

$$\chi_1^2 = 20, \quad \chi_2^2 = 66.^{48}$$

Результаты обследования приведены в табл. 9—15. Слова сгруппированы по частям речи, в пределах каждой части речи они располагаются в порядке возрастания величины критерия χ^2. В левой части каждой таблицы находятся слова, у которых численное значение критерия не выходит за пределы однопроцентного доверительного интервала, в правой части располагаются слова, значения χ^2-критерия которых выходят за указанные выше пределы.

Как показывают табл. 9—15, у 72-х словоформ численные значения критерия χ^2 выходят за граничные пределы χ_1^2 и χ_2^2. Это свидетельствует, что употребление около 45% обследованных единиц не описывается схемой Бернулли — их употребление зависит от появления соседних единиц текста.

Обозначим (в %) долю словоформ, для которых значения критерия χ^2 выходят за пределы доверительного интервала, символом ϕ и будем рассматривать эту величину в качестве численной оценки контекстных связей текста, или, иными словами, меры комбинаторно-статистических ограничений, задаваемых нормой.

[48] См.: В. И. Романовский. Применение математической статистики в опытном деле. М.—Л., 1947.

Как мы только что выяснили для английских публицистических текстов, $\phi = 45\%$. При этом необходимо учитывать, какова величина ϕ в разных лексико-грамматических классах (методы расчета здесь те же, что и относительно текста в целом). Как показывает табл. 8, наибольший процент словоформ, имеющих связанное употребление, дают существительные и глаголы. При этом среди существительных связанное употребление обнаруживают словоформы, относящиеся к политической, экономической и административной тематике (company, council, government, minister, policy, president и т. п. — см. табл. 9—15). Существительные, не имеющие явно выраженного терминологического характера (building, days, people, house, life, home), показывают независимое употребление, описываемое схемой Бернулли. Среди глаголов связанное употребление обнаруживают вспомогательные и полувспомогательные словоформы (табл. 9—15). Напротив, знаменательные глагольные словоформы дают значения χ^2, попадающие в интервал $\chi_2^2 — \chi_1^2$, что говорит об их независимом употреблении. В других частях речи картина менее ясна: связанное употребление показывают такие словоформы, как he, I, much, own, military, already, not, here, her, his, it, me, she (табл. 9—15).

Таблица 8

Величина φ относительно разных грамматических классов слов

№№	Грамматические классы слов	ϕ, %
1	Существительные	50.0
2	Глаголы	58.5
3	Прилагательные	30.0
4	Местоимения	41.6
5	Наречия	36.8
6	Служебные слова	29.6
7	Числительные	33.3

Результаты только что проведенного исследования наводят на мысль, что статистическое нормирование текста может осуществляться по-разному. Одним словоформам норма предписывает независимое с вероятностной точки зрения употребление; вероятность появления других связана с определенными условиями.

Детально проверить это предположение можно с помощью такого статистического приема, который оказался бы, по сравнению с дисперсионным критерием, более чувствительным к особенностям нормирования текста. Одним из таких приемов является исследование колебаний частот по критерию «три сигмы» (3σ).

Таблица 9

Независимое и зависимое употребление именных словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
force	23.3	minister	77
public	33	court	79
man	33.5	government	84.2
part	39	Mr.	86
board	39.4	company	87
Dr.	47	policy	93
building	48.5	President	95
days	49.2	committee	95
people	50	meeting	96
time	51	council	97
house	54.3	London	100
office	54.3	hospital	105
day	56.5	city	109
life	58	Britain	111
home	62.5	party	119
state	64.5	Sir	147
months	67	labour	195
men	70.5	Mrs.	202
race	70.8	cent	247
secretary	73.7	police	342
night	75.6	industry	417

Таблица 10

Независимое и зависимое употребление глагольных словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
put	32	may	68.6
might	34.2	said	69.5
take	40.6	been	70
took	45.2	can	70
given	45.6	must	70
called	46	being	71.2
show	48.6	held	71.5
make	50	has	73
had	50.6	made	75.2
go	54	live	80
could	54.5	do	86
added	62.5	should	88.5
		have	113
		are	117
		is	127
		last	159
		be	206

Таблица 11

Независимое и зависимое употребление адъективных словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
few	25.3		
many	28.5	own	68
long	33.4	other	71.2
most	36	military	93
general	44	more	95
next	51.4	new	99
American	53.7	British	184
all	55		
high	55		
great	55.3		
national	58.5		
good	64.8		
some	65		

Таблица 12

Независимое и зависимое употребление адвербиальных словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
off	26.4	much	16.8
also	35	already	66.5
even	36.7	here	68.2
however	40.8	not	109
down	46.5	back	133
then	46.5	there	136
now	47.2	as	152
how	51		
still	51.5		
out	54.3		
so	62		
only	63.3		

Таблица 13

Независимое и зависимое употребление словоформ числительных

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
three	31	four	69
one	34	five	76.5
first	43		
second	51.5		

Таблица 14

Независимое и зависимое употребление местоименных словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
its	30	these	75.8
no	42.8	her	82
those	44	that	83
them	44.3	his	84
same	51	it	97.2
another	52.5	our	113
my	55.6	me	116
him	56.2	she	132
they	59.2	he	134
both	60	I	265
each	61		
any	64		
their	64.6		
this	66		
such	41		

Таблица 15

Независимое и зависимое употребление словоформ

Словоформы, показывающие независимое употребление	$20 \leqslant \chi^2 \leqslant 66$	Словоформы, показывающие зависимое употребление	$\chi^2 < 20$ или $\chi^2 > 66$
before	27.3	an	69.5
from	27.3	and	70
on	30.3	at	74
because	35	than	75.7
during	35	in	77
or	36.5	if	78
about	39	the	94
over	41	per	156
for	43		
through	43.7		
against	46.6		
since	47		
but	50.7		
by	51		
into	51		
a	54.6		
after	55		
between	55		
of	63		

§ 2. Критерий «три сигмы» и характер нормированности употребления словоформ

Если сравнивать частоты употребления словоформ в двух выборках длиной в N_1 и N_2 словоупотреблений, то колебания частот для тех словоформ, употребление которых подчиняется схеме Бернулли, будет описываться биномиальным распределением

$$P(F_1, N_1) = \frac{N_1!}{F_1!\,(N_1 - F_1)!}\, p^{F_1}(1-p)^{N_1 - F_1},$$
$$P(F_2, N_2) = \frac{N_2!}{F_2!\,(N_2 - F_2)!}\, p^{F_2}(1-p)^{N_2 - F_2}. \tag{26}$$

Если $N_1 \to \infty$ и $N_2 \to \infty$, а $F_1 \gg 1$ и $F_2 \gg 1$, то равенства (26) имеют вид нормального распределения:

$$P(F_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-\frac{(F_1 - \bar{F}_1)^2}{2\sigma_1^2}},$$
$$P(F_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\, e^{-\frac{(F_2 - \bar{F}_2)^2}{2\sigma_2^2}}, \tag{27}$$

где $\bar{F}_1 = p \cdot N_1 = \sigma_1^2$ и $\bar{F}_2 = p \cdot N_2 = \sigma_2^2$.

Для нашей задачи удобнее пользоваться не абсолютными, а относительными частотами

$$f_1 = \frac{F_1}{N_1} \text{ и } f_2 = \frac{F_2}{N_2}. \tag{28}$$

Если употребления словоформ будут совпадать со схемой Бернулли, то величины f_1 и f_2 будут описываться распределением Гаусса с параметрами: математическое ожидание

$$\bar{f} = \frac{\bar{F}}{N} = p$$

и дисперсией

$$\sigma^2(f) = \sigma^2\left(\frac{F}{N}\right) = \frac{1}{N^2} \cdot \sigma^2(F) = \frac{\bar{F}}{N^2}. \tag{29}$$

Таким образом, некоторая словоформа, употребление которой совпадает со схемой Бернулли, будет иметь следующие распределения вероятностей ее относительных частот:

$$P(f_1) = \frac{1}{\sqrt{2\pi\sigma^2(f_1)}}\, e^{-\frac{(f_1 - p)^2}{2\sigma^2(f_1)}} \tag{30}$$

Таблица 16

Применение критерия «три сигмы» относительно трех английских прилагательных

№№	Словоформы	f_1	f_2	$\sigma(f_1)$	$\sigma(f_2)$	$\lvert\alpha\rvert=\lvert f_1-f_2\rvert$	$\sigma(\alpha)=\sqrt{\sigma^2(f_1)+\sigma^2(f_2)}$	$\frac{\lvert\alpha\rvert}{\sigma(\alpha)}$
1	best	0.36 10^{-3}	0.30 10^{-3}	0.60 10^{-4}	0.55 10^{-4}	0.06 10^{-3}	0.81 10^{-4}	0.7
2	official	0.46 10^{-3}	0.18 10^{-3}	0.59 10^{-4}	0.62 10^{-4}	0.28 10^{-3}	0.80 10^{-4}	3.5
3	small	0.35 10^{-3}	0.31 10^{-3}	0.59 10^{-4}	0.56 10^{-4}	0.04 10^{-3}	0.81 10^{-4}	0.5

для выборки N_1 и

$$P(f_2)=\frac{1}{\sqrt{2\pi\sigma^2(f_2)}}e^{-\frac{(f_2-p)^2}{2\sigma^2(f_2)}}$$

для выборки N_2.

Введем величину $f_1-f_2=\alpha$, которая также будет случайной величиной, описывающей колебания разности речевых частот употребления одной и той же словоформы в двух разных выборках. В соответствии с законом композиции эта величина будет описываться распределением Гаусса

$$p(\alpha)=\frac{1}{\sqrt{2\pi\sigma^2(\alpha)}}e^{-\frac{\alpha^2}{2\sigma^2(\alpha)}} \qquad (31)$$

с математическим ожиданием

$$\bar{\alpha}=f_1-f_2=p-p=0,$$

дисперсией

$$\sigma^2(\alpha)=\sigma^2(f_1)+\sigma^2(f_2) \qquad (32)$$

и среднеквадратическим отклонением

$$\sigma(\alpha)=\sqrt{\sigma^2(f_1)+\sigma^2(f_2)}.$$

Опираясь на распределение $p(\alpha)$, мы можем ввести критерий «три сигмы», который с достаточно большой вероятностью показывал бы, является ли употребление интересующей нас словоформы независимым или связанным. Сущность этого критерия состоит в следующем. Разности $f_1-f_2=\alpha$ могут отличаться от нуля либо в силу случайных статистических колебаний, либо вследствие каких-то существенных (не статистических, но лингвистических) причин. В первом случае колебания α не превосходят $3\sigma(\alpha)$ (вероятность того, что из-за случайных колебаний величина α отклонится от нуля на 3σ или более чем на 3σ, равна здесь 0.003). Во втором случае имеем $|\alpha|\geqslant 3\sigma(\alpha)$.[49] Исходя из этих условий,

[49] Н. В. Смирнов, И. В. Дунин-Барковский. Курс теории вероятностей и математической статистики для технических приложений. М., 1965, стр. 142—145.

можно считать, что словоформы, которые показывают $|\alpha|<3\sigma(\alpha)$, употребляются в согласии со схемой Бернулли, т. е. независимо от появления слева от них в тексте других словоформ; напротив, словоформы, дающие $|\alpha|\geqslant 3\sigma(\alpha)$, выпадают из схемы Бернулли и имеют связанное употребление.

Расчеты, связанные с применением критерия трех сигм, показаны в табл. 16. Из трех обследованных прилагательных два — best и small — дают $|\alpha|<3\sigma$, что свидетельствует об их независимом употреблении; словоформа же official, у которой $|\alpha|>3\sigma$, подчиняется, очевидно, нормам связанного употребления.

Аналогичные расчеты были проведены относительно словоформ, принадлежащих к различным грамматическим классам. Результаты этих расчетов приведены в табл. 17.

Таблица 17

Применение критерия «три сигмы» к словоформам разных грамматических классов

№№	Части речи	Общее число обследованных словоформ	Число словоформ, для которых $\lvert\alpha\rvert<3\sigma$	% словоформ, имеющих независимое употребление
1	Существительные	90	45	50.0
2	Глаголы	87	69	79.3
3	Прилагательные	71	52	73.3
4	Местоимения	29	19	65.5
5	Наречия	21	17	80.9
6	Класс служебных слов	27	22	81.4

Как показывает сводная таблица, наименьший процент независимого употребления показывают именные словоформы: только половина существительных идет по схеме Бернулли, в то время как вторая половина подчиняется, очевидно, нормам связанного употребления. Каковы же лингвистические приметы именных словоформ второй группы? На этот вопрос дает ответ табл. 18, в которой показаны существительные, имеющие $|\alpha|<3\sigma$ (левая колонка), и имена, дающие $|\alpha|\geqslant 3\sigma$ (правая колонка).

Нетрудно заметить, что большинство словоформ, входящих в правую колонку, представляет собой имена существительные терминологического (общественно-политического, экономического, административного) характера.[50] Это наблюдение согласуется

[50] Терминологичность словоформы определялась с помощью лингвопсихологического теста, состоявшего в опросе группы преподавателей английского языка. Словоформы, которые были квалифицированы большинством опрошенных как термины, помечены в табл. 18—23 звездочкой.

Таблица 18

Применение критерия «три сигмы» к именам существительным

Словоформы, дающие $\lvert \alpha \rvert < 3\sigma$	Словоформы, дающие $\lvert \alpha \rvert \geqslant 3\sigma$
year, years, people, night, week, day, public,* home, man, country, state,* part, way, life, action,* report,* school, cause, months, conference,* street, members,* office,* board, days, miles, half, statement,* numbers, times, end, service, war,* top, morning, final, wife, chief, member,* children, Dr., building, West, South, countries	president,* states,* union,* world,* minister,* meeting,* market,* cent,* chairman,* council,* negro, set,* city, leader,* men, secretary,* business,* group,* leaders,* workers, area,* congress,* police,* officials,* race,* issue,* order,* minutes, policy,* power,* defence,* League,* court,* government,* company,* committee,* London, Mrs., Kennedy,* match,* house,* time, association,* cup *

с результатами применения дисперсионного критерия в § 1, который также сигнализировал о связанном употреблении в английских публицистических текстах существительных терминологического значения. Аналогичную картину показывают и прилагательные, у которых словоформы специально-терминологического значения имеют обычно $\lvert \alpha \rvert \geqslant 3\sigma$ (см. табл. 19). Эти данные полностью согласуются с результатами капитального исследования К. Ф. Лукьяненкова, изучавшего распределение существительных и прилагательных в английских научно-технических текстах. Как обнаружил К. Ф. Лукьяненков, при определенных нормировках внутрисерийных выборок существительные «нейтрального» значения дают нормальное распределение; напротив, «все без исключения именные словоформы, не подчиняющиеся нормальному закону, в научно-техническом тексте имеют терминологическое значение».[51] Аналогичная картина «наблюдается как при заданных, так и в измененных условиях исследования» и у прилагательных.[52] Статистическое своеобразие терминологических словоформ обнаруживается и в ходе применения непараметрических критериев, в частности метода ранговой корреляции.[53]

Различия в статистическом поведении именных и адъективных словоформ в публицистических и научно-технических текстах, очевидно, не случайны. Можно предполагать, что для словоформ

* Существительные специального (общественно-политического) значения.

[51] К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического текста. . ., АКД. стр. 24.

[52] Там же, стр. 25.

[53] См.: М. Г. Зореф, Р. Г. Пиотровский. Можно ли статистически определить термин? В сб.: Тезисы докладов XXIV областной научно-технической конференции, посвященной Дню радио и Дню связиста (14—17 апреля 1969 г.), Л., 1969, стр. 250.

Таблица 19

Применение критерия «три сигмы» к именам прилагательным

Словоформы, дающие $\lvert \alpha \rvert < 3\sigma$	Словоформы, дающие $\lvert \alpha \rvert \geqslant 3\sigma$
all, more, other, many, political,* good, long, few, great, own, next, little, far, better, common,* young, each, later, high, local,* right, old, best, able, early, small, federal,* big, large, possible, free, former, full, main, past, round, several, less, total, important, short, whole, hard, recent, annual,* certain, higher, latest,* royal,* strong, general, some	new, united,* national,* foreig,* white, military,* official,* prime,* central,* western,* special,* civil,* British, international,* economic,* industrial,* nuclear,* democratic,* earlier

общего (нетематического) значения норма языка задает биномиальное (а при $N \to \infty$ нормальное) *распределение*. В *тех же* случаях, когда именная или адъективная словоформа обозначает предмет, понятие или качество, непосредственно относящееся к конкретному содержанию или тематике текста, общеязыковая норма употребления этой словоформы нарушается: ее статистическое поведение определяется уже содержательной стороной данного текста или совокупности аналогичных текстов.

Отличную от именных форм картину показывают остальные части речи (см. табл. 20—23). Особый интерес представляет соотношение независимо ($\lvert \alpha \rvert < 3\sigma$) и зависимо ($\lvert \alpha \rvert \geqslant 3\sigma$) употребляемых глагольных словоформ. В первую группу попало большинство знаменательных глаголов, во вторую вошли вспомогательные, полувспомогательные формы, а также знаменательные словоформы said, told, held, said, announced.

Таблица 20

Применение критерия «три сигмы» к глагольным формам

Словоформы, дающие $\lvert \alpha \rvert < 3\sigma$	Словоформы, дающие $\lvert \alpha \rvert \geqslant 3\sigma$
was, had, were, would, made, could, make, like, did, get, put, take, called, go, went, left, added, stopped, leave, evoid, keep, came, asked, found, took, show, see, lost, taken, might, given, going, come, won, give, making, become, plan, say, agreed, got, does, brought, help, used, reported, began, gave, think, wont, *visit, seen, sent, following,* look, including, met, meet, move, leading, appeared, finish, expected, know, heard, died, cost, test	is, be, will, said, been, told, held, announced, says, do, should, have, are, has, being, must, can, may

Таблица 21

Применение критерия «три сигмы» к местоимениям

Словоформы, дающие $\lvert x\rvert < 3\sigma$	Словоформы, дающие $\lvert x\rvert \geqslant 3\sigma$
they, him, my, which, who, what, its, we, them, those, my, this, you, their, me, her, him, our, each	these, that, I, his, it, me, she, he, any, that

Таблица 22

Применение критерия «три сигмы» к наречиям

Словоформы, дающие $\lvert x\rvert < 3\sigma$	Словоформы, дающие $\lvert x\rvert \geqslant 3\sigma$
when, only, now, today, down, even, just, very, then, yesterday, so, also, here, still, out, off	already, not, back, there

Таблица 23

Применение критерия «три сигмы» к служебным словам

Словоформы, дающие $\lvert x\rvert < 3\sigma$	Словоформы, дающие $\lvert x\rvert \geqslant 3\sigma$
in, to, at, with, for, on, by, but, or, about, after, over, into, if, before, under, during, until, while, of, from, than	the, and, a, an, but

Аналогичные результаты мы получили при использовании дисперсионного критерия (см. § 1), сходную картину показало исследование К. Б. Бектаева и К. Ф. Лукьяненкова (см. ниже).

Очевидно, и здесь имеют место различия в функционировании нормы языка. Однако определить, в чем состоят эти различия, опираясь только на семантическую группировку глагольных словоформ, пока не удается. Для этого нужны новые гипотезы и разыскания.

§ 3. Нормирование грамматических единиц

Проведенные исследования обнаружили действие нормы языка на употребление лексики; однако остается неясным, распространяется ли вероятностно-статистическое воздействие нормы на выбор грамматических форм.

Чтобы ответить на этот вопрос, мы исследовали на том же материале поведение залогово-временных и именных глагольных форм, используя при этом метод «определения статистической ошибки частот».[54] Эта ошибка определялась из функции распределения вероятностей для частоты употребления *того или иного* времени.

Методика эксперимента состоит в следующем. Весь текст длиной в 200 тыс. словоупотреблений разбивается на 40 порций по 5000 словоупотреблений в каждой. Определяется частота встречаемости F_i данного времени в каждой порции.

Наилучшей оценкой частоты употребления временной формы глагола в этом случае является среднее арифметическое *из всех* F_i:

$$\bar{F} = \frac{1}{n}\sum_{i=1}^{n} F_i, \tag{33}$$

статистическая ошибка которого определяется выражением

$$\sigma(\bar{F}) = \frac{1}{\sqrt{n\,(n-1)}}\sqrt{\sum_{i=1}^{n}(F_i - \bar{F})^2}. \tag{34}$$

Обе эти величины для временных форм действительного и страдательного залогов показаны в табл. 24 и 25.

Таблица 24

Средние арифметические значения частот употребления временных форм глагола (действительный залог)

	Indefinite	Continuous	Perfect	Perfect Continuous
Present	115.4 ±4.38	8 ±1.105	20 ±1.265	1.1 ±0.2
Past	156.4 ±6.76	5.7 ±0.51	9.7 ±0.8	0.45 ±0.1
Future	15.85 ±1.55	0.8 ±0.25	0.15 ±0.08	
Future in the Past	8.95 ±1.08	0.15 ±0.1	0.45 ±0.1	

Теперь проверим, применимо ли распределение Гаусса (частный вид нормального распределения) для описания частоты употребления какого-либо выбранного времени. С этой целью сопоставим величины экспериментальной ошибки, найденной по формуле (34), с предсказываемым (гауссовским) распределением величин ошибки:

$$\sigma(\bar{F})_{\text{Гаусс}} = \sqrt{\frac{\bar{F}}{n}} \tag{35}$$

[54] Ср.: В. И. Романовский. Элементарный курс математической статистики. М.—Л., 1939, стр. 118—170.

Таблица 25

Средние арифметические значения частот употребления временных форм глагола (страдательный залог)

	Indefinite	Continuous	Perfect
Present	17.7 ±0.308	1.45 ±0.33	5.5 ±0.63
Past	25.6 ±1.47	0.3 ±0.4	2.75 ±0.5
Future	4 ±0.52	—	0.1 ±0.063
Future in the Past	2.5 ±0.24	—	0.05 ±0.044

Результаты сравнения в виде отношения $\sigma(\bar{F})/\sigma(\bar{F})_{\text{Гаусс}}$ приводятся в табл. 26 и 27.

Таблица 26

Отношение величины экспериментальной ошибки к величине ошибки Гаусса (действительный залог)

	Indefinite	Continuous	Perfect	Perfect Continuous
Present	4.4/2.4	1.1/0.6	1.3/1.0	0.2/0.2
Past	6.8/2.7	0.5/0.5	0.8/0.7	0.1/0.1
Future	1.5/0.9	0.2/0.1	0.1/0.6	—
Future in the Past	1.1/0.6	0.1/0.1	0.1/0.1	—

Таблица 27

Отношение величины экспериментальной ошибки к величине ошибки Гаусса (страдательный залог)

	Indefinite	Continuous	Perfect
Present	0.3/0.9	0.3/0.2	0.6/0.5
Past	1.5/1.1	0.4/0.1	0.5/0.2
Future	0.5/0.4	—	0.1/0.02
Future in the Past	0.2/0.3	—	0.04/0.05

Наблюдающиеся в табл. 26 и 27 расхождения между величинами экспериментальной ошибки $\sigma(\bar{F})$ и ошибкой, предсказываемой распределением Гаусса $\sigma(\bar{F})_{\text{Гаусс}}$, можно объяснить двояким образом. Расхождение можно отнести за счет случайных колебаний величины $\sigma(\bar{F})$, т. е. распределение Гаусса применимо для описания частот употребления временных форм (нулевая гипотеза). Расхождение между $\sigma(\bar{F})$ и $\sigma(\bar{F})_{\text{Гаусс}}$ не

случайно, тогда распределение Гаусса неприменимо (альтернативная гипотеза). Проверить эти гипотезы можно с помощью критерия χ^2. Известно, что случайная величина $(n-1)\times \times \frac{\sigma(\bar{F})}{\sigma(\bar{F})_{\text{Гаусс}}} = \chi^2$ подчиняется χ^2-распределению. Это позволяет подсчитать численное значение вероятности

$$P(\chi^2 \geqslant \chi_0^2) : P(\chi^2 \geqslant \chi_0^2) = \int_{\chi_0^2}^{\infty} P(\chi^2)\, d\chi^2. \qquad (36)$$

Если величина критерия $P(\chi^2 \geqslant \chi_0^2) < 0.05$, то расхождение между теоретической и экспериментальной ошибками не случайно и, следовательно, частная формула распределения Гаусса неприменима.

Таблица 28

Определение величины критерия $P(\chi^2)$ (действительный залог)

	Indefinite		Continuous		Perfect		Perfect Continuous	
	χ^2	$P(\chi^2)$	χ^2	$P(\chi^2)$	χ^2	$P(\chi^2)$	χ^2	$P(\chi^2)$
Present	62.7	$\leqslant$0.001	62.7	$\leqslant$0.001	28.6	~0.1—0.5	19	~0.50
Past	117.8	$\leqslant$0.001	19.0	~0.50	14.7	0.8—0.7	19	~0.50
Future	57.0	$\leqslant$0.001	3268	$\leqslant$0.001	24.7	~0.20	—	—
Future in the Past	60.8	$\leqslant$0.001	18.5	~0.50	19	~0.50	—	—

Таблица 29

Определение величины критерия $P(\chi^2)$ (страдательный залог)

	Indefinite		Continuous		Perfect	
	χ^2	$P(\chi^2)$	χ^2	$P(\chi^2)$	χ^2	$P(\chi^2)$
Present	1.9	>0.90	51.3	$\leqslant$0.001	32.3	0.05—0.02
Past	32.3	0.05—0.02	304	$\leqslant$0.001	117.8	$\leqslant$0.001
Future	30.4	0.05	—	—	178.6	$\leqslant$0.001
Future in the Past	11.9	0.90	—	—	13.3	~0.80

Из табл. 28 и 29 видно, что в большинстве случаев для временных форм глагола $P(\chi^2)$ значительно меньше 0.05. Эти формы глагола не подчиняются нормальному распределению (частный вид распределения Гаусса).

Аналогичный анализ был проведен и для неличных форм глагола (инфинитив, герундий, причастие). Его результаты показаны в табл. 30 и 31.

Таблица 30

Средние арифметические значения

	F	$\sigma(F)$
Инфинитив	103.05	3.82
Причастие	149.75	3.82
Герундий	17.45	1.14

Таблица 31

Определение величины экспериментальной ошибки к величине ошибки Гаусса

	$\sigma(F)/\sigma(F)_{\text{Гаусс}}$
Инфинитив	3.82/2.27
Причастие	3.82/2.73
Герундий	1.14/0.93

Результаты расчетов, приведенные в табл. 28 и 29, показывают, что употребление таких форм, как времена группы Indefinite, Present и Future Continuous, не подчиняется нормальному распределению, или, иными словами, не является статистически независимым. Другая глагольная форма — Past Continuous и группа времен Perfect — показывает независимое употребление.

Таким образом, и на грамматическом материале обнаруживается вероятностно-статистический характер функционирования нормы языка. Однако здесь так же, как и в лексике, норма функционирует неодинаково: для одних единиц она задает нормальное распределение и независимое употребление, для других — диктует иные статистические схемы, например цепи Маркова. Весьма вероятно, что между этими типами функционирования нормы в лексике и грамматике существует определенная связь, — например, зависимое употребление английских вспомогательных и полувспомогательных глаголов связано, по всей вероятности, с зависимым употреблением некоторых залоговых и временных форм. Однако, чтобы ответить на все эти вопросы, необходимы дополнительные исследования.

Итак, все примененные нами критерии настойчиво говорят в пользу существования некоторого «эталона» статистического построения и упорядочения текста. Этот эталон, охватывающий вероятности употребления отдельных лингвистических единиц, распределения этих вероятностей, а также включающий вероятностные схемы речи, является тем, что Э. Косериу интуитивно определил как «норму», находящуюся между нестатистической «системой» и порождаемым ею текстом («речью»).

Норма, которую следует рассматривать как сумму контекстных ограничений, накладывающихся на бесконтекстную «грамматику», задана в лингвистическом сознании членов коллектива, пользующихся данным языком (соответственно его разновидностью, функциональным стилем, подъязыком). Однако общие свойства этой нормы, в том числе ее статистические признаки, можно обнаружить только путем исследования текста, — в частности, его статистической структуры. Этой задаче и посвящено большинство статей первой части настоящего сборника.

Е. Б. Бектаев и Е. Ф. Лукьяненков

О ЗАКОНАХ РАСПРЕДЕЛЕНИЯ ЕДИНИЦ ПИСЬМЕННОЙ РЕЧИ

§ 1. Задачи исследования

1.1. В настоящей статье решаются две задачи: лингвостатистическая и лексикологическая.

Лингвостатистическая задача включает следующие вопросы:

1) создание универсальных схем автоматического построения вероятностно-статистических моделей распределения лингвистических единиц в тексте;

2) определение характера распределения словоформ всех грамматических классов слов (кроме междометий) и трехсловных сочетаний (триад) разной грамматической природы в английском научно-техническом тексте по судовым механизмам;

3) выяснение оптимальных условий исследования распределений лингвистических единиц в тексте.

Поставленная лингвостатистическая задача согласуется с принципом отбора единиц естественного языка в базовый язык и единиц естественного подъязыка в базовый подъязык; [1] она исходит из далеко не полной изученности характера распределения лингвистических единиц в тексте, а также из гипотезы о распределении лингвистических единиц, выдвинутой В. М. Калининым,[2] Г. Херданом,[3] Р. М. Фрумкиной.[4] Эта гипотеза не проверялась на боль-

[1] См.: Р. Г. Пиотровский. Экстралингвистические и внутриязыковые вопросы при переработке текста в системе «человек — электронно-вычислительная машина». В сб.: Вопросы социальной лингвистики. Л., 1969.

[2] См.: В. М. Калинин. О статистике литературного текста. ВЯ, 1964, № 1, стр. 123—127.

[3] См.: G. Herdan. Quantitative Linguistics. London, Butterworths, 1964, p. 61 etc.; The Advanced Theory of Language as Choice and Chance. Berlin—Heidelberg—New York, 1966, p. 201 etc.

[4] Р. М. Фрумкина. О законах распределения слов и классов слов. В сб.: Структурно-типологические исследования, М., 1962, стр. 124—133.

ших массивах текста; выводы, к которым пришли Б. Н. Головин,[5] Л. Е. Машкина,[6] Т. А. Микерина,[7] О. А. Нехай [8] в своих исследованиях по данному вопросу, не могли быть достаточно полными. Причина этого кроется в том, что при ручном способе работы и даже при использовании ЭВМ, не имеющей вывода данных на широкую печать, сам метод проб и ошибок, на который здесь опираются исследователи, и связанный с ними поиск оптимальной разбивки и нормировки исследуемого материала накладывают значительные ограничения как на выбор вариантов задаваемых условий исследования, так и на объем обследуемого материала, а это, в свою очередь, сужает возможности для выводов.

И наконец, лингвостатистическая задача отвечает также и нуждам методики преподавания иностранных языков, в частности целям оптимизации учебного процесса; существует мнение, согласно которому объективными критериями отбора активной части лингвистических единиц (как лексических, так и грамматических) для активного или пассивного усвоения учащимися или для составления учебников и учебных пособий должны служить частота, стабильная частота и наличность лингвистических единиц *в сознании носителя языка*.[9]

Лексикологическая задача состоит в попытке определить лингвостатистическим путем некоторые объективные признаки терминологичности/нетерминологичности лингвистических единиц в тексте.

1.2. Гипотезы

При постановке задачи мы исходили из допущения, что законы распределения лингвистических единиц в тексте нам неизвестны. На этом основании задаются следующие гипотезы:

1) распределение словоформ разных статистических групп частотного списка в текстах подчиняется разным законам;

[5] Б. Н. Головин. Опыт вероятностно-статистического изучения некоторых явлений истории русского литературного языка XIX—XX вв. ВЯ, 1965, № 3, стр. 135—136.

[6] Л. Е. Машкина. О статистических методах исследования лексико-грамматической дистрибуции (на материале публицистических текстов политической тематики современного немецкого языка). КД (рукоп.), Белорусский гос. унив. им. В. И. Ленина, Минск, 1968.

[7] Т. А. Микерина. Некоторые статистические приемы лексико-морфологического описания функционального стиля. КД (рукоп.), ЛГУ, 1967.

[8] О. А. Нехай. Статистика и автоматический анализ текста (на материале английских текстов по электронике). КД (рукоп.), Белорусский гос. унив. им. В. И. Ленина, Минск, 1968.

[9] См., например, работы Г. Гугенейма, Р. Мишеа, Г. Мюллера, К. Хойпеля и других в сб.: *Методика преподавания иностранных языков за рубежом*, М., 1967.

2) трехсловные сочетания разных статистических групп частотного списка подчиняются разным законам;

3) словоформы, принадлежащие к разным грамматическим классам слов, имеют разные законы распределения в тексте;

4) трехсловные сочетания разной грамматической природы подчиняются разным законам распределения.

Исследованию были подвергнуты эмпирические распределения 300 словоформ различных грамматических классов слов и 300 трехсловных сочетаний (триад), извлеченных из английского подъязыка судовых механизмов. Эти распределения проверялись относительно следующих теоретических распределений: нормального закона (распределения Гаусса), логарифмически-нормального и распределения Пуассона.

§ 2. Принципы отбора лингвистических единиц для анализа

2.1. Словоформы

Поисковые работы предшественников показали, что характер распределения словоформ претерпевает значительные изменения при переходе от наиболее частых словоформ, занимающих начальный участок списка, к словоформам, находящимся в средней зоне частотного словаря. В связи с этим исследованию были подвергнуты все словоформы начального участка, а затем использовался выборочный анализ, причем интервал между исследуемыми словоформами постепенно увеличивался (см. табл. 1). Словоформы отбирались в диапазоне достоверных единиц частотного списка, полученного от текста длиной в 400 тыс. словоупотреблений.[10]

Таблица 1

i	Порядок отбора словоформ из частотного списка	Всего словоформ из зоны	f_i^*
1—150	сплошь	150	0.5814
151—595	через 5 номеров	100	0.7501
600—950	» 10 »	35	0.8099
950—1200	» 25 »	10	0.8380
1200—1497	номера 1200, 1248, 1301, 1350, 1497	5	0.8631

В таблице: i — порядковый номер словоформы в частотном списке; f_i^* — относительная накопленная частота последней словоформы зоны.

[10] См.: К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического текста с помощью электронно-вычислительной машины (подъязык судовых механизмов). КД (рукоп.), Белорусский гос. унив. им. В. И. Ленина, Минск, 1969, стр. 117—122.

Таблица 2

Соотношение анализируемых словоформ разных грамматических классов слов и единиц искусственных языков

Словоформы знаменательных слов				Словоформы служебных слов				Словоформы искусственных языков	
1	грамматические связи класс известен	грамматические связи омонимия	всего	2	грамматические связи класс известен	грамматические связи омонимия	всего	3	всего
Существительное	81	37	117	Артикль	3	9	3	Цифры, формулы, индексы	1
Прилагательное	25	27	52	Предлог/частица	7		16		
Местоимение	4	5	9	Союз	4		4	Буквенные сокращения	1
Глагол	46	24	71						
Наречие	10	16	26						
Итого	166	109	275	Итого	14	9	23	Итого	2

При составлении частотного списка за единицу обследования была принята словоформа — цепочка букв, заключенная между двумя соседними пробелами. Лексико-грамматическая омонимия не учитывалась. Цифры, математические и другие формулы, индексы и т. п. считались одной словоформой и обозначались индексом *x*. Буквенные сокращения представлены символом *z*.

У 182 словоформ принадлежность к грамматическому классу можно было легко и однозначно определить по морфологическому составу слова. Для оставшихся 118 словоформ конверсионная омонимия была устранена путем соотнесения каждой из них с тем грамматическим классом, который указан в словаре первым.[11] Список словоформ *см. ниже (стр. 82)*. Соотношение словоформ различных грамматических классов, а также единиц искусственных языков показано в табл. 2.

2.2. Трехсловные сочетания

Отбору триад для анализа по законам распределений предшество-

[11] См.: Webster's New Colligiate Dictionary. London, G. Bell & Sons, Lmd., Springfield (Mass.), G. & C. Merriam Co, 1960.

вал этап построения с помощью ЭВМ по специальным процедурам четырех частотных списков трехсловных сочетаний.[12] Каждый список был получен от обследования текста объемом 400 тыс. словоупотреблений. Для этой цели в качестве опорных элементов (ядер) триад в каждом случае брались 130 наиболее частых единиц из частотного списка словоформ, полученного также от текста длиной в 400 тыс. словоупотреблений.

Словарь трехсловных сочетаний первого типа представляет собой частотный список триад, полученный путем сплошного анализа текста. Здесь в качестве левых ядерных элементов задаются все словоупотребления обследуемого текста. У триад второго типа центральными ядерными элементами выступают 130 наиболее частых единиц частотного списка *словоформ*. Триады третьего типа образованы от наиболее частых словоформ существительного (ядром служит правый элемент). Четвертый тип триад строится аналогично второму типу и отличается от него тем, что в качестве ядер здесь выступают 130—150 наиболее частых словоформ, выполняющих атрибутивную функцию: имена существительные в единственном числе в функции препозитивного определения имени существительного, прилагательные, причастия настоящего и прошедшего времени в препозиции и постпозиции к имени существительному.[13]

Из частотного списка триад 2-го типа взято 275 первых единиц в зоне абсолютных частот $F_i \geqslant 30$; из списка трехсловных сочетаний 4-го типа было отобрано девять единиц, не обнаруженных в предыдущем списке в зоне частот $F_i \geqslant 30$; три триады взяты из частотного списка сочетаний 3-го типа, не обнаруженных в двух предыдущих списках в пределах $F_i \geqslant 30$, и 13 триад отобрано из списка словосочетаний 1-го типа по тому же принципу. В списке анализируемых сочетаний, который мы приводим ниже (см. стр. 92 и сл.), триады расположены по типам в последовательности 2, 4, 3, 1. Соотнесенность триад с разными группами по признаку принадлежности центральной ядерной словоформы

[12] См.: А. В. Зубов, К. Ф. Лукьяненков, Р. Г. Пиотровский. Статистическое описание текста с помощью ЭВМ. В сб.: Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ. Тезисы докладов и сообщений, Изд. ЛГУ, 1966, стр. 49—50; А. В. Зубов, К. Ф. Лукьяненков, Р. Г. Пиотровский, Э. Н. Хотяшов. Лексико-статистическое описание текста на электронно-вычислительных машинах. СР, стр. 108—119; К. Ф. Лукьяненков. Некоторые схемы реализации комплекса лингвостатистических задач в системе «человек — электронно-вычислительная машина». В сб.: Частотные словари и автоматическая переработка лингвистических текстов. Тезисы докладов 2-й межвузовской конференции 4—6 апреля 1968, Минск, 1968.

[13] Подробно о типах триад см.: К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического текста . . .

Таблица 3

Соотношение анализируемых словосочетаний по грамматическим классам центральных элементов

Триады с ядерными словоформами					
знаменательных слов		служебных форм		единиц искусственных языков	
1	всего	2	всего	3	всего
Существительное	45	Артикль	52	x	25
Прилагательное	9	Предлог/частица	78		
Местоимение	15	Частица	2	z	17
Глагол	32	Союз	21		
Наречие	4				
Итого:	105		153		42

к различным грамматическим классам слов иллюстрируется в табл. 3.

§ 3. Общий алгоритм исследования законов распределений

3.1. Принципы формирования совокупностей текста и предварительные замечания

Выборочная совокупность строится из отрезков текста (серий или внутрисерийных выборок). Количество серий некоторой длины (K), которым представлена совокупность текста некоторого объема (α), обозначим через N_K^α.

Рассматриваемые схемы исследования характера распределения лингвистических единиц в тексте рассчитаны на сравнение соответствия эмпирических распределений теоретическим законам по критерию χ^2. Использование этого критерия связано с выполнением следующего условия: количество серий, которыми представлена анализируемая совокупность элементов, должно быть не менее 50, т. е. $N_K^\alpha \geqslant 50$.[14] Для выполнения этого условия при длине общей выборочной совокупности текста в 400 тыс. словоупотреблений задаем следующие объемы внутрисерийных выборок: $K=1000$, 2000, 4000, 8000 и 16 000 словоупотреблений. Тогда весь анализируемый текст может быть представлен в виде схемы формирования совокупностей текста, которую мы приводим ниже (см. рис. 1).

[14] См.: Дж. Э д н и Юл, М. Дж. К е н д э л. Теория статистики. Русск. пер. М., 1960, стр. 526—527; В. Е. Г м у р м а н. Введение в теорию вероятностей и математическую статистику. М., 1966, стр. 319.

На рис. 1 латинскими буквами обозначены конкретные массивы анализируемого текста. Каждая буква соответствует длине текста в 50 тыс. словоупотреблений.

Примеры чтения схемы: $N_{1000}^{A}=50$ — обследуется массив текста A объемом (α) 50 000 словоупотреблений, представленный 50-ью внутрисерийными выборками длиной (K) в 1000 словоупотреблений каждая; $N_{2000}^{ABCD}=100$ — обследуется массив текста ABCD объемом (α) 200 000 словоупотреблений. Массив представлен 100 внутрисерийными выборками при $K=2000$ словоупотреблений.

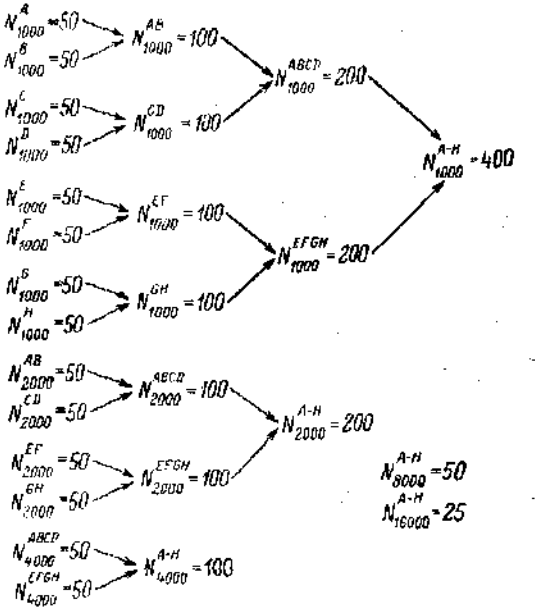


Рис. 1. Схема формирования совокупностей текста.

Из схемы видно, что при заданных значениях $K=1000$, 2000, 4000 и 8000 словоупотреблений условие $N_K^\alpha \geqslant 50$ может быть выполнено для 26 вариантов совокупностей текста разной длины. $N_K^\alpha=25$ при $K=16\,000$ словоупотреблений принято как исключение из условия. Таким образом, распределение каждой обследуемой единицы проверяется по 27 вариантам объемов текста при пяти разных объемах внутрисерийных выборок.

Для удобства дальнейшего описания схем исследования распределений лингвистических единиц в тексте даем определения некоторых понятий.

1) Любой из заданных объемов текста, состоящий из внутрисерийных выборок одной и той же длины, будем называть п р о-

межуточной выборочной (рабочей) совокупностью.

2) Независимыми выборочными (рабочими) совокупностями будем называть одинаковые объемы текстов, составленные из внутрисерийных выборок одной и той же длины.

3) Общей выборочной совокупностью будем именовать текст, представленный одной или несколькими независимыми совокупностями, сумма объемов которых составляет 400 тыс. словоупотреблений.

4) F_i — абсолютная частота, с которой обследуемая единица (словоформа или словосочетание) встретилась в серии текста некоторой заданной длины K в словоупотреблениях.

5) F'_i — абсолютная частота, с которой обследуемая единица встретилась в некотором количестве серий (внутрисерийных выборок) $m_{F'_i}$. Значения, принимаемые F'_i, будем называть классами частот $m_{F'_i}$.

6) $m_{F'_i}$ (абсолютная частота) — количество внутрисерийных выборок, в которых обследуемая единица встретилась с частотой F'_i.

Интересующим нас признаком исследования являются численные значения, принимаемые переменной величиной $m_{F'_i}$ по классам F'_i. Другими словами, нас интересуют частоты $m_{F'_i}$ — количество внутрисерийных выборок, в которых анализируемая лингвистическая единица встретилась с частотой F'_i в обследуемом объеме выборочной совокупности, составленной из N^{α}_{K} внутрисерийных выборок.

3.2. Основные этапы построения машиной схемы исследуемых распределений

Алгоритм исследования строился таким образом, чтобы решение поставленной задачи осуществлялось как единый непрерывный процесс от ввода текста и анализируемых единиц в оперативную память машины до вывода данных, требующих минимальной ручной обработки. За основными этапами процесса обработки информации можно проследить по принципиальной блок-схеме (см. рис. 2).

1. В Б2 анализируемые единицы в коде М-2 вводятся с перфоленты (ПЛ) по специальной программе в оперативную память машины. В Б3 производится перекодировка их из кода М-2 в новый код.

Буквенная информация при представлении ее в машине изображается в виде цепочек цифр. Например, по коду М-2 букве А в машине соответствует цифровой код 111 000 в двоичной системе счисления. Код М-2 не дает возможности упорядочения буквенной информации по алфавиту, поэтому в процессе работы производится перекодирование каждого символа в «новый» код. По новому коду, например, букве А соответствует цепочка цифр 100 000, В — 100 001, С — 100 010 и т. д. в двоичной системе счисления. Другими словами, алфавит по новому коду представляется по убыванию значений цепочек цифр, что обеспечивает возможность сортировки буквенной информации по алфавиту.

В Б3 на каждую единицу заводится машинная карточка. С целью ускорения поиска анализируемых единиц в тексте их машинные карточки сортируются в Б4 по алфавиту и запоминаются в МОЗУ.[15] Чтобы убедиться в надежности ввода и записи их в оперативной памяти, анализируемые единицы в алфавитном порядке выводятся на устройство печати (УПч), а быстродействующий печатающий механизм (БПМ) выдает количество запомненных единиц. В МОЗУ анализируемым единицам придаются номера от 0 до max.

В качестве единиц обследования одновременно могут выступать и словоформы, и словосочетания.

2. В Б5 осуществляется ввод текста сериями по 1000 словоупотреблений. В Б6 из каждой серии отбираются абсолютные частоты (F_i) анализируемых единиц. Частоты запоминаются (накапливаются) в оперативной памяти машины в виде неранжированных статистических рядов. При этом БПМ в целях контроля выдает номер серии, количество разных единиц, отобранных из серии, их накопленную частоту и контрольную сумму (КС) введенного текста. Эти процедуры выполняются последовательной работой Б5, Б6 и Б7. Цикл переработки одной порции текста повторяется до просмотра 25 серий.

Информация, полученная по каждой обследуемой единице из массива длиной в 25 серий, запоминается в девяти ячейках памяти машины. В первой ячейке регистрируются номер массива, номер единицы и ее частота в первой серии. Остальные ячейки занимают частоты по сериям от № 2 до № 25 по горизонталям слева направо (см. табл. 4). В таблице 4 представлен неранжированный ряд частот предлога with серий от № 276 до № 300.

Данные, полученные от текста длиной в 25 серий, по всей группе обследуемых единиц в МОЗУ записываются в виде табл. 5. Пустые клетки заполняются значениями F_i по каждой обследуемой единице для каждой серии.

3. После обработки 25 серий информация в виде, накопленном в МОЗУ, записывается в Б8 на МЛ и сохраняется для последующей сортировки в Б11.

4. Для полной обработки всей выборочной совокупности процедуры, описанные выше в пунктах 2 и 3, повторяются. В круго-

[15] МОЗУ — магнитное оперативное запоминающее устройство (оперативная память).

Таблица 4

№ ячейки	№ массива	№ единицы	Частота
372	12	145	11
№№ ячеек	Частоты		
373	3	6	5
374	3	7	11
375	9	10	8
376	11	5	11
377	12	9	8
400	7	13	3
401	9	13	4
402	7	4	11

Таблица 5

№№ единиц	№№ серий												
	1	2	3	4	5	6		20	21	22	23	24	25
1													
2													
...													
300													

вом цикле работают блоки 5, 6, 7, 8. После обработки 25 серий Б8 передает работу программы на Б9, и далее с Б10 осуществляется переход на Б5 для переработки очередного массива текста длиной в 25 серий.

В конце этого этапа информация по всему объему текста 400 тыс. словоупотреблений представлена на МЛ в виде табл. 5 шестнадцатью последовательно расположенными массивами.

5. В Б11 осуществляется сортировка информации на МЛ по номеру обследуемой единицы.

На этом этапе часто́ты по каждому номеру обследуемой единицы располагаются последовательно по номерам от 1 до 400. Информация по всей обследуемой совокупности текста принимает на МЛ вид табл. 6. Массивам по 25 серий придаются номера от 1 до 16.

Таким образом, для каждой обследуемой единицы мы получаем на МЛ неранжированный ряд частот по сериям 1000 словоупотреблений для всей выборочной совокупности объемом 400 серий.

6. В Б12 формируются таблицы типа 1 (см. стр. 64—66). Таблицы этого типа представляют собой неранжированные статисти-

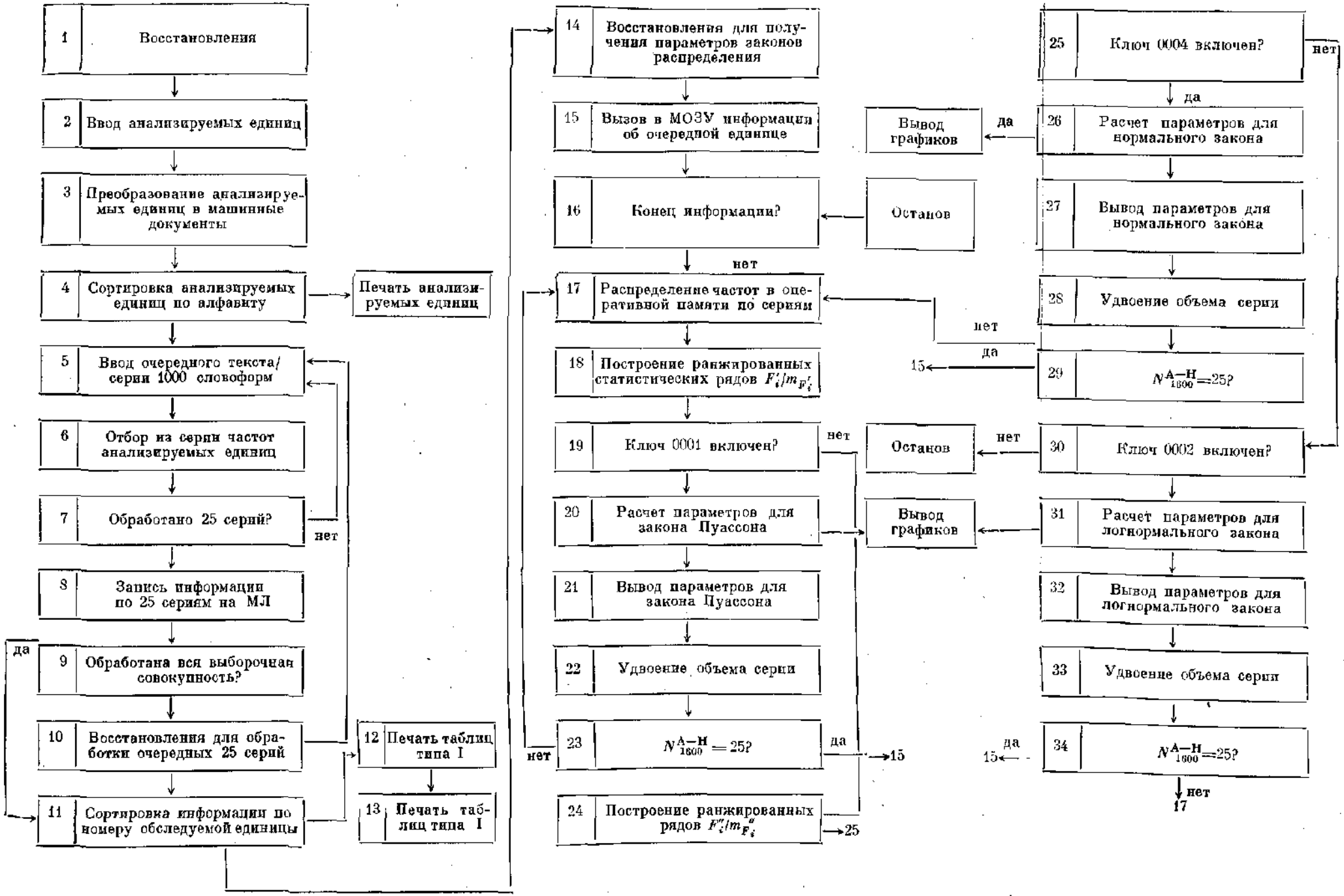


Рис. 2. Принципиальная блок-схема 1.

Таблица 6

№№ единиц	№№ серий												
	1	2	3	4	5	6	7	8	...	397	398	399	400
1													
2													
⋮													
300													

ческие ряды частот для различных объемов выборочных совокупностей при разной длине внутрисерийных выборок, заданных схемой, представленной на рис. 1.

В МОЗУ вызывается вся информация об очередной единице. В ячейках памяти машины частоты по сериям длиной в 1000 словоупотреблений располагаются, как показано в табл. 6, от первой серии до 400-й. *Последовательным объединением каждых двух* соседних массивов по 25 серий получаем восемь выборочных совокупностей по 50 серий ($N^{A}_{1000}=50$, $N^{B}_{1000}=50$ и т. д.).[16] Объединением каждых четырех массивов по 25 серий формируем четыре совокупности по 100 серий ($N^{AB}_{1000}=100$, $N^{GD}_{1000}=100$ и т. д.). Аналогичным путем формируются все варианты (N^{a}_{1000}) *объемов выборочных совокупностей*.

Чтобы построить статистические неранжированные ряды для выборочных совокупностей при других объемах внутрисерийных выборок, предварительно суммируются частоты F_i каждых двух, четырех, восьми или 16 соседних серий. Например, при N^{a}_{2000} всю обследуемую совокупность текста представляют уже не 16, а 8 исходных массивов по 25 серий. Другими словами, весь обследуемый текст представлен 200 сериями по 2 тыс. словоупотреблений и т. д.

Вывод таблиц на устройство печати осуществляется подключением в общий процесс стандартной программы построения таблиц. Блок 12 блок-схемы 1 представлен в развернутом виде на принципиальной блок-схеме 2 (рис. 3). Переход от вывода одного варианта N^{a}_{x} к другому осуществляется с помощью ключей и внесением в программу незначительных изменений с ПУ ЭВМ.

7. В Б13 формируются таблицы типа II.[17] Для вывода таблиц этого типа используется та же стандартная программа, что и для вывода таблиц типа I.

---

[16] Подробное описание таблиц этого типа см. ниже, § 4, стр. 63. N^{A}_{1000} — количество внутрисерийных выборок по 1000 словоупотреблений в массиве текста А объемом 50 серий (см. рис. 1).

[17] Подробное описание таблиц этого типа см. ниже, стр. 67—71.

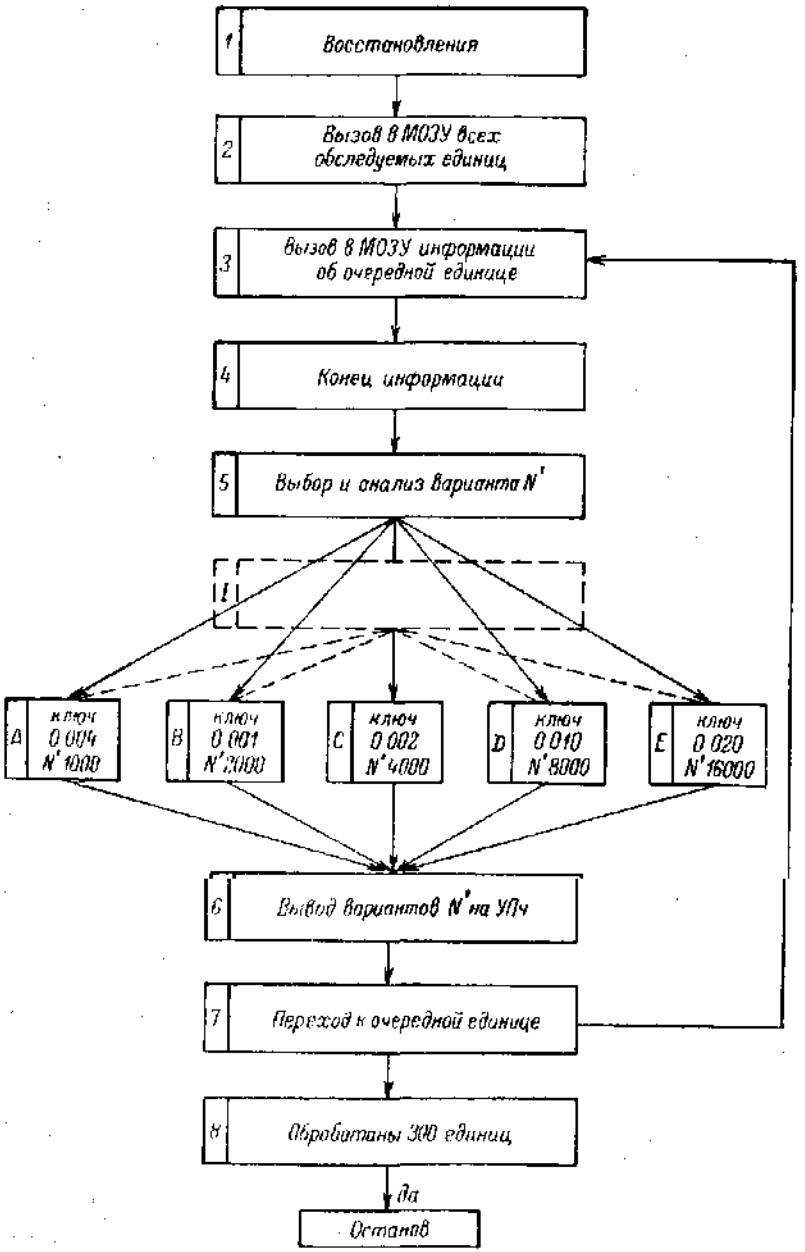


Рис. 3. Принципиальная блок-схема 2.

После Б5 блок-схемы 2 в работу одним из ключей вводится дополнительная подпрограмма (см. блок *I*). Таким образом, Б13 представляет собой несколько усложненный вариант Б12 (см. блок-схему 1, рис. 2).

Согласно схеме на рис. 1 для каждой обследуемой единицы по всем вариантам N'_K подсчитывается количество серий, в которых анализируемая единица встретилась с одной и той же частотой F_i. Сумма таких серий дает численное значение величины $m_{F'_i}$ (ср. выше, стр. 54), а разные значения, принимаемые частотой F_i, представляют собой неранжированный ряд частот F'_i. Здесь же определяется размах вариации ($F'_{\min}$ и $F'_{\max}$). В пределах размаха вариации часто́ты F'_i сортируются в оперативной памяти по возрастанию от F'_0 до $F'_{\max}$. Путем «привязки» частот $m_{F'_i}$ к своему классу F'_i обеспечивается построение ранжированных (эмпирических) вариационных рядов частот $F'_i/m_{F'_i}$ для текста длиной в N'_K внутрисерийных выборок. Такой ряд имеет следующий вид:

$$\begin{array}{cccccc} F'_0 & F'_1 & F'_2 & F'_3 & \ldots & F'_{\max} \\ m_{F'_0} & m_{F'_1} & m_{F'_2} & m_{F'_3} & \ldots & m_{F'_{\max}} \end{array}$$

(ср. ниже с табл. типа II, стр. 67—71).

Вывод таблиц по вариантам N'_K осуществляется введением в работу соответствующих ключей (см. блоки A, B, C, D, E блок-схемы 2, рис. 3).

Следует указать, что таблицы типа I и типа II необходимы лишь для контроля за ходом выполнения задачи до Б11 включительно, а также для визуального наблюдения за построенными эмпирическими вариационными рядами, с тем чтобы принимать решения для корректировки методики исследования. Если в этом нет необходимости, то блоки 12, 13 исключаются из общей программы, не нарушая процесса. В этом случае работает Б11 с переходом непосредственно на Б14.

8. Блоки 14—18 блок-схемы работают аналогично блокам 12, 13. На Б18 заканчивается та цепочка блоков, которые являются общими для всех трех схем исследования распределений. Далее для каждой схемы строится особая программа.

В Б20 определяются статистические параметры для закона Пуассона, строятся теоретические вариационные ряды и сравниваются с эмпирическими (подробно см. ниже, стр. 71 и сл.). Б20 упрощенно приводится ниже (см. рис. 4).

Определение параметров для закона Пуассона по одной обследуемой единице осуществляется последовательным вводом в работу блоков 14, 15, 16, 17, 18, 19, 20, 21 — для вариантов N'_{1000}. Для других вариантов N'_K работают последовательно в круговом цикле блоки 22, 23, 17, 18, 19, 20, 21.

Для обработки очередной обследуемой единицы процесс повторяется, начиная с Б15.

9. Б24 является общим для нормального и логарифмически-нормального законов. Здесь построенные в Б18 дискретные статистические ряды $F'_i/m_{F'_i}$ приводятся к интервалу, формируются непрерывные статистические ряды F''_i и $m_{F''_i}$,[18] строятся теоретические ряды и сравниваются с эмпирическими (см. ниже, стр. 74 и сл.).

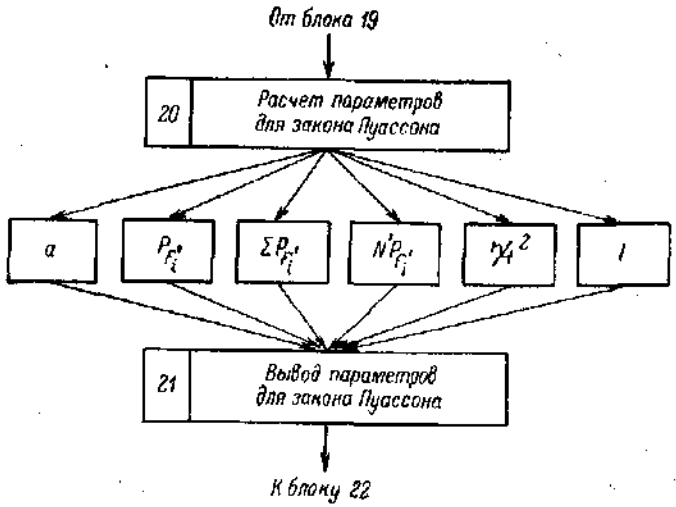


Рис. 4. Блок 20 блок-схемы 1.

При определении статистических параметров для нормального закона по одной обследуемой единице сначала работают последовательно блоки 14, 15, 16, 17, 18, 24, 25, 26, 27 (для N^{α}_{1000}), а затем блоки 28, 29, 17, 18, 24, 25, 26, 27 (для последующих вариантов N^{α}_{K}).

Блок 26 блок-схемы 1 в развернутом виде приводится ниже (см. рис. 5).

10. При определении статистических параметров для логарифмически-нормального распределения в цикле работает последовательность блоков 14, 15, 16, 17, 18, 24, 30, 31, 32, а затем — 33, 34, 17, 18, 24, 30, 31, 32. Блок 31 приводится ниже (см. рис. 6).

11. Вывод графиков кривых распределений производится подключением к блокам 20, 26, 31 стандартной программы построения графиков (подробно о графиках см. ниже, стр. 80 и сл.).

[18] F''_i и $m_{F''_i}$ — то же, что и F'_i, $m_{F'_i}$ (см. выше, стр. 54 и 59), но после приведения ряда к интервалу. Подробно о приведении ряда к интервалу см. ниже, стр. 70 и сл.

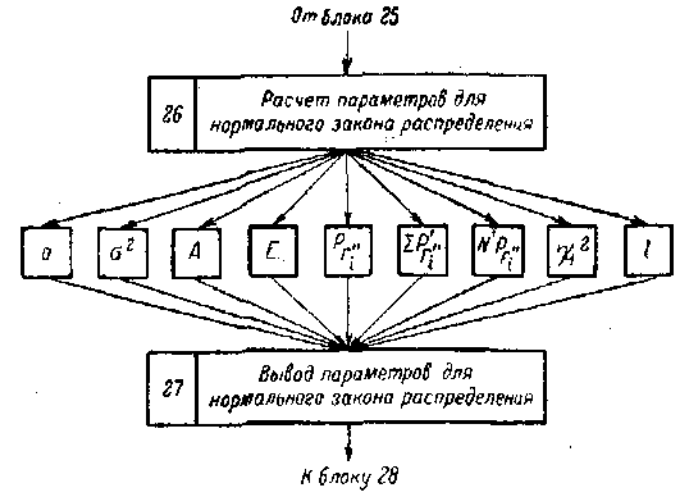


Рис. 5. Блок 26 блок-схемы 1.

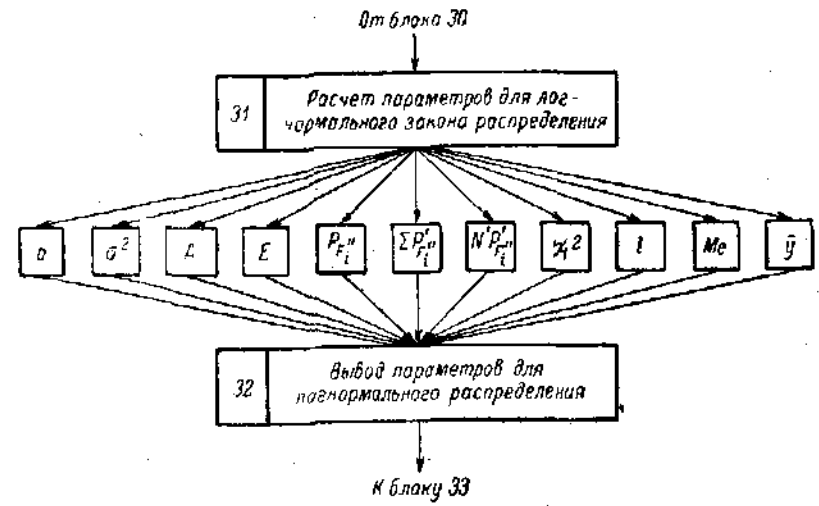


Рис. 6. Блок 31 блок-схемы 1.

§ 4. Статистические вариационные ряды

Поскольку текст, написанный на естественном языке, представляет собой некоторый марковский процесс, появления словоформ и словосочетаний в тексте, строго говоря, не являются независимыми событиями. Другими словами, наша переменная величина, определяющая закон распределения единицы текста, является зависимой случайной величиной. Теоретически же распределение Гаусса (нормальный закон) и распределение Пуассона распространяются лишь на независимые случайные величины.

Ввиду сложности использования в лингвостатистике математического аппарата для изучения зависимых случайных событий, например с применением марковских цепей, исследователи обычно делают допущение о независимости появления лингвистических единиц в тексте, например словоформ, словосочетаний и т. п. Это допущение по Мандельброту предполагает, что, во-первых, «моделирующий текст является стационарным случайным процессом. . .», во-вторых, не учитываются вероятностные связи между элементами, т. е. считается, что вероятностная модель полностью задана.[19] Такое допущение делаем и мы. Правомерность этого допущения связана с результатами работ А. А. Маркова, С. Н. Бернштейна и А. Я. Хинчина о полном выяснении вопроса применимости закона больших чисел к зависимым случайным величинам, «если только сильная зависимость существует между случайными величинами с близкими номерами, а зависимость между слагаемыми с далекими номерами достаточно слаба».[20]

Хотя, принципиально, появление слов в тексте в какой-то степени диктуется стилем изложения, содержанием описываемых объектов (явлений) (экстралингвистическими факторами) и т. п., вероятностные связи между отдельными словами через определенное количество шагов значительно слабеют.[21] Еще меньшую взаимозависимость имеют словоформы и словосочетания, принадлежащие к различным внутрисерийным выборкам,[22] извлеченным из текста с интервалами в 10—20 страниц или взятым из разных статей различных журналов или из разных книг. Это дает нам право утверждать, что применяемая методика исследования в достаточной степени строга и математически оправдана.

[19] Подробно см.: Е. А. Калинина. Изучение лексико-статистических закономерностей на основе вероятностной модели. СР, стр. 65—66.

[20] См.: А. И. Карасев. Основы математической статистики. М., 1962, стр. 92.

[21] Некоторые экспериментальные данные, полученные методами теории информации применительно к письменной речи, показывают, что далеко действующие статистические связи между словами очень слабо проявляются и практически затухают через четыре-пять слов. См., например: Р. Г. Пиотровский. Информационные измерения языка. Л., 1968, стр. 57—58; N. G. Burton and J. C. K. Licklider. Long-Range Constraint in the Statistical Structure of Printed English. The American Journal of Psychology, vol. LXVIII, № 4, December 1955, pp. 650—653. Ср. также некоторые другие данные: Ю. И. Левин. Знаки, язык и математика (о некоторых вопросах современной математики и кибернетики). М., 1965, стр. 449; Murray Aborn, Herbert Rubinstein and Theodor D. Sterling. Sources of Contextual Constraint upon Words in Sentences. Journal of Experimental Psychology, vol. 57, № 3, 1957, pp. 178—179. Попытку лингвистического обоснования независимости появления в тексте трехсловных сочетаний, в которых в качестве центральных ядерных элементов выступают служебные слова, именные и глагольные словоформы, см. в работе: О. А. Нехай, ук. соч., стр. 24—25.

4.1. Неранжированные статистические ряды частот

На прилагаемых ниже таблицах типа I (стр. 64) представлены неранжированные эмпирические ряды частот указательного местоимения this для различных объемов внутрисерийных выборок (ср. с табл. 6, также см. п. 6, стр. 54—57). Рис. 7 представляет собой фрагмент машинного результата тех же выборок.

В верхнем горизонтальном ряду таблицы печатаются анализируемая единица и номер, приданный ей в оперативной памяти ЭВМ; в крайнем ряду слева по вертикали латинскими буквами обозначены обследуемые массивы текстов (промежуточные выборочные совокупности, ср. со схемой 1). В каждых двух горизонтальных рядах слева направо даются частоты (F_i) обследуемых единиц в каждой внутрисерийной выборке от номера первого до номера, равного N^a_K. Крайний вертикальный ряд справа (SF_i) занимает сумма абсолютных частот единицы в N^a_K внутрисерийных выборках.[23] Внизу таблицы справа печатается абсолютная частота единицы в общей выборочной совокупности.

Лист 1 представляет собой ряды для N^a_{1000}, лист 2 — N^a_{2000} и т. д. Ряды листа 1 являются исходными для получения рядов листа 2, в свою очередь лист 2 служит исходным для листа 3 и т. д. Увеличение объема внутрисерийной выборки производится путем суммирования частот соседних серий. Объемы промежуточных выборочных совокупностей формируются слева направо при выполнении условия $N^a_k \geqslant 50$ согласно схеме 1. Эти ряды являются общими в построении неранжированных эмпирических рядов для трех исследуемых законов распределений. Этапы построения их машиной описаны выше в § 3, стр. 52 и сл.

[22] Исходная внутрисерийная выборка (микровыборка или малая выборка) длиной в 1000 словоупотреблений в нашем случае отбиралась через 10—20 страниц монографии. Подробно о формировании выборочной совокупности см.: К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического текста. . .

[23] Точность работы программы и машины проверяется сравнением суммы частот N^a_K серий с абсолютной частотой единицы в частотном списке, полученном от объема текста, равного N^a_K внутрисерийным выборкам.

Лист 1

Таблица типа I (N^{a}_{1000})

THIS 119 118

A	10	5	5	4	10	5	7	3	4		12	2	4	3		3	2	4	9	17	12	7	10	11	2	
A	8	11	12	5		13	6	3	3	8	4	6	3	2	6	1	3	12	9	8	6	7	5	7	1	SF_i=300
B	3	6	3	2	6	1	7	5	5	6	1	6	6	5	4	5	10		11	15	8	6	3	6	6	
B	4	10	9	7	8	19	17	7	12	8	8	7	7	9	4	5	5	2	4	4	15	6	12	5	1	SF_i=331
C	4	2	7	4	5	6	10	10	16	14	19	6	8	6	7	9	5	3	4	5	3	6	6	9	9	
C	4	5	5	7	3	7	14	8	3	3	4	4	4	4	5	3	3	9	1	9	9	3	6	3	1	SF_i=310
D	7	7	3		6	5	4	5	11	7	5	4	8	5	8	2	8	7	5	5	4	7	7	11	2	
D	5	3	7	5	5	1	4	8	2	9	4	6	11	6	5	5	9	3	7	6	11	6	3	7	6	SF_i=287
E	9	9	6	2	4	3	4	1		3	7	4	1	6	12	3	5	2	4	9	8	10	7	8	3	
E	10	6	12	5	7	5	10	4	2	6	7	10	12	11	5	12	7	6	11	3	1	2	12	4	8	SF_i=308
F	7	9	3	6	6	9	9	5	11	5	9	8	4	10	11	8	10	8	3	4	7	10	7	9	9	
F	7	9	12	4	10	3	5	3	3	6	11	9	5	3	4	9	4	11	7	7	8	3	8	5	7	SF_i=350
G	6	6	5	8	7	9	11	5	4	13	7	6	6	2	2	2	4	7	4	8	19	8	12	10	13	
G	4	9	4	3	2	7	3	5	6	5	9	2	3	13	7	6	6	6	2	5	3	1	8	6	7	SF_i=316
H	7	15	3	4	7	10	8	9	12	4	5	8	9	6	8	8	7	5	6	8	11	9	7	7	3	
H	19	6	6	3	5	2	10	6	5	10	8	3	7	3	4	1	3	8	7	3	10	3	4	8	10	SF_i=339

S=2541

Лист 2

Таблица типа I (N^{a}_{2000})

THIS 119 118

A	15	9	15	10	4	14	7	3	6	26	19	21	10	23	5	19	6	12	9	8	4	21	14	12	8	SF_i=300
B	9	5	7	12	11	7	11	9	10	26	14	9	10	19	15	36	19	16	14	13	10	6	19	18	6	SF_i=331
C	6	11	11	20	30	25	14	16	8	9	9	15	13	10	10	21	11	7	8	9	6	10	18	9	4	SF_i=310
D	14	3	11	9	18	9	13	10	15	10	11	18	7	10	10	5	10	13	17	11	14	10	17	9	13	SF_i=287
E	18	8	7	5	3	11	7	15	7	13	18	15	13	18	12	15	6	13	22	16	19	17	14	14	12	SF_i=308
F	16	9	15	14	16	17	14	19	18	7	17	16	16	21	14	8	6	17	14	7	13	18	15	11	12	SF_i=350
G	12	13	16	16	17	13	8	4	11	12	27	22	17	13	5	10	11	14	5	20	12	8	8	9	13	SF_i=316
H	22	7	17	17	16	13	15	16	12	14	20	14	22	12	8	12	11	18	9	7	4	15	13	7	18	SF_i=339

S=2541

Лист 3

Таблица типа I (N^{A-H}_{8000})

THIS 119 118

ABCD	49	26	72	57	35	51	20	41	59	53	85	48	41	85	47	47	49	33	45	41	47	46	35	55	49	SF_i=1228
EFGH	38	36	53	58	57	54	52	61	61	67	45	53	48	62	35	79	40	45	52	57	56	70	43	38	53	SF_i=1313

S=2541

Лист 4

Таблица типа I (N^{A-H}_{16000})

THIS 119 118

A—H	77	129	86	70	112	133	127	94	82	86	93	90	87	89	115	106	122	112	101	97	119	97	113	113	91	SF_i=2541

S=2541

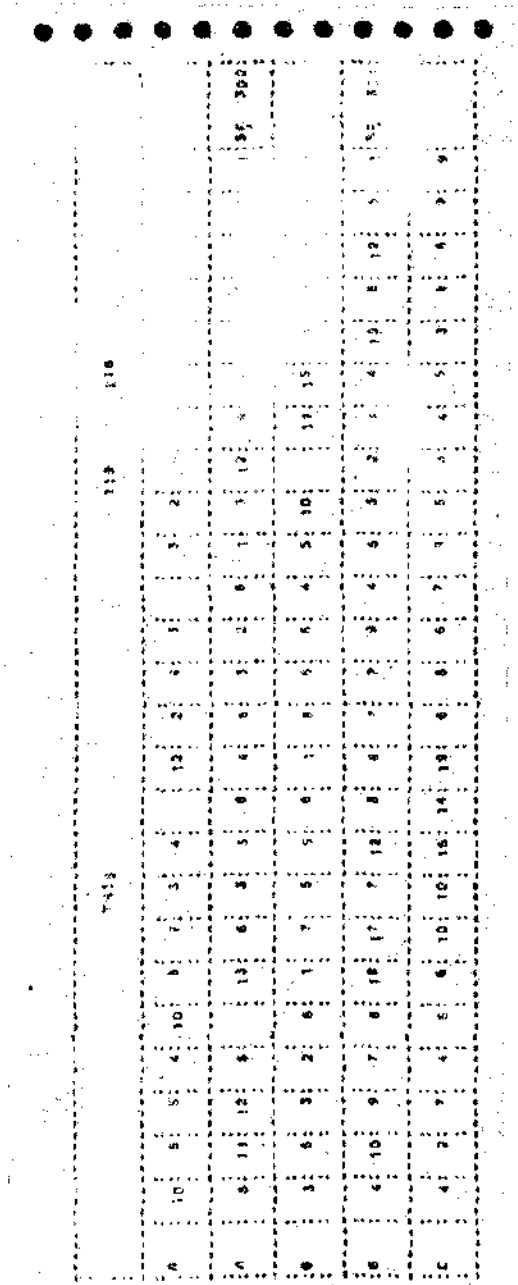


Рис. 7. Таблица типа 1 (N^{*}_{1000}). Лист 1. Неранжированные эмпирические ряды частот указательного местоимения this для различных объемов внутрисерийных выборок (фрагмент машинного результата).

4.2. Построение ранжированных статистических рядов частот

4.2.1. Закон Пуассона

Для проверки соответствия распределения словоформ и триад в текстах закону Пуассона строятся дискретные вариационные ряды.

Путем внутренней сортировки частот (F'_i) классов исследуемого признака по возрастанию от 0 до максимума и «привязки» к ним соответствующих значений $m_{F'_i}$ описанные выше неранжированные ряды преобразуются в ранжированные (ср. выше, стр. 59). Такие ряды представлены ниже таблицами типа II: лист 1 для N^{*}_{1000}, лист 2 для N^{*}_{2000}, лист 3 для N^{*}_{4000}.

Лист 1

Таблица типа II (N^{*}_{1000})

THIS 119

F'_i	A	B	C	D	E	F	G	H	AB	CD	EF	GH	ABCD	EFGH	A—H
0	3	1		1	1				4	1	1		5	1	5
1	2	3	2	1	3		1	1	5	3	3	2	8	5	13
2	4	2	1	3	4		5	2	6	4	4	8	10	12	22
3	7	3	9	4	5	7	4	7	10	13	12	11	23	23	46
4	5	5	8	5	6	5	5	4	10	13	11	9	23	20	43
5	5	7	6	11	4	5	5	4	12	17	9	9	29	18	47
6	4	9	6	6	5	3	9	5	13	12	8	14	25	22	47
7	4	5	4	9	5	7	6	7	9	13	12	13	22	25	47
8	3	4	2	4	3	5	4	8	7	6	8	12	13	20	33
9	2	2	6	2	3	9	3	3	4	8	12	6	12	18	30
10	3	2	2		4	4	1	5	5	2	8	6	7	14	21
11	2	1		4	2	4	1	1	3	4	6	2	7	8	15
12	4	2			5	1	1	1	6		6	2	6	8	14
13	1						3		1			3	1	3	4
14			2							2			2		2
15		2						1	2			1	2	1	3
16			1							1			1		1
17	1	1							2				2		2
19		1	1				1	1	1	1		2	2	2	4
Итог	50	50	50	50	50	50	50	50	100	100	100	100	200	200	400

В таблицах в верхнем горизонтальном ряду указываются обследуемая единица и номер, приданный ей в оперативной памяти машины; латинские буквы — индексы промежуточных выборочных совокупностей; вертикальный ряд (F'_i) слева —

Таблица типа II (N^{a}_{5000})

F'_i	A	B	C	D	E	F	G	H	AB	CD	EF	GH	ABCD	EFGH	A–H
0															
3	1			1	1				1	1	1		2	1	3
4	2		1		1		1	1	2	1	1	2	3	3	6
5	1	1		1	1		2		2	1	1	2	3	3	6
6	2	2	2		1	1			4	2	2		6	2	8
7	1	2	1	1	3	2		3	3	2	5	3	5	8	13
8	2		2		1	1	3	1	2	2	2	4	4	6	10
9	2	3	4	3		1	1	1	5	7	1	2	12	3	15
10	2	3	3	6			1		5	9		1	14	1	15
11		2	3	3	1	1	2	1	2	6	2	3	6	5	13
12	2	1			2	1	3	3	3		3	6	3	9	12
13		1	1	3	3	1	4	2	1	4	4	6	5	10	15
14	2	2	1	2	1	4	1	2	4	3	5	3	7	8	15
15	2	1	1	1	3	2		2	3	2	5	2	5	7	12
16		1	1		1	4	2	2	1	1	5	4	2	9	11
17				2	1	3	2	2		2	4	4	2	8	10
18		1	1	2	3	2		2	1	3	5	2	4	7	11
19	2	3			1	1			5		2		5	2	7
20			1				1	1		1		2	1	2	3
21	2		1			1			2	1	1		3	1	4
22					1		1	2			1	3		4	4
23	1								1				1		1
25			1							1			1		1
26	1	1							2				2		2
27							1					1		1	1
30			1							1			1		1
36		1							1				1		1
Итог	25	25	25	25	25	25	25	25	50	50	50	50	100	100	200

частота, с которой анализируемая единица встретилась в $m_{F'_i}$ внутрисерийных выборках. Нижний горизонтальный ряд (графу «итог») занимают заданные значения N^{a}_{K} (количество внутрисерийных выборок); в вертикальных рядах вниз от индексов промежуточных выборочных совокупностей печатаются значения $m_{F'_i}$ — количество внутрисерийных выборок, в которых анализируемая единица встретилась с частотой (F'_i) класса исследуемого признака. Пустые клетки соответствуют нулевым значениям $m_{F'_i}$.[24]

[24] При чтении таблиц типа II следует строго придерживаться схемы 1 (рис. 1), так как таблицы содержат пустые графы, не заданные схемой, и ряды $N^{a}_{K}=25$, которые при расчете параметров распределений для объемов серий менее 16 тыс. словоупотреблений не учитывались и выведены потому, что используемая программа составлялась для исследований по социологии, где этот вариант был необходим.

Таблица типа II (N^{a}_{1000})

TH1S 119

F'_i	A	B	C	D	E	F	G	H	AB	CD	EF	GH	ABCD	EFGH	A–H
0															
10									1				1		1
11												1		1	1
12									1		1	1	1	2	3
14										1	2		1	2	3
15										1		1	1	1	2
16										1			1		1
17									2	3		1	5	1	6
18									3	2	1		5	1	6
19									2		1		2	1	3
20										1	2	2	1	4	5
21										1			1		1
22										2	1		2	1	3
23									2	1	1	2	3	3	6
24									3	1	2	1	4	3	7
25									3	3		4	6	4	10
26									1		1		1	1	2
27										3	1	1	3	2	5
28											1	3		4	4
30									1	2	1	2	3	3	6
31										2	3		2	3	5
32									1		1	1	1	2	3
33									1		2	1	1	3	4
34									1			1	1	1	2
35											1	1		2	2
36											1	1		2	2
37											1			1	1
38											1			1	1
40									2				2		2
49												1		1	1
55									1	1			2		2
Итог									25	25	25	25	50	50	100

Следует отметить, что ранжированные статистические ряды, как они представлены в таблицах типа II, не являются полными. Дело в том, что если частоте некоторого класса F'_i соответствуют нулевые значения $m_{F'_i}$ по всем вариантам объемов анализируемых совокупностей текста, то это значение F'_i и соответствующие ему нулевые значения $m_{F'_i}$ исследуемого признака не печатаются в целях экономии бумаги.[25] Однако при

[25] Например, за частотой F'_{17} сразу следует F'_{19} (см. табл. типа II, л. 1, стр. 67).

расчете параметров закона распределения в оперативной памяти ЭВМ присутствуют полные статистические ряды $F'_i/m_{F'_i}$ от F'_0 до F'_{max} с шагом $\Delta F'_i$, равным единице. Параметры закона распределения определяются исходя из построенных полных эмпирических рядов частот.

Для данных таблиц типа II выполняется равенство $\Sigma F'_i m_{F'_i} = F_i$ (частоте обследуемой единицы в частотном списке, полученном от объема текста в N^{Σ}_K внутрисерийных выборок).[26]

4.2.2. Нормальное и логарифмически-нормальное распределения

Для этих законов, имеющих место в случае непрерывной случайной величины, строятся непрерывные вариационные ряды.

Построенные в Б18 [27] дискретные ряды для закона Пуассона в Б24 преобразуются в непрерывные. С этой целью производится деление эмпирического ряда с шагом, равным единице, на некоторое число укрупненных классов. Количество таких классов (r) определяется в зависимости от значения величины N^{Σ}_K по табл. 7.[28]

Таблица 7

N^{Σ}_K	25	50	100	200	400
r	5	7	9	10	12

Непрерывность ряда достигается установлением интервала (шага разбиения $\Delta F''_i$).[29] Интервал для каждой обследуемой единицы в пределах одного N^{Σ}_K задается равным, а величина его определяется путем деления разности между максимальным и минимальным значениями F'_i на число классов (r):[30]

$$\Delta F''_i = \frac{F'_{max} - F_{min}}{r}.$$

В качестве примера приводим элементарные расчеты по преобразованию дискретного ряда в непрерывной для словоформы this при $N^{A-H}_{2000} = 200$ (см. табл. типа II, лист 2, стр. 68)

$$\Delta F''_i = \frac{36-3}{10} = 3.3.$$

[26] Контроль за построением ранжированных рядов осуществляется сравнением $\sum F'_i m_{F'_i}$ с соответствующей ей суммой частот в таблицах типа I и с данными частотного списка.

[27] См. блок-схему 1, стр. 54. Блоки 14, 15, 16, 17, 18 выполняют ту же задачу, что и блоки 12, 13 (см. п. 8, стр. 59).

[28] См.: П. Ф. Рокицкий. Биологическая статистика. Минск, 1964, стр. 20.

[29] Обращаем внимание читателя, что F''_i и $m_{F''_i}$ то же, что и F'_i, $m_{F'_i}$, но после приведения рядов $F'_i/m_{F'_i}$ к интервалу (см. сноску 18 на стр. 60, а также стр. 59).

[30] См.: Н. В. Смирнов, И. В. Дунин-Борковский. Курс теории вероятностей и математической статистики. М., 1965, стр. 125.

Округлением полученного значения в большую сторону принимаем $\Delta F''_i = 4$. Тогда частота исходного класса, приведенного к интервалу ряда, определяется как

$$F''_0 = F'_0 + \frac{\Delta F''_i}{2} = 3 + \frac{4}{2} = 5$$

(ср. со значениями F''_0 и $\Delta F''_i$, полученными на машине при выводе графиков для $N^{A-H}_{2000} = 200$, стр. 80).

Частота очередного класса отличается от предыдущего на величину интервала $\Delta F''_i$:

$$F''_1 = F''_0 + \Delta F''_i = 5 + 4 = 9,$$
$$F''_2 = F''_1 + \Delta F''_i = 9 + 4 = 13,$$

и так до конца рядов $F'_i/m_{F'_i}$.

Для получения ряда $m_{F''_i}$ суммируем значения частот $m_{F'_i}$ в пределах каждого интервала:

$$m_{F''_0} = m_{F'_0} + m_{F'_1} + m_{F'_2} + m_{F'_3} = 3 + 6 + 6 + 8 = 23,$$
$$m_{F''_1} = m_{F'_4} + m_{F'_5} + m_{F'_6} + m_{F'_7} = 13 + 10 + 15 + 15 = 53,$$

и так до конца рядов $F'_i/m_{F'_i}$.

В результате приведенный к интервалу ряд для словоформы this при $N^{A-H}_{2000} = 200$ принимает вид табл. 8.

Таблица 8

F''_i	5	9	13	17	21	25	29	33	37
$m''_{F''_i}$	23	53	55	44	18	4	2	0	1

(ср. с данными графика на стр. 53).

Сформированные непрерывные вариационные ряды $F''_i/m_{F''_i}$ на печать ЭВМ не выводятся, а сохраняются в оперативной памяти машины и используются для определения параметров нормального и логарифмически-нормального законов распределений.

Контроль за работой программы и машины осуществляется путем построения вручную с помощью таблиц типа I и II непрерывного ряда, расчета параметров закона и сравнения их с параметрами, полученными с помощью ЭВМ.

§ 5. Статистические характеристики и теоретические вариационные ряды

5.1. Закон Пуассона

Случайная величина дискретного типа, распределенная по закону Пуассона, характеризуется лишь одним параметром λ, численно равным математическому ожиданию случайной величины, причем дисперсия ее также равна λ.

В нашей работе мы определяли среднее арифметическое (a), являющееся аналогом (моделью) математического ожидания:

$$\bar{x} = \frac{\sum_i x_i n_i}{n} = a = \frac{\sum_i F'_i m_{F'_i}}{N_K^{\alpha}}.$$

Вероятности определялись из формулы

$$P_m \simeq \frac{\lambda^m}{m!} e^{-\lambda} \simeq P_{F'_i} = \frac{a^{F'_i}}{F'_i!} e^{-a}.$$

5. 2. Нормальный и логарифмически-нормальный законы распределения

Случайная величина, распределенная по нормальному закону, как вытекает из

$$\varphi_{(n)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

характеризуется двумя параметрами a и σ, поэтому были прежде всего определены оценки этих характеристик.

1) Среднее арифметическое случайной величины для нормального закона распределения:

$$a = \frac{\sum_i x_i \cdot n_i}{n} = \frac{\sum_i F''_i m_{F''_i}}{N_K^{\alpha}}.$$

2) Среднее арифметическое для логарифмически-нормального закона распределения:

$$a' = \frac{\sum_i m_{F''_i} \cdot \lg F''_i}{N_K^{\alpha}}.$$

3) Соответственно дисперсии для нормального и логнормального распределений определяются из

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n} = \frac{\sum_i (F''_i - a)^2 m_{F''_i}}{N_K^{\alpha}},$$

$$(\sigma')^2 = \frac{\sum_i (\lg F''_i - a')^2 m_{F''_i}}{N_K^{\alpha}}.$$

Математическое ожидание (центральный момент первого порядка) и дисперсия (центральный момент второго порядка), наиболее часто используемые статистические характеристики, описывают важные стороны распределения: степень разбросанности значений переменной величины и положение кривой на графике. Для более подробного описания эмпирических распределений мы обратились к моментам более высокого порядка.

Через моменты третьего (M_3) и четвертого (M_4) порядков определяются коэффициенты асимметрии (A) и эксцесс случайной величины (E):

$$M_3 = \frac{\sum_i (x_i - \bar{x})^3 n_i}{n} = \frac{\sum_i (F''_i - a)^3 m_{F''_i}}{N_K^{\alpha}},$$

$$M'_3 = \frac{\sum_i (\lg F''_i - a')^3 m_{F''_i}}{N_K^{\alpha}}.$$

Третий центральный момент имеет размерность куба случайной величины. Для придания ему безмерной характеристики его делят на куб среднего квадратического отклонения. Полученная величина носит название коэффициента асимметрии:

$$A = \frac{M_3}{\sigma^3},$$

$$A' = \frac{M'_3}{(\sigma')^3}.$$

Эксцессы случайных величин (E), или показатели плосковершинности и туповершинности кривой распределения, определяются следующим образом:

$$E = \frac{M_4}{\sigma^4}; \quad E' = \frac{M'_4}{(\sigma')^4};$$

где:

$$M_4 = \frac{\sum_i (x_i - \bar{x})^4 n_i}{n} = \frac{\sum_i (F''_i - a)^4 m_{F''_i}}{N_K^{\alpha}},$$

$$M'_4 = \frac{\sum_i (\lg F''_i - a')^4 m_{F''_i}}{N_K^{\alpha}}.$$

Плотности вероятностей вычислялись по следующим формулам.

1) Для нормального закона распределения:

$$P_{F''_i} = \frac{\Delta F''_i}{\sigma\sqrt{2\pi}} e^{-\frac{(F''_i - a)^2}{2\sigma^2}};$$

2) Для логарифмически-нормального закона:

$$P'_{F'_i} = \frac{\Delta F''_i \lg e}{\sigma' F''_i \sqrt{2\pi}} e^{-\frac{(\lg F''_i - a')^2}{2(\sigma')^2}}.$$

Для этого закона, кроме того, определялась мода

$$M_e = a' - 2.3026\,(\sigma')^2$$

и среднее случайной величины

$$\bar{y} = a' + 1.1513\,(\sigma')^2.$$

Теоретические ряды частот для трех законов распределений строились по произведениям $N''_K P_{F'_i}$ для закона Пуассона, по $N''_K P_{F''_i}$ для нормального закона распределения и $N''_K P'_{F''_i}$ — для логарифмически-нормального.[31]

§ 6. Условия проверки гипотез о законах распределения словоформ и трехсловных сочетаний в тексте

6.1. Сравнение эмпирических рядов частот с теоретическими

Выдвинутые нами гипотезы определили вид и форму предполагаемых эмпирических распределений. Нахождение эмпирических законов распределений свелось к определению основных параметров, характеризующих функцию распределения этих законов: a — для закона Пуассона, a и σ^2 — для нормального и логарифмически-нормального распределений.

Приняв параметр a за математическое ожидание и σ^2 — за дисперсию соответствующих распределений, мы определили теоретические функции распределений (или плотности вероятностей). Другими словами, на основании опытных данных составили теоретические законы распределений.

Дальнейшей задачей исследования явилась проверка правильности поставленной гипотезы. С этой целью было сделано сравнение построенных эмпирических рядов частот ($m_{F'_i}$ и $m_{F''_i}$), характеризующих распределение рассматриваемых признаков в выборочной совокупности, с теоретическими $\left(N''_K P_{F'_i}, N''_K P_{F''_i}\right.$ и $\left. N''_K P'_{F''_i}\right)$, характеризующими признаки в генеральной совокупности. Для этой цели был использован критерий согласия χ^2 (хи-квадрат) Пирсона.

[31] Расчет параметров законов, построение теоретических вариационных рядов и сравнение их с эмпирическими производилось в блоках 20, 26, 31 (см. блок-схему I, рис. 2). Подробно о процедурах сравнения см. ниже.

Применительно к закону Пуассона формула критерия имеет вид:

$$\chi^2 = \sum_i \frac{(x_i - np_i)^2}{np_i} = \sum_i \frac{\left(m_{F'_i} - N''_K P_{F'_i}\right)^2}{N''_K P_{F'_i}}$$

с $(l - 1)$ степенями свободы.[32]

Относительно нормального и логарифмически-нормального законов распределений формулы соответственно записываются как

$$\chi^2 = \sum_i \frac{\left(m_{F''_i} - N''_K P_{F''_i}\right)^2}{N''_K P_{F''_i}}$$

и

$$(\chi')^2 = \sum_i \frac{\left(m_{F''_i} - N''_K P'_{F''_i}\right)^2}{N''_K P'_{F''_i}}$$

с $(l - 2)$ степенями свободы.

При использовании критерия согласно χ^2 для проверки некоторой гипотезы относительно вероятностей особое место занимает вопрос величины математических ожиданий np. Б. Л. ван дер Варден [33] указывает, что в литературе можно встретить замечания, в которых авторы допускают применение этого критерия лишь в том случае, если величины математических ожиданий np больше или равны 5 или 10. Он пишет далее, что исследователи, занимавшиеся этим вопросом, пришли к «более оптимистическому заключению». Хорошее согласие между теоретическими и эмпирическими вероятностями достигается и при значениях $np < 5$ и даже равных единице, если имеется большое число классов.[34]

В нашем исследовании не предполагалось наличия такого количества классов для большинства обследуемых единиц и в особенности для трехсловных сочетаний. В связи с этим было обусловлено, что при вычислении $\Sigma\chi^2$ и определении количества степеней свободы к р а й н и е члены теоретических рядов объединяются до $N''_K P_{F'_i}$, $N''_K P_{F''_i}$ и $N''_K P'_{F''_i} \geqslant 5$, а соответствующие им значения частот ($m_{F'_i}$ и $m_{F''_i}$) эмпирических рядов сводятся в один класс.

Результаты сравнений вместе со статистическими характеристиками законов распределений выдавались на УПч ЭВМ для закона Пуассона в виде формы вывода 1 (стр. 76—77), формы

[32] l — количество членов статистического ряда после группировки значений $N''_K P_{F'_i}$ до $\geqslant 5$ и суммирования соответствующих им значений эмпирических частот $m_{F'_i}$ (см. ниже).

[33] Б. Л. ван дер Варден. Математическая статистика. Русск. пер. М., 1964, стр. 286.

[34] Б. Л. ван дер Варден (ук. соч., стр. 275) приводит пример с десятью классами.

Форма вывода 1

```
takes
133          A                0.18      0.01      0.999      1000
 41.76        8.19            2
takes
133          B                0.20      0.00      0.999      1000
 40.93        9.01            2
takes
133          C                0.18      0.73      0.999      1000
 41.76        8.23            2
takes
133          D                0.24      0.74      0.975      1000
 39.33        9.44            2
takes
133          E                0.26      0.01      0.998      1000
 38.65       11.32            2
takes
133          F                0.34      1.90      0.999      1000
 35.59       14.41            2
takes
133          G                0.08      0.00      0.997      1000
 46.16                        1
takes
133          H                0.22      0.00      0.998      1000
 40.13        9.80            2
takes
133          AB               0.13      0.00      0.999      1000
 82.70       17.20            2
takes
133          CD               0.21      0.06      0.999      1000
 81.06       18.93            2
takes
133          EF               0.30      1.26      0.999      1000
 74.09       25.92            2
takes
133          GH               0.15      0.00      0.999      1000
 86.07       13.88            2
takes
133          ABCD             0.20      0.05      0.999      1000
163.75       36.24            2
takes
133          EFGH             0.22      0.87      1.000      1000
159.70       40.30            2
takes
133          A—H              0.21      2.78      1.000      1000
323.42      68.73  7.85        3
takes
133          AB               0.38      0.02      0.993      2000
 34.19       15.46            2
takes
133          CD               0.42      0.11      0.999      2000
 32.85       17.10            2
takes
133          EF               0.60      1.11      0.999      2000
 27.44      16.46  6.08        3
takes
133          GH               0.30      0.15      0.996      2000
 37.04       12.78            2
takes
133          ABCD             0.40      0.04      0.999      2000
 67.03      26.81  6.08        3
takes
133          EFGH             0.45      0.36      0.999      2000
 63.76      28.96  7.53        3
takes
133          A—H              0.43      0.32      0.999      2000
130.75      55.57  13.66       3
takes
133          ABCD             0.80      0.33      0.991      4000
 22.47      17.97   9.11       3
takes
133          EFGH             0.90      0.04      0.999      4000
 20.33      18.30  11.36       3
takes
133          A—H              0.85      0.14      0.999      4000
 42.74      36.33  15.44  5.46 4
takes
133          A—H              1.70      0.31      0.998      8000
  9.13      15.53  13.20  12.05 4
takes
133          A—H              3.40      3.10      0.976     16000
  8.49       5.47   4.65   5.82 4
```

Форма вывода 2

```
this
118          A        3.00      6.60     15.21    0.57    0.48  0.958  3.02    1000
  6.53      13.28     14.94     13.13                                    4
. . . . . .
. . . . . .
this
118          GH       2.00      7.04     12.52    1.14    2.20  0.953  3.81    1000
  8.18      15.59     21.60     21.74     15.90   12.83                   6
. . . . . .
. . . . . .
this
118          A—H      2.00      5.82     11.98    0.75    0.80  0.977 15.36    1000
 22.38      50.08     80.26     92.12     75.71   44.56   18.78  7.08   8
. . . . . .
. . . . . .
this
118          A—H      4.00     13.14     29.18    0.74    1.26  0.973  3.43    2000
 18.99      44.06     58.08     45.78     20.50    6.17                   6
. . . . . .
. . . . . .
this
118          EFGH     5.00     26.96     59.56    0.26    0.20  0.983  0.19    4000
 14.12      15.39     12.53      7.11                                    4
. . . . . .
. . . . . .
this
118          A—H      9.00     51.04    163.33    0.73    0.67  0.967  1.20    8000
 15.54      14.04     11.29      7.47                                    4
this
118          A—H     13.00    102.50    283.92    0.11    1.20  0.952  3.32   16000
  8.06       7.70      8.06                                              3
```

Форма вывода 3

being 17	A—H	1.00	0.115	0.14	0.05	1.15	1.08	21.27	0.16	0.31	1000
176.04	124.69	60.28	30.69	16.74	9.69	5.89	3.73	6.05		9	
. . . . .											
. . . . .											
being 17	EFGH	2.00	0.41	0.11	0.01	1.03	1.08	6.94	0.17	0.54	2000
47.63	34.73	14.48	6.33	5.79						5	
. . . . .											
. . . . .											
being 17	ABCD	3.00	0.70	0.13	0.36	1.23	0.996	11.70	0.41	0.85	4000
16.30	15.94	8.64	8.91							4	
. . . . .											
. . . . .											
being 17	A—H	4.00	1.06	0.05	0.49	0.36	0.96	2.31	0.94	1.12	8000
5.24	15.06	12.55	7.71	7.92						5	
being 17	A—H	6.00	1.37	0.02	0.12	1.13	0.962	2.83	1.32	1.40	16000
12.65	6.06	5.50								3	

вывода 2 (стр. 77) для нормального и формы вывода 3 (стр. 78) для логарифмически-нормального распределений. Рис. 8 представляет собой фрагмент машинного результата тех же операций.

В форме вывода 1 в верхнем горизонтальном ряду печатается обследуемая единица. Во втором ряду слева направо даются: номер обследуемой единицы в оперативной памяти машины, индекс обследуемой промежуточной совокупности (ср. со схемой на рис. 1, и таблицами типа I, II, стр. 64, 69), значение средней арифметической a; далее идут значения $\Sigma\chi^2$, l, $\Sigma P_{F_i'}$ и объем внутрисерийной выборки. Последний горизонтальный ряд содержит значения $N_K^{\alpha} P_{F_i'}$ после объединения по $N_K^{\alpha} P_{F_i'} \geqslant 5$.

В верхнем ряду формы вывода 2 печатается обследуемая единица; во втором ряду по горизонтали слева направо следуют: номер обследуемой единицы в оперативной памяти машины, индекс промежуточной выборочной совокупности, $\Delta F_i''$, a, σ^2, A, E, $\Sigma P_{F_i''}$, $\Sigma\chi^2$, l и объем внутрисерийной выборки. Нижний горизонтальный ряд занимают значения $N_K^{\alpha} P_{F_i''}$ после группировки для выполнения условия $N_K^{\alpha} P_{F_i''} \geqslant 5$.

В форме вывода 3 во втором ряду по горизонтали после индекса выборочной совокупности печатаются значения $\Delta F_i''$, a', $(\sigma')^2$, A', E', $\Sigma P_{F_i'}'$, $\Sigma(\chi')^2$, M_e, $\bar{y}$ и объем микровыборки.

По закону Пуассона и нормальному распределению получены данные в виде описанных выше форм вывода для 300 словоформ и 300 триад. Для логарифмически-нормального распределения

выведены данные для 150 первых словоформ частотного списка. Всего, согласно схеме на рис. 1, получено около 35 тыс. значений $\Sigma\chi^2$ и более 70 тыс. строк параметров законов распределений.

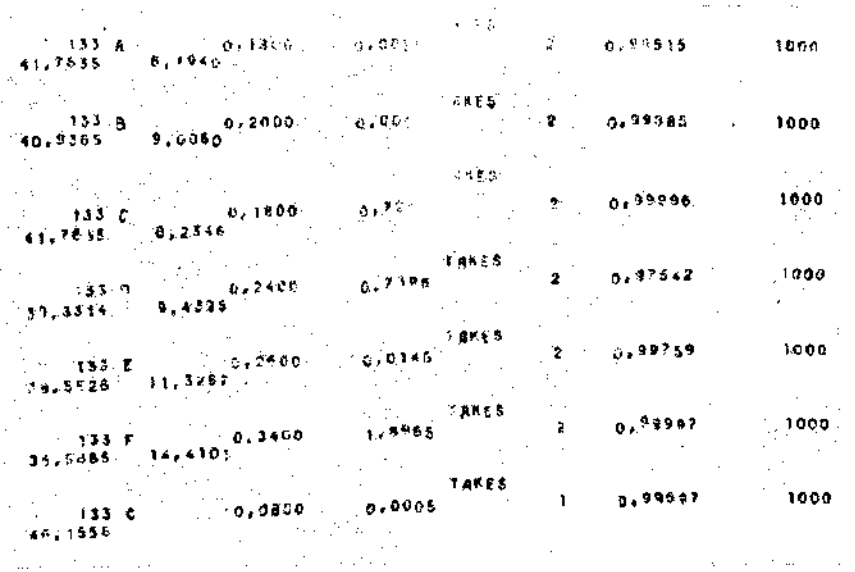


Рис. 8. Форма вывода 1 (начало).

6.2. Условие соответствия 1

В настоящей статье ввиду большого объема полученных данных *мы не задавались целью дать оценку всем полученным статистическим характеристикам.* По этой причине остановимся лишь на результатах сравнения степени соответствия эмпирических распределений теоретическим законам по критерию согласия χ^2 Пирсона.

Задаемся тремя уровнями значимости (0.1, 0.05, 0.01) и считаем, что, если при $(l-1)$ степенях свободы для закона Пуассона и $(l-2)$ для нормального и логарифмически-нормального распределений полученные значения $\Sigma\chi^2$ равны или меньше табличных,[35] то эмпирические распределения соответствуют теоретическим законам с различной степенью приближения, характеризующейся одним из заданных уровней значимости.

С результатами сравнения, представленными в виде форм вывода 1, 2, 3, работать трудно; поэтому для упрощения про-

[35] Б. Л. ван дер Варден, ук. соч., стр. 408; А. И. Карасев. Основы математической статистики, стр. 348.

цедуры наблюдения сводим (при выполнении условия 1) все полученные значения $\Sigma\chi^2$ в пять отдельных таблиц: каждая для одного закона и типа анализируемых единиц текста. Образец одной из них представлен ниже в табл. 9.

Соответствие эмпирических распределений словоформ и триад теоретическим законам также хорошо подтверждается или отвергается путем сопоставления эмпирических и теоретических кривых на графиках, построенных машиной. Проиллюстрируем это на нескольких примерах для нормального закона распределения.

В графиках, которые мы приводим ниже (рис. 9—10), на оси ординат откладываются значения теоретических $\left(N_K^a P_{F_i''}\right)$ и эмпирических $(m_{F_i''})$ частот, отмеченные соответственно цифрами 1 и 2. Эти значения отпечатаны в двух колонках цифр слева до объединения статистических рядов по $N_K^a P_{F_i''} \geqslant 5$ (условие применения критерия согласия χ^2 см. выше, стр. 75). Ряд $N_K^a P_{F_i''}$ после объединения по $N_K^a P_{F_i''} \geqslant 5$ для варианта $N_{2000}^{\text{А—Н}}$ словоформы this можно также найти в форме вывода II (стр. 77).

На оси абсцисс откладываются значения частот F_i''. Этот ряд легко восстанавливается из значений первой частоты (F_0'') ряда и интервала разбиения $\Delta F_i''$ (подробный расчет рядов F_i'' и $m_{F_i''}$ для варианта $N_{2000}^{\text{А—Н}} = 200$ словоформы this см. выше — стр. 71, 77). За точностью выполнения машиной всех процедур формирования рядов частот F_i'' и $m_{F_i''}$ для варианта $N_{2000}^{\text{А—Н}} = 200$ словоформы this можно проследить по данным частотного списка, табл. типа I (стр. 64—65), табл. типа II (стр. 67—69), табл. 8 (стр. 71), в которой приводятся результаты ручной проверки, формы вывода 2 (стр. 77), наконец, по данным графика 1 (рис. 9).

Поскольку формат бумаги для вывода результатов на АЦПУ постоянный, для каждого графика автоматически определяется масштаб относительно ее ширины. При этом за исходные величины принимаются максимальные значения $N_K^a P_{F_i''}$. Отсюда следует, что графики для разных вариантов N_K^a несовместимы, так как у каждого графика ордината остается постоянной и всегда равна ширине бумаги или значению $N_K^a P_{F_i''} = \max$.

По данным основной таблицы (здесь не приводится), для словоформы this соответствие частот $m_{F_i''}$ и $N_K^a P_{F_i''}$ по критерию χ^2, согласно условию 1, наблюдается для всех объемов внутрисерийных выборок и выборочных совокупностей, за исключением варианта $N_{1000}^{F} = 50$. Несоответствие эмпирических характеристик распределения с теоретическими для этого последнего варианта хорошо видно на графике 1. Графическое представление распределения терминологической словоформы turbine, несмотря

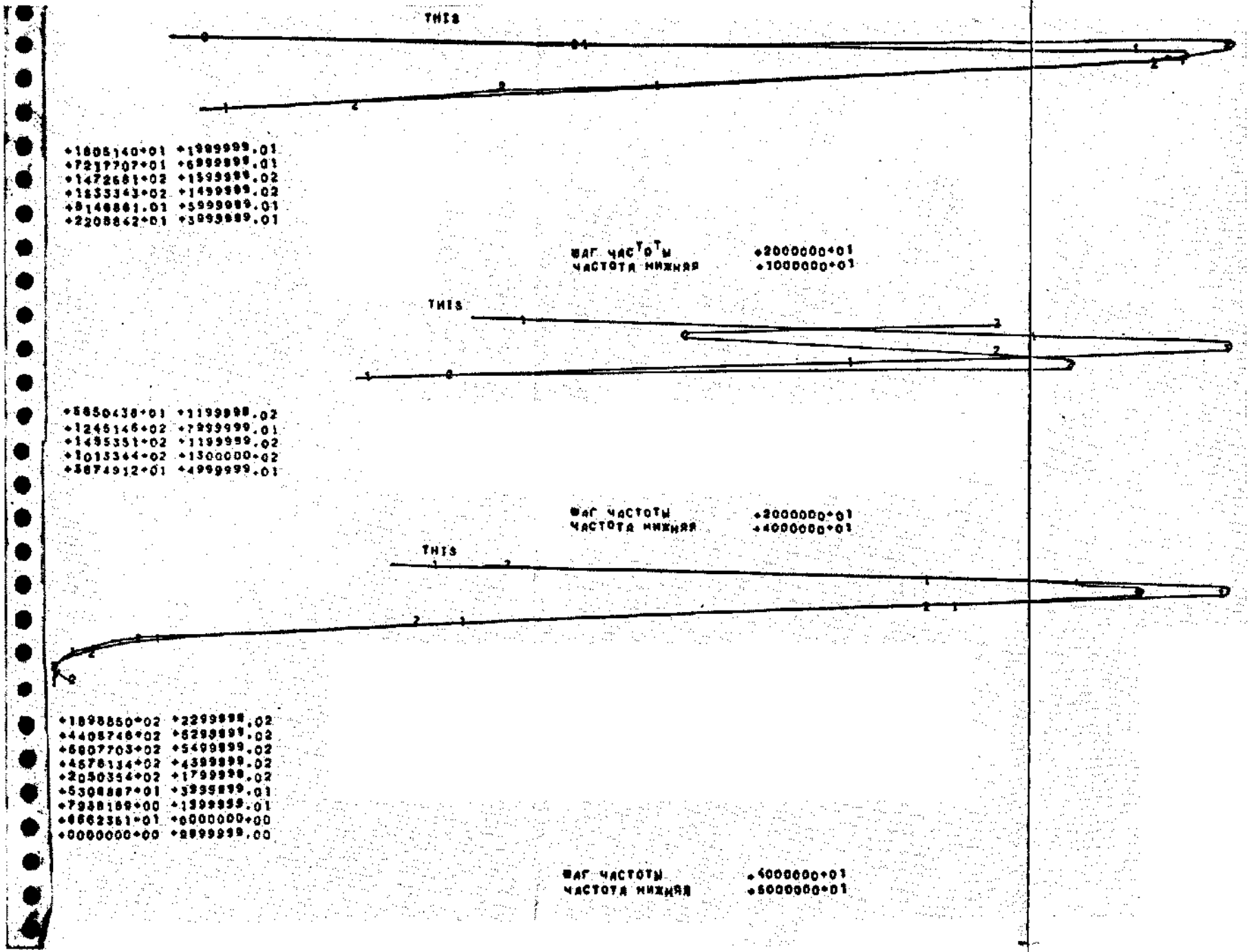


Рис. 9. График 1.

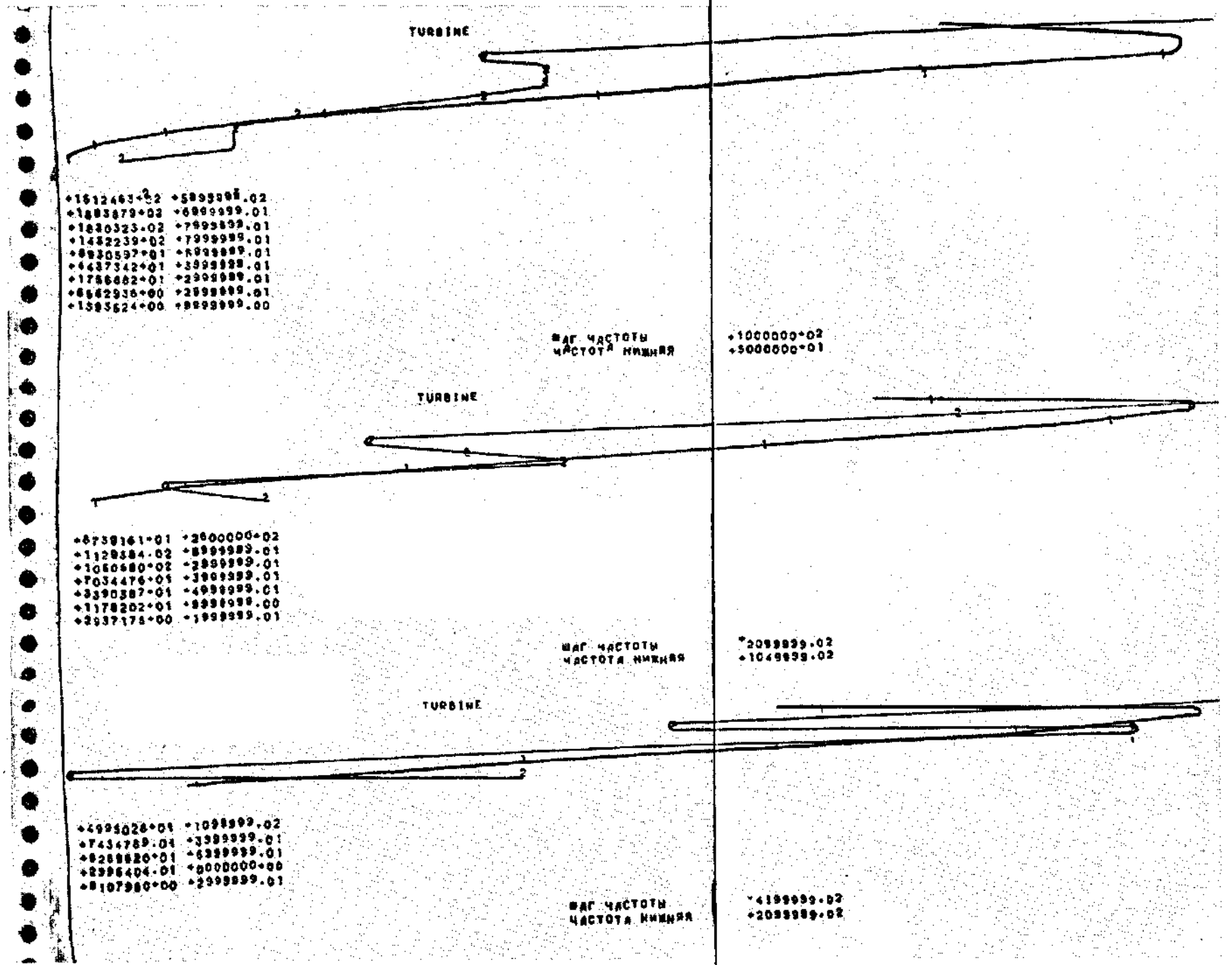


Рис. 10 График 2.

на ее большое значение F_i в частотном списке, показывает полное несовпадение эмпирического и теоретического распределений для всех 27 вариантов нормировок, включая и наибольший объем внутрисерийной выборки. При больших расхождениях значений $m_{F_i''}$ и $N_X'' P_{F_i''}$ точки, соответствующие максимальным значениям $m_{F_i''}$, выходят за пределы формата бумаги.

6.3. Условие соответствия 2

Напомним, что задачей исследования является анализ распределений лингвистических единиц (словоформ и триад) во всей выборочной совокупности текста объемом 400 тыс. словоупотреблений. Для каждой обследуемой единицы статистические характеристики определялись по 27 вариантам выборочных совокупностей при пяти разных объемах внутрисерийных выборок. Естественно, что соответствие эмпирических данных теоретическим по всем вариантам и для всех единиц мы получить не можем. Отсюда возникает вопрос: в каком случае считать, что эмпирические распределения соответствуют теоретическим законам? Для преодоления этого затруднения и окончательного решения вопроса вводим дополнительное условие соответствия 2.

Эмпирические распределения соответствуют теоретическим законам в том и лишь только в том случае, если условие 1 (стр. 79) выполняется для варианта, охватывающего общую выборочную совокупность, или для группы вариантов независимых выборочных совокупностей. При этом сумма их объемов также должна составлять весь выборочный текст длиной в 400 тыс. словоупотреблений.

Выполнение этого условия позволяет свести все пять таблиц в два легкообозримых списка сравнения трех законов распределений: один для словоформ (стр. 82 и сл.), другой для триад стр. 92 и сл.).

§ 7. Общий анализ полученных данных и оптимальные условия исследования распределений словоформ и трехсловных сочетаний в тексте

7.1. Эмпирические распределения словоформ относительно двух законов — Пуассона и нормального

Ниже приводится список 1 (стр. 82—88), в котором представлены результаты исследования распределений словоформ в текстах английского подъязыка судовых механизмов по трем законам.

В списке словоформ: i — порядковый номер словоформ в частотном списке; F_i — абсолютная частота словоформы в частотном списке. Грамматические классы словоформ даются по толковому

Список анализируемых словоформ

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
1	the	art	38028	—	+	+
2	of	prep	16630	—	+	+
3	and	conj	11025	—	+	+
4	to	prep/adv	9484	—	+	+
5	in	prep/adv/adj	9010	—	+	+
6	is	v	8764	—	+	+
7	a	art	8322	—	+	+
8	x		7151	—	+	+
9	z		5290	—	+	+
10	be	v	4208	—	+	+
11	for	prep/conj	3705	—	+	+
12	are	v	3675	—	+	+
13	by	prep/adv	3203	—	+	+
14	with	prep	3087	—	+	+
15	that	pron/adj/conj	2644	—	+	+
16	at	prep	2603	—	+	+
17	as	adv/conj/pron/prep	2573	—	+	+
18	this	pron/adj	2566	—	+	+
19	on	prep/adv/adj	2526	—	+	+
20	which	pron/adj	2426	—	+	+
21	it	pron	2414	—	+	+
22	from	prep	2156	—	+	+
23	steam	n/v	2080	—	+	+
24	engine	n	1947	—	+	+
25	or	conj	1920	—	+	+
26	y		1895	—	+	+
27	turbine	n	1703	—	—	+
28	pressure	n	1699	—	+	+
29	an	art	1409	+	+	+
30	water	n/v/adj	1327	—	+	+
31	has	v	1272	—	+	+
32	fuel	n/v	1252	—	+	+
33	will	v	1234	—	+	+
34	oil	n/v/adj	1194	—	+	+
35	valve	n/v	1168	—	—	—
36	was	v	1068	—	—	+
37	air	n/v/adj	1056	—	+	+
38	have	v	1032	—	+	+
39	not	adv	1029	+	+	+
40	when	adv/conj/pron	1009	—	+	+
41	through	prep/adv	1005	—	+	+
42	been	v	988	—	+	+
43	may	v	971	—	+	+
44	boiler	n	953	—	—	+
45	these	pron/adj	911	—	+	+
46	cylinder	n	902	—	—	+

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
47	one	adj/n/pron	900	+	+	+
48	two	adj/n	898	—	+	+
49	can	v/n	889	—	+	+
50	if	conj	824	—	+	+
51	temperature	n	815	—	—	+
52	engines	n	775	—	—	+
53	used	v	772	+	+	+
54	power	n	759	—	—	+
55	each	adj/pron/adv	756	—	+	+
56	type	n/v	748	—	+	+
57	into	prep	732	—	+	+
58	should	v	723	—	—	+
59	all	adj/adv/n	714	+	+	+
60	gas	n/v	714	—	—	—
61	system	n	714	—	+	+
62	there	adv	711	+	+	+
63	than	conj	689	+	+	+
64	so	adv/conj/pron	686	+	+	+
65	pump	n/v	680	—	—	+
66	between	prep/adv/pron	667	+	+	+
67	main	n/adj	661	—	+	+
68	heat	n/v	660	—	—	+
69	per	prep	659	—	+	+
70	exhaust	v/n	654	—	—	+
71	but	prep/conj/adv	636	+	+	+
72	design	v/n	626	—	+	+
73	also	adv	617	+	+	+
74	being	v/n	612	—	+	+
75	piston	n	610	—	—	+
76	must	v/n	603	—	—	+
77	its	pron	599	+	+	+
78	turbines	n	597	—	—	+
79	were	v	593	—	+	+
80	speed	n/v	590	—	+	+
81	high	adj/adv	584	—	+	+
82	more	adj/adv	572	+	+	+
83	would	v	570	—	—	+
84	such	adj/pron	562	+	+	+
85	valves	n	560	—	+	+
86	combustion	n	548	—	+	+
87	other	adj/pron/adv	539	+	+	+
88	some	adj/pron	530	+	+	+
89	any	adj/pron	521	+	+	+
90	shaft	n	514	—	—	+
91	flow	v/n	501	—	—	+
92	up	adv prep	496	+	+	+
93	only	adj/adv	495	+	+	+

Список 1 *(продолжение)*

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
94	control	v/n	491	−	−	+
95	made	v	476	+	+	+
96	end	n/v	472	−	+	+
97	large	adj	471	−	+	+
98	no	adv	455	+	+	+
99	about	adv/prep	449	+	+	+
100	shown	v	449	+	+	+
101	they	pron	433	+	+	+
102	bearing	n	425	−	−	+
103	blades	n	425	−	−	−
104	operation	n	425	−	+	+
105	same	adj/adv/pron	423	+	+	+
106	use	v/n	421	+	+	+
107	then	adv/adj	413	−	+	+
108	very	adj/adv	412	+	+	+
109	lower	adj/adv	407	+	+	+
110	fitted	v	405	−	+	+
111	ship	n/v	400	−	+	+
112	gear	n/v	398	−	+	+
113	first	adj/adv	395	+	+	+
114	out	adv/prep/adj	391	+	+	+
115	pumps	n/v	386	−	−	+
116	efficiency	n	384	−	−	+
117	required	v	384	+	+	+
118	reactor	n	376	−	−	−
119	under	prep/adv/adj	376	+	+	+
120	during	prep	374	+	+	+
121	where	adv/conj	373	+	+	+
122	diesel	n	367	−	−	−
123	low	adj/adv	365	+	+	+
124	small	adj	364	+	+	+
125	boilers	n	358	−	−	+
126	side	n/adj/v	357	−	+	+
127	velocity	n	356	−	−	−
128	conditions	n	355	−	+	+
129	operating	v/n	352	−	+	+
130	necessary	adj	348	+	+	+
131	their	pron	348	+	+	+
132	possible	adj	343	+	+	+
133	number	n/v	342	+	+	+
134	steel	n/adj/v	342	−	+	+
135	cent	n	339	−	+	+
136	time	n/v/adj	333	−	+	+
137	bearings	n	332	−	−	+
138	both	adj/pron/conj	330	+	+	+
139	thus	adv	330	−	+	+

Список 1 *(продолжение)*

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
140	nozzle	n	325	−	−	+
141	tubes	n	323	−	−	−
142	marine	adj/n	322	−	−	+
143	given	v	320	+	+	+
144	after	adv/adj/conj	319	+	+	+
145	maximum	n/adj	319	−	+	+
146	feed	v/n	313	−	+	+
147	machinery	n	313	−	−	−
148	top	n/v	312	−	+	+
149	work	v/n	310	−	−	+
150	above	adv/prep/adj	309	+	+	+
151	since	adv/prep/conj	308	+	+	
156	over	adv/prep/adj	298	+	+	
161	most	adj/adv	291	+	+	
166	provided	v/conj	281	−	+	
171	cylinders	n	269	−	−	
176	amount	v/n	265	+	+	
181	means	n/v	258	+	+	
186	casing	n	253	−	−	
191	arrangement	n	246	−	+	
196	weight	n/v	241	+	+	
200	without	adv/prep	237	+	+	
205	types	n	232	−	+	
210	level	n/adj/adv	226	−	−	
215	ratio	n	222	−	−	
220	output	n	220	−	−	
225	vessel	n	217	−	−	
230	open	adj/v	210	−	+	
235	parts	n/v	209	+	+	
240	stroke	n/v	208	−	−	
246	four	n/adj	202	+	+	
251	compression	n	198	−	−	
256	size	n	195	+	+	
261	now	adv/conj	191	+	+	
266	drive	v/n/adj	187	+	−	
271	well	n/adv	185	+	+	
276	thrust	v/n	183	−	−	
281	greater	adj	179	+	+	
286	increased	v	175	+	+	
291	pressures	n	172	+	+	
296	carried	v	170	+	+	
301	way	n	169	+	+	
305	below	adv/prep	166	+	+	
310	closed	v	164	+	+	
315	moving	v	164	−	−	
320	consists	v	160	+	+	
325	generator	n	159	−	−	

Список 1 *(продолжение)*

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
330	stages	n	159	−	−	−
335	lever	n	157	−	−	−
340	suction	n	156	−	−	−
345	wheel	n/v	154	−	−	−
350	special	adj	152	+	+	+
355	see	v	149	+	+	+
360	developed	v	145	+	+	+
364	second	adj/n/v	144	+	+	+
368	ports	n	140	+	+	+
372	vacuum	n	139	−	−	−
377	specific	adj	137	−	−	−
381	considered	v	135	+	−	−
386	approximately	adv	132	+	+	+
391	change	v/n	131	+	+	+
395	scavenging	v/n	130	−	−	−
400	hand	n/adj/v	129	+	+	+
405	heating	n/v	126	+	−	−
410	using	v	125	+	+	+
414	materials	n	123	+	−	−
419	period	n	122	+	+	+
423	passes	n/v	121	+	+	+
427	modern	adj	119	+	+	+
431	what	pron/adj/adv	118	+	−	−
435	done	v	116	+	−	−
439	do	v	115	+	+	+
443	proper	adj	115	−	−	−
447	capacity	n	114	+	+	+
451	located	v	113	+	−	−
455	failure	n	112	+	+	+
459	corrosion	n	111	−	−	−
463	heavy	adj	111	+	+	+
467	rods	n	110	+	+	+
471	indicated	v	109	+	−	−
475	gearing	n	108	+	−	−
479	scavenge	v	108	+	−	−
483	contact	n/v	108	+	+	+
487	reduce	v	107	+	+	+
491	stress	v/n	106	−	−	−
495	course	n	104	+	+	+
499	hot	adj	104	+	−	−
501	short	adj	103	+	+	+
505	hours	n	102	−	−	−
509	series	n	101	+	−	−
513	mixture	n	99	−	−	−
517	common	adj	98	+	+	+
521	problem	n	98	+	+	+
525	manoeuvring	v/n	97	+	+	+

Список 1 *(продолжение)*

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
529	assembly	n	96	+	+	+
533	factors	n	95	+	+	+
537	filter	n	94	−	−	−
541	great	adj	93	+	+	+
545	require	v	93	+	+	+
549	crank	n/adj	92	−	−	−
555	view	n/v	92	+	+	+
559	critical	adj	91	+	+	+
563	seen	v	90	+	+	+
567	timing	n	89	−	−	−
571	particular	adj	88	+	+	+
575	arrangement	n	86	+	−	−
580	indicator	n	86	+	−	−
584	takes	v	86	+	+	+
587	automatic	adj	85	+	+	+
591	pinion	n	85	+	−	−
595	few	adj	84	+	+	+
610	circuit	n	81	+	−	−
620	thickness	n	81	+	−	−
630	resulting	v	80	+	−	−
640	reading	n/v	79	−	−	−
650	slightly	adv	78	+	−	−
660	cleaning	n/v	76	−	−	−
670	compared	v	75	+	−	−
680	connecting	v	74	+	−	−
690	spaces	n	74	+	−	−
700	occur	v	73	+	−	−
710	might	v/n	72	+	−	−
721	degrees	n	70	−	−	−
731	shut	v/adj/n	70	+	−	−
733	already	adv	69	+	+	+
740	recent	adj	69	+	−	−
750	life	n	68	+	+	+
760	immediately	adv	67	+	+	+
770	clutch	v/n	66	−	−	−
780	limited	v	66	+	+	+
790	current	adj/n	65	−	−	−
800	production	n	65	+	−	−
810	associated	v	63	+	+	+
820	rapid	adj	63	+	+	+
830	sizes	n	62	+	+	+
840	measured	v	61	+	−	−
850	noted	v	60	+	+	+
860	entire	adj	59	+	+	+
870	according	adj	68	+	+	+
880	anlysis	n	57	−	−	−
890	shafts	n	57	−	−	−

Список 1 *(продолжение)*

i	Словоформа	Грамматический класс	F_i	Закон Пуассона	Нормальный закон	Логнормальный закон
900	enough	adj/adv	56	+	+	
910	slight	adj	56	+	+	
920	making	v/n	55	+	—	
930	content	adj/v	54	+	—	
940	proved	v	54	+	+	
950	horisontal	adj	53	+	—	
975	navy	n	52	+	+	
1000	tip	n/v	51	+	—	
1024	practically	adv	49	+	+	
1048	craft	n	47	+	—	
1073	left	adj/v	46	+	+	
1097	introduction	n	45	+	+	
1122	round	adj	44	+	+	
1147	shafting	n	43	+	—	
1172	building	n/v	41	+	—	
1200	yet	adv/conj	41	+	—	
1248	necessity	n	39	+	+	
1301	providing	v/conj	37	+	—	
1350	day	n	35	+	—	
1497	indicates	v	31	+	—	

английскому словарю.[36] (В этой связи см. также описание табл. 2, стр. 50). Знаком «+» отмечены словоформы, отвечающие условиям соответствия 1, 2 эмпирических распределений теоретическим законам; знаком «—» указаны значения $\Sigma\chi^2$, превышающие табличные для уровня значимости 0.01.

Рассмотрим полученные данные по двум законам (Пуассона и нормальному), не учитывая пока лексико-грамматических характеристик словоформ. При этом выясняется, что из 300 отобранных для обследования словоформ эмпирические распределения 157 единиц соответствуют закону Пуассона, а 143 не подчиняются этому закону. Нормальному закону соответствуют распределения 192 словоформы, а 108 единиц не согласуются с законом. При этом из 300 единиц 42 словоформы подчиняются только закону Пуассона, в то время как распределения 77 словоформ согласуются только с нормальным законом. Распределения 115 словоформ подчиняются и тому и другому закону одновременно. Эмпирические распределения 66 словоформ остались неизвестными.

[36] См. прим. 11 на стр. 50.

Таблица 9

Результаты сравнения степени соответствия эмпирических рядов частот теоретическим по критерию согласия χ^2 Пирсона (словоформы по нормальному закону распределения)

			$N'_{1000}=50$									$N'_{1000}=100$				$N'_{1000}=200$		$N'_{1000}=400$	$N'_{2000}=50$				$N'_{2000}=100$		$N'_{2000}=200$	$N'_{4000}=50$		$N'_{4000}=100$	$N'_{8000}=50$	$N'_{16000}=25$
i	Обследуемые единицы	F_i	A	B	C	D	E	F	G	H	AB	CD	EF	GH	ABCD	EFGH	A—H	AB	CD	EF	GH	ABCD	EFGH	A—H	ABCD	EFGH	A—H	A—H	A—H
1																													
2	of	16630	×	+	+	+	+	+	+	+	+	+	+	+	×	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
83	would						v						v	+												v			
. . .																													
300																													

В таблице: i — порядковый номер словоформы в частотном списке; F_i — абсолютная частота словоформы в частотном списке; N — количество внутрисерийных выборок заданного объема; латинские буквы — индексы выборочных совокупностей; знак + — значение $\Sigma\chi^2$, равное или меньшее табличного $cl-2$ степенями свободы при уровне значимости 0.1; знак × — значение $\Sigma\chi^2$ при уровне значимости 0.05; знак v — значение $\Sigma\chi^2$ при уровне значимости 0.01; пустые клетки соответствуют значениям $\Sigma\chi^2$, превышающим табличные.

На первый взгляд создается довольно неопределенная картина, и судить по этим данным о закономерностях распределения словоформ в тексте трудно.

Проанализируем полученные данные с учетом порядковых номеров и абсолютных частот словоформ, при которых их эмпирические распределения согласуются/не согласуются с теоретическими законами Пуассона и Гаусса.

Из 60 первых словоформ частотного списка закону Пуассона соответствуют лишь 5 единиц, в то время как нормальному — 50 словоформ. До номера 296 сохраняется все то же преобладание в количестве словоформ, подчиняющихся нормальному закону, а с номера 300 начинает наблюдаться обратное явление. Это хорошо видно, если анализируемый список словоформ представить в виде табл. 10.

Таблица 10

i	П	Н	i	П	Н
1—30	1	29	301—435	21	15
31—60	4	21	439—555	23	16
61—90	12	22	559—790	24	10
91—120	15	23	800—1497	29	15
121—150	11	21			
151—296	17	20		157	192

Табл. 10 представляет собой список анализируемых единиц, разделенный на 10 зон с шагом 30 словоформ. В первом слева вертикальном ряду указаны зоны по номерам словоформ в частотном списке, во втором — количество словоформ, подчиняющихся закону Пуассона в пределах каждой зоны, в третьем — то же для нормального закона распределения.

С точки зрения двух рассматриваемых законов распределений словоформ в текстах английского подъязыка судовых механизмов частотный список, на основании полученных результатов при заданных нами условиях исследования,[37] можно разделить на четыре зоны:

1) от номера 1 до 60—90 — зона действия нормального закона распределения;

2) от номеров 60—90 до 250—300 — смешанная зона с преобладанием словоформ, отвечающих условиям нормального закона распределения;

3) от номеров 250—300 до 1500 — смешанная зона с преобладанием словоформ, подчиняющихся закону Пуассона;

[37] *Здесь и далее заданными предполагаются условия, описанные выше* (стр. 52—54). См. также схему 1 (рис. 1), гистограмму 1 (рис. 11) для словоформ и гистограмму 2 (рис. 12) для триад.

4) примерно от номера 1500 и далее — зона, отвечающая условиям закона Пуассона.

Дальнейший анализ списка 1 дает возможность обнаружить еще два довода в пользу сделанных выше выводов. В связи с этим отметим, что:

а) все 77 словоформ, распределения которых отвечают условиям только нормального закона, находятся в зоне порядковых номеров 1—230; при этом первая словоформа, подчиняющаяся только закону Пуассона, имеет порядковый номер 266 (см. drive), т. е. все словоформы, соответствующие только закону Пуассона, находятся в зоне $i \geqslant 266$;

б) большинство (43 из 66) словоформ, которые не подчиняются ни одному из двух законов, находятся в диапазоне действия смешанной зоны с преобладанием словоформ, отвечающих условиям нормального закона распределения. Это, по-видимому, должно свидетельствовать о том, что выделенная зона нормального распределения не имеет существенного отношения к закону Пуассона, и, следовательно, предположение, сделанное выше относительно зон действия двух законов распределения (Пуассона и нормального) при заданных условиях наблюдения можно, вероятно, считать правильным.

Из высказанных выше соображений следуют три общих вывода.

1) При заданных условиях наблюдения обнаруживается зависимость между порядковым номером словоформы (соответственно ее абсолютной частотой) и видом закона распределения (см. выше, пп. 1—4, стр. 89—90).

2) Эта зависимость позволяет сформулировать следующие рекомендации к методике исследования характера распределения лингвистических единиц в научно-техническом тексте: в аналогичных заданным нами условиях наблюдения исследование распределений словоформ относительно нормального закона следует, вероятно, осуществлять в зоне порядковых номеров частотного списка 1—300, а по закону Пуассона началом зоны, по-видимому, нужно считать порядковые номера 250—300.

3) И наконец, анализ основных таблиц дает основание предполагать, что при увеличении объема общей выборочной совокупности, т. е. с увеличением частот словоформ, условная нижняя граница зоны действия нормального закона будет смещаться в сторону увеличения порядковых номеров словоформ в частотном списке, отодвигая в свою очередь в ту же сторону верхнюю границу зоны действия закона Пуассона. И наоборот, с уменьшением объема общей выборочной совокупности (с соответственным уменьшением частот) условная нижняя граница зоны действия нормального закона будет смещаться в сторону уменьшения

порядковых номеров, отодвигая в ту же сторону верхнюю границу зоны действия закона Пуассона. Это, разумеется, хорошо согласуется с известными в теории вероятностей и математической статистике соотношениями обоих законов.

7.2. Эмпирические распределения словоформ и логарифмически-нормальный закон

Изложенные выше соображения относительно нормального закона полностью относятся и к логарифмически-нормальному закону. Из 150 подвергнутых анализу словоформ первых номеров частотного списка распределения 142 отвечают условиям последнего, причем 8 единиц, не подчиняющихся логнормальному закону, относятся к 66 словоформам, эмпирические распределения которых не согласуются ни с одним из первых двух законов.

7.3. Эмпирические распределения трехсловных сочетаний и законы Пуассона и Гаусса

В списке 2 (стр. 92 и сл.) приводим результаты исследования характера распределения 300 триад английского подъязыка судовых механизмов относительно законов Пуассона и Гаусса. Здесь знаком «+» отмечены триады, отвечающие условиям соответствия 1,2 (см. стр. 79 и сл.) при уровне значимости 0.1, знаком «×» — при уровне значимости 0.05 и знаком «v» — при 0.01. Знак минус указывает, что значения $\Sigma\chi^2$ превышают табличные для уровня значимости 0.01. Триады расположены по типам в последовательности 2, 4, 3, 1 (о типах триад см. выше, стр. 51 и сл.). 11 триад, отобранных из частотного списка первого типа,[38] имеют два значения F_i; первое — для $N^{A}_{1000} = 50$ и второе — для $N^{A-H}_{1000} = 400$. Второе значение было определено по таблицам типа I. У двух последних триад вторые значения F_i остались неизвестными, так как таблицы для этих словоформ на печать ЭВМ не выводились.

Закону Пуассона соответствуют распределения 272 единиц, нормальному — 114. Из 272 только закону Пуассона подчиняются 163, а из 114 только с нормальным распределением согласуются лишь 5 сочетаний. Распределения 23 единиц не соответствуют ни одному из заданных теоретических законов.

При анализе результатов исследования распределений триад относительно этих двух законов обнаруживается следующая особенность. Если эмпирические распределения триад в текстах отвечают условиям закона Пуассона в основном по большинству или даже по всем объемам внутрисерийных выборок, то для нормального закона, которому, как видно из списка, и подчи-

[38] Частотный список триад этого типа был получен от массива А.

Список анализируемых триад

i	Триады	F_i	Закон Пуассона	Нормальный закон
1	x z △	543	—	+
2	of x z	372	—	v
3	△ it is	358	+	+
4	in z x	319	—	v
5	z x △	319	—	—
6	x per cent	316	—	v
7	x and x	276	—	—
8	at x z	272	—	—
9	△ in the	252	+	+
10	x z and	223	—	v
11	to x z	212	+	×
12	of the engine	211	—	—
13	x to x	207	—	—
14	△ z x	192	+	v
15	shown in z	190	+	—
16	△ if the	179	+	—
17	end of the	179	+	×
18	△ this is	160	+	+
19	about x z	149	+	—
20	△ when the	141	+	v
21	of the steam	136	—	—
22	so that the	134	+	+
23	part of the	125	+	—
24	the use of	124	+	×
25	is x z	122	—	—
26	x z x	121	—	—
27	x z at	116	v	v
28	and x z	112	+	v
29	there is a	108	+	×
30	it will be	106	+	+
31	△ there are	105	+	—
32	△ this is	98	+	+
33	figure x z	97	—	—
34	side of the	97	+	+
35	of the turbine	96	+	—
36	z x shows	94	+	v
37	z x z	93	—	—
38	as shown in	90	+	+
39	x z of	89	+	v
40	is shown in	88	+	—
41	of the cylinder	88	+	—
42	the engine △	88	+	v
43	top of the	87	+	—
44	z at x	86	—	—
45	in which the	85	+	×
46	one of the	85	+	—
47	of the fuel	82	+	v
48	△ there is	80	+	—
49	due to the	80	+	×
50	in the case	80	+	—

i	Триады	F_i	Закон Пуассона	Нормальный закон
51	the engine is	78	+	v
52	and x △	77	+	—
53	of the main	77	+	—
54	bottom of the	76	+	—
55	the end of	76	+	v
56	z x and	76	+	—
57	△ the main	75	+	v
58	up to x	75	+	—
59	this type of	74	+	—
60	from x to	72	+	×
61	the gas turbine	72	—	—
62	there is no	72	+	—
63	△ it was	71	+	—
64	pressure in the	71	+	+
65	△ as the	68	+	+
66	△ the engine	68	+	—
67	△ it will	67	+	—
68	consists of a	67	+	v
69	z and x	67	+	v
70	△ in addition	66	+	+
71	of the boiler	66	+	—
72	it may be	64	+	v
73	the turbine △	64	+	—
74	of about x	63	+	+
75	of the z	63	+	—
76	pressure of x	63	+	+
77	temperature of the	63	+	+
78	△ the steam	62	+	—
79	because of the	62	+	—
80	of the piston	62	+	—
81	that of the	62	+	+
82	x z the	62	+	—
83	and it is	61	+	+
84	in the cylinder	61	—	—
85	x is the	61	+	—
86	x z per	61	+	—
87	some of the	60	+	+
88	the boiler △	59	+	—
89	the design of	58	+	×
90	means of a	57	+	+
91	portion of the	57	+	—
92	see z x	57	—	—
93	shown in figure	57	—	—
94	the steam drum	57	+	—
95	the z turbine	57	+	—
96	x and the	57	+	—
97	at the same	55	+	v
98	connected to the	55	+	v
99	it is not	55	+	+
100	x z bore	55	+	—
101	△ the first	54	+	v
102	△ the fuel	54	+	—

Список 2 *(продолжение)*

i	Триады	F_i	Закон Пуассона	Нормальный закон
103	△ the oil	54	+	—
104	a x z	54	+	—
105	z and the	54	+	—
106	the steam is	53	+	—
107	to the turbine	53	+	—
108	x z in	53	+	v
109	△ in a	52	+	+
110	△ on the	52	+	—
111	the temperature of	52	+	—
112	it has been	51	+	—
113	of the ship	51	+	—
114	operation of the	51	+	+
115	speed of the	51	+	—
116	the x z	51	+	—
117	x z to	51	+	+
118	in the boiler	50	+	—
119	of the oil	50	+	—
120	the speed of	50	+	—
121	△ at the	49	+	+
122	of the valve	49	+	—
123	attached to the	48	+	—
124	by y △	48	+	—
125	in the steam	48	+	—
126	of this type	48	+	+
127	water in the	48	+	+
128	△ for the	47	+	v
129	△ in order	47	+	—
130	△ it has	47	+	—
131	cent of the	47	+	—
132	of the reactor	47	+	—
133	△ the z	46	+	—
134	△ they are	46	+	+
135	have to be	46	+	—
136	is necessary to	46	+	—
137	it is possible	46	+	—
138	of the air	46	+	—
139	△ with the	45	+	×
140	as a result	45	+	—
141	end of the	45	+	—
142	there are two	45	+	—
143	to x per	45	+	—
144	about x per	44	+	—
145	in the z	44	+	—
146	is about x	44	+	—
147	is x △	44	+	—
148	it is necessary	44	+	—
149	of the exhaust	44	+	—
150	and in the	43	+	+
151	it should be	43	+	v
152	the flow of	43	+	—
153	to the engine	43	+	—
154	velocity of the	42	+	—

Список 2 *(продолжение)*

i	Триады	F_i	Закон Пуассона	Нормальный закон
155	can be obtained	42	+	—
156	the combustion chamber	42	+	v
157	the exhaust gases	42	+	—
158	the pressure in	42	+	—
159	to x △	42	+	—
160	x z is	42	+	v
161	bolted to the	41	+	—
162	is to be	41	+	—
163	of the y	41	+	—
164	out of the	41	+	+
165	x z for	41	+	+
166	△ for example	40	+	+
167	△ if a	40	+	+
168	△ the two	40	+	—
169	according to the	40	+	—
170	are shown in	40	+	—
171	in the engine	40	+	+
172	more than x	40	+	+
173	of the gas	40	+	—
174	of the water	40	+	—
175	parts of the	40	+	+
176	so as to	40	+	—
177	the turbine is	40	+	—
178	value of x	40	—	—
179	x of the	40	+	—
180	y and y	40	+	—
181	flow through the	39	+	—
182	the cylinder △	39	+	—
183	the engine and	39	+	—
184	the operation of	39	+	—
185	the water level	39	—	—
186	approximately x z	38	+	—
187	at the top	38	+	×
188	figure x shows	38	+	—
189	in the same	38	+	—
190	of the same	38	+	—
191	that it is	38	+	v
192	△ from the	37	+	—
193	△ the pressure	37	+	×
194	△ this type	37	+	+
195	a pressure of	37	+	+
196	half of the	37	+	—
197	in such a	37	+	+
198	is used to	37	+	—
199	may be used	37	+	+
200	mounted on the	37	+	—
201	position of the	37	+	×
202	surface of the	37	+	v
203	temperature of x	37	+	+
204	△ during the	36	+	+
205	△ in some	36	+	—
206	△ the exhaust	36	+	—

i	Триады	F_i	Закон Пуассона	Нормальный закон
207	efficiency of the	36	+	×
208	is of the	36	+	—
209	led to the	36	+	+
210	less than x	36	+	v
211	passes through the	36	+	—
212	the engine room	36	+	—
213	the high temperature	36	+	—
214	which can be	36	+	×
215	is fitted with	35	+	—
216	length of the	35	+	+
217	of the blade	35	—	—
218	of the pump	35	—	—
219	the steam turbine	35	+	—
220	weight of the	35	+	×
221	x or x	35	+	v
222	△ the turbine	34	+	—
223	are of the	34	+	v
224	at the bottom	34	+	+
225	between x and	34	+	—
226	design of the	34	+	—
227	in the fuel	34	+	—
228	it can be	34	+	—
229	movement of the	34	+	v
230	of the two	34	+	v
231	of which is	34	+	v
232	pressure of the	34	+	—
233	x z with	34	+	—
234	△ as a	33	+	—
235	△ it should	33	+	—
236	△ one of	33	+	—
237	can be used	33	+	—
238	of x to	33	+	v
239	top and bottom	33	+	—
240	△ the upper	32	+	v
241	directly to the	32	+	—
242	passing through the	32	+	—
243	pressure and temperature	32	+	v
244	steam at x	32	+	—
245	the turbine and	32	+	—
246	used in the	32	+	—
247	based on the	31	+	—
248	be used △	31	+	—
249	driven by a	31	+	—
250	is in the	31	+	—
251	must be taken	31	+	—
252	of the rotor	31	+	—
253	than x z	31	+	—
254	the boiler is	31	+	—
255	the same time	31	+	—
256	the y △	31	+	—
257	x where x	31	—	—
258	△ the power	30	+	×

i	Триады	F_i	Закон Пуассона	Нормальный закон
259	at the end	30	+	—
260	development of the	30	+	—
261	fitted in the	30	+	—
262	in this case	30	+	—
263	in x △	30	+	—
264	it must be	30	+	—
265	marine diesel engine	30	—	—
266	one or more	30	+	v
267	size of the	30	+	—
268	so that it	30	+	×
269	the same as	30	+	—
270	to the main	30	+	×
271	upper and lower	30	+	—
272	use of a	30	+	—
273	will be seen	30	+	—
274	x is a	30	+	—
275	z and z	30	+	v
4	the bottom of	79	+	—
7	the top of	70	+	—
13	the development of	49	+	—
15	the weight of	44	+	—
24	is known as	38	+	+
35	of moving blades	31	—	—
38	is possible to	30	+	×
39	is provided with	30	+	—
41	the cooling water	30	+	—
35	the lubricating oil	40	+	×
41	the forward end	36	+	—
51	the cruising turbine	32	—	—
18	by means of	24—151	+	+
20	in order to	23—126	+	+
34	per cent of	16—72	+	—
41	the amount of	15—103	+	+
45	the effect of	15—58	+	v
56	a number of	12—97	+	v
64	as follows △	11—66	+	×
78	a row of	10—24	+	—
79	a series of	10—39	+	+
88	the case of	10—85	+	v
113	the purpose of	9—22	+	—
117	△ since the	8	+	v
90	the number of	10	+	v

няется значительное их число (114), соответствие наблюдается в подавляющем большинстве случаев лишь для серий объемом 8 или 16 тыс. словоупотреблений (см. ниже гистограмму 1, рис. 11).

7.4. *Оптимальные условия исследования распределений словоформ и словосочетаний (триад) в научно-техническом тексте*

7.4.1. Условия исследования распределений словоформ относительно нормального закона

Данные основных таблиц (образец см. выше, стр. 80, 89) о соответствии эмпирических распределений словоформ теоретическим законам Пуассона и Гаусса можно обобщенно представить в виде гистограммы 1 (рис. 11). В гистограмме на оси абсцисс нанесены объемы внутрисерийных выборок и соответствующие им объемы выборочных совокупностей, обозначенные латинскими буквами согласно схеме 1 (стр. 53). На оси ординат — порядковые номера единиц от номера 1 до номера последней обследуемой единицы частотного списка. Порядковые номера показывают границы, начиная с которых анализируемые единицы подчиняются законам лишь при указанных объемах внутрисерийных выборок. Так, например, фигура (зона I), образованная ломаной жирной линией, соответствует распределению словоформ по закону Пуассона. Фигура (зона II), образованная пунктирной линией, соответствует распределению словоформ по нормальному закону. Этому закону подчиняются распределения словоформ от номера 1 до номера 22 при всех заданных объемах внутрисерийных выборок и выборочных совокупностей; начиная примерно с номера 22—30, распределения словоформ соответствуют этому закону при длине микровыборки 4 тыс. словоупотреблений и более и т. д.

Дальнейший анализ основных таблиц и гистограммы 1 приводит нас к следующим выводам.

1) В большинстве случаев между порядковым номером (соответственно — абсолютной частотой) словоформы, эмпирическое распределение которой соответствует *нормальному закону*, и объемами внутрисерийных выборок и выборочных совокупностей существует определенная зависимость: если номер первой подчиняющейся закону единицы частотного списка [39] принять за начало отсчета, то с дальнейшим увеличением порядкового номера (уменьшением абсолютной частоты) увеличивается как длина внутрисерийной выборки, так и объем выборочной совокупности.

2) При исследовании распределений словоформ относительно нормального закона вопрос объема внутрисерийных выборок и связанный с ним вопрос объема выборочных совокупностей нельзя решать однозначно. Вероятно, делать это целесообразно двумя путями. Если выборочная совокупность подготовлена, то на основании ее объема можно решить вопрос объема внутрисерийной выборки, и наоборот. И первое и второе зависят от обследуемой зоны частотного списка.

[39] Первая подчиняющаяся данному закону единица не обязательно должна занимать первый порядковый номер в частотном списке (см. гистограмму 1, рис. 11).

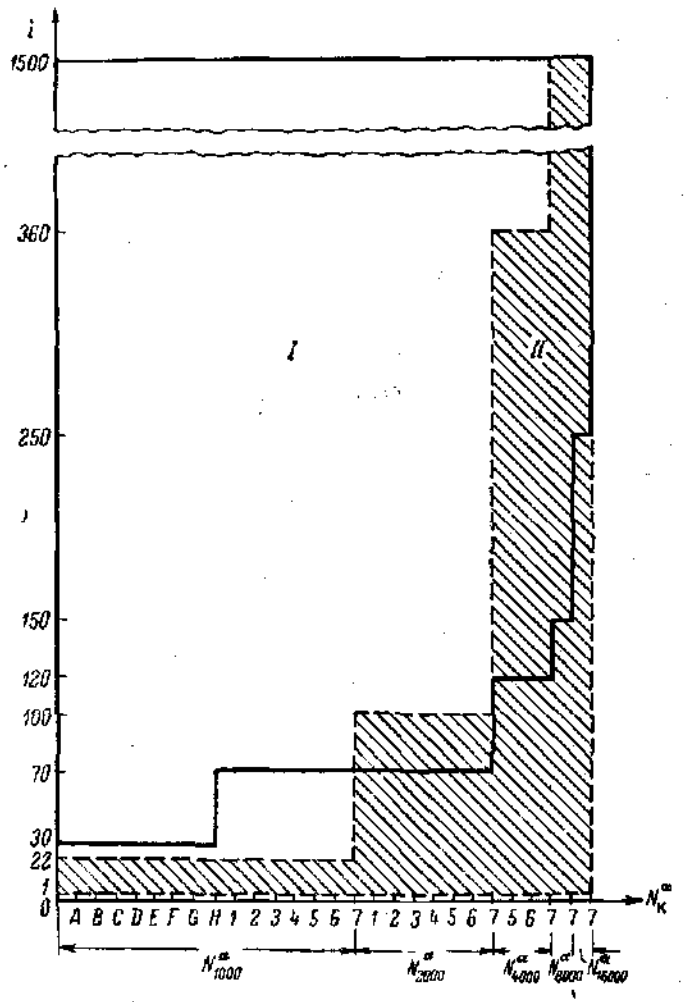


Рис. 11. Гистограмма 1.

Условные обозначения

1 — AB;
2 — CD;
3 — EF;
4 — GH;
5 — ABCD;
6 — EFGH;
7 — A—H.

3) Полученные результаты дают возможность уточнить в виде рекомендаций два общих вывода, сделанных выше в пп. 1,2:

а) при длине общей выборочной совокупности 400 тыс. словоупотреблений для обследования словоформ по нормальному закону целесообразно брать несколько объемов серий, при этом минимальным объемом серии может служить текст длиной в 4 тыс. словоупотреблений;

б) наименьшим объемом промежуточной выборки следует принять текст длиной в 200 тыс. словоупотреблений; такой вывод также согласуется с условием применения критерия χ^2, согласно которому $N_K^a \geqslant 50$ (подробнее см. выше, стр. 52);

в) для обследования словоформ по нормальному закону текст длиной в 400 тыс. словоупотреблений является минимальным объемом общей выборочной совокупности; следовательно, дальнейшее обследование распределений словоформ относительно этого закона целесообразно делать на текстах большей длины — 1 млн словоупотреблений и более.

7.4.2. Условия исследования распределений словоформ относительно закона Пуассона

Основные таблицы и гистограмма 1 показывают, что словоформы от номера 1 до номера 30 с законом Пуассона вообще не согласуются, а начиная примерно с порядкового номера 30 до 70 подчиняются закону лишь при объеме серии 1 тыс. словоупотреблений и объеме выборочной совокупности 50 тыс. словоупотреблений. В зоне порядковых номеров 70—120 соответствие уже наблюдается при всех объемах выборочных совокупностей для N_{1000}^a и N_{2000}^a, а в зоне рангов 120—150 распределения словоформ соответствуют закону Пуассона как при всех предыдущих значениях N_K^a, так и при объеме серии 4 тыс. словоупотреблений; от номера 150 до 250 к указанным выше объемам внутрисерийных выборок добавляется и серия длиной в 8 тыс. словоупотреблений и т. д. Опираясь на эти данные, можно сделать предположение, что при длине внутрисерийной выборки менее 1 тыс. словоупотреблений распределения словоформ в зоне порядковых номеров $i \leqslant 22$ также будут соответствовать закону Пуассона; это предположение подкрепляет тезис о перемещении границ зон действия законов распределений (см. выше, стр. 90).

Частные выводы и рекомендации к методике исследования распределений лингвистических единиц в тексте сводятся к следующему:

1) в качестве минимального объема внутрисерийной выборки при исследовании словоформ зоны $1\,500 \geqslant i \geqslant 30$ относительно закона Пуассона следует принять текст длиной в 1 тыс. словоупотреблений, поэтому 50 тыс. словоупотреблений являются достаточным объемом промежуточной выборки;

2) для получения достоверных данных, а также для сравнения однородности распределений по нескольким вариантам независимых выборок желательно, чтобы общая выборочная совокупность была представлена несколькими промежуточными выборками по 50 тыс. словоупотреблений;

3) заданные нами условия исследования распределений словоформ относительно закона Пуассона являются избыточными.

7.4.3. Условия исследования распределений словоформ относительно логарифмически-нормального закона

Принципиальных особенностей в распределении словоформ по логарифмически-нормальному закону в сравнении с нормальным не выявлено. Как уже отмечалось, логнормальному закону соответствуют распределения 142 из 150 подвергнутых анализу словоформ. В их число входит большинство словоформ, не подчиняющихся ни нормальному закону, ни закону Пуассона. Вероятно, этот закон наиболее тонко фиксирует характер распределения лингвистических единиц текста. Однако здесь же обнаруживается и другая особенность закона, на наш взгляд, связанная с первой. Будучи наиболее чувствительным «фиксатором» колебаний вероятностей, он вместе с тем сильно реагирует на резкие колебания последних, обнаруживая при этом отклонения $\Sigma P'_{F_i''}$ от единицы в бóльшую или меньшую сторону. Такие колебания наблюдаются, как правило, в вариантах малых объемов внутрисерийных выборок: 1 тыс., 2 тыс. словоупотреблений, а с уменьшением i и при 4 тыс. Причину этого надо искать, как говорят математики, в «малом» объеме счета.[40] В связи с этим при обследовании распределений словоформ относительно логнормального закона распределения следует использовать микровыборки длиной не менее 8—10 тыс. словоупотреблений.

7.4.4. Условия исследования распределений трехсловных сочетаний относительно закона Пуассона

Выводы и рекомендации к методике исследования распределений триад в тексте относительно закона Пуассона делаем на основании гистограммы 2 (рис. 12) и списка 2 (стр. 92 и сл.).

На гистограмме зона I есть область соответствия распределений триад закону Пуассона, а зона II — область соответствия нормальному закону в зависимости от порядкового номера и объемов внутрисерийных выборок и выборочных совокупностей текста.

На гистограмме 2 можно обнаружить существование зависимости между порядковыми номерами (соответственно — абсолют-

[40] См.: W. G. Cochran. Some Difficulties in the Statistical Analysis of Replicated Experiments. Empire Journal of Experimental Agriculture, vol. VI, № 22, Oxford University Press, April, 1938, p. 171.

ными частотами) триад, распределения которых согласуются с законом Пуассона, и объемами внутрисерийных выборок и выборочных совокупностей текста. Действительно, распределения триад начальных номеров частотного списка ($i = 1 - 15$) прояв-

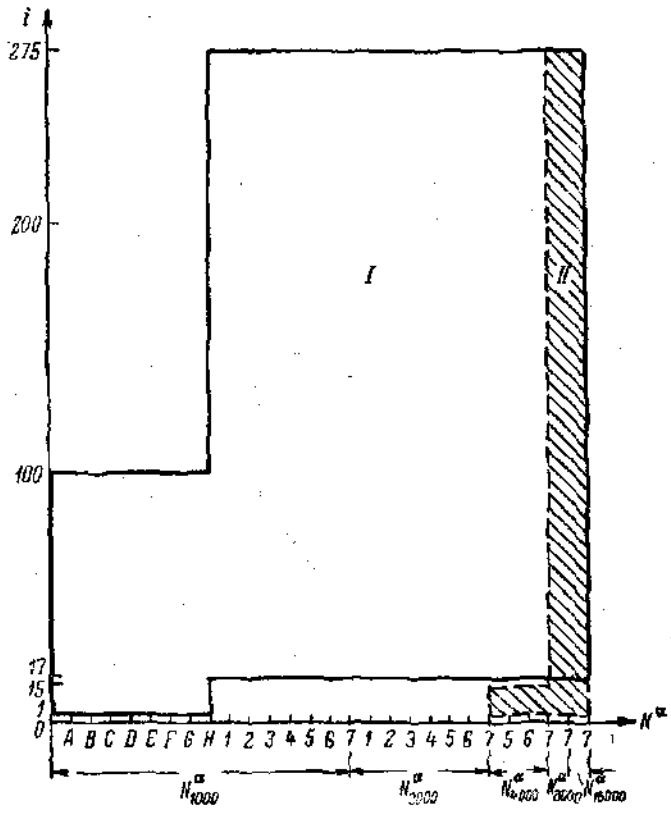


Рис. 12. Гистограмма 2 (условные обозначения те же, что и на рис. 11).

ляют соответствие закону Пуассона только при объеме серии 1 тыс. словоупотреблений и длине промежуточной выборки 50 тыс. словоупотреблений. При порядковом номере от 15 до 100 триады подчиняются закону уже при всех заданных вариантах нормировок текста. С дальнейшим увеличением порядкового номера триады в частотном списке соответствие закону Пуассона наблюдается также при всех заданных вариантах нормировок, но за исключением варианта промежуточной совокупности длиной в 50 тыс. словоупотреблений, составленной из серий объемом 1 тыс. словоупотреблений. На основании данных гистограммы 2 приходим к следующим частным выводам:

1) триады начала частотного списка ($i \leqslant 15$), полученного от обследования текста длиной в 400 тыс. словоупотреблений, могут быть обследованы относительно закона Пуассона на выборке объемом 50 тыс. словоупотреблений при длине серии 1 тыс. словоупотреблений;

2) начиная примерно с порядкового номера 15 минимальным объемом серии также может служить текст длиной в 1 тыс. словоупотреблений, но при объеме выборочной совокупности 100 тыс. словоупотреблений;

3) оптимальные условия исследования распределений триад порядковых номеров $i \geqslant 15$ относительно закона Пуассона могут быть созданы при объеме промежуточной выборочной совокупности 100 тыс. словоупотреблений, составленной из серий длиной в 2 тыс. словоупотреблений; при этом для получения достоверных данных необходимо обследовать несколько промежуточных совокупностей текста по 100 тыс. словоупотреблений;

4) заданные условия исследования распределений триад английского научно-технического подъязыка судовых механизмов относительно закона Пуассона являются избыточными.

7.4.5. Условия исследования распределений трехсловных сочетаний относительно нормального закона

Основные таблицы и гистограмма 2 (рис. 12) позволяют сделать следующие выводы относительно методики исследования распределений триад по нормальному закону:

1) минимальным объемом внутрисерийной выборки при исследовании распределений триад относительно нормального закона нужно принять текст длиной в 16—20 тыс. словоупотреблений;

2) для исследования распределений триад относительно этого закона следует увеличить объем общей выборочной совокупности до 1 млн словоупотреблений и более, при этом желательно иметь данные по нескольким промежуточным совокупностям текста (ср. выше, стр. 101, п. 2).

7.4.6. Общие выводы

В настоящем параграфе мы подробно проанализировали результаты исследования характера распределений 300 словоформ и 300 трехсловных сочетаний английского научно-технического подъязыка судовых механизмов и тем самым рассмотрели условия исследования, при которых эмпирические распределения анализируемых лингвистических единиц соответствуют/не соответствуют трем законам — нормальному, логнормальному и закону Пуассона. Как по словоформам, так и по триадам для каждого закона

были сделаны частные выводы. Общие выводы, к которым мы пришли в результате такого подробного анализа, состоят в следующем.

1) Изменение условий исследования приводит к замене одного теоретического закона (которому соответствуют эмпирические распределения лингвистических единиц) другим, в частности происходит переход от закона Пуассона к нормальному распределению, и наоборот.

2) Поставленные лингвистические задачи (проверка гипотез о распределениях словоформ разных грамматических классов слов и триад разной грамматической природы и в связи с этим попытка определить некоторые объективные признаки терминологичности лингвистических единиц в тексте) могут получить освещение только с точки зрения некоторых статистических условий наблюдения, в которых поставленные цели выполняются. Например, вывод о том, что распределения словоформ некоторого грамматического класса слов подчиняются только данному закону и т. п., может быть сделан лишь относительно определенных условий исследования.

Эти вопросы будут рассмотрены ниже.

§ 8. Результаты исследования эмпирических распределений словоформ относительно нормального закона по грамматическим классам слов

8.1. Полученные нами данные показывают, что при заданных условиях исследования (см. выше, стр. 52 и сл.) и условиях соответствия 1,2 (см. выше, стр. 79 и сл.) однозначное решение о характере распределения словоформ в тексте можно принять относительно наречий (с некоторыми отступлениями), предлогов и частиц, союзов, артиклей и единиц искусственных языков. Их распределения соответствуют нормальному закону (см. табл. 11).

В таблице знаком «+» отмечено количество словоформ, эмпирические распределения которых соответствуют закону, знаком «—» — не соответствующие закону.

Теперь обратимся к результатам исследования распределений словоформ других грамматических классов слов.

8.2. Существительные и гипотеза об объективном признаке «терминологичность/нетерминологичность» лингвистической единицы в тексте

Всего исследованию была подвергнута 121 словоформа имени существительного. При заданных условиях наблюдения распределения 53 словоформы соответствуют нормальному закону, не согласуются с законом 68 словоформ. При дальнейшем анализе полученных данных явно обнаруживаются две особенности.

Таблица 11

Существительное	+	53	121
	—	68	
Прилагательное	+	42	49
	—	7	
Местоимение	+	8	9
	—	1	
Глагол	+	40	70
	—	30	
Наречие	+	24	26
	—	2	
Артикль	+	3	3
	—	—	
Предлог/частица	+	16	16
	—	—	
Союз	+	4	4
	—	—	
Единицы искусственных языков	+	2	2
	—	—	
Итого:		300	300

С одной стороны, распределения словоформ существительного подчиняются, за некоторым исключением, нормальному закону только при объемах внутрисерийных выборок 8 тыс. и/или 16 тыс. словоупотреблений. С другой стороны, совершенно очевидно, что все без исключения именные словоформы, не подчиняющиеся нормальному закону, в научно-техническом тексте имеют терминологические значения,[41] иначе совокупность обследуемых текстов по отношению слов, употребляемых в терминологическом значении, *является неоднородной*.

На основании этих особенностей (ср. также 8.1) формулируем следующую гипотезу: в определенных условиях наблюдения характер распределения лингвистической единицы может стать объективным признаком «терминологичность/

[41] Подробнее см.: К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического текста. . .

нетерминологичность» этой единицы в тексте. В частности, несоответствие эмпирического распределения теоретическому закону свидетельствует о том, что для данной лингвистической единицы в анализируемом тексте характерно терминологическое значение, и наоборот, соответствие закону указывает на ее нетерминологичность.[42]

Всего, как уже отмечалось, обследованию была подвергнута 121 словоформа существительного. При заданных условиях исследования признак «терминологичность» наблюдается у 53 словоформ.

Допустим, что исследование распределений словоформ проводится при всех заданных нормировках текста, кроме объемов серий 8 тыс. и 16 тыс. словоупотреблений. Будем в дальнейшем называть их изменеными условиями исследования (наблюдения). В этих условиях признак «терминологичность» можно проследить у 113 терминологических словоформ имени существительного.

Из проведенного выше анализа следуют два дополнительных вывода.

1) Для достижения аналогичных (поставленных нами) лингвистических целей исследование распределений существительных относительно нормального закона целесообразно проводить при всех заданных вариантах нормировок текста, исключая серии длиной 8 и 16 тыс. словоупотреблений.

2) При исследовании распределений существительных в иных целях отрезки текста длиной в 8—16 тыс. словоупотреблений могут служить в качестве внутрисерийных выборок наименьшего объема. Отсюда вытекает, что текст длиной в 400—500 тыс. словоупотреблений может выступать в качестве выборочной совокупности наименьшего объема, следовательно дальнейшее исследование этого вопроса целесообразно провести на тексте длиной 1 млн словоупотреблений и более.

8.3. Имена прилагательные и гипотеза об объективном признаке «терминологичность / нетерминологичность» лингвистической единицы в тексте

Прилагательные, распределения которых согласуются с нормальным законом, относятся к общеупотребительной лексике. Из 42 единиц исключение может, вероятно, составить лишь одна словоформа automatic, которая соответствует закону при самом низком уровне значимости (об уровнях значимости см. выше, стр. 79). Словоформы, не подчиняющиеся закону, относятся

[42] Выдвинутая гипотеза согласуется с выводом, к которому мы пришли выше (см. стр. 104, п. 2).

к терминологической лексике. Эти данные позволяют сделать следующие выводы:

1) у адъективных словоформ признаки «терминологичность/нетерминологичность» наблюдаются как при заданных, так и в измененных условиях исследования;

2) выводы, сделанные выше относительно методики исследования распределений существительных, имеют силу и для прилагательных.

8.4. Глагольные и адвербиальные словоформы

Выше (стр. 104) было отмечено, что все наречные словоформы подчиняются, за некоторым исключением, нормальному закону. Однако следует добавить, что все они принадлежат к словам общеупотребительной лексики, поэтому здесь следует говорить только о признаке «нетерминологичность».

При анализе результатов исследования распределения глагольных словоформ выявляется сильная тенденция к выполнению двух частей гипотезы.

Эмпирические распределения как глаголов, так и наречий, несмотря на их сравнительно большие частоты, подчиняются нормальному закону при больших объемах (8—16 тыс. словоупотреблений) внутрисерийных выборок и выборочных совокупностей.

8.5. Общие выводы

В § 8 мы провели анализ результатов исследования распределений словоформ разных грамматических классов слов относительно нормального закона. По каждой части речи были сделаны частные выводы. Основные общие выводы по этому параграфу можно сформулировать следующим образом:

Слова, принадлежащие к частям речи:

1) наречия, местоимения, предлоги и частицы, союзы, артикли, а также единицы искусственных языков в английском научно-техническом тексте по судовым механизмам распределяются по нормальному закону; при этом вторая часть выдвинутой гипотезы (признак «нетерминологичность») выполняется регулярно как в заданных, так и в измененных условиях исследования;

2) первая часть выдвинутой гипотезы (признак «терминологичность») последовательно выполняется для всех подвергнутых анализу терминологических словоформ имени существительного в измененных условиях наблюдения;

3) у адъективных словоформ признаки «терминологичность» и «нетерминологичность» наблюдаются как в заданных, так и в измененных условиях исследования;

4) при анализе эмпирических распределений глагольных словоформ выявляется сильная тенденция к выполнению двух частей гипотезы

§ 9. Исследование эмпирических распределений словоформ относительно закона Пуассона по грамматическим классам слов (существительные, прилагательные, глаголы, наречия, служебные слова и единицы искусственных языков)

9.1. Выводы по частям речи и вопрос об объективном признаке «терминологичность» лингвистической единицы в тексте

Результаты исследования этого вопроса представлены в табл. 12.

1) Здесь однозначное решение можно, за некоторым исключением, принять относительно распределений наречий, артиклей, предлогов и частиц, союзов и единиц искусственных языков:

Таблица 12

Результаты исследования распределения словоформ разных грамматических классов слов относительно закона Пуассона при заданных условиях наблюдения

Существительное	+	45	121
	—	76	
Прилагательное	+	41	49
	—	8	
Местоимение	+	4	9
	—	5	
Глагол	+	41	70
	—	29	
Наречие	+	24	26
	—	2	
Артикль	+	1	3
	—	2	
Предлог / частица	+	4	16
	—	12	
Союз	+	1	4
	—	3	
Единицы искусственных языков	+	—	2
	—	2	
Итого:		300	300

а) распределения наречий согласуются с законом Пуассона (аналогичное явление наблюдается и при нормальном законе, см. выше, стр. 104);

б) распределения артиклей, предлогов и частиц, союзов, единиц искусственных языков и буквенных сокращений имеют сильную тенденцию не подчиняться закону Пуассона (это явление противоположно тому, какое мы наблюдали для нормального закона, см. выше, стр. 104).

2) Заданные условия исследования распределений словоформ всех грамматических классов слов являются избыточными: начиная с объема внутрисерийной выборки 1000 словоупотреблений изменение как длины серии, так и объема выборочной совокупности не влияет на конечный результат для большинства словоформ.

3) При заданных условиях исследования признак «терминологичность» наблюдается у 75 из 113 именных словоформ. В измененных условиях указанный признак можно проследить уже для 87 словоформ: для распределений 13 терминологических словоформ условие соответствия 2 (см. выше, стр. 81) в новых условиях исследования не выполняется. К этим словоформам относятся следующие: parts, ports, hand, failure, rods, gearing, contact, indicator, pinion, circuit, thickness, content, tip.

4) Если отказаться от уровня значимости 0.01 (об уровнях значимости см. выше, стр. 79), то признак «терминологичность» будет проявляться в более выраженной форме.

9.2. Объективные показатели терминологичности лингвистической единицы в тексте

В § 8 было выяснено, что в определенных условиях исследования признак «терминологичность» для нормального закона наблюдается по всей группе словоформ имени существительного и прилагательного. Аналогичное явление, но не в такой сильной форме, наблюдается и для закона Пуассона (см. раздел 9.1). В этой ситуации целесообразно, по-видимому, сравнить группы словоформ разных грамматических классов слов, чтобы определить, для каких из них указанный признак выявляется как для нормального закона, так и для закона Пуассона. Анализ таких групп позволяет выделить 95 словоформ существительного, прилагательного и глагола (см. список 3), для которых с точки зрения интуитивной оценки в научно-техническом тексте характерны терминологические значения.

Дальнейшие исследования характера распределения лингвистических единиц в тексте, возможно, позволят дать более точную интерпретацию описанным явлениям. Однако уже наши наблюдения дают право предполагать, что в определенных условиях исследования отсутствие согласования эмпири-

Список 3

i	Существительные	F_i	i	Существительные	F_i
23	steam	2080	205	types	232
24	engine	1947	210	level	226
27	turbine	1703	215	ratio	222
28	pressure	1699	220	output	270
30	water	1372	225	vessel	217
32	fuel	1252	235	parts	209
34	oil	1194	240	stroke	208
35	valve	1168	251	compression	198
37	air	1056	325	generator	159
44	boiler	953	330	stages	159
46	cylinder	902	335	lever	157
51	temperature	815	340	suction	156
52	engines	775	345	wheel	154
54	power	759	368	ports	140
56	type	748	372	vacuum	139
60	gas	714	400	hand	129
61	system	714	455	fialure	112
65	pump	680	459	corrosion	111
67	main	661	467	rods	110
68	heat	660	475	gearing	108
75	piston	610	483	contact	107
78	turbines	597	513	mixture	99
80	speed	590	537	filter	94
85	valves	560	549	crank	92
86	combustion	548	567	timing	89
90	shaft	514	580	indicator	86
96	*end*	472	594	*pinion*	85
102	bearing	425	610	circuit	81
103	blades	425	620	thickness	81
104	operation	425	640	reading	79
111	ship	400	660	cleaning	76
112	gear	398	721	degrees	70
115	pumps	386	890	shafts	57
116	efficiency	384	930	content	54
118	reactor	376	1000	tip	51
122	diesel	367			
125	boilers	358			
126	side	357			
127	velocity	356			
128	conditions	355			
134	steel	342			
135	cent	339			
136	time	333			
137	bearings	332			
140	nozzle	325			
141	tubes	323			
145	maximum	319			
147	machinery	313			
148	top	312			
171	cylinders	269			
186	casing	253			
191	arrangement	246			

i	Прилагательные	F_i
142	marine	322
377	specific	137
490	current	65

i	Глаголы	F_i
94	control	501
276	thrust	183
395	scavenging	130
491	stress	106
770	clutch	66

ческих распределений с теоретическими законами и Пуассона, и Гаусса (нормальным распределением) в большинстве случаев служит показателем терминологичности как словоформы (или группы словоформ), так и словосочетания.

§ 10. О выдвинутых гипотезах

Посмотрим, насколько подтверждаются гипотезы, выдвинутые нами в начале статьи (см. стр. 48).

1) Эмпирические распределения словоформ разных статистических групп частотного списка подчиняются разным законам. Начало списка соответствует исключительно нормальному закону. За этой зоной следует смешанная зона с преобладанием словоформ, подчиняющихся нормальному закону, а затем наблюдается обратное явление с постепенным переходом в зону соответствия распределений словоформ закону Пуассона (подробно см. стр. 88 и сл.).

2) 90% проанализированных словосочетаний в английском научно-техническом тексте по судовым механизмам распределяется по закону Пуассона.

3) Служебные слова (артикли, предлоги, частицы, союзы) и единицы искусственных языков распределяются в английском научно-техническом тексте только по нормальному закону.

4) Исследование различий в распределении служебных и терминологических единиц имеет важные теоретические и практические последствия.

Во-первых, пуассоновское и нормальное распределение служебных слов, а также слов и выражений «стертой» семантики, говорит о том, что текст, на котором проверяется это распределение, является статистически однородным по отношению к этим единицам. Наоборот, для слов и выражений, обладающих ярким терминологическим значением, неподчиняемость этим законам является сигналом к неоднородности текста по отношению к этим лексическим единицам.

Во-вторых, пуассоновское и гауссовское распределения служебных лексических единиц в однородной выборке свидетельствует, что марковские связи этих единиц очень слабы: эти единицы практически выступают в качестве независимых элементов текста. Напротив, отсутствие такого распределения у терминологических слов указывает на их марковские связи, за которыми стоят достаточно сильные семантико-грамматические валентности.

Указанные особенности распределения служебных и терминологических единиц текста выступают в качестве формального распознающего аппарата, который может быть использован для выделения семантически нагруженных (ключевых или доминантных) слов и выражений, опираясь на которые можно осуществлять автоматическое реферирование текста, семантический машинный перевод, диагностику душевных заболеваний и пр.

Для более полного решения вопросов, поставленных в настоящей статье, необходимо фронтально обследовать большие массивы текстов разных подъязыков, с тем чтобы получить частотные списки слов, словосочетаний, а также грамматических связей и уже затем исследовать законы распределений всех (или основной части) единиц частотных списков. Выполнение этих задач возможно лишь в условиях промышленного (т. е. массового) использования электронно-вычислительной техники в лингвистических исследованиях. В связи с этим следует отметить, что с точки зрения автоматической переработки текстов, написанных русским или латинским алфавитом, описанная методика исследования является универсальной. Схемы обеспечивают широкие возможности изменения приемов исследования непосредственно в ходе проведения эксперимента на машине. Алгоритм решения задачи и предусмотренные процедуры обращения к данным, перерабатываемым машиной, обеспечивают тесное взаимодействие и «взаимопонимание» человека и машины. Привлечение ЭВМ для выполнения таких задач позволяет снять все ограничения, связанные при ручном способе работы с поиском оптимальной разбивки и нормировки исследуемого материала, а также с необходимостью выполнения астрономического объема вычислений и приведения результатов на бумаге к форме, удобной для восприятия человеком. Отсюда неизбежно вытекает, что вряд ли целесообразно делать попытки исследовать распределения лингвистических единиц вручную и даже с привлечением счетно-аналитических машин или малых ЭВМ, не имеющих вывода на широкую печать.

* * *

Пользуемся случаем, чтобы выразить самую глубокую благодарность Р. Г. Пиотровскому и Э. Н. Хотяшову за консультации и ценные советы.

Т. Г. Гачечиладзе и Т. П. Цилосани

ОБ ОДНОМ МЕТОДЕ ИЗУЧЕНИЯ СТАТИСТИЧЕСКОЙ СТРУКТУРЫ ТЕКСТА

1. ОБЩАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Предлагаемый метод анализа статистической структуры текста является комбинацией двух известных методов — метода анализа пробелов [1] и метода Фукса [2] и соответствует модели, согласно которой процесс образования любой анализируемой статистической структуры представляется в виде суперпозиции двух процессов — абсолютно случайного и абсолютно детерминированного.

Суть метода заключается в следующем. По некоторому признаку фиксируются пары элементов; элементы, находящиеся между фиксированными, считаются пробелами. Таким образом, текст принимает следующий вид:

$$---[\alpha_1]---[\alpha_2]--[\beta_1][\alpha_3]--[\beta_2]-[\alpha_4][\beta_4]--------[\beta_5]-----[\alpha_5]---$$
$$--[\alpha_6]------[\beta_6]--- \qquad (*)$$

Рассмотрим конкретный случай распределения пробелов между $[\alpha]$ и $[\beta]$. С этой целью подвергаем текст (*) следующему преобразованию:

$$---[\alpha_1]------[\beta_1][\alpha_2]--[\beta_1][\alpha_3]--[\beta_2]\,v_1\,[\alpha_4][\beta_4]\,v_2\,[\alpha_5]----------$$
$$--[\beta_5][\alpha_6]------[\beta_6]--- \qquad (**)$$

Обратив внимание на индексы при $[\alpha]$ и $[\beta]$, легко заметить, каким образом текст (**) получен из (*). Буква v_1 означает, что в (**) опущены символы $[\beta_3]$ и все тире до $[\alpha_4]$ и т. д. Комплекс, состоящий из $[\alpha]$, ближайшего от него справа $[\beta]$ и расположенных между ними тире, условно назовем «словом».

Элементы в тексте не расположены беспорядочно. Всякое отклонение от чисто стохастического распределения указывает на наличие некоторой структуры. С целью выявления статистической

[1] V. H. Ingve, Institute of Radio Engineers Translations, vol. I, pt. 2, № 3, 1956.

[2] Теория передачи сообщений. М., 1957.

структуры мы будем изучать распределение длин «слов» (под длиной подразумевается число тире в «слове»). На длину «слова» влияют соседние «слова» и опущенные символы.

Для математического описания процесса образования таких «слов» используется обобщенная модель Фукса. Эта модель рассмотрена в ряде наших работ[3] (ср. также стр. 118, прим. 4). Приведем окончательную формулу, описывающую вероятностную организацию процесса словообразования:

$$F(i) = \exp[-(\bar{l} - A)] \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) \frac{(\bar{l} - A)^{i-\nu}}{(i-\nu)!} \varphi_\nu(A, \bar{l}, i), \quad (1)$$

где $\bar{l}$ — средняя длина слова, $(\varepsilon_\nu - \varepsilon_{\nu+1})$ — весовые множители, характеризующие вклад абсолютно детерминированного процесса в суммарный; совокупность параметров $\{\varepsilon_\nu\}$ называется лингвистическим спектром процесса словообразования. Этот спектр раскрывает важную психологическую сторону процесса и непосредственно связан с искомой статистической структурой. $A = \sum_{\nu=1}^{\infty} \varepsilon_\nu$ (считаем, что ряд сходится), а

$$\varphi_\nu(A, \bar{l}, i) = \frac{1}{2} \int_{-1}^{+1} (t+1)^{i-\nu} \exp[-(\bar{l} - A)t] dt =$$

$$= \frac{e^{(\bar{l}-A)}}{2(\bar{l}-A)^{i-\nu+1}} \left\{ \Gamma(i-\nu+1) - e^{-2(\bar{l}-A)} \sum_{s=0}^{i-\nu} \binom{i-\nu}{s} \times \right.$$

$$\left. \times [2(\bar{l}-A)]^{i-\nu-s} \Gamma(s+1) \right\}. \quad (2)$$

В последней формуле функция Γ — это эйлеров интеграл второго рода. Нетрудно показать, что имеет место следующее рекуррентное соотношение:

$$\varphi_k(\alpha) = -\frac{2^{k-1} e^{-\alpha}}{\alpha} + \frac{k}{\alpha} \varphi_{k-1}(\alpha).$$

Из формулы (1) распределение Фукса получается при $\varphi_\nu \to 1$. Распределение Фукса описывает процесс образования так называемых небескомпонентных «слов», в то время как обобщенное распределение (1) описывает процесс образования «слов» более общего класса. С помощью множителей $\varphi_\nu(A, \bar{l}, i)$ происходит учет влияния опущенных символов и соседних «слов».

Если среднюю длину «слова» брать из эксперимента, то задача определения статистической структуры печатной информации

[3] Т. Гачечиладзе, Г. Церцвадзе, Г. Чикоидзе, Тр. Инст. электроники, автоматики и телемеханики АН Грузинской ССР, т. II, 1961; Н. Бокучава, Т. Гачечиладзе, Тр. Тбилисского гос. унив., т. 103, 1965.

сводится к нахождению соответствующего ε-спектра. Множество «слов» разбивается на классы (соответственно особенностям их образования), статистические веса которых равны $(\varepsilon_\nu - \varepsilon_{\nu+1})$. Например, «слова», принадлежащие первому классу (статистический вес класса равен $\varepsilon_1 - \varepsilon_2$), характеризуются тем, что в каждом из этих «слов» имеется один ключевой элемент, управляющий их образованием. «Слова», принадлежащие второму классу (статистический вес класса равен $\varepsilon_2 - \varepsilon_3$), содержат два ключевых элемента и т. д.

Номер класса является структурным рангом «слова». При определении ε-спектра $F(i)$ и $\bar{l}$ считаются заданными из эксперимента. Нахождение спектра можно произвести двумя способами.

1) Из эксперимента определяются моменты распределения (1) и приравниваются к теоретическим значениям. Получается сложная бесконечная система трансцендентных уравнений:

$$M_1 = \left.\frac{\partial G(y, \varepsilon_0, \varepsilon_1, \ldots)}{\partial y}\right|_{y=1}$$

$$M_2 = \left.\frac{\partial^2 G(y, \varepsilon_0, \varepsilon_1, \ldots)}{\partial y^2}\right|_{y=1} + \left.\frac{\partial G(y, \varepsilon_0, \varepsilon_1, \ldots)}{\partial y}\right|_{y=1}, \ldots, \quad (3)$$

где производящая функция, соответствующая распределению (1), имеет следующий вид:

$$G(y, \varepsilon_0, \varepsilon_1, \ldots) = \frac{1}{2} \frac{\exp\{2(\bar{l}-A)(y-1)\} - 1}{(\bar{l}-A)(y-1)} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) y^\nu. \quad (4)$$

Моменты, вычисляемые по формуле (4), являются функциями $\bar{l}$ и $\{\varepsilon_\nu\}$. Например, для первых трех моментов имеем:

$$M_1 = \bar{l}, \quad M_2 = \bar{l}^2 + \bar{l} - \left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 - 2\sum_{k=1}^{\infty} \varepsilon_k + 2\sum_{k=1}^{\infty} k\varepsilon_k,$$

$$M_3 = \bar{l}^3 + 3\bar{l}^2 + \bar{l} + 2\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^3 + 3\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 -$$

$$- 2\sum_{k=1}^{\infty} \varepsilon_k - 3\bar{l}\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 - 6\bar{l}\sum_{k=1}^{\infty} \varepsilon_k + 3\sum_{k=1}^{\infty} k^2\varepsilon_k +$$

$$+ \left[6\left(\bar{l} - \sum_{k=1}^{\infty} \varepsilon_k\right) + 1\right] \sum_{k=1}^{\infty} k\varepsilon_k. \quad (4')$$

2) Сравниваются экспериментальные и теоретические значения $F(i)$.

Ниже рассмотрим второй способ и приведем приближенные формулы для нахождения ε-спектра.

Предположим, что спектр таков: $\varepsilon_1 \neq 0, \ldots, \varepsilon_n \neq 0, \varepsilon_{n+1} = \ldots = 0$. Тогда из (1) получим n уравнений

$$F(1) = \exp[-(\hat{l}-A)]\left[(\varepsilon_0-\varepsilon_1)\frac{\hat{l}-A}{1!}\varphi_1 + (\varepsilon_1-\varepsilon_2)\varphi_0\right],$$

$$F(2) = \exp[-(\hat{l}-A)]\left[(\varepsilon_0-\varepsilon_1)\frac{(\hat{l}-A)^2}{2!}\varphi_2 + \right.$$

$$\left. + (\varepsilon_1-\varepsilon_2)\frac{\hat{l}-A}{1!}\varphi_1 + (\varepsilon_2-\varepsilon_3)\varphi_0\right]. \qquad (5)$$

. . . . . .

Если n подобрать таким, что $\sum_{i=0}^{n} F(i) \approx 1$ (с подходящей степенью точности), и если условно считать A известным, то относительно ε-параметров получается система линейных уравнений. Решая ее, получим:

$$\varepsilon_i = \varepsilon_i(A).$$

Подставляя значение $\varepsilon_1(A)$ в уравнение

$$F(0) = \exp[-(\hat{l}-A)][1-\varepsilon_1(A)]\varphi_0, \qquad (6)$$

найдем численное значение A. Далее, подставляя это число в исходную систему, получим числовые значения для ε-спектра.

Этот путь нахождения ε-спектра нам представляется более удобным, так как по существу приходится решать одно сложное трансцендентное уравнение. Заметим, что этот способ является приближенным, но он дает возможность с любой степенью точности приблизиться к истинному значению ε-спектра.

Перепишем систему (5) в более удобной для нас форме:

$$\begin{aligned}
&x_1\varphi_0 = B_1\\
&\tfrac{\alpha}{1!}x_1\varphi_1 + x_2\varphi_0 = B_2\\
&\tfrac{\alpha^2}{2!}x_1\varphi_2 + \tfrac{\alpha}{1!}x_2\varphi_1 + x_3\varphi_0 = B_3\\
&\ldots\ldots\ldots\ldots\\
&\tfrac{\alpha^{n-1}}{(n-1)!}x_1\varphi_{n-1} + \ldots + x_n\varphi_0 = B_n,
\end{aligned} \qquad (7)$$

где $\alpha = \hat{l} - A$, $x_i = \varepsilon_i - \varepsilon_{i+1}$ и $B_i = F(i)e^{\alpha} - \frac{\alpha^i}{i!}\varphi_i x_0$.

Нетрудно убедиться, что решение этой системы таково:

$$x_s = \varphi_0^{s-1}[\varphi_0^{s-1}B_s + (-1)^{s-1}B_1\beta_{s-1}], \qquad (8)$$

где для β_{s-1} имеем следующее рекурентное соотношение:

$$\beta_{s-1} = \frac{\alpha}{1!}\varphi_1\beta_{s-2} - \frac{\alpha^2}{2!}\varphi_2\varphi_0\beta_{s-2} + \sum_{j=3}^{s-2}(-1)^{j+1}\frac{\alpha^j}{j!}\varphi_j\varphi_0^{j-2}\beta_{s-(j+1)} \qquad (9)$$

и

$$\beta_1 = \frac{\alpha}{1!}\varphi_1.$$

Трансцендентное уравнение для определения A примет вид:

$$F(0) = \exp[-(\hat{l}-A)]\left(1 - \sum_{k=1}^{n} x_k\right)\varphi_0. \qquad (10)$$

Интересно рассмотреть следующие предельные случаи: $2\alpha \gg 1$ и $2\alpha \ll 1$.

1) $2\alpha \gg 1$. В этом случае

$$\varphi_k \approx \frac{e^{\alpha}}{2\alpha^{k+1}}\Gamma(k+1).$$

Общее решение системы (7) в этом приближении имеет следующий вид:

$$x_s = 2\alpha[F(s) - F(s-1)], \qquad (11)$$

причем

$$\varepsilon_1 = 2\alpha[F(n) - F(0)]. \qquad (12)$$

В последней формуле $n = i_{\max}$, где $i_{\max}$ соответствует максимуму распределения. Это можно оправдать следующими соображениями. Во-первых, из условия $x_{n+1} = 0$ вытекает, что в нашем приближении $F(n+1) = F(n)$, а это может иметь место как раз вблизи максимума. Во-вторых, при больших α поведение φ_k таково, что оно сглаживает эффект множителя $\frac{\alpha^k}{k!}$.

Кроме того, количество слагаемых в теоретической формуле для $F(i)$ определяется индексом i, и, следовательно, возрастание $F(i)$ при приближении i к $i_{\max}$ будет в основном происходить за счет добавления слагаемых x_i. Отсюда видно, что в нашем приближении ожидаемый спектр должен быть таким: $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{i_{\max}} \neq 0$, а все остальные $\varepsilon_i = 0$. Подставляя значение ε_1 из формулы (12) в уравнение (10), получим:

$$\alpha = \frac{1}{2F(n)}. \qquad (13)$$

2) $2\alpha \ll 1$. В этом случае $\varphi_k \approx \frac{2e^{\alpha}}{\alpha^{k-1}}\Gamma(k+1)$, $(k>0)$ и $\varphi_0 = e^{\alpha}$.

Общее решение системы (7) в этом приближении будет иметь следующий вид:

$$x_s = F(s) - 2\alpha\sum_{k=0}^{s-1} F(k) \qquad (14)$$

и

$$\varepsilon_1 = \sum_{s=1}^{n} x_s = \sum_{k=1}^{n} F(k) - 2\alpha \sum_{k=1}^{n} \sum_{j=0}^{k-1} F(j), \tag{15}$$

а для α получаем:

$$\alpha = \frac{1}{2} \frac{1 - \sum_{k=0}^{n} F(k)}{\sum_{k=0}^{n} \sum_{j=0}^{k-1} F(j)}; \tag{16}$$

n мы можем определить из условия $x_{n+1}=0$, которое приводит к следующему уравнению:

$$F(n+1) = \sum_{k=1}^{n} F(k) \frac{1 - \sum_{k=0}^{n} F(k)}{\sum_{k=1}^{n} \sum_{j=0}^{k-1} F(j)}. \tag{17}$$

Из этого уравнения можно определить n методом проб и ошибок. Все необходимые таблицы для приближенного определения ε-спектров имеются в работе, указанной в примечании.[4]

II. КОНКРЕТНЫЕ МОДЕЛИ

Ниже рассматриваются конкретные примеры использования общей математической модели.

Все численные данные относятся к грузинскому языку.

1. Распределение слогов по словам

Процесс словообразования из слогов относится к классу небесконпонентных. Поэтому $\varphi_k \to 1$. Как показано в нижеприводимой работе,[5] соответствующий спектр таков: $\varepsilon_1=1$, $\varepsilon_2=\varepsilon_3=\ldots=0$ и $\bar{l}=2.542$. Распределение принимает следующий вид:

$$F(i) = \frac{l^{-1.542}\,(1.542)^{i-1}}{(i-1)\,!}.$$

Энтропия этого распределения $S=2.280$ bit.

В табл. 1 приведены теоретические и экспериментальные значения $F(i)$.

[4] Отчет теоретического отдела пробл. лаб. физ. кибернетики Тбилисского гос. унив., т. 6, 1967.

[5] Г. Церцвадзе, Г. Чикоидзе, Т. Гачечиладзе, Сообщ. АН Грузинской ССР, т. XXII, № 6, 1959.

Таблица 1

	1	2	3	4	5	6	7	8	9
Экспериментальные данные	0.238	0.312	0.224	0.146	0.061	0.015	0.003	0.001	0.000
Теоретические данные	0.214	0.331	0.255	0.131	0.051	0.016	0.004	0.001	0.000

2. Распределение звуков по слогам

Так же как и в случае распределения слогов по словам, *процесс слогообразования из звуков принадлежит классу небесконпонентных*. Как выяснилось,[6] хорошее приближение дают первые три члена ε-спектра.

Определение ε-параметров происходит с помощью моментов распределения (1). Экспериментальные значения первых трех моментов таковы:

$$M_1 = 2.328;\ M_2 = 6.074;\ M_3 = 17.670.$$

Используя эти значения, нужные нам уравнения получим из системы (4′):

$$\varepsilon_2^3 + 2\varepsilon_3\varepsilon_2 + (\varepsilon_3^2 - 2\varepsilon_3 - 0.675) = 0$$
$$\varepsilon_2^3 + (3\varepsilon_3 - 4.992)\,\varepsilon_2^2 + (3\varepsilon_3^2 - 12.985\varepsilon_3 + 6)\,\varepsilon_2 + (\varepsilon_3^3 - 7.992\varepsilon_3^2 + 14.985\varepsilon_3 + 3.764) = 0;$$

ε-спектр в данном случае таков:

$$\varepsilon_1 = 1;\quad \varepsilon_2 = 0.836;\quad \varepsilon_3 = 0.204;\quad \varepsilon_4 = \varepsilon_5 = \ldots = 0.$$

Наличие такой структуры слогообразования указывает, что слоги с точки зрения их образования из звуков делятся на отдельные группы. В нашем случае их три.

I. Слоги, состоящие не менее чем из одного звука
II. » » » » » » двух звуков.
III. » » » » » » трех »

Соответствующие статистические веса равны:

для группы I $\rho_1 = \varepsilon_1 - \varepsilon_2 = 0.164$;
» » II $\rho_2 = \varepsilon_2 - \varepsilon_3 = 0.633$;
» » III $\rho_3 = \varepsilon_3 = 0.204$.

[6] См.: Г. Церцвадзе, Г. Чикоидзе, Т. Гачечиладзе. Сообщ. АН Грузинской ССР, т. XXII, № 6, 1959; Th. Gatschetschiladze, G. Tsertsvadze, G. Tshikoidze, Grundlagenstudien aus Kybernetik, B. 3, 1962.

Таким образом, распределение звуков по слогам описывается формулой:

$$F(i) = e^{-0.288}\left[0.164\,\frac{(0.288)^{i-1}}{(i-1)!} + 0.633\,\frac{(0.288)^{i-2}}{(i-2)!} + 0.204\,\frac{(0.288)^{i-3}}{(i-3)!}\right] \quad (***).$$

В табл. 2 приведены результаты, полученные из эксперимента и на основании формулы (***).

Таблица 2

	1	2	3	4	5	6	7
Экспериментальное $F(i)$	0.105	0.553	0.269	0.058	0.012	0.002	0.001
Теоретическое $F(i)$	0.135	0.562	0.325	0.086	0.009	0.001	0.000

Принятие во внимание большего числа членов ε-спектра обеспечит, очевидно, бóльшую точность.

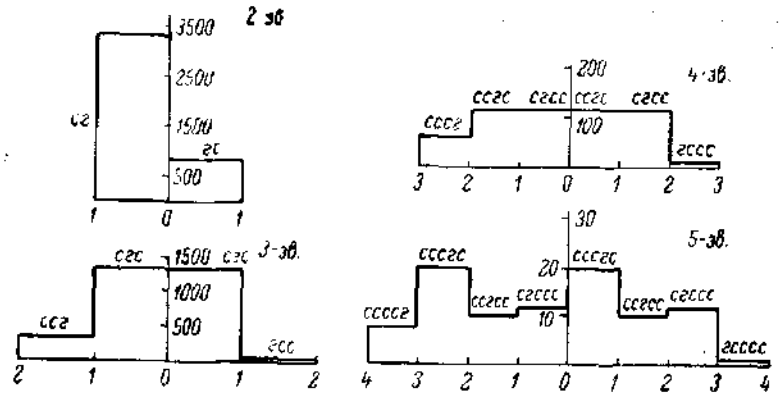


Рис. 1. Гистограммы распределений слогов.

Разбиение слогов на три группы не соответствует известным традиционным классификациям.[7] Ниже приводятся числовые данные (см. табл. 3 и рис. 1), позволяющие распределить реальные слоги по трем указанным группам:

$$\text{гр.}(\rho_1) = 1 \text{ гласн.} + 2 \text{ закр.}$$
$$\text{гр.}(\rho_2) = 2 \text{ откр.} + 3 \text{ откр.}$$
$$\text{гр.}(\rho_3) = 3 \text{ гарм.} + 3 \text{ закр.} + 4 \text{ гарм.}$$
$$\rho_1^{\text{эксп.}} = 0.081 + 0.080 = 0.161$$
$$\rho_2^{\text{эксп.}} = 0.527 + 0.041 = 0.568$$
$$\rho_3^{\text{эксп.}} = 0.008 + 0.193 + 0.009 = 0.210$$

[7] Н. Бокучава, Т. Гачечиладзе, К. Николадзе, Т. Цилосани, Тр. Тбилисского гос. унив., т. 103, 1965.

Таблица 3 (средние значения)

Вид слогов	1 гласн.	2 откр.	2 закр.	3 откр.	3 закр.	3 гарм.	4 откр.
Число слогов	514	3375	535	267	1253	47	57
Статистический вес	0.081	0.527	0.080	0.041	0.193	0.008	0.009

Таблица 3 (*продолжение*)

Вид слогов	4 закр.	4 гарм.	5 откр.	5 закр.	6 откр.	6 закр.
Число слогов	235	55	8	41	20	9
Статистический вес	0.037	0.009	0.001	0.006	0.002	0.001

3. Распределение букв по словам

Распределение букв по словам в грузинском языке изучено в работе Г. Церцвадзе и Т. Гачечиладзе.[8] Как выяснилось, условие $P_l(O) = l^{-\alpha l}$ выполняется довольно точно. В табл. 4 приводятся усредненные данные для отдельных букв.

В табл. 5 приводятся данные о распределении отдельных букв по словам.

Исключение составляют три буквы: ა(a), ი(i) и გ(g), см. табл. 6а.

Для этих букв необходимо пользоваться обобщенным распределением. В качестве примера рассмотрим букву ი (i). В этом случае $l = 0.612$; следовательно, α можно считать малой величиной. Воспользовавшись уравнением (17), находим $n \approx 2$; ε-спектр получается таким:

$$\varepsilon_0 = 1,\ \varepsilon_1 = 0.480,\ \varepsilon_2 = 0.092,\ \varepsilon_3 = \varepsilon_4 = \ldots = 0.$$

Соответствующее распределение записывается в виде:

$$F(i) = 0.961\left[0.520\,\frac{(0.040)^i}{i!}\varphi_i + 0.387\,\frac{(0.040)^{i-1}}{(i-1)!}\varphi_{i-1} + 0.092\,\frac{(0.040)^{i-2}}{(i-2)!}\varphi_{i-2}\right].$$

В таблице (*) приведены данные опыта и результаты вычислений *по последней формуле*.

[8] Тр. Инст. электроники, автоматики и телемеханики АН Грузинской ССР, т. 1, 1960.

Таблица 4

i / r	0	1	2	3	4	5
ა (a)	0.300	0.428	0.240	0.055	0.007	0.001
ბ (b)	0.881	0.109	0.010			
გ (g)	0.858	0.132	0.010			
დ (d)	0.729	0.233	0.037	0.002		
ე (e)	0.625	0.301	0.065	0.008	0.001	
ვ (v)	0.806	0.178	0.016			
ზ (z)	0.970	0.029	0.002			
თ (t)	0.843	0.135	0.022			
ი (i)	0.502	0.391	0.098	0.009		
კ (k)	0.935	0.061	0.001			
ლ (l)	0.822	0.168	0.009			
მ (m)	0.741	0.229	0.029	0.001		
ნ (n)	0.808	0.174	0.016	0.001		
ო (o)	0.772	0.201	0.024	0.001		
პ (p)	0.988	0.011	0.001			
ჟ (ž)	0.998	0.002				
რ (r)	0.664	0.327	0.009			
ს (s)	0.757	0.222	0.020	0.001		
ტ (t)	0.959	0.039	0.002			
უ (u)	0.885	0.104	0.010			
ფ (p)	0.978	0.021	0.001			
ქ (k)	0.945	0.050	0.002			
ღ (y)	0.973	0.026				
ყ (q)	0.954	0.045	0.001			
შ (š)	0.915	0.082	0.003			
ჩ (č)	0.961	0.038	0.001			
ც (c)	0.919	0.079	0.002			
ძ (ʒ)	0.981	0.018	0.002			
წ (ę)	0.968	0.031	0.002			
ჭ (č)	0.981	0.019				
ხ (x)	0.902	0.093	0.005			
ჯ (ǯ)	0.982	0.017				
ჰ (h)	0.983	0.017				

Таблица 5

r	Среднее по авторам	$r-\bar{r}$	p_1	M_2
ა	1.049	0.350	0.469	0.818
ბ	0.133	0.878	0	0.133
გ	0.153	0.861	0.151	0.314
დ	0.312	0.733	0.108	0.300
ე	0.457	0.631	0.152	0.434
ვ	0.210	0.811	0.105	0.199
ზ	0.032	0.970	0.064	0.028
თ	0.179	0.835	0.230	0.125
ი	0.612	0.543	0.348	0.491
კ	0.067	0.932	0.044	0.065
ლ	0.188	0.827	0.122	0.173
მ	0.290	0.748	0.142	0.270
ნ	0.210	0.811	0.069	0.205
ო	0.253	0.779	0.098	0.243
პ	0.013	0.990	0	0.013
ჟ	0.002	1.000	0	0.002
რ	0.345	0.705	0.316	0.245
ს	0.265	0.771	0.156	0.240
ტ	0.042	0.961	0.094	0.012
უ	0.126	0.878	0	0.126
ფ	0.022	0.980	0	0.022
ქ	0.054	0.951	0	0.054
ღ	0.027	0.970	0.010	0.027
ყ	0.046	0.951	0.026	0.046
შ	0.088	0.914	0.048	0.085
ჩ	0.040	0.961	0	0.040
ც	0.083	0.923	0.045	0.081
ძ	0.022	0.980	0	0.022
წ	0.034	0.970	0	0.034
ჭ	0.020	0.980	0	0.020
ხ	0.103	0.905	0.024	0.102
ჯ	0.017	0.980	0.020	0.017
ჰ	0.017	0.980	0.017	0.016

Таблица 6

Теоретическое распределение букв по словам

i \ l	0	1	2	3	4	5
ა	0.350	0.367	0.193	0.067	0.018	0.004
ბ	0.878	0.114	0.007			
გ	0.861	0.131	0.010			
დ	0.733	0.229	0.036	0.004		
ე	0.631	0.289	0.066	0.010	0.001	
ვ	0.811	0.170	0.018	0.001		
ზ	0.970	0.031				
თ	0.835	0.149	0.013	0.001		
ი	0.543	0.333	0.102	0.021	0.003	
კ	0.932	0.062	0.002			
ლ	0.827	0.155	0.015	0.001		
მ	0.748	0.217	0.031	0.003		
ნ	0.811	0.170	0.018	0.001		
ო	0.779	0.197	0.025	0.002		
პ	0.990	0.013				
ჟ		0.002				
რ	0.705	0.243	0.042	0.005		
ს	0.771	0.204	0.027	0.002		
ტ	0.961	0.041	0.001			
უ	0.878	0.111	0.007			
ფ	0.980	0.022				
ქ	0.951	0.051	0.001			
ღ	0.970	0.026				
ყ	0.951	0.044	0.001			
შ	0.914	0.080	0.003			
ჩ	0.961	0.038	0.001			
ც	0.923	0.077	0.003			
ძ	0.980	0.021				
წ	0.970	0.033	0.001			
ჭ	0.980	0.019				
ხ	0.905	0.093	0.005			
ჯ	0.980	0.017				
ჰ	0.980	0.016				

Таблица 6а

	ა (a)	ი (i)	ღ (r)
$P_i(0)$	0.300	0.502	0.664
$l^{-\bar{l}}$	0.350	0.543	0.705

Таблица*

$F(i)$	i				
	0	1	2	3	4
Экспериментальное	0.503	0.391	0.098	0.009	0.000
Теоретическое	0.520	0.429	0.165	0.080	0.000

Аналогичные результаты получаются и для ა(a) и ღ(r).

Несмотря на это исключение, мы пользуемся формулой (1).

Сначала был рассмотрен простой случай: $\varepsilon_0=1$, $\varepsilon_1=\varepsilon_2=\ldots=0$. Имеем:

$$W_l(i)=\frac{\bar{l}^i}{i!}e^{-\bar{l}}. \qquad (18)$$

Результаты, полученные с помощью этой формулы, приведены в табл. 6.

Рассмотренный выше ε-спектр не обеспечивает хорошего приближения. Это очевидно из следующих формул:

$$\varepsilon_1(\bar{M}_2)\sqrt{\bar{l}-\bar{M}_2} \qquad =0.152,$$

где

$$\bar{l}=\frac{1}{33}\sum_{\lambda=1}^{33}\sum_i iW_\lambda(i) \qquad =0.167,$$

$$\bar{M}_2=\frac{1}{33}\sum_{\lambda=1}^{33}\sum_i(i-\bar{l}_\lambda)^2W_\lambda(i) \qquad =0.144.$$

Из табл. 4, 5 и 6 очевидно, что для некоторых букв, в частности для ა, ბ, ე, ჟ, ვ, ი, ო, ს, ჰ, ღ и ხ, нужно вычислить пара-

Таблица 7

	১ (a)	৪ (g)	৫ (d)	৯ (e)	৩ (v)
a	—0.08	0.72	0.71	0.70	0.71
b	—0.06	0	0	—0.02	—0.01
c	0	0	0	0.02	0
M_3	0.541	0.148	0.273	0.400	0.209
ε_1	0.97	0.00	0.10	0.24	0.16
ε_2	0.28	0.00	0	0.02	0.01

Таблица 7 (*продолжение*)

	০ (i)	৭ (o)	৯ (m)	৯ (r)	৳ (s)
a	0.61	0.70	0.70	0.63	0.71
b	—0.01	0	0	—0.01	0
c	0.01	0	0	—0.01	0
M_3	0.311	0.146	0.231	0.110	0.198
ε_1	0.38	0.14	0.14	0.34	0.14
ε_2	0.02	0	0	0.01	0

метры ε_1 и ε_2. Для их вычисления имеем следующую систему уравнений:

$$\varepsilon_1^2 + 2\varepsilon_2\varepsilon_1 + (\varepsilon_2^2 + 2\varepsilon_2 - \bar{l} + M_2) = 0,$$

$$\varepsilon_1^3 + \frac{3}{2}(2\varepsilon_2 - 1)\varepsilon_1^2 + 3(\varepsilon_2^2 - 2\varepsilon_2)\varepsilon_1 +$$

$$+ \frac{1}{2}(\bar{l} + 2\varepsilon_2^3 - 9\varepsilon_2^2 + 6\varepsilon_2 - M_3) = 0,$$

где M_2 и M_3 — второй и третий моменты распределения (18). Значения M_2 приводятся в табл. 5, а M_3 — в табл. 7.

Для ε_2 имеем следующее уравнение:

$$\varepsilon_2^3 + a\varepsilon_2^2 + b\varepsilon_2 + c = 0,$$

где

$$a = 0.72 - 0.90\bar{l} + 0.90M_1,$$

$$b = 0.24\bar{l}^2 - 0.48\bar{l}M_2 + 0.24M_2^2 - 0.24\bar{l} + 0.36M_2 - 0.12M_3,$$

$$c = -0.02\bar{l}^3 + 0.06\bar{l}^2M_2 - 0.06\bar{l}M_2^2 + 0.02M_2^3 + 0.04M_2^2 +$$
$$+ 0.005M_3^2 + 0.02\bar{l}^2 - 0.03M_2M_3 - 0.06\bar{l}M_2 + 0.02\bar{l}M_3.$$

Численные значения этих коэффициентов и соответствующие решения ε_1, ε_2 для букв ১, ৪, ৫, ৯, ৩, ০, ৭, ৯, ৯, ৳ приводятся в табл. 7.

Для остальных букв простой ε-спектр представляет хорошее приближение. Для вышеперечисленных десяти букв распределение записывается в следующем виде:

$$W_l(i) = l^{-(\bar{l}_l - \varepsilon_1^l - \varepsilon_2^l)}\left[\frac{(\bar{l}_l - \varepsilon_1^l - \varepsilon_2^l)^i}{i!}(1 - \varepsilon_1^l) + \right.$$

$$\left. + \frac{(\bar{l}_l - \varepsilon_1^l - \varepsilon_2^l)^{i-1}}{(i-1)!}(\varepsilon_1^l - \varepsilon_2^l) + \frac{(\bar{l}_l - \varepsilon_1^l - \varepsilon_2^l)^{i-2}}{(i-2)!}\varepsilon_2^l\right].$$

Результаты вычислений приведены в табл. 8.

Таблица 8

l	i						
	0	1	2	3	4	5	6
১	0.30	0.44	0.21	0.06	0.02	0.00	0.00
৪	0.858	0.132	0.010	0.00	0.00	0.00	0.00
৫	0.729	0.23	0.04	0.00	0.00	0.00	0.00
৯	0.623	0.30	0.06	0.00	0.00	0.00	0.00
৩	0.806	0.17	0.02	0.00	0.00	0.00	0.00
০	0.50	0.40	0.097	0.008	0.00	0.00	0.00
৯	0.82	0.17	0.009	0.00	0.00	0.00	0.00
৯	0.74	0.23	0.03	0.00	0.00	0.00	0.00
৯	0.66	0.33	0.01	0.00	0.00	0.00	0.00
৳	0.77	0.21	0.03	0.00	0.00	0.00	0.00

Таблица 9

	১	২	৪	৫	৯	৩	৭	৮	০	৯	৯
S_l	0.57	0.18	0.20	0.30	0.38	0.23	0.06	0.21	0.44	0.10	0.22
$\bar{P}_l$	0.190	0.023	0.028	0.057	0.083	0.038	0.006	0.032	0.111	0.012	0.034
	৯	৬	৯	৩	৯	৯	৳	৩	৯	৯	৯
S_l	0.29	0.23	0.26	0.02	0	0.33	0.27	0.08	0.18	0.04	0.08
$\bar{P}_l$	0.053	0.038	0.046	0.002	0	0.063	0.048	0.008	0.023	0.004	0.010
	৯	৯	৩	৯	৳	৯	৯	৯	৳	৯	৯
S_l	0.06	0.08	0.13	0.08	0.13	0.04	0.06	0.04	0.13	0.04	0.04
$\bar{P}_l$	0.005	0.008	0.016	0.007	0.015	0.004	0.006	0.004	0.019	0.003	0.003

Энтропия отдельных букв:

$$S_l = -\sum_i W_l(i) \log W_l(i).$$

Полученные результаты приведены в табл. 9.
Полная энтропия:

$$S = -\sum_l \sum_i W_l(i) \log W_l(i) = 18.31 \text{ bit}.$$

Энтропию, соответствующую формуле (18), можно рассматривать как функцию l_l. Соответствующая кривая приведена на рис. 2. Там же показано большинство экспериментальных точек.

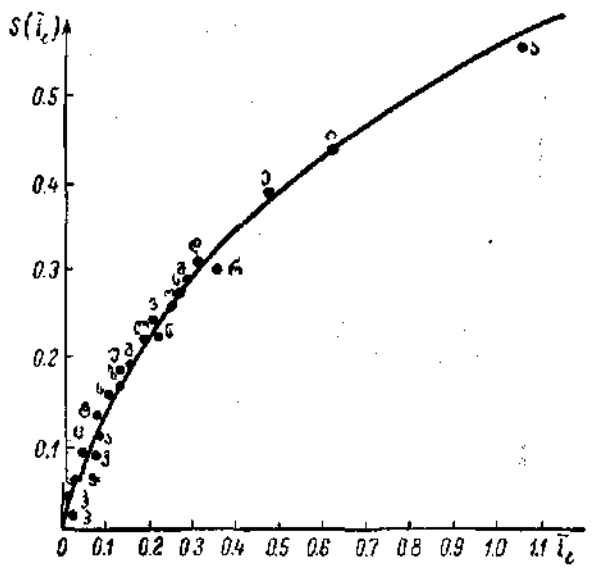


Рис. 2. Энтропия грузинских букв.

Средняя относительная частота появления буквы l (без учета промежутка между словами) будет:

$$\bar{P}_l = \frac{i_l}{\sum_l i_l}.$$

Для грузинского языка $\sum_l i_l = 5.507$.

Значения $\bar{P}_l$ приведены в табл. 9.

4. Статистические связи между существительным (С) и глаголом (Г)

В работе, указанной в прим. 3, были подсчитаны распределения пробелов между всевозможными парами двух частей речи [С] и [Г]:

$$[С]-[С],\ [С]-[Г],\ [Г]-[Г],\ [Г]-[С].$$

В качестве элементов [С] принимались слова, являющиеся существительными; субстантивированные прилагательные и причастия, а также любые другие слова, выступающие в роли существительного, к классу [С] нами не относились. Такой же чисто формальный морфологический подход осуществлялся и при подсчете [Г]: в этот класс включены только личные формы глагола.

Ниже приводятся соответствующие усредненные экспериментальные данные и теоретические результаты, полученные на основе общей модели.

В табл. 10 для каждой комбинации отведены 2 строки. Первая содержит экспериментальные данные. Вторая содержит численные результаты попытки описать экспериментальные данные формулой, где некоторые из $\varepsilon \neq 0$.

а) [С]—[С]. Средняя вероятность встречи [С]: $\bar{P}_С = 0.258$. Средняя длина пробела [С]—[С]: $l_{СС} = 2.415$; ε-спектр: $\varepsilon_1 = \varepsilon_2 = \ldots = 0$. Распределение:

$$F(i) = e^{-2.415} \frac{(2.415)^i}{i!} \varphi_0(0;\ 0.258;\ i).$$

б) [Г]—[Г]. Средняя вероятность встречи [Г]: $\bar{P}_Г = 0.180$. Средняя длина пробела [Г]—[Г]: $l_{ГГ} = 3.965$; ε-спектр: $\varepsilon_1 = 1$, $\varepsilon_2 = 0.5$; $\varepsilon_3 = \varepsilon_4 = \ldots = 0$. Распределение:

$$F(i) = e^{-2.465}\left\{\frac{0.5(2.465)^{i-1}}{(i-1)!}\varphi_1(1.5;\ 3.965;\ i) + \right.$$

$$\left. + 0.5\frac{(2.465)^{i-2}}{(i-2)!}\varphi_2(1.5;\ 3.965;\ i)\right\}.$$

в) [С]—[Г]. Средняя длина пробела [С]—[Г]: $l_{СГ} = 2.322$; ε-спектр: $\varepsilon_1 = \varepsilon_2 = \ldots = 0$. Распределение:

$$F(i) = e^{-2.322}\frac{(2.322)^i}{i!}.$$

г) [Г]—[С]. Средняя длина пробела [Г]—[С]: $l_{ГС} = 2.323$; ε-спектр: $\varepsilon_1 = \varepsilon_2 = \ldots = 0$. Распределение:

$$F(i) = e^{-2.323}\frac{(2.323)^i}{i!}.$$

Таблица 10

		Длина пробела														
		0	1	2	3	4	5	6	7	8	10	11	12	13	14	15
[С] — [С]	P_{CC}	0.200	0.199	0.174	0.135	0.087	0.052	0.041	0.028	0.016	0.009					
	$P^{теор}_{CC}$	0.205	0.197	0.178	0.146	0.108	0.071	0.042	0.021	0.009						
[Т] — [Т]	P_{TT}	0.003	0.102	0.190	0.171	0.147	0.099	0.078	0.064	0.034	0.015	0.014	0.007	0.009	0.005	0.006
	$P^{теор}_{TT}$	0	0.051	0.152	0.225	0.223	0.165	0.098	0.048	0.020						
[С] — [Т]	P_{CT}	0.234	0.198	0.145	0.197	0.087	0.059	0.043	0.029	0.012	0.010					
	$P^{теор}_{CT}$	0.213	0.204	0.181	0.146	0.106	0.068	0.039	0.020							
[Т] — [С]	P_{TC}	0.219	0.238	0.157	0.106	0.070	0.055	0.040	0.030	0.016	0.016					
	$P^{теор}_{TC}$	0.213	0.204	0.181	0.144	0.106	0.068	0.039	0.020							

5. Статистические связи между отдельными синтаксическими элементами

С целью установления статистических зависимостей, существующих между отдельными членами предложения, были обработаны тексты грузинских авторов XIX—XX вв.[9]

Анализируемые тексты были представлены в закодированном виде. Используемые обозначения таковы: a — сказуемое, b — подлежащее, c — дополнение, d — обстоятельство, p — определение. Остальные синтаксические элементы во всех случаях рассматривались в качестве пробелов. Таким образом, текст приобретал следующий вид:

$$da - d - cep - bda - \cdot - c - ba - - ca \text{ и т. д.} \qquad (\cdot\ \cdot)$$

Точками, как и в реальном тексте, друг от друга отделялись отдельные предложения.

Далее по отдельным предложениям подсчитывались относительные частоты появления различных пар элементов (например: $[a - a]$, $[a - b]$, $[a - c]$ и т. д.) с учетом длины интервалов между ними.

Под длиной интервала между двумя фиксированными синтаксическими элементами подразумевается число других элементов, находящихся между ними. Например, в первом предложении $(\cdot\ \cdot)$ длина интервала между a и b (для конструкции $[a - b]$) равна 7.

Исследуемые тексты были обработаны в следующей последовательности: 22, 50, 80 и 102 страницы каждого автора. Максимальная длина интервала в наших расчетах не превышала 6.

Для проверки однородности и достаточности объема текстов подсчитывались последовательности частичных сумм энтропии по формуле

$$S_r = \frac{1}{r} \sum_{k=1}^{r} S_k.$$

Изучение этих последовательностей частичных сумм показало, что тексты в достаточной степени однородны.

Результаты анализа приведены в табл. 11, где применяются следующие обозначения: S_N — энтропия, соответствующая объему текста в N страниц; P_i — относительная частота конструкции; i — длина интервала; $\bar{l}$ — среднее значение по распределению.

[9] См.: Т. Цидосанп, Тр. Тбилисского гос. унив., серия кибернетики, т. II, 1968.

Таблица 11

P_i	i									
	0	1	2	3	4	5	6	Σ	ī	S_N стр
(*a* — *a*) ср	0.048	0.197	0.260	0.197	0.121	0.075	0.039	0.938	2.403	0.740
(*a* — *b*) ср	0.437	0.263	0.123	0.074	0.038	0.025	0.012	0.972	1.079	0.613
(*a* — *c*) ср	0.298	0.279	0.177	0.107	0.050	0.027	0.018	0.948	1.397	0.680
(*a* — *d*) ср	0.432	0.232	0.141	0.070	0.051	0.031	0.014	0.973	1.170	0.641
(*a* — *p*) ср	0.456	0.260	0.145	0.069	0.030	0.018	0.011	0.994	1.040	0.605
(*b* — *b*) ср	0.046	0.312	0.216	0.146	0.105	0.060	0.025	0.919	2.061	0.684
(*b* — *a*) ср	0.346	0.283	0.146	0.083	0.049	0.035	0.021	0.965	1.321	0.676
(*b* — *c*) ср	0.259	0.232	0.177	0.121	0.071	0.045	0.028	0.933	1.626	0.728
(*b* — *d*) ср	0.417	0.225	0.148	0.096	0.055	0.025	0.015	0.981	1.244	0.653
(*b* — *p*) ср	0.322	0.217	0.168	0.101	0.075	0.040	0.021	0.944	1.482	0.707
(*c* — *c*) ср	0.187	0.267	0.197	0.134	0.079	0.054	0.022	0.940	1.781	0.736
(*c* — *a*) ср	0.602	0.174	0.099	0.053	0.027	0.017	0.010	0.982	0.783	0.524
(*c* — *b*) ср	0.243	0.304	0.198	0.105	0.046	0.034	0.021	0.951	1.495	0.689
(*c* — *d*) ср	0.372	0.267	0.160	0.093	0.046	0.031	0.020	0.989	1.325	0.673
(*c* — *p*) ср	0.304	0.282	0.169	0.108	0.052	0.034	0.018	0.967	1.430	0.692
(*d* — *d*) ср	0.266	0.267	0.190	0.112	0.066	0.046	0.020	0.967	1.597	0.719
(*d* — *a*) ср	0.511	0.193	0.110	0.068	0.042	0.025	0.014	0.963	0.992	0.594
(*d* — *b*) ср	0.306	0.308	0.176	0.097	0.046	0.025	0.014	0.972	1.344	0.673
(*d* — *c*) ср	0.239	0.291	0.195	0.113	0.072	0.038	0.023	0.971	1.636	0.721
(*d* — *p*) ср	0.340	0.268	0.161	0.098	0.058	0.032	0.015	0.972	1.366	0.683
(*p* — *p*) ср	0.318	0.249	0.159	0.103	0.061	0.038	0.025	0.953	1.460	0.703
(*p* — *a*) ср	0.168	0.352	0.216	0.133	0.067	0.035	0.018	0.989	1.734	0.705
(*p* — *b*) ср	0.580	0.138	0.112	0.060	0.045	0.022	0.012	0.969	0.905	0.551
(*p* — *c*) ср	0.561	0.161	0.107	0.069	0.040	0.025	0.014	0.977	0.951	0.573
(*p* — *d*) ср	0.276	0.261	0.167	0.113	0.074	0.036	0.018	0.942	1.518	0.702

	S_{21}	$\frac{1}{2}(S_{21}+S_{50})$	$\frac{1}{3}(S_{21}+S_{50}+S_{80})$	$\frac{1}{4}(S_{21}+S_{50}+S_{80}+S_{100})$
(*a* — *a*)	0.740	0.740	0.741	0.740
(*a* — *b*)	0.617	0.575	0.601	0.613
(*a* — *c*)	0.682	0.671	0.675	0.680
(*a* — *d*)	0.687	0.642	0.641	0.641
(*a* — *p*)	0.638	0.649	0.617	0.605
(*b* — *b*)	0.557	0.645	0.668	0.684
(*b* — *a*)	0.675	0.669	0.673	0.676
(*b* — *c*)	0.722	0.719	0.727	0.728
(*b* — *d*)	0.718	0.685	0.672	0.653
(*b* — *p*)	0.696	0.706	0.707	0.707
(*c* — *c*)	0.739	0.728	0.734	0.736
(*c* — *a*)	0.515	0.526	0.523	0.524
(*c* — *b*)	0.618	0.659	0.678	0.689
(*c* — *d*)	0.722	0.703	0.687	0.673
(*c* — *p*)	0.671	0.694	0.693	0.692
(*d* — *d*)	0.702	0.702	0.713	0.719
(*d* — *a*)	0.574	0.604	0.594	0.594
(*d* — *b*)	0.660	0.675	0.672	0.673

Таблица 11 *(продолжение)*

	S_{21}	$\frac{1}{2}(S_{21}+S_{50})$	$\frac{1}{3}(S_{21}+S_{50}+S_{80})$	$\frac{1}{4}(S_{21}+S_{50}+S_{80}+S_{100})$
(*d* — *c*)	0.705	0.711	0.715	0.721
(*d* — *p*)	0.634	0.669	0.683	0.683
(*p* — *p*)	0.683	0.712	0.707	0.703
(*p* — *a*)	0.706	0.715	0.703	0.705
(*p* — *b*)	0.528	0.583	0 561	0.551
(*p* — *c*)	0.564	0.596	0.579	0.573
(*p* — *d*)	0.676	0.691	0.699	0.702

П. Б. Невельский и М. Д. Розенбаум

УГАДЫВАНИЕ ПРОФЕССИОНАЛЬНОГО ТЕКСТА СПЕЦИАЛИСТАМИ И НЕСПЕЦИАЛИСТАМИ

В опытах по угадыванию незнакомого текста информант, даже обладающий самым высоким лингвистическим чутьем и языковой культурой, не может дать идеального в полном смысле этого слова предсказания, которое во многом зависит от его индивидуальных особенностей. Для того чтобы улучшить результаты угадывания и приблизить их к результатам идеального предсказания, выработан ряд приемов лингвистического, психологического и математического порядка: [1] специальный подбор испытуемого, угадывание небольших объемов текста, использование вспомогательного словарно-статистического аппарата, специальная вероятностно-лингвистическая коррекция результатов угадывания, коллективное угадывание и др.

Однако мера контекстной обусловленности [2] как количество информации, которую испытуемый может извлечь в среднем из одной буквы незнакомого текста, одновременно оценивает и статистическую информацию, связанную с сочетаемостью лингвистических единиц, и прагматическую информацию, характеризующую опыт, и другие индивидуальные особенности угадчика.

Цель приближения угадывания к идеальному предсказанию как раз и заключается в том, чтобы отделить в ходе эксперимента статистическую информацию текста от прагматического информационного «шума».[3]

Но то, что для объективных информационных измерений языка является шумом, выступает как полезная информация, если мы хотим оценить субъективную энтропию и субъективную избыточность какого-то специального текста для определенного испытуемого или определенной группы испытуемых.

Дело в том, что один и тот же текст может содержать разное количество информации для субъекта в зависимости от степени владения языком, прошлого опыта и целого ряда психологических факторов.[4]

[1] К. Шеннон. Работы по теории информации и кибернетике. Русск. пер. М., 1963, стр. 669—686; Р. Г. Пиотровский. Информационные измерения языка. Л., 1968, стр. 44—47.

[2] См.: Р. Г. Пиотровский, ук. соч., стр. 64—65.

[3] Там же, стр. 102.

I

Задача одного из двух проведенных экспериментов заключалась в сопоставлении субъективных оценок энтропии и избыточности специального делового текста (относящегося к очень узкому профессиональному подъязыку), получаемых при угадывании этого текста специалистами и неспециалистами в данной области, а также специалистами, в равной мере знакомыми с подъязыком текста, но в разной мере знакомыми с самим текстом, который им приходится использовать в процессе своей трудовой деятельности.

Таким образом, помимо информационных измерений неизвестного текста была сделана попытка произвести информационные измерения текста, в какой-то мере известного.

В качестве специального делового подъязыка был взят язык документации, проходящей и обрабатывающейся в финансовом отделе одного из рудоуправлений Донецкой области. Текст объемом в 1100 букв был взят из «Журнала-ордера по затратам на производство». Журнал этот представляет собой развернутый формуляр, ежедневно, из месяца в месяц, заполняемый цифровой информацией. Он содержит большое число рядов и колонок, наименования которых, отпечатанные типографским шрифтом, остаются неизменными в течение нескольких лет; меняется же только цифровая информация. Эти наименования и были взяты в качестве текста для угадывания.

Первые 50 букв этого текста (из 1100) предъявлялись испытуемым до опыта по угадыванию; следующие 50 букв они угадывали для тренировки, и результаты этих опытов не учитывались. Собственно эксперимент заключался в последовательном угадывании следующей тысячи букв. После каждого угадывания, которое фиксировалось, испытуемому сообщалось, какой была буква

[4] См.: П. Б. Невельский и Р. Г. Пиотровский. О применении метода угадывания лингвистического текста в психологии. Материалы III всесоюзного съезда Общ. психологов СССР, т. I, М., 1968, стр. 303—305; А. В. Напалков. Алгоритмический анализ сложных форм работы мозга. Переработка информации человеком, М., 1966, стр. 56—65 (XVIII Международный психологический конгресс. Симпозиум 18); П. Б. Невельский. Субъективная энтропия текста как ненадежность угадывания. Пробл. языкознания. Доклады и сообщения советских ученых на X международном конгрессе лингвистов. (Бухарест, 28.VIII—2.IX.1967). М., 1967, стр. 193—197.

в тексте, и он приступал к следующему угадыванию, и так далее до тысячной буквы. Таким образом, измерялась средняя условная энтропия 101—1100 порядка.

Алфавит букв, включая пробел, был равен 32 (*е* и *ё*, *ь* и *ъ* не различались); алфавит ответов был равен 33, так как испытуемый мог отказаться от ответа, и такой отказ считался буквой. Ниже приводится часть предъявлявшегося текста:

С КРЕДИТА СЧЕТОВ В ДЕБЕТ СЧЕТОВ С ДО [с-до, сальдо] НА НАЧ МЕСЯЦА ДОБЫЧА СЕВЕР Р К [р-к, рудник] ЮЖН Р К ДАЛЬНИЙ Р К ВСКРЫША ДОФ ИТОГО ЖДЦ ПЕРЕВОЗКА ПОГРУЗКА ЖДЦ ДАЛЬНИЙ УСЛУГИ ШПАЛОПРОПИТКА ПАРОКОТЕЛЬНАЯ ВОДА УМЯГЧЕННАЯ ВОДА СЫРАЯ ЖД КРАНЫ ОТВАЛЬНОЕ Х ВО [х-во, хозяйство] ГАЗВОДА ЦРМЦ КИП ЭЦ ЭЛЕКТРОПОДСТАНЦИЯ СВЯЗЬ КИСЛОРОДНАЯ ЦКДМ РАБОТА ОСНОВ ЦЕХОВ ПАРОКОТ СМУ КОМПРЕССОРА [именительн. падеж множественн. числа] ЦКАТ ДРОБЛЕНИЕ КОНВЕЙЕР СКЛАД ЗАПАС ПО СБОРУ ЛОМА.

Как легко заметить, текст этот не является вполне связным. Встречается много обособленных слов, которые по своей информационной нагрузке, видимо, ближе к словам вне текста. Информативность здесь значительно увеличена за счет всевозможных сокращений слов. Если «перевести» этот текст на обычный русский язык, то он увеличился бы в 2—3 раза. Для неспециалистов в данной области текст этот вообще очень мало понятен.

Опыты проводились по методике, разработанной одним из авторов настоящей статьи.[5] Субъективная энтропия текста рассматривалась здесь как ненадежность угадывания, как ненадежность передачи информации с входа на выход, от буквы текста к ответу испытуемого. Ненадежность передачи информации — это условная энтропия входа, когда выход известен, условная энтропия текста, когда известен ответ испытуемого. Другими словами, это мера той неопределенности, которая остается в тексте после угадывания, или то количество информации, которое испытуемый не в состоянии извлечь из текста, используя знание предшествующей последовательности букв этого текста и весь свой прошлый опыт, все свои знания:

$$H_{суб} = H_y(x) = H(x, y) - H(y), \tag{1}$$

где $H_y(x)$ — ненадежность угадывания, $H(x, y)$ — совместная энтропия входа и выхода, текста и угадывания (диграммы «буква текста (x_i) — буква в ответе испытуемого (y_j)»), $H(y)$ — энтропия выхода (угадывания).

[5] П. Б. Невельский, ук. соч.

Так как энтропия угадывания

$$H(y) = -\sum p(y_j) \log_2 p(y_j), \tag{2}$$

где $p(y_j)$ — наблюдаемая вероятность, т. е. относительная частота или частость появления j-той буквы в ответе испытуемого, а суммирование ведется по всем буквам алфавита ответов, и совместная энтропия текста и угадывания

$$H(x_i, y_j) = -\sum p(x_i, y_j) \log_2 p(x_i, y_j), \tag{3}$$

где $p(x_i, y_j)$ — совместная наблюдаемая вероятность появления диграммы «буква текста (x_i) — буква в ответе испытуемого (y_j)», и суммирование ведется по всем возможным диграммам, то

$$H_{суб} = -\sum p(x_i, y_j) \log_2 p(x_i, y_j) + \sum p(y_j) \log_2 p(y_j). \tag{4}$$

Измерялся еще один информационный параметр — количество переданной информации с входа на выход, от текста к ответу испытуемого

$$T(x, y) = H(x) + H(y) - H(x, y). \tag{5}$$

Здесь $H(x)$ — энтропия на входе, безусловная энтропия текста первого порядка, H_1, вычисленная по наблюдаемым вероятностям появления букв в предъявляемом для угадывания тексте:

$$H(x) = -\sum p(x_i) \log_2 p(x_i), \tag{6}$$

где $p(x_1)$ — наблюдаемая вероятность (т. е. частость или относительная частота) i-той буквы текста.

Соотношения рассмотренных информационных параметров представлены на рисунке (см. рисунок). $T(x, y)$ — это количество информации, которое испытуемый может извлечь из текста, используя знание непосредственно предшествующей части текста и свой прошлый опыт. Это не что иное, как предельная контекстная обусловленность, которая одновременно учитывает и статистическую информацию, связанную со знанием угадчиком сочетаемости лингвистических единиц, и прагматическую информацию, характеризующую его прошлый опыт.

$$T(x, y) = H(x) - H_{суб}, \tag{7}$$

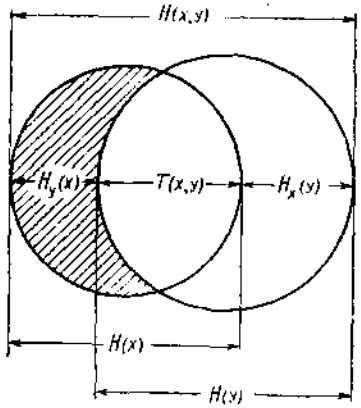
или, другими словами,

$$T(x, y) = I_1 - I_\infty \tag{8}$$

Переданная, или, точнее, извлеченная из текста, информация $T(x, y)$ несколько меньше предельной контекстной обусловленности в такой форме, как она обычно применяется:

$$K_\infty = I_0 - I_\infty.^6 \qquad (9)$$

Различия эти связаны с уменьшением K_∞ на величину $I_0 - I_1$. Дело в том, что информационный анализ,[7] который лежит в основе нашей методики угадывания,[8] предполагает на входе безусловную энтропию не нулевого, а первого порядка, т. е. энтропию алфавита с учетом вероятностей появления каждого символа этого алфавита.



Соотношения информационных параметров.

Для оценки рассмотренных информационных параметров составляется матрица, в которую заносятся частоты совместного появления букв на входе и на выходе, в тексте и в ответе (табл. 1).

Краевые суммы внизу по горизонтали дают частоты букв в тексте, а краевые суммы справа по вертикали указывают частоты букв в ответах испытуемого. Если угадывался текст длиной $N = 1000$ букв или 10 испытуемых угадывают по 1000 букв и $N = 10\,000$, то частоты легко превращаются в наблюдаемые вероятности (относительные частоты). При этом их просто нужно считать тысячными или десятитысячными долями.

Табл. 2 представляет матрицу вероятностей. Различаются эти матрицы в частности тем, что общая сумма в нижнем правом углу в одном случае равна N — числу угадываний (или букв в тексте, если опыты проводились с одним испытуемым), а в другом — единице. Если вычисления производятся по приведенным выше формулам и с применением таблицы значений $-p \log_2 p$,[9] нужно пользоваться матрицей вероятностей.

При обработке данных следует учитывать, что наблюдаемые нами вероятности отличаются от своих истинных значений и приближаются к ним только в случае бесконечно большого числа опытов (N); поэтому вычисленные нами оценки энтропии всегда

[6] См.: Р. Г. Пиотровский. Информационные измерения языка, стр. 65.

[7] См.: W. R. Garner and H. W. Hake. The Amount of Information in Absolute Judgements. Psychological Review, vol. 58, 1951, pp. 446—459.

[8] См.: П. Б. Невельский. Субъективная энтропия текста как ненадежность угадывания.

[9] А. В. Зубов. Таблица значений $p \log_2 p$. СР, стр. 231—258.

Таблица 1

Матрица частот

	x_1	x_2	x_3	...	x_i	...	x_k	$n(y)$
y_1	n_{11}	n_{21}	n_{31}	...	n_{i1}	...	n_{k1}	$n(y_1)$
y_2	n_{12}	n_{22}	n_{32}	...	n_{i2}	...	n_{k2}	$n(y_2)$
y_3	n_{13}	n_{23}	n_{33}	...	n_{i3}	...	n_{k3}	$n(y_3)$
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
y_j	n_{1j}	n_{2j}	n_{3j}	...	n_{ij}	...	n_{kj}	$n(y_j)$
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
y_m	n_{1m}	n_{2m}	n_{3m}	...	n_{im}	...	n_{km}	$n(y_m)$
$n(x)$	$n(x_1)$	$n(x_2)$	$n(x_3)$	...	$n(x_i)$	...	$n(x_k)$	N

Таблица 2

Матрица вероятностей

	x_1	x_2	x_3	...	x_i	...	x_k	$p(y)$
y_1	p_{11}	p_{21}	p_{31}	...	p_{i1}	...	p_{k1}	$p(y_1)$
y_2	p_{12}	p_{22}	p_{32}	...	p_{i2}	...	p_{k2}	$p(y_2)$
y_3	p_{13}	p_{23}	p_{33}	...	p_{i3}	...	p_{k3}	$p(y_3)$
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
y_j	p_{1j}	p_{2j}	p_{3j}	...	p_{ij}	...	p_{kj}	$p(y_j)$
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
.	.	.	.	...	.	...	.	.
y_m	p_{1m}	p_{2m}	p_{3m}	...	p_{im}	...	p_{km}	$p(y_m)$
$p(x)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	...	$p(x_i)$	...	$p(x_k)$	1.00

ниже своего истинного значения и нуждаются в поправке, связанной с величиной выборки (N) и с распределением небольших величин (n, p) на большое число категорий. Такая поправка была предложена Миллером и Мэдоу.[10] Она соответствует отношению числа степеней свободы к величине 2 ($\log_e 2$) $N \simeq 1.3863\ N$.

В опытах участвовали 20 испытуемых, представляющих четыре разнородные группы.

В п е р в у ю группу, из двух человек, входили специалисты (бухгалтеры рудоуправления) со средним образованием и большим стажем работы по специальности, изо дня в день работающие с «Журналом» и вписывающие цифровую информацию в графы этого журнала, текст из которого они угадывали; другими словами, это носители специального подъязыка, хорошо знакомые с текстом.

В т о р а я группа, из четырех человек, отличалась тем, что специалисты здесь готовили цифровые данные для указанного «Журнала» и тоже были знакомы с текстом, но в значительно меньшей степени.

В т р е т ь е й группе, тоже из четырех человек, специалисты совсем не были знакомы с угадываемым текстом, но так же хорошо, как и другие, были знакомы со специальным подъязыком этого текста.

Ч е т в е р т а я группа, из десяти человек, состояла из пионерских работников со средним образованием (учащихся педагогических курсов при Харьковском ГПИ), не знакомых ни со специальным подъязыком, ни с текстом. Такой состав испытуемых давал нам возможность, с одной стороны, сравнить субъективную энтропию и избыточность специального текста для носителей данного подъязыка, в разной мере знакомых с текстом и не знакомых с ним вовсе; с другой стороны, мы могли сравнить эти параметры для неспециалистов и специалистов (носителей специального подъязыка) в случае, когда текст для тех и других был неизвестным.

Опыты проводились индивидуально, хотя методика позволяет проводить их коллективно, одновременно с группой испытуемых (в этом случае нужно только следить, чтобы все испытуемые записали свой ответ до того, как им будет сообщена буква, которую они угадывали). Опыты по угадыванию 1000 букв продолжались около двух часов, и на принятие одного решения о продолжении текста испытуемый затрачивал в среднем около 7 секунд. При этом он, конечно, мог использовать только предшествующую часть текста и свой прошлый опыт.

Результаты опытов представлены в табл. 3. Здесь не указан параметр $H(x)=I_1$ (6), значение которого для данного текста составляло 4.41 дв. ед. Это больше обычно принимаемого для русских текстов значения $I_1=4.22$, потому что значение $H(x)$ в нашем эксперименте представляло энтропию первого порядка не для начальных букв слова, а для текста в целом. Все значения количества переданной при угадывании информации, $T(x, y)=I_1-I_\infty$, в этой таблице меньше контекстной обусловленности, $K_\infty=I_0-I_\infty$, на 0.59 дв. ед.

Предельная энтропия на букву одного и того же незнакомого профессионального текста составляла для специалистов 1.2 дв. ед., а для неспециалистов — 2.9, избыточность соответственно — 76 и 42%, а количество информации, извлеченное в среднем из одной неизвестной буквы текста, — 3.2 и 1.5 дв. ед. Отрывки протоколов опытов, проведенных по угадыванию неизвестного текста специалистом и неспециалистом, приведены в табл. 4 и 5.

[10] G. A. M i l l e r. Note on the Bias of Information Estimates. In: H. Q u a s t l e r (Ed.). Information Theory in Psychology. Glencoe (Ill), 1955, pp. 95—100.

Таблица 3

Результаты опытов по угадыванию специального текста

Группы испытуемых	Испытуемые	Энтропия текста в дв. ед. $H_{\text{суб}}$	Избыточность в % R	Переданная информация $T(x, y)$	Ранг r
I. Специалисты, хорошо знакомые с текстом	1) Х-ов	0.49	90.2	3.92	20
	2) П-ва	0.58	88.4	3.83	19
	среднее	0.53	89.4	3.88	19.5
II. Специалисты, менее знакомые с текстом	1) Ш-ва	0.82	83.6	3.59	18
	2) З-ко	0.87	82.6	3.54	17
	3) Д-на	0.88	82.4	3.53	16
	4) Ф-ин	0.97	80.6	3.44	15
	среднее	0.89	82.2	3.52	16.5
III. Специалисты, не знакомые с текстом	1) И-ов	1.16	76.8	3.25	14
	2) Я-ко	1.18	76.4	3.23	13
	3) Б-ва	1.20	76.0	3.21	12
	4) О-юк	1.22	75.6	3.19	11
	среднее	1.19	76.2	3.22	12.5
IV. Неспециалисты	1) М-пк	2.08	58.4	2.33	10
	2) В-на	2.80	44.0	1.61	9
	3) Г-ко	2.81	43.8	1.60	8
	4) О-ва	2.88	42.4	1.53	7
	5) М-ко	2.90	42.0	1.51	6
	6) О-ра	2.94	41.2	1.47	5
	7) С-ко	3.00	40.0	1.41	4
	8) К-ва	3.01	39.8	1.40	3
	9) Д-ко	3.16	36.8	1.25	2
	10) Ш-да	3.18	36.4	1.23	1
	среднее	2.88	42.4	1.53	5.5

Таблица 4

Часть протокола опытов по угадыванию специального текста
(испытуемый Я-ко, специалист, не знакомый с текстом)

Шаги текста	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620
Текст	**о**	б	щ	е	**р**	у	д	н	щ	ч	н	△	р	а	с	х	о	д	ы	△
Ответы	**з**	б	щ	е	**з**	у	д	н	и	ч	н	△	р	а	с	х	о	д	ы	△
Шаги текста	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640
Текст	**и**	т	о	г	о	△	з	**а**	△	**м**	**е**	**е**	**и**	**к**	△	**и**	**т**	**о**	**г**	**о**
Ответы	—	т	о	г	о	△	з	**а**	△	**м**	**е**	**о**	**п**	**ц**	△	—	**т**	**о**	**г**	**о**
Шаги текста	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660
Текст	△	**е**	△	**н**	**а**	**ч**	△	**г**	**о**	**д**	**а**	△	**е**	**а**	**л**	**ь**	**д**	**о**	△	**н**
Ответы	△	**е**	△	**н**	**а**	**ч**	△	**г**	**о**	**д**	**а**	△	**е**	**а**	**л**	**ь**	**д**	**о**	△	**н**
Шаги текста	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680
Текст	**я**	△	**н**	**а**	**ч**	**а**	**л**	**о**	△	**м**	**е**	**е**	**и**	**ц**	**а**	△	**в**	**е**	**п**	**о**
Ответы	**а**	△	**н**	**а**	**ч**	**а**	**л**	**о**	△	**м**	**е**	**е**	**я**	**ц**	**а**	△	**в**	**е**	**п**	**о**
Шаги текста	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700
Текст	**м**	**о**	**г**	**а**	**т**	**а**	**м**	**а**	**т**	**е**	**р**	**и**	**а**	**л**	**и**	△	**т**	**а**	**р**	**а**
Ответы	**м**	**а**	**г**	**а**	**т**	**а**	**м**	**а**	**т**	**е**	**р**	**и**	**а**	**л**	**и**	△	**т**	**а**	**р**	**а**

Таблица 5

Часть протокола опытов по угадыванию специального текста
(испытуемая О-ва, неспециалист)

Шаги текста	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620
Текст	**о**	—	**л**	**е**	**р**	у	д	н	**и**	**ч**	н	**а**	**р**	а	е	х	**о**	д	**ы**	△
Ответы	—	**б**	**л**	**а**	—	у	д	н	**а**	**к**	н	**а**	△	—	—	—	—	д	△	△
Шаги текста	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640
Текст	**н**	**т**	△	о	о	△	**з**	**а**	△	**м**	**е**	**е**	**л**	**ц**	△	**н**	**т**	о	**г**	**о**
Ответы	—	о	△	г	о	△	—	—	—	**г**	**е**	△	**я**	**ц**	△	**г**	—	о	**г**	**о**
Шаги текста	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660
Текст	△	**е**	△	**н**	**а**	**ч**	△	**г**	**о**	△	△	△	**е**	**а**	**л**	**ь**	**д**	**о**	△	**н**
Ответы	△	—	△	**м**	**о**	**м**	△	—	△	△	—	△	—	**м**	—	—	**н**	—	△	—
Шаги текста	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680
Текст	**а**	△	**н**	**а**	**ч**	**а**	**л**	**о**	△	**м**	**е**	**е**	**я**	**ц**	**а**	△	**в**	**е**	**п**	**о**
Ответы	**а**	△	**н**	**а**	**ч**	**а**	**т**	**ь**	△	—	**е**	**а**	**я**	**ц**	**а**	△	**г**	—	—	—
Шаги текста	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700
Текст	**м**	**е**	**г**	**а**	**т**	**а**	**м**	**а**	**т**	**е**	**р**	**и**	**а**	**л**	**ы**	**е**	**т**	**а**	**р**	**а**
Ответы	—	**н**	**г**	**а**	**т**	**а**	—	**е**	**т**	**е**	**р**	**и**	**а**	**л**	**ы**	**е**	—	—	—	**н**

Часто́ты правильных ответов, соответствующих буквам текста, в матрицах частот располагаются по диагонали n_{11}, n_{22}, n_{33} . . .; частоты отказов, обозначенных в таблицах прочерком (—), располагаются в нижней строке матриц, а ошибочные ответы, выделенные шрифтом, — в остальных клетках матриц.

Значимость полученных различий при сравнении четырех групп испытуемых проверялась по критерию Крускала и Уоллиса,[11] $p < 0.001$.

Опыты показали, что длительная ежедневная работа с одним и тем же текстом в течение многих месяцев и даже лет не приводит к его буквальному запоминанию (в противном случае субъективная энтропия текста равнялась бы нулю). Текст и в этом случае содержал достаточно много неопределенности — около 500 дв. ед. на 1000 букв. С точки зрения советских специалистов по психологии [12] это объясняется тем, что в основное содержание цели деятельности работника входил не этот лингвистический текст (и во всяком случае не его написание), а та изменяющаяся информация, которая относилась к отдельным смысловым кускам этого текста.

Полученные результаты дают основание полагать, что информационные измерения неизвестного специального текста могут служить показателями профессионального опыта и что информационные измерения знакомого смыслового текста могут служить показателями степени его запоминания.

II

Еще один эксперимент был проведен с двумя другими профессиональными «текстами». Первый относился к химическому, второй — к финансовому подъязыку. Тексты эти отличались тем, что были составлены из отдельных терминов, т. е. попросту являлись списками слов (табл. 6, 7). К сожалению, мы не имели возможности отобрать слова с учетом их частотности, но слова в обоих списках были общеизвестными.

Таблица 6

Химические термины

реактив	*глицерин*
водород	*нафталин*
азот	*окисление*
раствор	*щелочь*
аммиак	*ацетон*
сероводород	*анализ*
кислород	*фильтрование*
бензол	*кристаллизация*
хлороформ	*выпаривание*
эфир	*разбавление*

Таблица 7

Финансовые термины

отпуск	*ведомость*
счет	*накладная*
рапорт	*разнарядка*
аккредитив	*требование*
подотчет	*договор*
расчет	*шифр*
кредит	*аванс*
ордер	*баланс*
фонд	*чек*
поставщик	*заказ*

[11] Ж. Мост. Статистические предвидения и решения на предприятии. М., 1966, стр. 229.

[12] См.: П. И. Зинченко. Непроизвольное запоминание. М., 1961.

Термины в наших «текстах», синтаксически не связанные с другими терминами, являются, конечно, словами вне контекста; вместе с тем они связаны лексически, так как относятся к специальному подъязыку, и логически, поскольку входят в объем понятия «химический термин» или в объем понятия «финансовый термин». Кроме того, здесь имеются и морфологические ограничения, так как все слова являются именами существительными в именительном падеже единственного числа.

Оба списка предъявлялись для угадывания десяти химикам и десяти финансовым работникам. В качестве химиков выступала группа из десяти студентов V курса Харьковского педагогического института, получающих квалификацию преподавателей биологии и химии. Их можно считать специалистами с высшим образованием, но не имеющими практического опыта работы по специальности. Кроме того, нужно отметить, что химия для них — это лишь вторая специальность.

В качестве финансовых работников в опытах принимала участие группа из десяти бухгалтеров со средним образованием и многолетним профессиональным стажем.

Методика опытов в этом эксперименте отличалась тем, что угадывание начиналось с первой буквы. Из полученной средней оценки энтропии на одну букву легко можно было вычислить этропию всего списка терминов и в среднем на одно слово.

Первый список химических терминов содержал 180 букв, второй, финансовых терминов, — 152.

Мы здесь пользовались только матрицами частот, куда заносили данные опытов, проведенных со всеми десятью испытуемыми каждой группы, и N при этом равнялось 1800 или 1520.

В этом случае, когда N не равняется 1000 или 10 000, пользоваться формулами (2), (3), (6) и таблицами значений $-p \log_2 p$ весьма затруднительно. Поэтому вместо указанных формул мы использовали производные от них формулы, приведенные У. Дж. Мак-Джиллом: [13]

$$H(x) = \log_2 N - \frac{1}{N} \sum n(x_i) \log_2 n(x_i), \qquad (10)$$

$$H(y) = \log_2 N - \frac{1}{N} \sum n(y_j) \log_2 n(y_j), \qquad (11)$$

$$H(x, y) = \log_2 N - \frac{1}{N} \sum n(x_i, y_j) \log_2 n(x_i, y_j), \qquad (12)$$

где n — часто́ты, а N — общее число отдельных опытов (угадываний). Для обработки данных здесь нужна таблица значений $n \log_2 n$. Такая таблица при $1 \leqslant n \leqslant 1000$ была составлена

[13] См.: W. I. Mc Gill. Multivariate Information Transmission. Psychometrica, vol. 19, 1954, pp. 97—116.

в Харьковском университете на ЭВМ «Урал-2». Большего объема таблица приведена в книге Ф. Эттнива.[14] Можно также пользоваться таблицей значений $n \log_e n$,[15] но при этом нужно иметь в виду, что результаты будут получены не в двоичных, а в натуральных единицах информации, которые в 1.44269... раза больше двоичных.

Опыты по угадыванию двух списков проводились один за другим с каждым испытуемым отдельно. Каждый опыт занимал 35—40 минут, т. е. в среднем около 7 секунд на принятие одного решения.

Результаты опытов представлены в табл. 8. Небольшая длина списков не позволила оценить количество информации на входе, $H(x)$ (6), (10), и вместе с этим количество переданной с входа на выход информации, $T(x, y)$ (5). Поэтому вместо $T(x, y)$ в таблице приведена контекстная обусловленность K_∞(9). Что касается энтропии на выходе (энтропии угадывания), $H(y)$, то здесь число угадываний, которое составляло 1520 и 1800, давало возможность применять формулу (11) и формулу (12) для оценки $H_{суб}$, а также поправку на величину выборки по Миллеру и Мэдоу.[16]

Все информационные параметры, полученные при угадывании профессиональных терминов, оказались лучше у специалистов в данной области, чем у неспециалистов; химические термины — более избыточными и содержащими меньше информации для химиков, финансовые — для бухгалтеров (табл. 8).

На полученные результаты оказала влияние различная средняя длина слов в списках. Как известно, более длинные слова содержат обычно больше информации на слово и меньше информации на букву. Как видно из табл. 8, разность между энтропией на слово в двух списках у химиков, для которых профессиональные термины были длиннее, оказалась в два раза меньше, чем у бухгалтеров, для которых профессиональные термины были короче. В то же время разность между энтропией на букву (а также между контекстной обусловленностью и избыточностью) у химиков, наоборот, была в три раза больше, чем у бухгалтеров.

Сравнение различий всех информационных параметров, полученных при угадывании химиками и бухгалтерами химических и финансовых терминов, позволяет предположить, что в полученных данных отражается больший профессиональный опыт бухгалтеров по сравнению с химиками.

Отсюда следует, что специальное обучение и профессиональная деятельность уменьшают количество субъективной информации, содержащейся в профессиональных терминах, и увеличивают

[14] F. Attneave. Application of Information Theory to Psychology: A Summary of Basis concepts, Methods and Results. New York, Holt, 1959.
[15] См.: С. Кульбак. Теория информации и статистика. М., 1967.
[16] См.: G. A. Miller, ук. соч.

Таблица 8

Результаты опытов по угадыванию профессиональных терминов

Показатели	Группа химиков		Группа бухгалтеров		Разность между показателями профессиональных и непрофессиональных терминов		Разность между показателями у специалистов и неспециалистов	
	химические термины	финансовые термины	химические термины	финансовые термины	у химиков	у бухгалтеров	химические термины	финансовые термины
1) Длина слова в буквах	9.00	7.60	9.00	7.60	1.40	—1.40	0	0
2) Энтропия на букву $H_{суб}$ в дв. ед.	1.84	2.45	2.16	1.96	—0.61	—0.20	—0.32	—0.49
3) Избыточность R в процентах	63.2	51.0	56.8	60.8	12.2	4.0	6.4	9.8
4) Энтропия на слово в дв. ед.	16.6	18.6	19.4	14.9	—2.0	—4.5	—2.8	—3.7
5) Контекстная обусловленность K_{∞} в дв. ед.	3.16	2.55	2.84	3.04	0.61	0.20	0.32	0.49

их субъективную избыточность так, что эти параметры могут служить показателями профессионального опыта.

Наши надежды на то, что в результате экспериментов по угадыванию специального текста специалистами и неспециалистами в данной области можно будет как-то разделить меру контекстной обусловленности на статистическую и прагматическую информацию, не оправдались.

Контекстная обусловленность K_{∞} и информация, переданная от текста к ответу угадчика, $T(x, y)$ отражают в этих экспериментах статистическую и прагматическую информацию, не отделяя одну от другой. Эти два вида информации по-прежнему выступают слитно.

Однако, если в экспериментах, где принимаются все меры для приближения угадывания к идеальному предсказанию, доля прагматической информации как компонента контекстной обусловленности сводится к минимуму, то в экспериментах, где такие меры не принимаются вовсе, наоборот, контекстная обусловленность содержит минимум статистической информации, так как полученные данные о количестве извлеченной из текста информации отражают не столько статистику сочетаемости лингвистических единиц, сколько усвоение субъектом характеристик такой сочетаемости, его субъективные знания, субъективные вероятностные

оценки относительно элементов языка, сложившиеся в результате прошлого опыта.

Таким образом, не только прагматический компонент контекстной обусловленности, но и статистический ее компонент характеризуют в наших экспериментах, главным образом опыт угадчика.

Вместе с тем эксперименты подтвердили возможность использования измерения субъективных информационных параметров языкового текста для оценки их влияния на различные аспекты субъективной деятельности человека по отношению к этому тексту, в частности на его запоминание. До введения информационных мер не было единого критерия, который мог бы оценить степень запоминания различных по своему содержанию и структуре осмысленных (а также бессмысленных) сообщений. Можно думать, что информационные измерения смыслового материала позволят установить влияние информационных параметров на восприятие, скорость чтения, понимание, обучение, принятие решений и другие виды деятельности.

Следует отметить, что употреблявшееся нами понятие прагматической информации как отражения прошлого опыта субъекта не является общепринятым. В своем точном значении прагматическая информация характеризует ценность сообщения для его получателя. Однако этот вопрос в нашей статье не рассматривался, и речь шла об объективной и субъективной энтропии текста, объективной и субъективной его избыточности.

Для человека как для получателя лингвистических и других сообщений объективной избыточности просто не существует, если ему не известны те ограничения разнообразия, которые имеются в сообщении и вообще в объективном мире. Поэтому сообщение на неизвестном языке может субъективно восприниматься как случайный набор символов. Вместе с тем, воспринимая какие-нибудь последовательности внешних воздействий, человек (или животное) пытается исследовать скрытые от него ограничения и закономерности. В этом основа ориентировки в окружающем мире, приспособления к изменению внешних условий и познания объективного мира. Обучение человека и животных возможно вообще лишь тогда, когда «окружающая среда обнаруживает некоторые ограничения разнообразия».[17] Память нужна только в том случае, «если внешняя среда устроена так, что будущее часто п о в т о р я е т прошлое; если бы события будущего часто были бы п р о т и в о п о л о ж н ы, память была бы невыгодна».[18]

Творческое мышление человека с точки зрения понятия избыточности возможно только потому, что при решении информационных задач всегда имеется та или иная система ограниче-

[17] У. Р. Э ш б и. Введение в кибернетику. Русск. пер. М., 1959.

[18] У. Р. Э ш б и. Принципы самоорганизации. Русск. пер. В кн.: Принципы самоорганизации, М., 1966, стр. 314—343.

ний, что при отсутствии ограничений единственно возможной тактикой является полный упорядоченный перебор и что в этом случае мозг гения будет решать задачи так же, как и мозг голубя.[19]

Субъективная избыточность отражает степень познания человеком объективных ограничений и закономерностей, и эта степень познания тем больше, чем меньше различия между избыточностью объективной и субъективной.

[19] См.: А. В. Напалков. Алгоритмический анализ сложных форм работы мозга.

В. П. Григорьев

ПРЕДВАРИТЕЛЬНЫЕ ИТОГИ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ ПОЭТИКИ ИСПАНСКОГО НАРОДНОГО РОМАНСА

Использование статистических методов в анализе поэтического текста уже перестало быть новостью в литературоведении. Назовем, в частности, работы Ю. И. Левина и З. Г. Минц.[1] Известны теоретические обоснования возможностей применения подобных методов исследования поэтического текста у Р. Якобсона, Ю. М. Лотмана и других. Кроме того, в принципе литературовед может использовать частотный словарь любого законченного литературного произведения или совокупности произведений. Однако ни сам выбор текста, ни методика его исследования до сих пор еще никем не мотивировались, что понятно при современном состоянии структурного литературоведения.

В нашем случае текст выбран исходя из тех соображений, что для статистического исследования наилучшим мог бы явиться текст с максимально развитыми внутритекстовыми отношениями и минимально развитыми внетекстовыми отношениями. Идеальным примером является средневековый текст, еще лучше — средневековый текст фольклорного характера. Так как язык в статистических исследованиях не играет принципиальной роли, нами выбран испанский народный романс, до сих пор представляющий загадку во многих своих жанровых и языковых проявлениях.[2]

Нашу работу значительно облегчает то обстоятельство, что к настоящему времени внетекстовые связи средневекового текста в значительной мере выявлены.

[1] Ю. И. Левин. О некоторых чертах плана содержания в поэтических текстах. В сб.: Структурная типология языка, М., 1966; З. Г. Минц, Л. А. Аболдуева, О. А. Шишкина. Частотный словарь «Стихов о прекрасной даме» А. Блока и некоторые замечания о структуре цикла. Тр. по знаковым системам. Уч. зап. Тартуского гос. унив., т. 3, 1967.

[2] S. G. Morley. Spanish Ballad Problems. Univ. of California publ., 1925; R. Menéndez Pidal. Romancero hispánico. Madrid, 1953, 2 vols.

51 PORQUE	22	3643	+2955003-02
52 ERA	21	3664	+2820685-02
53 HABLADO	21	3685	+2820685-02
54 NON	21	3706	+2820685-02
55 PARA	21	3727	+2820685-02
56 YA	21	3748	+2820685-02
57 SANCHO	20	3768	+2686366-02
58 DIOS	19	3787	+2552048-02
59 GRADO	19	3806	+2552048-02
60 MAL	19	3825	+25520 02
61 ALLI	18	3843	+24177 02
62 ESTADO	18	3861	+24177 02
63 TU	18	3879	+24177 02
64 ARIAS	17	3896	2263
65 DIEGO	17	3913	283
66 SIN	17	3930	283
67 UNA	17	3947	
68 ESE	15	3962	
69 FUE	15	3977	
70 MIS	15	3992	+2 02
71 OS	15	4007	+2 02
72 SENOR	15	4022	+2 02
73 VAN	15	4037	+2014775-02
74 GONZALO	14	4051	+1880456-02
75 MANFRA	14	4065	+1880456-02
76 QUIEN	14	4079	+1880456-02
77 CUATRO	13	4092	+1746138-02
78 DOS	13	4105	+1746138-02
79 ELLA	13	4118	+1746138-02
80 SER	13	4131	+1746138-02
81 TAN	13	4144	+1746138-02
82 TODO	13	4157	+1746138-02
83 VELLIDO	13	4170	+1746138-02
84 CORTES	12	4182	+1611820-02
85 DICHO	12	4194	+1611820-02
86 HE	12	4206	+1611820-02
87 NOS	12	4218	+1611820-02
88 SERA	12	4230	+1611820-02
89 CONDE	11	4241	+1477501-02
90 CONDES	11	4252	+1477501-02
91 CONSEJO	11	4263	+1477501-02
92 HONRADO	11	4274	+1477501-02
93 LES	11	4285	+1477501-02
94 LLAMADO	11	4296	+1477501-02
95 RODRIGO	11	4307	+1477501-02
96 SEA	11	4318	+1477501-02
97 SOY	11	4329	+1477501-02
98 YA	11	4340	+1477501-02
99 CASTILLA	10	4350	+1343183-02
100 CINCO	10	4360	+1343183-02

Два фрагмента текста, сделанных на ЭВМ «Минск-22» (начало текста не приводится, так как в первые десятки слов обычно входят служебные слова, не рассматриваемые автором).

+4863216+00	+2105591-01	+2720627+01
+4921423+00	+2112531-01	+2741752+01
+4949630+00	+2112531-01	+2762877+01
+4977837+00	+2112531-01	+2784002+01
+5006044+00	+2112531-01	+2805128+01
+5034251+00	+2112531-01	+2826253+01
+5061114+00	+2028655-01	+2846540+01
+5089535+00	+1943921-01	+2865979+01
+5112155+00	+1943921-01	+2885419+01
+5137676+00	+1943921-01	+2904857+01
+5161853+00	+1858286-01	+2923440+01
+5189050+00	+1858286-01	+2942023+01
+5210208+00	+1858286-01	+2960606+01
+5233942+00	+1771698-01	+2978323+01
+5255876+00	+1771698-01	+2996040+01
+5278710+00	+1698-01	+3013757+01
+5301544+00	+1698-01	+3031474+01
+5321602+00	+1633-01	+3047428+01
+5341660+00	+1633-01	+3063382+01
+5361907+00	+1595433-01	+3079337+01
+5382135+00	+1595433-01	+3095291+01
+5402283+00	+1595433-01	+3111245+01
+5422431+00	+1595433-01	+3127200+01
+5441235+00	+1505622-01	+3142256+01
+5460040+00	+1505622-01	+3157312+01
+5478844+00	+1505622-01	+3172368+01
+5498306+00	+1414585-01	+3186514+01
+5513767+00	+1414585-01	+3200660+01
+5531228+00	+1414585-01	+3214806+01
+5548690+00	+1414585-01	+3228952+01
+5566151+00	+1414585-01	+3243098+01
+5583613+00	+1414585-01	+3257244+01
+5601074+00	+1414585-01	+3271389+01
+5617192+00	+1322230-01	+3284612+01
+5633310+00	+1322230-01	+3297834+01
+5649429+00	+1322230-01	+3311056+01
+5665547+00	+1322230-01	+3324279+01
+5681665+00	+1322230-01	+3337501+01
+5698444+00	+1228444-01	+3349785+01
+5711215+00	+1228444-01	+3362070+01
+5725990+00	+1228444-01	+3374354+01
+5740765+00	+1228444-01	+3386639+01
+5755540+00	+1228444-01	+3398923+01
+5770315+00	+1228444-01	+3411208+01
+5785090+00	+1228444-01	+3423492+01
+5799865+00	+1228444-01	+3435777+01
+5814640+00	+1228444-01	+3448061+01
+5829415+00	+1228444-01	+3460345+01
+5842847+00	+1133099-01	+3471678+01
+5856279+00	+1133099-01	+3483007+01

Два фрагмента. *(продолжение)*.

51 HA	18	2955	+2861233-02
52 MIS	18	2973	+2861230-02
53 MORO	18	2991	+2861230-02
54 PADRE	18	3009	+2861230-02
55 PARA	18	3027	+2861230-02
56 ES	17	3044	+2702273-02
57 ESTA	17	3061	+2702273-02
58 OS	17	3078	+2702273-02
59 PALABRAS	17	3095	+2702273-02
60 VIDO	17	3112	+2702273-02
61 YA	17	3129	+2702273-02
62 AQUESTO	16	3145	+2543315-02
63 DOS	16	3161	+2543315-02
64 TE	16	3177	+2543315-02
65 TU	16	3193	+2543315-02
66 VELLIDO	16	3209	+2543315-02
67 ALONSO	15	3224	+2384358-02
68 ARMAS	15	3239	+2384358-02
69 CABALLERO	15	3254	+2384358-02
70 CABALLEROS	15	3269	+2384358-02
71 CABALLO	15	3284	+2384358-02
72 DESQUE	15	3299	+2384358-02
73 AQUEL	14	3313	+2225401-02
74 DIJO	14	3327	+2225401-02
75 DIOS	14	3341	+2225401-02
76 GONZALO	14	3355	+2225401-02
77 SON	13	3368	+2068444-02
78 CONDES	12	3380	+1907486-02
79 GRAN	12	3392	+1907486-02
80 HIJOS	12	3404	+1907486-02
81 LES	12	3416	+1907486-02
82 RIEPTO	12	3428	+1907486-02
83 TAL	12	3440	+1907486-02
84 TRES	12	3452	+1907486-02
85 DICEN	11	3463	+1748529-02
86 ELLOS	11	3474	+1748529-02
87 MAL	11	3485	+1748529-02
88 TRAIDOR	11	3496	+1748529-02
89 DIAS	10	3506	+1589572-02
90 DOLFOS	10	3516	+1589572-02
91 LUEGO	10	3526	+1589572-02
92 ORDONEZ	10	3536	+1589572-02
93 PUES	10	3546	+1589572-02
94 VIEJO	10	3556	+1589572-02
95 DIA	9	3565	+1430615-02
96 DUQUE	9	3574	+1430615-02
97 ENTRE	9	3583	+1430615-02
98 ESE	9	3592	+1430615-02
99 HIJA	9	3601	+1430615-02
100 SER 6	9	3610	+1430615-02

Два фрагмента . . . (*продолжение*).

+4697186+00	+2137688-01	+2608813+01
+4725798+00	+2137685-01	+2630190+01
+4754410+00	+2137688-01	+2651567+01
+4783023+00	+2137688-01	+2672944+01
+4811635+00	+2137688-01	+2694320+01
+4838658+00	+2038631-01	+2714707+01
+4865681+00	+2038631-01	+2735093+01
+4892703+00	+2038631-01	+2755479+01
+4919726+00	+2038631-01	+2775866+01
+4946749+00	+2038631-01	+2796252+01
+4973771+00	+2038631-01	+2816638+01
+4999205+00	+1938382-01	+2836022+01
+5024638+00	+1938382-01	+2855406+01
+5050071+00	+1938382-01	+2874790+01
+5075504+00	+1938382-01	+2894174+01
+5100937+00	+1938382-01	+2913557+01
+5126781+00	+1836864-01	+2931926+01
+5148624+00	+1836864-01	+2950295+01
+5172468+00	+1836864-01	+2968663+01
+5196312+00	+1836864-01	+2987032+01
+5220155+00	+1836864-01	+3005401+01
+5243999+00	+1836864-01	+3023769+01
+5266253+00	+1733993-01	+3041139+01
+5288507+00	+1733993-01	+3058449+01
+5310761+00	+1733993-01	+3075789+01
+5333015+00	+1733993-01	+3093129+01
+5353679+00	+1629673-01	+3109426+01
+5372754+00	+1523791-01	+3124664+01
+5391829+00	+1523791-01	+3139902+01
+5410904+00	+1523791-01	+3155141+01
+5429979+00	+1523791-01	+3170378+01
+5449054+00	+1523791-01	+3185615+01
+5468128+00	+1523791-01	+3200853+01
+5487203+00	+1523791-01	+3216091+01
+5504689+00	+1416217-01	+3230253+01
+5522174+00	+1416217-01	+3244416+01
+5539659+00	+1416217-01	+3258578+01
+5557144+00	+1416217-01	+3272740+01
+5573040+00	+1306798-01	+3285808+01
+5588936+00	+1306798-01	+3298876+01
+5604832+00	+1306798-01	+3311944+01
+5620727+00	+1306798-01	+3325012+01
+5636623+00	+1306798-01	+3338080+01
+5652519+00	+1306798-01	+3351148+01
+5666825+00	+1195347-01	+3363101+01
+5681131+00	+1195347-01	+3375055+01
+5695437+00	+1195347-01	+3387008+01
+5709743+00	+1195347-01	+3398962+01
+5724050+00	+1195347-01	+3410915+01
+5738356+00	+1195347-01	+3422869+01

Два фрагмента . . . (*продолжение*).

Из комплекса еще не изученных вопросов поэтики испанского романса нами выбрано соотношение его строк: нечетной и четной — в современном написании строки восьмисложного романса, или первой и второй половины строки старинного шестнадцатисложного романса.[3]

Дело в том, что нечетные строки испанского романса не имеют рифмы или ассонанса, четные же имеют общий ассонанс. Современный читатель (а также все поэтики) воспринимают метрическую структуру романса именно как соотношение ассонированных (четных) и свободных (нечетных) строк, хотя, вероятно, первоначальная структура романса могла быть иной.

Предполагалось, что соотношение нечетных и четных строк дадут также выходы в поэтическую лабораторию хуглара: выбор слов в каждом из видов строк, закономерности перестановок слов в пределах одной стихотворной фразы и т. д.

Анализ одного отдельно взятого романса не может прояснить вопроса, так как внутренние характеристики соотношения строк изменчивы и часто зависят от темы романса. Полученные результаты были бы малопоказательны и не могли бы служить основанием для общих заключений. Предполагалось, что быстродействующая счетная машина сможет дать лучшие результаты, основанные на анализе большого объема поэтического текста.

ЭВМ «Минск-22» проанализировала поэтический текст всего цикла романсов о Сиде. Полученные результаты представляют собой два частотных словаря первой и второй половины поэтических строк романсов (см. рисунок). Кроме списка слов по убывающей величине, машина дала значения абсолютной частоты (F), накопленной абсолютной частоты (F^*), относительной частоты (f), накопленной относительной частоты (f^*), информации (I) и накопленной информации (I^*). Обработке подверглись 1315 строк поэтического текста общим объемом в 13 636 слов. Первая половина строк романсов состоит из 6291 слова, из них не повторяются 1458. Для второй половины строк соответственно — 7445 и 1645. Отношение неповторяющихся слов к общей массе слов по строкам: 4.3 — в первой половине и 4.5 — во второй, т. е. практически одинаково.

Во второй половине строки (далее — в четной строке) общая масса слов на 1154 слова больше, чем в первой (далее — в нечетной), что при дальнейших расчетах следует принимать во внимание во избежание ошибок.

Большее количество словоформ [4] в четных строках при их метрическом равенстве с нечетными позволяет утверждать, что

[3] Так, например, даны используемые нами романсы в издании: F. I. Wolf y C. Hofmann. Romances viejos castellanos, ed. 2-a, t. 1. Biblioteca clásica, t. CCVIII. Antología de poetas líricos, t. VIII, Madrid, s. a.

[4] Перед нами частотный словарь словоформ, так как даются отдельные значения всех форм слова (множественное или единственное число, парадигмы

четные строки должны в принципе содержать более короткие слова. Реально это достигается за счет увеличения количества служебных слов и глагольных форм.

Интуитивно считалось, что смысл четной строки романса как бы в том, что она развивает мысль нечетной и «дотягивает» ее до ассонанса (В. Лемке, Р. Веббер, Л. Пфандль, М. Пидаль и другие). Это тем более удобно, что ассонанс не предъявляет к авторам таких требований, как рифма. На его месте может быть любое слово, включая служебное.

Сравнение таблиц частотного словаря четной и нечетной строк помогает точнее понять механизм ассонанса в испанском романсе. Действительно, словоформы, наиболее часто встречаемые в ассонансе, преобладают в четной строке. Но это лишь одна сторона проблемы. Другая сторона в том, что словоформы, используемые в ассонансе, имеют в испанском языке относительно закрепленное место в предложении. Следовательно, отнесение этих строк в конец строфы автоматически перестраивает обычный порядок слов в новый — «поэтический». Однако для доказательства этого положения нужно другое, новое исследование. Здесь же укажем на результаты дистрибуции no, son, non по строкам:

no	в строке	с	ассонансом	(четной)	на	8-м	месте
no	»	»	без	ассонанса	(нечетной)	»	15-м	»
son	»	»	с	ассонансом	(четной)	»	42-м	»
son	»	»	без	ассонанса	(нечетной)	»	77-м	»
non	»	»	с	ассонансом	(четной)	»	54-м	»
non	»	»	без	ассонанса	(нечетной)	»	168-м	»

Значительное увеличение количества приведенных слов в строке с ассонансом в первую очередь объясняется тем, что они сами используются для создания ассонанса, т. е. становятся более «поэтическими» от места в стихе, а также потому, что вызывают в памяти слушателей известные поэтические штампы с ними (необходимейшее условие поэтичности в средневековой поэтике).

Например, ассонанс с no:

...no merecés culpa, no («Романс о Фернан Гонсалесе»);
...no casar contigo, no («Романс о холодном источнике»).

Количество таких стихов можно умножить не только за счет примеров из романсов, но также за счет других жанров испанской поэзии.

Еще более доказательно распределение причастий прошедшего времени и слов на -ado (ассонанс ao наиболее част в испанском

глагола и т. д.). К сожалению, по техническим причинам не получена вторая пара частотных словарей, где слова выявились бы как целые понятия, вне зависимости от их позиционных изменений.

романсе). В четной строке мы насчитываем 385 подобных слов, причем один раз употреблены 105 слов, два раза — 40, три раза — 21 и т. д.; используемых же многократно слов на -ado очень немного (2 слова — 9 и 11 раз, по одному слову — 18 и 21 раз). Распределение очень показательное. По-видимому, пресловутое «безразличие» к разнообразию в ассонансе не совсем отвечает действительной практике певца-хуглара.

В нечетной строке мы встречаемся только с 25 случаями употребления словоформ на -ado (33 словоупотребления), причем большинство из них не являются причастиями прошедшего времени. Итак, соотношение — 385 и 33. Направленность совершенно ясная. Исполнители романсов сознательно относили все эти формы в четную строку, заставляя тем самым перестраиваться определенным образом нечетную. Кроме того, необходимо учитывать, что причастия на -ado обладают и большой протяженностью; в строке всего восемь слогов, и два из них падают только на одно окончание причастия. Еще одно статистическое подтверждение того известного факта, что информативность поэтического слова зависит прежде всего от его места в стихе.

Несколько менее показательны словоформы на -ido — соответственно 87 и 28. Мы исключаем отсюда форму vido, соответствующую архаической форме перфекта (совр. vió).[5] Подобные же рассуждения можно с различной статистической характеристикой отнести к словоформам на -ada, -ara и, в меньшей степени в нашем цикле, к словам на -aba. В общей сложности все перечисленные формы составляют почти 20% словоупотреблений четных строк.

Сравнение частотных таблиц четной и нечетной строк показывает, что слова, отнесенные хугларом на конец четной строки, а также остальные слова — наполнители этой строки находились в каких-то определенных соотношениях со словами, образующими нечетную строку. Смысл этого соотношения проясняется тем неожиданным для исследователя фактом, вскрываемым полностью только с помощью статистического обследования больших массивов текста, что таблицы четных и нечетных строк практически повторяют в тех или иных пропорциях одна другую. Практически мы не можем привести почти ни одного значимого для романса слова, которое бы не было продублировано в обеих строках.

В табл. 1 приводим наиболее употребительные глаголы.

В принципе слова, наиболее существенные для понимания специфики романса, т. е. слова, несущие бóльшую информацию, сосредоточены в нечетных, зачинательных строках, в то время как в четных преобладают слова более общего характера.

[5] Любопытно отметить, что «формальное» звучание слова не заслонило собой для певца романса его «содержательности». Почти все случаи употребления слова фиксируются только в нечетной строке.

Таблица 1

№№	Словоформы	Частота употребления		№№	Словоформы	Частота употребления	
		в нечетной строке	в четной строке			в нечетной строке	в четной строке
1	ser	115	228	9	hablar	25	54
2	haber	67	231	10	tener	25	35
3	decir	65	45	11	oir	24	11
4	hacer	47	45	12	matar	22	16
5	ir	46	49	13	mandar	17	10
6	morir	45	21	14	tomar	13	19
7	ver	40	14	15	llamar	7	20
8	dar	31	65				

В нечетных мы видим глаголы со значением: «сражаться», «бить», «жить», «умирать», «убивать», «приказывать», «видеть» и т. п. В то же время в четных больше слов типа: «дать», «брать», «иметь». Активное: 'сказать' (decir) — в нечетных, более пассивное: 'говорить' (hablar) — в четных и т. д.

Еще более показательно значительное преобладание в четных строках (в 2—3 раза) вспомогательных глаголов ser, haber, которые зачастую лишены собственной семантики и выступают лишь как формальные наполнители строки.

Почти такая же картина и в распределении существительных, прилагательных, наречий и т. д. (см. табл. 2).

Таблица 2

№№	Словоформы	Частота употребления		№№	Словоформы	Частота употребления	
		в нечетных строках	в четных строках			в нечетных строках	в четных строках
1	rey	119	77	10	caballo	22	38
2	don	70	56	11	hijo	19	24
3	Cid	58	34	12	padre	18	27
4	todos	39	49	13	traidor(es)	18	20
5	conde(s)	35	23	14	palabras	17	4
6	mas	33	47	15	campo	5	34
7	caballero(s)	30	12		и даже		
8	Diego	25	17	16	Gonzalo	14	14
9	Sancho	23	20	17	Urraca	7	7

Обращаем внимание, что и существительные в определенной мере подвержены той же закономерности, что отмечена выше у глаголов. Активные участники событий, лица, непосредственно дей-

ствующие, чаще фиксируются в нечетных строках; в то время как аксессуары, указания на место действия, различного рода детали, даже существенные для понимания действия, находятся обычно в четных строках. Однако делать какие-либо далеко идущие заключения материал еще не позволяет, нужны дополнительные исследования.

К сожалению, пока остались неиспользованными величины f и f^*. В частности, величина f^* еще раз подтверждает некоторую однотипность построения строк романса. Табл. 3 показывает покрываемость текста.

Таблица 3

% покрываемости текста	Словоформы		% покрываемости текста	Словоформы	
	нечетной строки	четной строки		нечетной строки	четной строки
25	11	11	90	829	901
50	63	55	95	1144	1272
75	322	314	99	1396	1570
80	435	437			

Что же касается величин информации I, I^*, то их использование хотя и заманчиво, но пока еще не сулит особых перспектив: информация в языке и в художественном произведении различается в принципе. Отметим все же, что в среднем информативность четной строки несколько ниже, чем нечетной. Интуитивная оценка селективной информации обеих частей строки романса свидетельствует о том же. Сам характер сообщения в четной строке предопределяет ее меньшую информативность.

На первый взгляд это выглядит парадоксальным. Ассонанс, как и рифма, поэтически должен быть более информативен. Так и обстоит дело в действительности (см. пример с ассонансом по). Но научить машину вычислять информацию связей не только внутритекстовых, но и внетекстовых пока не удается, хотя в принципе материал испанского романса может позволить и это.

Подводя предварительные итоги статистического исследования проблем испанского романса, необходимо подчеркнуть следующее.

1) Многие выводы, полученные в процессе статистического эксперимента, строго говоря, могут быть теоретически выведены и интуитивно, что, однако, не умаляет значение статистических исследований, в особенности для будущего.

2) Соотношения между первой и второй половиной (четной и нечетной строкой) шестнадцатисложной строки романса еще требуют дополнительного исследования; однако уже сейчас ясно, что их зависимость определена не только наличием ассонанса, изменением порядка слов и возникающим отсюда «поэтизмом» синтаксиса в четных строках, но и твердым соотношением лексических единиц в каждой половине строки, порядок и выбор которых определялись как поэтикой жанра, так и индивидуальными особенностями его создателей и исполнителей.

3) Поэтика испанского романса по своей сути близка к статистике: ее идеал — «среднестатистическое». Талант исполнителя не должен заслонять собой фактического материала, данного традицией; поэтическая информация четных строк уравновешивается статистической информацией нечетных, частая немотивированность начала или конца романса нивелируются известностью его основных положений слушателям и т. д. Все это дает дополнительные основания если не проверять «математикой гармонию», то, по крайней мере, познавать гармонию с помощью и математики.

П. М. Алексеев

ЧАСТОТНЫЕ СЛОВАРИ АНГЛИЙСКОГО ЯЗЫКА И ИХ ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

За сравнительно недолгий срок существования квантитативная лингвистика в целом и ее важнейшая отрасль — статистическая лексикография — приобрели уже некоторую историю. Высказываются даже различные мнения о приоритете того или иного языка или страны. Д. Харкин, например, считает, что первым частотным словарем был список китайских иероглифов, изданный сто лет назад.[1] А. Жюйан настаивает на первенстве румынского языка;[2] Р. Мишеа напоминает, что частотные индексы к религиозным текстам составлялись еще в средние века,[3] а Дж. Юл относит зарождение лексической статистики к началу новой эры.[4]

И все же только на пороге XX в. появляется серьезная потребность в частотных словарях; первым таким словарем принято считать монументальный труд немца Ф. Кединга и тысячи его помощников.[5]

Многочисленные конкордансы, словоуказатели к текстам отдельных авторов или произведений следует относить к особой категории [6] хотя бы потому, что по внешнему виду и по своим целям они отличаются от собственно частотных словарей.

Первые частотные списки английского языка уступают словарю Ф. Кединга и по объему использованного материала и по размерам словника, однако количественным исследованиям именно английского языка обязана лингвостатистика своим становлением и дальнейшим развитием.

[1] W. Gamble. Two Lists of Selected Characters Containing All in the Bible and Twenty Seven Other Books. Shanghai, 1861; см.: D. Harkin. The History of Word Counts. «Babel», VIII, 3, 1957.

[2] A. Juilland, P. M. Edwards, I. Juilland. Frequency Dictionary of Rumanian Words, The Hague, 1965.

[3] Р. Мишеа. Словари основной лексики. В сб.: Методика преподавания иностранных языков за рубежом, М., 1967.

[4] G. U. Yule. The Statistical Study of Literary Vocabulary. Cambridge, 1944, pp. 7—8.

[5] F. W. Kaeding. Häufigkeitswörterbuch der deutschen Sprache. Steglitz bei Berlin, 1897—1898.

[6] Р. М. Фрумкина. Статистические методы изучения лексики. М., 1964, стр. 5, 7—8.

Ниже будут рассмотрены некоторые известные в литературе частотные словари английского языка.

1) J. Knowles. The London Point System of Reading for Blind. London, 1904.[7]

В этой книжке священник Дж. Ноулз предложил разработанную им систему чтения для слепых, в которой использовались данные статистического анализа словоформ в текстах общей длиной 100 тыс. словоупотреблений, взятых из Библии и произведений различных авторов. Он приводит частотный [8] список 353 словоформ, встретившихся от 8131 до 25 раз и занимающих более 3/4 всего текста. Других статистических данных не сообщается.

Список Дж. Ноулза считают первым частотным словарем английского языка.

2) R. C. Eldridge. Six Thousand Common English Words. Niagara Falls, 1911.

Управляющий небольшой фабрикой, в числе рабочих которой были иммигранты, за два года расписал 250 статей из четырех номеров воскресных провинциальных газет общей длиной 43 989 словоупотреблений. Первоначальной целью было выявление лексического минимума, который должен был помочь рабочим скорее овладеть разговорной речью. В процессе анализа текста Р. Элдридж понял широкие возможности применения частотных словарей, о которых он и сообщил в предисловии к книге. Каждый год в Америку переселяется полмиллиона людей, совершенно не владеющих английским языком. Неоценимую помощь окажут им статистические словари-минимумы. Количественные данные о словоупотреблениях и графемах будут использованы при рационализации английской орфографии; параллельные частотные словари различных языков позволят создать универсальный словарь-полиглот и послужат возникновению международного языка. Частотные словари будут способствовать более эффективному управлению колониальными предприятиями. . .

Свои данные Р. Элдридж представил в четырех частотных списках всех словоформ, зарегистрированных в каждой из газет. В пятом, сводном, списке помещены все 6002 словоформы с частотами от 4290 по 1. Последним приведен список Дж. Ноулза.

Словарь Р. Элдриджа, остававшийся до недавнего времени почти единственным, где представлены все обнаруженные в тексте словоформы, даже самые редкие, был использован при формулировке Дж. Ципфом своего «закона».[9]

[7] По Р. Элдриджу, Л. Эйрзу (1915) и другим источникам.

[8] Здесь и далее термин «частотный список» означает, что лексические единицы упорядочены по убыванию частот. В «алфавитно-частотном» они размещены по алфавиту (частоты также указаны).

[9] См., например: G. Zipf. The Psycho-Biology of Language. Cambridge, 1935 (1968), II. The Form and Behavior of Words.

3) L. P. Ayres. The Spelling Vocabulary of Personal and Business Letters. New York, 1913.[10]

Обследованы 2000 писем, из которых по первому слову со строки сделана выборка объемом 23 629 словоупотреблений; в их числе «словарных» слов оказалось 2001. В списке даны 542 слова с частотой 6 и выше.

4) W. A. Cook, M. V. O'Shea. The Child and His Spelling. Indianapolis, 1914.[11]

Здесь приводятся алфавитно-частотные списки 5200 словарных слов, зарегистрированных в 200 тыс. словоупотреблений из переписки тринадцати взрослых корреспондентов. Делается попытка некоторого анализа данных; 186 словоформ использованы всеми авторами, 577 — большинством. Имеются таблицы наращения словаря на каждую тысячу словоупотреблений.

5) L. P. Ayres. A Measuring Scale for Ability in Spelling. New York, 1915.

Выборка (100 тыс. словоупотреблений) составлена, как и в предыдущей работе Л. Эйрза, по одному слову со строки из писем 2500 авторов. Приводится частотный список первой тысячи словарных слов, встретившихся не реже 12 раз. Дается также алфавитный список, но без указания частот. Общее количество разных слов не сообщается.

6) 16 Spelling Scales. Teachers College Record, vol. 21, 4, 1920.[12]

Алфавитно-частотный список второй и третьей тысячи слов по материалам Л. Эйрза.

7) E. L. Thorndike. The Teacher's Word Book. New York, 1921 (1927).[13]

Это первый из серии больших частотных словарей, создававшихся в течение 25 лет профессорами Колумбийского университета Э. Торндайком и И. Лорджем. В варианте 1921—1927 гг. Э. Торндайк публикует 10 тыс. наиболее «важных» для обучения языку слов в их традиционной словарной форме; отдельные входы имеют лишь нерегулярные и некоторые регулярные (например, being и building как существительные) словоформы. Наречия на -ly также считаются словоформами и сводятся к «исходному» прилагательному.

Важность слова определяется по его частоте и наличию в определенном количестве источников (всего их 41). Коэффициентом важности слова является его «credit-number», который и указывается при каждом слове в списке.[14] Например, коэффициент 49 и выше означает, что слово относится к первой тысяче самых важных слов, от 29 по 48 — ко второй тысяче и т. д. Кроме этого коэффициента, первые 5000 слов снабжены показателем того, в какую тысячу и полутысячу от начала они входят. Индекс «4а» соответствует первым пятистам словам из четвертой тысячи.

Общий объем выборки 4 565 000 словоупотреблений; из них 3 млн из Библии и английских классиков, 625 тыс. из детской литературы, 300 тыс. из учебников, 500 тыс. из переписки, 90 тыс. из газет и 50 тыс. из книг по домоводству, сельскому хозяйству и ремеслам.

Свой словарь Э. Торндайк предназначал как пособие при обучении чтению. В этом его принципиальное отличие от работ, исследовавших лексику корреспонденции, т. е. письменно-активный словарь.

Г. Дьюи подвергает резкой критике систему «коэффициентов важности» Э. Торндайка, говоря, что она искажает представление об употребительности слов, которую следует определять по абсолютным частотам.

Общее количество разных слов, обнаруженных в выборке, не указывается. Отсутствуют и другие сведения, которые позволили бы использовать данные словаря как материал для дальнейших статистических исследований.

8) W. N. Andersen. Determination of a Spelling Vocabulary Based upon Written Correspondence. University of Iowa Studies in Education, vol. VII, 1, 1921.[15]

Исследованы 361 184 словоупотребления из 3723 частных и деловых писем. Обнаружены 9223 разные словоформы, из которых приводятся 3087 с частотой не ниже 5 и встретившиеся не менее чем в трех группах текстов из шести (деловых, профессиональных, частных и т. д.).

9) G. Dewey. Relativ Frequency of English Speech Sounds. Cambridge, 1923.

Г. Дьюи проанализировал выборку из 13 различных «жанров» английской письменной речи общим объемом 100 тыс. словоупотреблений в целях усовершенствования орфографии. Книга напечатана, между прочим, по предложенной им системе упрощенного написания многих слов, в том числе и слова relative как relativ. Это написание отражено в заглавии и нередко смущает читателя, встречающего его в ссылках и лишенного возможности непосредственно ознакомиться с книжкой Г. Дьюи.

[10] По Л. Эйрзу (1915), Г. Дьюи и Э. Хорну (1926).
[11] По Г. Дьюи и Э. Хорну (1926).
[12] По Г. Дьюи.
[13] Издание 1927 г. отличается только тем, что в списки добавлено несколько новых слов.
[14] Методика расчета «коэффициента важности» в книге не дается. Она приводится в работе: E. L. Thorndike. Word Knowledge in the Elementary School. Teachers College Record, September, 1921.
[15] По Э. Хорну (1926).

Автор приводит частотный и алфавитно-частотный списки 1027 самых частых словоформ; 1131 «корневое» слово размещено в аналогичных списках. Кроме частот, «корневые» слова снабжены указанием на количество сведенных к ним разных словоформ, причем каждое такое слово объединяет не только словоформы, но и производные слова. Общее число обнаруженных разных словоформ равно 10 161; что касается «корневых» слов, то подобных сведений о них не дается.

Книга Г. Дьюи примечательна в двух отношениях. Во-первых, она явилась первым в строгом смысле частотным словарем (см. выше о недостатках словаря Э. Торндайка) английского языка «в целом», хотя и основанным на умеренной выборке. Во-вторых, через 25 лет она послужила материалом для разработки К. Шенноном основ теории информации [16] и явилась как бы мостом между языкознанием и современными точными науками.

10) E. H o r n. The Vocabulary of Bankers' Letters. English Journal, vol. XII, 6, June, 1923.[17]

Приводится 2623 разные словоформы, встретившиеся в 1125 письмах банкиров по деловым вопросам. Объем выборки — 67 581 словоупотребление. Эта работа была пробой для последовавшего обширного анализа лексики писем, которым руководил Э. Хорн.

10a) E. H o r n. A Basic Writing Vocabulary. 10 000 Words Most Commonly Used in Writing. Iowa City, 1926.

Повышенный интерес к словарю переписки возник не случайно. Многих исследователей интересовал вопрос об активном словаре и количественных соотношениях между его единицами. Поэтому до появления одного из крупнейших (на сегодня третьего по объему выборки) частотных словарей английского языка Э. Хорна было составлено не менее полутора десятков словарей и списков слов и словоформ, используемых в письмах.

После того как Э. Хорн составил свой первый словарь и подготовил (но не опубликовал) второй вариант, куда собрал данные всех словарей переписки и свои собственные, был создан специальный фонд для изучения лексики корреспонденции в широких масштабах. Результаты этого исследования и вошли в большой словарь Э. Хорна.

Номинально объем выборки равен 5 136 816 словоупотреблениям.[18] Последовательность ее наращения была такой: вначале исследовалось 67 851 словоупотребление, затем, объединив данные разных списков, довели объем выборки до 864 334. Из дальнейшего

[16] К. Ш е н н о н. Работы по теории информации и кибернетике. Русск. пер. М., 1963, статьи: «Математическая теория связи» (1948), «Предсказание и энтропия печатного английского текста» (1951).

[17] По Э. Хорну (1926).

[18] Но не 15 000 тыс., как утверждает А. Роберте (A. R o b e r t s. A Statistical Linguistic Analysis of American English. The Hague, 1965) и вслед за ним В. М. Андрющенко (ВЯ, 1968, № 5, стр. 144).

анализа были исключены 372 самые частые словоформы. Когда анализ текстов был завершен, появилась необходимость все же восстановить частоты этих словоформ, что и было сделано экстраполяцией по выборке в 864 тыс. Оцененный таким образом окончательный объем выборки оказался равным 5 136 816 словоупотреблениям.

Инструкция по отбору словоформ в опубликованный список состоит из тридцати правил, но главным критерием были частота и распространенность словоформы по источникам. Коэффициент употребительности (Э. Хорн называет его, как и Э. Торндайк, «credit-number») определяется как произведение частоты на квадратный корень из числа словарей-источников, регистрирующих эту словоформу (всего таких словарей, составленных коллегами Хорна и другими авторами, было 65). Словарь Э. Хорна, таким образом, указывает не фактически наблюденные, а «оцененные» с помощью вышеприведенной методики частоты.[19]

11) Twenty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, 1925, pp. 186—198.[20]

Здесь приводятся данные о частотах слов в записях детской речи объемом 600 тыс. словоупотреблений.

12) N. R. F r e n c h, C. W. C a r t e r, W. K o e n i g. Words and Sounds of Telephone Conversations. The Bell System Technical Journal, vol. IX, 2, 1930.

Первый частотный словарь английской разговорной речи. Составлен путем выборочной регистрации слов при прослушивании телефонных разговоров. В течение первой недели регистрировались имена существительные (500 разговоров), во вторую — глаголы (500 разговоров), в третью — прилагательные и наречия (500 разговоров). Для предлогов, союзов, местоимений и артиклей прослушивалось по 150 разговоров.

Общий объем выборки составил 79 300 словоупотреблений. Обнаружено 2822 словоформы или 2240 словарных слов.[21] Н. Френч и его соавторы впервые приводят сведения об удельном весе в словаре и речи грамматических классов слов и вообще различают омонимию частей речи. Публикуется 737 словарных слов с частотами от 3990 (впереди стоит местоимение I; в других частотных словарях ранг 1 имеет артикль the) по 5, расположенных по убыванию частот. Слова с частотой 5 представлены не все, так как здесь критерием отбора было также количество разговоров, в которых встретилось слово.

[19] Суммирование этих «частот» и даст называемую А. Робертсом величину, но это уже не будет объемом выборки.

[20] По Э. Хорну (1926).

[21] Это почти единственный из известных частотных словарей английского языка, в котором указывается объем словника как в словоформах, так и в словах. См. также словарь Ф. Дж. Шонелла (№ 20, здесь на стр. 168).

Список Н. Френча, Ч. Картера и У. Кёнига популярен не только потому, что это первый словарь разговорной речи. В связи с тем что приведенные в нем 737 слов занимают 96% всех словоупотреблений (что и естественно; в письменной речи этот процент занимают несколько тысяч словоформ), эти цифры постоянно используются как довод в пользу лингвостатистики.

13) E. L. Thorndike. A Teacher's Word Book of 20 000 Words Found Most Frequently and Widely in General Reading for Children and Young People. New York, 1931.[22]

Это расширенный вариант словаря Э. Торндайка. Приводится 20 тыс. самых «важных» с точки зрения частоты и распространенности словарных слов. Количество источников равно 200.

14) I. Lorge, E. L. Thorndike. A Semantic Count of English Words. New York, 1938.[23]

Начиная с 1934 г. Э. Торндайк и И. Лордж работают над семантическим частотным словарем английского языка. В течение шести лет несколько сотен исполнителей, получив каждый по части Большого Оксфордского словаря, отмечали в текстах длиной 5 млн словоупотреблений те слова, которые входили в данную часть словаря, регистрируя также их значения. Таким способом готовился список слов, где указывались не только их частоты, но и частоты значений.

Весь текстовой материал состоял из двух равных выборок — основной и контрольной — по 2.5 млн словоупотреблений каждая из Британской энциклопедии, журналов, учебников, романов, справочников. В начале работы 500 самых частых слов из подсчетов были исключены, однако к концу было признано необходимым восстановить частоты значений этих слов, поскольку они оказались наиболее многозначными. Восстановление упущенного оказалось сложной и трудоемкой работой, поэтому не все слова из пятисот получили полную информацию о частотах значений.

К 1944 г. был изготовлен семантический словарь для выборки в 5 млн словоупотреблений.

15) I. Lorge. The Semantic Count of the 570 Commonest English Words. New York, 1949.

В этой работе И. Лордж дает краткую историю составления семантического словаря и методику восстановления частот значений 570 слов, исключенных первоначально из подсчетов. Приводится алфавитно-частотный список этих слов с указанием при каждом слове: 1) кода (номера значения) в Оксфордском словаре; 2) абсолютной частоты слова и относительных частот каждого значения в промилях как результата деления частоты значения

[22] По Э. Торндайку и И. Лорджу (1944) и А. Робертсу.
[23] По Э. Торндайку и И. Лорджу (1944) и И. Лорджу (1949).

на частоту слова; 3) количества источников (из 29) для слова и каждого значения.

Словарь занимает около 400 страниц машинописного текста. Для слова come, например, указывается 166 значений, go — 172 значения.

16) E. L. Thorndike, I. Lorge. The Teacher's Word Book of 30 000 Words. New York, 1944.[24]

Этой книгой Э. Торндайк и И. Лордж завершают свой многолетний труд по созданию крупнейшего в мире частотного словаря. Объем выборки равен приблизительно 18 млн словоупотреблений.

Слова в их «исходной» форме расположены по алфавиту. Каждое слово имеет при себе пять индексов: первая цифра от 1 до 49 означает, сколько раз в среднем на миллион словоупотреблений встретилось это слово. Слова, встретившиеся на миллион свыше 50 раз, имеют вместо цифры индекс A, а слова с частотой 100 и более на миллион обозначены индексом AA. Следующие четыре цифры показывают абсолютные частоты слова в четырех выборках по 4.5 млн словоупотреблений каждая (из которых и составлен словарь). Это словарь Э. Торндайка (1931), журнальный словарь Л. Лорджа (опубликован не был), словарь Э. Торндайка по текстам для детей и юношества и семантический словарь.

Хотя в общем абсолютные частоты по четырем выборкам даются почти для каждого слова, некоторые слова имеют индекс *M*, что значит вхождение в число самых частых.

Словарь Э. Торндайка и И. Лорджа, оставаясь по сей день самым большим по выборке частотным словарем, не дает сведений об абсолютных частотах в общей выборке и не сообщает в с е х частот в отдельных выборках, что сильно сужает сферу его применения.

Стремление составителей сэкономить время и силы за счет исключения из анализа высокочастотных слов нередко подводит их и либо заставляет повторно проделывать работу по анализу текста, либо снижает достоверность данных о частотах, когда они даются не в фактическом, а в оцененном виде.

В книге приводится также алфавитно-частотный список около семисот самых частых слов с указанием их частот в журнальном словаре И. Лорджа и семантическом словаре Лорджа—Торндайка. Имеется алфавитный список (без частот) самых частых 500 слов и вторых 500 из полного словаря. Помещены два аналогичных списка из словаря Э. Торндайка (1931).

Работы Э. Торндайка и И. Лорджа практически монополизировали статистические исследования письменной английской речи. С 1926 по 1945 г. выходили преимущественно только их словари.[25]

[24] Неоднократно переиздавался стереотипно.
[25] Четырехъязычный словарь Э. Итон в разных изданиях не является оригинальным. Он компилирует словари Э. Торндайка, Ф. Кединга, М. Бьюкенена и Дж. Вандер Беке.

17) H. Fairbanks. The Quantitative Differentiation of Samples of Spoken Language. Psychological Monographs, 1944, vol. 56, pp. 19—38.[26]

Выборка 30 тыс. словоупотреблений, произнесенных студентом-новичком, которому предлагают дать объяснение поговоркам. Приведены только 100 самых частых словоформ.

18) H. D. Rinsland. A Basic Writing Vocabulary of Elementary School Children. New York, 1945.

Это обширное исследование активного словаря письменной речи учеников средней школы. Материалом служили школьные сочинения, письма, рассказы, стихи, экзаменационные сочинения, а также некоторое количество записей разговоров. Всего проанализировано 100 тыс. источников, выборка составила 6 012 359 словоупотреблений. Это второй по величине частотный словарь английского языка после словаря Э. Торндайка и И. Лорджа. Из 25 632 разных словоформ опубликована 14 571 с частотами не менее 3; словоформы размещены в алфавитном порядке. Указаны частоты общие и для каждой из восьми выборок, соответствующих восьми классам школы.

Словарь Х. Ринсленда один из немногих словарей, оформленных так, что его данные можно использовать и для дальнейшего статистического анализа: приводятся все абсолютные частоты (кроме 1 и 2). Алфавитный список, правда, требует некоторого времени для извлечения самых частых словоформ, но эти труды окупаются добротным материалом.

19) J. W. Black, M. Ausherman. The Vocabulary of College Students in Classroom Speeches. Columbus, 1955.[27]

Целью словаря является отобрать лексикон молодых людей студенческого возраста и оценить его по выборке, равной 288 152 словоупотреблениям из 607 выступлений 274 студентов в аудитории.

20) F. J. Schonell, I. G. Middleton, B. A. Shaw. A Study of the Oral Vocabulary of Adults. Brisbane, 1956.

Словарь составлен для усовершенствования методики обучения иммигрантов в Австралии английскому языку. Были записаны на магнитофоне интервью (в 1500 различных ситуациях с 3000 различных информантов-рабочих) общей длиной 512 647 словоупотреблений, что дало 12 611 разных словоформ и 4539 «заглавных» слов. К этим словам сводились как словоформы, так и однокоренные слова. Таким образом, количество исходных грамматических форм сильно преуменьшено.[28] Приводится 1000 самых частых

[26] По Д. Хауэсу (1966).

[27] По А. Робертсу.

[28] Отношение числа различных слов к числу разных словоформ оказывается здесь равным 0.36. По словарю разговорной речи Н. Френча этот коэффициент равен 0.79, а для письменной речи (по данным словарей пере-

«заглавных» слов в алфавитном порядке с частотами не менее 21. Под каждым словом дается общая частота слова, перечень объединяемых им словоформ и производных слов и их частоты.

21) M. West. A General Service List of English Words. London, 1953 (Eleventh impression 1969).

Частотные семантические словари Э. Торндайка и И. Лорджа (1938 и 1949 гг.) были отпечатаны на гектографе и, по-видимому, не получили широкого распространения. Кроме того, они регистрировали все значения по Оксфордскому словарю с указанием только номеров значений, без их расшифровки. Поэтому практическое использование их данных в повседневной работе преподавателя вызывает известные трудности.

От этого недостатка свободен семантический словарь, изданный М. Уэстом. Работы по составлению этого словаря велись параллельно с исследованиями Э. Торндайка и И. Лорджа и финансировались частично из одних и тех же источников. Оба исследования имели целью способствовать обучению английскому языку как иностранному и планировались в то время, когда единственным способом повышения эффективности обучения языку считался научно обоснованный отбор словаря, в данном случае опираясь на статистику употреблений слов в тексте.

В 1936 г. был опубликован предварительный список наиболее важных для усвоения слов,[29] который и лег в основу словаря М. Уэста. И. Лордж предоставил данные своего семантического списка, и в окончательном варианте М. Уэст отразил также статистику значений. В отличие от словаря И. Лорджа и Лорджа—Торндайка здесь указываются не промили значений, а проценты. Значения сгруппированы, как говорит М. Уэст, без какой-либо строгой системы. Второстепенные и малоупотребительные значения не включены. Критериями отбора слов и значений, кроме частоты, были также легкость или трудность запоминания слова в данном значении, необходимость (с методической точки зрения), стилистический уровень (предпочтение отдавалось словам и значениям, не относящимся к высокому или низкому стилям) и нейтральность с точки зрения эмоционального употребления слов и значений.

В словарь помещены 2000 слов; частоты значений определялись по выборке 5 млн словоупотреблений. Слова приводятся в алфавитном порядке. Некоторые наиболее частые слова не имеют при себе абсолютной частоты (слово the, например, совсем не получило указания на частоту и на частоту значений), либо снабжены «оцененной» частотой (частота в выборке 2.5 млн была удвоена). При не-

писки, на которые ссылается Э. Хорн, и научно-технического стиля) он находится в пределах 0.66—0.68.

[29] The Interim Report on Vocabulary Selection. London, 1936.

которых частотах имеется помета, указывающая на то, что они основаны на выборке неизвестного объема.

Словарь М. Уэста пользуется большой популярностью и переиздавался с 1953 по 1969 г. одиннадцать раз.

22) D. Howes. A Word Count of Spoken English. Journal of Verbal Learning and Verbal Behavior, vol. 5, № 6, 1966.

Д. Хауэс, заинтересованный в данных о частотах языковых явлений в разговорной речи, проанализировал выборку объемом 250 тыс. словоупотреблений, записанных на магнитофон. Общая выборка состоит из трех частей — две по 100 тыс. и одна 50 тыс. словоупотреблений. Первая образована записью 20 интервью по 5000 словоупотреблений со студентами медицинского института, вторая — таким же количеством интервью с пациентами больницы (не страдающими умственными дефектами и острыми заболеваниями). Третья, контрольная, выборка делалась по записям речи одного человека. Всего таким способом исследованы 50 записей от 41 информанта.

Разных словоформ обнаружено 9699, все они приводятся в двух алфавитно-частотных списках: первый регистрирует частоты от 2 и выше. Каждой словоформе приписаны три величины — частота в общей выборке, частота в первой и частота во второй. Во втором списке, алфавитном, приведены все словоформы с частотой 1; при каждой словоформе указан индекс 1,2 или 3 в зависимости от того, в какой выборке она встретилась.

Информанту сообщали, что данные необходимы для научного анализа речи. Экспериментатор начинал беседу вопросом о семье, специальности и т. д. по заранее составленной программе. После того как информант начинал чувствовать себя свободно, начиналась запись. Исчерпав тему или заметив скованность информанта, исследователь выключал аппарат и предлагал другую тему.

Никаких исключений из подсчетов не делалось.

Словарь Д. Хауэса является уникальным источником сведений об устной английской речи. Это понимал и автор, поспешив опубликовать списки, не дожидаясь результатов их анализа и предоставив возможность заинтересованным читателям самостоятельно исследовать эти данные.

23) H. Kučera and W. N. Francis. Computational Analysis of Present-day American English. Providence, 1967.

В отличие от составителей большинства частотных словарей Г. Кучера и У. Френсис не имели целью определение «базового» лексикона, но ставили задачей всесторонний количественный анализ репрезентативной выборки из современной английской письменной речи в США.

Объем исследованных текстов составил 1 014 232 словоупотребления; количество минимальных выборок равно 500, по 2000 словоупотреблений каждая. Все использованные тексты опубликованы в одном, 1961-м, году. Они сгруппированы по 15 жанрам художественной, научной, научно-популярной, религиозной и т. д. прозы. Тексты выбирались случайным способом по каталогам крупнейших библиотек и книжных магазинов.

Каждый текст набивался на перфокарты, а с них переносился на магнитную ленту для обработки на электронной вычислительной машине ИБМ. Запись на ленте сохраняется и позволяет обращаться к текстам сколько угодно раз. Составлены микрословари для каждого из текстов и для выборок в 10, 50, 100, 250 тыс. и 1 млн словоупотреблений.

В книге приводятся частотный и алфавитно-частотный списки всех 50 406 разных словоформ, обнаруженных в общей выборке. Этим словарь Г. Кучеры и У. Френсиса отличается от почти всех других частотных словарей английского языка, в которых обычно публикуется либо только алфавитно-частотный список (кроме словаря Р. Элдриджа, где дается полный частотный список), либо небольшая часть того и другого (Г. Дьюи, Л. Эйрз).

Таблицы распределения частот для выборок разных размеров содержат практически всю количественную информацию, необходимую для лингвостатистического анализа данных словаря.

Словарь сопровождается серьезным, хотя еще и не всесторонним (дальнейшие публикации обещаны авторами) анализом выборки. Отдельная глава описывает распределение частот словоформ; приводятся графики и таблицы зависимости «ранг—частота», сообщаются значения параметров уравнения прямолинейной регрессии для разных выборок.

Специальные главы посвящены анализу длины словоформ и предложений; выводы подкрепляются большим количеством таблиц и графиков.

В книге излагается еще одно важное статистическое исследование — анализ логарифмически нормальной модели распределения частот словоформ и возможность предсказывать объем словаря путем интерполяции и экстраполяции данных рассмотренной выборки.

* * *

Настоящий обзор охватывает, по-видимому, большинство зарубежных частотных словарей английского языка,[30] о которых имеются сведения в доступной автору литературе.

Для всех этих словарей характерно типичное в американской статистической лексикографии отсутствие учета омонимии частей речи. Служебные омонимы различаются только в семантических словарях, а омоформы в пределах одной части речи не регистри-

[30] Как уже отмечалось, словоуказатели к произведениям отдельных авторов здесь не рассматриваются.

руются нигде. За редким исключением публикуется небольшая часть словника, не приводятся сведения о всех частотах, иногда не указывается даже количество разных лексических единиц, обнаруженных в текстах. Все это значительно снижает ценность словарей как источника статистического материала.

Эти недостатки объясняются, по-видимому, тем, что составители словарей ставили перед собой сравнительно узкие, хотя и достаточно важные цели — отбор лексического минимума для обучения языку, выявление типичных ошибок при написании слов и разработку методики их исправления и предупреждения и т. п. И только когда перед лингвостатистикой встали более общие проблемы, связанные с изучением количественных характеристик языковых явлений и с применением этих данных при автоматической обработке языковой информации, выявилась потребность в частотных словарях на более строгой статистической основе. Возникла проблема организации выборки в соответствии с правилами статистики, вопрос о единице подсчетов и о необходимости выделения лексико-грамматических и грамматических омонимов (омографов).

Эти соображения принимаются во внимание участниками группы «Статистика речи» при составлении частотных словарей; изготовленные к настоящему времени в этой группе частотные словари английского языка и будут перечислены ниже.

Методика анализа текста в общем стандартна: объем выборки для каждого словаря определяется в 200 тыс. словоупотреблений при ручном исполнении. Каждый словарь охватывает узкую область (подъязык) одного из функциональных стилей английского языка. Минимальная выборка принята размером 1000 словоупотреблений из одного текста (кроме словаря публицистики). Учитывается омография частей речи и словоформ в пределах одной части речи (с некоторыми отклонениями, — например, иногда составитель считает целесообразным выделять имена существительные в препозиции к другим существительным как самостоятельный грамматический класс). Цифры, формулы и другие нелингвистические элементы из подсчетов исключаются, но в последних работах (см. настоящий сборник) они учитываются как отдельные словоформы. В большинстве случаев для каждой минимальной выборки составляется алфавитно-частотный список словоформ; в дальнейшем информация из списков переносится на карточки.

1) П. М. Алексеев. Частотный словарь английского подъязыка электроники. СР.

В выборке длиной 200 тыс. словоупотреблений обнаружены 10 582 разные словоформы. Для удобства сопоставления с данными зарубежных частотных словарей здесь приводятся словоформы без

различения омографов — 2240 с частотами не менее 10 (из 9782 разных).[31]

2) Л. А. Турыгина. Частотный словарь английских и американских газетных текстов. СР.

Выборка — 100 тыс. словоупотреблений, приведены 500 словоформ с частотами не менее 21.

3) Т. А. Микерина. Частотный словарь английского подъязыка судостроения.

Выборка — 200 тыс. словоупотреблений. Обнаружено 16 120 разных словоформ.[32]

4) К. Ф. Лукьяненков. Частотный словарь английского подъязыка судовых механизмов.

Выборка — 400 тыс. словоупотреблений. Омографы не выделялись. Словарь составлен на ЭВМ «Минск-22».[33] Разных словоформ около 13 000.

5) В. В. Колесникова. Частотный словарь английского подъязыка геологии нефти и газа.

Выборка — 200 тыс. словоупотреблений. Разных словоформ 11 848 (см. стр. 206—214).

6) А. А. Заманский. Частотный словарь английского подъязыка терапии.

Выборка — 100 тыс. словоупотреблений. Разных словоформ 9 011 (см. стр. 223—228).

7) В. М. Вагабова. Частотный словарь английского подъязыка переработки нефти и газа.

Выборка — 200 тыс. словоупотреблений. Разных словоформ 12 293 (см. стр. 197—205).

8) А. Д. Борисевич. Частотный словарь английского подъязыка строительных материалов.

Выборка — 50 тыс. словоупотреблений. Разных словоформ 4 846.

9) В. В. Гончаренко. Английский частотный словарь по полупроводникам.

Выборка 300 тыс. словоупотреблений. Разных словоформ 12 125. Словарь составлен на ЭВМ «Минск-22». Тексты подвергнуты индексации до ввода в машину (см. стр. 179—190).

[31] Полностью алфавитно-частотный словарь словоформ с различением омографии приводится в кандидатской диссертации П. М. Алексеева «Частотный словарь английского подъязыка электроники» (Рукоп. ЛГУ, 1964).

[32] Т. А. Микерина. Некоторые статистические приемы лексико-морфологического описания функционального стиля (на материале английских текстов по судостроению). КД (рукоп.), ЛГУ, 1967.

[33] К. Ф. Лукьяненков. Лексико-статистическое описание английского научно-технического стиля с помощью ЭВМ (подъязык судовых механизмов). КД (рукоп.), Белорусский гос. унив., Минск, 1969.

10) М. М. Будман. Частотный словарь английского подъязыка автомобилестроения.[34]

Выборка 200 тыс. словоупотреблений. Разных словоформ 10 941.

ПРИЛОЖЕНИЕ I

Таблица

Параметры уравнения $y=ax+b$, по данным Г. Кучеры и У. Френсиса

Длина выборки	γ (a)	lg B (b)	σ
10 тыс.	−0.8380	2.734	0.0838
50 »	−0.9345	3.567	0.0752
100 »	−1.0220	4.113	0.0792
250 »	−1.1350	4.848	0.0834
1 млн	−1.1700	5.445	0.1091
1 » (без рангов 1—10)	−1.1720	5.451	0.1087

Если рассматривать сопоставляемые выборки в равных условиях (например, равные по объему, но различные качественно, или одинаковые качественно, но различные по объему), то обнаружится, что γ зависит по крайней мере от двух величин — прежде всего от объема словаря и, во-вторых, от частоты самого частого слова. Эта общая зависимость проявляется на различном материале по-разному. В речи ребенка или в патологически ненормальной речи частоты небольшого числа слов (словоформ) очень велики, а общий набор разных слов небольшой.

Параметр γ (угол наклона графика $F=\varphi(r)$) зависит также от количества словоформ с малыми частотами и от того, какой удельный вес они имеют в выборке и словаре. По данным Г. Кучеры, при выборке (N) 2000 словоупотреблений на слова с частотой (F) 1 падает 24.55% текста и 60.77% словаря, при $N=10\,000$ словоупотреблений соответственно 19.33% и 64%, при $N=100\,000$ — 7.08% и 51.71%, при $N=250\,000$ — 4.54% и 47.97%, при $N=1\,000\,000$ — 2.25% и 44.72%.

При значительном увеличении выборки наклон будет возрастать; параметр γ возрастет тоже. Его, следовательно, можно считать показателем стиля только при сравнении выборок с о п о с т а в и м о г о объема.[35]

Применимость «закона Ципфа» обычно связывают с определенной зоной распределения ранг—частота. Утверждается, в частности,[36] что он действует только в области рангов $50 < r < 1500$, далее начинаются большие группы слов с одинаковой частотой («площадки», или «ступеньки», в графическом представлении). Возникает вопрос: какова мера величины этих площадок, или что значит «большая группа слов с одинаковой частотой»? Логарифмический масштаб для частоты 1 при выборке $N=2000$ (здесь словоформ с этой частотой 491) уделяет гораздо больше места, чем при $N=1\,000\,000$ (словоформ с частотой 1 здесь 22 543, т. е. в 46 раз больше). Отчетливо видные на графиках площадки образуются не ранее, чем при разнице в 2—3 вторых знака в значении логарифма. Для $N=2000$ это соответствует частотам 7—8 и числу словоформ при них 8 и 5, тогда как для $N=1\,000\,000$ — частоте 25 и числу словоформ с этой частотой 129. Если говорить в этом смысле о нижней границе пригодности закона Ципфа, то она для $N=2000$ находится на $F=8$, т. е. $r=32$, а для $N=100\,000$ — на $F=25$, т. е. $r=4148$. Величина эта уже намного больше, чем обычно предсказывается.

Уравнение прямой линии $y=ax+b$ рассматривается всеми, в том числе и самим Г. Ципфом, для выборок н е б о л ь ш о г о объема. В распоряжении Г. Ципфа были выборки в 44 тыс. (словарь Р. Элдриджа) и 260 тыс. (словоуказатель М. Хенли) словоупотреблений. Однако закон Ципфа применяют не только для спрямления данной эмпирической кривой по данной выборке, но и для предсказания распределения «ранг—частота» при выборке, стремящейся к бесконечности, т. е. в языке в целом или по крайней мере в генеральной совокупности текстов, однородных с теми, которые исследовались для данной выборки.

При большом увеличении выборки кривая будет не только все более увеличивать угол наклона описывающей ее прямой, но сама будет все более отличаться по внешнему виду от прямой линии. А это значит, что закон Ципфа как уравнение $y=ax+b$ перестает действовать и эмпирическое распределение ранг—частота придется описывать криволинейными графиками и соответствующими им уравнениями.

ПРИЛОЖЕНИЕ II

Используемая Дж. Кэрролом[37] методика расчета объемов словаря описана им недостаточно детально, чтобы можно было ее изложить здесь или делать суждения о ее строгости. Однако такие смелые прогнозы для выборок, увеличенных в 10 и 100 раз и даже до бесконечности, делаются впервые (табл. 1).

Вопрос о потенциальном объеме словаря увеличенных во много раз выборок является, естественно, крайне важным, но прежде

[34] Уч. зап. Калининского ГПИ им. М. И. Калинина, т. 54, Калинин, 1969.

[35] Р. М. Фрумкина. Статистические методы изучения лексики, стр. 36—37.

[36] Там же, стр. 37.

[37] См.: H. Kučera and W. N. Francis. Computational Analysis of Present-day American English. Providence, 1967.

всего в теоретическом плане. С практической точки зрения было бы не менее важным предсказать объем той части словаря, которая будет обладать достоверными частотами.

Таблица 1

Наблюденные и ожидаемые величины объема словаря для разных размеров выборки, по данным Дж. Кэррола (словарь Г. Кучеры и У. Франсиса)

Размер выборки	Объем словаря	
	наблюденный	ожидаемый
1	(1)	1
100	—	87
1999	808	1063
10 051	3009	3576
101 566	13 706	15 865
253 538	23 655	26 218
1 014 232	50 406	50 963
10 000 000	—	118 624
100 000 000	—	206 309
∞	—	340 193

Пусть достоверной считается частота $F=35$ по уравнению

$$F=\frac{z_\rho^2}{\delta^2},$$

т. е. при $Z_\rho=1.96$ (для доверительного уровня 0.95) и $\delta=0.33$. Тогда в выборке $N=2000$ с частотой $F \geqslant 35$ окажется 6 словоформ, $N=250\,000$ (т. е. при увеличении в 125 раз) их будет 796 и т. д. (ср. табл. 2). В среднем размер достоверного словника находится

Таблица 2

Количество словоформ с частотой $F \geqslant 35$, по данным частотных словарей английского языка

Словарь	Выборка	Количество словоформ с $F \geqslant 35$
Г. Кучера	2 000	6
	250 000	796
	1 000 000	3157
Электроника	50 000	165
	100 000	420
	200 000	835
Судостроение	50 000	184
	100 000	377
	200 000	686
Публицистика	100 000	289

в почти точной пропорциональной зависимости от величины выборки. Иными словами, если мы захотим получить заданный размер достоверной части словаря при увеличении выборки, мы должны умножить число достоверных словоформ в исходной выборке во столько раз, во сколько увеличиваем выборку.

При достоверной частоте $F=35$, чтобы получить удовлетворяющий нас словник объемом в 50 000 словоформ, мы должны довести выборку Г. Кучеры до 15 700 000 словоупотреблений.

В табл. 3 представлены расчеты достоверной части словаря при увеличении выборки из текстов по электронике, судостроению и др.

Таблица 3

Ожидаемый объем выборки для обеспечения заданного количества словоформ с частотой не менее 35

Для словаря	Необходимый объем словаря			
	10 000	20 000	30 000	50 000
	ожидаемый объем выборки			
Г. Кучера	3 140 000	6 280 000	9 420 000	15 700 000
Электроника	2 398 000	4 796 000	7 194 000	11 990 000
Судостроение	2 916 000	5 832 000	8 748 000	14 580 000
Публицистика	3 460 000	6 920 000	10 380 000	17 300 000

Используется уравнение $N_0=\frac{L_0}{L}\cdot N$, где N_0 — ожидаемая выборка, L_0 — заданный объем словаря с $F \geqslant 35$, L — исходный объем словаря с $F \geqslant 35$, N — исходная выборка.

ПРИЛОЖЕНИЕ III

Дополнительные сведения о частотных словарях английского языка, составленных в группе «Статистика речи» к 1971 г.

1) А. Д. Борисевич, В. С. Крисевич. Частотный словарь словоформ английского подъязыка строительных материалов. В кн.: Статистика текста, т I. БГУ, Минск, 1969.

Выборка 200 тыс. словоупотреблений. Разных словоформ 9447. Текст обработан на ЭВМ Минск-22.

2) В. А. Букович. Частотный словарь английского подъязыка электронно-вычислительной техники. В кн.: Статистика текста. . .

Выборка 100 тыс. словоупотреблений. Разных словоформ 10 185. Текст обработан на ЭВМ Минск-22.

3) Н. В. Ем. Частотный список английского подъязыка физической химии. В кн.: Статистика текста. . .

Выборка 50 тыс. словоупотреблений. Разных словоформ 5177. Текст обработан вручную.

4) Р. С. Мелик-Гусейнова. Частотный словарь английских текстов по физике твердого тела (стр. 191—196 наст. сб.).

Выборка 100 тыс. словоупотреблений. Разных словоформ 5542. Текст обработан вручную.

5) Е. С. Тарасова. Частотный словарь английских текстов по виноделию и виноградарству (стр. 215—222 наст. сб.).

Выборка 100 тыс. Текст обработан вручную.

6) М. Н. Боркун. Частотный словарь трехчленных сочетаний в подъязыке английской публицистики. В кн.: Статистика текста . . .

7) М. В. Данейко. Статистика глагольных словосочетаний в английских текстах по радиоэлектронике. В кн.: Статистика текста. . .

8) В. А. Соркина. Статистика именных словосочетаний в английских текстах по радиоэлектронике. В кн.: Статистика текста. . .

9) А. Д. Борисевич. Частотный словарь трехсловных сочетаний, полученный с помощью ЭВМ на материале английских строительных текстов. В кн.: Статистика текста. . .

10) В. А. Ноздрина. Частотный список атрибутивных сочетаний в английских текстах по радиоэлектронике. В кн.: Статистика текста. . .

11) Э. М. Добрускина. Частотный список именных словосочетаний английского подъязыка сельскохозяйственного машиностроения (в печати).

Частотный словарь наиболее употребительных словосочетаний в текстах по электронике на английском языке представлен в канд. дисс. О. А. Нехай: Статистика и автоматический анализ текста (БГУ, Минск, 1968).

Составлены и готовятся к печати частотные словари английских текстов по физике элементарных частиц, по различным разделам физики твердого тела, по кинотехнике, стоматологии, хирургии сердца.

В. В. Гончаренко

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ПОЛУПРОВОДНИКАМ

Для составления частотного словаря по полупроводникам была подобрана и обработана американская и английская научная и техническая литература, представляющая собой статьи из журналов и материалы VII Международного симпозиума по физике полупроводников. Материалы подбирались по схеме:

Наука	50%	Техника	50%
1) физика полупроводников	40%	1) технология получения и обработка полупроводниковых материалов и пленок	10%
2) материаловедение	10%	2) полупроводниковая электроника	20%
		3) схемы и устройства с применением полупроводниковых приборов	20%

На ЭВМ «Минск-22» было получено 6 частотных и 6 алфавитно-частотных словарей: по физике полупроводников на 120 тыс. словоупотреблений, по материаловедению — на 30 тыс., по технологии получения и обработке полупроводниковых материалов и пленок — на 30 тыс., по полупроводниковой электронике — на 60 тыс., по схемам и устройствам с применением полупроводниковых приборов — на 60 тыс. и общий частотный словарь. Основой словаря послужили тексты объемом 300 тыс. словоупотреблений. Прежде чем внести тексты на перфоленту, они были заиндексированы по системе, разработанной в минской группе «Статистика речи». Автором были внесены незначительные дополнения в систему индексации: глаголы be и have (во всех формах) при употреблении в качестве самостоятельных или связочных глаголов помечаются индексом «v». Эти же глаголы как вспомогательные приводятся без индекса. Индекс A (a) указывает на ат-

трибутивное, c — союзное, d — адвербиальное, g — герундиальное, S (s) — субстантивное и v — глагольное употребление словоформы. Элементы устойчивых сочетаний соединены дефисом.

В результате обработки частотного словаря был получен частотный список объемом в 12 125 словоупотреблений с покрытием около 99%.

Ниже приводится частотный список словоформ с частотой до 20.

Частотный список словоформ

i	Словоформы	F
1	the	28303
2	x	15971
3	of	12408
4	and	7317
5	in	7049
6	a	6159
7	for	3273
8	is—v	3243
9	z	2828
10	to	2742
11	is	2706
12	by	2655
13	with	2420
14	at	2293
15	this	2289
16	be	2120
17	that—c	1908
18	y	1882
19	from	1643
20	which	1585
21	on	1448
22	an	1268
23	as—d	1247
24	it	1221
25	was	1210
26	are	1203
27	can	1029
28	were	978
29	or	946
30	are—v	944
31	current	851
32	these	848
33	not	809
34	been	779
35	shown	720
36	we	687
37	used	686
38	than	678
39	region	655
40	will	643
41	voltage	631

Продолжение

i	Словоформы	F
42	has	614
43	where	592
44—45	have, to be	578
46	surface—A	553
47	high—A	543
48	if	535
49	between	532
50	temperature	526
51	two	518
52	field	501
53	when	467
54	may	465
55	also	456
56	value	447
57	all	438
58	one—numeral	431
59	was—v	428
60	observed	420
61	more	411
62	obtained	409
63	results	401
64	measurements	395
65—66	effect, given	386
67	current—A	384
68	energy—A	382
69	time	381
70	into	377
71—72	density, low	374
73	since—c	371
74—75	energy, samples	370
76—77	effects, figure	363
78	other	362
79	states	359
80	very—d	352
81	however	351
82	carrier	349
83	data	343
84	only—d	341
85	but	339
86—87	surface, then	337
88	band—A	329
89	made	326
90—91	concentration, small	324
92	experimental	320
93	about—d	318
94	electron—A	317
95	values	316
96	resistance	314
97	its	313
98	same	311
99—100	case, measured	309
101	due to	308

Продолжение

i	Словоформы	F
102—104	junction, some—A, there	307
105—106	electrons, through	304
107—108	diodes, range	303
109	shows—v	301
110	device	294
111—112	sample, silicon	288
113	such	286
114	would	284
115	circuit	280
116	no—A	276
117	thermal	275
118	carriers	274
119—120	applied, layer	273
121	X-type—A	268
122	over	267
123	structure	266
124	devices	264
125	should	262
126	characteristics	261
127	thus	258
128	each	257
129	must	255
130	frequency	253
131—133	diode, model, silicon—A	252
134—135	system, temperature—A	251
136	base—A	248
137	that	245
138	large	242
139	rate	241
140	charge	239
141	different	237
142	lower	236
143	found	234
144—145	point, under	233
146—147	band, their	231
148	magnetic	229
149	negative	225
150	transistors	224
151—152	level, they	223
153—156	as—c, dependence, emission, therefore	222
157—158	conditions, transistor	219
159	capacitance	218
160	crystals	216
161—162	determined, temperatures	215
163	higher	214
164	process	213
165—166	most, properties	212
167—168	curves, material	211
169	second—A	210
170—172	after, curve, field—A	209
173	junction—A	208
174—175	less, possible	206

Продолжение

i	Словоформы	F
176	charge—A	205
177	diffusion—A	204
178	approximately	203
179—180	equation, materials	200
181—182	conduction—A, electric	198
184—185	analysis, were—v	196
186	per	195
187	any—A	194
188	single	193
189	during	191
190	theory	190
191	has—v	189
192	voltage—A	187
193—194	oxide, potential—S	182
195—197	factor, method, noise—A	181
198	within	179
199	described	178
200	one	177
201—202	near, use	176
203	using—d	175
204—105	could, thickness	174
206	above	173
207—208	mobility, three	172
209—211	both—c, effective, type	171
212—213	our, total	170
214—215	following, function	169
216—218	parameters, various, X-type	167
219—220	both, several	165
221—222	junctions, power	164
223—226	change, diffusion, resistivity, so-that	163
227—228	distribution, thin	162
229—230	impurity—A, required	161
231—233	bulk—A, electrical, operation	160
234—236	radiation—A, techniques, time—A	159
237—238	Germanium, here	158
239—240	films, taken	157
241—242	area, those	156
243	levels	155
244	below	154
245	emitter—A	153
246—248	because, crystal, experiments	152
249—252	bias—A, number, pressure, seen	151
253—256	being, crystal—A, length, new	150
257—258	collector—A, optical	149
259—260	contact, radiation	148
261—266	bias, source, technique, times, while—c, width	147
267—269	across, first—A, paper	146
270—274	as-a-function-of, expected, increase, magnitude, much	145
275—276	holes, many	144
277—279	coefficient, gate—A, power—A	143
280—281	reported, table	142

Продолжение

i	Словоформы	F
282	constant	141
283—285	breakdown, edge, noise	140
286—288	although, calculated, spectrum	139
289	absorption—A	138
290—292	bands, doped, velocity	137
293—294	assumed, important	136
295—298	oscillations, positive, typical, volts	135
299	simple	134
300—302	circuit—A, compared, does	133
303—305	because-of, diode—A, line	131
306	detectors	130
307—308	currents, epitaxial	129
309—311	similar, usually, without	128
312—314	associated, fields, have—v	127
315—320	components, drift—A, free, known, material—s, state	126
321—324	agreement, circuits, independent, such-as	124
325—327	condition, greater, in-order	123
328	ratio	122
329—330	becomes—v, pulse—A	121
331	control	120
332—333	behavior, now	119
334—338	energies, order, side, wafer, work	118
339	impurities	117
340—344	contacts, form, necessary, result, valence—A	116
345—347	cases, emission—A, good	115
348	present—A	114
349	that-of	112
350—357	difference, discussed, efficiency, larger, layers, physical, room—A, voltages	111
358—359	constant—s, detector	110
360—363	increased, input—A, peak, regions	109
364	lines	108
365—366	changes, reverse—A	107
367—369	consentrations, device—A, either—c	106
370—372	constant—A, drain—A, film—A	105
373—374	electrode, increases—v	104
375—378	available, conductivity, occurs—v, using	103
379—381	channel, film, transitions	102
382—388	experiment, measurement, recombination—A, section, theoretical, threshold—A, well	101
389—392	absorption,, average—A, direction, produced	100
393—400	along, avalanche—A, controlled, four, output, pulse, semiconductors, solution	99
401	significant	98
402—406	considered, gap, reduced, response, signal	97
407—410	gain, linear, relatively, series	95
411—412	fixed, particles	94
413—421	base, breakdown—A, light—s, maximum—A, operating, part, substrate—s, thick, unit	93
422—425	elements, growth, methods, sample—A	92

Продолжение

i	Словоформы	F
426—431	application, depletion—A, exciton—A, mass, primary, show—v	91
432—435	active, frequency—A, hole—A, transistor—A	90
436—440	acceptor—A, processes, surfaces, upon, zero	89
441—448	certain, even—d, formed,presence, studies, to-obtain, wafers	88
449—455	diffused, equations, expression, for-example, microwave—A, previously, variation	87
456—461	corresponding, decrease, do, fact, grown, smaller	86
462—464	frequencies, points, respectively	85
465—471	conventional, first—d, hence, integrated, intensity, threshold, types	84
472—477	defects, electron, emitter, equipment, long, similar-to	83
478—480	additional critical, electronic	82
481—487	calculations, decreases—v, donor—A, position, scattering—s, structures, zone	81
488—494	appears—v, before, damage, information, instability, performance, quite	80
495—500	based, component, having, space—charge—A, spectra, zero—A	79
501—509	amplifier, edge—A, further, Germanium—A, metal—A, rates, rather, systems, vacuum—A	78
510—516	above—d, injected,maximum, output—A, problem, showed—v, terms	77
517—525	another, crystal—s, equal-to, have—v, interest, limited, minimum, peak—A, plotted	76
526—531	centers, detailed, directly, ignition—A, ohmic, phonon—A	75
532—539	channel—A, had, indicated, introduced, mechanism, potential, recombination, up-to	74
540—548	amount, intrinsic, mode, particular, proportional-to, pure, resulting, sputtering, with-respect-to	73
549—554	cannot, constants, essentially, substrates, term, whose	72
555—560	applications, dielectric, gate, gives—v, had—v, indicates—v	71
561—569	interface, lattice—A, ligth—A, nearly, pulses, so, to-determine, treatment, way	70
570—576	densities, discussion, generally, performed, sensitive, somewhat, substrate—A	69
577—580	impedance, related, saturation—A, sputtered	68
581—590	as-well-as, cent, conductance, equilibrium previous, semiconductor, set—s, valley, wave—A, wide	67
591—596	a-few, derived, ions, irradiation, particularly, shape	66
597—601	characteristic, cycle, defined, functions, multiplication	65
602—609	basic, characteristic—s, in-addition, minority—A, noted, state—A, study	64
610—617	again, conduction, design, donors, external, forward, stress, until	63

Продолжение

i	Словоформы	F
618—629	about, areas, employed, equal, exposure, infrared, mounted, size, slope, space—A, still, zone—A	62
630—640	best, biased, chemical, contamination, dependent, end, follows—v, increasing, injection, presented, variations	61
641—654	apparent, approximation, at-least, comparison, compounds, deposited, difficult, failure—A, generated, hot, longer, plot, relation, tube	60
655—659	illustrated, lifetime, pairs, pressure—A, relationship	59
660—674	a-number-of, atoms, carried-out, equivalent, experimentally, geometry, indium, phase, present, recently, resistance—A, resistivity—A, secondary, waves	58
675—685	actual, done, easily, is—m, predicted, prepared, requirements, see—v, situation, transition—A	57
685—696	considerable, containing, deposition, distance, element, enough, exposed, negligible, oxide—A, protons, represents—v, test—A	56
697—707	barrier—A, collisions, desired, dislocations, excess—A, parameter, placed, spontaneous, technology, them, useful	55
708—719	acceptors, calculation, cells, completely, group, just, manner, nature, often, phase—A, signal—A, to-produce	54
720—729	capacitance—A, center, coefficients, mechanisms, planar, portion, product, removed, sufficiently, true	53
730—744	arrangement, assumption, barrier, coil, deep, detector—A, developed, direct—A, generator, ionized, might, period, strong, sufficient, uniform—A	52
745—757	achieved, collector, control—A, diagram, interface—A, internal, lasers, procedure, relaxation—A, solid—A, special, stable, suitable	51
758—768	contact—A, flow, investigated, ionization—A, lowest, minutes, operated, relative, specimens, standard—A, tunnel—A	50
769—785	assuming—d, caused, corresponds-to, did, doping, earlier, etched, impurity, increasing—g, induced, initial—A, majority, minima, oscillation, target, upper, valleys	49
768—798	chosen, considerably, degradation, difference, flux, gave—v, heating—s, indicate—v, industry, microns, recent, transient, work—A	48
799—818	better, bound, capacitor, complete—A, deposition—A, depth, etching—g, evaporated, evidence, extremely, highly, included, influence, neutron—A, photon—A, saturation, shift, to-make, varies—v, years	47
819—835	charged, clearly, consistent, development, error, fabricated, factors, leakage—A, limit, neglected, percent, problems, proposed, reverse, source—A, to-provide, units	46

Продолжение

i	Словоформы	F
836—860	added, advantages, almost, diamond—A, electrodes, finally, formation, heat—A, heavily, identical, in-terms, ion—A, line—A, longitudinal, oxidation, phosphorus, plane, reduction, resolution, sites, speed, step, to-have, transmission—A, water	45
861—876	common, concentration—A, consists-of, details, diameter, drop, gauges, glass, injection—A, normally, pattern, readily, reasonable, wave, wavelength, what	44
877—895	antimonide, axis, configuration, designed, example, features, in-general, introduction—A, investigation, laser—A, little, net—A, observation, particle—A, parts, possibility, proper, short, strain	43
896—919	according-to, action, angle, beam, boundary—A, connected, field-effect—A, forward—A, fundamental, interaction, next, note—i, pressures, provides—v rapidly, results-in, separation, shallow, stage, type—A, uniformity, using—g, usual, variable	42
920—932	absence, always, amplitude, emitted, fast, final—A, he, leads—v, mass—A, minimum—A, production, quantum, solutions	41
933—960	advantage, after—c, aluminum, bombardment, causes—v, computed, donor, evident, followed, gas—A, hole, introduction, others, potential—v, potentials, practical, primarily, reached, resistor, resistors, selected, sequence, sligthly specimen, strength, throughout, transport—A, yields—v	40
961—981	allows—v, amplification, approach, assembly, close-to, cyclotron—A, figures, involved, lithium, major—A, open—A, orientation, photocurrent, probably, roughly, sections, sides, steps, substrate, us, varied	39
982—1007	boundary, central, charges, depends-on, distributed, drain, excited, expressed, full, further—A, glass—A, later, loss, main, mean—A, modified, on-the-other-hand, requires, spectral, splitting—S, sputtering—S, test, to-form, transverse, written	
1008—1029	accuracy, decay, degree, down, equipped, established, gas, generation, importance, laser, liquid—A, path, provided, purpose, quantities, rapid, rise, rotation, sensitivity, series—A, sum, volume	37
1030—1044	ambient, chamber, chip, clear, contribution, domains, excitons, exists—V, lead—A, metal, negative-resistance—A, prior-to, resonance, seems—V, tube—A	36
1045—1070	amplifier—A, analyzed, assumptions, attributed, back—A, bottom, by-means-of, copper, copper—A, estimated, inductance, input, itself, load, load—A, local, mentioned, process—A, produces—v, pump, sources, thickness—A, too, transition, uniform, whether	35

i	Словоформы	F
1071—1097	adjusted, against, alpha, appropriate—A, atom, bulk, changes—v, edges, etch, examination, examined, except, excitation, five, gallium, half, heating—g, heavy, how, increased—v, probability, program, relative—A, reliability, simply, spin-orbit—A, sputtering—g	34
1098—1123	able, automatic, cathode, composition, design—A, determination, direct, equivalent, evaporation, explained, general—A, immediately, masses, modulation, nitrogen, oxygen, perpendicular-to, plasma, scattering, slice, slices, subsequent, symmetry, to-measure, to-reduce, valley—A	33
1124—1155	activation—A, among, assume—v, become—v, combination, comparable, compound, correlation, detected, especially, fairly, interband—A, interesting, interpretation, ionization, layer—A, machine, models, monolithic, motion, needed, occur—v, out, pits, production—A, shall, sharp—A, steady-state—A, storage—A, tested, tests, understanding—S	32
1156—1191	acceptor, atmosphere, before—c, capacitors, compensated, consideration, coupled, determining—G, differential, entire, evaporation—A, expressions, fraction, furthermore, gold—A, heated, hydrogen, in-detail, narrow, obtain—V, only—A, onto, original, oscillator, proton—A, quantitative, radiative, reason, rise—A, slow, strongly, surface-state—A, to-give, traps, treated, vapor	31
1192—1228	accumulation—A, accurate, around, arsenide, boundaries, capacity, complex, complex—A, dipole—A, directions, drifted, frequently, furnaces, general, gradient, greatly, hours, inversion—A, irradiated, mechanical, melt—S, normalized, observations, opposite, phenomena, phonons, quartz—A, reflection—A, role, saturated, short—A, sign, specific, stability, strain—A, stream, target—A	30
1229—1264	affected, alloy—A, atomic, bridge, characterized, continuous, cut, delay—A, discharge, double, drift, ground—A, groups, ion, located, measure, measuring—G, mode—A, patterns, processing—G, rate—A, represented, requirement, resistivities, resonance—A, significantly, static, to-explain, transconductance, treatments, uniaxial, uniformly, unit—A, use—v, varying—G, water—A	29
1265—1305	alloys, antimony, away, care, cell, clean, closely, considerations, core—A, defect—A, displacement, electrically, feedback—A, flat—A, fluctuations, gaps, give—v, in-addition-to, in-fact, inside, kept, localized, losses, micron, occur—v, onset, phenomenon, point—A, polished, preparation, rectifiers, reduces—v, remaining, simplified, thermally, thin-film—A, third, together, top—A, transit—A, tungsten	28

i	Словоформы	F
1306—1349	appropriate, boron, called, collection—A, concept, conductivity—A, cross—A, decay—A, distributions, domain—A, doping—G, drawn, fabrication, his, isolated, lattice, level—A, limits, means, measuring, obvious, once, oscilloscope, parallel—A, peaks, plate, pulsed, pump—A, quality, quantity, reference—A, reflected, sometimes, square, square—A, to-increase, toward, transfer, typically, understood, vacuum, variety, wavelengths	27
1349—1392	absolute, absorbed, addition, allowed, amplifiers, anisotropy, apparatus, become—i, block, capable, carrier, close—A, decreased, depending, differential—A, doping—S, effect—A, environment, expansion, gallium—A, high-field—A, ideal, including—G, indeed, individual, initial, integral, let, likely, luminescence, maintained, matrix—A, no, perfection, phosphorus—A, resistive, set, slowly, speed—A, ten, those-of, trapping—G, visible, whereas	26
1393—1434	account, annealing—S, avalanche, bridge, contributions, degenerate, A, desirable, diffusions, dioxide, dissipation, either, electrode—A, equilibrium—A, feature, grain—A, great, helium, hour, identified, include—v, index, indium—A, investigations, leading, light, mathematical, measurement—A, of-course, polarization, radius, rectifier, reliable, research, seconds, shorter, stabilized, suggested, to-avoid, to-maintain, towards, valid, vapor	25
1345—1484	already, broad, coil—A, complicated, consequently, consider—v, content, coupling—S, current-voltage—A, description, dispersion, errors, evaluation, exactly, existence, expect—v, far, find—v, gold, heigth, highest, illustrates—v, inch, means—v, metals, mobilities, numerical, ohms, operations, optimum, plots, profile, published, remains—v, resistor—A, resulted-in—v, similarly, small-signal—A, sodium, spacing—S, spot, stored, stress—A, substantially, suggest—v, suggests—v, system—A, thereby, to-find, who	24
1485—1538	actually, alloy, alloyed, anode, assumes—v, attempt, beyond, binding, capability, cause—v, ceramic, change—v, circuitry, coupling, density—A, depleted, devided, etch—A, exhibit—v, exist—v, exponential, face, faces, flowing, function—A, furnace, heat, height—A, indirect, initially, knowledge, lightly, limiting, modes, molecular, namely, neutral, neutrons, occurred—v, off, outer, polishing, poor, produce—v, reflectivity, require—v, six, theoretically, to-eliminate, to-prevent, to-show, trapped, via, widths	23

Продолжение

i	Словоформы	F
1539—1588	acts—v, apparently, authors, bismuth, choice, computer—A, conducting, continuously, diaphragm, dimentions, dispersion—A, domain, early, etching—S, every, existing, explanation, extensive, fluoride, illumination, includes—v, increase—v, involving, latter, mask, microscope, nuclear, oxidizes, oxygen—A, papers, partial, picture, preamplifier, properly, pumps, purity, question, realized, referred-to, simultaneously, starting, to-study, tubes, twice, vervus, wafer—A, waveform, why, X-ray—A, yet	22
1589—1645	adjustable, angles, annealing—G, appreciable, as-a-result-of, bottom—A, closed, compensation, concerned, crystallographic, damage—A, demostrated, describes—v, diffraction—A, distribution—A, dry—A, efficient, effort, eliminated, elsewhere, encapsulation, evaluated, flows—v, fully, functional, growth—A, implies—v, indicated—v, inductive, interpreted, like—d, limitations, liquid, makes—v, network, origin, oxidation—A, oxides, phonon, photons, plane—S, primary—A, principal, purposes, reaches—v, separated, shapes, sheet—A, steady, testing—S, thicknesses, to-calculate, to-use, transport, unless, velocity—A, yield—A	21
1646—1703	accomplished, accurately, alloying—G, analogous, appreciably, attached, be, boron—A, briefly, conclusion, degeneracy, difficulties, dynamic, excitation—A, expansion—A, extended, extent, fabrication—A, formula, found—v, gap—A, harmonic, hydrostatic, indicating—d, instead-of, linearly, location, magnetoresistance, mainly, motor, natural, nickel, obtaining—g, permits—v, planes, pointed-out, polishing—g, processing, produced—v, pumping, qualitatively, reasonably, relativistic, reported—v, rocking, scale, second, slight, solved, spark—A, to-achieve, to-minimize, transistorized, tunneling, up, wavelength—A, zinc	20

Р. С. Мелик-Гусейнова

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ФИЗИКЕ ТВЕРДОГО ТЕЛА

Приводимый ниже (см. табл. 1) частотный словарь английского подъязыка физики твердого тела составлен на материале английских научно-технических текстов по физике твердых диэлектриков и полупроводников общим объемом в 100 тыс. словоупотреблений. При составлении частотного словаря учитывались лексико-грамматические и грамматические омографы.

В отличие от предыдущих словарей [1] при подсчете учитывались как отдельные словоформы, так и цифры, формулы, символы, абревиатуры, таблицы и рисунки.

Обследованные тексты распределены следующим образом (в % от общего числа текстов):

1) теория твердых диэлектриков и полупроводников — 9%;
2) химическая связь и кристаллические поля — 6%;
3) упругие и тепловые свойства, фононные спектры — 7%;
4) электрические и оптические свойства — 49%;
5) магнитные свойства — 21%;
6) взаимодействие излучений с твердыми диэлектриками и полупроводниками — 8%.

Распределение количества словоформ при частотах 20 см в табл. 2.

[1] См.: СР.

Таблица 1

Частотный список словоформ

i	Словоформы	F
1	the	9161
2	C*	4560
3	of	4217
4	S**	3156
5	and	2620
6	Form.***	2593
7	in *prp*	2545
8	AB****	2500
9	a	1870
10	is *l*	1286
11	for *prp*	1269
12	NP*****	1238
13	to *prp*	1184
14	by *prp*	935
15	with	880
16	at	819
17	that *cj*	707
18	to *part*	692
19	as *adv*	680
20	this	650
21	from	556
22	is *aux*	551
23	be *aux*	481
24	on *prp*	440
25	temperature	435
26	it	423
27	an	416
28	which *pron rel*	401
29	energy	400
30	are *l*	359
31	was *aux*	348
32	were *aux*	312
33	we	306
34	are *aux*	305
35	be *l*	295
36	field	293
37	can *mod*	285
38	not	256
39	these	253
40	shown	233
41	crystal *n*	212
42	results *n*	202
43	band *n*	201
44	where *adv rel*	200

* C — цифры.
** S — символы.
*** Form. — формулы.
**** AB — сокращения.
***** NP — имена собственные.

Таблица 1 (*продолжение*)

i	Словоформы	F
45	measurements	198
46—48	or, two *num*, was *l*	196
49	value *n*	195
50	values *n*	192
51	function *n*	190
52—53	between *prp*, given	189
54	used *pII*	180
55	if *cj*	178
56	been *aux*	171
57	crystals	169
58—59	electron, Fig.*	168
60	equation	165
61—62	only *adv*, than	159
63	obtained *pII*	148
64	all *a*	146
65	experimental	145
66—67	may, sample *n*	144
68	has *aux*	142
69	will *aux*	141
70	have *aux*, *pres indef*	140
71	but *cj*	139
72	case *n*	138
73	were *l*	137
74—76	conductivity, electrons, using *pI*	136
77	samples *n*	135
78	figure *n*	134
79—80	lattice, magnetic	133
81	effect *n*	131
82	also	130
83—85	about *adv*, one *num*, when *adv cj*	129
86	same *a*	125
87—88	data, density	124
89—90	measured *pII*, model *n*	118
91	curve *n*	117
92	then *adv*	116
93	small	113
94—96	coefficient *n*, due *a*, observed *pII*	112
97—98	range *n*, where *adv cj*	111
99	time *n*	110
100	surface *n*	109
101—103	since *cj*, temperatures, that *pron dem*	108
104—105	applied *pII*, there *adv*	106
106	made *pII*	105
107	carrier	104
108—109	low *a*, table *n*	103
110	region	100
111—114	high *a*, however *cj*, mobility, single *a*	99
115—116	is *not*, voltage	98
117	found *pII*	96
118—119	order *n*, terms *n*	95

* Fig. — рисунки.

Таблица 1 (*продолжение*)

i	Словоформы	F
120—121	carriers, such *a*	92
122	conduction	91
123—126	curves *n*, dependence, pressure, specimen	88
127	state *n*	87
128—129	absorption, transition	86
130	concentration	85
131—132	other *a*, structure	84
133	current *n*	83
134—136	no *pron*, states *n*, would *aux*	81
137—139	constant *n*, large *a*, more *adv*	80
140—145	determined *pII*, different, electric, into, respectively, theory	79
146—147	free *a*, material *n*	78
148—149	one *n*, under *prp*	77
150—152	as *cj*, present *a*, spin *n*	76
153	very *adv*	75
154—156	excitation, interaction, through *prp*	74
157—162	axis, calculated *pII*, crystal *a*, form *n*, some *a*, which *pron cj*	73
163—164	effects *n*, thermal	72
165—171	because, effective, fields, level *n*, number *n*, point *n*, thus	71
172—173	could, out *adv*	70
174—175	edge *n*, equations	68
176—181	assumed *pII*, each *a*, peak *n*, per, their, zero	67
182	direction	66
183—187	atoms, described *pII*, incident *a*, intensity, should *mod*	65
188—192	charge *n*, distribution, light *n*, scattering *vbl n*, wave *n*	64
193—198	change *n*, experiments *n*, impurities, optical, rate *n*, valence	63
199—202	conditions *n*, frequency, similar, therefore	62
203—204	current *a*, method	61
205—207	after *prp*, has *not*, lower *a*	60
208—211	factor, first *a*, impurity, its	59
212—214	here, points *n*, those	58
215—217	cell, possible *a*, result *n*	57
218—219	*taken*, that *pron rel*	56
220	within *prp*	55
221—225	constant *a*, general *a*, higher *a*, mass *n*, system	54
226—227	saturation, solution	53
228—234	corresponding *pI*, line *n*, must *mod*, over *prp*, recombination, they, total *a*	52
235—236	functions *n*, have *not*, *pres indef*	51
237—239	along *prp*, approximately, velocity	50
240—248	energies, our, parameters, room *n*, so *adv*, so *cj*, T,* vacuum, variation	49
249—257	behavior, emission, levels *n*, over *adv*, relaxation, scattering *pI*, symmetry, term *n*, three *num*	48

* T — таблицы.

Таблица 1 (*продолжение*)

i	Словоформы	F
258—261	gives *v*, lines *n*, magnetude, peaks *n*	47
262—269	considered *pII*, electrical, exciton, independent *a*, polarization, section *n*, theoretical, unit	46
270—272	any *pron*, spectrum, up *adv*	45
273—276	both *pron*, known *pII*, resistance, well *adv*	44
277—285	becomes, expected *pII*, following *a*, growth, part *n*, seen, sites *n*, specimens, while *cj*	43
286—295	agreement, breakdown, center *n*, concentrations, defined *pII*, expression, silicon, simple *a*, type *n*, work *n*	42
296—304	above *prp*, direct *a*, equal *a*, holes *n*, parallel *a*, properties, relative *a*, see *imp*, several *a*	41
305—311	above *adv*, analysis, dielectric *a*, limit *n*, much *adv*, phase, show *pres indef*	40
312—319	anisotropy, approximation, below *prp*, centers *n*, during, layer *n*, now *adv*, resistivity	39
320—326	both *cj*, heat *n*, materials, molecules, processes *n*, pure, technique	38
327—333	diffusion, equilibrium, fact, necessary *a*, phonon, scattering *ger*, wavelength	37
334—343	activation, always, dipoles, minima, moment, near *adv*, photocurrent *n*, radiation, treatment, volume	36
344—352	electrode, even *adv*, first *adv*, ion, position *n*, reported *pII*, thickness, traps *n*, use *n*	35
353—363	angle *n*, calculations, cases *n*, composition, discussion, exchange *n*, intrinsic, ions, loss, presence, written	34
364—373	cubic, dipole, gap, hole *n*, lead *n*, negative *a*, occurs, paper *n*, ratio, typical	33
374—387	addition, are *not*, average *a*, bands *n*, compared *pII*, comparison, condition *n*, contribution, follows, four *num*, ionized *pII*, pair *n*, semiconductor, smaller	32
388—398	does *aux*, domain, force *n*, length, linear, noise *n*, plotted *pII*, process *n*, quantum, side *n*, way	31
399—412	arrangement, associated *pII*, atom, changes *n*, containing *pI*, corresponds, grown, increases *v*, integrals, obtain *inf*, procedure, trap *n*, valley, various *a*	30
413—429	absolute, according *adv*, although, amplitude, calculation, coefficients, coupling *vbl n*, difference, diode, discussed *pII*, drift *n*, hence *adv*, maximum *n*, minimum, particular *a*, photoconductivity, plane *n*	29
430—444	atomic, carried *pII*, components, contact *n*, donor, increasing *pI*, irradiation, normal *a*, phonons, second *a*, showed *past indef*, space *n*, upon *prp*, valid, zone *n*	28
445—462	apparatus, being *aux*, *pI*, component *n*, constants, each *pron*, good *a*, less *a*, molecular, orientation, oxide, parameter, perpendicular *a*, positive *a*, power *n*, solid *n*, susceptibility, threshold, transitions	27
463—475	assumption, be *not*, either *cj*, frequencies, isotropic, junction, larger *a*, might *mod*, noted *pII*, relation, spectral, tube *n*, vector	26
476—490	appropriate *a*, avalanche, determine *inf*, essentially, example, experiment *n*, film *n*, germanium, group *n*, local *a*, luminescence, melt *n*, rather, reduced *pII*, solid *a*	25

Таблица 1 (*продолжение*)

i	Словоформы	F
491—511	acceptor, acoustic, again, completely, densities, derived *pII*, error, have *not*, *inf*, indicate *pres indef*, involved *pII*, latter, lifetime, magnetization, mean *a*, methods, neglected *pII*, previously, spherical, spontaneous, sufficiently, too	24
512—532	above *a*, across *prp*, assuming *pI*, axes, being *l*, *pI*, cannot, diffraction, find *pres indef*, generally, generation, gradient, greater, had *aux*, *past indef*, heating *ger*, hexagonal, indicated *pII*, less *adv*, potential *a*, size *n*, spins *n*, usual *a*	23
533—560	against, area, before *prp*, capacitor, certain, collisions, depth, determining *ger*, dimensions, excess, excited *pII*, feature *n*, interactions, intermediate *a*, ionization, localized *pII*, matrix, momentum, monochromator, real *a*, represents, response, separation, series *n sing*, set *n*, solutions, transmission, usually	22
561—589	additional, appears, assume *pres indef*, based *pII*, below *adv*, bridge *n*, diamond *n*, electronic, end *n*, formula, initial *a*, integration, measurement, microwave, mobilities, nature, numerical, photon, pulse *n*, respect *n*, rotation, site *n*, slightly, specific *a*, spectra, sum *n*, tensor, vapor *n*, was *not*	21

Таблица 2

Распределение количества словоформ при $F \leqslant 20$

i	m	F	i	m	F
590—609	20	20	1044—1128	85	10
610—650	41	19	1129—1233	105	9
651—692	42	18	1234—1346	113	8
693—722	30	17	1347—1478	132	7
723—759	37	16	1479—1679	201	6
760—807	48	15	1680—1917	238	5
808—859	52	14	1918—2249	332	4
860—913	54	13	2250—2750	501	3
914—967	54	12	2751—3543	793	2
968—1043	76	11	3544—5542	1999	1

В. М. Вагабова

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ПЕРЕРАБОТКЕ НЕФТИ И ГАЗА

Материалом статистического описания английских научно-технических текстов по переработке нефти и газа служили тексты из английских, американских и канадских журналов по указанной теме (всего 22 названия). Общий объем обследованного материала составляет 200 тыс. словоупотреблений.

Ниже (см. табл. 1) приводится частотный список словоформ с учетом лексико-грамматических и грамматических омографов, а также список распределения количества словоформ с абсолютными частотами $F \leqslant 20$ (см. табл. 2).

Таблица 1

Частотный список словоформ

i	Словоформы	F
1	the	13944
2	C*	8108
3	of	8025
4	and	5062
5	in *prp*	4497
6	a	3931
7	AB**	3183
8	to *prp*	2565
9	to *part*	2213
10	for	2192
11	NP***	1999
12	is *aux*	1846
13	is *l*	1515
14	by *prp*	1450
15	as *adv*	1440
16	be *aux*	1384
17	this	1382

*C — цифры.
**AB — сокращения.
***NP — имена собственные.

i	Словоформы	*F*
18	with	1361
19	from	1260
20	which *pron rel*	1203
21	at	1201
22	on *prp*	1059
23	or *cj*	1006
24	that *pron dem*	985
25	it	953
26	are *aux*	880
27	gas *n*	739
28	S*	718
29	an	705
30	are *l*	654
31	will *aux*	642
32	process, *n*	610
33	can *mod*	583
34	oil *n*	580
35	these	561
36	catalyst	542
37	that *cj*	525
38	was *aux*	504
39	plant *n*	479
40	not	450
41	has *aux*	440
42	refinery	433
43	unit	430
44	be *l*	417
45	than	416
46	been *aux*	413
47	have *aux*	398
48	temperature *n*	385
49	about *prp*	377
50	pressure	373
51	hydrogen	372
52	control *n*	364
53	may *mod*	358
54	sulphur	353
55	high *a*	348
56	other *a*	346
57	feed *n*	341
58	system	331
59	but *cj*	327
60	also *adv*	325
61	we	316
62—63	gasoline, new	312
64	product *n*	304
65	into	303
66	two	296
67	more *a*	294
68	there	291

*S — символы.

i	Словоформы	*F*
69—70	is *not*, were *aux*	286
71	fuel	284
72	some *a*	279
73	when *cj*	275
74	only *cj*	272
75	water *n*	271
76—77	Form,* its	264
78	such *a*	263
79	operation	262
80	time *n*	261
81	liquid *n*	259
82	equipment	253
83—84	operating *pI*, products *n*	242
85	capacity	236
86	units	235
87	shown	232
88	all *a*	230
89	where *cj*	229
90	production	228
91	out *adv*	227
92	if	226
93	that *pron dem*	225
94	crude *n*	214
95	first *a*	213
96—97	low *a*, then	211
98—99	now, stream *n*	210
100	through *prp*	208
101	however *cj*	206
102	type	205
103—104	data, must *v mod*	204
105	one *num*	202
106—107	rate *n*, was *l*	199
108	made *pII*	196
109—110	conditions, per	195
111—113	flow *n*, heat *n*, oils *n*	194
114—115	one *pron*, required *pII*	193
116—117	carbon, computer *n*	192
118	acid *n*	191
119	have *not*	188
120	large	185
121	table	184
122	during	183
123	most *a*	180
124—125	catalytic, used *pII*	179
126—127	reaction, total *a*	178
128	design *n*	177
129	because	176
130	processes *n*	174
131	between *prp*	173
132—134	point *n*, reactions, steam *n*	171

*Form. — формулы.

Таблица 1 (*продолжение*)

i	Словоформы	F
135	ethylene *n*	170
136	content *n*	167
137	so	166
138	they	165
139—140	many *a*, range *n*	163
141—142	has *not*, member	162
143	temperatures *n*	161
144	available	160
145	over *prp*	159
146—147	being *pI*, *aux*, fuels *n*	157
148	higher *a*	156
149	very	154
150	air *n*	153
151—152	after *prp*, small	152
153—154	column, should *aux*	148
155	plants *n*	146
156	yield *n*	145
157	vapor *n*	141
158	both *a*	139
159—162	produced *pII*, their, up *adv*, year	138
163—165	gases *n*, ratio, use *n*	137
166	possible	136
167	now	134
168	well *adv*	132
169—170	lower *a*, some *a*	131
171	thus	130
172—174	any *pron*, conversion, sulfide	129
175—176	tons, under *prp*	128
177—178	naphtha, solution	127
179—180	addition, volume	126
181	three	124
182	could	123
183—185	material *n*, would *mod*, yields *n*	122
186—187	tank *n*, value *n*	121
188—191	area, cracking *vbl n*, systems, years	120
192—193	amount *n*, petroleum	118
194—196	costs *n*, hydrocarbons, less *a*	117
197	hydrocarbon	116
198	catalysts	115
199	given	114
200—206	all *a*, before *prp*, commercial, materials, operations, problem, would *aux*	113
207—208	important, quality	112
209—210	case, several	111
211	light *a*	110
212—214	necessary, phase, storage *n*	109
215—217	chemical *a*, million, safety	108
218—219	construction, results *n*	107
220	natural	106
221—224	are *not*, fire *n*, obtained *pII*, propane	105
225—228	heavy, level *n*, percent, test, *n*	104
229—230	method, removal	103

Таблица 1 (*продолжение*)

i	Словоформы	F
231	produce *inf*	102
232	while	101
233—236	development, feedstock, normal, since *cj*	100
237—242	distillation, effect *n*, increased *pII*, maximum *a*, shows *v*, within *prp*	99
243	molecular	98
244—245	major *n*, you	97
246—247	various, work *n*	96
248—251	cost *n*, either *cj*, methane, types	95
252—253	activity, consumption	94
254—256	another *a*, cracking *pI*, weight	93
257—260	even *adv*, furnace, normally, part *n*	92
261—265	conventional, example, hydrocracking *vbl n*, removed *pII*, section	91
266—269	due *a*, increase *n*, information, line *n*	90
270—271	composition, refineries	89
272—279	demand *n*, following *pI*, further *a*, main, overhead *a*, requirements, size *n*, stock *n*	88
280—283	found *pII*, greater *a*, therefore, usually	87
284—292	cycle *n*, different, figure *n*, here, propylene, second *num*, should *mod*, stocks *n*, desired *pII*	86
293—298	approximately, generally, our, processing *pI*, relatively, residual *n*	84
299—301	basis, taken, without *prp*	83
302—305	areas, caustic, octane, problems	82
306—310	bottom *n*, energy, present *a*, recovery, separation	81
311—313	equilibrium, good *a*, reduced *pII*	80
314—319	designed *pII*, order *n*, regeneration, similar, since *cj*, tower *n*	79
320—322	application, better *a*, continuing *pI*	78
323—324	although, reformer	77
325—331	based *pII*, carried *pII*, distillate *n*, facilities, formation, recycle *n*, result *n*	76
332—336	components, factors, places *n*, refining *pI*, sample n	75
337—338	be *not*, long *a*	74
339—348	ammonia, any *a*, general *a*, had *aux*, provide *v*, service *n*, standard, still *adv*, top *n*, typical	73
349—352	end *n*, industry, much *a*, tube	72
353—358	bed, charge *n*, coke *n*, mixture, power, valve	71
359—362	added *pII*, field, tray, uses *v*	70
363—368	analysis, down *adv*, extraction, film, points *n*, tubes	69
369—377	acetylene, component, considered *pII*, corrosion, digital, output, provided *pII*, vacuum, values *n*	68
378—383	combustion, developed *pII*, fluid *n*, fractions, properties, viscosity	67
384—387	as *cj*, factor, final *a*, vessel	66
388—395	been *pII*, *aux*, include *v*, jet *n*, minimum *n*, nitrogen, oxygen, period, sources *n*	65
396—398	certain, continues *v*, loss	64
399—410	applications, cent, compared *pII*, economic, few *a*, investment, particularly, program, reactions, reduction, research *n*, step *n*	63

Таблица 1 (*продолжение*)

i	Словоформы	F
411—415	company, expected *pII*, job, little *a*, metal *n*	62
416—420	cases, compounds *n*, off *adv*, rings *n*, surface *n*	61
421—429	converted *pII*, efficiently, hydrogenation, market *n*, model *n*, recent, remove *inf*, solvent, thermal	60
430—437	average *n*, boiling *pI*, his, installed *pII*, pumps *n*, space *n*, study *n*, stay *n*	59
438—443	does *aux*, electrical, heater, sieve, them, was *v not*	58
444—450	changes *n*, diameter, directly, might *mod*, mixed *pII*, plate, up *adv*	57
451—463	alkylation, experience *n*, feedstocks, form *n*, four, function *n*, instruments, maintenance, motor *n*, often, raw *a*, reforming *vbl n*, tanks *n*	56
464—468	alumina, do *aux*, improved *pII*, operator, valves	55
469—477	additional, companies, concentration, controlled *pII*, cracked *pII*, last *a*, means *n*, methods, special	54
478—485	above *prp*, absorber, best *a*, direct *a*, hot *a*, hydrocracking *vbl n*, significant, what *pron rel*	53
486—493	according, centrifugal, densely, give *inf*, installation, more *adv*, project *n*, short	52
494—503	almost, asphalt *n*, boiler, constant *a*, crudes, estimated *pII*, fraction, mechanical, show *v*, those	51
504—512	built *pII*, cat, complete *a*, dioxide, needed *a*, pressures, removed *pII*, see *v*, signal *n*	50
513—525	again, automatic, butane, condition *n*, employed *pII*, far *adv*, he, initial *a*, most *n*, near *a*, single *a*, stage *n*, structure	49
526—540	amine, basic, completely, how *adv rel*, improvement, increase *inf*, increasing *pI*, least *n*, make *v*, some *pron*, steel *n*, treated *pII*, turbine, velocity, vessels	48
541—557	calculated *pII*, change *n*, differential *a*, done, had *past indef not*, hydrocracking *ger*, operate *v*, oxide, performance, rates *n*, refiners, scale *n*, separated *pII*, tests *n*, times *n*, too *adv*, upon *prp*	47
558—573	balance *n*, below *prp*, effluent *n*, drop *n*, engineering *pI*, including *pI*, is *mod*, mercaptan, next *a*, presence, reduce *v*, reported *pII*, specifications, suitable, techniques, what *pron interrog*	46
574—585	above *adv*, controller, course, diesel *n*, every, isobutane, meet *inf*, pipe *n*, pure, recently, run *pII*, specific	45
586—597	additive, combined *pII*, enough *adv*, ethane, fact, formed *pII*, hours, latter *n*, operated *pII*, selected *pII*, supply *n*, use *v*	44
598—614	along *prp*, butadiene, by-product, carbonyl, combination, degree, gravity, heavier *a*, levels *n*, loop, months, pipeline, portion, primarily, probably, rather *adv*, wide	43
615—637	analog *n*, aqueous, aromatics, calculations, contains, elements, equations, exchangers, future *n*, gives, great, inlet *a*, involved *pII*, large, need *n*, particular, petrochemical *n*, pilot *a*, price, regenerator, separator, sieves, until	42

Таблица 1 (*продолжение*)

i	Словоформы	F
638—651	advantage, amounts *n*, around *prp*, base *n*, benzene, depending *pI*, early *a*, fed *pII*, half, installations, laboratory, located *pII*, source, variables *n*	41
652—668	absorption, actual, features *n*, inspection, limited *pII*, machine, measured *pII*, mixtures, on *adv*, polymer, principle, simple, state *n*, such, sufficient, take *v*, though	40
669—685	analyzer, considerable, electronic, evaluation, expansion, full, growth, indicated *pII*, industrial, life, olefins, partial, plus, pump *n*, resulting *pI*, site, streams *n*	39
686—701	aromatic, condensate *n*, diagram, fresh, individual *a*, manufacture *n*, material *a*, mercaptans, middle *n*, occur *v*, synthesis, towers, trays, treatment, using *ger*, your	38
702—714	applied *pII*, comparison, consideration, consists, day, desirable, difficult, distribution, economical, equation, known, lines *n*, waste *a*	37
715—734	accident, advantages, against *prp*, common *n*, essentially, fixed *pII*, free *a*, heating *pI*, hour, hydrocracking *pI*, load *n*, lubricating *pI*, percentage, physical, reactors, set *n*, severe, shell, T,* were *v*	36
735—754	approach *n*, compressors, equivalent *n*, especially, ethylbenzene, exchange *n*, form *inf*, glycol, olefin, proper, purity, quantity, ring *n*, selectivity, set *pII*, side *n*, studies *n*, supplied *pII*, synthetic, who *pron rel*	35
755—782	achieved *pII*, American *a*, back *adv*, carbonate, charged *pII*, contact *n*, critical, determined *pII*, distillates, filter *n*, functions *n*, heated *pII*, included *pII*, just *adv*, optimum, organic, permit *v*, potential *a*, primary *a*, prior, rapid, refrigeration, regenerated *pII*, selection, stability, sulfuric, vapors *n*, zone	34
783—801	assumed *pII*, atmospheric, contain *inf*, converter, except *prp*, fire *v*, input, interest *n*, local, once, other *pron*, overall *a*, oxidation, require *inf*, result *v*, run *n*, solid *a*, terms, today	33
802—830	barrel, blend *n*, chloride, circulation, completed *pII*, concerned *a*, days, did *v aux*, equal, feet, frequency, had *mod*, keep *inf*, kerosene, location, national, over *adv*, personnel, quite, requires, savings *vbl n*, separate *a*, series, six, smaller *a*, specification, stages *n*, technique, thousands	32
831—844	code *n*, cooled *pII*, exchanger, grade *n*, group *n*, highly, meeting *n*, mole *n*, operators, pound *n*, practical, rapidly, sweetening *pI*, world *n*	31
845—869	acrilonitrile, availability, benefits *n*, by *adv*, chemicals, cracking *ger*, delivered *pII*, each *n*, domestic, follows, furnaces, hydrocracker, includes, obtain *inf*, passed *pII*, past *a*, placed *pII*, polyethylene, prevent *v*, pyrolysis, quantities, residue, secondary, services *n*, volumes	30

*T — таблицы.

Таблица 1 (*продолжение*)

i	Словоформы	*F*
870—903	alkylate, already, aluminium, becomes, big, cannot, coker, continuously, desulfurization, do *not*, expensive, extent *n*, followed *pII*, importance, indicate *v*, indicates, introduced *pII*, kept *pII*, limits *n*, lube, makeup, moles, more *n*, occurs, passes *v*, readily, reason *n*, remaining *pI*, represents, results *v*, severity, station *n*, treating *pI*, variable *n*	29
904—928	brought *pII*, capacities, chamber, chosen, close *a*, effective, explosion, facility, having *pI not*, held *pII*, less *adv*, higher *a*, nearly, noise, original *a*, others, paraffins, presented *pII*, processing *ger*, provides, question *n*, reduces *v*, speed *n*, technology, trend *n*	28
929—956	appear *l*, bulk, columns, contractor, effects *n*, feeds *n*, figures *n*, finally, fractionation, isomerization, itself, later *adv*, like *a*, methyl, numbers *n*, outlet, parts *n*, permits, position, protection, requirement, showed *past indef*, situation, sour, spent *pII*, together, useful, whether	27
957—986	able, associated *pII*, automatically, being *pI not*, bottoms *n*, complex *a*, computers, considerably, cooling *pI*, daily, described *pII*, electric, go *v*, involves, lead *n*, light *n*, longer *a*, matter *n*, mild *a*, month, only *a*, previously, p-xylene, relative, response, samples, seen, substantial, throughout *prp*, usual	26
987—1014	ability, accidents, blended *pII*, characteristics, complex *n*, damage *n*, department, easily, find *inf*, increases *v*, iron *a*, items, lean *a*, management, meter *n*, minimize *inf*, modified *pII*, persons, piping *ger*, projects *n*, refiner, salt *n*, shift *n*, split *pII*, strength, true, varies, were *l*	25
1015—1058	active, adequate, allows, alone, always, among *prp*, analyzers, avoid *v*, become *inf*, capable, coal, condensed *pII*, consumed *pII*, current *n*, detail, economically, emulsion, entirely, established *pII*, flash *n*, gasolines, immediately, leaving *pI*, losses, maintained *pII*, manner, manual, open *a*, out *prp*, paper *n*, pneumatic, polymerization, purpose, reflux, reforming *pI*, satisfactory, sequence, slightly, sometimes, successful, tables, takes, throughput, Western	24
1059—1100	achieve *inf*, action, ambient, burned *pII*, carry *inf*, central, completion, compound *n*, consider *v*, contents, control *inf*, controllers, defined *pII*, delivery, device, engineer *n*, engineers *n*, existing *pI*, finished *pII*, foam, hand *n*, length, likely, limit *n*, listed *pII*, maintain *v*, metals, modern, monoxide, performed *pII*, processed *pII*, produces, R,* related *pII*, relationship, report *n*, serve *v*, stored *pII*, thickness, transfer *n*, treating *ger*, truck	23

*R — реакции.

Таблица 1 (*продолжение*)

i	Словоформы	*F*
1101—1138	accurate, adding *ger*, additives, atmosphere, clay *n*, closed *pII*, compositions, condenser, contaminants, details, discussed *pII*, efficient, excessive, heating *ger*, improvements, index, industries, intermediate, Isomax, jobs, markets *n*, molecules, names *n*, only *adv*, oxides, possibility, purposes *n*, reasons *n*, reforming *ger*, shutdown, signals *n*, slot, steps *n*, supply *inf*, toward, united *pII*, wells *n*, zeolitic	22
1139—1191	another *a*, attractive, been *aux*, boundary, butene, by-products, centre, chains *n*, choice, cobalt *a*, commercially, connected *pII*, continue *inf*, convergence, covered *pII*, currently, developments, devices, desulphide, elemental, enters *v*, excess *n*, fairly, fires *n*, flows *v*, imports *n*, largely, need *v*, nickel *a*, observed *pII*, old, older, outside *adv*, panel, planned *pII*, previous, progress *n*, real, resistance, respect *n*, rich, room, said *past indef*, stable *a*, started *pII*, third, through *adv*, trucks, unless, us, vary *v*, way, widely	21

Таблица 2

Распределение количества словоформ при $F \leqslant 20$

i	*m*	*F*	*i*	*m*	*F*
1192—1251	60	20	2094—2265	172	10
1252—1318	67	19	2266—2445	180	9
1319—1384	66	18	2446—2699	254	8
1385—1444	60	17	2700—2948	249	7
1445—1528	84	16	2949—3288	340	6
1529—1639	111	15	3289—3776	488	5
1640—1731	92	14	3777—4423	647	4
1732—1828	97	13	4424—5333	910	3
1829—1952	124	12	5334—7217	1884	2
1953—2093	141	11	7218—12293	5076	1

В. В. Колесников

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ГЕОЛОГИИ НЕФТИ И ГАЗА

Приводимый ниже (см. табл. 1) частотный список словоформ составлен на материале обследованных английских научно-технических текстов по геологии нефти и газа общей длиной в 200 тыс. словоупотреблений.

При статистическом анализе учитывались лексико-грамматические и грамматические *омографы*.

Распределение количества словоформ по частоте $F \leqslant 20$ дано в табл. 2.

Выбор текстов производился по следующей схеме (в % от общего числа текстов):

1) происхождение нефти и природного газа и формирование залежей — 15%;
2) методы поисков и разведки — 10%;
3) нефтепромысловая геология — 10%;
4) геология месторождений — 65%.

Таблица 1

Частотный список словоформ

i	Словоформы	*F*
1	the	15138
2	of	8900
3	C*	8046
4	NP**	7767
5	and	5975
6	in *prp*	5543
7	AB***	4233
8	a	3374
9	to *prp*	2255
10	is *l*	1602

* C — цифры.
**NP — имена собственные.
***AB — сокращения.

Таблица 1 (*продолжение*)

i	Словоформы	*F*
11	to *part*	1568
12	oil *n*	1511
13	by *prp*	1454
14	on *prp*	1345
15	from	1334
16	this	1254
17	for *prp*	1206
18	with	1122
19	is *aux*	1105
20	at	1104
21	as *adv*	1047
22—23	be *aux*, it	831
24	S*	809
25	gas *n*	772
26	that *cj*	758
27	are *l*	740
28	was *aux*	736
29	or	714
30	are *aux*	705
31	which *pron rel*	682
32	an	666
33	has *aux*	660
34	area	654
35	basin	606
36	well *n*	590
37	these	557
38	field	536
39	have *aux inf*	516
40	will *aux*	506
41	been *aux*	502
42	not	494
43	were *aux*	464
44	be *l*	427
45	than	412
46	feet	410
47	but *cj*	408
48	miles	404
49	more *adv*	381
50	wells *n*	368
51	reservoir	366
52	production	352
53	sand *n*	345
54	two *num*	336
55	sandstone	334
56	may	333
57—59	can *mod*, formation, one *num*	332
60	there *adv*	311
61	water *n*	307
62	data	306
63	drilled *pI*	302

*S — символы.

Таблица 1 (*продолжение*)

i	Словоформы	F
64	exploration	296
65	its	278
66	petroleum	277
67	between *prp*	276
68	other *a*	271
69	porosity	268
70	was *l*	267
71	permeability	265
72	rocks *n*	262
73	about *adv*	261
74	only *adv*	258
75	new	251
76—78	also, discovery, fields	247
79	found *pII*	246
80	shale	244
81	depth	239
82	time *n*	236
83	lower *a*	235
84	into	234
85	areas	226
86	during	223
87—88	drilling *vbl n*, part *n*	220
89	beds *n*	219
90	section *n*	217
91	per	216
92	figure *n*	213
93—94	rock *n*, sediments	212
95	all *a*	206
96—97	development, structure	205
98	limestone *n*	204
99	they	198
100	structural	194
101—102	upper *a*, which *a*	191
103	that *pron rel*	190
104—105	both *pron*, marine *a*	189
106	first *a*	186
107—109	company, reserves *n*, some *a*	183
110	no *pron*	180
111—112	their, within *prp*	178
113	analysis	177
114	is *not*	176
115—117	some *pron*, such *a*, surface *n*	173
118	each	171
119	total *a*	170
120	would *aux*	168
121	shown	167
122—123	large *a*, present *a*	166
124—125	been *not*, year	165
126	three *num*	164
127	samples *n*	163
128	years	158
129	over *prp*	157

Таблица 1 (*продолжение*)

i	Словоформы	F
130—132	because, now *adv*, test *n*	156
133—134	major *a*, stratigraphic	155
135—138	however *cj*, if *cj*, source, used *pII*	154
139	sands *n*	153
140—141	drilling *pI*, made *pII*	152
142	through *prp*	151
143—144	has *not*, zone	150
145	should *mod*	147
146	hydrocarbons	146
147	very *adv*	145
148	have *not inf*	143
149—152	companies, high *a*, several *a*, top *n*	142
153	map *n*	140
154—156	out *adv*, same *a*, seismic	139
157	any *pron*	137
158	study *n*	135
159	table *n*	134
160—161	after *prp*, along *prp*	133
162	many *a*	131
163—164	could, work *n*	130
165—166	reef *n*, structures	129
167	small *a*	127
168—170	known *pII*, those, were *l*	126
171—173	age *n*, Cretaceous *a*, unit	125
174	formations	124
175—178	distribution, group *n*, million, producing *pI*	123
179	geological	122
180—185	fault, near *adv*, trend *n*, type, western, where *adv rel*	121
186—187	county, up *adv*	120
188	completed *pII*	118
189—192	industry, low *a*, must *mod*, that *pron dem*	115
193—197	as *cj*, deposition, generally, important, natural	114
198	possible	113
199—200	had *aux*, percent	112
201—204	number *n*, sandstones, shows *v*, thickness	111
205—207	facies *n sing*, most *n*, we	110
208—209	hole *n*, well *adv*	108
210	sedimentary	107
211—213	information, northern, sample *n*	106
214—217	above *prp*, most *a*, offshore *a*, pressure	105
218—219	basins, core *n*	104
220—221	another *pron*, conditions	103
222—225	coal *n*, dolomite, general *a*, zones	102
226—227	much *adv*, organic	101
228	location	100
229—233	being *not pI*, four *num*, geologic, similar, system	99
234—237	interest *n*, regional, results *n*, under *prp*	97
238—240	base *n*, pore *n*, then *adv*	96
241—242	different, recent	95
243—244	productive, so *adv*	94
245—247	bar *n*, content *n*, when *cj*	93
248—250	before *prp*, last *a*, pay *n*	92

Таблица 1 (*продолжение*)

i	Словоформы	F
251—255	approximately, line *n*, reservoirs, shales, thin *a*	91
256—258	accumulation, Devonian *a*, trends *n*	90
259—264	below *prp*, dry *a*, exploratory, probably, most *adv*, sea	89
265—269	all *n*, deposits *n*, less *adv*, origin, size *n*	88
270—272	gravity, great, located *pII*	87
273	operations	86
274—277	available, considered *pII*, rate *n*, since *prp*	85
278	end *n*	84
279—282	eastern, he, north *n*, produced *pII*	83
283—286	been *l*, being *aux pI*, few *n*, region	82
287—291	acre, central, crude *a*, Mississipian, volume	81
292—296	due *a*, member, shelf, southern, state *n*	80
297—301	associated *pII*, average *a*, carbon, deposited *pII*, good *a*	79
302—307	Form.,* main *a*, middle *a*, river, sequence, series *n sing*	78
308—312	be *not*, geology, interval, more *a*, second *a*	77
313—316	based *pII*, east *adv*, mile, south *adv*	76
317—321	amount *n*, are *not*, early *a*, initial *a*, properties	75
322	using *pI*	74
323—325	column, land *n*, range *n*	73
326—330	history, methods, side *n*, still *adv*, units	72
331—334	although, depositional, north *adv*, various *a*	71
335—342	individual *a*, lease *n*, material *n*, shallow *a*, techniques, tests *n*, traps *n*, while *cj*	70
343—346	discoveries, higher *a*, lake, relatively	69
347—353	discovered *pII*, equipment, even *adv*, recovery, them, waters *n*, west *n*	68
354—361	cent, deep *a*, gulf *n*, holes *n*, normal *a*, reported *pII*, result *n*, survey *n*	67
362—365	acres, barrels *n*, indicates, was *not*	66
366—372	east *n*, environment, given, least *n*, limited *pII*, our, presence	65
373—379	across *prp*, here, limestones, obtained *pII*, occur *inf*, types, wildcat	64
380—383	continental *a*, salt *n*, south *n*, taken	63
384—389	drill *inf*, off *prp*, Paleozoic *a*, problems, studies *n*, surveys *n*	62
390—394	about *prp*, far *adv*, his, period, south *a*	61
395—401	carbonate, geologists, just *adv*, Pennsylvanian, program *n*, relationship, until *cj*	60
402—411	acreage, anticline, common *a*, expected *pII*, have *aux*, *pres indef*, hydrocarbon, increase *n*, order *n*, salt *a*, standard *n*	59
412—417	developed *pII*, five *num*, indicated *pII*, local *a*, place *n*, use *n*	58
418—429	basal, control *n*, cross *a*, depths, direction, faults, geologist, northwest *adv*, original *a*, Permian *a*, up *prp*, usually	57

*Form. — формулы.

Таблица 1 (*продолжение*)

i	Словоформы	F
430—439	basis, characteristics, clay, commercial *a*, evidence *n*, factors, flank *n*, parts *n*, platform, thus	56
440—450	completion, contains, form *n*, greater, oils *n*, point *n*, pool *n*, position, thick *adv*, throughout *prp*, west *adv*	55
451—464	bottom *n*, changes *n*, counties, drilling *ger*, example, few *a*, formed *pII*, mapped *pII*, porous, primary, search *n*, subsurface *n*, tertiary, thick *a*	54
465—475	abandoned *pII*, flow *n*, government, grains *n*, had *not past indef*, half *n*, research *n*, sediment, shows *n*, success, vertical *a*	53
476—488	appears, coast *n*, consists, detailed *pII*, determined *pII*, district, further *a*, migration, nature, showing *pI*, values *n*, west *a*, where *adv cj*	52
489—493	day, north *a*, subsurface, *a*, though *cj*, toward	51
494—501	concession, containing *pI*, fact, problem, rather, recently, significant, space *n*	50
502—510	active, almost, basement, case *n*, certain, described *pII*, features *n*, fracture *n*, quartz	49
511—521	black *a*, down *adv*, heavy, logs *n*, nearly, often, operator, portion, see *inf*, southeast *adv*, strata	48
522—532	already, chemical, contact, derived *pII*, lies *v*, matter *n*, out *prp*, pattern *n*, play *n*, potential *a*, six *num*	47
533—546	accumulations, activity, addition, change *n*, energy, estimated *pII*, geophysical, grain *n*, interpretation, occurs, particular, percentage, prospects *n*, temperature	46
547—559	does *aux*, edge *n*, entire *a*, future *a*, later *a*, mud, old *a*, ratio, said *past indef*, show *inf*, slightly, value *n*, world	45
560—571	bodies *n*, defined *pII*, find *inf*, having *not pI*, including *pI*, limits *n*, long *a*, paper *n*, proved *pII*, reefs *n*, times *n*, without *prp*	44
572—584	additional, analyses *n*, axis, body *n*, fine *a*, maximum *n*, Mesozoic *a*, necessary, occurence, operation, others *n*, stratigraphy, union	43
585—598	according *pI*, better *a*, cores *n*, done, erosional, fluid *n*, is *mod*, mountain *n*, perylene, required *pII*, sedimentation, surface *a*, terms *n*, therefore	42
599—617	American *a*, big, block *n*, commonly, country, degree, deeper *a*, determine, did *aux*, either, fluids, intervals, log *n*, method, numerous, other *n*, Pacific, relative, technique	41
618—638	abundant, anticlines, billion, channel, composed *pII*, considerable, controlled *pII*, correlation, cost *n*, dip *n*, figures *n*, horizon, late *a*, likely *adv*, little *adv*, mainly, might *mod*, offshore *adv*, province, stage *n*, too	40
639—656	around *prp*, condensate, current *a*, deposit *n*, dolomitic, foot, larger *a*, light *a*, make *inf*, measured *pII*, project *n*, quite, secondary, since *cj*, storage, tectonic, third *a*, true	39
657—669	British *a*, Canadian *a*, established *pII*, exposed *pII*, following *pI*, minimum, possibly, previously, provide *inf*, separated *pII*, southwestern, underlying *pI*, way	38
670—689	carried *pII*, complete *a*, economic, have *not*, *pres indef*,	37

Таблица 1 (*продолжение*)

i	Словоформы	F
	largely, level *n*, life, overlying *pI*, rig *n*, says, show *pres indef*, trap *n*, unconformity, uplift *n*, valley, variation, were *not*, who *pron rel*, wildcats, yet	
690—708	among, amounts *n*, classification, comparison, Devonian *n*, difficult, drilled *past indef*, extensive, feature *n*, gray *a*, mineral *a*, occur *pres indef*, Paleocene, pools *n*, process *n*, produce *inf*, reservation, solution, sources	36
709—735	anomalies, anticlinal, best *a*, Cambrian *a*, Eocene, fine-grained, first *adv*, following *a*, found *past indef*, held *pII*, highly *adv*, importance, itself, Jurassic *a*, knowledge, latter, minor *a*, much *n*, next *a*, older, only *a*, particularly, potential *n*, revisions, sandy, texture, today *adv*	35
736—755	apparently, average *n*, basic, character *n*, conglomerate, crude *n*, date *n*, effect *n*, electric, indicate *pres indef*, lithologic, operators, permit *n*, previous, relationships, restricted *pII*, single *a*, site, variations, weight *n*	34
756—770	argillaceous, compared *pII*, curve *n*, encountered *pII*, factor, floor *n*, groups *n*, increased *pII*, indicate *inf*, open *a*, pressures, service, strike *n*, substantial, test *inf*	33
771—787	applied *pII*, dark, east *a*, extent, favorable, imperial *a*, maps *n*, observed *pII*, presented *pII*, prospect *n*, purpose, red *a*, residue, scale *n*, systems, whereas, younger	32
788—814	always, anomaly, called *pII*, complex *n*, continuous, course *n*, diameter, dioxide, extension, French *a*, frontier, lateral, meters, months, noted *pII*, outcrop *n*, perhaps, pigments, plugged *pII*, ranges *n*, sections *n*, shell *n*, specific, square *a*, together, what *pron interrog*, whether *cj*	31
815—835	analysed *pII*, cement *n*, characterized *pII*, drillstem, environments, evaluation, include *inf*, maximum *a*, members, own *a*, possibility, processes *n*, rapid *a*, reached *pII*, recognized *pII*, regions, report *n*, run *inf*, seen, siltstone, southwest *adv*	30
836—861	alkanes, apparent, began, capacity, dome *n*, downdip *a*, early *adv*, except *prp*, fractures *n*, geosyncline, less *a*, measurements, methane, mines *n*, past *a*, permeable, possibilities, projected *pII*, published *pII*, relation, run *pII*, selected *pII*, successful, total *n*, what *pron cj*, wide *a*	29
862—890	action, adjacent, again, agreement, closely, closure *n*, components, contain *inf*, contain *pres indef*, cored *pII*, curves *n*, domes *n*, examination, horizons, includes, locally, locations, management, Miocene, northeast *adv*, peak *n*, poor *a*, primarily, prior, quantities, seeps *n*, seven *num*, spaces *n*, subsidence	28
891—925	Australian *a*, believed *pII*, bitumen, cemented *pII*, computer, costs *n*, decrease *n*, directly, dolomites, drill *n*, earlier *a*, enough, exist *inf*, give *inf*, inorganic, interesting, isolated *pII*, lithology, mine *n*, molecular, nine *num*, nitrogen, paraffins *n*, parallel *a*, pipeline, Precambrian *a*, principal *a*, seems, separate *a*,	27

Таблица 1 (*продолжение*)

i	Словоформы	F
	sometimes, soon, southeastern, strong *a*, technical, upon *prp*	
926—956	abundance, author, be *mod*, become *inf*, cannot, cases *n*, class *n*, compounds *n*, contours *n*, contractor, conventional, cut *pII*, do *aux pres indef*, differences, discussed *pII*, earth, erosion, fairly, flowed *past indef*, forms *n*, joint *a*, laboratory, marked *pII*, matrix, modern *a*, northwest *a*, objectives, pipe *n*, rates *n*, record *n*, tons	26
957—990	abandonment, added *pII*, bed *n*, belt *n*, beyond *prp*, completely, corporation, cubic, currently, difference, divided *pII*, eight *num*, equal *a*, estimates *n*, excellent, every, firms *n*, gentle *a*, highest *a*, horizontal *a*, law, northeast *a*, northeastern, oilfield, outcrops *n*, patterns *n*, physical, Precambrian *n*, reduced *pII*, reserve *n*, resources *n*, target, typical, venture *n*	25
991—1028	absence, attempt *n*, became, buried *pII*, city, coarse, complex *a*, concessions, conducted *pII*, covered *pII*, digital, entirely, experience *n*, extreme *a*, followed *pII*, going *pI*, immediately, included *pII*, iodide, limit *n*, little *a*, massive, over *adv*, provinces, ranging *pI*, respectively, result *inf*, short *a*, showed *past indef*, smaller *a*, so *cj*, subsequent, suggests, updip *a*, varies, very *a*, vugs, working *pI*	24
1029—1065	accurate, ago, arch *n*, below *adv*, beneath *prp*, blocks *n*, calcite, carbonates, caused *pII*, collected *pII*, demand *n*, department, effects *n*, essentially, existing *pI*, extend *inf*, extends, fossil *n*, fragments, greatly, green, include *pres indef*, increasing *pI*, lack *n*, largest *a*, ocean, overlain *pII*, permeabilities, points *n*, pristane, producers, rapidly, recovered *pII*, shape *n*, tool *n*, ultimate, variety	23
1066—1104	actual, Appalachian *a*, appearance, brown *a*, burial, displacement, distributions, do *inf*, excess, filled, fraction, illustrates, increase *inf*, indicated *past indef*, injection, island, lines *n*, liquid *n*, mapping *vbl n*, materials, Oligocene, operating *pI*, Ordovician *a*, picture *n*, prepared *pII*, price *n*, producer, readily, relating *pI*, represents, ridge *n*, rigs *n*, silty, sorted *pII*, statistical, sufficient *a*, term *n*, theory, use *inf*	22
1105—1144	able, against, anthraxolite, are *mod*, becomes, bore *n*, broad *a*, brought *pII*, calcareous, clearly, compaction, construction, contract *n*, crest *n*, deltaic, density, engineers, folding *vbl n*, how *adv interrog*, involved *pII*, isopachous, lands *n*, like *adv*, minutes *n*, northwestern, overall *a*, quality, ranges *v*, reached *past indef*, recorded *pII*, represented *pII*, saturation, September, shield *n*, slope *n*, somewhat *adv*, stages *n*, suggested *pII*, vicinity, where *adv interrog*	21

Таблица 2

Распределение количества словоформ при F ⩽ 20

i	m	F	i	m	F
1145—1192	48	20	1958—2112	155	10
1193—1234	42	19	2113—2286	174	9
1235—1299	65	18	2287—2500	214	8
1300—1363	64	17	2501—2771	271	7
1364—1434	71	16	2772—3126	355	6
1435—1527	93	15	3127—3570	444	5
1528—1615	88	14	3571—4191	621	4
1616—1712	97	13	4192—5125	934	3
1713—1824	112	12	5126—6547	1422	2
1825—1957	133	11	6548—11848	5301	1

Е. С. Тарасова

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ВИНОДЕЛИЮ И ВИНОГРАДАРСТВУ

Частотный словарь составлен на материале английских текстов по виноделию и виноградарству. Были использованы тексты из журналов: «Enology and Viticulture», «Phytopathology», «Food Technology», «Vine and Wines» за 1967—1968 гг.

Общая длина обследованных текстов равна 100 тыс. словоупотреблений.

Процентное распределение текстов по тематике следующее:

1) виноделие — 56%;
2) соки — 12%;
3) обработка виноградников — 16%;
4) фитопатология — 16%.

Ниже приводится частотный список наиболее употребительных словоформ (включая частоту 10), покрывающих 83.6% всех словоупотреблений. Перечень и расшифровку сокращений см. на стр. 179—180.

Частотный список словоформ

i	Словоформы	F
1	the	7577
2	x	4748
3	of	4743
4	and	3217
5	in	2633
6	y	2004
7	a	1720
8	with	1060
9	was	1035
10	to	1033
11—12	for, were	938
13	z	881

Продолжение

i	Словоформы	F
14	by	866
15	from	790
16	at	751
17	on	601
18	this	544
19	that—c	501
20	or	472
21—22	is—v, wines	440
23	is	439
24	as—d	422
25	be	399
26	it	387
27	are	347
28	not	340
29	these	338
30	wine	318
31—32	each, used	307
33	which	302
34	was—v	301
35	acid	293
36	juice	277
37	an	265
38	been	233
39	vines	229
40	grapes	227
41	fermentation	224
42	during	216
43	two	212
44—45	all, temperature	211
46	but	202
47	total	195
48	samples	194
49	are—v	192
50—51	after, than	185
52	were—v	179
53	per	176
54	also	174
55	grape—a	173
56	time	172
57	table	171
58	other	162
59	fruit	159
60	made	154
61—62	more, varieties	153
63—64	effect, results	151
65	same	150
66—67	no, one—v	149
68—69	berries, content	143
70	that	141
71—72	has, to be	140
73—75	have, only, some	139
76—77	between, values	134

Продолжение

i	Словоформы	F
78	concentration	133
79	method	131
80—81	methods, under	130
82	treatment	129
83	there	128
84	can	127
85	determined	125
86—87	then, yeast—a	124
88—89	about—d, most	123
90—91	water, would	122
92—93	may, wine—a	121
94	into	119
95—96	acids, different	117
97	solution	116
98—99	they, through	114
100	growth	110
101	their	108
102—104	concentrate, however, three	106
105	data	105
106—108	shown, treatments, vine	104
109	color	102
110	found	101
111	days	99
112—113	if, present	98
114—116	alcohol, could, since—c	97
117—118	rate, several	96
119—120	acid—a, white	95
121—122	conditions, its	94
123—124	high, temperatures	93
125—126	produced, sample	92
127	system	91
128—129	obtained, very	90
130	amount	89
131	fermentation—a	88
132—133	had, materials	87
134	higher	86
135	added	85
136—137	initial, when	84
138	pigments	83
139—140	as, malo-lactic	82
141—142	period, yeast	81
143—145	acetoin, less, shoots	80
146—147	had—v, reported	79
148—149	apple—a, production	78
150	following	76
151—153	lower, red, study	75
154	acidity	74
155	number	73
156	first	72
157	years	71
158—161	four, quality, studies, various	70
162—164	differences, malic, paper	69

Продолжение

i	Словоформы	F
165—168	analysis, low, prepared, procedure	68
169—170	those, variety	67
171—172	both—c, will	66
173—175	heating—c, storage, weight	65
176—180	clusters, described, diacetyl, stored, sugar	64
181—182	increase, test	63
183—184	above, any	62
185—187	among, both, seedlings	61
188—193	color—a, containing, effects, possible, process, sugars	60
194—197	amounts, ethyl—a, measured, solids	59
198—203	fruit—a, many, removed, seeds, taken, we	58
204	treated	57
205—211	changes, commercial, should, significant, small, use, vinegars	56
212—214	firgure, species, times	55
215—221	concentrations, difference, greater, must—s, tartaric, to-determine, volume	54
222—225	substances, until, where	53
226—233	before, decay, density, given, lots, over, table—a, volatile	52
234—237	development, flavor, shows—v, value	51
238—244	berry—a, did, distilled, medium, reduced, series, sirup	50
245—246	being, showed—v	49
247—254	generally, grown, juice—a, product, similar, sugar—a, tartrates, work	48
255—257	fermented, using—d, usually	47
258—266	analyses, control, determination, experiments, formation, found—v, inoculated, large, relatively	46
267—268	further, tartrate	45
269—276	heated, out, pressure, stability, tests, thus, virus, without	44
277—286	because-of, especially, lastic, new, rates, reduction, second, so, solutions, storage—a	43
287—296	applied, certain, either, full, leaves, major, paper—a, range, tanks, when—c	42
297—302	hours, infected, months, nematodes, pigment, relative	41
303—315	approximately, canes, cells, discussion, drying—g, increased, levels, minutes, nitrogen, pomage, season, vineyard	40
316—327	degree, dry, factors, five, good, heat—a, maturity, phenolic, presence, standard—a, such, vineyards	39
328—337	bacteria, dioxide, much, normal, original, size, soil, soluble, type, year	38
338—351	alcohol—a, although, concentrated, containers, harvest, have—v, immediately, important, inches, later, material, must, specific, tartrate—a	37
352—362	again, appears—v, due-to, even, index, juices, liter, necessary, previously, protein—a, set—s	36
363—374	available, characteristics, control—a, dried, experimenting—g, filtered, free, known, studied, such-as, within	35

Продолжение

i	Словоформы	F
375—389	another, area, browning—g, developed, effective, filtration, mixed, mixture, noted, plus, removal, significantly, stakes, succinic, weeks	34
390—401	a-few, acetic, adjusted, held, lot, organic, parts, plants, point, products, reaction, yield	33
402—411	additional, amino, chemical, comparison, manner, optical, placed, six, sodium—a, them	32
412—433	basis, because, bloom, cluster, column, compounds, concentrates, done, equal, extract, final—a, followed, increased—v, long, prior-to, region, required, root—a, seed—a, spot, systems, tested	31
434—449	according-to, air, anthocyanins, bacterial, considered, except, fermentations, isolated, little, one, our, reagent, reducing, selected, tank, tapes	30
450—464	addition, calculated, experiment, formed, individual, measurements, moisture—a, observed, reported—v, show—v, single, to-produce, strains, whether, wood	29
465—478	controlled, grape, has—v, laboratory—a, might, oxidation, portion, potassium—a, pressure—a, proteins, regions, room—a, separated, slightly	28
479—492	air—a, allowed, capacity, day, does, early, enough, expected, green, he, modified, protein, techniques, upon	27
493—511	buds, chromatographic, compared, consisted-of—v, cultures, grapevines, influence, natural, occurred—v, operation, papers, percentage, rapid, related, still, sulfur—a, tables, water—a, well	26
512—536	absorbance, apples, as-well-as, at-least, chromatography, clear, considerable, culture, decrease, end, equipment, examined, extracts, fresh, class—a, larger, problems, recovered, relation, therefore, tissue, to-have, trained, washed, while—c	25
537—555	action, apparatus, application, average, collected, composition, continuous, crushed, damage, form, fractions, groups, importance, media, often, sawdust, submerged, technique, vacuum	24
556—581	average—a, carbon—a, carried, chromatograms, crop, expressed, fruiting, glucose, grapefruit—a, greatly, indicates—v, intervals, last, left, lines, maple—a, pads, part, presented, probably, pruning—g, rapidly, response, separation, so-that, spots	23
582—613	below, best, change, complete, constant, contained—v, decay—a, evident, extent, factor, fermenting, filter—a, foam—a, inch, loss, maximum, moisture, nearly, powder, practical, ranged—v, reactions, relationship, roots, shatter, sprays, sufficient, suggested, though, to-give, to-remove, variental	22
614—641	after—c, alcoholic, along, apparent, bleaching, considerably, contents, direct, directly, film, foam, fraction, general, growing, have—v, in-order, marked, organisms, plant, previous, procedures, quite, reconstituted, rows, solvent—s, ultraviolet, variations, young	21

i	Словоформы	*F*
642—679	active, activity, applications, bands, based, container, conversion, ethanol, fact, gas—a, group, growing—g, here, incidence, industry, just, leaf, length, mean, nitrogen—a, nylon, off, orange—a, planted, preservative—s, problem, pure, radiation, patio, surface, symptoms, ten, to-obtain, variation, yeasts, winery, yields	20
680—706	against, cluster, consistent, cooled, difficult, diluted, earlier, filled, filter, fractional-blending, harvested, in-addition, increasing, indicated—v, interest, must—a, musts, packed, particulary, residual, throughout, training—g, unit, up-to, vine—a, weighed, yellow	19
707—743	acetaldehyde, aroma—a, berry, better, buffer, case, cases, chromatogram, extracted, feet, frost, ftor—a, ground, highest, injured, maximum—a, measurement, nature, near, now, operations, particular, practices, preparation, reagents, remaining, retained, smaller, stocks, strain, tannin, test—a, unless, up, vacuum, whole, wide	18
744—794	absorption, age, almost, areas, associated, before—c, components, covered, curves, date, desirable, determining—g, diethyl—a, excess, exposure, few, gave—v, heat, indicated, investigation, laboratory, light, light—a, limited, liquid, machine, malate, means, model, optimum—a, past, periods, plastic, population, precipitation, produced, reason, recently, report, resulting, rot, slight, solids—a, subsequent, standard, sucrose, to-prevent, together, top, vigorous, viruses	17
795—839	acre, adding—g, cell—a, cent, cold, common, conducted, controls, corresponding, decreased—v, desired, determinations, direction, distillate, divided, do, every, first—d, gives—v, identical, injury, installation, microorganisms, minor, old, oxygen, potassium, quantities, rather, received, recent, recovery, resistance, seen, sherry, stage, titratable, too, using, varied—v, vigor, while, wineries, wires	16
840—882	around, behaviour, blending—g, blue, carbon, characteristic, clarity, closed, connected, employed, evaluation, for-example, gallons, grapevine—a, harvest—a, heat-exchanger, identification, included, indicate, intensity, involved, kept, late, mold—a, muscat—a, nonvolatile, phenol—a, pollen, powders, power, processing—g, packed, reds—s, responsible, satisfactory, shoot—a, short, sound—a, special, stages, sterile, tubes, twice	15
883—943	absence, affected, analytical, appeared—v, aroma, commonly, complex, concentration—a, cooling—g, correlation, dissolved, distribution, down, eluted, excessive, extraction, flask, flavor—a, flow—a, foams, frozen, great, highly, identified, inoculation, inoculum, invert, maintained, maturation, measure, measu-	14

i	Словоформы	*F*
	ring—g, mixing—g, nylon—a, once, onto, opalscent, phase, phenolics, phenols, produce—v, pruning, readily, refrigeration, resin, result, resulted-in—v, sampled, see—i, skins, somewhat, source, spurs, stainless-steel—a, stream—a, stems, strongly, temperature—a, trail, upper, vitis, widely	
944—1009	acetate, always, apparently, aqueous; as-a-result, became—v, blendblended, bloom—a, cane, casein, cellar—a, condition, continuously, counts, decreased, definite, densities, designed, enzymatic, equipped, favorable, having, head, hour, investigated, mold, mosaic, occur—v, percentages, pigment—a, place, plant—a, planting—g, plot, poor, preliminary, producing, pruned, quantitative, readings, refractometer, saturated, severe, simple, sirups, solvent, spread, stake, strong, third, thoroughly, tips, to-make, training, trellis, trellis—a, true, typical, vapor, varied, vegetative, vintage, weekly, whites—s	13
1010—1076	aging—g, already, assay, bath, bleaching—g, bottles, cannot, climatic, closely, combination, completely, completion, compound—s, consists-of—v, contact—a, derived, delution, distillation, dryness, eight, entire, established, fermenter, follows, fructose, fruits, greatest, half, hadrolysis, in-fact, in-general, itself, lactobacilli, lateral, laterals, longer, mature, methanol, numbers, of-course, others, overall, pattern, plot, portions, position, pounds, processed, produce—v, production—a, programm, rather-than, recorded, respectively, significance, spur, steam, stock, stream, toward, trials, trunk, usual, utilization, vinifera, washing—g, whereas	12
1077—1146	about, add—v, ammonia—a, anthocyanidins, appearance, appreciable, ascorbic, band, base, believed, character, charcoal, chromatographed, commercially, comparable, crushing—g, dark, dormant, dose, equilibrium, exchange, flowers, food, harvesting, how, increasing, incubated, judges, latter, leaf—a, least, line, making—g, malvidin, markedly, meter, mixture, next, none, organism, ownrooted, passed, pots, preservation, preservatives, processes, reports, ripening—g, salts, secondary, sensory, shoot, side, spectra, spotted, spray—a, stated—v, sulfate, suspension, synthetic, to-avoid, to-stand, trend, uniform—a, units, useful, variability, visual, weights, who	11
1147—1248	ablity, absorption—a, agent, aid, ammonia, apart, as—s, atmospheric, authentic, brown, capable, carbohydrates, chamber, champagne, combinations, constituents, cost, counted, crop—a, cultural, cycle, degrees, dependent, depending—g, detachment, encountered, equivalent, evaporator, evidence, except-that, expensive, experience, fluorescence, frequently, grade, his, hot, humidity, hundred, increase—i, infection, introdu-	10

Продолжение

i	Словоформы	F
	ced, investigators, irrigated, is—m, lesser, likely, mentioned, microbial, neutral, nor, normally, nutrients, obvious, occur, oil, older, oven, panel, performed, presence, pressed, properties, proposed, pulp, quantity, reach, reasons, release, removing—g, representative, respective, screen, similarly, since, sites, spectrophotometer, spores, sprayed, statistical, step, suitable, sunlight, tannin, tendency, testing—g, to-increase, to-maintain, transmission, tube, two-dimentional, using—g, valve, variables, vinegar, vinegar—a, virus, wash, way, winter	

А. А. Заманский

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ ТЕКСТОВ ПО ТЕРАПИИ

На материале английских текстов по терапии общим объемом в 100 тыс. словоупотреблений составлен приводимый ниже частотный список английского подъязыка медицины (см. табл. 1).

В табл. 2 приводится распределение количества словоформ с $F \leqslant 20$.

Материал исследования извлекался из английских и американских журналов по медицине за 1963—1968 гг. (всего 11 названий) по специально составленной схеме дозировки текстов.

При составлении частотного словаря учитывалась лексико-грамматическая и грамматическая омонимия.

Таблица 1

Частотный список словоформ

i	Словоформы	F
1	the	6949
2	of	4848
3	and	3011
4	in	2890
5	fig.	2174
6	a	1866
7	names	1398
8	with	1123
9	to *prp*	1096
10	for	718
11	was *aux*	697
12	to *part*	664
13	by	656
14	is *aux*	566
15	patients	548
16	that *cj*	539
17	or	510
18	on	467

Таблица 1 (*продолжение*)

i	Словоформы	F
19	at	465
20	as	449
21	is	442
22	an	436
23	be	425
24	from	424
25	blood	414
26	this *pron*	402
27	after	400
28	was	399
29	were *aux*	376
30	not	365
31	these	357
32	it	327
33	which	295
34	been	283
35	there *introd*	276
36	patient	270
37	but	267
38—39	normal, was *not*	265
40	disease	260
41—42	heart, per	243
43	had *not*	238
44	cases	233
45	had *aux*	231
46	cardiac	230
47	two	218
48	were *l*	216
49	during	213
50	one *num*	211
51	group	210
52	et al	208
53	have *aux*	197
54	than	193
55	infarction	187
56—57	may, pressure	182
58	are *aux*	179
59	only	171
60	table	162
61	when	161
62—63	results *n*, three	155
64	are	153
65—66	their, those	153
67	more	152
68	coronary	151
69—71	other, showed *v*, they	146
72—73	treatment, who	145
74—75	between, left *pII*	143
76	therapy	140
77	acute	139
78	this *pron*	138
79—80	each, were *not*	137

Таблица 1 (*продолжение*)

i	Словоформы	F
81	also	135
82	case	134
83—84	onset, serum	133
85	study *n*	130
86—87	rate, within	127
88—89	clinical, has *aux*	125
90—91	hours, significant	123
92	years	122
93—94	days, evidence	121
95—97	fig, is, not, such	120
98—99	mg, minutes	119
100—101	changes *n*, found *pII*	118
102—104	four, hour, we	117
105	he	116
106	right	115
107—109	levels, liver, myocardial	112
110—111	increased *pII*, time	111
112—114	any, into, some	110
115	first *num*	109
116	day	108
117	oxygen	107
118	ventricular	106
119—120	before, small	102
121	examination	101
122—125	acid, activity, age, subjects	100
126	although	99
127	she	98
128—131	arterial, because, increase *n*, obtained *pII*	97
132—135	control *n*, mean *a*, weeks, without	96
136—139	effect *n*, however, period, severe	95
140—141	groups, present	94
142—143	hospital, pulmonary	93
144	method	91
145	level	90
146	known	88
147—148	then, studies *n*	87
149	both *pron*	86
150—151	could, those	85
152—155	about, his, made *pII*, response	84
156—157	given, shown	82
158—163	can, glucose, if, output, value *n*, used *pII*	81
164—165	failure, seen	79
166—167	symptoms, Values *n*	77
168—169	as *adv*, diagnosis	76
170	following	75
171—174	aortic, measured *pII*, men, units	73
175—178	its, kg, minute, over	72
179—181	physical, total, usually	72
182—187	have *not*, plasma, renal, secretion, six, since	71
188—190	difference, insulin, taken	70
191—196	control *n*, history, less, months, range *n*, reported *pII*	69

Таблица 1 (*продолжение*)

i	Словоформы	F
197—203	associated *pII*, effects, high, hypertension, man, methods, noted *pII*	68
204—205	second, test *n*	67
206—208	first *a*, revealed *v*, same	66
209—215	artery, chest, five, shock, stenosis, trial, weight	65
216—217	rhythm, volume	64
218—222	arteries, being, block *n*, daily *adv*, diet	63
223—224	times, treated *pII*	62
225	concentration	61
226—233	area, blood-pressure, intravenous, measurements, presence, similar, tachycardia, very	60
234—238	cm., considered *pII*, findings, initial, systolic	59
239	admission	58
240—243	did, lower *a*, often, previous	57
244—247	addition, change *n*, dose *n*, low, pattern	56
248—252	abnormal, administration, condition, series, various	55
253—258	degree, determined *pII*, later *adv*, might *v*, thus, would	54
259—262	due, observed *pII*, odema, through	53
263—265	fat *n*, incidence, so	52
266—271	apex, body, bladder, blood-volume, drug, that *pron*	51
272—276	anemia, cell, chronic, mitral, possible	50
277—282	cholesterol, died *v*, fastors, flow *n*, large, studied *pII*	49
283—290	below, different, data, described *pII*, injection, medical, tests *n*, until	48
291—297	attacks, even, lysozyme, mortality, oil, raised *pII*, state *n*	47
298—304	compared *pII*, diastolic, hour, mm., peripheral, urine, well	46
305—307	potassium, slight, will	45
308—312	acids, cause *n*, factor, platelet, sinus	44
313—320	abnormalities, g., her, pressures, previously, size, therefore, whether	43
321—326	attack, developed *v*, either, much, number, use *n*	42
327—333	admitted *pII*, both *cj*, every, hearts, higher, kwashiorkor, upper	41
334—337	early *a*, function *n*, improvement, maintained *pII*	40
338—344	active, negative, positive, reduced *pII*, remaining, several, year-old	39
345—358	above, course, catheter, dietary, differences, discussion, drugs, duration, greater, had *pII*, here, next, recorded *pI*, whom	38
359—368	again, areas, carried *pII*, cerebral, death, gastric, lesions, murmur, muscle, propranolol	37
369—377	aged, among, another, cells, certain, complete *a*, occurred *v*, using, should	36
378—382	estimated *pII*, instances, report *n*, venous, significantly, specific	35
383—387	alone, arrest *n*, capacity, rise *n*, under	34
388—395	average *a*, concentrations, eight, electrocardiagram, rates *n*, twice, week, names'	33

Таблица 1 (*продолжение*)

i	Словоформы	F
396—408	amount, aneurysm, diabetes, fibrillation, least, probably, reserpine, usual, seven, show *v*, surface, technique	32
409—416	content, except, fatty, features, few, increased *v*, ischemic, women	31
417—441	according, apparent, approximately, atrial, constant, does, end *n*, fluid, found *v*, has *not*, immediately, intravenously, laboratory, litres, marked *pII*, nine, observations, pacing *ger*, performed *pII*, procedure, relatively, vessels, slightly, system, taste *n*	30
442—458	conditions, containing, dosage, doses, important, insufficiency, lung, mild, new, now, patient's, remained *v*, seems, type, vascular, whole, where	29
459—477	additional, became, determination, generally, included *pII*, infusion, marasmus, moderate particularly, pernicious, placebo, position, primary, recent, upon, ventricle, side, work *n*, year	28
478—493	arrhythmias, control, decrease, decreased *pII*, fasting *prp*, frequently, hematocrit, none, produce, ratio, regimen, signs *n*, single, stickiness, throughout, tolerance	27
494—521	adhesiveness, almost, asthma, basis, demonstrated *pII*, development, deaths, episode, first, frequency, having, hepatic, including, infection, intervals, main, patients', peak, pulse, produced *pII*, risk *n*, see, samples, shows, solution, them, third, typical	26
522—534	action, aorta, beats *n*, considerable, distribution, done, ejection, performance, periods, renogram, result *n*, view *n*, serial	25
535—559	about *prp*, always, analysis, axial, congestive, criteria, direct *a*, electrocardiographic, established *pII*, examined *pII*, fell, gave, ischemia, lead, lesion, life, must, resistance, respectively, reduction, valvular, seconds, ten, transaminase, water	24
560—582	absence, accompanying, administered *pII*, atherosclerotic, breathing *ger*, cent, circulation, conduction, count *n*, effective, frequent, gm, hypertensive, sodium, still, subsequently, systemic, tissue	23
583—612	characteristic, common, congenital, defect, determinations, diseases, dyspnea, elevated *pII*, episodes, figure, general, index, indicated *pII*, linolenic, lymphoedema, most *adv*, old, population, prolonged *pII*, receiving, recovery, relation, repeated *pII*, resuscitation, up *prp*, saturated *pII*, significance, tendency, wall, x-ray	22
613—637	blood-sugar, brain, consisted *pII*, controlled *pII*, deficit, diabetic, example, excretion, grade, great, impairment, intake, investigation, lipid, manifestations, metabolism, measurement, prior, received *v*, relationship, reports *n*, respiratory, return *n*, second *n*, suggested *pII*	21

Таблица 2

Распределение количества словоформ при $F \leqslant 20$

i	m	F	i	m	F
639—676	37	20	1348—1464	116	10
677—720	43	19	1465—1606	141	9
721—762	41	18	1607—1775	168	8
763—821	58	17	1776—1987	211	7
822—870	48	16	1988—2216	228	6
871—950	79	15	2217—2581	364	5
951—1018	67	14	2582—3087	505	4
1019—1183	164	13	3088—3816	728	3
1184—1250	66	12	3817—5071	1254	2
1251—1347	96	11	5072—9011	4039	1

М. Г. Зорсф

ЧАСТОТНЫЙ СЛОВАРЬ НЕМЕЦКИХ ТЕКСТОВ ПО ЭЛЕКТРОНИКЕ

В настоящей статье приводятся результаты статистического обследования немецких текстов по электронике, заимствованных из журналов: «Radio und Fernsehen» (Berlin), «Elektronik» (München), «Der Elektroniker» (Aarau), «Radio-Service (Zürich)», «Funkschau» (München), «Automatik» (Hamburg), «Elektronische Datenverarbeitung» (Braunschweig), «Das Elektron» (München), «Internationale elektronische Rundschau» (Berlin), за 1965—1968 гг.

Использованные источники относятся к следующим темам: 1) электронные лампы; 2) полупроводники; 3) автоматика; 4) приборы; 5) телевидение; 6) вычислительная и информативная техника.

Объем всей выборки составляет 200 текстов по 1000 словоупотреблений из каждого текста, всего извлечено 20 405 разных словоформ.

Лексико-грамматические омографы рассматривались как разные единицы частотного словаря.

Результаты показаны в частотном списке словоформ (табл. 1). Кроме того, дан список зависимости номера и частоты при $F \leqslant 19$ (табл. 2).

Функциональная зависимость между частотой словоформы и ее номером вычислялась по известной формуле:

$$F = Nk(i + \rho)^{-\gamma},$$

где F — абсолютная частота, N — объем выборки, k, ρ, γ — константы зависимости, i — ранг словоформы. Значения констант для нашей выборки следующие: $k=0.10$; $\rho=1.59$; $\gamma=0.99$.

Достоверную часть списка (начиная с $\hat{\delta} \neq 0.3$) составляют 510 первых словоформ с накопленной относительной частотой $f^*=0.634265$. На одну словоформу приходится в среднем 10.44 дв. ед. информации.

Параметры зависимости, информационные и доверительные оценки получены на ЭВМ «Минск-22».

Таблица 1

Частотный список словоформ

i	Словоформы	F	i	Словоформы	F
1	der *art*	8837	51	die *pron*	467
2	die *art*	8638	52	zum	464
3	und	5459	53	können	426
4	in	3201	54	am	425
5	von	2842	55	noch	409
6	ist	2644	56	wie *cj*	396
7	mit	2569	57	Spannung	391
8	den	2489	58	dann	390
9	des	2469	59	hat	385
10	eine	2187	60	da	375
11	bei	1991	61	Schaltung	368
12	das *art*	1960	62	wenn	352
13	für	1921	63	zwei	345
14	werden	1838	64	etwa	340
15	wird	1804	65	muss	337
16	ein	1561	66	wurde	330
17	zu *part*	1534	67	vom	325
18	auf *prp*	1426	68	beim	306
19	im	1420	69	beiden	296
20	dem	1373	70	sehr	293
21	durch	1347	71	aber	291
22	sich	1333	72	zeigt	282
23	Bild	1259	73	damit	269
24	einer	1132	74	jedoch	255
25	als	1100	75	also	260
26	sind	1051	76	liegt	247
27	an	1024	77	hier	244
28	man	1015	78	unter *prp*	234
29	dass	946	79	V	230
30	auch	941	80	diesselbe	229
31	Z (*назв. лампы*)	936	81	er	225
32	aus	886	82	während *prp*	224
33	einem	863	83	Röhre	218
34	kann	825	84	vor *prp*	212
35	nicht *part*	772	85	je	209
36	zur	768	86	Transistoren	209
37	es	740	87	Strom	208
38	oder	734	88	haben	202
39	einen	729	89	Wert	197
40	über	714	90	Transistor	195
41	nach	687	91	sein *v*	192
42	diese	666	92	dieses	190
43	so *adv*	654	93	mehr	190
44	nur *adv*	555	94	ohne	186
45	dieser	548	95	ergibt	185
46	bis	539	96	verwendet	185
47	sie	536	97	soll	181
48	um *prp*	536	98	dadurch	180
49	eines	531	99	nun *adv*	176
50	zwischen	478	100	so *cj*	176

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
101	alle	175	153	Ausgang	108
102	lässt	175	154	gibt	108
103	z. B.	175	155	Verwendung	108
104	Zeit	175	156	arbeitet	107
105	dabei	174	157	Kondensator	107
106	erreicht	167	158	Grösse	106
107	Verstärker	166	159	Aufbau	105
108	Widerstand	165	160	Betrieb	105
109	daher	163	161	Verfahren	104
110	besonders	156	162	Werte	104
111	besteht	156	163	deshalb	102
112	wieder	153	164	hohe	101
113	sowie	152	165	Form	100
114	erfolgt	151	166	Hilfe	100
115	Relais	151	167	Impulse	100
116	dies	149	168	Widerstände	100
117	Gerät	148	169	Bauelemente	99
118	Bereich	147	170	das *pron*	99
119	Schaltungen	147	171	Weise	99
120	diesen	146	172	dessen	98
121	Beispiel	143	173	Spannungen	98
122	wurden	135	174	ausserdem	97
123	ab	130	175	kleiner	97
124	Geräte	130	176	neue	97
125	deren	129	177	Technik	97
126	keine	129	178	Transistors	97
127	bereits *adv*	128	179	direkt	96
128	gegenüber *prp*	128	180	Verstärkung	96
129	Teil	127	181	wir	96
130	Röhren	126	182	Entwicklung	95
131	Frequenz	125	183	folgenden	95
132	Leistungen	122	184	jeder	95
133	Diode	120	185	Eingang	94
134	nur *cj*	120	186	ihre	94
135	sondern *cj*	120	187	Stufe	93
136	beträgt	119	188	andere	92
137	der *pron*	119	189	beide	92
138	Katode	118	190	dazu	92
139	erhält	117	191	entsprechend	92
140	immer	117	192	ersten	92
141	Eigenschaften	116	193	schon	92
142	müssen	116	194	zunächst	92
143	denen	115	195	grosse	91
144	lassen	115	196	verschiedenen	91
145	einzelnen	114	197	zwar	90
146	anderen	113	198	Dioden	89
147	ebenfalls	113	199	elektrischen	89
148	Fall	113	200	elektronischen	89
149	möglich	112	201	Frequenzen	89
150	Signal	112	202	gleichzeitig	89
151	drei	109	203	seine *pron*	89
152	wie *adv*	109	204	Basis	88

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
205	**bestimmt**	88	257	ergeben	71
206	Gerätes	88	258	erreichen	71
207	Anwendung	87	259	Kapazität	71
208	war	87	260	meist	71
209	bestimmten	86	261	Verbindung	71
210	dargestellt	86	262	Vergleich	71
211	dient	86	263	Verstärkers	71
212	Funktion	86	264	Bedeutung	70
213	hohen	85	265	bezeichnet	70
214	Impuls	85	266	Firma	70
215	Art	84	267	relativ	70
216	gegen	84	268	entspricht	69
217	grossen	84	269	Multivibrator	69
218	Messung	83	270	sei	69
219	Anzahl	82	271	somit	69
220	Daten	82	272	zugeführt	69
221	heute	82	273	ein *отл глаг.*	68
222	Schalter	81	274	d. h.	67
223	seiner *pron*	81	275	neben	67
224	Steuerung	81	276	Zustand	67
225	Verfügung	81	277	Einfluss	66
226	weitere	81	278	stellt	66
227	gleichen	80	279	vier	66
228	erst	79	280	Vorteil	66
229	liegen	79	281	allem	65
230	Potential	79	282	Anforderungen	65
231	Elektronik	78	283	beispielweise	65
232	Stereo	78	284	Fernsehen	65
233	an *отл глаг.*	77	285	kleine	65
234	grösser	77	286	kleinen	65
235	neuen	77	287	stark	65
236	Typ	77	288	Stelle	65
237	weil	77	289	verbunden	65
238	wesentlich	77	290	Anode	64
239	Abhängigkeit	76	291	Anordnung	64
240	erzeugt	76	292	ganz	64
241	Reihe	76	293	gilt	64
242	wegen	76	294	jeweils	64
243	erforderlich	75	295	positiven	64
244	Gleichspannung	75	296	sollen	64
245	infolge	75	297	was *pron*	64
246	Sender	75	298	Wechselspannung	64
247	Thyristor	75	299	bekannt	63
248	einige	74	300	beschrieben	63
249	Thyristoren	74	301	ihrer	63
250	elektrische	73	302	Kanal	63
251	folgende	73	303	Radio	63
252	Ausgangspannung	72	304	Stufen	63
253	elektronische	72	305	allgemeinen	62
254	Prinzip	72	306	Darstellung	62
255	Abb.	71	307	Elektronen	62
256	bleibt	71	308	Energie	62

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
309	etwas	62	361	Anlage	55
310	geht	62	362	bestimmte	55
311	gross	62	363	C.	55
312	Masse	62	364	Eingangswiderstand	55
313	parallel	62	365	entwickelt	55
314	allen	61	366	geeignet	55
315	Aufwand	61	367	Kondensatoren	55
316	Genauigkeit	61	368	Kurve	55
317	gleich *adv*	61	369	Null	55
318	jedem	61	370	System	55
319	oft	61	371	verschiedene	55
320	praktisch	61	372	Zusammenhang	55
321	Signale	61	373	auf *отл глаг.*	54
322	solchen	61	374	Grenzfrequenz	54
323	Anwendungen	60	375	grosser	54
324	gleiche	60	376	heisst	54
325	hoher	60	377	maximal	54
326	mm.	60	378	mehrere	54
327	Tabelle	60	379	weiter	54
328	zweite	60	380	Amplitude	53
329	zweiten	60	381	Anschluss	53
330	Abmessungen	59	382	Feld	53
331	angegeben	59	383	Gleichrichter	53
332	Anzeige	59	384	Motor	53
333	erste	59	385	proportional	53
334	Gebiet	59	386	Silizium	53
335	geschaltet	59	387	Temperatur	53
336	Länge	59	388	weniger	53
337	liefert	59	389	Aufgabe	52
338	negativen	59	390	gesteuert	52
339	sowohl	59	391	gezeigt	52
340	steht	59	392	Gl.	52
341	Filter	58	393	innerhalb	52
342	gegeben	58	394	kein	52
343	genau	58	395	KHz	52
344	hierbei *adv*	58	396	Typen	52
345	Hz.	58	397	Emitter	51
346	Möglichkeit	58	398	gesperrt	51
347	negative	58	399	Gitter	51
348	positive	58	400	ihr	51
349	sein *pron*	58	401	Punkt	51
350	solche	58	402	technische	51
351	bisher *adv*	57	403	wirkt	51
352	kommt	57	404	besitzt	50
353	Berechnung	56	405	eingestellt	50
354	Empfindlichkeit	56	406	Frequenzbereich	50
355	Höhe	56	407	gewählt	50
356	jetzt	56	408	Messungen	50
357	klein	56	409	MHZ	50
358	möglich *adv*	56	410	nämlich *adv*	50
359	Möglichkeiten	56	411	Oszillografen	50
360	Zählen	56	412	Richtung	50

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
413	Verlustleistung	50	465	setzt	46
414	ausgeführt	49	466	wo	46
415	bedeutet	49	467	aller	45
416	führt	49	468	dort	45
417	Leistungen	49	469	dritte	45
418	links	49	470	Durchmesser	45
419	notwendig	49	471	eingesetzt	45
420	Programm	49	472	jede	45
421	stehen	49	473	Lage	45
422	technischen	49	474	letzten	45
423	UKV	49	475	rechts	45
424	verwendeten	49	476	sollte	45
425	vorhanden	49	477	Verlauf	45
426	Änderung	48	478	zusätzliche	45
427	dafür *adv*	48	479	ausser	44
428	dagegen	48	480	befindet	44
429	dar	48	481	begrenzt	44
430	darf	48	482	darauf	44
431	Dimensionierung	48	483	entsprechenden	44
432	Fehler	48	484	erkennen	44
433	Industrie	48	485	Innenwiderstand	44
434	Jahren	48	486	Praxis	44
435	Lebensdauer	48	487	seit *prp*	44
436	Regelung	48	488	unabhängig	44
437	Spannungsteiler	48	489	Verhältnis	44
438	stets	48	490	Abstand	43
439	tritt	48	491	Aufzeichnung	43
440	üblichen	48	492	Belastung	43
441	allerdings	47	493	einfach	43
442	ändert	47	494	fliesst	43
443	benötigt	47	495	Informationen	43
444	einmal	47	496	Kollektor	43
445	Empfänger	47	497	Steuerspannung	43
446	genannt	47	498	zusammen	43
447	höhere	47	499	besondere	42
448	Parameter	47	500	Einsatz	42
449	Vorteile	47	501	Erhöhung	42
450	weit	47	502	Flip-Flop	42
451	welche	47	503	gelangt	42
452	Zahl	47	504	gerade	42
453	Zündung	47	505	Grund	42
454	angeschlossen	46	506	Heizspannung	42
455	Anodenspannung	46	507	insbesondere	42
456	auftreten	46	508	selbst *pron*	42
457	Ausführung	46	509	verwenden	42
458	benutzt	46	510	wesentlichen	42
459	doch *cj*	46	511	andererseits	41
460	einfache	46	512	bekannten	41
461	Frage	46	513	erfolgen	41
462	Hand	46	514	fast	41
463	Potentiometer	46	515	Folge	41
464	Schwierigkeiten	46	516	gemessen	41

i	Словоформы	F	i	Словоформы	F
517	geringe	41	569	ermöglicht	37
518	Gleichstrom	41	570	ferner	37
519	Grenzen	41	571	folgt	37
520	konnte	41	572	grundsätzlich	37
521	Kontakt	41	573	Heft	37
522	kurz *adv*	41	574	Jahr	37
523	Lösung	41	575	Schicht	37
524	mittels	41	576	steigt	37
525	Modell	41	577	Übergang	37
526	Wahl	41	578	u. s. w.	37
527	ab *prp*	40	579	vermeiden	37
528	Arbeit	40	580	W.	37
529	ebenso *adv*	40	581	waren	37
530	einfachen	40	582	Wicklung	37
531	gelegt	40	583	Wirkunsweise	37
532	Geräten	40	584	abhängig	36
533	Kennlinie	40	585	Anlagen	36
534	mechanische	40	586	ausschliesslich	36
535	schliesslich	40	587	Bildröhre	36
536	Vorgänge	40	588	Ende	36
537	wäre	40	589	finden	36
538	behandelt	39	590	gleich *adv*	36
539	bringt	39	591	gute *a*	36
540	entsprechende	39	592	hängt	36
541	erforderliche	39	593	her	36
542	erhalten	39	594	Konstant	36
543	Falle	39	595	Lautsprecher	36
544	Fällen	39	596	möglichst *adv*	36
545	geringen	39	597	selbst *part*	36
546	grössere	39	598	sogenannten	36
547	ihren	39	599	Stunden	36
548	magnetische	39	600	viel	36
549	schaltet	39	601	viele	36
550	Weg	39	602	würde	36
551	Wiedergabe	39	603	bewirkt	35
552	Wirkungsgrad	39	604	digitalen	35
553	Zeichen	38	605	enthalten	35
554	Bauelementen	38	606	entweder	35
555	Betriebspannung	38	607	Material	35
556	digitale	38	608	mechanischen	35
557	Erzeugung	38	609	recht	35
558	führen	38	610	Regler	35
559	gesamte	38	611	steuert	35
560	Grössen	38	612	verstärkt	35
561	Halbleiter	38	613	vielen	35
562	höheren	38	614	wiederum	35
563	unmittelbar	38	615	zeigen	35
564	worden	38	616	Zuverlässigkeit	35
565	beginnt	37	617	Bausteine	34
566	betrachtet	37	618	bestehen	34
567	Einstellung	37	619	Eingangspannung	34
568	erhöht	37	620	elektrisch	34

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
621	entstehen	34	673	normalen	32
622	Gehäuse	34	674	Pal	32
623	grösseren	34	675	sonst	32
624	gut *a*	34	676	Sperrspannung	32
625	häufig	34	677	wodurch	32
626	Lastwiderstand	34	678	Zeile	32
627	machen	34	679	Zündspannung	32
628	Probleme	34	680	Band	31
629	schnell	34	681	beschriebene	31
630	übertragen	34	682	bilden	31
631	weiteren	34	683	Dämpfung	31
632	wichtigsten	34	684	einfacher	31
633	Widerstandes	34	685	Einführung	31
634	aus *om глаг.*	33	686	erfüllt	31
635	bezogen	33	687	Faktor	31
636	gelten	33	688	gesamten	31
637	höher	33	689	hergestellt	31
638	Konstruktion	33	690	Hersteller	31
639	lange	33	691	kleinen	31
640	liegenden	33	692	leicht *adv*	31
641	mittleren	33	693	Leitungen	31
642	nimmt	33	694	Licht	31
643	ob *cj*	33	695	m. A.	31
644	pro *prp*	33	696	Seite	31
645	sogennante	33	697	sogar *adv*	31
646	solcher	33	698	Untersuchungen	31
647	Speicher	33	699	Vorgang	31
648	Stellung	33	700	weitgehend	31
649	völlig	33	701	angeordnet	30
650	Widerständen	33	702	beider	30
651	zeigte	33	703	berücksichtigt	30
652	zurück	33	704	daraus	30
653	zusätzlich	33	705	Ebene	30
654	Abschnitt	32	706	elektronischer	30
655	automatisch	32	707	fällt	30
656	beschriebenen	32	708	gebildet	30
657	betrieben	32	709	geeignete	30
658	bietet	32	710	Gegensatz	30
659	Decoder	32	711	genügend	30
660	digital	32	712	Gründen	30
661	einigen	32	713	hoch *adv*	30
662	Elemente	32	714	jedes	30
663	erforderlichen	32	715	Leitung	30
664	Ergebnisse	32	716	Magnetfeld	30
665	Gegenkopplung	32	717	mindestens	30
666	gut *adv*	32	718	möglichst *adv*	30
667	handelt	32	719	Oszillator	30
668	inden *cj*	32	720	Schaltkreise	30
669	Kreis	32	721	Speisespannung	30
670	macht	32	722	Stand	30
671	meisten	32	723	Temperatur	30
672	nahezu	32	724	treten	30

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
725	Vorraussetzung	30	777	Antennen	27
726	Ablenkung	29	778	Bandbreite	27
727	analog	29	779	Batterie	27
728	ausgelegt	29	780	befinden	27
729	beeinflusst	29	781	cm.	27
730	bildet	29	782	durchgeführt	27
731	eignet	29	783	geführt	27
732	erscheint	29	784	gehalten	27
733	erzeugen	29	785	geliefert	27
734	ganze	29	786	gemäss	27
735	genannten	29	787	hinter	27
736	Geschwindigkeit	29	788	könnte	27
737	Hannover	29	789	Lampe	27
738	Herstellung	29	790	leitend	27
739	Impulsen	29	791	Linie	27
740	kommen	29	792	Mitte	27
741	leicht *adv*	29	793	näher	27
742	magnetischen	29	794	oben	27
743	nächsten	29	795	spricht	27
744	Oberfläche	29	796	Übertragung	27
745	positiv *adv*	29	797	uns	27
746	Regelstrecke	29	798	Abgriff	26
747	Schirm	29	799	anderer	26
748	Serie	29	800	Änderungen	26
749	Signals	29	801	annähernd	26
750	solange	29	802	Ansteuerung	26
751	Ströme	29	803	Beitrag	26
752	Stromversorgung	29	804	besitzen	26
753	Werten	29	805	braucht	26
754	Analogrechner	28	806	Brücke	26
755	automatische	28	807	denn *cj*	26
756	berechnet	28	808	Fertigung	26
757	Bestimmung	28	809	fest	26
758	bestückt	28	810	Fläche	26
759	Drehzahl	28	811	Forderungen	26
760	Elektroden	28	812	gelangen	26
761	Endstufe	28	813	gemacht	26
762	findet	28	814	Grenze	26
763	gering	28	815	integrierten	26
764	Grössenordnung	28	816	jeweiligen	26
765	Halbwelle	28	817	Kapazitäten	26
766	Induktivität	28	818	negativ *adv*	26
767	Information	28	819	Steuergerät	26
768	Kombination	28	820	üblich	26
769	Ladung	28	821	versehen	26
770	Mikrofon	28	822	vorliegenden	26
771	natürlich	28	823	Wege	26
772	rund	28	824	Zenerdiode	26
773	Spannungsabfall	28	825	Arbeitsweise	25
774	Winkel	28	826	auftritt	25
775	angezeigt	27	827	Bedingung	25
776	Antenne	27	828	beeinflussen	25

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
829	beliebig	25	881	gewünschten	24
830	Berücksichtigung	25	882	höherer	24
831	Einbau	25	883	ihnen	24
832	einerseits	25	884	Kern	24
833	eingegangen	25	885	lediglich	24
834	eingestellten	25	886	legt	24
835	einstellbar	25	887	liefern	24
836	erwähnt	25	888	liegende	24
837	erzielt	25	889	Mass	24
838	Frequenzgang	25	890	meistens	24
839	getrennt	25	891	mittlere	24
840	Gewicht	25	892	Netz	24
841	Gruppe	25	893	neuer	24
842	hierfür *adv*	25	894	Phase	24
843	Integration	25	895	Problem	24
844	kaum *adv*	25	896	Punkte	24
845	logischen	25	897	Rechner	24
846	«Mikki»	25	898	Rechnung	24
847	nachdem	25	899	sehen	24
848	negativer	25	900	spezielle	24
849	niedrigen	25	901	starke	24
850	obwohl	25	902	steuern	24
851	seinen	25	903	Teile	24
852	Sicherheit	25	904	Tuner	24
853	sobald	25	905	um *от глаг.*	24
854	sofort	25	906	unten	24
855	«Source»	25	907	untergebracht	24
856	Steilheit	25	908	Verhalten	24
857	trotz *prp*	25	909	verhältnismässig	24
858	unterschiedlichen	25	910	vollständig	24
859	verändert	25	911	Wärme	24
860	Verhältnisse	25	912	wesentliche	24
861	Zählers	25	913	Zeilen	24
862	Abweichungen	24	914	Zeitkonstante	24
863	Achse	24	915	allgemein	23
864	allgemeine	24	916	angewendet	23
865	auftretenden	24	917	Basisstrom	23
866	ausgerüstet	24	918	Bedingung	23
867	äusseren	24	919	Betriebssicherheit	23
868	beachten	24	920	Bezeichnung	23
869	darin	24	921	Beziehung	23
870	EAY	24	922	bleiben	23
871	Eingänge	24	923	Entladung	23
872	eingebaut	24	924	erkennt	23
873	entstehende	24	925	erlaubt	23
874	erster	24	926	erläutert	23
875	«Gate»	24	927	gehören	23
876	Gatter	24	928	gesteuerten	23
877	gebaut	24	929	gleichen	23
878	geben	24	930	hoch *adv*	23
879	geöffnet	24	931	Induktion	23
880	gespeist	24	932	Jahre	23

Таблица 1 (*продолжение*)

i	Словоформы	F	i	Словоформы	F
933	Kondensators	23	985	verhalten	22
934	lieferbar	23	986	voll *adv*	22
935	Literatur	23	987	Zweck	22
936	*Messgerät*	23	988	*allein*	21
937	Mittelwert	23	989	Analogrechners	21
938	Netzteil	23	990	Anodenstrom	21
939	notwendige	23	991	beruht	21
940	obere	23	992	besonderen	21
941	Raum	23	993	DDA	21
942	Schichten	23	994	Deutschland	21
943	seien	23	995	«Drain»	21
944	speziell	23	996	eigentlichen	21
945	ständig	23	997	eingeschaltet	21
946	Stromdichte	23	998	einiger	21
947	umfasst	23	999	Einleitung	21
948	unterscheiden	23	1000	entnommen	21
949	Welle	23	1001	Ergebnis	21
950	wiederholt	23	1002	erheblich	21
951	Wirkung	23	1003	exakt	21
952	Abgleich	22	1004	Feldstärke	21
953	Arbeitspunkt	22	1005	gekennzeichnet	21
954	Ausführungen	22	1006	gewissen	21
955	«CEE»	22	1007	Gleichspannungen	21
956	darstellen	22	1008	grösserer	21
957	Datenverarbeitung	22	1009	hin	21
958	deutschen	22	1010	ihrem	21
959	Einheit	22	1011	Integrierer	21
960	Einrichtung	22	1012	interessant	21
961	elektronisch	22	1013	Interesse	21
962	ermittelt	22	1014	jeden	21
963	gestattet	22	1015	Kasseten	21
964	*gewisse*	22	1016	*Klirrfaktor*	21
965	Halbleitertechnik	22	1017	Konstante	21
966	hinsichtlich	22	1018	Kontrolle	21
967	Impulses	22	1019	Lampen	21
968	Kraft	22	1020	Maschinen	21
969	Kreise	22	1021	mehreren	21
970	Leitfähigkeit	22	1022	Messe	21
971	linear	22	1023	Messgeräte	21
972	Markt	22	1024	Messstelle	21
973	Methode	22	1025	Messtechnik	21
974	Methoden	22	1026	Messverfahren	21
975	nennt	22	1027	normale	21
976	Oszillograf	22	1028	Preis	21
977	Quelle	22	1029	Produkt	21
978	Rundfunk	22	1030	sicher	21
979	Service	22	1031	Sollwert	21
980	später	22	1032	stellen	21
981	stellte	22	1033	Steuergitter	21
982	Stromes	22	1034	Struktur	21
983	Übertemperatur	22	1035	Träger	21
984	unterhalb *prp*	22	1036	Übertrager	21

Таблица 1 (*продолжение*)

i	Словоформы	*F*	*i*	Словоформы	*F*
1037	Verbesserung	21	1069	Kollektrostrom	20
1038	wichtige	21	1070	Kurven	20
1039	Wicklungen	21	1071	kurze	20
1040	zehn	21	1072	kurzer	20
1041	Zeitpunkt	21	1073	Multivibratoren	20
1042	Angaben	20	1074	München	20
1043	angebracht	20	1075	Nachrichtentechnik	20
1044	angepasst	20	1076	Nachteil	20
1045	annehmen	20	1077	Nachteile	20
1046	Anschlüssen	20	1078	Netzspannung	20
1047	aufweisen	20	1079	neu	20
1048	Aussteuerung	20	1080	oberen	20
1049	Auswahl	20	1081	Punkten	20
1050	Batteriespannung	20	1082	Schaltbild	20
1051	bestimmen	20	1083	Schwingung	20
1052	betragen	20	1084	Speisung	20
1053	bezüglich	20	1085	speziellen	20
1054	Digitalrechner	20	1086	Spule	20
1055	Drossel	20	1087	Störungen	20
1056	dürfte	20	1088	Strecke	20
1057	Effekt	20	1089	Systems	20
1058	Emission	20	1090	umgekehrt	20
1059	Erde	20	1091	Verfahrens	20
1060	erfordert	20	1092	vermieden	20
1061	ermöglichen	20	1093	Volt	20
1062	Faktoren	20	1094	vorhandenen	20
1063	geeigneten	20	1095	Vorspannung	20
1064	gehört	20	1096	weiss	20
1065	Gleichrichtung	20	1097	wirken	20
1066	Gleichung	20	1098	zulässige	20
1067	Halbleitern	20	1099	Zwecke	20
1068	Impedanz	20	1100	zweier	20

Таблица 2

Распределение количества словоформ при $F \leqslant 19$

i	*m*	*F*	*i*	*m*	*F*
1101—1180	80	19	2268—2509	242	9
1181—1246	66	18	2510—2841	332	8
1247—1323	77	17	2842—3267	426	7
1324—1403	80	16	3268—3775	508	6
1404—1504	101	15	3776—4510	735	5
1505—1628	124	14	4511—5635	1125	4
1629—1763	135	13	5636—7407	1772	3
1764—1899	136	12	7408—11117	3710	2
1900—2050	151	11	11118—20405	9288	1
2051—2267	217	10			

Э. М. Гаспарова

ЧАСТОТНЫЙ СЛОВАРЬ НЕМЕЦКИХ ТЕКСТОВ ПО СЕЛЬСКОХОЗЯЙСТВЕННОМУ МАШИНОСТРОЕНИЮ

Исследуется лексическая статистика в немецких текстах по сельскохозяйственному машиностроению. Объем выборки — 100 тыс. словоупотреблений. В процессе исследования разграничивались случаи грамматической омонимии (между словами различных частей речи), поэтому словоформы в списке там, где это необходимо, снабжены индексами, показывающими грамматическую принадлежность.

Имя существительное индексом нигде не снабжено, поскольку в данном случае показателем служит написание с заглавной буквы.

В результате исследования получен частотный список словоформ (см. табл. 1).

Таблица 1

Частотный список словоформ

i	Словоформы	*F*
1	der *art*	4250
2	die *art*	3811
3	und	2967
4	in	1621
5	mit *prp*	1484
6	den *art*	1394
7	von	1136
8	des *art*	1121
9	ist	1070
10	werden	1023
11	für	984
12	eine *art*	900
13	bei *prp*	859
14	das *art*	818
15	auf *prp*	786
16	wird	696
17	zu	649

Таблица 1 (*продолжение*)

i	Словоформы	F
18	sind	645
19	dem *art*	636
20	auch	634
21	durch *prp*	626
22	sich	622
23	im	608
24	ein *art*	578
25	daß	561
26	nicht	527
27	die *pron*	465
28	als *cj*	448
29	an *prp*	430
30	es	423
31	kann	402
32	einer *art*	401
33	zur	391
34	oder	382
35	nach *prp*	352
36	einem *art*	332
37	einen *art*	302
38	man	296
39	diese	295
40	nur	294
41	zum	290
42	aus *prp*	288
43	bis *prp*	284
44	über *prp*	266
45	sie	265
46	beim	252
47	noch	245
48	so *adv*	238
49	können	231
50	hat	229
51	Schlepper	226
52	am *prp*	220
53	aber	213
54	zu *prp*	209
55	wie *adv*	208
56	dieser	202
57	wurde	191
58	Bild	189
59—60	das *pron*, wenn	188
61	vom	183
62	dann	180
63	haben	170
64	um *cj*	169
65	er	168
66	der *pron*	167
67	PS	155
68	muß	153
69	vor *prp*	151
70	Maschinen	145

Таблица 1 (*продолжение*)

i	Словоформы	F
71	Maschine	144
72	zwei	137
73	besonders	135
74	dabei *adv*	133
75	unter *prp*	131
76	Mähdrescher	128
77	hier	126
78—79	wurden, zwischen *prp*	123
80	da *cj*	122
81	Einsatz	120
82—83	sehr, um *prp*	119
84	wir	117
85	sein *v*	115
86	eines *art*	114
87—88	alle, ohne *prp*	113
89	so *cj*	112
90	Landwirtschaft	107
91	damit *adv*	105
92	also	102
93	dadurch *adv*	101
94	war	98
95—96	Leistung, soll	97
97—98	diesem, sowie *cj*	91
99	schon	90
100—102	etwa, je *adv*, neue	88
103—104	große, jedoch	87
105	Boden	86
106—107	Arbeit, dieses	83
108	möglich *attr*	82
109	heute	81
110—111	bereits, müssen	80
112	mm	79
113	ausgerüstet	74
114—115	bzw, immer	73
116—117	an *pref*, Geräte	72
118—120	beiden, läßt, wieder *adv*	71
121—124	bisher, nun, sollte, zu *adv*	70
125	diesen	69
126—129	besteht, erfolgt, neuen, sondern	68
130	außerdem	67
131	allen	66
132—135	Entwicklung, keine, Verwendung, weil	63
136—138	allem, m, Zeit	62
139—141	großen, kommt, Z. B.	61
142—143	Geräteträger, mehr *adv*	60
144	neben *prp*	59
145—146	Motoren, Pflug	58
147	zeigt	57
148—150	anderen, eingesetzt, Schleppers	55
151—155	Arbeiten, dazu, drei, Ernte, Rüben	54
156—160	angebracht, cm, entsprechend *adv*, während *cj*, während *prp*	53
161—166	Abb., auf *pref*, Jahren, Teil, verwendet, vorhanden	52

i	Словоформы	F
167—168	hohe, mehr *a*	51
169—173	deshalb, ein *pref*, finden, seine, Traktor	50
174—175	Praxis, weitere	49
176—177	Firma, seiner	48
178—187	arbeiten, aus *pref*, Betrieb, gibt, gleichzeitig, jeder, lassen, Typen, usw., Verfügung	47
188—191	andere, denen *cj*, ebenfalls, Hilfe	46
192—197	ab *pref*, erreicht, ihre, Messer, seit *prp*, verschiedenen	45
198—202	deren *cj*, ganz *adv*, hohen, liegt, uns	44
203—209	einige, ergibt, Hand, jetzt, landwirtschaftlichen, Motors, waren	43
210—213	angeordnet, Antrieb, Seite, stehen	42
214—218	besitzt, beträgt, bleibt, oft, vier	41
219—224	einmal, Frage, Gerät, Getriebe, hinter, Mähdreschers	40
225—235	Arbeitsbreite, Ausführung, doch *cj*, Druck, gegen *prp*, leicht *adv*, liegen, Möglichkeit, Pflügen, sowohl, wäre	39
236—245	größere, oberen, Pumpe, Regelhydraulik, solche, verbunden, Verfahren, verschiedene, wesentlich *adv*, zwar *adv*	38
246—251	beide, besondere, dagegen, einzelnen, Konstruktion, Vorteile	37
252—259	DM, dort, geliefert, Jahr, notwendig *attr*, ob, Stellung, Weise	36
260—265	erst, ersten, ihrer, kg, Lenker, recht *adv*	35
266—274	Ausrüstung, daher, darf, Ende, gute, Lage, sei, selbst *pron*, Zugkraft	34
275—284	dem *pron*, erforderlich *attr*, fast, Form, ha, handelt, infolge *prp*, km/h, Pfluges, weiter *adv*	33
285—296	Anlage, bringt, darauf, Forderungen, Getreide, kaum, kommen, RS, Verluste, Vorderachse, vorn *adv*, wegen	32
297—310	Behälter, Betriebe, Dreschtrommel, Fällen, folgende, Form, gleichen, Industrie, meist, oben, Produktion, Schleppern, Stand, steht	31
311—324	Abstand, Art, Bauer, bekannten, denn, Fahrer, geben, gegeben, Gewicht, Jahre, Richtung, sollten, stark *adv*, unteren	30
325—343	arbeitet, Aufwand, Betrieben, durchgeführt, ermöglicht, etwas *pron*, geeignet, gezeigt, gleiche, Hydraulik, Luft, Schütz, Schwierigkeiten, sollen, technische, Type, verhältnismäßig *adv*, was *pron*, zunächst	29
344—362	aller, Bedeutung, bestimmten *attr*, darüber, Drehzahl, eingebaut, gegenüber *prp*, geht, konnte, Landmaschinenindustrie, letzten, macht, möglichst, Rahmen, seinen, Stroh, Traktoren, unserer, worden	28
363—377	Anbau, Arbeitstiefe, beispielsweise, Bodenbearbeitung, entwikelt, Forderung, größeren, hierbei, hydraulische, Mähdrusch, sein *pron*, sogar, Teile, Typ, zusätzlich	27
378—393	allerdings, am *part*, Bedingungen, Böden, Bodens, damit *cj*, Feldhächsler, ferner *adv*, gebaut, Grund, machen, Reihe, Schneidwerk, U/min, vorgesehen, Wagen	26
394—414	angebaut, Aufgabe, Automatisierung, bestimmte *attr*, dürfte, Geschwindigkeit, gleich *adv*, gut *adv*, ihr, jeweils, Mechanisierung, Möglichkeiten, Räder, Reinigung, stellt, Trommel, viel *adv*, vielen, weit *adv*, wobei *cj*, Zylinder	25
415—437	Abbildung, allein, Anwendung, Arbeitszylinder, Ausführungen, Bauart, Bauern, befindet, dient, hinten, ihren, jede,	24

i	Словоформы	F
	jedem, konnten, Kraft, Landtechnik, Mähen, neuer, relativ *adv*, Steile, technischen, Wirtschaftlichkeit, Zapfwelle	
438—465	bekannt *attr*, Bereich, deutschen, Erhöhung, erreichen, Fahrgeschwindigkeit, Fall, Funktion, gebracht, Geräten, gesamte, gilt, hinsichtlich, Höhe, insbesondere, kleinen, links, Lösung, luftgekühlten, Mähdreschern, Mähwerk, Oberlenker, Pflege, Prozent, Regelung, Tatsache, verwenden, wo *cj*	23
466—488	Bedingungen, Betätigung, bringen, dafür, entstehen, Firmen, gerade *adv*, Gerätes, großer, ihm, jeweiligen, Kartoffeln, kleine, Kupplung, landwirtschaftliche, Markt, schaltet, sehen, sonst, vor *pref*, Wert, wesentlichen, zeigen	22
489—510	ab *prp*, außer, bietet, bisherigen, dessen *cj*, direkt *adv*, Durchmesser, ebenso, gehalten, Getreideernte, groß *attr*, Grubber, hatte, hydraulisch *adv*, LPG, mindestens, Nachteile, Öl, seinem, Verbindung, VEB, wobei *cj*	21
511—539	Ackerschlepper, Anteil, Beispiel, bestehen, darin *adv*, entsprechende, entspricht, ergeben, erhalten, erste, Feld, genügend *adv*, hinaus *adv*, höhere, Kolben, Kosten, ohne *cj*, rechts, Sämaschine, Schlepperfahrer, schwere, sofort, solchen, Traktors, Transport, unten, Untersuchungen, Vergleich, vorhandenen	20
540—566	Abhängigkeit, Acker, Anordnung, benötigt, Bezeichnung, bleiben, Dreschmaschinen, Energieträger, Erfahrungen, Fahrersitz, fest *adv*, gewährleistet, GmbH, hergestellt, kein, könnte, leicht, Leistungen, schaffen, Schwad, Steigerung, Technik, voll *adv*, Vorteil, Vorwärts-Zusammenhang, zusätzliche	19
567—608	Achsen, angeboten, angetrieben, Arbeitskräfte, bekannte, Claas, DDR, dienen, Egge, Einbau, einfach *adv*, einfache, einfachen, Einstellung, entsprechenden, ermöglichen, Falle, Gehäuse, gestattet, geworden, gut, guten, Hersteller, Interesse, ja, Landwirt, leichte, Lösungen, Maße, mittels, montiert, natürlich *adv*, Programm, Schnittbreite, setzt, stellen, üblichen, viele, wahlweise, weniger *adv*, Zahl, zweiten	18
609—644	Achse, bedeutet, ccm, Drillmaschine, durchaus *adv*, einzusetzen, entsteht, Ergebnisse, erzielt, früher *adv*, gelagert, gemacht, Geräteträgers, geringe, gezogen, Größe, höheren, ihn, Kraftheber, kurz *adv*, Landmaschinenbau, lediglich *adv*, Maß, Reihen, richtige, Samen, Schalter, spezifischen, ständig *adv*, System, teilweise, verringert, Vorrichtung, Wirkung, wodurch *adv*, Zinken	17
645—686	Anbaugeräte, Anhänger, Arbeitsgeschwindigkeit, Austrittsöffnung, benutzt, besitzen, besonderen, besser *attr*, betätigt, Bewegung, Bremsen, eingestellt, entweder *cj*, fällt, Fahrzeug, folgt, Garben, gering *attr*, geringen, gleicher *attr*, Heben, hoher, ihrem, Maßnahmen, München, Preis, obwohl; schnell *adv*, schweren, Seiten, seitlich, selbstfahrende, serienmäßig *adv*; Siebkette, stets, u. a., Verschleiß, Vollerntemaschinen, weniger *attr*, Werte, würde, zugleich	16
687—745	abhängig *attr*, anders, angebauten, Arbeitsaufwand, ausgeführt, automatischen, Bedarf, dargestellt, davon, deutsche, Deutschland, Dieselmotoren, d. h., Drehzahlen, Druschgut,	15

Таблица 1 (*продолжение*)

i	Словоформы	F
	Einrichtungen, einschließlich, erfolgen, erkennen, Erntegutes, Fahrgestell, Fendt, Folge, Förderung, führt, genau *adv*, geschaffen, gewünschten, größer *attr*, Haspel, häufig *adv*, heißt, Hinterachse, hoch *a*, hydrostatische, konstant *adv*, Korntank, läuft, Mais, mehrere, moderne, nämlich *adv*, Problem, Scheiben, schließt, Schwadbilden, stufenlos *adv*, technisch *adv*, Technologie, trägt, trotzdem, unsere, versehen, Versuche, viel *attr*, Voraussetzungen, Wege, wirkt, Zukunft-	
746—804	allgemeinen, Änderungen, angepaßt, Anzahl, Arbeiterproduktivität, Ausstellung, befestigt, bewährt, bewährten *attr*, Breite, Dieselmotor, DLG-Ausstellung, Drusch, einigen, erfordert, erhält, erscheint, festgestellt, Fläche, Flächen, führen, Gebiet, Gebläse, gefunden, geöffnet, geschieht, Geschwindigkeiten, gewissen, gleichmäßig *adv*, Gruppe, Hälfte, Handarbeit, hierzu, ingesamt; jeden, kleineren, Lenkung, mechanischen, modernen, neu *attr*, öffnet, Probleme, Qualität, Rührleitung, schließlich, Senken, treten, Verhältnis, vermeiden, völlig *adv*, voneinander *adv*, Wasser, Weg, welche *pron*, wesentliche, wichtig *attr*, Wunsch, zu *pref*, zweite	14
805—885	ähnlich *a*, AG, alles, alten, anderer, andererseits, angehoben, Arbeitsbreiten, Arbeitsgang, Aufbau, Aufbaugeräten, Aufgaben, Bearbeitung, berücksichtigt, bestimmt *v*, bewegt, den *pron*, dürfen, Einfluß, eingeführt, enthält, Erde, Erfolg, Erntegut, erwähnen, Fahrtrichtung, fallen, Feder, festzustellen, Frühjahr, ganze, ganzen, gar, Gelände, gelegt, genannten, genaue, genug *adv*, geringer *attr*, größerer, größte, Halle, Hebel, Heu, Hof, Hubkraft, Köpfe, Korn, Kühler, kurze, Messern, nehmen, Neuentwicklung, nimmt, normalen, notwendige, Öffnung, Pflugkörper, Prüfung, PS-Schlepper, Regel, Reifen, Rückwärtsgängen, Schutz, Schwaddrusch, Schwader, selbstverständlich, später *adv*, starken, stationären, stehenden, trotz, überhaupt, verschiedener, Verschlußlappen, vorhandene, weiteren, Werkzeuge, wichtigsten, wiederum, Wirkungsgrad	13
886—961	abgefederten, Anfang, Anforderungen, Arbeitskräften, aufgenommen, ausschließlich, automatisch *adv*, automatische, befinden, Belastung, Berücksichtigung, billiger *attr*, bis *cj*, Durchführung, eignet, einiger, Erfindung, erhält, erstmalig, Erzeugnisse, Fahrers, Fahrt, findet, folgenden, Fragen, gefüllt, gelöst, geschaltet, geschlossen, geschützt, gestellt, gewährleisten, Grundlage, grundsätzlich, günstigen, hauptsächlich, heutigen, Jahres, Kabine, Kombination, Korb, Kraftbedarf, Lageregelung, lange *adv*, Leitung, Mähbalken, Mann, meistens, Minuten, nachdem *cj*, neuerdings, Neuheit, Pflegearbeiten, Raum, Scheibenbremsen, selbstfahrendem, sicher *adv*, sieht, sogenannten, sogenannte, solcher, starke, Umfang, Varianten, Verbesserung, vergangenen, Verhältnisse, verhindern, vermieden, verstellbar *attr*, was *cj*, wenigen *attr*, Welle, ziehen, zurück *pref*, Zwecke	12
962—1066	Ablage, allgemein *adv*, angegeben, angeordneten, angewendet, Anlagen, Aufnahme, ausgestattet, Ausnutzung, außerordentlich, Bauweise, begrenzt, Beschädigungen, bessere, bewegen,	11

Таблица 1 (*продолжение*)

i	Словоформы	F
	braucht, breite, Bundesrepublik, ca, da *adv*, danach, daraus, Dreschkorb, Dreschmaschine, eigene, «Ein», eingeschaltet, Einrichtung, einsetzen, Einsparung, einzigen, elektrische, entsprechen, erfaßt, erforderliche, Fahr, fahren, Fortschritt, gefertigt, Gefahr, geführt, gegenüber *pref*, Gehäuses, gesamten, Gestänge, gleichmäßige, Halme, halten, Hang, her *adv*, Holder, Hubwelle, hydrostatischen, indem *cj*, innerhalb *adv*, Kauf, keinen, Klappe, kombinierten, kommenden, Körner, Krafthebers, Kraftregelung, langen, Last, laufen, leichten, leisten, Masse, mechanisch *adv*, meisten, Mittel, miteinander *adv*, mittlere, Motorleistung, MTS, Nachteil, normale, nunmehr, praktisch *adv*, Riemenscheibe, rotierenden, Rückwärtsgang, sagen, Schäden, Schare, Sicherheit, somit *adv*, sozialistischen, Speicher, speziell *adv*, Steuerventil, Stunde, tritt, übrigen, verringern, vorgenommen, weiterhin, wichtige, Winkel, wirtschaftliche, wollen, würden, Ziel, zusammen *adv*	
1067—1196	Abmessungen, agrotechnischen, Allradschlepper, Anbaugerüst, angesehen, Anschluß, auftreten, ausgebildet, ausgeschaltet, ausreichend *adv*, Aussaat, bäuerlichen, bedingt, Bergung, besonderer, Bestellung, besten, Bewertung, bewirkt, Blockbauweise, Bodenbearbeitungsgeräte, breiten, Eggen, einschaltet, einwandfrei *adv*, Einzelheiten, Elektroenergie, engen, entsprechender, erfüllen, erhöhten *attr*, ermittelt, ersetzt, erster, Europa, Fahrerkabine, Federung, fest *pref*, festen, Flächenleistung, fördert, französische, fünf, Furche, Gang, gearbeitet, gefahren, gefördert, gehen, gehört, gesenkt, gewählt, gleich *a*, günstig *a*, günstige, Handhabung, hingewiesen, Hinterräder, innerhalb *prp*, Instandsetzung, Konstruktionen, Kühlung, kurzen, Lader, Laders, Ländern, lange, leider, liegenden, Menge, Merkmale, Messungen, Methoden, mußte, müßte, nennen, neu *adv*, Neutral-Stellung, notwendigen, parallel *adv*, Pflanzen, Pflüge, Pflugfurche, rechnen, reicht, richtig, Sachs-Stamo, Saugöffnung, Schädlingsbekämpfung, Scheibenegge, Schleppertypen, Schneidspalt, sechs, senkrecht *adv*, Silomais, Situation, Sitz, sobald *cj*, Spur, Spurweite, Standardschlepper, Stärke, Staub, Stückzahlen, Systems, Tafel, Temperatur, Tiefe, übertragen, umfaßt, unmittelbar *adv*, unseren, Ursachen, USA, Verbrauch, Verhältnissen, verschiedensten, Voraussetzung, Vorderräder, vorliegenden, Wahl, weiteres, Weiterentwicklung, weitgehend *adv*, wodurch *cj*, Zeitrelais, Zugkraftregelung, zusammen *pref*, zusätzlichen, zweckmäßig *adv*	10
1197—1319	abgestimmt, Angaben, aus *adv*, ausreichender, außen, Ausstattung, Baureihe, beginnt, begonnen, bekanntlich *adv*, beliebig *adv*, benötigt, besser *adv*, besseren, bestimmen, bieten, bilden, bildet, Bodenfreiheit, Bremse, dar, Darstellung, Diskussion, DLG, Dreipunktaufhängung, drückt, eigentlich, eigentlichen, Einachsschlepper, entfällt, entwickelten *attr*, Entwicklunghilfe, Erfinder, erfolgte, erfolgern *v*, erhebliche, erhöhen, erkannt, Ermittlung, Erntemaschine, Fahrgeschwindigkeiten, Fahrzeuge, Fassungsvermögen, fehlen, Fertigung, festgelegt, Fingerbalken, Fremdkörper, Gebrauchtschlepper,	9

i	Словоформы	F
	gedrückt, Gegensatz, gehören, genannt, genommen, genügt, gewisse, Grenzen, größten, Gründen, hält, Hilfsmittel, hoch *adv*, hohem, Hubraum, interessant *attr*, jederzeit, Kette, konstante, Kräfte, Länge, leistungsfähige, Lenkhilfe, Maschinenfabrik, Massey-Ferguson, Material, Methode, nahezu, neuartige, obere, Odense, Öffnungen, oftmals, ökonomischen, optimale, Pflugarbeit, Prinzip, Rübenerntemaschinen, Rückschlagventil, rund, Saat, scheint, Schleifeinrichtung, Schlupf, schneller *adv*, Schnitt, Schüttler, Schwadmähen, schwierigen, Schwimmstellung, Sicht, soweit *cj*, Spannung, Strohbergung, Systeme, technischer, tun, unterscheiden, Ventil, Verbesserungen, Verringerung, verstellbare, Verstopfungen, verwendet, Vierradschlepper, vollen, vorgestellt, vorwiegend *adv*, wenig *adv*,	
1320—1480	will, zeigte, Zustand, Zweck, Zylinderkopf acht *num*, Ähren, ähnliche, Aggregate, Agri-Robot, alte, Anbringung, angetriebene, Anhängers, Anhängekupplung, Anpassung, Arbeitszeit, Armaturenbrett, Arbeitslosigkeit, aufeinander *adv*, Aufmerksamkeit, außerhalb, Baugruppen, bedeutend *adv*, befindliche, Beginn, Beleuchtung, beschrieben, beschriebenen, bestehenden, Betracht, betragen, Betriebes, betrug, bezeichnet, daran *adv*, derartige, Differentialsperre, Drehmoment, Dreschwerk, Drillmaschinen, Druschgutes, dz, dz/h, Einflüsse, Einführung, eingebaute, Einmann-Bedienung, empfehlen, entfallen, entscheidend *adv*, entwickelt, erheblich *adv*, erlaubt, Erntemaschinen, Erntegüter, Ernteverluste, ersetzen, Feldgemüsebau, feststellen, Folgen, folgender, Furchentiefe, Futter, gab, geeignete, geerntet, Gegenschneide, gelten, gesichert, Gestaltung, gesteuert, getragen, getrennt, Getriebegehäuse, guter, gewünschte, günstiger, harten, helfen, hierfür, höherer, Hydraulikanlage, Hydropumpe, ihnen, Institut, intensive, inzwischen, je *cj*, Ketten, Klasse, konstanter, konstruktive, Kontakte, Kurbelwelle, Landmaschinenindustrie; Linie, Liter, Mähgut, mechanische, Mensch, Messerantrieb, min., nichts, Notwendigkeit, Plattform, Platz, praktischen, Produkt, Produktivität, Räumen, Regler, Republik, Rolle, Rückwärtsgänge, rückwärts, Rührsaugleitung, Schlagleisten, Schlepperbau, Schneckenwalze, Schritt, Schützes, Schwaden, schwer *adv*, Schwingungen, selten, Selbstfahrer, Senkung, setzen, sitzt, stärkere, stärkerer *adv*, statt *cj*, Stellen, steigt, Steuerhebel, stumpfen, Tagen, Tiefgang, Triebachse, Übertragung, unbedingt, unterbrochen, Ursache, verändern, verbessert, verfügt, verhindert, volle, vordere, vorteilhaft *adv*, Vorzüge, wählen, Wartung, wassergekühlten, Weiterer, Welt, Wenden, wenige, Werkzeugen, wichtig *adv*, wirtschaflicher, Zufuhr, zuletzt, Zündkerzen, zunehmender	
1481—1720	abgenommen, Ackerschiene, Altschlepper, Amerika, amerikanischen, ändert, Anhänge-Drillmaschinen, Anlasser, annähernd *adv*, Annahme, anpassen, anschließend *adv*, arbeitenden, Arbeitsgänge, Arbeitsmaschinen, Arbeitsstellung, Arbeitsweise, aufgehängt, Aufsattel-Drillmaschinen, aufweisen, Ausland, ausländischen, ausreichende, äußerst,	

i	Словоформы	F
	Auswertung, Basis, Bearbeitet, Becherwerkes, bedeuten, befindlichen, Bereitstellungskosten, Berührung, Beschickung, beschänkt, Beschreibung, besonderes, Bestände, bestimmter, Betriebskosten, Besucher, betätigt, betrieben, bewältigen, bewußt, Bodenverhältnissen, Dach, Dänemark, darum *adv*, drehbar *adv*, Dreipunkt-Hecklader, Dreipunktgestände, Drechtrommeldrehzahl, Drosselventil, Dünger, Dungstreuer, durch *pref*, durchführen, durchzuführen, Ebene, Eggenbalken, eigenen, Eigengewicht, Eingriff, Einhaltung, Einsatzes, Einzug, Elektroindustrie, Enden, Erbsen, Ergebnis, erforderlichen, Erfüllung, erheblichen, Erkenntnis, erleichtern, Erreichen, exakte, Export, Fahrsilo, falls *cj*, Feldes, fließt, Forstwirtschaft, Fräse, frei *adv*, führte, Führung, geblieben, Gebr., gedacht, gefordert, geforderten, gegenwärtig *adv*, gekennzeichnet, Geräteträgern, gerecht, gereinigt, gepreßt, gesagt, gesetzt, gewesen, Grad, Grenze, großem, großes, Gut, Häcksler, hätte, häufiger, heraus *adv*, Hermann, Herstellerfirmen, Herstellung, hervor *pref*, hin *pref*, Hinsicht, hinteren, Hinweise, hinzu *adv*, hinzuweisen, Hockendrusch, Höhenverstellung, höher *adv*, Hubhöhe, hydraulischen, Ing., interessante, Italien, jährlich *adv*, jedes, KG, kleiner, Kleinschlepper, kleinste, Koppel, Körper, Kraftübertragung, kürzester, längere, Längsschwadablage, Lebensdauer, legen, Leistungsklassen, liefert, linken *attr*, Lüfter, luftgekühlte, manchmal, Maschinenbaus, mehreren, mittleren, mögliche, möglichen, Motorzapfwelle, mußten, Nutzungsdauer, öffnen, Pferde, praktische, Preissenkung, rechten, richten, richtig *adv*, Rodescheiben, Saugstücker, Saugstutzen, Schar, Schleppersitz, schließen, schneidet, Schräglauf, selbst *adv*, selbsttätig *adv*, senken, sonstige, sorgt, Speichers, sprechen, spricht, stand, stellte, Stengel, Streubreite, Strohpresse, Strom, Tag, Thermostat, tragen, Traktoristen, Trennung, Triebstoff, Triebwerk, u., überall, Übergang, übersehen, Umbau, umgekehrt *adv*, unabhängig *attr*, unabhängig *adv*, unmöglich *attr*, unterhalb *prp*, Unterschiede, unterschiedlich *adv*, unterschiedliche, Veränderung, verbesserte, verbreitet, verlangt, verstellbaren, verstellt, Versuchen, versucht, vertreten, verwendbar *attr*, vielleicht, vielmehr, vielseitige, vorderen, vorgesehenen, vorliegen, Vorsteuerschieber, Vorsteuerschiebers, vorwärts, Vorwärtsgängen, Walzen, wären, warum, Wasserpumpe, weist, welche *cj*, wenig *attr*, Werk, wichtigen, zeigten, Zerkleinerungserfolg, Zug, Zugwiderstand, Zusammenarbeit, zweifellos *adv*	
1721—2009	Abbau, Abfederung, abgelassen, abgelegt, abgeschaltet, Abschluß, abgesehen, ähnlichen, achten *attr*, Ackeroberfläche, Angebot, angehängte, Anhängegeräte, Anheben, Anschaffung, Antriebe, Arbeitskolben, Arbeitsplätze, Arbeitsqualität, Arbeitsverfahren, arbeitswirtschaftlichen, Aushebung, ausgedrückt, ausgeführten, ausgehoben, ausgestellt, Außenanschliff, äußere, äußeren, beachten, beachtet, beeinfussen, beeinflußt, Beendigung, behandelt, Behandlung, bei *pref*, «Belarus», belastet, benötigen, Benzin, Bereifung, Berge, berücksichti-	6

i	Словоформы	F
	gen, berücksichtigt, betätigen, betrachtet, beträchtlich *adv*, Betriebssicherheit, Beurteilung, bewährte, Beziehung, Bodenunebenheiten, Bodenzerkleinerung, brauchen, breit *a*, Claas-Mähdrescher, Cultitrac, daneben *adv*, darstellen, darstellt, David, deshalb *cj*, drehbar *a*, Dreipunkt-Anbau, Dreipunktgestänges, Dreipunkthydraulik, Druckluftanlage, Düngung, durchgesetzt, Durchsatz, dz/ha, eben, Ecken, Einachser, einer *num* einfacher, einfaches, eingesetzten, eingestellte, Einschalten, einschlägigen, Eintrittsöffnung, einwandfreie, einzeln, einzige, einzustellen, Elektromotoren, Energieaufwand, Entfernung, entwickeln, Erleichterung, Erprobung, erstmals, erwähnt, erwähnten *attr*, erweitern, erweitert, etc., europäischen, Fahren, faßt, Federn, Fehler, Feststellung, Frankreich, französischen, Frucht, früheren, Fuß, Ganggetriebe, gedreht, gekennzeichnet, gekommen, gelangen, geprüft, geringer *adv*, geringsten, geschliffen, gestaltet, Gesichtspunkt, gezogene, gleichem, greift, Grunde, Grünfutterholen, günstigste, halben, Hängen, Hanomag, hatten, Hebelarm, herausgebracht, Heutrocknung, Hilfskraftanlagen, hin *adv*, hingegen, hintere, höchste, hydraulisch *attr*, ihres, Ilo-Allzweck-Motor, innen, Jahrzehnten, John, kamen, Kartoffelvollerntemaschinen, Keilriemen, keineswegs, Kerze, kg/ha, könnten, Körnermais, Körnern, Kraftheber-Hubkraftkurve, Kraut, Kurbelgehäuse, laden, Lagerung, lang *adv*, langes, langsam *adv*, Laufe, legt, leistet, Leitungen, Mai, manchen, Maschinenbau, max., Menschen, Messe, Meter, Mist, Mitte, Mittelpunkt, Modell, moderner, Motordrehzahl, Motorschutzschalter, nachfolgend *adv*, nächsten, Nenndrehzahl, neues, niemals, nochmals, Nutzlast, Ölbehälter, Planung, Pritsche, Pumpenstrom, Punkt, rasch *adv*, Regeldruck, Regelhydraulik-Anlage, Relais, Reparaturkosten, richtigen, richtiger, Rodeeinrichtung, Roden, RTS, Rücksicht, Rüstzeiten, Sammel-Roder, sämtliche, Schaden, Schärfen, Scheibe, Schicht, Schlepperhydraulik, Schlitten, schmalen, Schmied, Schneiden, Schrägstellung, Schutzart, Schwerpunkt, sichern, Sicherungen, Siebe, sinkt, Sinne, sogen., Spurlockerer, Stalldungstreuer, Stamo, stärkeren, Steigleitung, Steine, Sternrädern, Steuerhebels, Steuerstrom, Streifen, stufenlose, Tandemschlepper, Tastschalter, Teilen, Tragschlepper, Transportarbeiten, treffen, Triebräder, Tubermatic, Überblick, Überlastung, überlegen, übernimmt, überprüfen, üblich *attr*, umfangreiche, Umstand, Umständen, ungünstigen, Unterlenker, unverändert *adv*, Variante, verbessern, Verbreitung, verstehen, verstellbar *adv*, Versuchsfräse, verursachen, verursacht, verwirklicht, verzichten, vielfach *adv*, vollkommen *adv*, waagerecht *adv*, weder *cj*, weitergehende, welcher *pron*, welches *cj*, wer *pron*, Widerstandslinie, Widerstandsresultierende, wirken, wirkende, wirtschaftlichen, Wünschen, zählen, zahlreichen, Zahnrad, Zapfen, zieht, Zuckerrübenernte, zurückzuführen, Zusammenfassung, zweckmäßigsten, Zylindern	
2010—2460	abgestellt, abhängt, absolut *adv*, Abständen, Ackerschleppers, Ackerwagen, ähnlich *adv*, Allradantrieb, Anbau, anerkannt,	5

i	Словоформы	F
	angegebenen, angeordnete, angeschafft, angeschlossen, angenommen, angetriebenen, Anhängeschiene, Anlaß, April, anstelle, Arbeitsgemeinschaft, Arbeitsgerät, Arbeitsgeräte, Arbeitskraft, Arbeitswiderstand, Arten, Aufbau-Drillmaschinen, aufgesetzt, aufgestellt, Aufhängung, Aufsattel-Drillmaschine, auftretenden, auftritt, Aufwandskennziffer, Ausdrusch, Ausfall, Ausfuhr, ausgeführte, ausgehend *adv*, ausgelegt, ausgenutzt, Auslastung, ausreichenden, Auswahl, auswirkt, auszurüsten, Auto, automatischer, Bauarten, bäuerliche, Baumuster, Bauprogramm, Bauteile, Beachtung, bedient, Bedienungsperson, beendet, Befestigung, beider, Beispiele, bekannter, bequem *adv*, Bernburg, berührt, beschäftigt, Beseitigung, Besitzer, bestehend *adv*, Bestimmung, Bestreben, betreffenden, Bevölkerung, bewährte, beziehen, beziehungsweise, Bindemäher, Blick, Bodenart, Bodenbearbeitungsmaschinen, Bodendruck, Bodenoberfläche, Bodenverhältnisse, brachte, Bunker, Co., Dammaufnahme, dänische, dargestellte, dauernd *adv*, dementsprechend, dennoch, derselben, desto *cj*, deutlich *adv*, Dicke, diente, direkte, Dr., drehbaren, Dreschen, dritte, drücken, Durchgang, Durchschnitt, Eigenschaften, eignen, ein-, ein *num*, Einachsanhänger, Einachs-Motorgeräten, Einssachchleppern, einerseits, einfach *attr*, eingegangen, eingehalten, eingespart, Einmann-Arbeit, einreihigen, Einsatzgrenzen, Einsatzmöglichkeiten, eintreten, elektrischen, Empfehlungen, empfiehlt, Endgetriebe, Energie, England, englischen, entfernt, enthalten, Entleerung, Erfordernissen, ergab, Ergebnissen, erledigen, erneut, Erntearbeiten, Ernten, Erreichung, Ersatz, erscheinen, Erwärmung, erwarten, erzielen, evtl., Fahrerstand, fährt, Fahrzeugen, Fahrzeuges, fand, Felde, Feldrand, Ferguson, feste, flacher, Förderband, Fördern, Forschung, Früchte, führten, Funktionen, Gartenbau, geändert, gebildet, geförderte, gelangt, Geld, Gelegenheit, Gelenkwelle, gemäht, gemessen, genannte, Gerätereihe, gering *adv*, geringem, geringere, geringeren, gesehen, Gesichtspunkte, gestalten, gestatten, gesteuerte, Getrieben, getrieben, getrennten, gewährleistet, Gewerkschaften, Gewichtes, gewiß, Gotland, Großbetriebe, Größen, größer *adv*, Grunbach, Grundlagen, gnundsätzlichen, Handhebel, hängigen, Hanglagen, hängt, Hauptsache, Heck, Heizöl, Herbst, hergestellte, hervorzuheben, hindurch, hinweg, höher *attr*, Hub, Hubkupplung, hydraulischer, Hydrostatik, ins, insofern, Instituts, International, internationale, jedenfalls, jeweilige, kam, Katalog, kaufen, Kenntnis, Kennziffern, Kippvorgang, kombiniert, Kraftstoffverbrauch, Kreis, Kulturen, kurz *attr*, kurzer, Laderbaugruppe, Landarbeit, Länder, längst, Lauf, laufende, laufenden, Laufrollen, Lanz, leichter *a*, leichteren, Lenkbremssystem, Lenkrad, ließ, Lockerung, Lohn, losen *attr*, lösen, luftgekühlt, Luftkühlung, machte, mag, Magneten, Mähmaschine, Mähtisch, mancher, Mängel, Markkleeberg, Maschinensystem, mechanisches, mehrerer, Meinung, Merkmal, Meßfeder, Meßlatten, mittlerer, möchte, Motormäher, Nachbearbeitung, neuartigen,	

Таблица 1 (*продолжение*)

i	Словоформы	F
	Neuentwicklungen, Neuerungen, neuesten, Neuheiten, niedrige, niedriger *attr*, niedriger *adv*, nötig *attr*, notwendig *adv*, Nr., Nutzung, o. a., Oberfläche, objektive, Obstbau, ökonomische, optimalen, Organisation, organisatorische, Person, Pferd, Pflugwiderstand, Pick, Potsdam-Bornim, Praktiker, Presse, pro Produzenten, Prof., PS-Typ, quer *adv*, Querschnitt, Rad, Rädern, Radstand, Rationalisierung, rechtzeitig *adv*, Regelanlage, Regelsteuergerät, regelt, reguliert, Reinigungsflüssigkeit, Reparaturen, Rübe, Rübenernte, Rührstutzen, Sachs, samt, sauber *adv*, saubere, Schaltbild, Schaltung, Schiene, Schlagkraft, Schlauch, schleift, Schlepperfirmen, Schlepprechen, Schmutz, Schnittenergiebedarf, schräg *adv*, schwerer *attr*, Seilzugegge, Seitenflächen, seitlichen, Serie, siebfähigen, Sinkstoffe, solange, Sonderausrüstung, sonstigen, Sorte, Spannungsabfall, späteren, spezielle, Spurreiβer, Stadt, Stall, Standpunkt, stark *attr*, starkem, starker, stehen *pref*, steigenden, Sterndreieckschalter, Steuerung, Stillstand, Stoppeln, Strebe, Streuung, Striegel, Stück, stumpfe, Stunden, Stundenleistung, Sturm, «Super», Tabelle, Tendenz, teurer, theoretisch *adv*, tiefe, Tiefeneinstellung, Tragkraft, trennen, Trocknung, Übereinstimmung, übliche, übrigens, UdSSR, um *pref*, umlaufenden, Umlegen, unserem, untere, Unternehmer, untersuchten, unterscheidet, unterschiedlichen, Untersuchung, Unterschied, unzureichend *adv*, Ventils, verändert, verbleibt, Vergrößerung, Verkettung, verlangt, Verlauf, Verlusten, vermag, Verriegelung, verringert, versorgt, Versorgung, verstärkt, verstellbarer, Versuchsgut, verwendete *attr*, verzeichnen, Verzögerung, Volkswirtschaft, voll *attr*, Wärme, Warmluft, wassergekühlte, Watt, Wegeventils, weisen, weiter *attr*, welchen *pron*, wenden, Werkzeug, wesentlicher *attr*, Westdeutschland, Widerstand, wieviel, wirkenden, wissen, womit *cj*, zehn, Zeitpunkt, Zerkleinerungserfolges, zueinander *adv*, zugeführte, Zugleistung, Zugöse, Zugwiderstandes, zulässigen, zumal *cj*, zumindest, zusammengefaβt, Zusatzeinrichtung, Zusatzeinrichtungen, zweier, zweireihige, Zwischenachsanbau	

Таблица 2

Распределение количества словоформ при $F \leqslant 4$

i	m	F
2461—3137	677	4
3138—4313	1176	3
4314—7030	2717	2
7031—17219	10189	1

С. Г. Чапля

ЧАСТОТНЫЙ СЛОВАРЬ ФРАНЦУЗСКИХ ТЕКСТОВ ПО НЕФТИ И ГАЗУ

При составлении приводимого ниже частотного словаря французского подъязыка нефти и газа была использована существующая в группе «Статистика речи» единообразная методика.[1]

В отличие от предыдущих частотных словарей нами подсчитывались также цифры (C), формулы (Form.), аббревиатуры (AB), символы (S), имена собственные (NP); лексико-грамматические и грамматические омографы учитывались как отдельные единицы словаря.

Материалом исследования служили французские научно-технические тексты из журналов 1958—1967 гг.: «Association française des techniciens du pétrole», «Industrie du pétrole», «Pétrole Informations», «Pétrole-Progrès», «Revue de l'Institut français du pétrole et annales des Combustibles liquides», «Revue pétroliére», «Techniques et applications du pétrole».

Материал подбирался по специально разработанной для данного подъязыка схеме дозировки текстов.

Общая длина проанализированных текстов составила 200 тыс. словоупотреблений.

В табл. 1 приводятся 1165 словоформ с частотами от 11 980 по 20 включительно. В табл. 2 приводится распределение количества словоформ по частотам при $F \leqslant 19$.

Сокращения

adj	— adjectif.	*nég*	— négation
adv	— adverbe.	*part*	— participe.
art	— article.	*pers*	— personnel.
aux	— auxiliaire.	*pl*	— pluriel.
c rég	— cas régime.	*pr*	— pronom.
conj	— conjonction.	*prép*	— préposition.
contr	— contracté.	*réfl*	— réfléchi
cop	— copule	*s*	— substantif.
dém	— démonstratif.	*sg*	— singulier.
f	— féminin.	*v*	— verbe.
m	— masculin.		

[1] См.: П. М. Алексеев. Частотные словари и приемы их составления. СР, стр. 62.

Таблица 1

Частотный список словоформ

i	Словоформы	F
1	de	11980
2	C	9524
3	la *art*	5368
4	NP	4598
5	à	4243
6	et	3934
7	les *art*	3843
8	d'	3688
9	le *art*	3478
10	S	3266
11	AB	3261
12	en *prép*	2810
13	du	2647
14	l' *art f*	2445
15	Form.	2271
16	une	2058
17	des *art contr*	2039
18	l' *art m*	2018
19	par	1976
20	dans	1907
21	un	1853
22	des *art*	1543
23	pour	1510
24	que	1404
25	est *v cop*	1338
26	au	1095
27	qui	1075
28	plus	1025
29	gaz *s sg*	1024
30	il	845
31	on	804
32	est *v aux*	780
33	sur	756
34	a *v aux*	736
35	pétrole	632
36	été *part*	589
37	ou	588
38	avec	565
39	se	518
40	cette	513
41	sont *v cop*	507
42	ont *v aux*	502
43	s' *pr réfl*	491
44	ces	487
45	qu'	486
46	hydrocarbures	455
47	être	436
48	production	424
49	aux	422
50	produits *s*	380

i	Словоформы	F
51—52	ce *adj dém*, sont *v aux*	378
53	ne	375
54	pas *nég*	372
55	température	351
56	peut	344
57	ce *pr dém*	336
58	n'	329
59	liquide *s*	326
60	huile	323
61	deux	310
62—63	mais, pression	288
64	très	278
65	point *s*	262
66	comme	256
67	acétylène	253
68	dont	252
69	eau	249
70	entre *prép*	244
71	ainsi	243
72	conditions	241
73	tableau	238
74	soit *v cop*	233
75	sous	230
76	réaction	222
77	c'	219
78	industrie	217
79	calorimétrie	215
80	elle	214
81	méthane	212
82	procédé *s*	210
83	nous	208
84	leur	205
85	en *particule*	204
86	tonnes	203
87—88	éthylène, hydrogène	202
89	forage	200
90—91	énergie, son	199
92	oxydation	192
93	formation	191
94	moins *adv*	186
95	même *adj*	184
96—97	autres, procédés *s*	182
98	huiles	178
99	l' *particule*	172
100—101	où, partie	171
102—103	cours, même *adv*	170
104	charge *s*	169
105	benzène	167
106	carbone	166
107	sans	164
108—109	raffinerie, si *conj*	162
110—111	données *s*, jusqu'	160

Таблица 1 (*продолжение*)

i	Словоформы	F
112—113	**bien, chaleur**	158
114	**français** *adj sg*	156
115	**tous**	155
116	**celle**	154
117	**donc**	153
118—120	**aussi, cas** *s sg*, **rapport**	152
121	**essence**	151
122	**puits** *s pl*	149
123	**millions**	146
124	**raffinage**	144
125	**naturel**	143
126—127	**autre, sa**	142
128	**quelques**	141
129—131	**a** *v*, **boue, ses**	140
132	**catalyseur**	139
133	**résultats**	138
134—135	**également, état**	137
136—137	**part** *s*, **rendement**	136
138—141	**carbures, ils, partir, produit** *s*	135
142	**synthèse**	134
143—145	**environ** *prép*, **nombre, peuvent**	133
146	**aromatiques** *s*	132
147	**mélange**	131
148—149	**capacité, permet**	130
150—152	**bitumes, première, tout**	127
153—155	**gisement, institut, oxygène**	126
156—158	**étude, installation, recherches**	125
159	**effet**	124
160—161	**fabrication, réactions**	123
162—163	**brut** *s*, **distillation**	122
164—165	**libre, structure**	121
166	**après**	120
167—168	**société, travaux**	119
169—171	**augmentation, en** *pr*, **équation**	118
172	**grande**	117
173	**y** *adv*	116
174—175	**aromatiques** *adj*, **exploitation**	115
176—179	**alors, construction, essais, propriétés**	114
180—181	**encore, groupe**	113
182—183	**acide faible** *adj*	112
184—185	**non, temps**	111
186	**vitesse**	110
187—188	**depuis, installations**	109
189—192	**asphaltènes** *s*, **composés** *s*, **valeurs, volume**	108
193—194	**caractéristiques** *s*, **consommation**	107
195—199	**développement, liquide** *adj*, **octane, premier, recherche** *s*	106
200—201	**étant, fait** *s*	105
202—205	**cet, chimique** *adj*, **surface, unité**	104
206—207	**combustion, type**	103
208	**viscosité**	102
209—213	**bitume, celui, elles, stockage, zone**	101
214—217	**fonction, injection, peu, utilisation**	100

Таблица 1 (*продолжение*)

i	Словоформы	F
218—219	**années, vers**	99
220	**leurs**	98
221—222	**méthode, mise** *s*	97
223—225	**gisements, mis** *part*, **raffineries**	96
226—227	**densité, quantité**	95
228—233	**doit, domaine, entropie, problème, réacteur, transport** *s*	94
234—235	**actuellement, air**	93
236—237	**divers, sera** *v cop*	92
238—239	**éthane, plusieurs**	91
240	**spectroscopie**	89
241—242	**composition, toluène**	88
243—248	**exemple, forages, gaz** *s pl*, **paraffines, solvant** *s*, **toutes**	87
249—253	**cependant, cracking, fluorescences, ordre, pays** *s pl*	86
254—259	**déjà, études, lorsque, propane, service, solution**	85
260—264	**année, contre, mer, nouvelles, trois**	84
265—270	**équilibre, était** *v cop*, **méthodes, possible, problèmes, suite**	83
271—272	**compte** *s*, **présence**	82
273—274	**faire, solide** *s*	81
275—277	**certains, emploi** *s*, **températures**	80
278—281	**façon, poids** *s pl*, **reforming, valeur**	79
282—283	**notamment, total** *adj*	78
284—286	**catalyseurs, nouvelle** *adj*, **suivant** *adj*	77
287—291	**analyse** *s*, **calcul, mesure** *s*, **polystyrène, seulement**	76
292—297	**ceci, ceux, poids** *s sg*, **revue** *s*, **système, utilisé** *part*	75
298—300	**car, enfin, teneur**	74
301—305	**nature, nouveau, pétroliers** *adj*, **raison, usine**	73
306—307	**obtenir, sera** *v aux*	72
308—309	**avoir, charbon**	71
310—311	**fond, près**	70
312—317	**ensemble** *s*, **extraction, heptane, indice, région, surtout**	69
318—321	**engineering, gazeux** *adj sg*, **marché** *s*, **puis**	68
322—325	**celles, condensat, différentes, fin**	67
326—332	**assez, chaque, donne, oxyde** *s*, **période, sociétés, traitement**	66
333—337	**fraction, ici, laquelle, naphtènes, obtenus** *part*	65
338—345	**débit** *s*, **faut, hydrogénation, l'** *pr pers*, **nord, phase, toujours, villes**	64
346—350	**différents, élevé** *part*, **en** *adv*, **toute, vapeur**	63
351—360	**dire, fractions, il y a, lieu, matière, nécessaire, opération, quantités, souffre** *s*, **thermique** *adj*	62
361—367	**activité, chimiques** *adj*, **fait** *v*, **le** *pr pers*, **particulièrement, prix** *s pl*, **vue** *s*	61
368—373	**atomes, avait** *v aux*, **cyclohexane, jour, mètres, série**	60
374—378	**cas** *s pl*, **distribution, donné** *part*, **laboratoire, zones**	59
379—384	**golfe, grâce, importance, isobutane, nouveaux, usines**	58
385—393	**an, base** *s*, **combustible, milieu, obtenu** *part*, **oeuvre, profondeur, réserves, trouve**	57
394—398	**goudron, paraffiniques** *adj*, **pendant, pipe-line, rapide**	56
399—405	**avons** *v aux*, **concentration, moyenne** *adj*, **opérations, rapidement, thermodynamiques** *adj* **types**	55
406—417	**d'abord, départ** *s*, **diverses, grand, matières, montre** *v*, **moteur, ont** *v*, **permettent, pertes, pétrolière, place** *s*	54
418—426	**carburants** *s*, **certaines, était** *v aux*, **déshydrogénation,**	53

Таблица 1 (*продолжение*)

i	Словоформы	F
	isopentane, légère, puissance, totale *adj*, **utilisés** *part*	
427—431	**fut** *v aux*, **mise** *part*, **normal, pétrochimie, relativement**	52
432—445	**appareil, auteurs, azote, brûleur, butadiène, circulation, électrique** *adj*, **fait** *part*, **hydrocarbure, important, importante, indices, moteurs, sud**	51
446—457	**afin, application, dernier, forme** *s*, **haute, moyens** *s*, **présente** *v*, **récemment, reste** *v*, **transformation, unités, ville**	50
458—468	**acides, aniline, charges, contenant, dérivés** *s*, **éléments, essences, fois** *s sg*, **lui, pétroles, propylène**	49
469—477	**appareils, butane, chaîne, chauffage, début** *s*, **souvent, technique** *adj*, **trop, utiliser**	48
478—485	**avant** *prép*, **beaucoup, c'est-à-dire, conversion, entropies, four, molécules, selon**	47
486—505	**agit, certain, chimie, contrôle** *s*, **demande** *s*, **diamètre,** double *adj*, **élevée** *part*, **enregistrement, fois** *s pl*, **hexane, isomères, manière, moléculaire, parmi, proportion, puits** *s sg*, **réalisation, structures, techniques** *adj*	46
506—520	**accroissement, acier, ans, anthracène, besoins, catalytique** *adj*, **concerne** *v*, **d'après, dernière, donnent, existe, générale** *adj*, **intérêt, niveau, possibilitées**	45
521—530	**centre, chaleurs, chantier, combustibles, là, mai, mêmes, seront** *v aux*, **seul, variation**	44
531—542	**deuxième, importantes, l'** *pr pers m*, **les** *pr pers*, **lubrifiants** *s*, **plupart, principaux, programme, réservoirs, second** *adj*, **technique** *s*, **voie**	43
543—560	**aromatisation, ayant, butène, diminution, ensuite, essentiellement, évolution, exploration, fonctionnement, fuel, liaison, monde, origine, passage, presque, principe, semble, techniques** *s*	42
561—568	**évidence, expérience, extension, figure, long, produit** *v*, **particulier, pétrolières**	41
569—580	**aide** *s*, **dépôts, détection, française, graissage, importants, maintenant** *adv*, **mouvement, pu, séparation, tableaux, tubes**	40
581—593	**industriels, large, mesures, milliards, nos, pourrait, quatre, réseau, résistance, sulfurique** *adj*, **tels, travail, voir**	39
594—611	**action, cela, chiffre, colonne, contact, durée** *s*, **formule, français** *adj pl*, **généralement, grandes, isooctane, longue, moyen** *s*, **nombreux, produire, réaliser, sécurité, utilise**	38
612—626	**avril, basse, constantes** *s*, **étudié, faibles, gasoil, lors, lorsqu', méthyle, néopentane, obtient, premiers, récupération, représente, styrène**	37
627—644	**arc, atteint** *part*, **courant** *s*, **diméthylpentane, équipement, groupes, isobutène, limite** *s*, **liquides** *s*, **lourds** *adj*, **polymère, premières, réservoir, revue** *part*, **saturés** *part*, **solvants** *s*, **taux** *s sg*, **utilisant**	36
645—659	**chaînes, comporte, cycle, effort, fut** *v cop*, **nécessaires, normaux, pétrochimique, plan, préparation, principales, schéma, tant, tel, telles**	35
660—679	**aucune, brut** *adj*, **casing, conduit** *v*, **contient, continue** *part*, **heure, industrielle, n-heptane, n-pentane, pur, qua-**	34

Таблица 1 (*продолжение*)

i	Словоформы	F
	lité, rotation, sensiblement, seule, supérieur, traces, transports, troisième, vient	
680—700	**ailleurs, aromatique** *adj*, **atteindre, but, butyne, condensats** *s*, **constituants** *s*, **cycliques** *adj*, **économique** *adj*, **étape, éthylbenzène, houille, industries, maltènes, phénols, pourcentage, pourra, pratiquement, tonnage, va, vinyle**	33
701—732	**activités, addition, apparaît, calorimétriques, côté, courbes, coût, dernières, effluent** *s*, **équations, eux, expansion, fuels, juin, marche** *s*, **molécule, mondiale, n-butane, n-octane, nombreuses, permettant, quant, quelque, réduction, sait, serait** *v cop*, **simple, situation, source, supérieure, toutefois, xylène**	32
733—754	**analyses, barils, carburant** *s*, **continu** *part*, **données** *part*, **durant** *prép*, **enregistreur, furent** *v aux*, **général** *adj*, **guerre, introduction, investissements, légers, lequel, libres, nous** *c rég*, **permettre, platine, points, prix** *s sg*, **régions, sol**	31
755—783	**agents, aliphatiques** *adj*, **applications, bruts** *adj*, **caoutchouc, chlorhydrique** *adj*, **complexe** *adj*, **création, décomposition, dès, difficile, difficultés, donnée** *part*, **économie, effectués, entièrement, frais** *s pl*, **mars, mieux** *adv*, **oléfiniques** *adj*, **pentane, pouvant, propène, quand, réalisations, régénération, supérieurs, t/an, tube**	30
784—807	**atmosphère, certaine, chlorure, classiques** *adj*, **conduite** *s*, **constante** *s*, **contiennent, degré** *s*, **d'environ, diméthylhexane, fluorescence, gamme, longueur, maximum, pétrolier, pipe-lines, question, réactionnel, rendements, rouge, sorte** *s*	29
808—825	**bas** *adj*, **chiffres, concernant, considérablement, constitue, directe, ébullition, effectue, effectuer, forte, heures, influence, lourd** *adj*, **naphténiques** *s*, **polyéthylène, solide** *adj*, **tension, thermiques** *adj*	28
826—846	**autant, champ, champs, cinq, comportement, craquage, doivent, donner, effectué, égale, montrent, notre, partielle, permis, présent** *adj*, **présentent, pressions, socle, sondage, spécifique, substance**	27
847—880	**avancement, cadre, carbonique, cause, centrales** *s*, **chambre, classique** *adj*, **conduites** *s*, **contamination, d'ailleurs, découverte** *s*, **déterminer, dit** *part*, **effets, évidemment, hydrogénée, intensité** *s*, **intéressant, intérieur** *s*, **liquéfiés, mettre, mois** *s pl*, **naphtalène, navires, opératoires** *adj*, **passe, polymérisation, projets, résidu** *s*, **s'** *conj*, **transformé, travers, triméthylpentane, xylènes**	26
881—921	**alimentation, aucun, avantage, avantages, avenir** *s*, **avaient** *v aux*, **butylbenzène, calories, choix** *s sg*, **constante** *adj*, **couche** *s*, **différence, échelle, effectuées, élastique** *adj*, **étaient** *v aux*, **exemples, exportation, facile, finis, fonctions, forme** *v*, **lignite** *s*, **loin, masse** *s*, **matériel** *s*, **navire, n-hexane, oil, outil** *s*, **phénol, phénomène, possibilité, prévue** *part*, **prototype, provenant, radicaux** *s*, **réalisée, richesse, rôle, variations**	25

Таблица 1 (*продолжение*)

i	Словоформы	F
922—963	accord, alcools, améliorer, assurer, au-dessus, augmente, bilan, centres, complexe *s*, compris, d'autres, dépend, directement, dispositif *s*, distance, éviter, fixation, gros, idée, largement, liquéfaction, méthylhexane, nettement, normale, or, outre, particulières, physiques *adj*, plateforme, pleine, pompage, potentiel *s*, prévu, principal *adj*, processus, qualités, stabilité, suivantes, tenu, tonnes/an, vanadium, varie	24
964—997	augmenter, celle-ci, changement, complète, conception, courbe, disposition, dû, élevés, étranger *s*, eu, fluide *s*, grandeur, grands, grès, laboratoires, liquide *adj*, montré, pays *s pl*, pieds, précision, primaire, principalement, réalisé, relative, reprise *s*, résines, roches, seconde *adj*, section, suffisamment, suivante, supérieures, telle	23
998—1037	actuelle, anodes, argile, carbure, citernes, conclusions, consiste, critique *adj*, découvertes *s*, effectuée, entrée *s*, éthyléniques *s*, évaporation, extrêmement, face, février, font, hydraulique *s*, kérosène, mécanique *adj*, méthylbutène, méthylpentane, nationale, nonane, obtenue, paraffiniques *adj*, parfois, pénétration, pose *v*, poussées *s*, pouvoir *v*, puisqu', refroidissement, spectroscopiques *adj*, systèmes, terre, turbine, utilisée, utilisées, vannes	22
1038—1086	absence, absolue, amélioration, américain, augmenté, auteur, bonne, calculs, comprend, considérer, convient, corps, cubes *s*, dispose, éthylpentane, expériences, facteurs, formations, fournit, fusion, haut, légères, majeure, mises *part*, modernes, moins *s*, ni, noyau, objet, polyvinyle, pompes, possibles, raisons, relation, résoudre, résulte, sable, sens, sismique *adj*, sortie *s*, substitution, tenir, traité, traitée, triméthylbutane, vaporisation, vente, vide, vu	21
1087—1165	alcool, altération, aluminium, ammoniac, atteint *v*, cabine, capacités, caractère, connue, conséquent *s*, constitué, contraire *s*, correspond, croît, d'autant, détail, déshydrocyclisation, détermination, développé, devient, doute *s*, économiques, électrochimique, élément, élimination, équipements, étaient *v cop*, étapes, facilement, favorables, fer, final *adj*, françaises, industrielles, jours, la *pr pers*, lesquelles, lesquels, litres, loi, lourde *adj*, matériels *s*, maximale, meilleure, mélanges, métal, méthanol, moléculaires, molybdène, nécessité *s*, offshore, oléoduc, paraffine, parrois *s pl*, passer, performances, pollution, porté, pourtant, précédent *adj*, probablement, projet, puisque, recyclage, respectivement, secteur, sélectivité, septembre, spéciaux, stade, superfractionnement, supplémentaire, suivants, suspension, turbines, valorisation, vitesse, voies, volumétrique	20

Таблица 2

Распределение количества словоформ при $F \leqslant 19$

i	m	F	i	m	F
1166—1200	35	19	2008—2177	170	9
1201—1254	54	18	2178—2354	177	8
1255—1314	60	17	2355—2616	262	7
1315—1379	65	16	2617—2953	337	6
1380—1467	88	15	2954—3371	418	5
1468—1550	83	14	3372—3963	592	4
1551—1636	86	13	3964—4908	945	3
1637—1744	108	12	4909—6703	1795	2
1745—1854	110	11	6704—11830	5127	1
1855—2007	153	10			

Л. С. Никитина

ИМЕННЫЕ ТРЕХСЛОВНЫЕ СОЧЕТАНИЯ В РУССКИХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТАХ

Нами была предпринята попытка статистического исследования трехсловных сочетаний общественно-политической информации, содержащейся в газетных текстах. Отбирались тексты по следующей тематике:

1) официальные сообщения;
2) заявления и письма глав правительств;
3) комментарии на международные темы;
4) доклады и выступления на съездах и пленумах;
5) обзор политических событий;
6) фельетоны политического характера.

Общий объем обследованного текста — 2500 тыс. словоформ. Всего из этого текста извлечено 200 тыс. трехсловных сочетаний.

Извлечение словосочетаний осуществлялось по следующей процедуре. Из 600 тыс. словоформ, которые были отобраны путем сплошной выборки из различных публицистических статей, напечатанных в газетах «Правда», «Известия», «За рубежом», «Комсомольская правда» за 1964—1966 гг., было выбрано 120 наиболее употребительных имен существительных. От этих существительных прибавлением двух словоформ слева были образованы трехсловные словосочетания.

Сочетания, полученные в результате такой процедуры, являются, как правило, осмысленными и синтаксически оформленными. Лишь немногие словосочетания (приблизительно 40 из приведенного списка, насчитывающего около 1870 сочетаний) оказывались неоформленными с семантико-синтаксической точки зрения.

Как при отборе опорных словоформ, так и при формировании словосочетаний (табл. 1) мы придерживались следующих правил:

1) самостоятельной словоформой считается любая последовательность букв, ограниченная двумя пробелами;

2) дефис принимается за букву, и, таким образом, слово, написанное через черточку, представляет собой одну словоформу;

3) географические названия, состоящие из двух или более слов, выделяются в самостоятельную лексическую единицу (*Германская Демократическая Республика*, *Союз Советских Социалистических Республик*) и учитываются как одна словоформа.

4) общепринятые сокращения [*и*] *т. д.*, [*и*] *т. п.* и др. рассматриваются как одна словоформа;

5) все числительные, написанные в текстах цифрами, обозначаются одним символом — *ц*.

Падежи существительных, прилагательных и местоимений указываются лишь в том случае, если их нельзя определить по самому словосочетанию. Например: отношения *и* братской дружбы.

В табл. 2 приводится частотное распределение количества словосочетаний с частотой 11 и менее.

Сокращения и условные обозначения

в	— винительный падеж.	*собств*	— имя собственное.
д	— дательный ».	*т*	— творительный падеж.
и	— именительный ».	*ц*	— цифра.
п	— предложный ».	Δ	— знак пробела.
р	— родительный ».	ΔΔ	— « абзаца.

Таблица 1

Частотный список трехсловных сочетаний

i	Словосочетания	*F*	*m*
1	в *ц* году	1819	1
2	Верховного Совета СССР	1704	1
3	Совета Министров СССР	1288	1
4	во всем мире	914	1
5	министр иностранных дел	887	1
6	Президиума Верховного Совета	877	1
7	Председателя *р* Совета Министров	752	1
8	*собств* △ член	742	1
9	Председатель Совета Министров	724	1
10	в настоящее время	625	1
11	△△ председатель	603	1
12	США *р* во Вьетнаме	577	1
13	Министра *р* иностранных дел	552	1
14—15	△ во время; иностранных дел СССР	539	2
16	Первый секретарь ЦК	511	1
17	войны *р* во Вьетнаме	497	1
18	Коммунистической партии *р* Советского Союза	495	1
19	△ Председатель Совета	482	1
20	во внутренние дела	472	1
21	*ц* съезда КПСС	465	1
22	то же время	459	1
23	Октябрьской Социалистической революции	430	1
24	△ первый секретарь	425	1
25	и рабочих партий	422	1
26	*собств* △ заместитель	408	1

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
27	мира и безопасности	402	1
28	*собств* △ председатель	401	1
29	секретарь ЦК КПСС	400	1
30	член Политбюро ЦК	399	1
31	заместитель Председателя Совета	388	1
32	△△ Советский Союз	386	1
33	Политбюро *р* ЦК КПСС	376	1
34	△△ министр	359	1
35	в последнее время	358	1
36	△△ правительство	354	1
37	безопасности *р* в Европе	343	1
38	в *нашей* стране	322	1
39—40	второй мировой войны; ~△ Генеральный секретарь	321	2
41	международного коммунистического движения	319	1
42	△△ президент	313	1
43	против вьетнамского народа	309	1
44—47	агрессии *р* во Вьетнаме; ~ Генеральной Ассамблеи ООН; ~△ в связи; ~△ секретарь ЦК	308	4
48	*собств* △ министр	304	1
49	△ заместитель председателя	301	1
50—51	△ от имени; ~ нераспространении ядерного оружия	290	2
52	△ первый заместитель	279	1
53	Пленума ЦК КПСС	273	1
54	в прошлом году	272	1
55	в то время	267	1
56	свободу и независимость	264	1
57	△ член ЦК	254	1
58	△ наша партия	249	1
59	△ *ц* съезд	246	1
60	с *ц* года	245	1
61	△ коммунистическая партия	244	1
62	△ Председатель Президиума	243	1
63	мира в Европе	236	1
64	развития *р* народного хозяйства	230	1
65	*собств* △ секретарь	228	1
66—67	Центрального Комитета КПСС; ~ Национального фронта освобождения	222	2
68	комитета коммунистической партии	219	1
69—70	в этом году; ~ фронта освобождения Южного Вьетнама	217	2
71	между нашими странами	214	1
72	△ за мир	211	1
73—74	депутаты *в* Верховного Совета; ~ с советской стороны	204	2
75	член ЦК КПСС	202	1
76	против американской агрессии	201	1
77—78	войну во Вьетнаме; ~△ советский народ	196	2
79	△ что правительство	194	1
80	и советского правительства	191	1
81	мирового коммунистического движения	188	1
82	Генеральный секретарь ЦК	187	1
83	△△ делегация	185	1

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
84—85	△△ заместитель; ~ в этой связи	183	2
86	△△ член	182	1
87	социализма и коммунизма	179	1
88—89	и рабочего движения; ~ других социалистических стран	178	2
90	в свое время	177	1
91—92	△△ партия; ~△ Советское правительство	174	2
93	между двумя странами	165	1
94	министром иностранных дел	163	1
95	*ц* съезд *и* КПСС	161	1
96	Председателем Совета Министров	160	1
97	△ заместитель министра	159	1
98—100	△△ борьба; ~ война во Вьетнаме; ~ первый заместитель Председателя	156	3
101	решений *ц* съезда	154	1
102	КПСС *р* △ председатель	151	1
103	за последнее время	150	1
104—105	△ что Советский Союз; ~ Первым секретарем ЦК	148	2
106	△ в стране	147	1
107	секретаря *р* ЦК КПСС	145	1
108—109	△△ представители; ~ члены *в* Политбюро ЦК	141	2
110	участников *р* Варшавского Договора	139	1
111—112	△△ секретарь; ~ дружбы и сотрудничества	138	2
113	Президиум *и* Верховного Совета	133	1
114—116	Первый заместитель Министра; ~ партии *р* и народа; другими социалистическими странами	132	3
117	июня *ц* года	129	1
118	секретарь Центрального Комитета	125	1
119—120	△ в борьбе; ~ Политбюро *р* △ секретарь	124	2
121—122	△ посол СССР; ~ со своей стороны	122	2
123	агрессию во Вьетнаме	121	1
124—125	Политического консультативного комитета; ~ Государственного комитета Совета	119	2
126	борьбе *п* против империализма	118	1
127—132	△ министр обороны; ~ Социалистической единой партии; ~ Французской коммунистической партии; ~ партии *р* и правительства; ~ с другой стороны; ~ в той стране	117	6
133	января *ц* года	116	1
134	Президиума ЦК КПСС	115	1
135—136	министерство *и* иностранных дел; ~ министров *р* иностранных дел	113	2
137—138	мировой социалистической системы; ~ Коммунистическая партия Советского Союза	112	2
139—140	Председателю Совета Министров; ~ (ТАСС) △ президент	111	2
141	по приглашению ЦК	110	1
142—144	Комитета Совета Министров; ~ объединенной рабочей партии; ~ против агрессии США	109	3
145—148	в ближайшее время; ~ науки и культуры; ~ всего советского народа; ~ народного хозяйства СССР	108	4
149—150	декабря *ц* года; ~ Генерального секретаря *р* ЦК	107	2

Таблица 1 (*продолжение*)

i	Словосочетания	*F*	*m*
151	Вьетнамской народной армии	106	1
152	с социалистическими странами	104	1
153—155	июля *ц* года; ~ министерства *р* иностранных дел; ~ и Советское правительство	102	3
156	ноября *ц* года	101	1
157—159	секретарем ЦК КПСС; ~ экономики и культуры; ~ борьбы вьетнамского народа	100	3
160—161	*ц* съезде КПСС; ~ бойцы Армии освобождения	99	2
162—163	Совет *и* Министров СССР; ~ Совете Министров СССР	98	2
164—166	марта *ц* года; ~ Центральный Комитет и КПСС; ~ социалистической рабочей партии	97	3
167—168	борьбы за мир; ~ △ Пленум *и* ЦК	94	2
169—171	△ коммунистической партии; ~ за свою свободу; ~ с одной стороны	93	3
172	материально-технической базы коммунизма	92	1
173—176	△ бойцы армии; ~ октября *ц* года; ~ △ наш народ; ~ с вьетнамским народом	91	4
177—179	△△ народ; ~комитет *и* коммунистической партии; ~ заместитель Председателя Президиума	90	3
180	Монгольской народно-революционной партии	89	1
181	△ член Президиума	88	1
182—185	Великой Отечественной войны; ~ мая *ц* года; ~ Итальянской коммунистической партии; ~ съезд *и* коммунистической партии	87	4
186—188	до *ц* года; ~ △ *ц* лет; ~ и других стран	86	3
189—190	вклад *в* в дело; ~ △ вьетнамский народ	85	2
191—196	соглашений *ц* года; ~ мира и дружбы; ~ Всемирного Совета мира; ~ героического вьетнамского народа; ~ мировой системы социализма; ~ член Президиума ЦК	84	6
197—199	*ц* съездом КПСС; ~ Национальностей Верховного Совета; ~ Первого секретаря *р* ЦК	83	3
200—203	дружбы между народами; ~ *ц* съезда партии; ~ между обеими странами; ~ Армии *р* освобождения Южного Вьетнама	82	4
204	Членом Политбюро ЦК	81	1
205—207	△ ЦК *и* КПСС; ~ при Совете Министров; ~ декабрьского Пленума ЦК	80	3
208—213	в нынешнем году; ~ мирного сосуществования государств; ~ и безопасности *р* народов; ~ свободы и независимости; ~ △ на основе; ~ Румынской коммунистической партии	79	6
214—220	августа *ц* года; ~ агрессии *р* американского империализма; ~ стран и народов; ~ Болгарской коммунистической партии; ~ КПСС *р* △ секретарь; ~ правящие круги и США; ~ в области экономики	78	7
221—225	февраля *ц* года; ~ отделом ЦК КПСС; ~ борьбе *п* за мир; ~ весь советский народ; ~ и внешней политики	77	5
226—228	внутренние дела *в* государств; ~ Организации *р* африканского единства; ~ за национальную независимость	76	3

Таблица 1 (*продолжение*)

i	Словосочетания	*F*	*m*
229—231	против империалистической агрессии; ~ для дела мира; ~ △ правительство *и* США	74	3
232—233	имени *р* Центрального Комитета; ~ Верховный Совет *в* СССР	73	2
234—235	в его борьбе; ~ Соверховный Совет *и* СССР	72	2
236—239	Председатель Государственного комитета; ~ и Совет *и* Министров; ~ и Совета Министров; ~ Председатель Государственного Совета	71	4
240—244	△ глава делегации; ~ ревизионной комиссии *р* КПСС; ~ в защиту мира; ~ в строительстве социализма; ~ дела *в* других стран	70	5
245—247	новой мировой войны; ~ борьбы против империализма; ~ с зарубежными странами	69	3
248—251	войне *п* во Вьетнаме; ~ империализма во Вьетнаме; ~ △ секретарь Президиума; ~ сентября *ц* года	68	4
252—254	партии *р* трудящихся Вьетнама; ~ консультативного комитета государств; ~ распространения *р* ядерного оружия	67	3
255—260	участниц *р* Варшавского Договора; △ Совет *и* Министров; ~ ЦК *р* коммунистической партии; ~ заместителем Председателя Совета; ~ министра *р* обороны СССР; ~ братских социалистических стран	66	6
261—265	Центрального Комитета партии; ~ помощь *в* и поддержку; ~ всех прогрессивных сил; ~ мира и социализма; ~ в развитии экономики	65	5
266—271	в этой борьбе; ~ △ Председателю Президиума; ~ во всех странах; ~ после *ц* съезда и сельского хозяйства; ~ в члены ЦК	64	6
272—276	Пленум *и* ЦК КПСС; ~ КПСС *р* △ министр; ~ в интересах мира; ~ заместители *и* Председателя Совета; ~ △ секретарем ЦК	63	5
277—279	за это время; ~ дела *в* других государств; ~ претворение *в* в жизнь	62	3
280—282	△ председатель комитета; ~ социализма и мира; ~ и культурного строительства	61	3
283—285	октябре *ц* года; ~ двух германских государств; ~ члена *в* Политбюро ЦК	60	3
286—292	контролю во Вьетнаме; ~ мира во Вьетнаме; ~ международного рабочего класса; ~ Пленум *и* Центрального Комитета; ~ приглашению ЦК КПСС; ~ за *ц* лет; ~ в других странах	59	7
293—300	апреля *ц* года; ~ международного рабочего движения; ~ *собств* и министр; ~ заместитель министра обороны; ~ △ председателю совета; ~ демократии *р* и социализма; ~ Академии *р* наук СССР; ~ всех социалистических стран	58	8
301—307	вывода американских войск; ~ в наше время; ~ △ во главе; ~ борьбой вьетнамского народа; ~ министр национальной обороны; △ государственный секретарь; ~ со всеми странами	57	7

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
308—315	вклад *и* в дело; ~ напряженности *р* в Европе; ~ приглашению Центрального Комитета; ~ народов всего мира; ~ △ заместители Председателя; ~ Коммунистической партией Советского Союза; ~ в этих странах; ~ с агрессией США	56	8
316—320	в текущем году; ~ *собств* от имени; ~ об укреплении мира; ~ руководством коммунистической партии; ~ *собств* и член [делегации]	55	5
321—329	делегация во главе; ~ партии *р* во главе; ~ в будущем году; ~ всех советских людей; ~ борцов *р* за мир; ~ (ТАСС) △ правительство; ~ для дальнейшего развития; ~ пятилетнему плану развития; △ посол Советского Союза	54	9
330—333	претворения в жизнь; ~ генеральный секретарь ООН; ~ «пролетарской культурной революции»; ~ отношений между Советским Союзом	53	4
334—341	декабре *ц* года; ~ советских обществ дружбы; ~ и общественных организаций; ~ съезда Коммунистической партии; △ *собств* председатель; ~ выступил Первый секретарь; ~ Союза Верховного Совета; ~ Герой Социалистического труда	52	8
342—350	обеспечения европейской безопасности; ~ обоих германских государств; ~ обстановки в Европе; ~ КПСС *р* △ заместитель; ~ героическим вьетнамским народом; героическому вьетнамскому народу; ~ партия и правительство; ~ △ за свободу; ~ между социалистическими странами	51	9
351—359	агрессия во Вьетнаме; ~ мир *в* во Вьетнаме; ~ организаций американских государств; ~ против американского империализма; ~ секретарю Центрального Комитета; ~ △ национальную независимость; ~ главы *р* Советского правительства; ~ СЕПГ *р* △ председатель; ~ Генеральным секретарем ЦК	50	9
360—364	между нашими народами; ~ приглашению Советского правительства; ~ *собств* и председатель; ~ внешней торговли СССР; ~ отраслей народного хозяйства	49	5
365—369	в строительстве коммунизма; ~ американских вооруженных сил; ~ отношений с Советским Союзом; ~ вооруженных сил США; ~ отчетном докладе ЦК	48	5
370—376	мае *ц* года; ~ ноябре *ц* года; ~ глава советской делегации; ~ и национальную независимость; ~ всех революционных сил; ~ депутатов *р* Верховного Совета; ~ из *ц* стран	47	7
377—384	в их борьбе; ~ в это время; ~ партии *р* и государства; ~ министр внутренних дел; ~ △ Совета Министров; ~ Национальным фронтом освобождения; ~ со стороны Советского Союза; ~ против арабских стран	46	8
385—392	после *ц* года; ~ борьбу против империализма; ~ председателя *р* Государственного комитета; ~ борьбу за мир; ~ в современном мире; ~ △ министры СССР; ~ во многих странах; ~ в дело укрепления	45	8

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
393—400	в Совете Безопасности; ~ Пленумов ЦК КПСС; ~ всех честных людей; ~ и Совет *в* Министров; ~ (ТАСС) △ председатель; ~ *собств* △ президент; ~ иностранных дел стран; ~ *ц* Пленума ЦК	44	8
401—409	системы *р* коллективной безопасности; ~ начале *ц* года; ~ (ТАСС) △ министр; ~ за укрепление мира; ~ генерального секретаря *р* ООН; ~ областной партийной организации; ~ Французская коммунистическая партия; ~ всех антиимпериалистических сил; ~ Союза и Совета	43	9
410—422	и международной безопасности; ~ △ что борьба; ~ войной во Вьетнаме; ~ январе *ц* года; ~ положения *р* в Европе; ~ борьба за мир; ~ поддержку и помощь; ~ и прогрессивных сил; ~ социалистических левых сил; ~ △ Председателем Совета между народами Советского Союза; ~ государственный секретарь США; ~ заведующего *р* отделом ЦК	42	13
423—438	протеста против агрессии; ~ прочь от Вьетнама; ~ после *ц* года; ~ *собств* △ делегация; ~ вкладом в дело; ~ стран Варшавского Договора; ~ докладе ЦК КПСС; ~ со всеми народами; ~ на этой основе; ~ секретарь коммунистической партии; ~ △ в поддержку; ~ △ господин председатель; ~ заявил Генеральный секретарь; ~ нашими двумя странами; ~ △ президент США; ~ директив *ц* съезда	41	16
439—452	что во время; ~ границ в Европе; ~ △ большое значение; ~ в ЦК *п* КПСС; ~ движения *р* на основе; ~ △ что президент; и демократических сил; ~ межамериканских вооруженных сил; ~ председатель Революционного Совета; ~ СССР *р* Маршал Советского Союза; ~ с другими странами; ~ в капиталистических странах; ~ Совета депутатов трудящихся; ~ развития сельского хозяйства	40	14
453—464	Советского Союза во главе; ~ и национально-освободительного движения; ~ △ члены делегации; ~ при ЦК КПСС; ~ председателя *р* кабинета министров; ~ о предоставлении независимости; ~ ограждении их независимости; ~ Национальный фронт *и* освобождения; ~ △ заместителем председателя; ~ и Председатель Совета; ~ с Председателем Совета; ~ △ правительство *и* ФРГ	39	12
465—478	во время войны; ~ империалистов *р* во Вьетнаме; ~ последние *в* два года; ~ конце *ц* года; ~ министру иностранных дел; ~ Советской партийно-правительственной делегации; ~ △ в Европе; ~ помощи *р* вьетнамскому народу; ~ △ что партия; ~ и внешнюю политику; ~ заявление *и* Советского правительства; ~ борьбе *п* за свободу; ~ отделами МИД СССР; ~ в деле укрепления	38	14

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
479—494	△ за время; ~ △ ц года; ~ после ц года; ~ и советского народа; ~ поддержку вьетнамского народа; ~ дружбу между народами; ~ и безопасность в народов; ~ и независимости р народов; ~ и Советского народов; ~ с заместителем председателя; ~ многосторонних ядерных сил; ~ революционных вооруженных сил; ~ отделом МИД СССР; ~ в своей стране; ~ Коммунистической партии р США; ~ что правительство и США	37	16
495—512	в годы войны; ~ войне д во Вьетнаме; ~ Соединенных Штатов во Вьетнаме; ~ делегации р во главе; ~ июне ц года; ~ сентябре ц года; ~ других социалистических государств; ~ и Советского государства; ~ министры и иностранных дел; ~ Пленума Центрального Комитета; ~ движения р за мир; ~ △ чтобы правительство; ~ депутат Верховного Совета; ~ сессия Верховного Совета; ~ решения и ц съезда; ~ решениями ц съезда; ~ Советов депутатов трудящихся; ~ △ секретаря р ЦК	36	18
513—538	сил в борьбе; ~ успехов в борьбе; ~ *собств* во время; ~ агрессии п во Вьетнаме; ~ агрессоров р во Вьетнаме; ~ осенью ц года; ~ в честь делегации; ~ Советской правительственной делегации; ~ △ советская делегация; ~ △ в дело; ~ комитета защиты мира; ~ прогресса и мира; ~ борьбе д вьетнамского народа; ~ борьбу вьетнамского народа; ~ генеральному секретарю ООН; ~ первым заместителем председателя; ~ эффективности р общественного производства; ~ △△ свободу; ~ развития производительных сил; ~ п всестороннего сотрудничества; ~ парламентской группы СССР; ~ поддержку со стороны; ~ государственного секретаря р США; ~ министр обороны США; ~ и всех трудящихся; ~△ Пленума ЦК	35	26
539—561	агрессии д во Вьетнаме; ~ военщины во Вьетнаме; ~ весной ц года; ~ выборах ц года; ~ феврале ц года; ~ в минувшем году; ~ в состав делегации; ~ в области культуры; ~ до ц лет; ~ широких народных масс; ~ △ что народ; ~ всего немецкого народа; ~ за свою независимость; ~ ц съездом партии; ~ в честь председателя; ~ *собств* в связи; ~ выступил Генеральный секретарь; ~ депутаты и Верховного Совета; ~ △ Маршал Советского Союза; ~ ц-го флота США; ~ иностранных дел ФРГ; ~ войск из Южного Вьетнама; ~ фронтом освобождения Южного Вьетнама	34	23
562—581	против грязной войны; ~ против арабских государств; ~ претворить в жизнь; ~ секретарь МГК КПСС; ~ члены и ЦК КПСС; ~ хозяйства р и культуры; ~ последние ц лет; ~ Председателя в Совета Министров; ~ борьбе д за мир; ~ в арабском мире; ~ жить в мире; ~ первого р в мире; ~ и советским народами; ~ сотрудничества между народами; ~	33	20

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
	в рамках НАТО; ~ Генеральная Асамблея ООН; ~ ц съезд и партии; ~ в ряде стран; ~ △ сельского хозяйства; ~ и член ЦК		
582—599	США и во Вьетнаме; ~ всех европейских государств; ~ министра в иностранных дел; ~ вмешательства р в дела; ~ Советская правительственная делегация; ~ секретаря в ЦК КПСС; ~ для всеобщего мира; ~ всему советскому народу; ~ △ в основе; ~ под руководством партии; ~ называемой «культурной» революции; ~ всех миролюбивых сил; ~ сотрудничества между Советским Союзом; ~ молодежных организаций СССР; ~ △ со стороны; ~ братскими социалистическими странами; ~ ответственные работники и ЦК; ~ члены и Политбюро ЦК	32	18
600—614	вопросам европейской безопасности; ~ августе ц года; ~ марте ц года; ~ жизненного уровня народа; ~ испытаний ядерного оружия; ~ нераспространения ядерного оружия; ~ ЦК и коммунистической партии; ~ заявление в Советского правительства; ~ с членом Президиума; ~ делегация Верховного Совета; ~ сессии р Верховного Совета; ~ государственного бюджета СССР; ~ правительств социалистических стран; ~ в социалистических странах; ~ со стороны США	31	15
615—635	вклад в в борьбу; ~ △ во Вьетнаме; ~ участники Варшавского Договора; ~ США р в Европе; ~ △ против империализма; ~ △ несколько лет; ~ течение в ц лет; ~ уже ц лет; ~ за дело мира; ~ из военной организации; ~ применения р ядерного оружия; ~ Фронта национального освобождения; ~ Болгарская коммунистическая партия; ~ и внутренней политики; ~ глава Советского правительства; ~ революционного рабоче-крестьянского правительства; ~ КПСС р △ Президиума; ~ сессии р Верховного Совета; ~ Союза писателей СССР; ~ правящих кругов США; ~ члена р Политбюро ЦК	30	21
636—666	системы р европейской безопасности; ~ Первой мировой войны; ~ △ три года; ~ декабре прошлого года; ~ министерстве иностранных дел; ~ миру в Европе; ~ △ особое значение; ~ придают большое значение; ~ ее Центрального Комитета; ~ работники ЦК КПСС; ~ трудящихся р всего мира; ~ весь наш народ; ~ и советский народ; ~ борьба вьетнамского народа; ~ братского вьетнамского народа; ~ государств и народов; ~ с советским народом; ~ делегация коммунистической партии; ~ Центральный Комитет и партии; ~ других братских партий; ~ против агрессивной политики; ~ вопросы и дальнейшего развития; ~ заместителя р Председателя Совета; ~ △ трудящихся р Советского Союза; ~ с помощью Советского Союза; ~ министр обороны	29	31

i	Словосочетания	F	m
	СССР; ~ трудящиеся нашей страны; ~ задач коммунистического строительства; ~ что *ц* съезд; ~ секретарь ЦК; ~ Первого секретаря *в* ЦК		
667—687	КПСС *р* во главе; ~ периодом прошлого года; ~ соглашения *и ц* года; ~ претворение *и* в жизнь; ~ секретарем Центрального Комитета; ~ Пленуме ЦК КПСС; ~ более *ц* лет; ~ на *ц* лет; ~ △ бывший министр; ~ *ц* стран мира; ~ △ в мире; ~ братскому вьетнамскому народу; ~ △ английское правительство; ~ заместителем Председателя Президиума; ~ и культурного сотрудничества; ~ материально-технической базы социализма; ~ других арабских стран; ~ народов обеих стран; ~ против социалистических стран; ~ по пути строительства; ~ членом Президиума ЦК	28	21
688—716	в результате агрессии; ~ в любое время; ~ войск из Вьетнама; ~ начале этого года; ~ совещания *р ц* года; ~ △ национально-освободительного движения; ~ строительстве новой жизни; ~ борьбе *п* против империализма; ~ секретарю ЦК КПСС; ~ *ц* съезду КПСС; ~ у советских людей; ~ других районах мира; ~ △ монгольский народ; ~ что советский народ; ~ и национальной независимости; ~ заявление *и* в связи; ~ делегации *р* Верховного Совета; ~ Президиуму Верховного Совета; ~ *собств* председатель Совета; ~ отношения *и* между Советским Союзом; ~ △ правительство *в* СССР; ~ сельского хозяйства СССР; ~ единства социалистических стран; ~ в арабских странах; ~ в некоторых странах; ~ в своих странах; ~ работе *п ц* съезда; ~ мартовского Пленума ЦК; ~ постоянного Президиума ЦК	27	29
717—746	прекращения американской агрессии; ~ против преступной агрессии; ~ за прекращение войны; ~ с народом Вьетнама; ~ конце прошлого года; ~ летом *ц* года; ~ в составе делегации; ~ △ наша делегация; ~ *собств* и заместитель; ~ Пленумом ЦК КПСС; ~ в дело мира; ~ солидарности *р* с народами; ~ странами и народами; ~ комитету Коммунистической партии; ~ *ц* съезду партии; ~ что его правительство; ~ *собств* △ представитель; ~ △ заявил президент; ~ пятилетнего плана развития; ~ Совета и Совета; ~ в пользу социализма; ~ Арабского Социалистического Союза; ~ глав *р* правительств стран; ~ в жизни страны; ~ решения *в ц* съезда; ~ повышения производительности труда; ~ на территории ФРГ; ~ Генеральному секретарю ЦК; ~ заведующий отделом ЦК; ~ фронт *и* освобождения Южного Вьетнама	26	30
747—776	борьбе *п* против агрессии; ~ △ Совет *и* Безопасности; ~ в общей борьбе; ~ его справедливой борьбе; ~ в мирное время; ~ войск во Вьетнаме; ~ действий во Вьетнаме; ~ положения *р* во Вьетнаме; ~ против национально-освободительного движения; ~ министра *р* внутренних дел; ~ обстановке *п* братской дружбы; ~ рабочего класса; ~ △ двадцать лет; ~ △ заявил министр; ~ на благо мира; ~ прогрессивных сил мира; ~ партия и народ; ~ сессии *р* Совета НАТО; ~ на принципиальной основе; ~ области *п* внешней политики; ~ советской внешней политики; ~ МНРП *р* △ председатель; ~ мир *в* и свободу; ~ Советского Союза в связи; ~ из социалистических стран; ~ народов наших стран; ~ в африканских странах; ~ ЦК *р* Партии трудящихся; ~ правящих кругов ФРГ; ~ развитии народного хозяйства	25	30

i	Словосочетания	F	m
777—810	национальной народной армии; ~ вон из Вьетнама; ~ против народа Вьетнама; ~ миру во Вьетнаме; ~ △ два года; ~ свои внутренние дела; ~ придают большое значение; ~ мне от имени; ~ агрессивных сил империализма; ~ Президиума Центрального Комитета; ~ Политбюро *и* ЦК КПСС; ~ во имя мира; ~ всех стран мира; ~ сил в мире; ~ уровня жизни народа; ~ △ членов *в* организации; ~ Национальный фронт *и* освобождения; ~ Генеральный секретарь партии; ~ СССР *р* △ председатель; ~ на пост Председателя; ~ от имени Президиума; ~ делам «культурной революции»; ~ независимость *в* и свободу; ~ всех левых сил; ~ и миролюбивых сил; ~ △ председателя *в* совета; ~ депутатами Верховного Совета; ~ заместителя *в* Председателя Совета; ~ председателем Государственного совета; ~ сотрудничества с Советским Союзом; ~ политической жизни *п* страны; ~ империалистической агрессии *р* США; ~ в работе съезда; ~ от имени ЦК	24	34
811—865	по вопросам безопасности; ~ △ эта борьба; ~ в своей борьбе; ~ в дело борьбы; ~ поддержку справедливой борьбы; ~ героического народа Вьетнама; ~ уйти из Вьетнама; ~ агрессией во Вьетнаме; ~ текущем финансовом году; ~ молодых независимых государств; ~ мире социалистического государства; ~ безопасность *и* в Европе; ~ безопасность *в* в Европе; ~ и в Европе; ~ сотрудничества в Европе; ~ имеет большое значение; ~ △ делегация КПСС; ~ *ц* съезд *в* КПСС; ~ уже несколько лет; ~ народов за мир; ~ безопасности *р* и мира; ~ комитет *и* защиты мира; ~ различных районах мира; ~ отношений между народами; ~ конференции *р* солидарности народов; ~ на благо народов; ~ организации *р* солидарности народов; ~ всем советским народом; ~ с героическим народом; ~ солидарности *р* с народом; ~ △ министерство *и* обороны; ~ Генеральной Ассамблее *п* ООН; ~ вокруг коммунистической партии; ~ славная Коммунистическая партия; ~ △ федеральное правительство; ~ △ господин президент; ~ по приглашению Президиума; ~ △ за развитие; ~ КПСС *р* и Совета; ~ и международного	23	55

i	Словосочетания	F	m
	сотрудничества; ~ Верховным Советом СССР; ~ коллегии *р* МИД СССР; ~ для обеих стран; ~ в нашей стране; ~ народов нашей страны; ~ делегаты *ц* съезда; ~ директивы *и ц* съезда; ~ работы *р ц* съезда; ~ Героя *р* Социалистического труда; ~ роста производительности труда; ~ жизненного уровня трудящихся; ~ развития и укрепления; ~ △ Политбюро *и* ЦК; ~ международным отделом ЦК; ~ члены *в* Президиума ЦК		
866—916	Комитета государственной безопасности; ~ роль *в* в борьбе; ~ агрессии *р* и войны; ~ протеста против войны; ~ всех иностранных войск; ~ △ некоторое время; ~ урегулирования во Вьетнаме; ~ СССР *р* во главе; ~ ноябре прошлого года; ~ осенью прошлого года; ~ с половиной года; ~ в юбилейном году; ~ решать свои дела; ~ обстановку в Европе; ~ для укрепления единства; ~ течение *в* многих лет; ~ на весь мир; ~ △ народы мира; ~ войны и мира; ~ *и* всего мира; ~ и упрочения мира; ~ △ корейский народ; ~ что вьетнамский народы; ~ всего нашего народа; ~ в интересах народа; ~ всех европейских народов; ~ и коммунистической партии; ~ комитета нашей партии; ~ руководители коммунистической партии; ~ ЦК *р* Трудовой партии; ~ от имени правительства; ~ партии *р* △ председатель; ~ ПОРП *р* △ председатель; ~ △ Указ *и* Президиума; ~ в члены Президиума; ~ *и* пищевой промышленности; ~ △ председателя *р* Совета; ~ приглашению Верховного Совета; ~ *о политике Советского Союза*; ~ партийно-правительственной делегации *р* Советского Союза; ~ по инициативе Советского Союза; ~ председатель Совета Союза; ~ государственном бюджете СССР; ~ народам арабских стран; ~ всех концов страны; ~ на севере страны; ~ работу *ц* съезд; △ делегаты съезда; ~ социалистического разделения труда; ~ всего народного хозяйства; ~ развитии сельского хозяйства	22	51
917—961	прекращение *в* американской агрессии; ~ укрепления европейской безопасности; ~ действия *р* в борьбе; ~ △ народ Вьетнама; ~ комитета *ц* государств; ~ мирного существования государств; ~ сплоченность *в* коммунистического движения; ~ министерства *р* внутренних дел; ~ за укрепление единства; ~ претворению в жизнь; ~ △ важное значение; ~ Ленинского Центрального Комитета; ~ правительство *и* и народ; ~ права *в* вьетнамского народа; ~ правительства *р* и народа; ~ и независимость *в* народов; ~ помощь *в* вьетнамскому народу; ~ центрального правления общества; ~ комитет *и* нашей партии; ~ председатель коммунистической партии; ~ съезде коммунистической партии; ~ народа и правительства; ~ по приглашению правительства; ~ на-	21	45

i	Словосочетания	F	m
	рода за свободу; ~ озабоченность *в* в связи; ~ что в связи; ~ и миролюбивых сил; ~ в депутаты Совета; ~ и председателем совета; ~ на заседании совета; ~ председателя *р* Государственного Совета; ~ △ между Советским Союзом; ~ внешней политики Советского Союза; ~ и научно-технического сотрудничества; ~ также посол СССР; ~ народов арабских стран; ~ и другими странами; ~ и социалистическими странами; ~ *положения р* в стране; ~ на юге страны; ~ разведовательного управления США; ~ решениях *ц* съезда; ~ Всеобщей конфедерации *р* труда; ~ △ всех трудящихся; ~ докладом выступил член [делегации]		
962—1022	народов в борьбе; ~ в состоянии войны; ~ и во время; ~ интервенции *р* во Вьетнаме; ~ последние *в* четыре года; ~ революции *р ц* года; ~ △ за дело; ~ за порученное дело; ~ о принципах деятельности; ~ в обстановке дружбы; ~ на укрепление единства; ~ агрессивной политики империализма; ~ народов против империализма; ~ политики *р* американского империализма; ~ докладе Центрального Комитета; ~ секретаря *р* Центрального Комитета; ~ протяжении многих лет; ~ течение *в* ряда лет; ~ уже много лет; ~ жизни *р* советских людей; ~ за прочный мир; ~ и укрепления *р* мира; ~ на страже мира; ~ укрепления *р* всеобщего мира; ~ этом районе мира; ~ △ румынский народ; ~ жизни *р* советского народа; ~ всех миролюбивых народов; ~ и безопасности *д* народов; ~ вооруженными силами НАТО; ~ борьбы за независимость; ~ строительства нового общества; ~ от ядерного оружия; ~ △ председатель партии; ~ Объединенной социалистической партии; ~ политики *р* нашей партии; ~ руководителей *р* коммунистической партии; ~ ленинской национальной политики; ~ нашей внешней политики; ~ визит *в* Председателю Президиума; ~ объем *и* промышленного производства; ~ Великой Октябрьской [социалистической] революции; ~ и в связи; ~ и председателю совета; ~ иностранным делам Совета; ~ △ правительство *и* Советского Союза; ~ Коммунистической партии *р* Советского Союза; ~ по пути социализма; ~ всеми социалистическими странами; ~ между европейскими странами; ~ с капиталистическими странами; ~ в обеих странах; ~ народного хозяйства страны; ~ трудящихся *р* нашей страны; ~ и коммунистического строительства; ~ △ агрессия США; ~ имени *р ц* съезда; ~ Швейцарской партии *р* труда; ~ отрасли *в* народного хозяйства; ~ и секретарь ЦК; ~ *ц* Пленум *и* ЦК; ~ *ц* Пленуме ЦК	20	61
1023—1094	управления *р* Советской Армии; ~ и европейской безопасности; ~ переднем крае борьбы; ~ вывод *в* аме-	19	72

i	Словосочетания	F	m
	риканских войск; ~ же самое время; ~ Отечественного фронта Вьетнама; ~ события *и* во Вьетнаме; ~ внутренние дела *и* государства; ~ мирное сосуществование *в* государств; ~ △ глава государства; ~ главы *р* советской делегации; ~ расследованию антиамериканской деятельности; ~ в Доме дружбы; ~ мир *в* в Европе; ~ претворяя в жизнь; ~ проводить в жизнь; ~ союз *и* рабочего класса; ~ Московского городского комитета; ~ за пять лет; ~ и упрочение *в* мира; ~ по поддержанию мира; ~ в капиталистическом мире; ~ место *в* в мире; ~ △ болгарский народ; ~ для нашего народа; ~ дружба между народами; ~ мира между народами; ~ и дружбы народов; ~ интересам всех народов; ~ с борьбой народов; ~ к советскому народу; ~ помочь вьетнамскому народу; ~ △ национальной независимости; ~ распространению ядерного оружия; ~ △ национального освобождения; ~ разрыве дипломатических отношений; ~ Генеральным секретарем партии; ~ Демократической левой партии; ~ что коммунистическая партия; ~ что наша партия; ~ и военную помощь; ~ △ американское правительство; ~ БКП *р* △ председатель; ~ РКП *р* △ председатель; ~ *собств* заместитель председателя; ~ и кооперирования производства; ~ вклад *и* в развитие; ~ вопросам дальнейшего развития; ~ укрепления *р* и развития; ~ ходе культурной революции; ~ народов за свободу; ~ президиума △ секретарь; ~ связей с Советским Союзом; ~ в знак солидарности; ~ других стран социализма; ~ европейских социалистических стран; ~ из братских стран; ~ наших двух стран; ~ отношений между странами; ~ △ в странах; ~ в наших странах; ~ других социалистических странах; ~ обстановка в стране; ~ развития нашей страны; ~ △ империалисты США; ~ грязной войны США; ~ с трибуны съезда; ~ научной организации *р* труда; ~ широких масс трудящихся; ~ труженики сельского хозяйства; ~ △ орган *и* ЦК; ~ на Пленуме ЦК		
1095—1162	в совместной борьбе; ~ и в борьбе; ~ на вьетнамскую войну; ~ мировой термоядерной войны; ~ △ долгое время; ~ героическим народом Вьетнама; ~ в течение года; ~ весной этого года; ~ за четыре года; ~ летом прошлого года; ~ *ц* финансовом году; ~ польской партийно-правительственной делегации; ~ в общее дело; ~ общества *р* советско-вьетнамской дружбы; ~ положение *и* в Европе; ~ всех областях жизни; ~ и культурной жизни; ~ культурного уровня жизни; ~ проведение *в* в жизнь; ~ колониализма и империализма; ~ деятельность *в* Центрального Комитета; ~ Пленуме Центрального Комитета; ~ за построение коммунизма; ~ кандидатов *в* блока коммунистов; ~ △ представители КПСС; ~ еще несколько лет; ~ последние *в* пять лет; ~ течение *в* нескольких лет; ~ ПОРП *р* △ министр; ~ за всеобщий мир; ~ мир *в* между народами; ~ в глазах народов; ~ после провозглашения независимости; ~ и министр обороны; ~ строительстве нового общества; ~ Совета Безопасности ООН; ~ комитета молодежных организаций; ~ первичных партийных организаций; ~ *ц*-й годовщины освобождения; ~ *ц* съезде партии; ~ всех коммунистических партий; ~ всю необходимую помощь; ~ △ новое правительство; ~ КПЧ *р* △ председатель; ~ с председателем Президиума; ~ вопросам культурной революции; ~ и Генеральный секретарь; ~ министр вооруженных сил; ~ палат Верховного Совета; ~ отношения *в* с Советским Союзом; ~ Франции и Советского Союза; ~ государственными флагами Советского Союза; ~ Франции *р* и Советского Союза; ~ что касается Советского Союза; ~ и братского сотрудничества; ~ всех стран социализма; ~ в Совет Союза; ~ международной политике *и* СССР; ~ агрессии *р* со стороны; ~ народы *и* наших стран; ~ и других странах; ~ власть *в* в стране; ~ и социалистического строительства; ~ агрессивных действий США; ~ государственного департамента США; ~ *в интересах* укрепления; ~ отделом пропаганды ЦК; ~ *ц* Пленума ЦК народа против агрессии; ~ проблем европейской безопасности; ~ проблемам европейской безопасности; ~ △ в борьбу; ~ в поддержку борьбы; ~ после окончания войны; ~ *собств* во Вьетнаме; ~ △ из года; ~ концу *ц* года; ~ других арабских государств; ~ об отказе государств; ~ принципах деятельности государств; ~ мирового революционного движения; ~ их внутренние дела; ~ △ участники договора; ~ существующих *р* в Европе; ~ △ укрепление *и* единства; ~ партии *р* пролетарского единства; ~ и духовной жизни; ~ строительства новой жизни; ~ имеет важное значение; ~ имеет огромное значение; ~ городского комитета КПСС; ~ △ пять лет; ~ за семь лет; ~ исполняется *ц* лет; ~ с лишним лет; ~ десятки *и* тысяч людей; ~ миллионов советских людей; ~ △ государственный министр; ~ △ Совету Министров; ~ всех народов мира; ~ движения *р* сторонников мира; ~ коммунистов *р* всего мира; ~ межамериканских сил мира; ~ героический вьетнамский народ; ~ борьбе *п* вьетнамского народа; ~ сотрудничество *и* между народами; ~ для всех народов; ~ дружбы наших народов; ~ с борющимся народом; ~ братскому советскому народу; ~ великому советскому народу; ~ в отношении НАТО; ~ части *и* противовоздушной обороны; ~ и других организаций; ~ политики *р*	18	68
1163—1252		17	90

i	Словосочетания	F	m
	коммунистической партии; ~ съезда нашей партии; ~ △ социалистическая партия; ~ объединенная рабочая партия; ~ △ на поддержку; ~ своей внешней политики; ~ △ если правительство; ~ что Советское правительство; ~ *собств* △ председателя; ~ на дальнейшее развитие; ~ △ дальнейшего развития; ~ в тесной связи; ~ делам вооруженных сил; ~ наших вооруженных сил; ~ постепенных межамериканских сил; ~ Московского городского Совета; ~ президиума Всемирного Совета; ~ честь *в* Председателя Совета; ~ отношения *и* с Советским Союзом; ~ сотрудничество *и* между Советским Союзом; ~ △ Герой Советского Союза; ~ завершения строительства социализма; ~ с французской стороны; ~ комитет *и* солидарности стран; ~ сотрудничества социалистических стран; ~ сплоченности *р* социалистических стран; ~ в латиноамериканских странах; ~ в освободившихся странах; ~ в развивающихся странах; ~ в *ц* странах; ~ во всей стране; ~ сейчас в стране; ~ демократических сил страны; ~ △ посол США; ~ морской пехоты США; ~ правящие круги *в* США; ~ представителей *р* конгресса США; ~ △ производительность *и* труда; ~ △ для укрепления ; ~ что правительство *и* ФРГ; ~ △ секретаря *в* ЦК; ~ △ члены ЦК; ~ деятельность *в* Политбюро ЦК; ~ с секретарем ЦК		
1253—1346	△ против агрессии; ~ отражения американской агрессии; ~ △ генерал армии; ~ △ это борьба; ~ в нашей борьбе; ~ народа в борьбе; ~ СССР *р* и борьбе; ~ новую мировую войну; ~ вывода израильских войск; ~ все это время; ~ борющимся народом Вьетнама; ~ Соединенных Штатов Америки во Вьетнаме; ~ к концу года; ~ последние *в* три года; ~ январе этого года; ~ в истекшем году; ~ и коммунистического движения; ~ и освободительного движения; ~ министерством иностранных дел; ~ НАТО *р* в Европе; ~ их *в* в жизнь; ~ претворены в жизнь; ~ ТПК *р* △ заместитель; ~ доклад *в* Центрального Комитета; ~ член Исполнительного комитета; ~ пути *д* строительства коммунизма; ~ *ц* тысяч коммунистов; ~ членом ЦК КПСС; ~ последние *в* десять лет; ~ △ полномочный министр; ~ премьер-министр и министр; ~ приняли участие министр; ~ *собств* и министра; ~ в пользу мира; ~ в условиях мира ; ~ для обеспечения мира; ~ за сохранение мира; ~ волю вьетнамского народа; ~ всего вьетнамского народа; ~ материального благосостояния народа; ~ стороне *п* вьетнамского народа; ~ дружбы с народами; ~ дела *в* других народов; ~ национально-освободительной борьбы народов; ~ освободительной борьбы народов; ~ партией и народом; ~ по-	16	94

i	Словосочетания	F	m
	мощь *и* вьетнамскому народу; ~ свободу и независимость; ~ министра *р* национальной обороны; ~ развитого социалистического общества; ~ и водородного оружия; ~ социалистических общественных отношений; ~ установления *р* дипломатических отношений; ~ Итальянской социалистической партии; ~ неонацистской национал-демократической партии; ~ против коммунистической партии; ~ против нашей партии; ~ и коммунистическая партия; ~ солидарность *в* и поддержку; ~ открыл Первый секретарь; ~ провожали Генеральный секретарь; ~ СЕПГ *р* △ секретарь; ~ против агрессивных сил; ~ визит *в* Председателю Совета; ~ между Председателем Совета; ~ премьера *р* Государственного Совета; ~ △ которую Советский Союз; ~ △ народы Советского Союза; ~ Советский комитет *и* солидарности; ~ мира и сотрудничества; ~ отношений и сотрудничества; ~ связей и сотрудничества; ~ мировая система социализма; ~ Советом Министров СССР; ~ из других стран; ~ из этих стран; ~ народов всех стран; ~ содружества социалистических стран; ~ сплоченность *и* социалистических стран; ~ трудящихся *р* всех стран; ~ с арабскими странами; ~ баз в странах; ~ в тех странах; ~ обстановку в стране; ~ сил в стране; ~ улучшения *р* условий труда; ~ △ представители трудящихся; ~ в интересах трудящихся; ~ США *р* и ФРГ; ~ министерства *р* сельского хозяйства; ~ Генерального секретаря *в* ЦК; ~ сентябрьского Пленума ЦК; ~ за освобождение Южного Вьетнама; ~ Фронт *в* освобождения Южного Вьетнама		
1347—1451	немедленного прекращения агрессии; ~ обеспечение *в* европейской безопасности; ~ укрепление *в* европейской безопасности; ~ стран в борьбе; ~ свою агрессивную войну; ~ агрессии *р* против Вьетнама; ~ войска *в* из Вьетнама; ~ событий во Вьетнаме; ~ что во Вьетнаме; ~ дружбы во главе; ~ апреле этого года; ~ выборов *ц* года; ~ границах *ц* года; ~ октябре прошлого года; ~ единства коммунистического движения; ~ США *р* в дела; ~ члены советской делегации; ~ △ партийно-правительственная делегация; ~ организации *р* Варшавского Договора; ~ членов *р* Варшавского Договора; ~ Общества *р* германо-советской дружбы; ~ сил в Европе; ~ претворять в жизнь; ~ имеет особое значение; ~ особо важное значение; ~ которой *п* от имени; ~ борьба против империализма; ~ врага *р* △ империализма; ~ интересы *в* рабочего класса; ~ заместитель председателя комитета; ~ Политбюро *р* Центрального Комитета; ~ делами ЦК КПСС; ~ члены *в* ЦК КПСС; ~ △ десять лет; ~ △ тысячи *и* людей; ~ СЕПГ *р* △ министр; ~ сессия Совета Министров; ~	15	105

i	Словосочетания	F	m
	делу укрепления мира; ~ народы *и* всего мира; ~ трудящиеся всего мира; ~ положение *в* в мире; ~ самого китайского народа; ~ свободолюбивого вьетнамского народа; ~ в борьбе народов; ~ всех свободолюбивых народов; ~ конференция солидарности народов; ~ национально-освободительного движения народов; ~ оказывать вьетнамскому народу; ~ военной организации *д* НАТО; ~ суверенитет *в* и независимость; ~ по вопросам обороны; ~ развития социалистического общества; ~ странами на основе; ~ только на основе; ~ социалистических производственных отношений; ~ политику коммунистической партии; ~ Португальской коммунистической партии; ~ наша Коммунистическая партия; ~ американской внешней политики; ~ вопросам внешней политики; ~ △ на помощь; ~ и материальную помощь; ~ △ члены правительства; ~ в заявлении *правительства*; ~ *в области* промышленности; ~ в дальнейшее развитие; ~ вопросы *в* дальнейшего развития; ~ государственного плана развития; ~ и социального развития; ~ путь *в* самостоятельного развития; ~ борцы за свободу; ~ поздравления *в* в связи; ~ заявил Первый секретарь; ~ всех демократических сил; ~ состоялось заседание Совета; ~ СССР *р* △ Совета; ~ связей между Советским Союзом; ~ и правительства *р* Советского Союза; ~ первой конференции *р* солидарности ; ~ в странах социализма; ~ за построение социализма; ~ Верховному Совету СССР; ~ внешней политики СССР; ~ заместители министров СССР; ~ Красного Полумесяца СССР; ~ Союза журналистов СССР; ~ △ социалистических стран; ~ народами наших стран; ~ народов социалистических стран; ~ в различных странах; ~ для небольшой страны; ~ других районах страны; ~ из этой страны; ~ каждой социалистической страны; ~ партии *р* и страны; ~ в деле строительства; ~ государственный департамент *и* США; ~ с помощью США; ~ открылся *ц* съезд; ~ *ц* △ съезд; ~ Советы *в* депутатов трудящихся; ~ для дальнейшего укрепления; ~ развитию сельского хозяйства; ~ и секретаря *р* ЦК; ~ сентябрьского Пленумов ЦК		
1452—1577	в справедливой борьбе; ~ их справедливой борьбе; ~ поддерживают справедливую борьбу; ~ помощь *в* в борьбе; ~ угрозы *р* новой войны; ~ отвода израильских войск; ~ Германии *р* во главе; ~ республики *р* во главе; ~ летом этого года; ~ мае этого года; ~ месяцев этого года; ~ от *ц* года; ~ плана *ц* года; ~ сентябре прошлого года; ~ в новом году; ~ всех социалистических государств; ~ вмешательство *в* в дела; ~ дня подписания договора; ~ что в Европе; ~ ее *в* в жизнь; ~ стороны *р* △ заместитель; ~ △ огромное значение; ~ Пленумов Центрального Комитета; ~ члены *и* Центрального Комитета; ~ дело *в* строительства коммунизма; ~ за победу коммунизма; ~ Центральному Комитету КПСС; ~ исполнилось *ц* лет; ~ на пять лет; ~ стороны *р* △ министр; ~ *собств* △ министра; ~ △ заместители министров; ~ заседании Совета Министров; ~ борцы за мир; ~ Всемирный Совет *и* Мира; ~ дела *р* всеобщего мира; ~ для укрепления мира; ~ на укрепление мира; ~ народами всего мира; ~ благосостояния советского народа; ~ дела *в* вьетнамского народа; ~ дело *в* вьетнамского народа; ~ освобождения немецкого народа; ~ от имени народа; ~ повышения благосостояния народа; ~ благо *в* наших народов; ~ и свободы *р* народов; ~ суверенитета и независимости; ~ борьбе *д* за независимость; ~ партий △ организаций; ~ учреждений и организаций; ~ запрещение *в* ядерного оружия; ~ установление *и* дипломатических отношений; ~ △ единство *и* партии; ~ нашей Коммунистической партии; ~ славной Коммунистической партии; ~ съезду коммунистической партии; ~ Южно-Африканской коммунистической партии; ~ △ правящая партия; ~ △ Советского правительства; ~ заявлении Советского правительства; ~ имени *р* Советского правительства; ~ △ боннское правительство; ~ △ лейбористское правительство; ~ ни одно правительство; ~ Совета и председатель; ~ *ц* июня председатель; ~ и постоянного президиума; ~ стимулирования промышленного производства; ~ экономического стимулирования производства; ~ торговли и промышленности; ~ вклад *в* в развитие; ~ государственном плане развития; ~ некапиталистический путь *в* развития; ~ борцов *р* за свободу; ~ борьбу за свободу; ~ борьбы за свободу; ~ борющихся *в* за свободу; ~ встречали Генеральный секретарь; ~ КПСС *р* Первый секретарь; ~ сказал Генеральный секретарь; ~ и антиимпериалистических сил; ~ советских вооруженных сил; ~ в рамках совета; ~ заседаний Верховного Совета; ~ Президиума Верховного Совета; ~ Президиумом Верховного Совета; ~ премьер Государственного Совета; ~ помощи *п* между Советским Союзом; ~ Афганистана и Советского Союза; ~ в отношении Советского Союза; ~ дружбы и солидарности; ~ и взаимовыгодного сотрудничества; ~ и технического сотрудничества; ~ налаживания общеевропейского сотрудничества; ~ для построения социализма; ~ Ленинского Коммунистического Союза; ~ секретарь правления Союза; ~ △ представитель СССР; ~ рыбного хозяйства СССР; ~ Совету Министров СССР; ~ специального образования СССР; ~	14	126

i	Словосочетания	F	m

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
1578—1713	студенческий совет *и* СССР; ~ народами обеих стран; ~ всеми европейскими странами; ~ власти *р* в стране; ~ в экономике страны; ~ жизни *п* нашей страны; ~ оборонной мощи *р* страны; ~ производительных сил страны; ~ министерство *и* обороны США; ~ прекращения агрессии США; ~ директивах *ц* съезда; ~ работой *ц* съезда; ~ и производительности *р* труда; ~ повышение *и* производительности труда; ~ внутренних дел ФРГ; ~ правящие круги *и* ФРГ; ~ отраслях народного хозяйства; ~ выступил секретарь ЦК; ~ округу секретаря *в* ЦК; ~ октябрьского Пленума ЦК; ~ отчетного доклада ЦК; ~ Первому секретарю ЦК; ~ △ народ Южного Вьетнама; ~ △ патриоты Южного Вьетнама последствий израильской агрессии; ~ председателю Совета Безопасности; ~ поддержку в борьбе; ~ что в борьбе; ~ против преступной войны; ~ вывод *и* израильских войск; ~ положение *и* во Вьетнаме; ~ августе прошлого года; ~ границ *ц* года; ~ договора *р ц* года; ~ заявлении *ц* года; ~ июле этого года; ~ первого социалистического государства; ~ и демократического движения; ~ министерство *и* внутренних дел; ~ министром внутренних дел; ~ мы имеем дело; ~ на укрепление дружбы; ~ и общественной жизни ; ~ исключительно важное значение; ~ говорить от имени; ~ агрессию американского империализма; ~ борющихся *р* против империализма; ~ колониальной системы империализма; ~ на заседании комитета; ~ приветствие *в* Центрального Комитета; ~ имени *р* ЦК КПСС; ~ под руководством КПСС; ~ секретари *и* ЦК КПСС; ~ тезисы *и* ЦК КПСС; ~ △ пятнадцать лет; ~ через *ц* лет; ~ всех прогрессивных людей ; ~ △ что министр; ~ КПЧ *р* △ министр; ~ выступающих *р* за мир; ~ в сохранении мира; ~ всех сил мира; ~ всех странах мира; ~ за судьбы мира; ~ и всеобщего мира; ~ по укреплению мира; ~ сил всего мира; ~ △ китайский народ; ~ в интересах народа; ~ дела *р* вьетнамского народа; ~ безопасности *р* всех народов; ~ внутренние дела *в* народов; ~ и монгольского народов; ~ интересах наших народов; ~ борющимся вьетнамским народом; ~ оказывают вьетнамскому народу; ~ поддержку вьетнамскому народу; ~ правительству и народу; ~ ядерных сил НАТО; ~ укрепления национальной независимости; ~ свою национальную независимость; ~ △ членов *в* ООН; ~ Совете Безопасности ООН; ~ Совета Министров организации; ~ и комсомольских организаций; ~ предприятий и организаций; ~ распространение *и* ядерного оружия; ~ верного сына *р* партии; ~ комитета Трудовой партии; ~ Политбюро *р* ЦК партии; ~ Суданской коммунистической партии; ~ ЦК *р* нашей партии; ~ в жизнь политики; ~ области *п* международной политики; ~ оказывать всестороннюю помощь; ~ △ председатель правительства; ~ заявления *р* Советского правительства; ~ МНРП *р* и правительства; ~ что американское правительство; ~ ВСРП *р* △ председатель; ~ Германии *р* △ председатель; ~ △ заместителя *р* председателя; ~ приняли участие представители; ~ △ бывший президент; ~ △ как президент; ~ КПЧ *р* △ президент; ~ Советского Союза △ Президиума; ~ и сельскохозяйственного производства; ~ влияние *и* на развитие; ~ и культурного развития; ~ и экономического развития; ~ великой культурной революции; ~ монгольской народной революции; ~ за мир, свободу; ~ △ членов *р* совета; ~ и председателя *р* совета; ~ сотрудничество *и* с Советским Союзом; ~ в адрес Советского Союза; ~ звания *р* Героя Советского Союза; ~ △ неделя солидарности; ~ единства и солидарности; ~ Советского комитета солидарности; ~ и экономического сотрудничества; ~ в условиях социализма; ~ за победу социализма; ~ к строительству социализма; ~ делам Совета Союза; ~ депутаты *в* Совета Союза; ~ Государственных комитетов СССР; ~ гражданской авиации *р* СССР; ~ с нашей стороны; ~ газеты *и* социалистических стран; ~ многих других стран; ~ обеих наших стран; ~ между всеми странами; ~ между этими странами; ~ отношений со странами; ~ в каждой стране; ~ внутренние дела *в* страны; ~ о положении страны; ~ военно-воздушных сил США; ~ вооруженные силы *и* США; ~ вооруженными силами США; ~ империалистические круги *и* США; ~ министра *р* обороны США; ~ начальников *р* штабов США; ~ прежде всего США; ~ от имени трудящихся; ~ усилия *в* для укрепления; ~ для сельского хозяйства; ~ области *п* сельского хозяйства; ~ развитию народного хозяйства; ~ △ Президиум *и* ЦК; ~ △ члены *р* ЦК; ~ второй секретарь ЦК; ~ майского Пленума ЦК; ~ с членом ЦК; ~ участие *в* секретарь ЦК; ~ приняли участие член [ЦК]; ~ сателлитов *р* из Южного Вьетнама по поводу агрессии; ~ заседание *и* Совета Безопасности; ~ проблема европейской безопасности; ~ народов в борьбе; ~ партий в борьбе; ~ поддержку героической борьбы; ~ △ после войны; ~ в случае войны; ~ в ходе войны; ~ угрозы *р* ядерной войны; ~ отводе израильских войск; ~ численность *и* американских войск; ~ урегулированию во Вьетнаме; ~ ГДР *р* во главе; ~ делегацию во главе; ~ декабря прошлого года; ~ марте этого года; ~ ноября прошлого года; ~ соглашениям *ц* года; ~ через два года; ~ молодых национальных государств; ~ △ правительство; ~ △ председатель правительства; ~ заявления *р* Советского правительства; ~ МНРП *р* и правительства; ~ что американское правительство; ~ ВСРП *р* △ председатель; ~ Германии *р* △ председатель; ~ △ заместителя *р* председателя; ~ приняли участие представители; ~ △ бывший президент; ~ △ как президент; ~ КПЧ *р* △ президент; ~ Советского Союза △ Президиума; ~ и сельскохозяйственного производства; ~ влияние *и* на развитие; ~ и культурного развития; ~ и экономического развития; ~ великой культурной революции; ~ монгольской народной революции; ~ за мир, свободу; ~ △ членов *р* совета; ~ и председателя *р* совета; ~ сотрудничество *и* с Советским Союзом; ~ в адрес Советского Союза; ~ звания *р* Героя Советского Союза; ~ △ неделя солидарности; ~ единства и солидарности; ~ Советского комитета солидарности; ~ и экономического сотрудничества; ~ в условиях социализма; ~ за победу социализма; ~ к строительству социализма; ~ делам Совета Союза; ~ депутаты *в* Совета Союза; ~ Государственных комитетов СССР; ~ гражданской авиации *р* СССР; ~ с нашей стороны; ~ газеты *и* социалистических стран; ~ многих других стран; ~ обеих наших стран; ~ между всеми странами; ~ между этими странами; ~ отношений со странами; ~ в каждой стране; ~ внутренние дела *в* страны; ~ о положении страны; ~ военно-воздушных сил США; ~ вооруженные силы *и* США; ~ вооруженными силами США; ~ империалистические круги *и* США; ~ министра *р* обороны США; ~ начальников *р* штабов США; ~ прежде всего США; ~ от имени трудящихся; ~ усилия *в* для укрепления; ~ для сельского хозяйства; ~ области *п* сельского хозяйства; ~ развитию народного хозяйства; ~ △ Президиум *и* ЦК; ~ △ члены *р* ЦК; ~ второй секретарь ЦК; ~ майского Пленума ЦК; ~ с членом ЦК; ~ участие *в* секретарь ЦК; ~ приняли участие член [ЦК]; ~ сателлитов *р* из Южного Вьетнама	13	136
1714—1872	по поводу агрессии; ~ заседание *и* Совета Безопасности; ~ проблема европейской безопасности; ~ народов в борьбе; ~ партий в борьбе; ~ поддержку героической борьбы; ~ △ после войны; ~ в случае войны; ~ в ходе войны; ~ угрозы *р* ядерной войны; ~ отводе израильских войск; ~ численность *и* американских войск; ~ урегулированию во Вьетнаме; ~ ГДР *р* во главе; ~ делегацию во главе; ~ декабря прошлого года; ~ марте этого года; ~ ноября прошлого года; ~ соглашениям *ц* года; ~ через два года; ~ молодых национальных государств; ~ △	12	159

Таблица 1 (*продолжение*)

i	Словосочетания	F	m
	тельственная делегация; ~ отношения *и* братской дружбы; ~ сотрудничества и дружбы; ~ что отношения *и* дружбы; ~ американского образа жизни; ~ выступать от имени; ~ агрессивным силам империализма; ~ агрессия американского империализма; ~ реакции *р* и империализма; ~ вокруг Центрального Комитета; ~ заседании Исполнительного комитета; ~ Катовицкого воеводского комитета; ~ поручению Центрального Комитета; ~ Ленинградского обкома КПСС; ~ △ министр культуры; ~ науки, техники, культуры; ~ △ шесть лет; ~ более двадцати лет; ~ последние *в* семь лет; ~ почти *ц* лет; ~ протяжении ряда лет; ~ △ сказал министр; ~ РКП *р* △ министр; ~ безопасности *д* и мира; ~ дело *в* укреплении мира; ~ для всего мира; ~ за упрочение мира; ~ обеспечения прочного мира; ~ от внешнего мира; ~ по обеспечению мира; ~ мест в мире; ~ обстановка в мире; ~ что в мире; ~ великий советский народ; ~ △ всего народа; ~ братского советского народа; ~ достижениями советского народа; ~ имени *р* советского народа; ~ мужественного вьетнамского народа; ~ государствами и народами; ~ и свободу народов; ~ братским советским народом; ~ со всем народом; ~ и советскому народу; ~ предоставления вьетнамскому народу; ~ представитель министерства обороны; ~ развития советского общества; ~ строительства коммунистического общества; ~ строительства социалистического общества; ~ в члены ООН; ~ Генеральной Ассамблеи Организации; ~ на общеевропейской основе; ~ к нормализации отношений; ~ развития добрососедских отношений; ~ △ руководители партий; ~ делегации *р* коммунистической партии; ~ политика нашей партии; ~ родной Коммунистической партии; ~ с председателем партии; ~ своей коммунистической партии; ~ здравствует Коммунистическая партия; ~ Румынская коммунистическая партия; ~ социалистическая рабочая партия; ~ области *и* внутренней политики; ~ миролюбивую внешнюю политику; ~ △ глава правительства; ~ четырех пунктов правительства; ~ △ военное правительство; ~ △ французское правительство; ~ △ японское правительство; ~ партии *р* и правительство; ~ был заместитель председателя; ~ выступил заместитель председателя; ~ первого заместителя *р* председателя ; ~ присутствовали заместитель председателя; ~ △ что представители; ~ △ когда президент; ~ △ Председателя *р* Президиума; ~ заместителю Председателя Президиума; ~ международных связей Президиума; ~ пост *в* Председателя Президиума; ~ *собств* Председатель Президиума; ~ честь *в* Председателя Президиума; ~ повышения эффективности производства; ~ роста промышленного производства; ~ с опытом работы; ~ △ па развитие; ~ реконструкции *р* и развития; ~ правительства *р* в связи; ~ Политбюро *р* и секретарь; ~ приняли участие секретарь; ~ *собств* и секретарь; ~ изо всех сил; ~ делегацию Верховного Совета; ~ июня Председатель Совета; ~ пост *в* Председателя Совета; ~ присутствовали Председатель Совета; ~ △ когда Советский Союз; ~народов △ Советский Союз; ~ вместе с Советским Союзом; ~ дружбы с Советским Союзом; ~ △ позиция Советского Союза; ~ при содействии Советского Союза; ~ △ страны *и* социализма; ~ для строительства социализма; ~ со странами социализма; ~ заседания *и* Совета Союза; ~ комиссий Совета Союза; ~ △ правительство *и* СССР; ~ председатели комитетов СССР; ~ председателя *р* Госплана СССР; ~ △ усилия *и* стран; ~ комитета солидарности стран; ~ коммунистов *р* всех стран; ~ ряда других стран; ~ Советского Союза со странами; ~ сотрудничества между странами; ~ в европейских странах; ~ в разных странах; ~ развитых капиталистических странах; ~ различных районов страны; ~ экономического развития страны; ~ агрессивной войны США; ~ министерства *р* обороны США; ~ от правительства США; ~ поводу агрессии США; ~ преступной агрессии *р* США; ~ здравствует *ц* съезд; ~ и водного хозяйства; ~ подъема сельского хозяйства; ~ в заявлении ЦК; ~ заместитель Председателя ЦК; ~ последующих Пленумов ЦК; ~ постановление *в* Пленума ЦК; ~ Президиума Политбюро *р* ЦК; ~ члены *и* Президиума ЦК; ~ △ заявил член ЦК; ~ находящийся здесь член [правительства]		

Таблица 1 (*продолжение*)

i	Словосочетания	F	m

Таблица 2

Распределение частот для низкочастотных словосочетаний

i	m	F	i	m	F
1873—2090	218	11	4260—5435	1176	5
2091—2551	261	10	5436—7413	1978	4
2352—2654	303	9	7414—10995	3582	3
2655—3034	380	8	10996—20741	9746	2
3035—3486	452	7	20742—67154	46413	1
3487—4259	773	6			

Часть II. АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

А. В. Зубов

ПЕРЕРАБОТКА ТЕКСТА ЕСТЕСТВЕННОГО ЯЗЫКА В СИСТЕМЕ «ЧЕЛОВЕК — МАШИНА»

§ 1. ВВЕДЕНИЕ

Как известно, в наше время возникла большая диспропорция между объемом вновь поступающей информации и способностью человека к ее переработке. Поток информации растет из года в год, а эффективность деятельности человека практически во всех сферах определяется его способностью к переработке информации, а не его энергетическими возможностями.[1] Такие задачи, как поиск информации в различных хранилищах, составление рефератов, обучение, перевод с одного языка на другой и т. п., заставляют человека искать новые пути для их решения. Одним из возможных способов ускорения процессов решения подобных задач является передача этих задач, или отдельных частей задач, электронно-вычислительным машинам. Однако это требует знания тех операций, которые выполняет человек при реализации соответствующих заданий. Другими словами, необходимо знать алгоритмы умственной деятельности человека.

Под алгоритмом мы будем понимать точное предписание о выполнении в определенном порядке некоторой системы операций для решения всех задач некоторого данного типа.[2]

Мозг человека можно рассматривать как сложнодинамическую систему с обратной связью. Возможны два подхода к изучению таких систем: макроподход и микроподход.

Если бы нас интересовали лишь результаты деятельности мозга, то достаточно было бы сформулировать интересующую нас проблему в терминах «черного ящика». Такой подход, игнорирующий внутреннее строение системы, называется макроподходом.[3]

[1] См.: А. Н. Колмогоров. Кибернетику — на службу коммунизму. М.—Л., 1966, стр. 5.

[2] Б. А. Трахтенброт. Алгоритмы и машинное решение задач. М., 1960, стр. 7.

[3] К. Е. Морозов. Математические модели в кибернетике. М., 1968, стр. 36.

Тогда, получив на выходе то, что дает человеческое поведение, мы были бы удовлетворены.

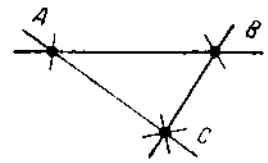
Однако для процесса познания и самопознания необходимо знать не только результат некоторой деятельности мозга, но и саму деятельность как процесс. При этом важно знать, из каких составных частей состоит система, как эти части связаны между собой, как они функционируют и т. д. Такой подход называется микроподходом.

Эти два способа моделирования переплетаются между собой, дополняют друг друга. Их взаимопроникновение проявляется по крайней мере в двух моментах:

1) изучая с помощью ЭВМ функции мозга, можно строить гипотезы о его структуре;

2) моделируя нейроны и создавая искусственные нейронные сети, мы изучаем функции этих нейронов и строим гипотезы о функционировании системы нейронов в целом.

Чтобы познать умственную деятельность человека как процесс, необходимо добиться не только тождества результатов на выходе у человека и автомата, но и тождества процессов у того и другого. И даже не тождества, а подобия. Если изобразить модель структуры мозга в виде схемы



где частички A, B, C,. . . представляют собой отдельные акты мышления, то задача сводится к тому, чтобы заменить тип материальности частичек, а связи между ними оставить те же самые. Происходит как бы проекция человеческого мышления на «мышление» автоматов, перевод одной и той же цепочки событий на два языка — язык «мышления» автоматов и язык мышления человека.

Таким образом, чтобы построить модель структуры мозга, необходимо воспроизвести все связи (например: AB, AC, BC,. . .) между отдельными актами мышления и затем, на основании информации об упорядочении этих связей в отдельных актах мышления, собрать структуру мозга. Следовательно, задача сводится к построению функциональных моделей актов мышления.

Из этой задачи мы выбираем следующие акты мышления при переработке человеком текстов естественного языка:

1) выделение наиболее частых единиц письменной речи (букв, буквосочетаний, словоформ, сочетаний словоформ);

2) определение значения иностранного слова в тексте при переводе;

3) определение значения связанной группы иностранных слов при переводе.

Наши задачи входят в общий цикл задач по моделированию деятельности мозга. В лингвистике сюда можно отнести исследования по установлению соответствий между алфавитами двух языков и выделению классов букв у Б. В. Сухотина.[4] В математике это алгоритмы эвристического поиска логического вывода (машинное доказательство теорем) у Н. А. Шанина, Ван Хао и Г. Гелернтнера[5] и программа решения задачи символического интегрирования.[6] Сюда можно отнести также известные работы А. Ньюэлла, Дж. Шоу и Г. Саймона по моделированию доказательства теорем символической логики (программа «логик — теоретик») и решению задач (программа General Problem Solver).[7] К тому же типу интеллектуальных задач относятся задачи по моделированию вербального обучения и формированию человеческих понятий.[8] Наконец, ряд моделей нервной системы, предложенных Е. Конорским, К. С. Лэшли, Р. Л. Берлем, У. Р. Эшби, Д. О. Хеббом, К. Г. Прибрамом и другими,[9] также вписываются в указанный цикл задач. Знание указанных выше процессов способствует уяснению природы формального и творческого мышления, внимания, памяти, способности к сравнению и суждению, природы обучения, перевода и т. п. Алгоритмы этих процессов находят и большое практическое применение. Как мы отмечали выше (стр. 286), знание их позволит ускорить процессы перевода, информационного накопления и поиска, дешифровки, реферирования, обучения и т. п. Практика использования этих алгоритмов рождает новые теоретические проблемы. Важнейшей из них

---

[4] Б. В. Сухотин. Экспериментальное выделение классов букв с помощью ЭВМ. В сб.: Проблемы структурной лингвистики, М., 1962, стр. 198—205.

[5] См., например: Ван Хао. На пути к механической математике. Кибернетический сб., вып. 5, 1962, стр. 114 и сл.; Г. Гелернтэер. Реализация машины, доказывающей геометрические теоремы. ВММ, стр. 145—165.

[6] Д. Слэйджл. Эвристическая программа, решающая задачи символического интегрирования в объеме первого курса университета. ВММ, стр. 204—219.

[7] А. Ньюэлл и Г. Саймон. GPS — программа, моделирующая процесс человеческого мышления. ВММ, стр. 113—145, 283—301.

[8] Э. Фейгенбаум. Моделирование вербального обучения. ВММ, стр. 301—317; Э. Хант и К. Ховленд. Машинная модель формирования человеческих понятий. ВММ, стр. 317—337.

[9] Ф. Джордж. Мозг как вычислительная машина. М., 1963, стр. 357—389.

является проблема взаимоотношения человека и автомата. Эта проблема имеет два аспекта:

1) симбиоз человека и машины;

2) распределение функций между человеком и автоматом.

Вопрос о симбиозе человека и машины предусматривает изучение свойств человека как звена системы, в которой человек в совокупности с техническими средствами участвует в выполнении работы. Второй вопрос требует решения задач гармонического сочетания человеческого и машинного факторов.[10]

Но, чтобы машина была надежным партнером человека в его умственной деятельности, необходимо, чтобы она представала перед ним в виде устройства, «очеловеченного» в максимальной степени, т. е. машины, которая способна к постоянному и непосредственному контакту с человеком.[11] Основной задачей при этом является задача создания первичного входного языка машины. Для машины всегда существует некоторый полный входной язык с тривиальной, но универсальной семантической функцией.[12] Это язык машинных программ в их истинном виде. Но в большинстве случаев задание машине возникает у человека на естественном языке, являющемся наиболее универсальным средством общения. Поэтому возникает необходимость расширения семантической функции ЭВМ с тем, чтобы она могла перерабатывать информацию, записанную на естественном языке. Сейчас эта проблема решается с помощью трансляторов — некоторых промежуточных языков, обладающих более широкой семантической функцией, чем язык программ, использующих крайне ограниченный запас слов естественного языка и элементарные правила их сочетаний. Существует около 1700 таких языков (например: АЛГОЛ, АЛГАМС, АЛГЕК, ИПЛ-V, КОБОЛ, ЛЯПАС, SLC-II, СИМУЛА СИМСКРИПТ, ФОРТРАН и др.). Но все эти языки имеют узкую проблемную ориентацию, они не приспособлены для обдумывания, постановки и нахождения возможных путей решения задачи, анализа неожиданных и незнакомых ситуаций, отысканий аналогий и т. п.[13] Нужен язык, достаточно полный относительно того класса элементов внешнего мира, который интересует человека. Помимо этого, он должен обладать достаточно развитой грамматической структурой, допускающей естественную равносмысленную перефразировку любой (в разумных пределах) грамматической конструкции, употребляемой человеком.

---

[10] А. Я. Лернер. Начала кибернетики. М., 1967, стр. 358.

[11] А. П. Ершов. Об одном виде контакта человека с машиной. «Семинар. Автоматизация мыслительных процессов». Киев, 1963, стр. 4.

[12] Основные определения, относящиеся к полноте входного языка и универсальной семантической функции, см. ниже, стр. 344—345.

[13] Д. Ю. Панов. О взаимодействии человека и машины. ВФ, 1967, № 1, стр. 48.

Но семантическую функцию ЭВМ нельзя расширить и за счет введения в машину всего естественного языка, поскольку это связано с двумя существенно важными ограничениями.

Первое состоит в сложности естественного языка. Второе — в ограниченности «памяти» ЭВМ.

Рассмотрим первое из этих ограничений. Известно, что любой достаточно развитый естественный язык содержит десятки тысяч лексических единиц и их сочетаний. Необозримо велико число их возможных комбинаций, образующих осмысленные предложения. Помимо этого, сам процесс порождения текста на основе данной языковой системы определяется многими частными факторами. Вряд ли целесообразно и возможно добиваться подробного и тщательного описания и моделирования всех этих частных факторов.

Переходя ко второму ограничению, приведем следующий пример. Элементарный подсчет [14] показывает, что даже для простейшего пословного автоматического перевода с русского языка на английский «память» машины должна содержать не менее 4.5×10^7 двоичных единиц информации. Эта величина соизмерима с суммарным объемом «памяти» ЭВМ средней мощности.[15] Для осуществления более или менее полного речевого общения в системе «человек—машина» понадобится автомат, обладающий «памятью» в десятки и сотни раз большей. Несмотря на быстрое развитие электронно-вычислительной техники вряд ли можно ожидать, что в ближайшие десятилетия для массовой переработки текста будут использованы ЭВМ с объемом «памяти» порядка 10^{10} дв. ед.

Указанные ограничения приводят к необходимости создания сокращенного, базового языка, который включает наиболее важные (частые и наиболее информационно-нагруженные) единицы языка. Отбор единиц в базовый язык осуществляется путем изучения соответствующего естественного языка с помощью приемов теории вероятностей и математической статистики. При этом исследованию подвергается не весь язык, а некоторый подъязык — совокупность языковых элементов и их отношений в текстах с ограниченной тематикой.

При вероятностно-статистическом описании конкретного подъязыка выделяются две задачи:

1) отбор наиболее важных единиц подъязыка и определение их информационных характеристик;

[14] Р. Г. Пиотровский. Базовые языки и отраслевой бинарный перевод. Сб.: «Язык и общество», Л., 1968.

[15] Ср.: Е. Т. Гавриленко. Автоматизация программирования и стандартные программы для вычислительных машин «Урал-2», «Урал-3», «Урал-4». М., 1966; А. П. Лук. Память и кибернетика. М., 1966; В. Ф. Ященко. Программирование для цифровых вычислительных машин М-220, М-20, БЭСМ-3М, БЭСМ-4. М., 1967; В. В. Шураков, В. В. Каплинский. Программирование учетно-статистических задач для ЭЦ ВМ «Минск-22». М., 1967.

2) построение правил, связывающих форму выделенных единиц с их содержанием.

Первая задача включает в себя выделение наиболее частых морфем, слов, словосочетаний, грамматических конструкций, а также определение структуры слова, распределение информационной нагрузки в нем, веса аналитической и флективной морфологий и ряд других вопросов.[16] Иными словами, на первом этапе строятся морфология, лексика и синтаксис базового языка.

В процессе решения второй задачи отыскиваются правила сочетаемости слов и классов слов, правила различения омонимии и многозначности лингвистических единиц и другие правила, позволяющие перейти от формы выделенных единиц к их содержанию. Таким образом, на втором этапе строится семантика подъязыка.

Из всего сказанного следует, что базовый язык строится как некоторое статистическое приближение к реальной лексико-грамматической системе подъязыка. Степень этого приближения зависит не только от того, насколько корректно были выбраны входящие в базовый язык единицы, но также от быстроты создания этого базового языка и от времени, в течение которого он без изменений используется в системе «человек—машина». Естественный язык, используемый человеком, является незамкнутой системой, способной к постоянному изменению и развитию. Любой язык, используемый машиной, представляет собой замкнутую систему. Хотя базовый язык и является вариантом естественного языка, но, будучи предназначенным для машины, он должен быть замкнутой системой. В результате такой несовмещенности базовый язык как бы отстает в своем развитии от естественного языка. И если создание базового языка затягивается на десятилетия, а затем еще несколько десятков лет он используется в общении человека и машины, то естественно, что разрыв между этим базовым языком и естественным языком становится все более заметным. Именно поэтому создание базового языка должно осуществляться в сжатые сроки с широким использованием самих же электронно-вычислительных машин. В соответствии со сказанным, цель нашей работы может быть сформулирована следующим образом: построить модели указанных выше (стр. 288) актов речевого поведения человека таким образом, чтобы они имитировали и результат и поведение человека при выполнении соответствующей умственной работы.

Методика нашей работы сводится к выполнению следующей последовательности операций: описываем общую схему речевого поведения человека, принимаем эту схему за общую схему работы

[16] См., например: Р. Г. Пиотровский. Информационное измерение языка. Л., 1968; Г. П. Богуславская. Энтропия английского печатного текста. АКД, Минск, 1966; А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживина. Энтропия русского языка. ВЯ, 1962, № 6.

машины, сообразуясь со спецификой отличия ЭВМ от человеческого мозга, строим алгоритм, затем программу. Реализация этих программ и анализ результатов позволяют судить о том, насколько верно эти построенные нами модели отражают действительные последовательности операций, совершаемых мозгом при выполнении той или иной умственной работы.

В соответствии с этим построено наше исследование. В первой части работы приводятся модели, имитирующие те исследовательские приемы, которыми фактически пользуется лингвист в процессе выделения наиболее частых единиц подъязыка. Особенностью приведенных здесь алгоритмов является то, что они моделируют те функции мозга, которые могут быть формализованы до конца. Действительно, «ручная» работа по дистрибуционно-статистическому и информационному описанию текстов представляет собой серию дискретных, эксплицитных, легко формализуемых операций. Поэтому наши машинные программы, построенные на основе этих алгоритмов, изоморфны «ручным» программам. Во второй и третьей частях работы дается ряд моделей, описывающих те формы умственной деятельности человека, при которых известную роль играет его интуиция, эвристические способности. В таких случаях моделируемые явления не могут быть формализованы полностью, ибо они представляют собой цепочки имплицитных операций. Однако с определенной степенью вероятности могут быть учтены и элементы эвристики. Для этого в «память» машины вводится ряд дополнительных сведений, отражающих те или иные особенности эвристического поведения.

Глава I. ОСНОВНЫЕ ЕДИНИЦЫ БАЗОВОГО ЯЗЫКА И ФУНКЦИОНАЛЬНЫЕ МОДЕЛИ ИХ ВЫДЕЛЕНИЯ

§ 2. Вероятностно-статистические методы в применении к языку

Как отмечалось выше, вопрос о выделении основных единиц базового языка сводится к вероятностно-статистическому исследованию соответствующих подъязыков.

Чтобы применить аппарат теории вероятностей и математической статистики к такому феномену, как язык, необходимо ввести ряд ограничений.

1. Язык — открытая система и система, находящаяся в постоянном изменении. Поэтому в качестве первого ограничивающего условия мы должны предположить, что на определенном этапе (время, в течение которого создается и используется базовый язык) изменений в языке не происходит.[17]

[17] H. P. Edmundson. A Statistican's View of Linguistics Models and Language-Data Processing. In: Automatic Translation of Languages. Venice, July, 1962, p. 151.

Таким образом, открытая система моделируется с помощью «закрытой» системы.

2. Всякий язык, обладающий письменностью, пользуется по крайней мере двумя субстанциями плана выражения — звуковой и графической. Наличие разных субстанций вносит некоторые различия и в форму выражения.[18] Не существует ни одного языка, в котором наблюдалось бы полное совпадение фонемного и буквенного алфавитов. Другое, существенное с информационной точки зрения, различие между письменной и устной формами заключается в том, что письменный текст прерывен, а устная речь — дискретно — непрерывна. Тем не менее, специфика письменной формы системы выражения языка как особым образом организованной знаковой системы является достаточным основанием для построения описательных грамматик, базирующихся только на зрительном восприятии речи.[19]

Ориентируясь в дальнейшем на машинную реализацию всех предлагаемых алгоритмов, мы в нашей работе будем рассматривать речь в ее письменной форме, ибо такая форма более приемлема для статистического исследования и для переработки на ЭВМ, которые по принципу своей работы являются дискретными цифровыми автоматами.[20]

3. Далее, любой естественный язык понимается как система, состоящая из упорядоченной совокупности различного рода элементов, каждый из которых имеет определенное и типовое множество возможных комбинаций с другими элементами системы. Все эти элементы взаимодействуют между собой в речи по очень сложным закономерностям, и практически невозможно учесть все множество факторов, участвующих в образовании того или иного сочетания языковых элементов. Однако далеко действующие вероятностные связи между единицами текста слабы и, кроме того, они быстро затухают (примерно на 3—4 шага текста от данного элемента).[21] Поэтому следующее допущение заключается в том, что на первом этапе исследования текст рассматривается как последовательность независимых друг от друга дискретных единиц. Основной единицей текста будем считать *словоупотребление*, под которым мы понимаем цепочку букв между двумя пробелами в тексте. Другими единицами текста являются буквы, их сочетания, а также сочетания нескольких словоупотреблений. Появление каждой такой единицы в тексте рассматри-

[18] Л. Ельмслев. Пролегомены к теории языка. Русск. пер. НЛ, вып. 1, 1960, стр. 305—318.

[19] См., например: Т. Н. Николаева. Письменная речь и специфика ее изучения. ВЯ, 1961, № 3; В. Мотш. К вопросу об отношении между устным и письменным языком. ВЯ, 1963, № 2; И. А. Бодуэн-де-Куртенэ. Об отношении русского письма к русскому языку. СПб., 1912, стр. 16.

[20] H. P. Edmundson, ук. соч., стр. 152.

[21] Г. П. Богуславская, ук. соч.

вается как последовательная реализация отдельных независимых испытаний.[22]

4. Никаким лингвистическим исследованием нельзя охватить все многообразие речи. Приходится для исследования брать некоторое конечное число текстов, характерных для того или иного подъязыка. В этом заключается следующее, четвертое ограничение. Все множество текстов, охватываемых данным подъязыком, назовем генеральной совокупностью. Исследователю в его наблюдениях не дана генеральная совокупность, и ему приходится изучать ее по некоторому конечному числу текстов, которое назовем выборочной совокупностью или просто выборкой. Число элементов в генеральной совокупности и в выборке назовем их объемами. Чтобы иметь право распространить на генеральную совокупность данные об изучении распределения интересующего нас признака в выборке, должны быть выполнены определенные условия:

а) выборка из генеральной совокупности должна быть произведена случайно, т. е. каждый ее член должен иметь одну и ту же вероятность попасть в выборку;

б) необходимо, чтобы выборка была произведена по возможности из однородной совокупности;

в) величина выборочной совокупности должна отвечать требованию достаточности, т. е. число входящих в выборку элементов не должно быть меньше некоторой определенной величины, определяемой теоретически или эмпирически.

Выполнение всех этих условий является гарантией того, что применяемые при анализе полученной статистической модели методы теории вероятностей и математической статистики дадут верные результаты.[23]

Рассмотрим, каким образом указанные требования выполняются при вероятностно-статистическом исследовании подъязыка. Начнем с выяснения вопроса о минимально необходимом объеме выборки.

Разработанные в математической статистике методы позволяют точно определить объем выборочной совокупности лишь в случае, если известен объем генеральной совокупности. Практически невозможно определить количество единиц, входящих в тот или иной подъязык. Поэтому для определения достаточности выборки используются различные косвенные приемы. Ряд авторов по-своему обосновывают взятые ими объемы текстов для построения частотных списков. Например, при построении словаря

[22] Т. А. Микерина. Некоторые статистические приемы лексико-морфологического описания функционального стиля. АКД, Изд. ЛГУ, 1967, стр. 5—6; P. L. Garvin. A Linguist's View of Language-Data Processing. NLC, 1963.

[23] А. И. Карасев. Основы математической статистики. М., 1962, стр. 161—163.

испанского языка Гарсия Ос руководствуется скоростью появления новых слов в равных отрезках текста и частотой вновь появляющихся слов. По его критериям минимальный объем выборки составляет 400 тыс. словоупотреблений. Г. Йоссельсон в своем словаре русского языка определяет выборочную совокупность, достаточную для лексико-статистического исследования, в 1 млн. словоупотреблений. Критерием для него является условие, чтобы в 90 случаях из 100 слова с частотой 0,00002 и 0,00001 не могли попасть в один и тот же разряд.

Помимо этих, чисто интуитивных приемов, находит применение ряд методов, использующих известные статистические законы распределения случайных величин. Наиболее содержательные из этих процедур приведены в работах В. М. Калинина и Р. М. Фрумкиной.[24]

Определив тем или иным способом объем выборочной совокупности, рассмотрим вопрос об однородности и случайности выборки. Каждый подъязык, будь то публицистика, электроника, виноделие и т. д., состоит в свою очередь из разделов, составляющих бóльшую или меньшую часть подъязыка.[25] Задача состоит в том, чтобы:

а) подобрать тексты таким образом, чтобы они отражали процентное содержание тематических разделов в подъязыке («дозировка текста»);

б) подобрать единичные тексты [26] так, чтобы они обеспечили однородность и случайность выборки.

Эти задачи решены в нескольких работах группы «Статистика речи». Наиболее полно этот вопрос представлен в работе О. А. Нехай.[27]

При обработке текстов на ЭВМ рассмотренные процедуры выполняются человеком в процессе подготовки текста для кодирования.

§ 3. Кодирование лингвистической информации для ввода в ЭВМ

Ввиду отсутствия читающих устройств информация, подлежащая обработке на ЭВМ, должна быть представлена в форме, доступной для прочтения машиной. При кодировании текста чаще

[24] В. М. Калинин. О статистике литературного текста. ВЯ, 1964, № 1; Р. М. Фрумкина. Статистические методы изучения лексики. М., 1964.

[25] К. Ф. Лукьяненков. К вопросу о методике составления частотного словаря английского подъязыка судовых механизмов с помощью ЭВЦ. В сб.: Материалы XVIII научно-теоретической конференции. Фонетика, фонология и статистика речи, Минск, 1966, стр. 30.

[26] Под «единичным» текстом мы понимаем текст, содержащий строго определенное небольшое число словоупотреблений (например, 1000), который воспринимается и перерабатывается машиной за один прием.

[27] О. А. Нехай. Статистика и автоматический анализ текста. КД (рукоп.), Минский ГПИИЯ, 1968, стр. 203—206.

всего используется двоичная система счисления. Это объясняется тем, что операции в этой системе чрезвычайно просты и это позволяет легко реализовать в технических устройствах известный принцип «всё» или «ничего» («да» или «нет»), основывающийся на этой системе счисления. Именно поэтому большинство электронно-вычислительных машин работают на таком принципе.[28]

Существует предположение, что в мозгу человека часть лингвистической информации закодирована по непрерывному принципу, а другая часть — по дискретному.[29] В связи с этим уместно возвратиться к вопросу о роли бинарного представления лингвистических отношений.

Следует различать дихотомию представления лингвистической структуры, дихотомию (бинарность) работы мозга, ЭВМ и, наконец, бинарность объективного строения системы.

Дихотомическое изображение различных лингвистических структур вызывается необходимостью возможно проще представить то или иное языковое явление. Например, такие операции, как анализ по непосредственно составляющим и порождение предложения в модели В. Ингве, осуществляются по дихотомическому принципу.[30]

С другой стороны, мозг человека, ЭВМ, лингвистическая система могут использовать не только бинарные, но и тернарные, тетрарные, пентадные и т. д. отношения. Примеры тому: типы фонологических систем у Н. Трубецкого или принцип работы ЭВМ «Сетунь», использующий не двоичную, а троичную систему счисления.[31] Часть лингвистов считает, что все эти отношения могут быть приведены к наиболее простым бинарным отношениям, другие же обращаются к помощи более сложных принципов моделирования.[32] Так, А. А. Реформатский, выступая против всеобщности принципа бинарности, говорит, что предметность может быть дана не только бинарно, но и в виде «цепочек» или пучков признаков. Ср. классические классификации индоевропейских согласных по месту их образования: p — t — k, b — d — g, (ph — tk — kh, bh — dh — gh).

Часто на первый взгляд дихотомические отношения при более внимательном рассмотрении оказываются не чисто дихотомическими. В пользу последней точки зрения говорит тот факт, что при решении одной и той же задачи на машинах, использующих различные системы счисления, получаются идентичные результаты. Тем не менее принцип бинарности находит в практике достаточно широкое применение. Рассмотрим одно из применений этого принципа — при осуществлении перевода текста, написанного на естественном языке, в текст, который может быть прочитан ЭВМ. Каждая буква, цифра и знак в двоичной системе счисления изображаются в виде набора двоичных цифр 0 и 1. Эти комбинации двоичных единиц с помощью специальных аппаратов переносятся на перфоленту или перфокарты, которые представляют собой особую субстанцию плана выражения (наряду со звуковой и графической). На рис. 1 приведен пример записи слова на одном из типов перфолент. Комбинации пробивок (перфораций), составляющие текст указанного примера, взяты из международного телеграфного кода М-2.[33]

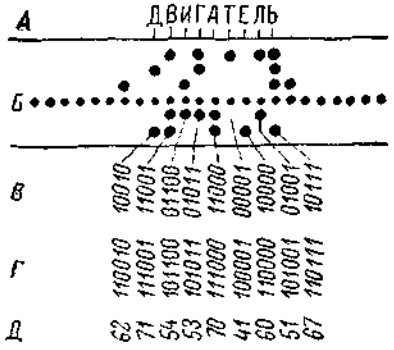


Рис. 1. Слово и его кодовое изображение.

А — слово естественного языка;
Б — то же слово в коде М-2 на перфоленте;
В — двоичная расшифровка кода М-2 на ленте;
Г — двоично-восьмеричное представление в «памяти» машины;
Д — условное восьмеричное представление слова.

§ 4. Автомат и человек — принципы переработки информации

Прежде чем переходить непосредственно к построению моделей, имитирующих речевое поведение человека, рассмотрим те принципы, которые лежат в основе переработки информации машиной и человеком.

Общая схема электронной вычислительной машины дана на рис. 2 (стр. 299). Проследим, как действует каждый блок этой схемы и сравним его с работой мозга при выполнении аналогичных операций.

---

[28] Г. С. Царалуа. *Об унифицированной длине машинного слова.* Тр. Тбилисского научно-исследовательского института приборостроения и средств автоматики, 1966, № 7.

[29] T. A. Sebeok. The Informational Model of Language and Digital Coding in Animal and Human Communication. NLC, 1963, p. 56.

[30] Г. Глиссон. Введение в дескриптивную лингвистику. Русск. пер. М., 1959; V. H. Yngve. A Model and a Hypothesis for Language Structure. Research Laboratory of Electronics and Department of Modern Languages Massachusetts Institute of Technology, Cambridge (Mass.), 1960.

[31] Н. С. Трубецкой. Основы фонологии. Русск. пер. М., 1960; Е. А. Жоголев. Система команд и интерпретирующая система для машины «Сетунь». Журнал вычислительной математики и математической физики, т. I, 1961, № 3.

[32] А. А. Реформатский. Дихотомическая классификация дифференциальных признаков и фонематическая модель языка. Сб. «Вопросы теории языка в современной зарубежной лингвистике», М., 1961, стр. 111; A. Martinet. Substance Phonique et Traits Distinctifs. Bulletin de la Société de Linguistique de Paris, vol. 53, p. 72.

[33] Полный список этого кода, с учетом некоторых диакритических знаков, приводится в табл. 1.

Таблица 1

Модифицированный код М-2

№№ комбинаций	Кодовые комбинации пробивок: 1-я позиция	2-я позиция	синхродорожка	3-я позиция	4-я позиция	5-я позиция	Буквы	Цифры	Диакритические знаки: знак	название
1	•	•	·				А (A)	—		
2	•		·		•	•	Б (B)	?		
3	•	•	·			•	В (W)	2	`	accent grave
4		•	·		•	•	Г (G)	Ш		
5	•		·		•		Д (D)	Х		
6	•		·				Е (E)	З	¯	долгота гласных
7		•	·	•	•	•	Ж (V)	=		
8	•		·			•	З (Z)	+		
9		•	·	•			И (I)	8	~	тильда
10	•	•	·		•		Й (J)	Ю		
11	•	•	·	•	•		К (K)	(		
12		•	·			•	Л (L)	)		
13			·	•	•	•	М (M)	.		
14			·	•	•		Н (N)	,		
15			·		•	•	О (O)	9	¨	tréma
16		•	·	•		•	П (P)	0		
17		•	·		•		Р (R)	4	ˇ	шипящие
18	•		·	•			С (S)	;		
19			·			•	Т (T)	5	^	accent circonflexe
20	•	•	·	•			У (U)	7	ç	cédille
21	•		·	•	•		Ф (F)	Э		
22			·	•		•	Х (H)	Щ		
23		•	·	•	•		Ц (C)	:		
24	•		·	•		•	Ы (Y)	6	'	мягкость согласных
25	•		·	•	•	•	Ь (X)	/		
26	•	•	·	•		•	Я (Q)	1	´	accent-aigu
27			·	•				пробел латынь русский		
28	•	•	·	•	•	•				
29			·					цифры		
30	•	•	·		•	•				

Устройства подготовки исходных данных. Как мы отмечали выше (стр. 295), отсутствие автоматов, способных читать текст прямо с листа, и тем более надежных устройств, «понимающих» звуковую форму речи, заставляет нас каким-то образом кодировать текст. Один из способов кодирования изложен выше в § 3. Такое кодирование осуществляется на различного рода устройствах подготовки исходных данных. Эти устройства позволяют кодировать со скоростью до 400 знаков в минуту. Практически скорость кодирования текста зависит от быстроты реакции человека, работающего на кодирующем аппарате, и в лучшем случае может составить 4000—5000 словоупотреблений за 7 часов работы. В понятие «быстрота реакции» входят, в частности, понятия «быстрота восприятия» и «быстрота воспроизведения». Быстрота восприятия зависит от того, на родном или не на родном языке кодируется текст, много ли в тексте знакомых слов. Эта реакция зависит от шрифта текста и т. п., а также от ряда чисто индивидуальных психологических качеств кодировщика.

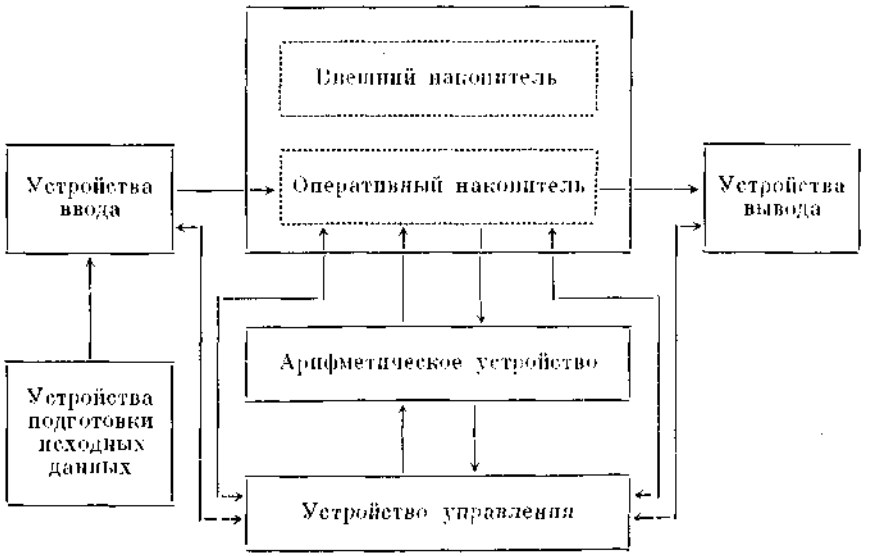


Рис. 2. Общая блок-схема ЭВМ.

Быстрота воспроизведения зависит, в частности, от того, насколько равновероятны воспринятые символы, а также от той номинальной информации (оцениваемой интуитивно), которая приходится на символ: чем большее число букв в алфавите, из кото-

рого отбирались символы, тем больше информации может воспроизвести человек.[34]

Ввод закодированной информации с перфолент и перфокарт в машину осуществляется через устройства ввода машины, которые преобразуют комбинации пробивок на носителях в комбинации токов и напряжений внутри машины. Вводимая информация располагается в строго определенных элементах (ячейках) «памяти» машины. При построении устройств ввода обычно не учитывается избыточность источников сообщений. Число двоичных единиц, вводимых в ЭВМ, во много раз, иногда на несколько порядков, превосходит количество информации, действительно содержащейся во введенном сообщении.[35] В связи с этим используемый в настоящее время побуквенный ввод не оправдывает себя. Более экономным было бы кодирование и ввод двухбуквенными и трехбуквенными сочетаниями.

Скорость ввода уже закодированной информации в машину колеблется от 100 до 1000 знаков в секунду (600—60000 дв. ед./сек.).

Опознавание при вводе информации в мозг человека происходит как переход от однобуквенного выбора к выбору из возможных слогов (чтение по складам), а затем к выбору из целых слов и даже наиболее вероятных сочетаний слов. По существу это не что иное, как этапы все более глубокого статистического кодирования текстовых сообщений.[36] Другое важное различие между вводом информации в машину и в мозг человека заключается в том, что запись информации в мозг осуществляется не в определенные пронумерованные ячейки, а производится по ассоциативному принципу — по связи вновь поступающих объектов с объектами, уже хранящимися в памяти.[37] Скорость восприятия информации человеком колеблется от 18 до 45 дв. ед./сек.[38] Последняя цифра характеризует скорость восприятия при чтении. Она зависит как от объема оперативной памяти человека, так и от скорости распространения импульса по нейрону.

Запоминающие устройства вычислительных машин, как правило, делятся на две части: оперативный (быстродействующий) накопитель и внешний накопитель.[39]

[34] П. Б. Невельский. Сравнительное исследование объема кратковременной и долговременной памяти. В сб.: XVIII международный психологический конгресс. Симпозиум 21, М., 1966, стр. 28.

[35] И. И. Цуккерман. О вводе информации в мозг и вычислительную машину. В сб.: Информация и кибернетика, М., 1967, стр. 199.

[36] Там же стр. 207.

[37] Л. П. Крайзмер. Хранение информации в кибернетических системах. В сб.: Информация и кибернетика, М., 1967, стр. 182.

[38] К. Штейнбух. Автомат и человек. Русск. пер. М., 1967, стр. 194, 300.

[39] Мы не используем здесь довольно часто употребляемые в литературе термины «оперативная память» и «внешняя память», желая тем самым подчеркнуть принципиальное различие между памятью человека и «памятью» машин. Термины эти мы будем употреблять применительно к видам памяти человека.

Основными характеристиками подобных устройств является их емкость и быстродействие.

Емкость запоминающего устройства измеряют обычно в словах (числах) заданной разрядности или в двоичных единицах.

Быстродействие накопителя определяется временем обращения — $T_{обр}$. Время обращения — это промежуток времени от момента посылки адреса, по которому должна быть записана или прочитана информация, до момента получения на выходе считанной информации.[40] Оперативные накопители электронно-вычислительных машин используют в своей работе самые различные принципы. Наибольшее распространение нашли магнитные оперативные запоминающие устройства (МОЗУ), принцип действия которых основан на явлении остаточного магнетизма. Емкость подобных накопителей невелика — от 4 до 64 тыс. ячеек ($10^5 \div 4 \cdot 10^6$ дв. ед.). Процесс накопления информации в МОЗУ можно рассматривать как транспортировку и замещение сообщения. Текст вводится в определенные фиксированные ячейки памяти. При записи любого нового числа стирается прежнее число, хранившееся в ячейке. Таким образом, оперативный накопитель, работающий на жестких непересекающихся связях, есть не более как элементарное хранилище некоторой суммы данных.[41] Время обращения к МОЗУ колеблется от тысячных долей секунды до десятых долей секунды.

Однако, учитывая, что стоимость хранения информации в МОЗУ достаточно велика, используют различного типа внешние накопители — магнитные ленты (МЛ), магнитные барабаны (МБ) и магнитные диски (МД).

Объем подобных накопителей достигает нескольких сот миллионов дв. ед. Время обращения в этом случае гораздо больше, чем в случае с оперативным накопителем. Оно может колебаться от десятых долей секунды до нескольких секунд.

В памяти человека выделяют оперативную память и постоянную память. Под объемом памяти человека понимается число единиц, которое человек может воспроизвести при одном повторении или в среднем на одно повторение. При кратковременном восприятии запоминаемого материала и немедленном и полном его воспроизведении число воспроизводимых единиц составляет

[40] Л. П. Крайзмер, ук. соч., стр. 130.

[41] В последние годы начали разрабатываться оперативные накопители, работающие на иных принципах. Сюда относятся, например, ассоциативные запоминающие устройства. Основное отличие таких устройств состоит в том, что извлечение информации из них производится не по заданному адресу, а по некоторым признакам самой информации. См., например, указанную работу Л. П. Крайзмер (стр. 169—171).

объем оперативной памяти. Объем оперативной памяти отражает способность моментального (непосредственного) сохранения. Объем постоянной памяти отражает способность не только сохранения, но и постепенного накопления информации и показывает, что человек может запомнить сверх того, что он уже запомнил.[42]

Природа запоминания в мозгу человека — это возникновение ассоциативных связей между новыми и старыми объектами информации. Не случайно человек лучше запоминает информацию, если в его мозгу уже есть определенное количество сходной информации. Например, человек, знающий несколько иностранных языков, легче усваивает новый язык. Разные исследователи по-разному оценивают объем оперативной памяти человека. Так, Г. Франк считает, что емкость ее равна 160 бит, другие же оценивают ее в 7 ± 2 слова ($180 \div 340$ бит).[43]

Природа процессов, происходящих в постоянной памяти человека, еще далеко не ясна. Существует несколько теорий, объясняющих механизм долговременного запоминания,[44] но в нашу задачу не входит их рассмотрение. Следует лишь отметить, что объем постоянной памяти не зависит от числа запоминаемых символов самого по себе, а определяется количеством информации, которое содержится в одном символе или во всей последовательности символов.[45] Величина этого объема оценивается по-разному, с различных точек зрения. Средние цифры, очевидно, определяются величиной порядка $10^{15} \div 10^{16}$ бит.[46] Среди других различий накопителей ЭВМ и памяти человека можно отметить следующие:

1) многократное считывание информации из накопителей ослабляет следы этой информации в них; многократное воспроизведение «запомненной» человеком информации закрепляет ее в памяти человека;

2) в исправно работающих накопителях величины входной и выходной информаций равны; общее количество воспроизводимой человеком информации всегда меньше количества принятой;

3) время хранения информации в накопителях машин практически неограниченно; при отсутствии тренировки информация, хранящаяся в памяти человека, постепенно забывается; период полураспада памяти человека, т. е. время, в течение которого человек

[42] П. Б. Невельский, ук. соч., стр. 26.
[43] См.: П. Б. Невельский. Объем памяти и количество информации. В сб.: Проблемы инженерной психологии, вып. 3, Л., 1965; Дж. А. Миллер. Магическое число семь, плюс или минус два. Русск. пер. в сб.: Инженерная психология, М., 1964.
[44] См., например: Д. Вулдридж. Механизмы мозга. Русск. пер. М., 1965, стр. 257—266; Л. П. Крайзмер, ук. соч., стр. 187—188.
[45] П. Б. Невельский. О скорости запоминания. ВП, 1967, № 1.
[46] А. П. Лук. Память и кибернетика. М., 1966, стр. 36; К. Штейнбух, ук. соч., стр. 81.

забывает половину полученной им информации, примерно равен 12 часам;[47]

4) для питания блоков накопителя машины емкостью порядка 10^7 дв. ед. расходуется несколько десятков или сотен ватт, в то же время мозг на все выполняемые им функции тратит лишь несколько ватт или десятков ватт;

5) лучшие накопители имеют величину удельной емкости порядка 10^3 бит/см^3, в то время как удельная емкость мозга[48] оценивается величиной 10^{12} бит/см^3.

Арифметическое устройство (АУ) машины — это основной узел, где осуществляется переработка информации.

Процесс переработки информации в АУ основан на применении правил математической логики[49] к элементам информации, представленным в виде комбинации двоичных цифр 0 и 1. Время выполнения этих операций определяет быстродействие машины. Наибольшее время быстродействия современных машин — $5 \cdot 10^6$ операций в секунду.[50]

С другой стороны, нервная система человека состоит из отдельных клеток различного типа. Нервные клетки объединяются под общим названием — нейрон. Одна часть нейрона, аксон, обладает свойствами обычного двухпозиционного переключателя, т. е. может работать по принципу «всё» (1) или «ничего» (0). Другая часть нейрона подчинена более сложным законам, регулируемым химическими или электрическими изменениями в окружающей ткани. Поэтому и принцип работы мозга по переработке информации несколько иной — он не сводится исключительно к простой комбинации сигналов «всё» или «ничего» («да» или «нет»). Большую роль при этом играют эвристические способности человека. Известно, что человек может принимать разумные решения даже тогда, когда не имеет всей необходимой информации. Очевидно, эти способности связаны с ассоциативным характером выполняемых мозгом операций, а также с тем фактом, что в мозгу человека заложено некоторое описание внешнего мира. Скорость переработки информации человеком определяется скоростью распространения импульса по аксону и характеризуется величиной порядка 100 м/сек. В процессе работы мозга происходит многократный переход от цифрового (дискретного) процесса к аналоговому (непрерывному), в то время как в машине действуют исключительно дискретные процессы.[51]

[47] Э. А. Якубайтис. Кибернетика и автоматизация труда исследователя. Вестник АН Латвийской ССР, Рига, 1967, № 2, стр. 13.
[48] Л. П. Крайзмер, ук. соч., стр. 186.
[49] Р. С. Гутер, Б. В. Овчинский, П. Т. Резниковский. Программирование и вычислительная математика. М., 1965, стр. 184—186.
[50] Дж. Амдаль, Дж. Блоу, Ф. Брукс. Архитектура системы IBM-360. Русск. пер. Кибернетический сб., НС, вып. 1, М., 1965, стр. 101.
[51] Д. Вулдридж, ук. соч., стр. 330; А. П. Лук, ук. соч., стр. 231.

Устройство управления (УУ) электронной машины служит для автоматического управления всеми устройствами машины в процессе вычисления. Оно обеспечивает режим работы машины в соответствии с программой решения задачи. При этом машине для работы необходима полная информация о всех деталях процесса, которым она управляет. Именно поэтому все операции машины эксплицитны. УУ является средством связи ЭВМ и внешнего мира. Через него осуществляется обратная связь и корректировка работы машины.

Как мы отмечали выше, мозг человека богат не одной лишь логикой. Операции, выполняемые им, по своему характеру имплицитны (правда, природа «имплицитного» аспекта работы мозга не вполне ясна). Как и у ЭВМ, у человека тоже существует обратная связь. Она осуществляется через органы чувств и самооценку своих решений.

Устройства вывода предназначены для вывода из машины результатов обработки в обозримом виде. Вывод может быть буквенным, цифровым или же закодированным на перфоленту и перфокарты. Лучшие устройства вывода позволяют выводить 10^2—10^6 бит/сек.

Воспроизведение информации человеком из его памяти может производиться без запроса (непроизвольное возникновение в сознании переживаний без внешнего возбуждения).[52] Скорость воспроизведения переменна. Она зависит от жизненной важности информации. Но, в отличие от ЭВМ, соотношение скорости переработки и скорости вывода обратно. При скорости переработки 12—45 дв. ед./сек. скорость вывода информации из мозга зависит от скорости устной речи, от темпа скорописи (стенографической или орфографической) оператора и т. п. Так, быстрота устной речи на родном языке составляет 60—200 слов в минуту (расчет осуществлялся относительно русского языка), а опытная машинистка может работать со скоростью 400—600 букв в минуту.[53]

§ 5. Принципиальные отличия в функционировании мозга и ЭВМ

Выбор машины

Обобщая сказанное в предыдущем параграфе, можно выделить ряд особенностей, характерных как для мозга, так и для ЭВМ.

I. Преимущества машины перед мозгом:

1) большая скорость выполнения операций;
2) точность исполнения конечного числа операций;
3) отсутствие потерь информации;

[52] Л. П. Крайзмер, ук. соч., стр. 184.

[53] А. В. Ярмоленко. Структуры и фазы многоязычия. В сб.: Проблемы общей и инженерной психологии, изд. ЛГУ, 1964, стр. 162—163.

4) возможность длительного хранения информации;
5) возможность быстрого извлечения необходимой информации.

II. Преимущества мозга перед ЭВМ:

1) наличие эвристических способностей;
2) большой объем памяти;
3) высокая надежность работы;
4) малый расход энергии для питания;
5) большая величина удельной емкости.

Эти различия между машиной и человеком и определяют общий характер лингвистических задач, которые должны решаться с помощью ЭВМ. Они также помогут нам ответить на ряд принципиальных вопросов. Целесообразно ли вообще применять ЭВМ в обработке лингвистической информации? Если да, то какие именно виды обработки можно поручить машине? Какие необходимы для этого машины?

Большая скорость выполнения элементарных операций, высокая степень точности, отсутствие потерь информации, долгий срок ее хранения — эти факторы являются решающими при ответе на первый вопрос. Да, машины вполне целесообразно применять для обработки лингвистической информации, которая характеризуется:

а) большим исходным объемом;
б) требованием точности анализа;
в) необходимостью возможно быстрого решения задачи;
г) требованием долгого хранения как исходных, так и промежуточных и конечных результатов.

Перечисленные пункты — это требования со стороны лингвистической информации к машине. Какие же требования предъявляет машина к лингвистической информации?

Первое и основное требование — полная формализация лингвистических задач.

Учитывая, что принцип работы ЭВМ построен на реализации условий «всё» или «ничего» («да» или «нет»), это требование сводится к тому, чтобы лингвистическая задача была разложена на конечное число элементарных операций. Вместе с тем формализация решения лингвистических задач — дело достаточно сложное и трудоемкое. Оно требует от исследователя большой лингвистической культуры, глубоко интуитивного проникновения в существо изучаемого лингвистического явления. Здесь, наконец, необходима определенная квалификация в области программирования, позволяющая сформулировать ход решения задачи в виде последовательности правил.[54]

Второе требование сводится к тому, чтобы все утверждения и формулировки, используемые в изложении лингвистической задачи, были недвусмысленными и полными.

[54] В. А. Москович. Автоматизация некоторых аспектов лингвистической работы. ВЯ, 1966, № 1, стр. 111.

И, наконец, вопрос о выборе машины. Некоторые исследователи по этому поводу заявляют, что «алгоритмы . . . разумнее разрабатывать безо всякой оглядки на технические возможности машин»,[55] т. е. нецелесообразно ориентироваться на конкретную машину, учитывать объем ее накопителей и другие характеристики. Здесь сразу же следует сказать, что такой подход с прагматической точки зрения совершенно неверен. Когда мы читаем, слушаем, запоминаем, переводим, решаем задачи — мы всегда подсознательно учитываем объем и способности нашей памяти. Например, если нам предложат расположить по алфавиту пять односложных слов, мы это сделаем «в уме» и назовем упорядоченную последовательность их. Если ту же работу необходимо проделать с двадцатью словами, то без карандаша и бумаги (внешний накопитель!) нам уже не обойтись. Таким образом, выбор машины для решения той или иной лингвистической задачи зависит в первую очередь от исходного объема информации и во вторую очередь — от сложности задачи.

Выше (стр. 290) мы отмечали, что в процессе построения базового языка выделяются две задачи.

I. Отбор наиболее важных единиц подъязыка и определение их информационных характеристик.

II. Построение правил, связывающих форму выделенных единиц с их содержанием.

В первой задаче в свою очередь можно выделить две подзадачи:

1) определение основных информационных свойств подъязыка;

2) отбор наиболее важных единиц этого же подъязыка.

Первая подзадача имеет вычислительный характер. Она предусматривает выявление некоторых общих лингвистических и информационных закономерностей текста (например, определение параметров так называемого закона Эсту—Ципфа—Мандельброта, определение характера распределения частот слов и т. д.). Здесь ЭВМ обрабатывает только цифровую информацию сравнительно небольшого объема. Поэтому реализация этой задачи особых трудностей не представляет — она может быть решена практически на любой ЭВМ.

Вторая подзадача состоит в выборе и упорядочении по определенным признакам различных единиц текста и их комбинаций. Решение этой задачи связано с машинным анализом больших массивов информации и требует применения машин с большой емкостью накопителей.

При построении правил, связывающих форму выделенных единиц с их содержанием (задача II), приходится также иметь дело с большими массивами текстовой информации. Но еще больший объем занимают в «памяти» машин соответствующие правила.

[55] И. А. Мельчук, Р. Д. Равич. Автоматический перевод. 1949—1963 гг. М., 1967, стр. 8.

Здесь уже важен не только объем накопителей, но и быстродействие машины.

Опыт работы группы «Статистика речи» показывает, что подобные работы могут быть выполнены с достаточно высоким коэффициентом рентабельности лишь на машинах, отвечающих следующим требованиям:[56]

а) машина должна обладать возможностью ввода и вывода алфавитной информации;

б) ввод текста должен осуществляться со скоростью не менее 800 символов в секунду;

в) оперативный накопитель должен иметь объем не менее 8000 ячеек по 6 символов в каждом слове;

г) внешний накопитель должен иметь не менее 600 000 ячеек;

д) машина должна обладать быстродействием не менее 1000 операций/сек.;

е) вывод должен осуществляться со скоростью не менее 100 символов/сек.

Этим условиям удовлетворяет ряд отечественных машин. Среди них: БЭСМ-4, БЭСМ-6, «Стрела», «Урал-4», «Урал-6», М-220, «Минск-22», «Минск-23», «Минск-32», «Раздан-3».

Среди зарубежных ЭВМ можно указать французскую «Гамма-60», американские IBM-360, IBM-7090, КАДАК, РАМАК, РЕЙКОМ.

§ 6. Выбор наиболее частых букв и буквосочетаний текста

6.1. Принцип решения задач в системе «человек—машина»

В общем виде процесс решения любой задачи в системе «человек—машина» можно представить состоящим из следующих этапов:[57]

а) постановка задачи специалистом;

б) математическая постановка задачи;

в) подготовка начальных данных для ввода в ЭВМ;

г) разработка метода решения;

д) разработка принципиального алгоритма решения задачи и представление его в виде блок-схемы;

е) программирование алгоритма;

[56] Р. Г. Пиотровский, А. В. Зубов, К. Ф. Лукьяненков, Э. Н. Хотяшов, А. И. Чапля. Статистическое исследование лексики и грамматики текста с помощью электронно-вычислительной машины. В сб.: Научная конференция. Проблемы синхронного изучения грамматического строя языка, М., 1965, стр. 145.

[57] В. А. Устинов. Применение ЭВМ в исторической науке. М., 1964, стр. 96.

ж) решение задачи на машине;

з) анализ и интерпретация полученных результатов.

Специалист, ставящий задачу (в нашем случае лингвист), должен в наиболее формализованном виде сформулировать задачу (см. стр. 305).

Математическая постановка задачи сводится к тому, что ищется способ представления исходных данных в числовом виде (ибо, как это видно из § 3, каждое слово текста представляет собой в машине определенное число). В поставленной задаче выделяются составные части, решение которых сводится к решению отдельных математических задач. Определяется последовательность операций, позволяющих по результатам решения частных задач получить окончательное решение.

Разрабатывая метод решения конкретной задачи, необходимо учитывать принципы решения уже известных задач, имеющих общее с поставленной задачей. Если задача допускает несколько решений, то нужно выбрать наиболее оптимальный вариант.

Под алгоритмом здесь мы понимаем точное предписание о выполнении в определенном порядке некоторой системы операций для решения всех задач некоторого данного типа.[58]

Система операций каждой конкретной машины называется системой команд данной машины. В систему команд входят операции арифметические, логические, операции управления и некоторые специальные операции. Отдельная команда несет в себе информацию об операции, подлежащей выполнению, и о величинах, над которыми должна быть выполнена эта операция, а также сведения о том, куда должен быть помещен результат выполнения операции. Обычно в операции участвуют две величины. В команде указываются не сами эти величины, а номера (адреса) тех ячеек, в которых величины расположены. Как правило, операции обозначаются двузначными восьмеричными числами от 00 до 77 со знаками «+» или «—». Адреса ячеек нумеруются четырехзначными восьмеричными числами, начиная с номера 0000.

Помимо алгорима как такового, мы будем выделять принципиальный алгоритм и подалгоритм.

Назовем принципиальным алгоритмом записанную в виде схемы последовательность алгоритмов, приводящую к решению определенной сложной задачи. Подалгоритм — это алгоритм решения отдельной простой задачи, входящей в состав сложной.

Алгоритмы и подалгоритмы можно записывать различными способами: специальными символами, операторами, блоками и т. п. В нашей работе мы используем запись алгоритмов в виде блоков. Блок — это часть алгоритма или подалгоритма, которая может быть представлена одной или несколькими (не более 10) командами машины.

[58] Б. А. Трахтенброт. Алгоритмы и машинное решение задач, стр. 7.

6.2. Принципиальный алгоритм выделения наиболее частых единиц письменной речи

Выше (стр. 290) мы отмечали, что исходной задачей при построении базового языка является задача выделения наиболее частых единиц подъязыка. Эта задача решается путем построения частотных списков соответствующих единиц.

Частотным списком тех или иных единиц текста называется упорядоченный список всех единиц этого текста, где каждой единице ставится в соответствие определенная характеристика — частота употребления в тексте. Под упорядочением единиц списка понимается такое их расположение, при котором на первом месте стоит единица с наибольшей частотой, за ней — с меньшей и т. д. Для единиц, обладающих одинаковой частотой, производится упорядочение по алфавиту, в котором записаны единицы.

Допустим, что мы имеем текст, содержащий N единиц и удовлетворяющий всем требованиям, изложенным в § 2. Каким же образом выделить из этого текста наиболее частые единицы? Рассмотрим прежде всего, как человек решает задачу построения частотного списка. Вполне естественно, что на первом этапе он «делит» текст на отдельные единицы и где-то фиксирует их. На следующем шаге производится подсчет одинаковых единиц, и, наконец, производится упорядочение подсчитанных единиц. Такая методика подробно описана во многих работах группы «Статистика речи».[59]

В общем виде алгоритм процесса построения частотного списка человеком можно представить следующим образом:

1) выделение единиц текста и их запись на «карточки»;
2) подсчет частоты встречаемости одинаковых единиц;
3) упорядочение всех единиц по частоте;
4) упорядочение единиц с одинаковой частотой по алфавиту.

Для получения достоверных результатов необходимо, чтобы величина объема выборки N была достаточно велика. Мозг человека, обладая большой надежностью, способен справиться (с помощью карандаша и бумаги) с любым объемом N. Другое дело, что на это уйдет много времени. Те ЭВМ, которые используются лингвистами сейчас, и, очевидно, те, которые будут в их распоряжении в течение ближайших 5—10 лет, не позволяют обработать сразу большой объем лингвистической информации. Это объясняется,

[59] См., например: П. М. Алексеев. Частотный словарь английского подъязыка электроники. АКД, Л., 1965; Л. И. Ешан. Опыт статистического описания научно-технического стиля румынского языка. АКД. Изд. ЛГУ, 1966; Т. А. Микерина. Некоторые статистические приемы лексико-морфологического описания функционального стиля; О. А. Нехай. Статистика и артоматический анализ текста. АКД; А. И. Чапля. Опыт статистического описания лексической комбинаторики. КД (рукоп.), Ленинградское отд. Инст. языкознания АН СССР, 1967.

во-первых, недостаточностью объема накопителей, и во-вторых — все еще невысокой степенью надежности ЭВМ при обработке больших массивов информации. Третья причина, не позволяющая обрабатывать сразу большой массив текста, связана с особенностью лингвистической информации. В ходе кодирования человеком этой информации на носителях (перфоленте и перфокартах) неизбежно появляется некоторое количество ошибок, которые машина, будучи пунктуальным счетчиком, но не обладая эвристическими возможностями исправлять эти ошибки, выдает на печать. Эти ошибки тем многочисленнее, чем больше массив текста. Поэтому целесообразнее обрабатывать на ЭВМ сразу не весь массив из N единиц, а какие-то его части. После проверки результатов обработки таких порций и исправления в них ошибок частичные частотные списки объединяются. Это позволит ослабить нагрузку на ЭВМ и выявлять ошибки на промежуточных этапах обработки, сократить их число в итоговом частотном списке.

Разделим весь текст, содержащий N единиц, на п е р в и ч н ы е м а с с и в ы, которые будем обозначать N_i ($i=1, 2, 3 \ldots k$). Тогда общая задача получения частотного списка из N единиц разделится на $k+1$ отдельную задачу: k одинаковых задач получения частотных списков из k первичных массивов и $(k+1)$-я — задача объединения первичных списков.

В свою очередь каждый массив N_i будем составлять из t э л е м е н т а р н ы х м а с с и в о в (порций), каждый из которых содержит n единиц, так, чтобы

$$N_i = n \cdot t.$$

В таком случае общий алгоритм построения частотного списка машиной можно представить в виде следующей блок-схемы (рис. 3, стр. 311).

Рассмотрим подробнее эту схему в применении к конкретной задаче, а именно — к задаче получения частотного списка букв и буквосочетаний.

6.3. Лингвистические аспекты задачи

Данные о вероятности появления отдельных букв и их сочетаний находят широкое применение в самых различных областях науки. Описание процесса чтения текста и переработки информации человеком требует знания характера влияния вероятностей появления букв и их комбинаций на скорость и точность восприятия информации. Читает ли человек, опираясь на статистически наиболее частые буквы, или наоборот — на наиболее редкие? Какой из вариантов чтения наиболее легок? Ответы на эти и им подобные вопросы связаны со знанием указанных выше вероят-

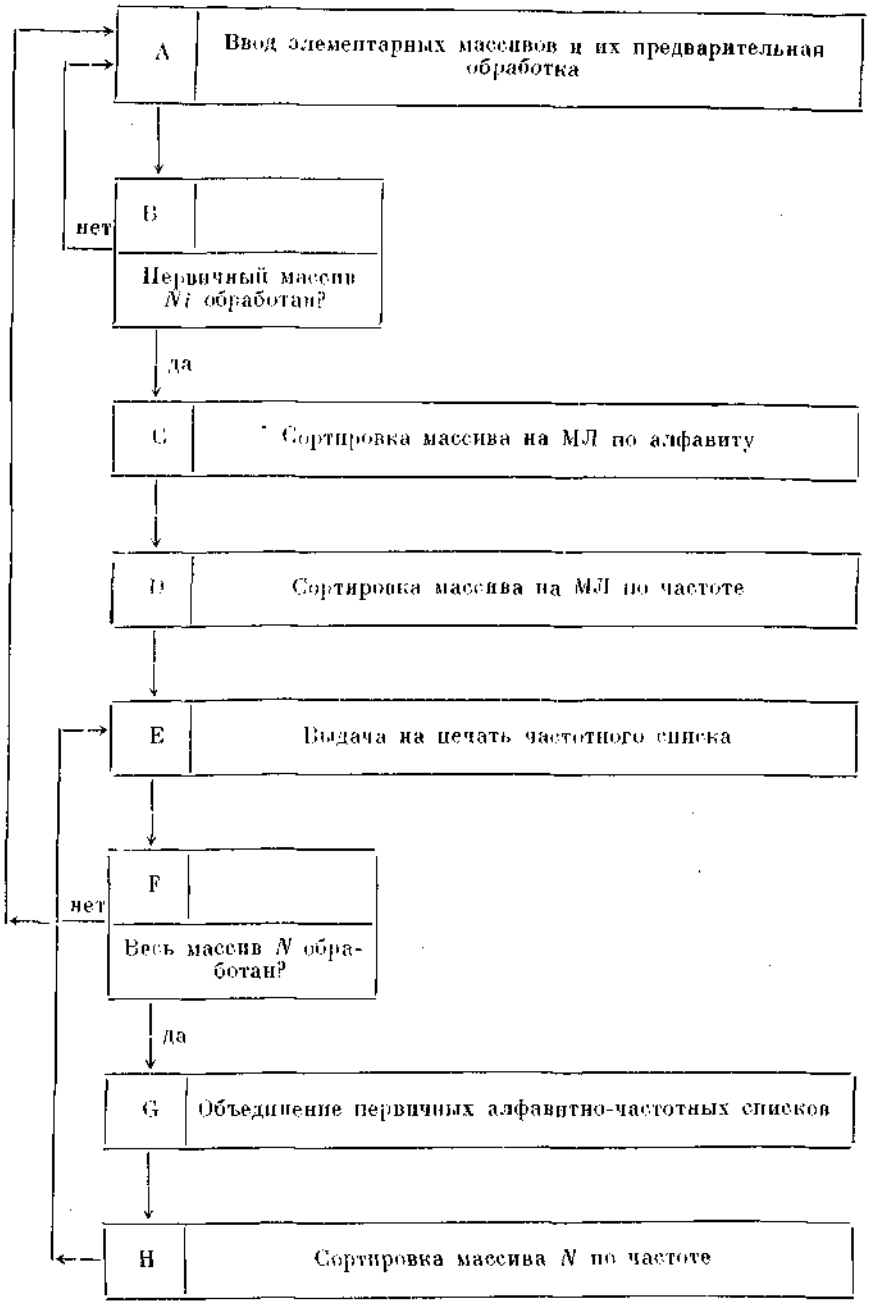


Рис. 3. Общий алгоритм построения частотного списка из N единиц.

ностей.[60] Дешифровка неизвестных ранее письменностей и языков часто осуществляется на сопоставлении различных единиц языков, таких как отдельные буквы, слоги, морфемы. Подобное сопоставление требует огромной затраты труда на изучение количественных характеристик текста. Частотность букв и их сочетаний используется и в исследованиях по определению информационных характеристик текста (расчет энтропии на 1-ю, 2-ю, 3-ю и т. д. букву текста).[61] Наконец буквенная статистика является основным условием для статистико-комбинаторного моделирования языков.[62]

Далее, изучение таких вопросов, как функциональная нагрузка фонем в словаре и тексте, частота встречаемости фонем и их сочетаний в словаре и тексте также связаны с исследованием функционального распределения графем в тексте. Используются эти данные и при изучении вопросов о пограничных сигналах, в определении границ морфем и слогов и т. д. Здесь, правда, нужно помнить о трудностях перевода графического представления звуков в фонемное.[63]

Итак, вопрос о частности употребления букв и их комбинаций играет большую роль в решении многих теоретических и прикладных задач.[64] Однако решение этого вопроса связано с исследованием больших объемов текста. Насколько это трудоемкая задача, видно из следующей формулы

$$N = k^n, \qquad (1)$$

где N — число возможных сочетаний по n букв из алфавита в k букв.[65] Например, для русского алфавита, содержащего 32 буквы, число двухбуквенных сочетаний будет $-32^2=1024$, трехбуквенных $-32^3=32768$, четырехбуквенных $-32^4=1048576$.

Вполне естественно, что не все они реализованы в языке, однако число их (особенно для больших значений n) достаточно велико. Именно поэтому для исследований, проводимых вручную, обычно брались небольшие выборки в 10 000—30 000 знаков [66]).

В приведенных работах, а также в ряде других [67] даются в основном сведения о распределении букв, двухбуквенных и трехбуквенных сочетаний, ибо для дальнейшего анализа пришлось бы исследовать огромный исходный материал.

В последние годы, с развитием вычислительной техники, появился ряд исследований по статистике буквосочетаний и сочетаний фонем, — исследований, выполненных с помощью автоматов.[68] Недостатком этих работ является то, что все они предназначены в основном для выполнения разовых работ. Программы не были универсальными. Выбирался какой-то один тип буквосочетаний из определенного текста.

Нами предлагается универсальный алгоритм (и соответственно программа), позволяющий определить частотность отдельных букв (в сплошном тексте, начальных букв слова, конечных букв слова), а также сочетаний букв в тех же позициях.[69]

Исходным материалом для нас является текст, содержащий N_i словоупотреблений, построенный с соблюдением всех правил, изложенных в § 2, и закодированный на перфоленте.

При этих предположениях решение задачи построения частотного списка букв и их комбинаций может быть сведено к выполнению следующих последовательных этапов:

1) из текста выбираются все необходимые n-буквенные сочетания $(n=1\div6)$;

2) подсчитывается частота встречаемости каждого сочетания;

3) полученные разные сочетания располагаются в частотно-алфавитный список.

---

[60] А. Н. Леонтьев. Е. П. Кричик. Переработка информации человеком в ситуации выбора. Сб. «Инженерная психология», изд. МГУ, 1964; С. Н. Кечхуашвили. Статистическое распределение графем и их информативная ценность. ВП, 1966, № 2, стр. 47—48.

[61] Г. П. Богуславская. Энтропия английского печатного текста; А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин. Энтропия русского языка. ВЯ, 1962, № 6; Р. Г. Пиотровский. Информационное измерение языка.

[62] См. сб.: Статистико-комбинаторное моделирование языков, Л., 1965.

[63] P. Denes. The Use of Computers for Research in Phonetics. Proceedings of the IX International Congress of Linguists, s'Gravenhage, 1964; G. Herdan. The Relation between the Functional Burdening of Phonemes and the Frequency of Occurence. «Language and Speech», 1958, 13. H. Kučera. Statistical Determination of Esotopy. Proceedings of the IX International Congress. . .

[64] В своей работе «Deutsche Sprachstatistik» (I—II, Hildesheim, 1964) известный немецкий лингвист Х. Майер показывает важность данных буквенной статистики при решении вопросов диалектальных исследований, реформы немецкой орфографии, восприятия речи и т. п.

[65] L. Doležel, J. Prucha. A Statistical Law of Grapheme Combinations. Prague Studies in Mathematical Linguistics, Prague, 1966.

[66] Н. В. Петрова. Энтропия французского печатного текста. Изв. АН СССР, серия литературы и языка, т. XXIV, 1965, вып. 1; Charles P. Bourne, Donald F. Ford. A Study of the Statistics of Letter in English Words. «Information and Control», vol. 4, № 1, New York, 1961.

[67] См., например: В. Н. Топоров. Материалы для дистрибуции графем в письменной форме русского языка. Сб.: Структурная типология языков. М., 1966; L. Doležel. Předběžný odhad entropie a redundance psané češtiny. Slovo a slovesnost, 24, 1963.

[68] Н. П. Елкина, А. С. Юдина. Статистика открытых слогов русской речи. В сб.: Вычислительные системы, Новосибирск, 1964, 14; Д. С. Лебедев, В. А. Гармаш. Статистический анализ трехбуквенных сочетаний русского текста. В сб.: Проблемы передачи информации, М., 1959, 2.

[69] Первоначальный вариант соответствующего алгорима приведен в докладе: А. А. Пиотровская, А. В. Зубов. Автоматическое определение вероятностей появления в тексте n-буквенных сочетаний. Вторая межвузовская конференция по частотным словарям и автоматической переработке текста, Минск, 1968.

Выборка из текста должна удовлетворять следующим условиям:

а) величина n изменяется от 1 до n, равного числу букв, помещающихся в одной ячейке накопителя машины; для ЭВМ «Минск-22» $n=1\div6$;

б) выборка может быть проведена из сплошного текста, который рассматривается как одно «большое» слово, с учетом всех знаков препинания; пробел между словами считается за букву; при этом считается общее количество буквосочетаний, пробелов, точек и остальных знаков (вместе);

в) выборка может быть проведена по первым n буквам каждого очередного словоупотребления текста; считается общее количество буквосочетаний и словоупотреблений;

г) выборка может быть проведена по последним n буквам каждого очередного словоупотребления; считается общее количество буквосочетаний и словоупотреблений текста.

Максимальный обрабатываемый объем N_i практически зависит от величины (в ячейках) внешнего накопителя машины и от используемой программы упорядочения информации. Применяемая в группе «Статистика речи» программа упорядочения [70] позволяет обрабатывать тексты объемом до нескольких миллионов словоупотреблений.

6.4. Математическая постановка задачи

1) В зависимости от типа выборки текст сегментируется на отдельные n-буквенные сочетания. Каждое сочетание включается в машинный документ (МД), состоящий из двух ячеек. В первой ячейке располагается n-буквенное сочетание, во второй — частота его употребления.

2) Массивы МД, образованные по каждому элементарному тексту, записываются во внешнем накопителе машины — на МЛ.

3) Массивы на МЛ упорядочиваются по первым ячейкам машинных документов, причем они располагаются так, чтобы на первом месте стоял МД с наименьшим содержимым первой ячейки, за ним — с бóльшим и т. д. (сортировка по алфавиту).

В зависимости от типа используемой программы упорядочения может возникнуть необходимость еще в одном переходе, не отраженном в общей блок-схеме (см. рис. 3). Дело в том, что программа упорядочения работает в простейшем случае с таким количеством массивов на МЛ, которое соответствует количеству элементарных массивов. В разных массивах могут встретиться одинаковые сочетания. Одни программы упорядочения работают так, что при сортировке по алфавиту частоты одинаковых сочетаний в разных порциях массивов на МЛ складываются между собой и само сочетание в упорядоченном массиве записывается лишь один раз с суммарной частотой. Другие программы упорядочения не складывают между собой частоты одинаковых сочетаний, а располагают эти сочетания одно под другим с их частотой в каждом элементарном массиве. В этом последнем случае нужен еще один переход, который бы суммировал частоты всех одинаковых сочетаний, стоящих в упорядоченном списке одно под другим. Назовем эту операцию «сжатием массива».

4) Сжимается упорядоченный массив, если в этом есть необходимость. При этом для МД, обладающих одним и тем же содержимым первых ячеек, необходимо сложить между собой содержимые вторых ячеек. Затем образовать новый МД, в первой ячейке которого записать один раз код n-буквенной единицы, а во второй — суммарную частоту всех одинаковых единиц.

5) Полученный на предыдущем этапе массив упорядочивается по вторым ячейкам МД так, чтобы на первом месте стоял МД с наибольшим значением содержимого 2-й ячейки, за ним — с меньшим и т. д. (сортировка по частоте). Для МД, обладающих одинаковым содержимым вторых ячеек, следует провести упорядочение по первым ячейкам, располагая эти МД так, чтобы на первом месте стоял документ с наименьшим содержимым первой ячейки.

6) Содержимое машинных документов, полученных на предыдущих этапах, расшифровывается путем вывода их на устройство печати. При выводе каждого МД необходимо указать его порядковый номер, содержимое первой ячейки, содержимое второй ячейки и суммарное значение содержимых вторых ячеек всех выведенных ранее МД, а также выводимого МД (абсолютная накопленная частота).

6.5. Разработка метода решения. Принципиальный алгоритм решения задачи

Задача определения частоты употребляемости букв и их сочетаний является для машины одной из простейших. Соответственно и методы решения этой задачи не отличаются разнообразием. В соответствии с математической постановкой задачи принципиальным алгоритмом для нас будет являться алгоритм, схемы которого приведены на рис. 17—20 (см. Приложение).

6.6. Программирование алгоритма

Программирование алгоритма заключается в том, чтобы расписать все его блоки в виде последовательности элементарных операций каждой машины. Такую определенную последовательность команд и называют п р о г р а м м о й р е ш е н и я з а д а ч и. При этом каждый блок подалгоритма заменяется одной или несколькими командами из системы команд конкретной ЭВМ.

[70] Э. Н. Х о т я ш о в. Универсальная программа сортировки для ЭВМ «Минск-22». Минск, 1965.

Именно на этом этапе решения лингвистических задач особенно заметно разделение труда между лингвистами и математиками-программистами. В этом отражается не только вполне естественное разделение квалификации между ними, но также деление между информацией о тексте, написанном на естественном языке, которая имеет силу безотносительно к машине, и техникой передачи этой информации к конкретной машине. До сих пор проблема общения человека и машины является нерешенной. Как показывает опыт работы группы «Статистика речи», наибольший эффект в этом отношении может быть достигнут, когда в одном человеке сочеталась бы лингвистическая подготовка с некоторой квалификацией в области программирования.

6.7. Решение задачи на машине. Анализ полученных результатов

После того как алгоритмы лингвистической задачи запрограммированы, наступает окончательный этап — решение задачи. Этот этап связан с реализацией всех программ на ЭВМ и включает два подэтапа:

1) отладка программы;

2) непосредственно решение задачи по отлаженной программе.

При подготовке к решению любой задачи неизбежны ошибки. Это или логические недоработки в процессе составления алгоритма, или ошибки в передаче переходов алгоритма в конкретных адресах машины, или, наконец, ошибки в процессе кодирования исходной информации и программы. Все эти несоответствия и выявляются в процессе отладки программы. При этом осуществляется обратная связь между машиной и человеком. Аварийные остановы — это сигналы машины о неполадках в программе, в ее отдельных устройствах и т. п. Это стимулирует вмешательство человека для корректировки заданной машине информации. Таким образом, создается избыточность, предохраняющая от искажений при переработке лингвистической информации в системе «человек—машина». Процесс отладки — это не что иное, как диалог человека и машины. Аварийные остановы машины — это сигналы машины о непонимании текста. Исправления, вносимые после таких остановов, представляют собой реакцию человека на запрос машины. Это как бы перефразировка исходного задания, направленная на установление взаимопонимания человека и машины.

Решение задачи по отлаженной программе есть автоматическая реализация построенной модели. Результаты этой реализации показывают, насколько построенная модель точно отражает те или иные свойства оригинала. Например, при рассмотрении функциональных моделей речевого поведения человека результаты видны непосредственно и их анализ не представляет никаких затруднений (см. ч. II и III, стр. 371 и далее). Иное дело, когда

рассматриваются статистические модели речи — разного вида частотные списки. Здесь связь между оригиналом (распределением вероятностей лингвистических единиц в подъязыке) и статистической моделью (частотным списком) непосредственно не видна. Приходится для этого получать ряд других характеристик частотных списков, как-то: величину покрываемости наугад взятого текста данным частотным списком, энтропию, приходящуюся на единицу текста, границы достоверности полученных значений частот, степень близости полученного эмпирического распределения к вычисленным иным путем теоретическим распределениям и т. д.

Методика анализа частотных списков и выяснения их основных лингвостатистических характеристик приводится в работе: А. В. Зубов, Э. Н. Хотяшов. Статистический анализ текста с помощью электронно-вычислительной машины. В сб.: Энтропия языка и статистика речи, Минск, 1966, стр. 137—166.

На рис. 4 и 5 представлены начала частотных списков буквосочетаний, полученных на машине по разработанным программам.

§ 7. Выбор наиболее частых словоформ [71] текста

7.1. Лингвистические аспекты задачи

Словарный состав языка трудно поддается формальному описанию, ибо функционирование каждого отдельного слова определяется обычно очень сложным комплексом причин (семасиологических, стилистических, грамматических), которые бывает трудно разложить на элементарные шаги. Однако данные о функционировании словаря необходимы как для собственно лингвистических исследований, так и для решения прикладных задач. При исследовании значительных массивов материала, в котором действует большое количество факторов, причем поведение каждого учесть невозможно, целесообразно применить вероятностно-статистическую методику. Как уже говорилось, с помощью этой методики удается выделить наиболее важные факторы, определяющие поведение изучаемого объекта. Частотные словари, являясь статистической моделью лексики, позволяют решить ряд проблем общей лексикологии — вопрос о ядре словаря, актив-

[71] Как и во всех остальных работах группы «Статистика речи», в данной работе, помимо словоупотребления (см. стр. 293), выделяются также словоформы и слова. Под словоформой мы понимаем полностью совпадающие словоупотребления. Слово выступает как некоторый класс семантически и грамматически связанных между собой словоформ. Таким образом, словоупотребление является единицей текста, словоформа — единицей частотного словаря, слово — единицей двуязычного, толкового, энциклопедического и т. п. словарей. То состояние основной единицы текста, когда она уже изъята из текста, но еще не включена в конечный частотный список, будем называть также словоупотреблением.

1	THE	549	549
2	ING	183	732
3	ION	146	878
4	AND	109	987
5	EAM	85	1072
6	ITY	81	1153
7	URE	68	1221
8	AGE	64	1285
9	HIS	54	1339
10	INE	50	1389
11	T&D	45	1434
12	HAT	44	1478
13	FOR	42	1520
14	ROM	42	1562
15	ENT	41	1603
16	LOW	39	1642
17	ARE	38	1678
18	WEN	36	1714
19	NCE	36	1750
20	OSS	35	1785
21	RGW	34	1819
22	LLY	32	1851
23	TES	30	1881
24	ERE	29	1910
25	ORK	28	1938
26	ESS	27	1965
27	KET	27	1992
28	ILL	25	2017
29	ITH	25	2042
30	LSE	25	2067
31	ITS	24	2091
32	NTS	24	2115
33	OWN	24	2139
34	SED	24	2163
35	ONS	23	2186
36	ICH	22	2208
37	SES	22	2230
38	ARV	21	2251
39	LAR	21	2272
40	TIC	21	2293
41	WER	21	2314
42	ADE	19	2333
43	BLE	19	2352
44	EEN	19	2371
45	GLE	19	2390
46	IAL	19	2409
47	IVE	19	2428
48	ORE	18	2446
49	ULD	18	2464
50	ATE	17	2481

Рис. 4. Фрагмент частотного списка трехбуквенных сочетаний по последним буквам словоформ (подъязык английских судовых механизмов).

151	ПЛОС	7	2842
152	ПОСТ	7	2849
153	ПРОБ	7	2856
154	ПРОТ	7	2863
155	РЕЖИ	7	2870
156	ТАК-	7	2877
157	ТОКИ	7	2884
158	ТОРЦ	7	2891
159	ЧЕРЕ	7	2898
160	ШИРО	7	2905
161	ЭТИХ	7	2912
162	АНАЛ	6	2918
163	ВДОЛ	6	2924
164	ВИДО	6	2930
165	ВЛИЯ	6	2936
166	ГРАН	6	2942
167	ИМПУ	6	2948
168	КРОМ	6	2954
169	ЛИБО	6	2960
170	МОГУ	6	2966
171	НАИБ	6	2972
172	НАЛИ	6	2978
173	НИЗК	6	2984
174	ОБЕИ	6	2990
175	ОКОЛ	6	2996
176	ОТНО	6	3002
177	ПЛЕЧ	6	3008
178	ПО-С	6	3014
179	ПОДВ	6	3020
180	ПОДС	6	3026
181	ПРАВ	6	3032
182	ПРИН	6	3038
183	СПОС	6	3044
184	СПРА	6	3050
185	СХЕМ	6	3056
186	ТАКО	6	3062
187	ТЕХН	6	3068
188	ЦЕЛЕ	6	3074
189	ЦОКО	6	3080
190	ЭНЕР	6	3086
191	АМПЛ	5	3091
192	БУДЕ	5	3096
193	В-СЛ	5	3101
194	ВЕРХ	5	3106
195	ВНЕШ	5	3111
196	ВЫХО	5	3116
197	ДИЭЛ	5	3121
198	ДОБА	5	3126
199	ДОПО	5	3131
200	ЕДИН	5	3136

Рис. 5. Фрагмент частотного списка четырехбуквенных сочетаний по первым буквам словоформ (подъязык русской электроники).

ном и периферическом словарях. Статистические данные о составе лексики открывают большие возможности для типологических исследований, для разграничения функциональных стилей языка и авторских стилей, а также в области исторической лексикологии.[72]

Что касается практического применения частотных словарей, то необходимо отметить их исключительную важность для автоматической обработки текстовой информации (автоматический перевод, реферирование и т. п.). Они являются лексической основой базовых языков (ср. выше, стр. 290). Можно также отметить их применимость к установлению авторства анонимных текстов и к методике обучения языкам.[73] Таким образом, диапазон задач, при решении которых в той или иной степени используются частотные словари, достаточно широк. Но, как мы отмечали выше (см. стр. 294), для получения сколько-нибудь достоверных данных необходимо, чтобы объем выборки был достаточно велик. Механическая, нетворческая работа по статистическому анализу больших объемов текста отнимает у исследователя месяцы и годы.

С развитием вычислительной техники вполне естественно возник вопрос о передаче таких механических операций машинам. Первые работы в этом направлении были связаны с использованием счетно-аналитических машин (САМ). При этом человек самостоятельно сегментировал текст на слова и переносил их на перфокарты. Дальнейший подсчет и сортировку производили САМы.[74]

Совершенствование электронно-вычислительных машин позволило еще более автоматизировать процесс получения частотных словарей. Первый словарь, полученный с помощью ЭВМ, был опубликован Г. Йоссельсоном. В последние годы некоторое число таких словарей получено как за рубежом,[75] так и у нас.[76]

Основной недостаток всех программ, по которым составлялись частотные словари, — отсутствие универсальности и направленности на большие исходные объемы информации.

[72] Ср.: Б. Н. Головин. О роли статистики в описании языковых и речевых стилей. В сб.: Частотные словари и автоматическая переработка лингвистических текстов, Минск, 1968; Л. Н. Засорина. Частотные словари и вопросы лексикостатистики. МКЧСА; О. В. Творогов. О применении частотных словарей в исторической лексикологии русского языка. ВЯ, 1967, № 2.

[73] Э. У. Бабаева. О статистических методах установления авторства анонимных текстов. МКЧСА, стр. 58—60; Г. А. Эрнв. Роль и место частотного словаря в процессе обучения языку. Там же, стр. 83—86.

[74] Р. Буза. Электротехника при механизации филологического анализа. Русск. пер. АЛ.

[75] H. Josselson. The Russian Word Count and Frequency Analysis of Grammatical Categories of Standart Literary Russian. Detroit, 1953.

[76] Наиболее крупные машинные частотные словари получены сотрудниками группы «Статистика речи». Это отраслевые словари русского, английского и французского языков из текстов по 400 тыс. словоупотреблений каждый, а также ряд немецких, испанских, казахских словарей, построенных на основе текстов, в 200, 100 и 50 тыс. словоупотреблений.

Мы предлагаем алгоритмы и программы, которые позволяют получить частотный словарь практически из любого текста, написанного кириллицей или латиницей, вне зависимости от его объема. Об эффективности программы, построенной по приводимым ниже алгоритмам, можно судить по следующему примеру.

Чтобы получить частотный словарь из выборки в 200 тыс. словоупотреблений, исследователю необходимо около двух лет интенсивной работы. Применение ЭВМ средней мощности, например «Минск-22», дает возможность выполнить всю работу за 2.5 месяца. При этом основная часть времени идет на кодирование текста (из расчета 4000 словоупотреблений за 7 часов работы). На сам процесс получения словаря на машине требуется всего лишь несколько часов (12 часов для ЭВМ «Минск-22», обладающей, как известно, малым быстродействием: 5000—6000 операций в секунду).

В общем виде задачу построения частотного словаря из текста, содержащего N_t словоупотреблений, можно представить состоящей из следующих отдельных этапов:

а) из текста выбираются все различные словоформы;

б) подсчитывается частота встречаемости каждой словоформы;

в) полученный массив словоформ располагается в частотно-алфавитный список.

Дополнительными условиями при этом являются следующие:

1) в качестве промежуточного результата — получить алфавитно-частотный список словоформ;

2) при выдаче словаря на печать указать порядковый номер i словоформы, ее абсолютную частоту F_i и абсолютную накопленную частоту F_i^* (при необходимости можно выдать на печать и другие данные).

7.2. Математическая постановка задачи

Словоупотребления текста содержат различное число букв, поэтому необходимо каким-то образом стандартизировать эту величину. Статистические исследования, проводимые в группе «Статистика речи» по разным языкам, показали, что почти во всех европейских языках длина слова не превышает 24 буквы. В связи с этим предполагается, что самое длинное словоупотребление текста не превышает такой величины.

Математически задача может быть сформулирована следующим образом:

1) текст сегментируется на отдельные словоупотребления, каждое словоупотребление включается в машинный документ (МД), состоящий из $m+1$ ячейки; в первых m ячейках располагается код словоупотребления, в ячейке $(m+1)$ — частота словоупотребления в порции текста;

51	THIN	53	12752
52	PROCESS	53	12810
53	DIFFUSION-A	53	12867
54	SPUTTERING	53	12924
55	VERY-D	53	12981
56	SPUTTERED	56	13037
57	TO-BE	55	13092
58	ALSO	54	13146
59	FILM	54	13200
60	JUNCTIONS	50	13250
61	LAYERS	49	13299
62	LOW	49	13348
63	SECOND-A	49	13397
64	WILL	49	13446
65	MADE	46	13492
66	OBTAINED	46	13538
67	PROPERTIES	46	13584
68	DIFFUSION	44	13628
69	ALL	43	13671
70	CONCENTRATION	43	13714
71	REGION	43	13757
72	SUBSTRATE-S	43	13800
73	OTHER	42	13842
74	HOWEVER	41	13883
75	WERE-V	41	13924
76	SUBSTRATE-A	40	13964
77	BUT	39	14003
78	DEPOSITION	39	14042
79	MEASUREMENTS	39	14081
80	TWO	39	14120
81	DURING	38	14158
82	DIAMOND-A	37	14195
83	GROWTH	36	14231
84	JUNCTION	36	14267
85	ONE-N	36	14303
86	TECHNIQUES	36	14339
87	USE	36	14375
88	BREAKDOWN	35	14410
89	DEPOSITION-A	35	14445
90	EXPERIMENTAL	35	14480
91	GERMANIUM	35	14515
92	MATERIAL-S	35	14550
93	RANGE	35	14585
94	THROUGH	35	14620
95	WHEN	35	14655
96	MATERIAL	34	14689
97	RATE	34	14723
98	RESISTIVITY	34	14757
99	SHOULD	34	14791
100	SHOWS-V	34	14825

Рис. 6. Фрагмент частотного списка словоформ подъязыка английской электроники.

2) массивы МД, полученные после преобразования каждой порции текста, записываются на МЛ;

3) массивы МД на МЛ упорядочиваются по содержимому первых m ячеек МД; на первое место при этом ставится МД, содержащий наименьшее значение содержимого указанных m ячеек, на второе — с бóльшим и т. д. (сортировка по алфавиту);

4) упорядоченный по алфавиту массив сжимается; при этом для МД, обладающих одним и тем же содержимым первых ячеек, слагаются содержимые их $(m+1)$-х ячеек; вместо группы машинных документов, обладающих указанным свойством, образуется новый МД, в первых m ячейках которого располагается общее содержимое m ячеек группы, а в ячейке $(m+1)$ — суммарная частота всех одинаковых словоупотреблений (сумарное содержимое $(m+1)$-х ячеек);

5) сжатый массив упорядочивается по $(m+1)$-м ячейкам машинных документов; на первое место ставится МД с наибольшим содержимым $(m+1)$-й ячейки, на второе — с меньшим и т. д.; для группы МД, имеющих одинаковое содержимое $(m+1)$-х ячеек, МД упорядочиваются по содержимому первых m ячеек; на первое место при этом ставится документ с наименьшим содержимым первых m ячеек, на второе — с бóльшим и т. д.;

6) содержимое машинных документов, полученных на предыдущем этапе, расшифровывается путем вывода их на устройство печати; при выводе каждого МД указывается порядковый номер его, содержимое первых m ячеек, содержимое $(m+1)$-й ячейки и суммарное значение содержимых $(m+1)$-х ячеек всех выведенных ранее машинных документов и выводимого МД (абсолютная накопленная частота).

7.3. Разработка метода решения

Из математической постановки задачи видно, что она сводится по существу к решению нескольких более простых задач:

1) преобразование каждого словоупотребления текста в МД;
2) сортировка МД по алфавиту;
3) сжатие массива МД;
4) сортировка МД по частоте и алфавиту;
5) вывод списка на печать.

Все эти задачи можно решать различными методами. Можно, например, совместить выполнение задач 1 и 2, применяя метод переполнения или метод частичных словников.[77] При этом из первоначальной перфоленты, содержащей исходный текст, получается некоторый упорядоченный по алфавиту подмассив и новая

[77] Г. Г. Белоногов, Р. Г. Котов. Автоматизированные информационно-поисковые системы. М., 1968, стр. 83—84; А. Ж. Т. Колин. Автоматическое составление словника. Русск. пер. АЛ.

перфолента со словами, не вошедшими в первый подмассив. Эта лента снова пропускается через машину. Получается второй подмассив и третья перфолента и т. д. Подобная процедура, на наш взгляд, нецелесообразна по следующим причинам:

а) как известно, устройства ввода и вывода работают в тысячи раз медленнее, чем сама машина; поэтому на многократный ввод перфоленты и вывод ее тратится очень много времени;

б) чем больше перфоленты участвует в работе, тем выше вероятность появления ошибок;

в) получаемые при такой работе малые по объему подмассивы требуют их объединения, чтобы получить единый алфавитный список; это связано со вводом в машину больших массивов перфолент или ручным объединением подмассивов; подобное объединение малых массивов становится задачей крайне трудоемкой для больших объемов информации (а именно для таких объемов и необходимо применять ЭВМ).

То же самое можно сказать и о методе, предлагаемом в работе английских исследователей.[78]

Делались попытки применить для получения алфавитного списка слов так называемый промежуточный накопитель — магнитные барабаны.[79] Однако длительность обработки (2—3 часа на 1000 словоупотреблений) приводит к мысли о нецелесообразности подобного метода.

Анализ указанных методов и наш опыт работы в группе «Статистика речи» позволяют предложить новый алгоритм, направленный в основном на использование внешнего накопителя — магнитных лент.

7.4. Разработка принципиального алгоритма решения задачи [80]

Принципиальный алгоритм получения частотного списка из текста, содержащего N_t словоупотреблений, приведен на рис. 17—21 (см. Приложение).

[78] Andrew D. Booth, L. Brandwood, J. Cleave. Mechanical Resolution of Linguistic Problems. London, 1958, pp. 35—43.

[79] О. В. Маслиева. Опыт применения ЭВМ для алфавитной классификации слов. МКЧСА.

[80] За основу алгоритма и программы выделения наиболее частых словоформ текста взяты алгоритм и программа, разработанные Э. Н. Хотяшовым (см. стр. 311). Основные отличия нашей программы от указанной заключаются в следующем:

1) в алгоритме «Ввод элементарных массивов и их предварительная обработка с записью на МЛ» нашей программы предусмотрена возможность обработки текстов, написанных не только латиницей, но и кириллицей, здесь, кроме того, предусмотрены проверки, не допускающие попадания в частотный список знаков препинания, а также любых цифр, случайно оказавшихся на перфоленте;

7.5. Программирование алгоритмов. Решение задачи на машине

Рассмотренные алгоритмы были запрограммированы для ЭВМ «Минск-22».

Программы проверялись на русском, английском, французском, немецком, испанском, латышском текстах, а также на текстах, написанных на некоторых тюркских и кавказских языках.[81]

На рис. 6 уже был приведен фрагмент из частотного списка, полученного по разработанной нами программе.

§ 8. Построение алфавитно-частотного обратного словаря

8.1. Лингвистические аспекты задачи

Обратный алфавитно-частотный словарь представляет собой список всех словоформ текста с указанием частоты употребления их в тексте, причем упорядочение этих словоформ осуществляется по алфавиту не первых, а конечных букв.

При этом словоформы с одинаковыми окончаниями оказываются собранными вместе. Такие списки дают важный материал для изучения флективной и агглютинативной морфологий, словообразования, этимологии и ряда других отраслей лингвистики.[82]

Наличие обратных статистических словарей по различным языкам является одним из условий для осуществления квантитативных оценок родства языков в области словообразования и флективной морфологии. Обратные словари помогают выявить омонимию (в том числе ложную омонимию) флексий, что крайне важно при составлении программ морфологического анализа и синтеза на ЭВМ.

Из других применений обратных словарей можно отметить их использование при составлении словаря рифм; составление обратного частотного списка словоформ обычно является необходимым условием при разного вида дешифровках текста. При наличии «прямых» и обратных частотных словарей можно значи-

2) в нашей программе применен принципиально иной алгоритм вывода частотно-алфавитного списка на печатающее устройство; он занимает гораздо меньшее число ячеек МОЗУ и предусматривает вывод не только величины абсолютной частоты F_i, но также относительной частоты f_i, относительной накопленной частоты f_i^*, антропии H_i и H_i^* каждой единицы; пользуясь нашим алгоритмом, можно выводить на устройство печати тексты, написанные и латиницей и кириллицей.

[81] К. Б. Бектаев, А. Джубанов, А. В. Зубов. Автоматическое построение частотных списков (прямого и обратного). Вестник АН Казахской ССР, 1968, 12; А. В. Зубов, А. И. Чапля. Опыт одновременного получения частотных списков и их статистических параметров. МКЧСА.

[82] Р. В. Бахтурина, И. А. Мельчук. [Рец. на]: M. L. Alinei. Dizionario Inverso Italino, The Hague, 1962. ВЯ, 1965, № 5.

тельно упростить алгоритм статистико-комбинаторного моделирования. Обратные словари находят широкое применение, поэтому работа по их составлению стала в последние годы очень интенсивной. Уже существует большое число обратных словарей по различным языкам.[83] Однако большинство этих словарей составлено «вручную», что потребовало большого объема механической работы, в то время как применение ЭВМ значительно экономит усилия лингвиста и сокращает время, необходимое для составления таких словарей. Объем работы, необходимый для получения обратного частотного словаря, несколько меньше по сравнению с тем, что требуется для обычного частотного словаря, вследствие того, что для первого не требуется сортировка по частоте (такое упорядочение никакой дополнительной информации не дает, а а лишь удлиняет время поиска в словаре).

При использовании ЭВМ для получения обратного словаря никакой дополнительной домашинной обработки текста (или прямого словаря) не требуется. Машина по специальной программе инверсирует каждое прочитанное слово. При использовании же для получения такого словаря счетно-аналитических машин (САМ) требуется двойное количество перфокарт.[84]

Процесс получения обратного частотного словаря состоит из следующих этапов:

а) из текста выбираются все различные словоформы;

б) подсчитывается частота встречаемости каждой слоформы;

в) полученный массив словоформ располагается в алфавитном порядке по их конечным буквам.

8.2. Математическая постановка задачи

Приведенное выше лингвистическое задание представляет собой алгоритм, по которому действовал бы лингвист, составляя «вручную» обратный частотный словарь. В математической постановке задачи мы несколько изменим указанную последовательность отдельных заданий. Это вызвано стремлением сделать все алгоритмы (программы) статистического анализа текстов в какой-то степени стандартными и универсальными. Такая стандартизация и универсальность алгоритмов состоит, в частности, во взаимозаменяемости блоков. Так, желательно, чтобы задачи составления обратного частотного словаря и обычного частотного словаря отличались одна от другой максимум на один-два подалгоритма. Из этих соображений мы формулируем рассматриваемую задачу следующим образом:

[83] См., например: И. А. Мельчук. [Рец. на]: A. Juilland. Dictionnaire Inverse de la Langue Française. London, 1965. ВЯ, 1966, № 6; H. H. Bielfeld. Rückläufiges Wörterbuch der russischen Sprache der Gegenwart. Berlin, 1958.

[84] Й. Штипдлова. Обратные словари. АЛ, стр. 89—90.

201	TENIA	2	691
202	CEREMONIA	1	692
203	PRECARIA	1	693
204	ESTUDIARIA	1	694
205	EXTRAORDINARIA	1	695
206	NECESARIA	1	696
207	CANCILLERIA	4	700
208	ARTILLERIA	1	701
209	BATERIA	2	703
210	MATERIA	1	704
211	FRIA	2	706
212	SATISFACTORIA	1	707
213	TRAYECTORIA	1	708
214	MAYORIA	1	709
215	RECORRIA	1	710
216	GARANTIA	1	711
217	COMPARTIA	1	712
218	AMNISTIA	1	713
219	CONSTITUIA	1	714
220	TODAVIA	1	715
221	BAJA	2	717
222	VENTAJA	2	719
223	ROJA	3	722
224	SURJA	1	723
225	LA	488	1211
226	INTERCALA	1	1212
227	ALLA	1	1213
228	BATALLA	1	1214
229	AQUELLA	2	1216
230	SENCILLA	1	1217
231	REJILLA	1	1218
232	ESPAÑOLA	3	1221
233	DIFERENCIARLA	1	1222
234	CONTARLA	1	1223
235	RESUMIRLA	1	1224
236	CONVERTIRLA	1	1225
237	ISLA	1	1226
238	VINCULA	1	1227
239	LLAMA	2	1229
240	PROBLEMA	5	1234
241	TEMA	1	1235
242	SISTEMA	6	1241
243	ESQUEMA	5	1246
244	ULTIMA	3	1249
245	MAXIMA	1	1250
246	PROXIMA	2	1252
247	CALMA	1	1253
248	AFIRMA	1	1254
249	MISMA	3	1257
250	CUBANA	3	1260

Рис. 7. Фрагмент обратного частотного словаря подъязыка испанской публицистики.

1) текст сегментируется на отдельные словоупотребления; каждое словоупотребление включается в машинный документ, состоящий из $(m+1)$ ячейки; в первых m ячейках располагается код слова, в ячейке $(m+1)$ — единица счета;

2) инверсируется код, расположенный в m ячейках каждого МД; иными словами, код слова преобразуется так, чтобы начало слова стало его концом, а конец — началом;

3) массивы МД, полученные после инверсирования каждого кода словоупотребления порции, записываются на МЛ;

4) массивы упорядочиваются на МЛ по алфавиту, т. е. по содержимому первых m ячеек МД; на первое место при этом ставится МД, содержащий наименьшее содержимое указанных m ячеек, на второе — бóльшее и т. д.;

5) упорядоченный массив сжимается;

6) расшифровывается содержимое машинных документов, полученных на предыдущих этапах; при выводе на устройство печати каждого МД инверсируется содержимое первых m ячеек (т. е. делается так, чтобы словоформа, как и ранее, читалась слева направо), указывается порядковый номер словоформы, содержимое $(m+1)$-й ячейки и суммарное значение содержимых $(m+1)$-ых ячеек всех выведенных ранее МД и выводимого МД.

8.3. Принципиальный алгоритм решения задачи

Принципиальный алгоритм получения обратного частотного словаря состоит из последовательности алгоритмов получения обычного частотного словаря (рис. 17, см. Приложение), исключая алгоритм Е.

Алгоритм «Инверсия словоупотребления», используемый на вводе текста и выводе обратного словаря, дается на рис. 22—23 (см. Приложение).

8.4. Программирование алгоритма. Решение задачи на ЭВМ

Алгоритмы построения обратного частотного словаря были запрограммированы для машины «Минск-22».

По этой программе были обработаны тексты латышского языка общим объемом в 200 тыс. словоупотреблений и небольшая часть текстов казахского языка.[85]

На рис. 7 приводится образец списков, выдаваемых машиной «Минск-22» по разработанным выше программам.

[85] К. Б. Бектаев, А. Джубанов, А. В. Зубов, ук. соч.

§ 9. Выделение наиболее частых n-словных сочетаний

9.1. Лингвистические аспекты задачи

Как мы уже отмечали, вероятностно-статистический отбор одной лишь лексики еще не решает вопроса о построении базового языка. В нем должны быть также учтены наиболее характерные для данного подъязыка особенности в сочетаемости (свободной и устойчивой) словоформ, а также многозначность и омонимия его единиц. Для получения таких данных строятся частотные списки различных n-словных сочетаний (именных, глагольных, атрибутивных и т. п.). Из этих словосочетаний извлекаются затем так называемые типовые контексты (т. е. наиболее типичные для данного языка грамматические и лексические схемы контактных дистрибуций), с помощью которых и решаются вышеназванные вопросы.[86]

Помимо этого данные о сочетаемости слов в текстах необходимы и при решении задач автоматического информационного поиска, автоматического реферирования, в задачах теории передачи сообщений, в построении научной методики обучения языкам, при составлении терминологических словарей и т. д.

Достаточно обширная и достоверная статистика n-словных сочетаний до сих пор не могла быть получена «вручную», поскольку это требует огромных усилий большого числа исследователей. Практически эта задача оказывается невыполнимой при использовании ручной выборки.

Применение ЭВМ в лингвистических исследованиях позволяет использовать их и в решении задач выделения наиболее частых n-словных сочетаний из больших массивов текста.[87]

По существу задача определения количественных оценок связей между словоупотреблениями в тексте сводится к изучению дистрибуции тех или иных элементов. Под дистрибуцией элемента понимается совокупность всех окружений, в которых этот элемент встречается. Иными словами, речь идет о выделении всех возможных позиций или употреблений данного языкового элемента (словоформы или слова) по отношению к употреблениям других элементов.[88]

Как отмечает Ю. С. Степанов, дистрибуция может быть мелкая, когда рассматривается окружение на глубину одного эле-

[86] См., например: И. К. Бельская. О принципах построения словаря для МП. ВЯ, 1959, № 3; А. И. Чапля. Опыт статистического описания лексической комбинаторики. КД, Л., 1967, стр. 102—153; А. Н. Шаранда. Статистическое выделение типовых контекстов и автоматический перевод. КД, Минск, 1968.

[87] Ср.: А. В. Зубов, К. Ф. Лукьяненков, Р. Г. Пиотровский, Э. Н. Хотяшов. Статистическое описание текста с помощью ЭВМ. СР.

[88] Z. S. Harris. Structural Linguistics. Chicago, 1961, pp. 15—16.

мента, и глубокая, когда учитываются несколько рядом стоящих элементов.[89]

В приводимых ниже алгоритмах рассматриваются дистрибутивные группы обоих типов. Каждый из типов мы условно делили на два класса:

1) сплошные дистрибутивные сочетания;
2) дистрибутивные сочетания с заранее заданным ядром (опорным словом).[90]

Назовем n-словным дистрибутивным сочетанием с ядром линейный отрезок текста, содержащий n словоупотреблений и состоящий из некоторого опорного слова, стоящего в сочетании на месте i, и его одностороннего, правого или левого, или двухстороннего окружения. Слова n-словного дистрибутивного сочетания условимся обозначать слева направо цифрами 1, 2, . . ., n. Тогда двухсловное сочетание запишется так: 1, 2, трехсловное: 1, 2, 3, шестисловное: 1, 2, 3, 4, 5, 6 и т. д.

В роли опорных слов обычно берется группа слов (словоформ) из уже составленного частотного словаря. Это либо самые частые слова (словоформы) такого словаря (начальный отрезок его в 100—150 единиц), либо самые частые существительные, глаголы, прилагательные, наконец служебные слова исследуемого подъязыка. В принципе в качестве опорного можно взять любое слово или группу слов, дистрибутивные сочетания которых желает изучить исследователь.

Опорное слово может занимать в n-словном дистрибутивном сочетании любое место, т. е. i может быть равно

$$i = 1, 2, 3, \ldots, n.$$

Словосочетание в таком случае называется так: n-словное сочетание с опорным словом i. Опорное слово обозначим номером с чертой над ним. Черта сверху цифры указывает на позицию опорного слова в условном обозначении n-словного сочетания.

Сплошное дистрибутивное сочетание при этом можно рассматривать как частный случай сочетания с опорным словом, при условии, что i=0. Различие возможных видов n-словных дистрибутивных сочетаний покажем на следующем примере.

Рассмотрим предложение: *Применение термина «модель» в лингвистике повторяет в основном его употребление в математике.* Пусть опорным словом является предлог *в*. Тогда в предложении можно выделить следующие двухсловные дистрибутивные сочетания:

1) *«модель» в*
2) *в лингвистике*
3) *повторяет в*
4) *в основном*
5) *употребление в*
6) *в математике*

В случаях 1, 3, 5 имеем сочетание типа (1, $\bar{2}$), а в остальных случаях — ($\bar{1}$, 2).

Аналогично выделяем трехсловные дистрибутивные сочетания:

7) *термина «модель» в*
8) *«модель» в лингвистике*
9) *в лингвистике повторяет*
10) *лингвистике повторяет в*
11) *повторяет в основном*
12) *в основном его*
13) *его употребление в*
14) *употребление в математике*
15) *в математике* △[91]

В случаях 7, 10, 13 имеем тип (1, 2, $\bar{3}$), в случаях 8, 11, 14 — тип (1, $\bar{2}$, 3) и, наконец, в случаях 9, 12, 15 — тип ($\bar{1}$, 2, 3).

Таким же образом можно строить четырех-, пятисловные и n-словные дистрибутивные сочетания с опорным словом *в*.

Сплошными двухсловными и трехсловными дистрибутивными сочетаниями (соответственно — (1, 2) и (1, 2, 3)) приведенного выше предложения будут:

1) △ *применение*
2) *применение термина*
3) *термина «модель»*
4) *«модель» в*
5) *в лингвистике*
6) *лингвистике повторяет*
и т. д.

1) △ *применение термина*
2) *применение термина «модель»*
3) *термина «модель» в*
4) *«модель» в лингвистике*
5) *в лингвистике повторяет*
6) *лингвистике повторяет в*
и т. д.

Задача выделения наиболее частых дистрибутивных словосочетаний состоит из следующей последовательности этапов:

а) из текста, содержащего N_t словоупотреблений, выбираются все n-словные дистрибутивные сочетания, опирающиеся на отдельное слово (словоформу) определенного класса, положение которого в сочетании фиксировано (i-Const);

б) для всех разных n-словных сочетаний подсчитывается частота употребления их в заданном тексте;

в) полученный массив n-словных сочетаний располагается в частотно-алфавитном порядке.

[89] Ю. С. Степанов. Основы языкознания. М., 1966, стр. 43.
[90] Ниже в данном параграфе под «словом» понимается словоформа.
[91] Знак △ здесь и в дальнейшем означает пробел между словами, а также начало и конец отдельного предложения.

9.2. Математическая постановка задачи

При чтении текста каждое словоупотребление записывается машиной в m рабочих ячеек (см. выше, стр. 321). И тогда для n-словного дистрибутивного сочетания требуется n групп по m ячеек. Будем обозначать их слева направо: $m_1, m_2, m_3, \ldots, m_i, \ldots, m_n$. Например, положив $m=4$ и $n=3$ и считая, что в каждой ячейке размещается не более 6 букв, сочетание под номером 11 (стр. 331) в машине запишется так:

п	о	в	т	о	р
я	е	т			

в					

о	с	н	о	в	и
о	м				

Как видно из приведенного чертежа, такое размещение n-словного дистрибутивного сочетания неэкономично. Бо́льшая часть МОЗУ при этом остается неиспользованной, и такая запись каждого дистрибутивного сочетания во много раз удлинит время обработки первичных массивов и уменьшит их объем.

Возможно и другое решение этого вопроса.

Назовем n_i-единицей текста n-словное дистрибутивное сочетание текста (i=0, 1, 2, . . ., n), представленное непрерывным образом. Например, для приведенного выше трехсловного дистрибутивного сочетания с опорным словом, стоящим на втором месте, соответствующая n_i-единица будет иметь вид 3_2-единицы и содержимое ее будет записано следующим образом:

п	о	в	т	о	р
я	е	т	△	в	△
о	с	н	о	в	н
о	м				

m_s

В этом случае общее количество ячеек m_s, занимаемое n_i-единицей, значительно сократится. Кроме того, при таком подходе к задаче процесс ее решения можно свести к уже рассмотренной задаче выделения наиболее частых словоупотреблений текста.

Допустим, что для исследования задана группа, состоящая из l опорных слов. Тогда математическая последовательность операций сводится к следующему:

1) каждое опорное слово из группы в l слов преобразуется так, чтобы оно было записано в определенное число ячеек m (одно и то же для всех слов); все преобразованные опорные слова размещаются в зоне ЯС;

2) выделяются очередные n словоупотреблений текста, каждое из них преобразуется так, чтобы оно записывалось в m ячейках МОЗУ; полученные n групп размещаются по m ячеек в зоне ДС оперативного накопителя;

3) последовательным сравнением содержимого группы ячеек m_i (i заранее задано условием задачи) из зоны ДС со всеми группами по m ячеек из зоны ЯС отыскивается совпадающая пара групп по m ячеек;

4) если такая пара найдена, то содержимое зоны ДС преобразуется в n_i единицу;

5) если такой пары нет, то в зоне ДС сдвигается содержимое всех групп по m ячеек на m ячеек влево, т. е. содержимое зоны ДС перемещается так, что содержимое группы m_2 оказывается в группе m_1, содержимое группы m_3 — в группе m_2 и т. д.; содержимое группы m_n замещается кодом очередного прочитанного словоупотребления текста;

6) каждая вновь полученная n_i-единица включается в машинный документ; в первые m_s ячеек этого МД помещается n_i-единица, в ячейку (m_s+1) — единица счета;

7) массивы МД, получаемые после обработки каждой порции текста, записываются на МЛ;

8) массивы МД упорядочиваются на МЛ по содержимому первых m_s ячеек МД; на первое место при этом становится МД с наименьшим содержимым этих ячеек, на второе — с бо́льшим и т. д.;

9) упорядоченный по алфавиту массив сжимается;

10) сжатый массив упорядочивается по содержимому (m_s+1)-х ячеек МД; на первое место ставится МД с наибольшим содержимым (m_s+1)-й ячейки, на второе — с меньшим и т. д.; для группы МД, имеющих одинаковое содержимое (m_s+1)-х ячеек, машинные документы в группе упорядочиваются по содержимому первых m_s ячеек; при этом на первое место ставится документ с наименьшим содержимым первых m_s ячеек, на второе — с бо́льшим и т. д.;

11) содержимое полученных МД расшифровывается путем вывода их на устройство печати; при выводе каждого документа указывается его порядковый номер, содержимое первых m_s ячеек, содержимое (m_s+1)-й ячейки и суммарное содержимое (m_s+1)-х ячеек всех выведенных ранее МД и выводимого машинного документа.

9.3. Разработка метода решения. Принципиальный алгоритм задачи

Впервые работы по выборке n-словных дистрибутивных сочетаний начали проводиться в группе «Статистика речи»; поэтому при разработке алгоритма решения этой задачи мы пользовались методом, выведенным сотрудниками этой группы.[92]

Принципиальный алгоритм выделения n-словных сочетаний представлен на рис. 24—25 (см. Приложение).

9.4. Программирование алгоритма. Решение задачи на машине

Как и все предыдущие, рассмотренный алгоритм был запрограммирован для машины «Минск-22». Построенная программа позволяет проводить следующие дистрибутивные выборки:

1) сплошных двухсловных сочетаний (1, 2);
2) двухсловных сочетаний по левому слову ($\bar{1}$, 2);
3) двухсловных сочетаний по правому слову (1, $\bar{2}$);
4) сплошных трехсловных сочетаний (1, 2, 3);
5) трехсловных сочетаний по левому слову ($\bar{1}$, 2, 3);
6) трехсловных сочетаний по правому слову (1, 2, $\bar{3}$);
7) трехсловных сочетаний по среднему слову (1, $\bar{2}$, 3);
8) четырехсловных сочетаний по второму слову слева (1, $\bar{2}$, 3, 4);
9) сплошных четырехсловных сочетаний (1, 2, 3, 4);
10) четырехсловных сочетаний по третьему слову слева (1, 2, $\bar{3}$, 4).

При изменении с пульта четырех команд можно также выбирать:

11) четырехсловные сочетания по левому крайнему слову ($\bar{1}$, 2, 3, 4);
12) четырехсловные сочетания по правому крайнему слову (1, 2, 3, $\bar{4}$).

Разработанные программы проверялись на текстах различных языков и различных объемов. В частности, по этим программам были получены частотные списки трехсловных сочетаний из английских текстов по радиоэлектронике объемом в 100 тыс. словоупотреблений, частотные списки трехсловных сочетаний из испанского текста также в 100 тыс. словоупотреблений для Ленин-

[92] См., например: Л. Е. Машкина. О статистических методах исследования лексико-грамматической дистрибуции. АКД, Минск, 1968; О. А. Нехай. Статистика и автоматический анализ текста. АКД; А. Н. Чапля, ук. соч.

101	DE SU	3	961
102	DEFENSA DE	3	964
103	DEL MOVIMIENTO	3	967
104	DEL PRESIDENTE	3	970
105	DEPARTAMENTO DE	3	973
106	DESPACHOS DE	3	976
107	DOCTOR W	3	979
108	EL COMUNICADO S	3	982
109	EL EXILIO	3	985
110	EL PRIMERO	3	988
111	EMBARGO-O	3	991
112	EN QUE-A	3	994
113	ENERO .	3	997
114	ENTRE W	3	1000
115	ES LA	3	1003
116	ESTA-D W	3	1006
117	GOBIERNO	3	1009
118	GOBIERNO [illegible]	3	1012
119	HOY DE	3	1015
120	LA FUERZA	3	1018
121	LA PRENSA	3	1021
122	LA SEMANA	3	1024
123	LA SITUACION	3	1027
124	LA SUSPENSION	3	1030
125	LA VISITA-S	3	1033
126	LOCAL-S DE	3	1036
127	LOS ASUNTOS	3	1039
128	LOS DOS	3	1042
129	LOS PARTIDOS	3	1045
130	MOVIMIENTO X	3	1048
131	MUY MALO	3	1051
132	NUCLEARES .	3	1054
133	PARA EL	3	1057
134	PARTE DE	3	1060
135	PRIMER MINI[illegible]O	3	1063
136	PRUEBAS NUC[illegible]RES	3	1066
137	QUE LAS	3	1069
138	QUE-A NO	3	1072
139	REUNION DE	3	1075
140	SINGO EMBARGO-O	3	1078
141	SUSPENSION DE	3	1081
142	UNA REUNION	3	1084
143	UNA SUSPENSION	3	1087
144	UNOS X	3	1090
145	W DE	3	1093
146	W NO	3	1096
147	W PARA	3	1099
148	W W	2	1102
149	. AL	2	1104
150	. ASI	2	1106

Рис. 8. Фрагмент частотного списка сплошных двухсловных сочетаний подъязыка испанской публицистики.

51	СВЯЗИ-И ПРИ ПОМОЩИ-П	2	129
52	СИСТЕМ В ГЕНЕРАТОРАХ	2	131
53	СИСТЕМА КОАКСИАЛЬНЫХ ФИДЕРОВ	2	133
54	ЧАСТОТ . В	2	135
55	ЧАСТОТАХ А	2	137
56	ЭЛЕКТРОННЫХ ПРИБОРОВ .	2	139
57	. А В-СЛУЧАЕ	1	140
58	. АНАЛОГИЧНЫЕ ФОРМЫ-И	1	141
59	. БОЛЬШОГО СРОКА	1	142
60	. БОЛЬШОЙ РАЗБРОС	1	143
61	. В ГЕНЕРАТОРАХ	1	144
62	. В НАСТОЯЩЕЙ-П	1	145
63	. В ПЕРВОЙ-П	1	146
64	. В ПЕРВОМ	1	147
65	. В ПОСЛЕДНЕМ	1	148
66	. В РАДИОПЕРЕДАТЧИКАХ	1	149
67	. В ТЕХНИКЕ	1	150
68	. В ТО	1	151
69	. ВАЖНЕЙШИМИ ИЗ	1	152
70	. ВАРИАНТЫ ПОДОБНЫХ	1	153
71	. ВВОДИМЫЕ В	1	154
72	. ВВОДИМЫХ В	1	155
73	. ВЕЛИЧИНА ЭТОГО	1	156
74	. ВЕСЬМА БОЛЬШАЯ	1	157
75	. ВИБРАТОР ЯВЛЯЕТСЯ	1	158
76	. ВИДОИЗМЕНЕНИЕМ ТОГО	1	159
77	. ВЛИЯНИЕ ВРЕМЕНИ	1	160
78	. ВЛИЯНИЕ ПОСЛЕДНИХ	1	161
79	. ВНУТРИ РАСПОЛОЖЕНА	1	162
80	. ВО ВНУТРЕННЕЙ-П	1	163
81	. ВОЗБУЖДАЮЩИХСЯ В	1	164
82	. ВОПРОС О	1	165
83	. ВТОРИЧНОЕ МАГНИТНОЕ	1	166
84	. ВТОРОЕ КОГДА	1	167
85	. ВТОРОЙ-П ЧАСТИ-П	1	168
86	. ГДЕ ДЛИНА	1	169
87	. ГДЕ ИМЕЕТСЯ	1	170
88	. ГРАНИЦА МЕЖДУ	1	171
89	. ДВА-ПОСЛЕДНИХ	1	172
90	. ДЛЯ ГЕНЕРАТОРОВ	1	173
91	. ДЛЯ ИХ	1	174
92	. ДЛЯ МАЛОМОЩНЫХ	1	175
93	. ДЛЯ НАСТРОЙКИ	1	176
94	. ДЛЯ ОТБОРА	1	177
95	. ДЛЯ УСТРАНЕНИЯ	1	178
96	. ДОЛЖНО БЫТЬ	1	179
97	. ЕГО РЕЗОНАНСНАЯ	1	180
98	. ЕСЛИ ВСЕ	1	181
99	. ЕСЛИ ДАВЛЕНИЕ	1	182
100	. ЕСЛИ ДИСК	1	183

Рис. 9. Фрагмент частотного списка сплошных трехсловных сочетаний подъязыка русской электроники.

1	OF THE ENGINE .	7	7
2	SHOWN IN FIGURE X	7	14
3	THE MAIN STEAM PIPING	4	18
4	WATER AND OIL PUMPS	4	22
5	. IT WILL BE	3	25
6	AND OIL PUMPS .	3	28
7	THE KINETIC ENERGY IS	3	31
8	. EACH CYLINDER UNIT	2	33
9	. FOUR-CYLINDER ENGINE WITH	2	35
10	. FOUR-CYLINDER ENGINES WITH	2	37
11	. FROM STEAM TABLES	2	39
12	. MAIN STEAM PIPING	2	41
13	. SCAVENGING AIR IS	2	43
14	. SIX-CYLINDER ENGINES WITH	2	45
15	. THE COOLING WATER	2	47
16	. THE ENGINE IS	2	49
17	. THE FUEL IS	2	51
18	. THE OIL AT	2	53
19	AND A VELOCITY X	2	55
20	AT A PRESSURE OF	2	57
21	AT RATED POWER .	2	59
22	AT THE TOP OF	2	61
23	DEPENDS ON OIL PRESSURE	2	63
24	EACH CYLINDER UNIT HAS	2	65
25	FROM THE POINT OF	2	67
26	IN MAIN STEAM PIPING	2	69
27	IN THE FIGURE .	2	71
28	IN THE ORDER OF	2	73
29	JACKET AND PISTON COOLING	2	75
30	OF STEAM FLOW .	2	77
31	OF STEAM FLOW BASIS	2	79
32	OF THE CYLINDER BARREL	2	81
33	OF THE MAIN STEAM	2	83
34	OF THE STEAM X	2	85
35	OF THE TURBINE MAY	2	87
36	POUND OF STEAM FLOW	2	89
37	THE AFT END OF	2	91
38	THE BLADE SPEED IS	2	93
39	THE COOLING WATER IS	2	95
40	THE FORWARD END OF	2	97
41	THE KINETIC ENERGY OF	2	99
42	THE LAST STAGE OF	2	101
43	THE LOWER PISTON .	2	103
44	TOP AND BOTTOM COVERS	2	105
45	TWIN-SCREW CARGO-LINER INSTALLATION.	2	107
46	WHEN THE BLADE SPEED	2	109
47	Z WITH CYLINDERS X	2	111
48	. A TEMPERATURE OF	1	112
49	. ALL ENGINES BUILT	1	113
50	. ALL ENGINES HAVE	1	114

Рис. 10. Фрагмент частотного списка четырехсловных сочетаний с опорным существительным на третьем месте слева (подъязык английских судовых механизмов).

градского ГПИ им. А. И. Герцена и небольшие списки из текстов трех дагестанских языков.[93]

Образцы частотных списков, получаемых по рассмотренным программам, приведены на рис. 8, 9 и 10.

§ 10. Объединение первичных частотных списков

После того как получено k первичных частотных списков, каждый из текста объемом в N_i словоупотреблений, возникает задача объединения этих списков (необходимость такого объединения определена выше, см. стр. 309—310).

Для решения этой задачи можно предложить три подхода:

1) объединение в ЭВМ невыведенных результатов;

2) объединение с реперфорацией первичных частотно-алфавитных списков;

3) комбинированное объединение.

10.1. Объединение в ЭВМ невыведенных результатов

Сущность этого приема заключается в том, что объединяются массивы, находящиеся на магнитных лентах. При этом используются не конечные частотно-алфавитные списки, а алфавитно-частотные списки, являющиеся промежуточным результатом основной задачи (алгоритмы С и Д, рис. 17; алгоритмы В и С, рис. 24; см. Приложение).

Вначале объединяются первичные алфавитно-частотные списки с двух лент, результат объединяется со списком на третьей МЛ и т. д. После объединения всех предыдущих МЛ и последней результат будет представлять собой алфавитно-частотный список всего массива в $N=\sum_{i=1}^{k} N_i$ словоупотреблений. Этот массив сортируется по частоте (алгоритм Е, рис. 17; алгоритм Д, рис. 24; см. Приложение) и выводится на устройство печати. Вывод осуществляется в соответствии с алгоритмами F (рис. 20) и АВ' для вывода обратного словаря (рис. 23, см. Приложение). В алгоритме программы, объединяющей два алфавитно-частотных списка, находящихся на магнитных лентах № 1 и № 2, можно выделить следующие основные подалгоритмы (рис. 26, см. Приложение).

Основным недостатком рассматриваемого метода объединения частотных списков буквосочетаний, словоформ и словосочетаний является то, что при объединении первичных массивов N_i в общий список вносятся все те ошибки, которые имелись в каждом частотном списке. Как показала практика, бо́льшая часть этих ошибок возникает в процессе кодирования текста. Это или пропуски букв в слове, или повторение одной и той же буквы, или, наконец, отсутствие пробела между двумя соседними словоупотреблениями текста. Такие единицы текста, содержащие ошибки (как правило, с частотой 1), оказываются в окончательном частотном списке и тем самым меняют истинное значение частот для некоторых единиц.

Для исправления конечного частотного списка, составленного по этому методу, человеку приходится «опознавать» искаженные единицы, вычеркивать их из списка и за счет этого увеличивать частоты соответствующих неискаженных единиц.

Рассмотренный алгоритм был запрограммирован для ЭВМ «Минск-22» и проверялся при объединении двух массивов по $N_i=50\,000$ словоупотреблений каждый для французских и испанских текстов.

10.2. Объединение с реперфорацией первичных частотно-алфавитных списков

По этому методу каждый из полученных k частотных списков после вывода на печать корректируется и заново кодируется на перфоленты или перфокарты. При этом, как показывает опыт, из текста объемом в $N_i=50\,000$ словоупотреблений получается частотный словарь на 4000—8000 словоформ (в зависимости от языка и подъязыка) и частотный список на 500—30 000 трехсловных сочетаний (в зависимости от количества опорных слов и их частоты). Перфорацию каждой текстовой единицы при этом осуществляют вместе с ее частотой употребления во взятом тексте в такой последовательности: признак цифр — частота — пробел — признак алфавита — единица текста — пробел — частота и т. д.

Полученные перфоленты (или перфокарты) снова вводят в машину. Обработка этих массивов повторяет процесс получения первичного списка из текста в N_i единиц. Все алгоритмы при этом аналогичны алгоритмам, приведенным на принципиальной схеме (рис. 17). Исключение составляет алгоритм A — «Ввод текстов и их предварительная обработка с записью на МЛ». В нем совершенно меняется подалгоритм АС (рис. 27, см. Приложение). По существу метод с реперфорацией — это объединение первичных списков в системе «машина—человек», где эвристические функции исправления ошибок переданы человеку, а трудоемкое задание пересчета и классификации исследуемых текстовых единиц — машине. По методу с реперфорацией был получен обратный частотный словарь латышского языка общим объемом в 200 тыс. словоупотреблений. Программа алгоритма А для вторичного ввода текста в виде «единица текста — частота ее» приводится на рис. 27 (см. Приложение).

[93] А. В. Зубов, А. И. Чапля. Опыт одновременного получения частотных списков. . . ; В. А. Ноздрина. Получение частотного списка с опорным прилагательным на базе текстов по радиоэлектронике. В сб.: VII научная студенческая конференция, посвященная 150-летию со дня рождения К. Маркса, 50-летию БССР и 50-летию ВЛКСМ, Минск, 1968.

Чтобы избежать недостатков двух рассмотренных методов объединения, можно применить третий, комбинированный метод объединения. При этом весь массив из первичных списков делится на две части

$$N_1 = \sum_{i=1}^{k/2} N_i \text{ и } N_2 = \sum_{i=k/2+1}^{k} N_i.$$

Каждый из массивов, N_1 и N_2, образуется по методу объединения в ЭВМ невыведенных результатов. Затем в том и другом массивах выявляются все ошибки, и эти исправленные списки N_1 и N_2 перфорируются заново. Их объединение осуществляется по методу с реперфорацией.

Как мы отмечали выше, при использовании метода с реперфорацией первичных частотных списков приходится вновь перфорировать довольно значительные порции информации (при больших N) и снова вводить их в машину для обработки. Но, как известно, и сам процесс кодирования, и ввод текста занимают значительную часть общего времени получения частотного списка (ср., например, стр. 321). Комбинированный метод позволяет свести к минимуму вновь перфорируемый материал и уменьшить число вводимых порций с $k\left(N = \sum_{i=1}^{k} N_i\right)$ до двух.

Если в процессе получения частотного списка из текста объемом в N словоупотреблений не требуется вывода на печать всех промежуточных частотно-алфавитных списков N_i, то при использовании метода объединения в ЭВМ невыведенных результатов можно не производить сортировку по частоте для некоторых (или всех) первичных списков N_i. Это несколько сократит общее время обработки списка из текста в N словоупотреблений.

Глава II. МОДЕЛИ ОПРЕДЕЛЕНИЯ ЗНАЧЕНИЯ ИНОСТРАННОГО СЛОВА ПРИ ПЕРЕВОДЕ

§ 11. Моделирование на ЭВМ интеллектуальной деятельности мозга

В проблеме создания моделей, имитирующих некоторые функции интеллектуальной деятельности мозга, следует различать три уровня:

1) принципиальная возможность такого моделирования;
2) возможность технической реализации модели;
3) практическая целесообразность моделирования.

Рассмотрим подробнее каждый из этих уровней.

1. Принципиальная возможность моделирования деятельности мозга. На первом уровне рассматриваемая проблема решается, исключительно исходя из имеющихся сегодня знаний фундаментальных законов природы, безотносительно к каким бы то ни было конкретным техническим воплощениям. Если познанные человеком законы природы не противоречат осуществлению того или иного проекта, необходимо признать его принципиально реализуемым.

Процессы человеческого мышления могут быть объяснены в терминах происходящих в мозгу электрических и химических процессов, в терминах нейронной организации мозга или в терминах процессов переработки информации. В следующих двух главах работы нас будет интересовать информационная сторона процессов человеческого мышления при переводе отдельных единиц и простейших совокупностей единиц письменной речи с иностранного языка на русский. Исходной предпосылкой такого подхода является то, что сложные процессы мышления строятся из элементарных процессов оперирования с символами (элементарных информационных процессов). Но все ли действия, выполняемые мозгом, можно представить в виде конечной последовательности операций? Человек, безусловно, является чрезвычайно сложной, но все же конечной системой ограниченного совершенства, что делает принципиально возможным ее моделирование и объективное изучение методами и средствами точного знания. И это естественно, ибо любая деятельность человека подчиняется только объективным законам природы. Как отмечает акад. А. Н. Колмогоров, принципы взаимодействия компонентов человеческой машины не могут быть принципиально невоспроизводимыми, иначе это означало бы, что ими управляет «жизненная сила», vis vitalis.[94]

Следовательно, можно допустить, что любая функция мышления может быть представлена в виде формальной непротиворечивой системы.[95] Тогда, в соответствии с известной теоремой существования Мак-Каллока—Питтса, [96] такая функция может быть реализована формальной нервной сетью, а значит воспроизведена машиной. «Таким образом, — пишет В. М. Глушков, — в настоящее время факт принципиальной возможности программирования на современных электронных цифровых машинах любых информа-

[94] См.: А. Н. Колмогоров. Жизнь и мышление как особые формы существования материи. В сб.: О сущности жизни, М., 1964, стр. 5.

[95] Известный американский ученый Джон фон Нейман в связи с этим заметил, что отрицать возможность логического описания любой функции нервной системы с помощью конечного числа слов — значит примкнуть к разновидностям логического мистицизма.

[96] У. Мак-Каллок, В. П. Питтс. Логическое исчисление идей, относящихся к нервной активности. Русск. пер. в сб.: Автоматы, М., 1956.

ционных моделей установлен не менее твердо, чем факт возможности разложения любого материального объекта на элементарные частицы. Важно еще раз подчеркнуть, что речь идет здесь именно о моделях любой (а не только математической) природы».[97] Другое дело, невозможно окончательно, «без остатка» смоделировать все функции ВНД (как единого целого), которыми она располагает в момент моделирования. Такое моделирование будет бесконечным, асимптотическим процессом. Можно все ближе и ближе подходить к «идеальной» модели мозга, но никогда ее не достигнуть, потому что, передав подобной модели какие-то одни из своих функций, как бы сложны они ни были, мозг обязательно воспользуется этим и сформулирует другие, еще более сложные функции.[98]

2. Возможность технической реализации моделей деятельности мозга. Еще 10—15 лет назад такой возможности не было. Электронно-вычислительные машины только начинали свою историю. Они имели очень ограниченный набор команд, малые объемы «памяти» и небольшое быстродействие. Но уже сейчас положение коренным образом переменилось. Существующие ЭВМ обладают достаточно гибкой логикой, включающей сотни команд, их «память» вмещает сотни миллионов машинных слов, и быстродействие возросло до нескольких миллионов операций в секунду.

Поэтому практически есть все условия для построения на ЭВМ самых сложных моделей функционирования мозга.

3. Практическая целесообразность моделирования деятельности мозга. Знание алгоритмов работы мозга необходимо как для многих теоретических изысканий, так и для решения ряда практических задач. Особенно резко встал этот вопрос сейчас в связи с некоторыми социальными процессами. Моделирование мыслительных функций мозга иногда приводит к поразительным практическим результатам. Так, например, функциональная модель работы мозга, реализуемая электронно-вычислительной машиной и дополненная сведениями, почерпнутыми из энцефалограмм, очень хорошо имитирует процессы умственных расстройств больного и позволяет не только ставить диагнозы, но и предсказывать будущий ход течения болезни.[99]

Как же создаются функциональные модели интеллектуальной деятельности мозга? В настоящее время еще не существует разработанных строгих процедур построения таких моделей. Общие правила при этом сводятся к следующему.

[97] В. М. Глушков. Гнесеологическая природа информационного моделирования. ВФ, 1963, № 10.

[98] А. А. Леонтьев. Языкознание и психология. М., 1966, стр. 27.

[99] К. Е. Морозов. Математические модели в кибернетике. М., 1968, стр. 10.

Исследователь начинает свою работу с того, что определяет область человеческого поведения, которая его интересует, например «самообучение», «перевод слова [100] с одного языка на другой», «решение геометрических задач». Затем он обычно сосредоточивает внимание на поведении в конкретной ситуации, например при распознавании значения иностранного слова. Для того чтобы создать предварительную модель поведения, исследователь должен иметь представление или ряд представлений о поведении, проявляющемся в данной задаче. Такое первоначальное представление может явиться плодом наблюдения (в нашем случае лингвистического) за поведением людей, результатом опроса людей о том, что они делают в процессе решения задачи, или же результатом размышления о типе устройства, которое потребовалось бы для решения данной задачи.

Следующим шагом является построение модели, т. е. составление программы для вычислительной машины, реализующей эти представления. В ходе этой работы исследователь сталкивается с недостаточностью своих первоначальных представлений. Он обнаруживает, что ему необходима дополнительная информация. Чтобы ее получить, он может заново проанализировать прежние данные или провести новые эксперименты. Так исследователь сталкивается с одним из преимуществ машинного моделирования — с требованием полноты и точности в описании работы мозга. После многих видоизменений и переделок первоначальной программы получается окончательная модель.

Затем исследователь ставит перед моделью ту же задачу, что и перед живым испытуемым, например задачу перевода слова (или слов) из текстов определенного подъязыка. По существу, модель можно рассматривать как «искусственный субъект», с которым повторяется эксперимент, проведенный над живым человеком. Поведение программы, т. е. ответы в различении значения иностранного многозначного слова, ответы в задаче обучения, шахматные ходы и т. п., сравнивается с поведением испытуемого (информанта). Это сравнение позволяет определить коэффициент перехода от модели к оригиналу. Таким коэффициентом является отношение числа правильных ответов модели к общему числу вопросов.

§ 12. О понимании текста человеком и машиной

Понимание текста есть первый шаг на пути к переводу.

Как же понимает текст человек и как его понимает машина?

Человеку дан текст, и он смотрит на первое слово этого текста. Каким образом он понимает значение данного слова, т. е. соотносит его с определенным объектом внешнего мира? Были проведены

[100] В первых трех параграфах этой части работы мы употребляем термин «слово» в его широком смысле, в частности вместо термина «словоформа».

многочисленные эксперименты по восприятию слов при чтении.[101] В частности, было показано, что слова при чтении воспринимаются не путем перебора отдельных букв, а как единое целое. Но остается неясным само понятие целостности слова. Предполагают, что целостное восприятие слова определяется сформированным в процессе обучения образом слова.[102]

Но что такое образ слова? Как он формируется? При решении этих вопросов выделяют два процесса: различение и узнавание.[103] Прежде всего, читая слово, человек выделяет какие-то формальные признаки его, которые отличают это слово от других слов, образы которых в виде определенного набора формальных дифференциальных признаков хранятся в его памяти. В свою очередь дифференциальные элементы слова выделяются путем их узнавания, т. е. сопоставлением этих элементов с уже имеющимися в памяти зрительно-слуховыми компонентами. Но одного различения элементов еще недостаточно для создания образа. Необходимо каким-то образом закрепить в сознании человека разрешенный порядок расположения отдельных лингвистических элементов относительно друг друга, в результате чего эти элементы начинают представлять собой некоторое новое единство. Эту задачу выполняют время и практика. Таким образом, для человека понять (узнать) слово — это значит выполнить операции, которые составляют это слово.[104]

Опознавание и осмысление человеком более крупных речевых единиц является сложным многоуровневым процессом. Чем сложнее единица, тем большее число характеристик влияет на ее восприятие. Например, при восприятии словосочетаний, помимо уже указанных особенностей восприятия слов, в действие вступают синтаксические связи между составляющими их словами; при восприятии фраз начинают преобладать фразовые структуры и т. д.[105]

Прежде чем ответить на вопрос, как понимает текст машина, дадим несколько дополнительных определений.

1) Если некоторому тексту можно сопоставить некоторый возможный (отмеченный) набор объектов или явлений внешнего мира, то будем говорить, что этот текст имеет смысл.

[101] См., например: В. И. Галунов. Некоторые особенности восприятия речи человеком. В сб.: XVIII Международный психологический конгресс. Симпозиум 23. Модели восприятия речи, М., 1966; Т. Г. Егоров. Психология овладения навыком чтения. М., 1963; Н. И. Жинкин. Механизмы речи. М., 1958; А. А. Леонтьев. Теоретические проблемы психолингвистического моделирования речевой деятельности. АДД. М., 1968.

[102] Е. И. Исенин. К вопросу о формировании образа слова. ВП, 1967, № 1, стр. 51.

[103] Н. И. Жинкин, ук. соч.

[104] Silvio Ceccato and Bruna Zonta. Human Translation and Translation by Machine. In: 1961. International Conference on Machine Translation of Languages and Applied Language Analysis, London, 1962.

[105] Р. Г. Пиотровский. Моделирование фонологических систем и методы их сравнения. Л.—М., 1966, стр. 354.

2) Сопоставление осмысленному тексту соответствующего элемента внешнего мира можно рассматривать как вычисление значения некоторой семантической функции.

3) Множество осмысленных текстов есть область определения семантической функции.

4) Семантика языка — это способ вычисления семантической функции для всех осмысленных текстов данного языка (подъязыка).

5) Областью значений семантической функции является множество элементов, описываемых данным языком.

6) Если область значений семантической функции охватывает всю совокупность интересующих нас элементов внешнего мира, то семантическая функция называется универсальной.

7) Язык, обладающий универсальной семантической функцией, называется полным.[106]

Теперь допустим, что в машину вложен некоторый базовый язык, обладающий своей семантической функцией. Все мыслимые действия машины при всех возможных вариантах загрузки ее запоминающего устройства будем считать совокупностью элементов внешнего мира. Таким образом, машина и ее полный базовый язык представляет собой замкнутую подсистему в открытой системе «человек—природа» (имеется в виду «словесное» описание мира). Базовый язык, являясь общим для машины и для человека, позволяет так ставить задание машине, что она воспринимает текст как некоторую программу. При этом человек точно знает, чего он хочет от машины.

При такой постановке начальных условий понимание машиной текста определяется на строго поведенческой основе, а именно: если после ввода текста в ЭВМ действия машины соответствуют тому, что ожидает человек, то считается, что машина правильно поняла текст.[107]

Рассмотрим теперь, как машина «воспринимает» слово. В отличие от восприятия слова человеком машина не «воспринимает», а опознает слово, так как она имеет дело с кодом слова, но не с образом. Опознать слово для ЭВМ — значит найти его код в «памяти» машины. Как отмечено выше, для машины совокупность элементов внешнего мира — это все мыслимые действия машины. Значит, процесс опознания слова сводится к выполнению строго определенной последовательности операций, приводящей к искомому коду. Опознавание более крупных рече-

[106] А. П. Ершов. Об одном виде контакта человека с машиной, стр. 6—7.

[107] У. Рейтман. Познание и мышление. Моделирование на уровне информационных процессов. Русск. пер. М., 1968, стр. 316; А. П. Ершов, ук. соч., стр. 7.

вых единиц сводится к выполнению определенного комплекса последовательностей операций.

Основное отличие восприятия человеком от опознавания машиной состоит в том, что человек способен работать при недостаточности информативных признаков, содержащихся в объекте, и вместе с тем человек способен использовать огромную информацию, которая в признаках объекта непосредственно не закодирована.[108] Опознование машиной объекта возможно лишь при задании ей полного набора признаков, характеризующих тот или иной объект.

§ 13. Неоднозначность единиц базового языка и методы ее автоматического устранения

Допустим, что нам удалось выделить наиболее частые, наиболее информативные и равномерно распределенные единицы, которые должны войти в базовый язык. Как правило, наиболее частые слова неоднозначны. Это в первую очередь относится к служебным словам и некоторым глаголам, которые могут выступать как вспомогательные.[109]

Все случаи неоднозначности (равноименности) А. А. Реформатский делит на 3 группы: [110]

1) полисемию;
2) омонимию (в нашей ситуации — омографию);
3) параллельные образования слов из одного источника, где нет ни полисемии, ни совпадения различных слов.

В свою очередь каждая из этих групп имеет по нескольку разновидностей. В частности, в группе омографов выделяют:

1) грамматические омографы;
2) лексико-грамматические омографы;
3) лексические омографы;
4) омоформы — различные аффиксы, совпадающие по написанию и различающиеся по функциям.

Подходы к проблемам различения неоднозначности весьма разнообразны. При этом спорным является не только вопрос о конкретном значении слова, но и о том, к какому типу неоднозначности оно относится.[111] Все методы различения неоднозначности условно можно разделить на две группы:

[108] А. Н. Леонтьев, Д. Ю. Панов. Психология и технический прогресс. В сб.: Философские вопросы физиологии высшей нервной деятельности и психологии, М., 1963, стр. 401.

[109] См., например: P. Guiraud. Problemes et Methodes de la Statistique Linguistique. Dordrecht—Holland, 1959, pp. 30—31.

[110] А. А. Реформатский. Введение в языковедение. М., 1967, стр. 75.

[111] См.: В. И. Абаев. О подаче омонимов в словарях. ВЯ, 1957, № 3; М. Г. Арсеньева, Т. В. Строева, А. П. Хазанович. Многозначность и омонимия. Изд. ЛГУ, 1966; В. В. Виноградов. Основные типы лексических значений слова. ВЯ, 1953, № 5.

1) логико-семантические;
2) формальные.

Критерии первой группы, как правило, основаны на изучении характера соотношения слов с реальной действительностью. Они не могут обеспечить достаточно четких, объективных правил, дающих возможность построить автоматически действующие модели разграничения неоднозначности, ибо используемые при этом методы базируются лишь на одной интуиции. Поэтому для указанной цели используются формальные методы, в частности комбинаторные и вероятностные.[112]

Возможность изучения значения формальными методами вытекает из того положения, что форма и значение языковой единицы образуют единство. Задавая условия, при которых одна из этих сторон отражает другую, можно значение изучать через форму и обратно — форму через значение.

С точки зрения нашей задачи совершенно не важно, имеем ли мы дело с омографией или полисемией. Омографы и многозначные слова будем считать величинами одного порядка.

За последние годы наметилось несколько путей различения неоднозначности с помощью комбинаторных и вероятностных приемов. Рассмотрим некоторые из них.[113]

Один из методов использует деление всех наук на 9 областей, из которых каждая включает 6—8 дисциплин. Каждому частному значению неоднозначного слова приписывается индекс той дисциплины, в текстах которой оно встречается с наибольшей вероятностью (эта вероятность определяется заранее при статистическом описании соответствующих текстов). Имея пословный перевод текста, можно подсчитать частоту индексов в переведенных текстах, а значит — определить, о каких дисциплинах идет речь в этих текстах. При повторном просмотре текстов частные значения, имеющие индексы не рассматриваемых в данном тексте дисциплин, вычеркиваются.

Другой метод сводится к тому, что выделяется ряд понятийных категорий; каждому слову приписывается номер соответствующей категории; многозначным словам приписывается несколько номеров. Чтобы определить частное значение неоднозначного слова, отыскивается среднее арифметическое понятийных номеров всех однозначных слов обрабатываемой фразы. Затем из нескольких номеров неоднозначного слова выбирается тот, который наиболее близок к найденному среднему арифметическому.

[112] Ю. Д. Апресян. К вопросу о структурной лексикологии. ВЯ, 1962, № 3; А. Я. Шайкевич. Распределение слов в тексте и выделение семантических полей. Иностранные языки в высшей школе, 1963, № 6; В. И. Перебейнос. Об использовании структурных методов для разграничения значений многозначного глагола. ВЯ, 1962, № 3.

[113] И. А. Мельчук, Р. Д. Равич. Автоматический перевод. . . , стр. 237—246.

Следующий метод выбора частного значения неоднозначного слова основан на синтаксических свойствах связанных слов. При этом для каждого неоднозначного слова выделяется несколько групп словосочетаний, включающих это слово. Дистрибутивные свойства таких словосочетаний описываются с помощью определенного числа синтактико-семантических критериев (одушевленность, число, падеж и т. п.).

В работах Мичиганского университета [114] для установления нужного частного значения слова используется специальный словарь неоднозначных слов, в котором различные переводы одного слова расположены в определенном порядке, позволяющем правильно выбирать перевод, используя контекст.

Среди других необходимо отметить методы, основанные на данных морфологии (анализ окончаний слов) и синтаксиса (анализ грамматических сочетаемостей слов).

Однако основным способом определения конкретного значения слова большинством лингвистов признается изучение и описание контекстуальных связей данного слова, в которых реализуется искомое его значение. Как отмечал В. В. Виноградов, «иметь разные значения для слова чаще всего значит входить в разные виды семантически ограниченных *фразеологических связей*. Значения и оттенки значения слова большей частью обусловлены его фразовым окружением».[115] Каждое такое «фразовое окружение», или «типовой контекст», слова ассоциируется с некоторым новым семантическим вариантом этого слова. Выделение типовых контекстов осуществляется, как правило, путем статистического анализа больших объемов текста определенного подъязыка. Для этой цели могут быть использованы алгоритмы, описанные в § 9 гл. I. Вполне естественно возникает вопрос о количестве единиц во фразовом окружении, т. е. о «мощности» контекстного окружения.[116]

Ряд исследований [117] показал, что цепочка слов, состоящая из одного предшествующего и одного последующего слова, вполне достаточна для определения значения слова (ядра), стоящего между этими словами. Но в некоторых случаях приходится использовать контекст большей мощности, включающий не только контактные, но и дистантные связи. Однако понятие контекста — очень широкое. Контекстом может быть абзац, глава, книга, тексты данной специальности. Поэтому возникает задача выделения определенных типов контекстов. Мы будем различать:

[114] Там же, стр. 243.

[115] В. В. Виноградов, ук. соч., стр. 17.

[116] А. А. Холодович. Опыт теории подклассов слов. ВЯ, 1960, № 1, стр. 37.

[117] A. Caplan. An Experimental Study of Ambignity and Context. «Mechanical Translation», vol. 2, № 2, November, 1955; M. Těšitelova. Об экономии высказывания (на материале омонимии словоформ имен существительных в чешском языке). In: Prague Studies in Mathematical Linguistics, 1966; А. И. Чапля, ук. соч.

1) макроконтекст — совокупность текстов данной тематики (подъязык);

2) микроконтекст — минимальное словесное окружение некоторого ядра (с использованием контактных и дистантных связей), достаточное для определения конкретного значения ядра.

Использование формальных методов для разрешения вопросов неоднозначности лингвистических единиц предполагает построение формальных правил, которые, независимо от того, как и кем они были получены, должны привести к одним и тем же результатам. Такой подход позволит использовать ЭВМ как для получения таких правил, так и для проверки справедливости их на новом материале.

Процедура разрешения неоднозначности через контекст осуществляется в два этапа. На первом этапе работает макроконтекст. Знание макроконтекста позволяет задавать в машинном словаре лишь те значения, которые с наибольшей вероятностью будут представлены в обрабатываемых ЭВМ текстах. Многозначность знаменательных слов сокращается благодаря макроконтексту приблизительно на 60—70%, чего нельзя сказать о грамматической омографии. Что касается служебных слов, макроконтекст снимает лексическую омографию, но ничего не дает для сужения их многозначности и разрешения грамматической омографии.

На втором этапе для устранения неоднозначности слов базового языка используются микроконтексты.

Можно предложить два пути использования микроконтекстов:

1) построение специального машинного словаря, содержащего наборы микроконтекстов для каждого многозначного слова;

2) в обычном машинном словаре при каждом многозначном слове дается несколько значений в зависимости от того, с каким словом оно входит в микроконтекст.

Первый путь использования микроконтекстов практически целесообразен лишь для микроконтекстов с контактными связями слов. Наличие в общем списке микроконтекстов с дистантными связями вызовет необходимость построения специальных алгоритмов для выделения таких микроконтекстов из анализируемого текста. Еще одним недостатком такого подхода является то, что списки микроконтекстов занимают в «памяти» машины значительный объем. Поэтому для автоматического анализа больших массивов текста, при использовании существующих ЭВМ, этот метод мало пригоден. Применение его для автоматического анализа отдельных слов подробно описано в работе О. А. Нехай.[118]

При использовании второго пути в словаре задаются не наборы микроконтекстов, а слова, входящие в них, с соответствующими переводами. Одно и то же слово может входить в разные микроконтексты. При использовании уже рассмотренного метода это

[118] О. А. Нехай, ук. соч., стр. 203—205.

слово повторяется столько раз, сколько микроконтекстов его содержат. Второй путь позволяет записывать такое слово лишь однажды. Возьмем, к примеру, многозначное французское слово feu. В зависимости от слова, стоящего слева, оно может «переводиться» по-разному:

feu	'огонь', 'стрельба'
mettre feu	'поджигать'
faire feu	'выстрелить'
faux feu	'осечка'

Не задавая этих микроконтекстов, можно в словаре однажды указать слово feu и дать при нем несколько переводов. При каждом значении этого слова специальным кодом указывается, к какому слову микроконтекста относится соответствующее значение. Таким кодом может быть, например, адрес этого слова в «памяти» машины. Этот путь, при разработке соответствующих кодов, может стать эффективным средством устранения неоднозначности при автоматическом анализе текстов любых объемов.

При использовании обоих путей устранения неоднозначности строят специальные алгоритмы для каждого неоднозначного слова. Необходимо отметить, что такие алгоритмы при использовании списка микроконтекстов будут проще, чем при задании нескольких значений для одного слова.

В следующих параграфах настоящей работы будет использован упрощенный вариант второго из рассмотренных методов.

§ 14. Модель разрешения человеком лексико-грамматической омографии слова zu

14.1. Лингвистические аспекты задачи

Прежде всего необходимо отметить, что рассматриваемый в последующих параграфах автоматический анализ осуществляется не в терминах какого-либо искусственного языка, а через метаязык, в роли которого выступает естественный русский язык. В качестве языка-объекта в данной задаче используется немецкий язык. Следовательно, такие понятия, как омонимия (омография), многозначность, идиоматичность и т. п., рассматриваются здесь с точки зрения соответствующей двуязычной ситуации. В частности, омонимичными и многозначными мы считаем лишь те словоформы языка-объекта, которые имеют несколько русских эквивалентов.

Лингвистическая информация, необходимая для построения алгоритма, была получена в результате анализа микроконтекстов немецкого подъязыка радиоэлектроники.[119]

[119] А. Н. Шарадда. Статистическое выделение типовых контекстов. . . , стр. 130—136.

Из общего числа в 3335 микроконтекстов zu в функции предлога встретилось в 1110 случаях (эмпирическая вероятность $p=0.333$), в функции частицы при инфинитиве — в 1801 случае ($p=0.540$). Далее, 175 раз ($p=0.052$) zu встретилось в конструкции с причастием I, 160 раз ($p=0.048$) — в качестве усилительной частицы; в 89 зафиксированных случаях ($p=0.027$) zu выступало как отделяемая приставка.

Дальнейший статистический анализ микроконтекстов (с учетом левой и правой позиций словоформ, входящих в микроконтекст) позволил выделить наиболее частые формальные признаки указанных выше пяти функций (см. табл. 2).[120]

Рассмотрим, каким образом человек, в мозгу которого заложены все эти выделители, распознаёт грамматическую функцию zu. Например, человек рассматривает следующую конструкцию: [121]

häufigsten zu findenden Ausgangselemente.

Здесь в словоформе за zu легко уловить признак причастия I. Но что значит «уловить»? Выделяет ли человек из словоформы findenden сразу -nden или же побуквенно с конца: -n, -en, -den? Процесс эвристического поведения человека не ясен: мы не можем рассматриваемое лингвистическое поведение человека описать с помощью конечного числа правил. Попытаемся достичь результата с помощью метода проб и ошибок.

Известно, что машина не обладает эвристическими способностями. Поэтому все наши мыслительные эксперименты мы будем строить, ориентируясь на правила математической логики, лежащие в основе работы ЭВМ. Допустим, что человек на первом шаге анализа выделил две последние буквы словоформы findenden. Пусть, далее, формальный признак *-en* сравнивается с формальными выделителями из табл. 2. Очевидно, что такое сравнение приведет человека к выводу, что рассматриваемая словоформа findenden представляет собой инфинитив и zu играет роль приинфинитивной частицы. Результат не верен. Отсюда следует, что проверка на признак инфинитива должна проводиться позже, нежели проверка на признак причастия I.

Рассмотрим еще один пример. Выясним, каким образом «мощность» микроконтекста влияет на точность анализа. Действуя «как машина», мы должны ограничить возможности нашей программы. Пусть одно из правил для определения zu в функции «предлог» записывается так: «Если в конструкции первая или вторая слово-

[120] Ср.: Е. В. Гулыга, Н. Ф. Натанзон. Грамматика немецкого языка. М., 1957.

[121] Под конструкцией мы будем понимать предложение или часть его, содержащую микроконтекст.

Таблица 2

Формальные выделители для разрешения лексико-грамматической омонимии zu

Функция в предложении	Формальные выделители					
	левая дистрибуция		перевод zu	правая дистрибуция		
	слова	признаки		слова	признаки	
Частица при инфинитиве	um ohne		'чтобы' 'без того, чтобы'	sein, seinem, tun	окончание инфинитива: -en, -rn, -ln	
Предлог	— bis		'к', 'в' 'вплоть до'	dem, den, der, diesem, dessen, dieser, einem, einer, jeden, jeder, denen, deren, dessen	*признак существительного*, на 1-м или 2-м месте, формула-F	
В конструкции с причастием I			'который' 'следует'		суффиксы причастия I и окончания существительных: -ende, -nden, -lnde, -rnde, -nder, -ndes	
Отделяемая приставка				und, oder, als	знаки препинания	
Усилительная частица	формальные выделители не обнаружены; берется как исключение после проверки признаков первых четырех случаев					

форма за zu есть существительное,[122] то zu — предлог». Часто правило это срабатывает верно. Например:

№ 32 rregler zu automatischen ttrocknungs;[123]
№ 82 bis zu vverfahren mit;
№ 952 zu diesem zzweck.

Однако в некоторых случаях результаты получаются иные. Рассмотрим две следующие конструкции:

№ 40 zu sehr niedrigen iimpulsfolgefrequenzen;
№ 966 nimmt stärker zu als fflinmechanik.

Каким формальным способом можно проверить указанное выше правило? Естественно, что необходимо взять две первые буквы словоформы, непосредственно следующей за zu, и проверить их на совпадение. Если первая после zu словоформа не есть существительное, то необходимо выполнить такое же сравнение со второй. В примере № 40 в результате двух указанных сравнений существительное обнаружено не будет. Нет в нем и признаков причастия I, отделяемой приставки и инфинитива (см. табл. 2). Остается допустить, что в рассматриваемой конструкции № 40 zu выступает в роли усилительной частицы. Очевидно, что такой вывод не верен. Проверка на признак существительного 3-й словоформы после zu дала бы верный результат: «предлог».

В примере № 966 при втором сравнении на признак «предлог» получим положительный ответ. Но и он не соответствует действительности. Если бы проверка формальных признаков «zu в функции отделяемой приставки» проводилась раньше, чем проверка признаков «zu в функции предлога», то ответ был бы правильным: «отделяемая приставка» (см. табл. 2, стр. 352).

Таким образом, микроконтекст одной и той же «мощности» в одних случаях оказывается недостаточным для анализа, в других же — он избыточен.

Приведенные примеры показывают, насколько эвристическое мышление человека отличается от дискретно-формального эксплицитного «мышления» автоматов, работающих на жестких, непересекающихся связях. Отсюда ясно, насколько полной, непротиворечивой и последовательной должна быть формализация того или иного языкового явления, чтобы его можно было с достаточной точностью смоделировать на ЭВМ.

---

[122] Формальным признаком существительного во всех моделях, относящихся к немецкому языку, является удвоенная первая буква соответствующей словоформы. Такое удвоение буквы осуществляется в процессе кодирования текста. Например, немецкие существительные Informationen, Veränderungen передаются соответственно как iinformationen, vveraenderungen.

[123] Для проверки построенной модели использовался набор конструкций. Номера конструкций представлены в таком порядке, в каком их выдавала на печатающее устройство ЭВМ «Минск-22».

Лингвистическое задание для снятия неоднозначности zu можно сформулировать следующим образом.

Задан текст, состоящий из N конструкций, в каждой из которых содержится словоформа zu. Задан набор формальных выделителей в виде словоформ или формальных признаков (табл. 2). Необходимо для каждой конструкции:

1) узнать, используя формальные выделители, какую грамматическую функцию выполняет словоформа zu;

2) определить (в соответствии с табл. 2) лексическое значение словоформы zu;

3) если zu входит в конструкции с инфинитивом или причастием, то дать перевод соответствующего инфинитива.

Пункт 3 этого задания заставил нас ввести в алгоритм задачи немецко-русский словарь. Этот словарь включал инфинитивы тех глаголов, которые входили в указанные выше конструкции с zu. Для упрощения задания каждому немецкому глаголу давался один наиболее частый русский перевод.

14.2. Математическая постановка задачи

Указанная выше лингвистическая задача может быть сведена к выполнению следующей последовательности заданий:

1) каждая словоформа немецко-русского словаря приводится к стандартной форме, содержащей m ячеек оперативного накопителя; каждая пара словоформ (немецкий глагол и его перевод) включается в машинный документ $МД_1$, содержащий $2m$ ячеек; в первых m ячейках размещается немецкая словоформа, во вторых — русская; массив документов $МД_1$ располагается в зоне С оперативного накопителя (или во внешнем накопителе);

2) каждое словоупотребление конструкции преобразуется к стандартному виду путем записи его в m ячейках МОЗУ;

3) каждое очередное содержимое указанных в пункте $2m$ ячеек передается в определенную зону накопителя — зону К; концом зоны К считается место занесения в нее кода точки (точкой заканчивается каждая конструкция);

4) последовательным сравнением содержимого m ячеек зоны К с кодом словоформы zu отыскивается и запоминается место этого кода в зоне К;

5) с использованием кодов формальных выделителей и содержимого ячеек зоны К отыскивается один из пяти возможных выходов программы;

6) расшифровывается содержимое зоны К и признаков, характеризующих найденный выход.

14.3. Подготовка начальных данных для ввода в ЭВМ

Текст в виде набора конструкций перфорируется на перфоленты или перфокарты в соответствии с методикой, изложенной в § 3.

В конструкцию при этом должно входить не более восьми словоупотреблений (считая точку за словоупотребление). Таким же образом кодируется и немецко-русский словарь. При этом вначале кодируется немецкая словоформа, затем ее перевод, новая немецкая словоформа и ее перевод и т. д. Словоформы одна от другой отделяются пробелами. Никаких иных разделительных знаков в этом массиве нет.

Формальные выделители записываются непосредственно в программе в упорядоченном коде У-1 (или ему подобном).

Например, окончания -en, -rn, -ln, -nden запишутся в «памяти» ЭВМ «Минск-22» так:

```
                  e n
0000 0000 4455
                  r n
0000 0000 6155
                  l n
0000 0000 5355
            n d   e n
0000 5543 4455
```

Каждый формальный выделитель имеет в оперативном накопителе свой адрес. Так, в приведенной ниже программе две словоформы, zu и diesem, размещаются соответственно в ячейках с номерами $\overset{zu}{0722}$ и 1003:

```
                  z u
0722) 7164 0000 0000
            d i   e s   e m
1003) 4350 4462 4454
```

14.4. Принципиальный алгоритм решения задачи

Несмотря на все трудности в процессе моделирования лингвистического поведения человека (см. стр. 340—343), именно это поведение является для нас основой для построения модели различения лексико-грамматической омографии словоформы zu. Мы моделировали процесс выполнения этого задания информантом, в роли которого выступал преподаватель кафедры фонетики Минского ГПИИЯ А. Н. Шаранда.

Принципиальный алгоритм задачи различения лексико-грамматической омографии словоформы zu состоит из нескольких алгоритмов (рис. 28—29, см. Приложение).

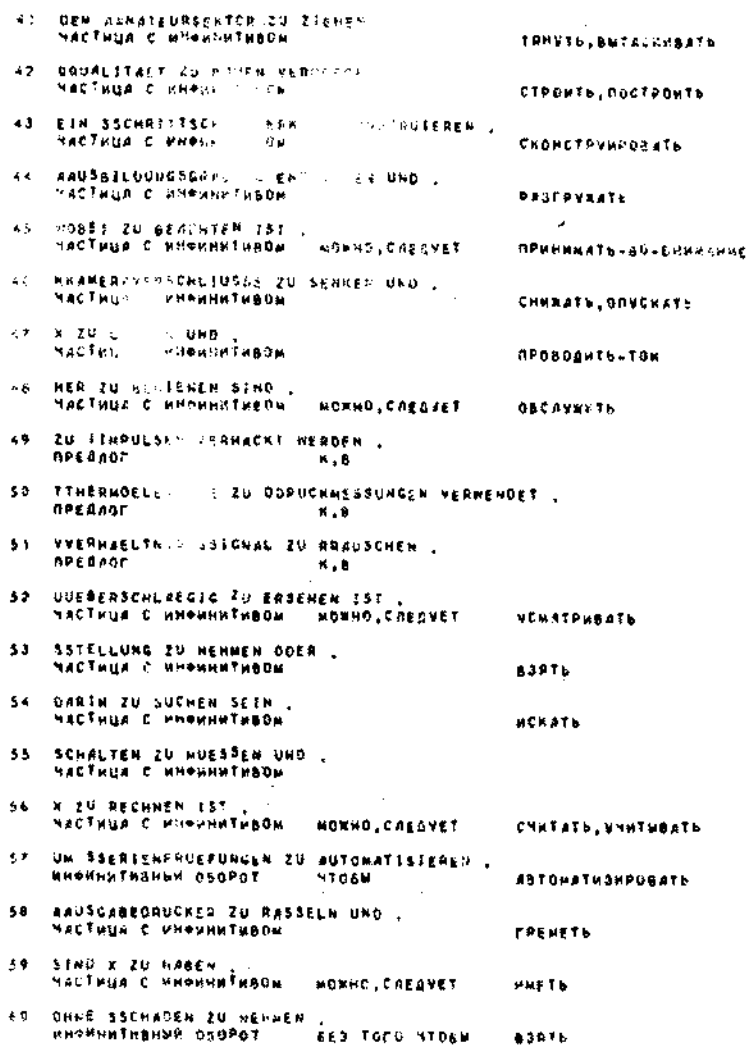


Рис. 11. Часть результатов машинного анализа слова zu.

Алгоритмы рассмотренной задачи были запрограммированы для машины «Минск-22». Для проверки нашей функциональной модели было взято более 1000 конструкций, каждая из которых содержала zu в той или иной функции. В процессе отладки программы диалог человека с машиной позволил устранить в программе ряд технических ошибок и логических неточностей.

Рассмотрим несколько примеров.

В процессе кодирования конструкций для анализа оператор не поставил после очередной конструкции точку. Машина в таком случае читает две конструкции как одну. Но по условию нашей задачи в конструкцию должно входить не более восьми словоупотреблений (стр. 355). Получающаяся «двойная» конструкция занимает часть основной программы, и машина останавливается, показывая тем самым, что в такой постановке текст ей непонятен. Человек реагирует на сигнал машины, исправляя ошибку оператора.

Другой пример. Проводя дихотомический поиск по немецко-русскому словарю, машина не нашла перевода для инфинитива какого-либо глагола и напечатала: «(не найдено)». Но программист точно знает, что эта словоформа в словаре есть. Тогда он в свою очередь задает вопрос машине, вызывая из словаря искомую словоформу и ее перевод. Чаще всего оказывается, что в немецкой словоформе пропущена одна буква или вместо одной буквы поставлена другая (опечатка при перфорации).

Часть результатов работы модели приведена на рис. 11.

14.6. Анализ результатов работы модели

Итак, используя элементарные операции электронной машины, мы построили действующую модель поведения человека при решении задачи различения лексико-грамматической омонимии слова zu. Возникает вопрос, насколько удачна эта модель? Так ли на самом деле действует человек при решении подобной задачи. Первой приближенной оценкой нашей функциональной модели может служить количество правильных ответов машины при проверке незнакомого текста. Из общего числа в 1016 конструкций, проверяемых полученной моделью, неверный результат был получен лишь в 30 случаях.[124] Но эта цифра ничего не говорит о том, какие участки модели работают наиболее надежно, а какие имеют наибольшее число логических ошибок. Такой анализ позволит нам в дальнейшем усовершенствовать разработанную здесь модель. Рассмотрим более подробно ошибки, допущенные машиной.

[124] А. Н. Шаранда, А. В. Зубов. Разрешение омонимии на ЭВМ. В сб.: Материалы XIX научно-теоретической конференции. Языкознание, Минск, 1967.

Наибольшее число ошибок (17) относится к случаям неразличения машиной zu в функциях «усилительная частица» и «предлог».

Основная причина этого — отсутствие формальных выделителей для zu в функции усилительной частицы.

Например, во всех следующих случаях:

№ 66 ein zu grosses mmas;
№ 322 viel zu viel zzeit;
№ 415 die zu hohe sstroeme

машина дала ответ: «предлог», ибо нашла на втором месте после zu признак существительного (удвоенная первая буква). Чтобы в дальнейшем избежать этой ошибки, можно отдельным списком задать словоформы, положение которых после zu характеризует его в функции усилительной частицы. Вторая причина смешения в процессе анализа тех же функций zu заключается в недостаточной «мощности» микроконтекста. В процессе построения модели мы сделали допущение (стр. 351), что существительное за zu может стоять не далее как на 2-м месте. В результате этого ряд конструкций был проанализирован неверно:

№ 963 um zu organisch konstruierten mmaschinen;
№ 995 das zu ganz neuen sschaltungen einlaedt.

Здесь в обоих случаях существительное находится на 3-м месте после zu. Машина, проверив на признак существительного первые две, следующие после zu, словоформы, дала в результате анализа ответ: «усилительная частица». Устранить подобные ошибки можно путем расширения микроконтекста.

Наконец необходимо отметить еще один тип ошибок, когда машина не различает zu в функции усилительной частицы и в функции приинфинитивной частицы. Рассмотрим следующие конструкции:

№ 414 einen zu kleinen oder,
№ 758 vor zu hohen sspannungen.

В результате анализа словоформ, стоящих непосредственно за zu, машина ответила, что в конструкциях № 414 и № 758 zu выступает в роли приинфинитивной частицы. На самом же деле здесь имеет место случайное совпадение конечных морфем. Для устранения в будущем подобных ошибок необходимо задать в пределах подъязыка список словоформ, которые оканчиваются на -en и могут стоять непосредственно за zu в функции усилительной частицы.

Только три ошибки из общего числа в 30 можно отнести к ошибкам техническим. Все они одного типа — ошибки в кодировании существительного. Как мы уже отмечали, признаком существительного является наличие удвоенной первой буквы словоформы. В процессе перфорации текста или словаря возможны случаи, когда первая буква существительного не повторяется. Это приводит к ошибкам в процессе анализа. Например:

№ 405 mmodell zu model umgestellt.

В первом употреблении существительное Model написано правильно, во втором — нет. В результате анализа такой конструкции машина, не обнаружив существительного и необходимых формальных выделителей, неправильно опознала zu, охарактеризовав ее как усилительную частицу.

Проведенный анализ позволяет сделать следующие рекомендации по усовершенствованию модели:

1) необходимо задать список словоформ, положение которых после zu распознаёт эту словоформу в функции усилительной частицы;

2) проверку на наличие признака «усилительная частица» проводить раньше, чем аналогичную проверку на признак «приинфинитивная частица»;

3) можно предположить, что, поскольку далекие статистические связи между словами практически затухают через 4—5 слов,[125] расширение глубины микроконтекста до 4—5 шагов вправо (и влево) даст практически полную распознаваемость конкретных словоформ, в том числе и zu; при этом практически не изменится ни загрузка «памяти» машины, ни объем программы, лишь слегка удлинится время анализа.

§ 15. Модель распознавания человеком лексического значения многозначного глагола faire

15.1. Лингвистические аспекты задачи

Приведенная выше программа является одной из простейших моделей речевого поведения, ибо число возможных выходов программы (число грамматических значений словоформы zu) невелико.

Попытаемся теперь построить более сложную модель, функционально имитирующую поведение человека при распознавании лексического значения французского глагола faire. Здесь, как и в предыдущем параграфе, в роли метаязыка выступает естественный русский язык, и все определения, относящиеся к многозначности, идиоматичности и т. п., понимаются в плане двуязычной ситуации.

В глагольной системе французского языка выделяется несколько слов с отвлеченным смысловым содержанием. Такие глаголы могут употребляться в значении многих других глаголов. Например, такие понятия, как «шагать», «идти», «бежать», «плыть», «лететь», «подниматься», «опускаться», «ползти» и т. д., могут быть во французском тексте переданы одним и тем же глаголом

[125] Г. П. Богуславская. Энтропия английского печатного текста, стр. 15.

aller без ущерба для ясности и точности передачи содержания. Вместе с тем следует помнить, что для указанных русских глаголов существуют и их эквиваленты во французском, соответственно: marcher, courir, nager, voler, monter, descendre, ramper и т. д. К другим подобным глаголам общего значения во французском языке относятся avoir, etre, faire, mettre, prendre, tenir, donner и др. Как отмечал С. Ульман, максимального абстрагирования в своей лексической структуре французский язык достигает именно в классе глаголов.[126]

Рассмотрим лингвистические особенности одного из наиболее абстрактных, многозначных и частотных французских глаголов — глагола faire.[127]

В качестве лингвистической основы для построения алгоритма были выбраны 9000 конструкций, взятых из произведений фран-

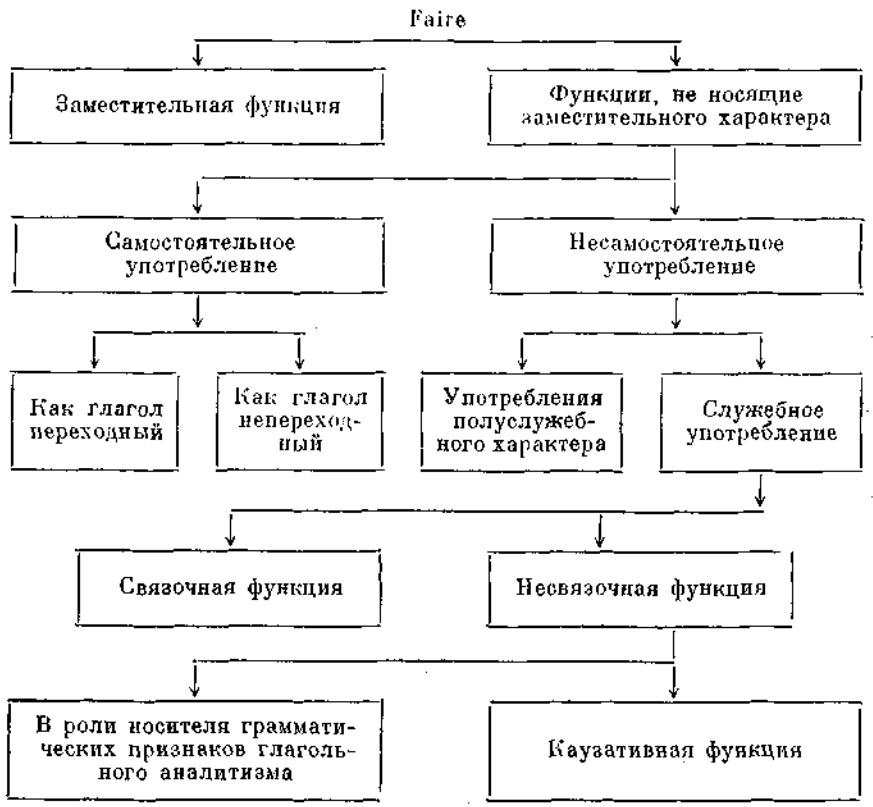


Рис. 12. Функции глагола faire.

[126] S. Ullman. Précis de sémantique française. Bern, 1959, p. 143.
[127] V. Brøndal. Les parties du discours. Copenhague, 1947, p. 145; G. E. Van der Beke. French Word Book. New York, 1934.

цузских писателей, а также из газет. На первом этапе исследования применением дистрибутивного анализа к этим конструкциям были выделены 23 дистрибуционные формулы.[128]

Дальнейший анализ полученных формул с точки зрения особенностей употребления в них исследуемого глагола позволил выделить следующие его функции [129] (см. рис. 12).

Рассмотрим формальные признаки, характеризующие то или иное функциональное употребление faire.

1. Основным формальным признаком употребления faire в заместительной функции (889 случаев, вероятность $p=0.12$) является возможность подстановки вместо этого глагола предыдущего глагола, глагольного выражения или целого отрезка высказывания, а также отсутствие прямого дополнения при faire. Помимо этого, faire в такой функции может быть замещен более конкретными по смысловому содержанию глаголами, например глаголом dire.

Примеры:

Je levai le bras comme fait (lève le bras) l'agent de police qui stoppe les voitures au carrefour;

«Tu ne regrettes pas ta promenade habituelle», fit (dit)-il en souriant.

Здесь в скобках даны возможные подстановки.

2. Использование faire в самостоятельном употреблении может быть охарактеризовано нижеследующими формальными признаками.

а) Как прямо-переходный глагол (997 случаев, $p=0.13$) он характеризуется возможностью проведения активно-пассивной трансформации и ее разновидностей (образованием пассивного инфинитива и заменой существительного объектным местоимением.

Примеры:

— активно-пассивная трансформация:

Le tailleur fait un costume;
Le costume est fait par le tailleur;

— образование пассивного инфинитива:

Il n'avait pas fait cinquante mètres. . .;
Il n'avait pas cinquante mètres à faire;

[128] Л. Ф. Кистанова. Функционально-семантическая характеристика глагола faire в современном французском языке. КД (рукоп.), Минский ГПИИЯ, 1966, стр. 19—27.
[129] Там же, стр. 46.

— замена существительного объектным местоимением:

Elle f a i t l e d e v o i r à domicile;

Elle l e f a i t.

б) Как самостоятельный непереходный (390 конструкций, $p=0.05$) глагол faire характеризуется отсутствием прямого дополнения, а также возможностью замены faire непереходным глаголом (например: agir), выражающим действие безотносительно к объекту.

П р и м е р ы:

J'avais cru bien f a i r e, non petit;
J'avais cru bien a g i r, mon petit.

В несамостоятельном употреблении глагола выделяются его употребления, где функционирование носит полуслужебный характер (640 конструкций, $p=0.084$), и его употребление в роли служебного глагола.

3. Трудно выделить какие-либо признаки употребления faire в полуслужебной функции. Можно лишь отметить, что те словосочетания, где faire выступает как глагол полуслужебный, характеризуются ограниченной свободой перемещения компонентов по отношению друг к другу, а также ограниченными возможностями синтаксических преобразований или трансформаций.

4. Основным формальным признаком служебного функционирования faire в роли глагола-связки (783 случая, $p=0.10$) является сочетаемость его с именем прилагательным или существительным, выполняющим роль предикатива и выражающим предикативный признак подлежащего; сюда же относятся невозможность, без нарушения смыслового инварианта, проведения активно-пассивной трансформации и ее разновидностей, а также возможность замены faire связочным глаголом.

П р и м е р ы:

Ca f a i t sérieux;

La lampe lui f a i t les cheveux blancs.

5. Употребление faire в так называемой «каузативной функции» (1680 конструкций, $p=0.22$) характеризуется его сочетаемостью с инфинитивом другого глагола; при этом отмечается фиксированный порядок следования компонентов и невозможность постановки между faire и инфинитивом какого-либо именного компонента.

П р и м е р:

Ma femme va me f a i r e la vie si je fais attendre son souper.

6. И, наконец, faire в роли носителя грамматических признаков глагольного аналитизма (1765 случаев употребления, $p=0.23$). Основной формальный признак — сочетаемость faire с именами существительными, называющими действие, а также трудность проведения активно-пассивной трансформации и замены существительного при глаголе объектным местоимением. В этом случае faire выступает средством вербализации имени существительного и по своей грамматической функции является эквивалентом глагольной словообразовательной морфемы.

П р и м е р:

I l f a i t l'inspection de la pièce (il inspecte la pièce).

Рассмотренные формальные признаки на самом деле являются формальными только для человека, обладающего эвристическими способностями, но не для машины. Пока что машина еще не может отличить осмысленную («отмеченную») фразу от грамматически правильной, но бессмысленной («неотмеченной»), и для нее не являются существенными [130] такие формальные признаки, как «возможность подстановки faire вместо другого глагола», «возможность проведения активно-пассивной (и любой другой) трансформации». Без предварительной разметки текста машина «не поймет» и такие признаки, как «сочетаемость с именем прилагательным или существительным, выполняющими роль предикатива», а также ряд аналогичных ситуаций.

Помимо этого, нельзя провести отчетливой формальной границы между различными функциями глагола faire. Между всеми указанными функциями этого слова существуют многочисленные переходные случаи. Поэтому установленные выше дифференциальные признаки функций не являются формальными для машины.

Каким же образом подойти к задаче формального различения значений faire при переводе? На первый взгляд кажется, что задача невыполнима. И в самом деле, в той постановке, как было указано выше, она действительно не может быть решена. Где же выход? Попробуем разобраться, как переводит многозначное слово человек. Прежде всего условимся, что будем ана-

[130] Термин «существенный формальный признак» мы понимаем в широком плане с учетом возможностей современных ЭВМ. Можно, конечно, составить для подъязыка набор структурных формул, характеризующих употребление какого-либо одного слова, например faire. Тогда можно формализовать проверку признака «возможность подстановки faire вместо другого глагола». Можно задать для отдельного слова подъязыка и список возможных трансформаций. Однако все это потребует использования ЭВМ с очень большой «памятью». Вполне естественно, что использование моделей лингвистического поведения человека, базирующихся на таких признаках, для автоматического анализа больших массивов текста практически невозможно.

лизировать работу человека, для которого французский язык не является родным, и что для перевода в некоторых случаях человек пользуется различного рода пособиями, справочниками и словарями. Такой подход объясняется не отсутствием информанта — носителя языка, а желанием приблизить возникающие при переводе ситуации к возможностям современных электронно-вычислительных машин. Известно, что носитель языка накапливает информацию об употреблении того или иного слова в течение всей своей жизни. Ассоциативное запоминание и эвристические способности носителя языка в настоящее время трудно смоделировать. Иное дело человек, который выучил второй язык. Обучение языку проходит в определенный ограниченный отрезок времени.[131] За этот отрезок времени человек запоминает определенное количество слов, правила их употребления, читает конечное число текстов на изучаемом языке. Этот процесс гораздо легче формализовать. Безусловно, что переводчик со временем начинает все больше пользоваться эвристическими приемами. Тем не менее, удовлетворяющие специалистов переводы научно-технических текстов чаще всего осуществляются людьми, для которых второй язык не стал родным. Как же действует такой переводчик? Практика показывает, что если то или иное многозначное слово встретилось в контексте, не относящемся к технической специальности переводчика, то он обращается к словарям и справочникам. При выборе значения из тематического словаря переводчик, очевидно, в первую очередь «подставляет» в переводимый текст те значения многозначного слова, которые встречались ему достаточно часто в течение его переводческой практики. В качестве основы такого предположения взята гипотеза, сущность которой заключается в том, что в механизмах, связанных с восприятием и воспроизведением речи, слова и другие элементы письменной речи кодируются и декодируются с учетом их вероятностей. Носитель языка не только интуитивно владеет вероятностными закономерностями, но и определенным образом использует их для оптимизации переработки информации.[132]

Но как расположены эти вероятностные единицы в памяти человека и каким образом они используются — однозначного ответа на эти вопросы пока нет.[133]

[131] Переводчик, конечно, учится языку всю жизнь в процессе своей работы. Мы здесь имеем в виду тот минимальный промежуток времени, в течение которого человек получает квалификацию переводчика (например, пятилетний срок обучения в языковом вузе).

[132] А. Н. Леонтьев, Е. П. Кричик. О некоторых особенностях процесса переработки информации человеком. ВП, 1962, № 6; Р. М. Фрумкина. Проблемы восприятия слов в зависимости от их вероятности. В сб.: Проблемы языкознания. Доклады и сообщения советских ученых на X Международном конгрессе лингвистов, М., 1967.

[133] Например, одна гипотеза предполагает, что высоковероятные единицы и их комбинации хранятся в долговременной памяти, а маловероятные — в оперативной. Другая гипотеза предполагает, что слово записывается в памяти человека вместе с его вероятностью. См., например: Р. М. Фрумкина. Проблемы восприятия слов. . . , стр. 93.

Задача значительно осложняется при изучении двуязычной ситуации. Где хранятся у человека «иностранные» слова и где — их эквиваленты на родном языке? Каким образом связаны между собой процессы отождествления родных и неродных слов? Для ответа на эти и им подобные вопросы понадобятся многие годы исследований лингвистов, психологов, физиологов и кибернетиков.

Итак, следуя тому принципу, который использует человек при переводе многозначного слова на родной язык, применим вероятностную методику к задаче перевода на русский язык французского слова faire. Первым шагом в этом направлении будет сужение рамок исследования. Будем брать не весь язык, а макроконтекст — определенный научно-технический подъязык. В результате этого число возможных значений полисемантичных слов, в том числе и faire, сокращается в два и более раз.[134] Далее, указанные выше функциональные значения faire дадут в подъязыке иное распределение, чем в языке в целом. Изменится их вероятность употребления. На первый план выдвинется несколько наиболее вероятных функций, характерных для данного подъязыка. Остальные употребления будут встречаться очень редко. Допустим, что во взятом подъязыке наиболее частыми оказались следующие значения и функции faire:

1) faire в каузативном значении;

2) faire в функции носителя грамматических признаков глагольного аналитизма;

3) faire как связка.

Рассмотрим, каким образом можно формализовать для машины эти значения и функции faire.

1. Наиболее формальными (с «машинной» точки зрения) являются признаки, характеризующие faire в каузативной функции. Их можно представить в виде следующей таблицы:

Функция faire в предложении	Формальные признаки		Перевод
	признак инфинитива	место инфинитива в конструкции	
Каузативная	*-er*, *-ir*, *-oir*, *-re*	*1-я 2-я или 3-я* словоформа за faire	'заставить что-либо сделать'

Однако и в этом случае имеется целая группа глаголов, в комбинации с которыми faire не переводится отдельным словом.

[134] Л. Ф. Кистанова, ук. соч.

Русский эквивалент относится ко всей конструкции «faire+инфинитив». Например:

faire appeler	'позвать, вызвать'
faire croire	'убеждать'
faire passer	'пропустить'

Применяя программы, подобные приведенным в § 9 гл. 1, можно выбрать из подъязыка конструкции с faire и затем, анализируя их, выделить микроконтексты, необходимые для определения правильного перевода сочетания «faire+инфинитив». Нас в данном случае интересует не определение той функции, которую выполняет faire в предложении, а его перевод на русский язык, поэтому каждый выделенный микроконтекст сопровождается наиболее вероятным переводом. В приведенной выше таблице, в графе «место инфинитива в конструкции», мы написали: «1-е, 2-е или 3-е место за faire». Возможно, конечно, что между faire и инфинитивом будет стоять и большее число слов, но вероятность такой ситуации очень мала.

2. Рассмотрим употребление faire в функции аналитичной морфемы глагольности и один из формальных признаков этой функции — сочетаемость faire с именами существительными, называющими действие. В зависимости от степени грамматизации faire в сочетании с именами существительными различают аналитические глагольные единицы, конструкции с полной грамматизацией faire и полуаналитические конструкции[135] с неполной грамматизацией всего сочетания, где глагол сохраняет лексическое значение, хотя оно и является достаточно ослабленным.

Первые из них представляют собой устойчивые словосочетания, характеризующиеся грамматической неделимостью и постоянством компонентов. Такие словосочетания закреплены в системе самого языка. Например:

faire retour	'возвращаться'
faire reproche	'упрекать'
faire plaisir	'доставлять удовольствие'

Частотные фразеологические словари французского языка показывают, что из общего числа глагольных аналитических единиц этого типа 30—35% приходится на фразеологизмы, образованные с глаголом faire.[136]

В полуаналитических конструкциях сочетания глагола faire с именами существительными представляют собой аналитические

[135] В. М. Жирмунский. Об аналитических конструкциях. В сб.: Аналитические конструкции в языках различных типов, М.—Л., 1965, стр. 32.

[136] З. Н. Левит. К проблеме изучения компонентов фразеологических единиц. НДВШ, Филологические науки, 1965, № 2, стр. 76—85.

эквиваленты слов, ибо по смысловому содержанию они соответствуют простым глаголам, основу которых составляет в большинстве случаев имя существительное, сочетающееся с глаголом faire. Например:

faire une visite	'посещать'
faire l'inspection	'инспектировать'
faire des calculs	'считать'

В подобных случаях faire выступает в качестве служебного элемента и перевод целиком определяется существительным. Такие образования не зафиксированы в словарях (в языке), а постоянно возникают в речи. Чтобы получить более или менее связанный перевод faire в функции аналитической морфемы глагольности необходимо:

а) задать отдельным списком наиболее частотные устойчивые сочетания «faire+существительное» — для аналитических конструкций;

б) задать отдельным списком наиболее частые существительные, которые определяют перевод сочетания «faire+существительные» в полуаналитических конструкциях.

3. Наконец, рассмотрим faire в функции связки. В этой функции faire в определенных условиях синтаксического употребления выступает как средство присоединения предикативного признака к подлежащему или дополнению. Статистический анализ показывает, что, как правило, при faire в функции связки употребляется существительное без артикля, прилагательное или причастие. Здесь, как и в двух предыдущих случаях, для получения перевода необходимо задать список наиболее вероятных существительных, прилагательных и причастий, при которых faire выступает в роли связки.

Вероятностная методика, используемая в данной работе, есть отражение некоторых приемов работы человека в условиях неполной, неясной, нечеткой лексико-грамматической информации о структуре языка. В таких случаях человек ставит частные задачи, подпроблемы, намечает частные цели. Такая разбивка основной проблемы может рассматриваться как первый алгоритмический шаг к эвристической программе.

Ставя в общей задаче перевода faire отдельные частные задачи и решая последние применением вероятностных методов, мы, как указывалось выше, не можем учесть всех факторов, влияющих на значение в процессе перевода. Рассмотрим те дополнительные ограничения, с учетом которых строилась модель.

Помимо трех указанных функций faire, мы включили в анализируемый текст ряд конструкций, в которых faire выступает в других функциях. Чтобы не загромождать модель слишком большим словарем, мы не включили в него слова, характеризующие перевод faire в соответствующих функциях, а снабдили

faire в этих функциях специальными пометами. Это, в частности, относится к тем случаям, когда faire используется в заместительной или самостоятельной функциях и переводится как 'делать', а также в случае, когда этот глагол используется в настоящем времени и, как правило, не переводится (или переводится как 'быть'). При построении модели мы исходим из предположения, что человек переводит не отдельными словами, а ищет в предложении все возможные микроконтексты переводимого слова. Если же никакого типового микроконтекста для взятого слова нет, то такое слово переводится однозначно, в зависимости от наличия той или другой пометы при нем.

Во всех остальных случаях, когда машина в результате анализа не находит ни одного из предусмотренных выходов, она выдает на печать слова: «не учтено». Это означает либо ошибку в тексте или в словаре, либо то, что faire употреблен в функции, которая не учитывалась в алгоритме анализа.

Таким образом, лингвистическую задачу перевода глагола faire на русский язык можно сформулировать так:

1) построить словарь микроконтекстов, характеризующих тот или иной перевод faire на русский язык;

2) в каждой очередной конструкции текста найти одну из форм глагола faire, найти входящие с ним в один микроконтекст слова́ конструкции и перевести микроконтекст на русский язык; задаются следующие формы глагола faire: faire, fait, fais, faits, faite, faisons, faites, font, ferai, feras, fera, ferons, ferez, feront, faisais, faisait, faisions, faisiez, faisaient, fis, fit, fîmes, fîtes, firent, ferais, ferait, ferions, feriez, feraient, fasse, fasses, fassions, fassiez, fassent, fisse, fisses, fît, faisant;

3) напечатать конструкцию, микроконтекст и перевод [137] глагола faire в этом микроконтексте на русский язык.

15. 2. Математическая постановка задачи

Сформулированное выше лингвистическое задание может быть представлено в виде следующей последовательности формальных операций:

1) каждый микроконтекст с переводом глагола faire преобразуется в стандартные машинные документы $Д_1$, каждый из которых содержит $2m$ ячеек оперативного накопителя; в первых m ячейках размещается код микроконтекста, во вторых — код перевода; все документы $Д_1$ располагаются в зоне М МОЗУ;

2) каждая форма глагола faire преобразуется к стандартному виду и записывается в m_1 ячейках оперативного накопителя; в МОЗУ образуется зона Ф, включающая 38 групп по m_1 ячеек ($m_1 \leqslant m$);

[137] В рассматриваемой задаче все формы глагола переводятся в инфинитиве.

3) каждое словоупотребление конструкции, используемой для проверки модели, преобразуется к стандартному виду и записывается в m ячейках запоминающего устройства;

4) из общего текста выделяется очередная конструкция; для этого каждое новое содержимое m ячеек (код словоупотребления конструкции) передается в определенную зону накопителя (зона К), содержащую n групп по m ячеек; признаком конца конструкции считается занесение кода точки в зону К;

5) последовательным сравнением содержимого зоны К с содержимым зоны Ф отыскивается и запоминается место группы m' в зоне К, первые m_1 ячеек которой совпадают с определенными m_1 ячейками группы Ф;

6) для каждой группы m_1 отыскивается в зоне К такая группа m, чтобы их общее содержимое совпадало с одним из документов $Д_1$ в зоне М;

7) расшифровывается содержимое зоны К и найденного документа $Д_1$.

15. 3. Подготовка начальных данных для ввода в ЭВМ

Для проверки модели был взят текст, содержащий около 500 конструкций из произведений французских авторов и из газет. Каждая порция текста содержала 50 конструкций, закодированных в коде М-2 на перфоленте. Предполагалось, что в конструкции может быть не более 16 словоупотреблений (считая точку за словоупотребление).

Для словаря микроконтекстов были выбраны наиболее часто употребляющиеся глаголы, имена существительные, прилагательные и наречия, которые в сочетании с faire давали определенный перевод. Пример построения такого словаря приведен в табл. 3.

Те конструкции, в которых faire не имел соответствующих микроконтекстов, содержали при faire специальные пометы. Помета в виде индекса ставилась через черточку за формой глагола faire и означала следующее:

faire-P — переводится как 'делать';
faire-N — не переводится.

При кодировании словаря микроконтекст отделялся от соответствующего перевода глагола faire точкой. Например, при кодировании словаря, представленного в табл. 3, словоформы записывались следующим образом:

appeler. позвать, вызвать. croire. убеждать . . . . . bilan. подвести итог. de la politique. заниматься политикой . . .

Словарь перфорировался двумя порциями на отдельную перфоленту.

Таблица 3

Часть словаря микроконтекстов для глагола faire

№№	Функция в конструкции	Левый компонент микроконтекста	Правый компонент микроконтекста	Перевод микроконтекста
1	Каузативная функция		appeler	'позвать', 'вызвать'
2			croire	'убеждать'
3			cuire	'сварить'
4			épouser	'женить'
5			rire	'рассмешить'
6			venir	'вызвать'
7			allusion	'намекать'
8	Носитель грамматических признаков глагольного аналитизма		attention	'обращать внимание'
9			bilan	'подвести итог'
10			de la politique	'заниматься политикой'
11			des essais	'испытывать'
12			du bien	'помогать'
13			l'achat	'закупать, покупать'
14			la lettre	'писать письмо'
15			idée	'составлять представление'
16			illusions	'заблуждаться'
17			ministre	'быть министром'
18			jeune	'выглядеть молодым'
19			chaud	'тепло'
20			petit	'становиться маленьким'
21			vite	'быть сделанным быстро'

Формальные признаки инфинитива записывались в оперативном накопителе так же, как это было сделано при реализации предыдущей задачи (см. стр. 355).

15. 4. Разработка метода решения и принципиальный алгоритм решения задачи

Как и в предыдущей задаче, для построения модели мы использовали информанта, в роли которого выступала старший преподаватель Минского ГПИИЯ Л. Ф. Кистанова.

Принципиальный алгоритм задачи может быть представлен в виде схемы (рис. 28, см. Приложение).

15. 5. Программирование алгоритмов. Решение задачи на машине

При создании программы по рассмотренному алгоритму мы приняли, что каждая форма глагола faire записывается в двух ячейках запоминающего устройства ЭВМ «Минск-22» ($m_1=2$).

Каждое слово конструкции записывалось в четырех ячейках ($m=4$). Машинный документ $Д_1$ содержал $2 \cdot 4 = 8$ ячеек.

При реализации программы текст обрабатывался порциями по 50 конструкций в каждой. На печатающее устройство машины выдавалась следующая информация:

1) конструкция с faire — 1-я строка печати;

2) микроконтекст с одной из форм глагола faire и перевод глагола в данном микроконтексте — 2-я строка печати.

Если в конструкции было два глагола faire, то в 3-й строке печатался второй микроконтекст и соответствующий перевод второго глагола. Часть результатов приведена на рис. 13.

15. 6. Анализ результатов работы модели

Полученные в результате машинной реализации нашей программы данные содержат ряд ошибок в переводе или отказов ЭВМ от перевода. Все эти ошибки можно сгруппировать следующим образом:

1) ошибки, связанные с недостаточностью лингвистического задания;

2) чисто механические ошибки, возникающие в процессе перфорации текста и словаря.

К числу ошибок первого типа относятся случаи, когда машина не нашла в словаре нужный микроконтекст.

Рассмотрим несколько примеров.

№ 294 pour le faire comprendre à l'autre.

Здесь машина выделила в следующем за faire слове comprendre признак инфинитива глаголов 3-й группы, но не нашла словоформы comprendre в словаре, ибо она не была задана.

№ 275 y ne faisaient pas la charité;
№ 361 il fait aussi des vers;
№ 496 à qui fait elle du tort.

Для этих трех примеров в словаре микроконтекстов не были заданы соответственно выражения: la charité, des vers, du tort. В результате анализа этих конструкций машина напечатала: «случай не учтен».

К ошибкам второго типа относится, например, следующий случай:

№ 206 semblable au paysan qui fait sa tournée dans son domaine.

В словаре вместо tournée ошибочно было задано tourneé. Результат анализа тот же самый: «случай не учтен». Из текста, содержащего 498 конструкций с глаголом faire, ошибочный ре-

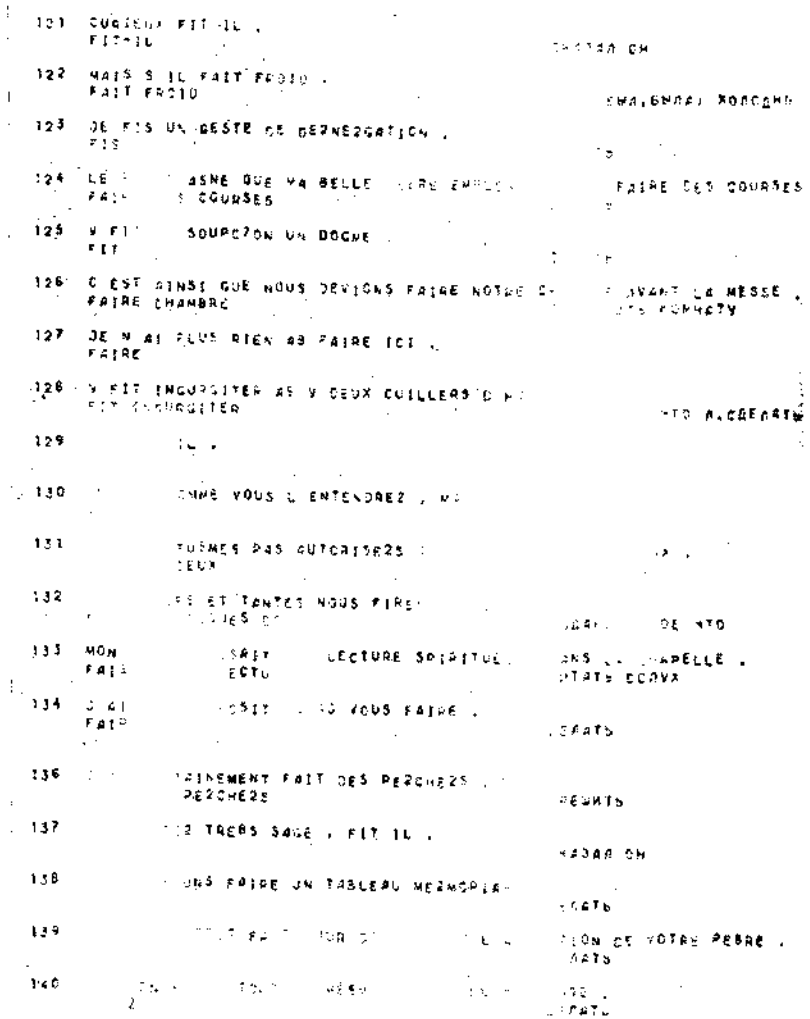


Рис. 13. Часть результатов автоматического анализа глагола faire.

зультат машина дала в 11 случаях, что составляет 2.2%. Такой невысокий процент ошибок свидетельствует о том, что используемая нами вероятностная методика является достаточно надежным способом определения значения многозначного слова при переводе.[138]

Глава III. ДВЕ МОДЕЛИ ПЕРЕВОДА СВЯЗАННОЙ ГРУППЫ СЛОВ

§ 16. Формализация процесса перевода

Перевод любой связанной группы слов с одного языка на другой можно характеризовать как определенное преобразование сообщения. Поэтому попытаемся определить то, что мы интуитивно понимаем под переводом, с точки зрения процесса коммуникации.[139]

Во всяком акте коммуникации есть передатчик сообщения (отправитель А) и приемник сообщения (адресат Б). В рассматриваемом нами случае коммуникации-перевода отправитель А и адресат Б пользуются для передачи сообщения различными системами. Поэтому в процесс коммуникации включается переводчик П, который является одновременно адресатом по отношению к отправителю А и отправителем по отношению к адресату Б. Логическую схему их взаимодействия можно представить следующим образом (рис. 14).

Здесь отправитель А передает некоторое сообщение C_1 о ситуации в действительности Д переводчику П. Последний, получив сообщение C_1, переходит к некоторой системе соответствий между исходным языком (ИЯ) и переводящим языком (ПЯ). Эта система соответствий с помощью ПЯ формирует сообщение C_2, которое и информирует адресата Б о ситуации в действительности Д.

Как видно из этой схемы, процесс перехода от одной системы языка к другой осуществляется без непосредственного обращения к ситуации, имеющей место в действительности. Конечно, когда используемая система соответствий устанавливалась, то учитывалась та действительность, те ситуации, которые отражают соответствующие категории в том и другом языке. Чрезвычайно важно, однако, что это — факт прошлого, а не самого процесса перевода.[140]

---

[138] Л. Ф. Кистанова, А. В. Зубов. К вопросу об автоматическом анализе семантики многозначного слова. В сб.: Проблемы изучения семантики языка. Научная конференция, посвященная 50-летию Днепропетровского гос. унив., Днепропетровск, 1968.

[139] И. И. Ревзин, В. Ю. Розенцвейг. Основы общего и машинного перевода. М., 1964, стр. 56—64.

[140] Там же, стр. 58—59.

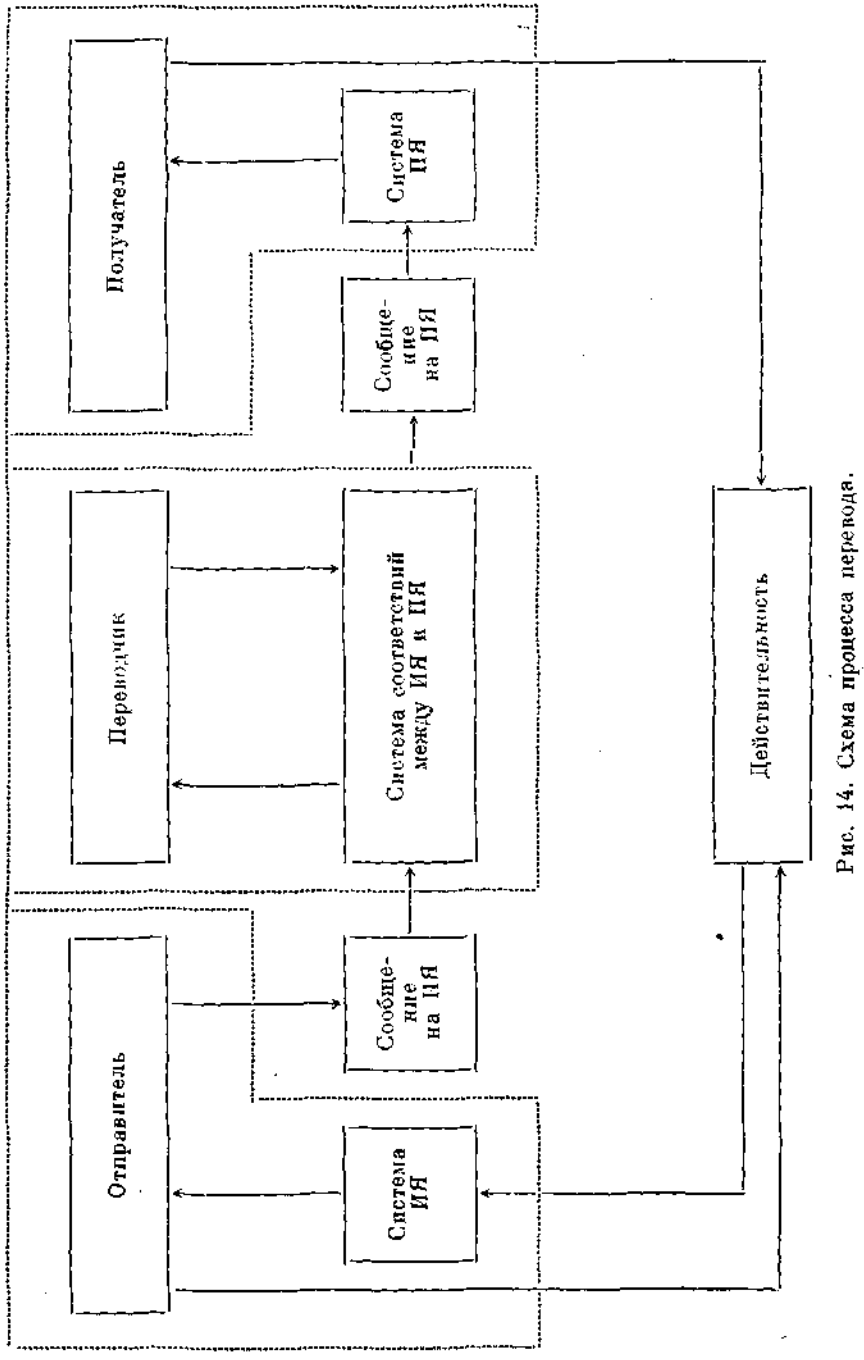


Рис. 14. Схема процесса перевода.

Таким образом, задача полной формализации перевода сводится к вопросу о том, можно ли для любых двух языков установить такое соответствие между лексическими и грамматическими категориями, которое предполагается схемой перевода. В задачу настоящей работы не входит рассмотрение теоретических вопросов, относящихся к этой проблеме. Стоит лишь отметить, что по этому вопросу существуют две противоположные точки зрения. Одна из них допускает возможность полной формализации и, значит, машинной реализации; [141] другая же говорит о несводимости одна к другой двух картин мира, описываемых разными языками.[142]

Тем не менее, несмотря на доводы, приводимые сторонниками второй точки зрения, автоматический перевод существует. Уже несколько лет отдел иностранной техники военно-воздушных сил США использует систему, которая переводит со скоростью 10 тыс. слов в час.[143] Правда, они переводят лишь заголовки статей без детального синтаксического и семантического анализа.

Сейчас наступила вторая стадия в развитии автоматического перевода. Основное внимание уделяется лингвистическим проблемам. Главная цель состоит теперь не в том, чтобы немедленно получить «полностью автоматический перевод высокого качества», о котором говорили в 50-е годы, а в том, чтобы создать систему, которая обеспечит незамедлительное улучшение качества переводов и в то же время позволит производить дальнейшее усовершенствование системы уже после ввода ее в эксплуатацию.[144]

Основное внимание в процессе автоматического анализа обращается в настоящее время на синтаксис. Он считается отправной точкой для решения большинства других проблем перевода.

Мы уже говорили во введении о важности автоматического анализа текстов для решения большого числа задач, вызванных современными социальными процессами (см. стр. 286). Остановимся подробнее еще на одном применении алгоритмов автоматического анализа. Речь пойдет о построении научной методики обучения языкам.

Еще Л. В. Щерба в процессе обучения языку выделял две основные цели:

1) первая цель состоит в том, чтобы изучающий умел понять текст *на иностранном языке, т. е.* чтобы ему была ясна функция каждой формы иностранного языка;

---

[141] Л. И. Жирков. Границы применимости машинного перевода. ВЯ, 1956, № 5; D. G. Hays. Linguistic Research of the RAND Corporation. Proceedings of the Natural Symposium in Machine Translation, London, 1961.

[142] И. Бар-Хиллел. Будущее машинного перевода. НДВШ, Филологические науки, 1962, № 4, Л. С. Бархударов, Г. В. Колшанский. К вопросу о возможностях машинного перевода. ВЯ, 1958, № 1.

[143] С. Перше. Машинный перевод: вторая стадия развития. Русск. пер. М., 1968, стр. 6.

[144] Там же, стр. 11.

2) вторая цель заключается в том, чтобы изучающий умел выражать мысли на иностранном языке, т. е. знал все формы, которые выражают данную мысль, и условия применения каждой из них.

Каждой из этих целей должна соответствовать своя грамматика. Грамматику первой цели Л. В. Щерба назвал пассивной, второй — активной. Как видно из этих определений, пассивная грамматика изучает функции значения строевых элементов данного языка, исходя из их форм, т. е. из внешней их стороны. Активная грамматика учит употреблению этих форм.[145] Большую практическую помощь при разработке разумной методики пассивной грамматики может оказать разбор готовых алгоритмов автоматического анализа, в которых имеется составленный по чисто формальным критериям полный список правил, необходимых для перевода достаточно простого текста. Конечно, не именно этими правилами следует руководствоваться при обучении языку. Это нецелесообразно хотя бы потому, что каждое правило обычной грамматики распадается при автоматическом анализе на ряд более элементарных, что важно для машины, но вовсе не нужно для человека. Здесь существенно лишь то, что при таком анализе дан перечень всех форм данного языка, с указанием возможных функций каждой формы и условий, при которых форма имеет ту или иную функцию.[146]

В последующих двух параграфах будет рассмотрен алгоритм перевода на русский язык немецких предложно-именных групп как один из простейших образцов автоматического анализа синтаксических отношений.

§ 17. Первая модель перевода немецких предложно-именных групп

17. 1. Лингвистические аспекты задачи

Известно, что всякое синтаксическое отношение слагается из трех элементов. Сюда входят два связанных между собой члена отношения и указание вида связи между ними. Вид связи может выражаться либо эксплицитно (например, предлогом или союзом), либо имплицитно одним или обоими членами отношения, либо же просто задаваться порядком слов.[147] Возникает задача выражения

[145] Л. В. Щ е р б а. Преподавание иностранных языков в средней школе. В сб.: Общие вопросы методики, М., 1947.

[146] Ср.: А. П. С и д е л ь к о в с к и й, Алгоритмический подход к анализу процессов обучения правомерен. ВП, 1962, № 6; А. А. З а л и з н я к. Опыт обучения англо-русскому переводу с помощью алгоритмов. В сб.: Питання прикладної лінгвістики. Тезі доповідей міжвузівської наукової конференции, Чернівці, 1960.

[147] С. П е р ш к е, ук. соч., стр. 12.

такой связи в виде конечного числа элементарных правил. Вполне естественно, что в первую очередь встает вопрос о том, как человек приходит к выводу о наличии той или иной связи между членами синтаксического отношения. Ведь если мы функционируем каким-то определенным способом, то это должно быть обусловлено эффективностью этого способа и его ценностью с точки зрения выживания. Но если это так, то каким образом и при каких допущениях мы можем убедиться, что именно этот вид деятельности эффективен?

Прежде всего рассмотрим один из простейших случаев синтаксической связи — связь между предлогом и существительным в немецкой предложно-именной группе.

Мы уже отмечали, что наиболее частые слова (словоформы) частотного словаря, включаемые в базовый язык, многозначны. Это относится, в частности, и к предлогам. Так, в научном немецком тексте в процессе анализа выделено 13 значений для предлога mit и 15 — для прелога an.[148] Для определения контекстного значения предлога могут быть использованы, как и выше, макроконтекст и микроконтекст. Сокращение числа значений через макроконтекст осуществляется на первом этапе чисто интуитивным путем, без обращения к машине. Из общей системы значений, зафиксированных в словаре, вручную выделяются лишь те лексические значения, которые используются в данном подъязыке. На следующем шаге, используя микроконтекст, отыскивается частное значение предлога в выбранной конструкции. Например, при переводе с английского языка на русский необходимым микроконтекстом для снятия лексико-грамматической омографии предлогов будет микроконтекст с одной словоформой в левой дистрибуции и двумя словоформами в правой дистрибуции.[149]

Однако предлоги относятся к служебным словам, и им свойственны некоторые особенности. Если у знаменательных частей речи лексическое значение является центром смысловой структуры, сохраняя свое внутреннее единство несмотря на разнообразие грамматических форм изменений, то у предлогов лексические и грамматические функции совпадают.[150] В каждом отдельном случае лексическое значение предлога конкретизируется посредством полнозначных (знаменательных) слов. Какова эта связь? От каких свойств существительного зависит значение предлога? Каким образом знаменательные слова, в частности существительное и его определители, получают окончания в зависимости от управляющего предлога? Каков механизм этих процессов?

[148] С. С. Б е л о к р и н и ц к а я. Принципы составления немецко-русского словаря многозначных слов для МП. МП, 1958.

[149] О. А. Н е х а й, ук. соч., стр. 19.

[150] В. В. В и н о г р а д о в. Русский язык. Грамматическое учение о слове. М., 1947, стр. 663.

Конечно, нелегко учесть все факторы, влияющие на связь предлога и существительного в предложно-именной группе. О преобладании тех или иных факторов можно судить, лишь строя модели перевода таких групп человеком. Сравнение нескольких таких моделей, при разных начальных предположениях, позволит выявить степень влияния различных элементов на связь между предлогом и существительным.

Поэтому, строя нашу первую модель перевода предложных групп, мы исходим из следующих предположений:

1) между предлогом и следующим за ним именем существительным существует свободная связь, таким образом, здесь не учитываются жесткие идиоматические связи между предлогом и существительным (т. е. идиоматичность предложного управления); как и при построении двух предыдущих моделей, идиоматичность здесь понимается в плане двуязычной ситуации; ее следует понимать как невозможность перевести комбинацию слов слово за словом (в этом случае получается бессмыслица), и переводить такую комбинацию приходится целиком;

2) каждый из предлогов немецкого языка имеет не более двух эквивалентов в русском языке;

3) если в конкретном подъязыке провести статистику сочетаний «предлог+существительное», то можно выделить определенное число наиболее частых предложных групп; возникает вопрос о том, достаточен ли такой микроконтекст для определения значения предлога, входящего в наиболее частые предложно-именные группы; поэтому каждой такой предложной группе дан однозначный перевод, вместе с которым она и располагается в оперативном накопителе машины (табл. 4); встретив такой микроконтекст в исследуемой конструкции текста, машина дает тот перевод, который мы ей задали; сопоставление же такого перевода с более широким контекстом конкретной конструкции позволит опреде-

Таблица 4

Часть наиболее частотных предложных групп с переводами

Немецкая предложная группа	Перевод предложной группы	Немецкая предложная группа	Перевод предложной группы
am Ausgang	'на выходе'	in dem Falle	'в случае'
auf dem Gebiet	'в области'	in der Regel	'по правилу'
auf die Anode	'на анод'	in der Tat	'на практике'
aus der Katode	'из катода'	mit dem Wert	'с величиной'
bei der Messung	'при измерении'	mit der Zeit	'со временем'
beim Drücken	'при нажатии'	nach der Methode	'по методу'
durch die Wahl	'путем выбора'	über die Diode	'через диод'
für den Fall	'для случая'	unter dem Einfluss	'под влиянием'
im Abschnitt	'на отрезке'	von der Form	'от вида'
im Laufe	'в ходе'	zum Teil	'частично'

Таблица 5

Немецкие предлоги и их перевод на русский язык

Группа	Предлог	Перевод	Артикль после предлога	Требуемый русский падеж
Е	unweit	'недалеко от'		род.
	mittels	'посредством'		»
	während	'во время'	der	»
	infolge	'вследствие'	des	»
	oberhalb	'над'	eines	твор.
	unterhalb	'под'	einer	»
	innerhalb	'внутри'		род.
	außerhalb	'вне'		»
	wegen	'из-за'		»
	statt	'вместо'		»
Д	an	'на, (в)'		предл.
	auf	'на (в)'	dem	»
	in	'в'	der	»
	hinter	'за'	den	твор.
	neben	'рядом'	einem	»
	über	'над'	einer	»
	unter	1) 'под'; 2) 'при'		1) твор.; 2) предл.
	vor	'перед'		твор.
	zwischen	'между'		»
А	aus	'из'	dem	род.
	außer	'кроме'	den	»
	bei	'при'	der	предл.
	mit	('с')	einem	твор.
	nach	1) 'по'; 2) 'после'	einer	1) дат.; 2) род.
	von	1) 'от'; 2) пассив		2) род.; 2) твор.
	zu	'к'		дат.
	gegenüber	'по отношению к'		»
В	vom	1) 'от'; 2) пассив		1) род.; 2) твор.
	zum	2) 'к'; 2) 'для'		1) дат.; 2) род.
	zur	1) 'к'; 2) 'для'		» »
	im	'в'		предл.
	am	'в (на)'		»
	beim	'при'		»
Д'	an	'к (на)'		дат.
	auf	'на (в)'	die	вин.
	in	'в'	den	»
	hinter	'за'	das	»
	neben	'рядом'	ein	»
	über	1) 'с'; 2) 'через'	eine	1) предл.; 2) вин.
	unter	'под'	einen	твор.
	vor	'перед'		вин.
	zwischen	'между'		»
С	durch	1) 'посредством' 2) 'через'		1) род. 2) вин.
	für	'для'		род.
	gegen	'против'	die	»

Таблица 5 (*продолжение*)

Группа	Предлог	Перевод	Артикль после предлога	Требуемый русский падеж
	um	1) 'о'; 2) 'вокруг'	den das	предл.
	wieder	'против'	ein	род.
	ohne	'без'	eine	»
	bis	'вплоть (до)'	einen	»
	entlang	'вдоль'		»
	ins	'в'		вин.

лить вероятность именно того перевода предложной группы, который мы задали машине; такая вероятность будет служить мерой достаточности микроконтекста для однозначного перевода предлога.[151]

Имя существительное, с которым сочетается предлог, может быть распространено за счет прилагательных, а также других частей речи. Рассмотрим отдельно вопросы, относящиеся к каждому члену предложной группы.

Предлог. Список предлогов, задаваемых машине для анализа, включает 42 наиболее употребительных немецких предлога (см. табл. 5). Каждый однозначный в данном подъязыке предлог имеет в таблице (и в «памяти» машины) единственный перевод и указание на то, каким русским падежом он управляет. Для многозначных предлогов указывается два наиболее частых значения (перевода) в данном подъязыке и указывается управление, характеризующее каждое из этих значений.

Имя существительное. В модели, предлагаемой нами, существительное предложной группы оформляется правильным грамматическим окончанием. Для этого в немецко-русском словаре-модели каждое немецкое существительное (заданное словоформой) сопровождается переводом в виде русской основы,[152] снабженной определенной грамматической информацией. Эта информация включает следующие сведения об основе:

1) род;
2) число;
3) тип основы (твердая или мягкая);

[151] Указанные микроконтексты выбраны в результате статистического анализа текстов по радиоэлектронике, проведенного А. Н. Шарандой. В нашей модели используется 176 наиболее частых предложно-именных групп. См.: А. Н. Шаранда. Статистическое выделение типовых контекстов. . .

[152] Под основой понимается часть слова, графически неизменяемая во всех его формах. Соответственно, окончанием называется часть слова, остающаяся после выделения из него основы. Ср.: О. С. Кулагина. О машинном переводе с французского языка на русский. Проблемы кибернетики, вып. 3. М., 1959, стр. 160.

Таблица 6

Типы формообразования существительных единственного числа

№№ типа	Падеж				
	род.	дат.	вин.	твор.	предл.
1	а	у	а	ом	е
2	а	у	—	ом	у
3	а	у	—	ом	е
4	а	у	а	ем	е
5	а	у	—	ем	е
6	а	ю	—	ем	е
7	я	ю	я	ем	е
8	я	ю	ь	ем	е
9	я	у	ь	ем	е
10	и	и	ь	ем	и
11	п	и	ь	ью	и
12	я	ю	я	ем	е
13	я	ю	я	ем	и
14	я	ю	й	ем	е
15	я	у	й	ем	е
16	я	ю	и	ем	и
17	а	у	о	ом	е
18	я	ю	я	ем	е
19	я	ю	е	ем	е
20	а	у	е	ем	е
21	я	ю	е	ем	и
22	ы	е	у	ой	е
23	ы	е	у	ей	е
24	и	е	у	ой	е
25	и	е	у	ей	е
26	п	е	ю	ей	е
27	и	и	ю	ей	и
28	и	и	я	ей	и
29	—	—	—	—	—

Таблица 7

Типы формообразования существительных множественного числа

№№ типа	Падеж				
	род.	дат.	вин.	твор.	предл.
1	ов	ам	ов	ами	ах
2	ов	ам	ы	ами	ах
3	ев	ам	ев	ами	ах
4	ев	ам	ы	ами	ах
5	—	ам	—	ами	ах
6	—	ам	ы	ами	ах
7	ов	ам	и	ами	ах
8	ев	ям	ев	ями	ях
9	ев	ям	и	ями	ях
10	ей	ям	ей	ями	ях

Таблица 7 (*продолжение*)

№№ типа	Падеж				
	род.	дат.	вин.	твор.	предл.
11	ей	ям	и	ями	ях
12	ей	ам	ей	ами	ах
13	ей	ам	и	ами	ах
14	ей	ям	ей	ьми	ях
15	ей	ям	ей	ями	ях
16	ей	ям	и	ьми	ях
17	—	ам	—	ами	ях
18	—	ам	и	ами	ах
19	—	ям	и	ями	ях
20	ий	ям	и	ями	ях
21	й	ям	и	ями	ях
22	ь	ям	ь	ями	ях
23	ь	ям	и	ями	ях
24	ов	ам	ов	ами	ах
25	ов	ам	а	ами	ах
26	—	ам	—	ами	ах
27	—	ам	а	ами	ах
28	ев	ям	ев	ями	ях
29	ев	ям	я	ями	ях
30	ей	ям	ей	ями	ях
31	ей	ям	я	ями	ях
32	й	ям	я	ями	ях
33	ий	ям	я	ями	ях
34	—	ам	—	ами	ях

Таблица 8

Типы формообразования прилагательных

№№ типа	Падеж				
	род.	дат.	вин.	твор.	предл.
1	ого	ому	ый	ым	ом
2	ой	ой	ую	ой	ой
3	его	ему	ий	им	ем
4	ой	ей	юю	ей	ей
5	ого	ому	ое	ым	ом
6	ых	ым	ые	ыми	ых
7	его	ему	ее	им	ем
8	их	им	ие	ыми	их
9	ых	ым	ые	ыми	ых
10	ых	ым	ие	ыми	ых
11	их	им	ие	ими	их
12	ях	им	ие	ими	их

4) тип формообразования;
5) тип синтеза основы и окончания.

Каждое существительное получает окончание в зависимости от управляющего предлога и от принадлежности существительного к тому или иному типу формообразования. Мы используем типы формообразования, предложенные в работе: З. М. Волоцкая, Т. Н. Молошная, Т. Н. Николаева. Опыт описания русского языка в его письменном виде. М., 1967. Они представлены в табл. 6 (для ед. числа) и в табл. 7 (для мн. числа).

Необходимость указывать тип синтеза вызвана тем, что в немецких научно-технических текстах широко используются сложные существительные, которые переводятся на русский язык с помощью различных языковых средств; чаще всего используются три нижеследующие схемы перевода.

1) Наиболее распространенная схема атрибутивного сочетания:

nnormalwiderstand [153] 'эталонн╳[154]---сопротивлен╳'

sspektralfunktion 'спектральн╳--функци╳'

Здесь первая немецкая основа передается по-русски с помощью прилагательного, вторая — с помощью существительного.

2) Схема словосочетаний с родительным беспредложным:

bbelichtungsbedingungen 'условн╳---освещения'
ccolectorwiderstand 'сопротивлени╳---коллектора'

3) Схема предложного словосочетания:

kkristalstrom 'ток╳---в кристалле'
rrueckstellung 'возврат╳--в -исходное положение'

В последних двух схемах окончание существительного получает первая русская основа.

Прилагательное. При решении задач морфологического синтеза целесообразно включать в класс имен прилагательных не только собственно прилагательные, но и те разряды слов, которые получают грамматическое оформление, однотипное с оформлением имен прилагательных. Сюда входят, в частности, некоторые числительные, существительные и местоимения. В нашем словаре каждое такое слово сопровождается следующей грамматической информацией:

1) признак прилагательного;
2) тип основы (твердая или мягкая);
3) тип формообразования.

[153] Напоминаем, что двойная первая буква немецкой словоформы свидетельствует о принадлежности этой словоформы к классу существительных.

[154] Знак ╳ (кто — там) условно показывает место в основе, куда должно быть присоединено окончание.

Однако известно, что выбор правильного окончания прилагательного определяется существительным, с которым согласовано прилагательное, и управляющим им предлогом.

Эта информация «отбирается» в процессе автоматического анализа у существительного (род, число) и у предлога (падеж).

При составлении таблицы формообразования прилагательных (табл. 8) учитывался лишь тип основы прилагательного.

Автоматический словарь. Под автоматическим словарем в широком смысле этого слова понимается некоторое автоматическое устройство, на вход которого наступает входной код X_k некоторого слова, а на выходе выдается необходимая для дальнейшего анализа информация Y_k об этом слове.[155]

Лингвистические вопросы организации такого словаря сводятся к тому, чтобы определить, какие единицы должны содержать входной код и какие — выходной. Существует много способов построения машинных словарей. В нашу задачу не входит разбор и сравнение этих методов. Достаточно полно эта проблема освящена в ряде работ.[156] Необходимо лишь отметить следующее. Высказывались мнения, что автоматический словарь — это просто перечень слов, построенных любым способом, и что его единственной функцией является идентификация каждой единицы входного языка. Такая точка зрения основана на не совсем верном понимании технологии организации словаря и поиска в нем. В машинные словари, строившиеся на основе соответствующих двуязычных технических словарей, включались, как правило, слова в канонической форме, т. е. так, как они даны в этих словарях. Однако опыт показал, что такой подход нерационален. Для переводов научно-технических текстов надежнее составлять машинные словари, основанные на статистическом изучении текстов соответствующего подъязыка. При этом в машинный словарь отбираются наиболее частые словоформы, встречающиеся в текстах данной тематики. Большое значение имеет также вопрос о том, как организован автоматический словарь. Практика показывает, что основная доля времени при машинном переводе с одного языка на другой уходит на поиск слов переводимого текста в словаре. Например, при переводе математических текстов с французского языка на русский около половины машинного времени уходило именно на этот поиск. Существует достаточное число различных способов организации машинного словаря. Наиболее подробно

[155] Л. А. Калужнин, Л. С. Стойкова, А. А. Стогний. О принципах построения машинных словарей. В сб.: Прикладная лингвистика и машинный перевод, Киев, 1962, стр. 7.

[156] См., например: Г. Г. Белоногов, Р. Г. Котов. Автоматизированные информационно-поисковые системы. М., 1968; стр. 32—36; К. И. Курбаков. Кодирование и поиск информации в автоматическом словаре. М. 1968; В. А. Вертель, Е. А. Вертель, В. С. Крисевич, Р. Г. Пиотровский, Л. И. Трибис. Автоматические словари в системе бинарного вероятностного перевода. В сб.: Инженерная лингвистика, Л., 1971, стр. 3—140.

этот вопрос освещен в указанной выше работе В. А. Вертель и других. В рассматриваемой задаче мы используем один из предлагаемых в названной работе способов — способ последовательного деления словаря пополам. Все коды словоформ немецкого языка располагаются во внешней «памяти» машины (на МЛ) в порядке возрастания их числовых значений, и затем осуществляется постепенное сужение области просмотра словоформ по следующему правилу. Сначала код искомой словоформы сравнивается с кодом словоформы, находящейся в середине словаря. Если произошло совпадение, то поиск считается законченным. Соответствующая русская основа и грамматическая информация к ней передаются в определенные ячейки оперативного накопителя. В противном случае, если код искомой словоформы $i^* > i_{L/2}$ (L — число словоформ в словаре, $i_{L/2}$ — адрес средней словоформы словаря), поиск следует продолжить в нижней половине словаря (от $L/2+1$ до L). Если же код искомой словоформы $i^* < i_{L/2}$, то поиск продолжается в верхней половине словаря (от 1 до $L/2-1$). Поиск словоформ в выбранной половине словаря продолжается также делением пополам. Процесс деления пополам осуществляется до момента нахождения нужной словоформы.

Максимальное время поиска при использовании этого способа поиска определяется формулой

$$T_{\max} = E(\log_2 L) + 1 \text{ [циклам]}.^{157}$$

Среднее время поиска может быть найдено по следующей формуле:

$$T_{\text{ср}} = \sum_{i=1}^{L} p_i \cdot t_i,$$

где p_i — вероятность появления (или относительная частота) i-й словоформы текста; t_i — время поиска i-й словоформы в словаре (в циклах).

Большое значение для скорости перевода имеет вопрос о способах обращения к словарю. Задача сводится к следующему. Как наиболее экономично поступить:

1) сначала найти перевод для каждой словоформы предложения, а потом перевести его?

2) сначала найти перевод для словоформ нескольких предложений, а затем последовательно переводить каждое из этих предложений?

[157] Цикл поиска — это совокупность арифметических и логических операций, необходимых для осуществления одной идентификации (одного сравнения заданного кода с кодом словоформы в машинном словаре). См.: К. И. Курбаков, ук. соч., стр. 102. $E(\log_2 L)$ означает, что берется целая часть от числа $\log_2 L$.

Особенно важен этот вопрос для автоматических словарей, хранящихся на магнитных лентах. Известно, что лентопротяжные механизмы, приводящие в движение эти ленты, работают в тысячи раз медленнее, чем МОЗУ машины. Многократное обращение к словарю на МЛ во много раз увеличивает время автоматической обработки и ведет к износу ленты.

Конечно, тот или иной способ обращения к словарю зависит от того, как построен сам словарь.

Как мы уже отмечали, словарь нашей модели упорядочен по алфавиту немецких словоформ. В процессе работы модели в оперативный накопитель машины вызывается порция текста, содержащая 40 немецких конструкций, в каждой из которых имеется предложно-именная группа. Машина автоматически выбирает из этого текста словоформы, относящиеся ко всем предложным группам порции, и располагает их в оперативном накопителе по алфавиту немецких словоформ. Получается выборочный словарь порции текста. После этого наложением выборочного словаря на автоматический словарь подъязыка, хранящийся на магнитной ленте, находятся переводы для всех словоформ выборочного словаря. Происходит наполнение выборочного словаря. После этого читается первая (вторая и т. д.) конструкция текста и для нее перевод словоформ ищется уже в наполненном выборочном словаре, находящемся в оперативном накопителе.

Отметим следующие особенности этого способа обращения к автоматическому словарю подъязыка:

1) в процессе построения выборочного словаря для порции текста в него включаются только разные словоформы всех предложных групп; этим значительно уменьшается общее число словоформ, которые необходимо искать в основном словаре подъязыка;

2) при наполнении выборочного словаря приходится лишь однажды обращаться к словарю подъязыка, хранящемуся на МЛ; таким образом, обращение к МЛ производится не для каждой словоформы конструкции, а для словоформ нескольких конструкций сразу;

3) поиск в выборочном словаре, находящемся в МОЗУ, осуществляется во много раз быстрее, чем в словаре на магнитной ленте.

И наконец, вопрос о той информации, которая должна сопровождать каждую словоформу в машинном словаре. Здесь также нет единого мнения. По существу для рассматриваемой нами модели перевода предложно-именных групп автоматический словарь и есть та система соответствий между двумя языками, о которой мы говорили выше, при описании логической схемы перевода (стр. 374). Точность перевода будет тем выше, чем точнее задана информация к каждому члену рассматриваемого синтаксического отношения. Мы уже определили ту информацию, которой

снабжается каждый член предложно-именной группы (стр. 380—383). Насколько она достаточна, покажет результат реализации модели. Только после получения первых пробных переводов можно будет определить, какая именно дополнительная информация должна быть введена в словарь в первую очередь.

При построении автоматического словаря нашей модели мы использовали идеи и методы, разработанные в группе «Статистика речи». В частности, наши тексты принадлежат к определенной тематике (немецкий подъязык электроники). Построенный путем статистического анализа текстов этого подъязыка автоматический словарь включает наиболее частые единицы его. Поскольку сложные немецкие существительные, как правило, переводятся более чем одной русской словоформой, то общее число ячеек, занимаемых под сложное слово и его перевод, превышает число ячеек, занимаемых обычной словоформой. В целях экономии объема оперативного накопителя мы выделили такие слова во вспомогательный словарь.

Таким образом, конструктивно автоматический словарь нашей первой модели состоит из трех частей:

1) основной словарь, включающий 950 немецких словоформ с их русскими эквивалентами;

2) вспомогательный словарь, включающий 108 немецких сложных слов с их переводами;

3) список наиболее частых немецких предложных групп и их перевод на русский язык.

Основной и вспомогательный словари построены так, что немецкая словоформа, соответствующая русская основа и информация к ней представляют собой одну запись, занимающую стандартное число ячеек оперативного накопителя. В основном словаре эта запись содержит 8 ячеек (4 ячейки для немецкой словоформы, 3 ячейки для русской основы и 1 ячейка для информации). Во вспомогательном словаре в каждую запись включается 6 ячеек для немецкой словоформы, 5 ячеек для русской основы и 1 ячейка для информации. Каждая частотная немецкая предложная группа занимает 4 ячейки МОЗУ. Столько же ячеек отводится для ее перевода.

Лингвистическое задание для построения модели с учетом всего вышесказанного можно сформулировать следующим образом:

1) из каждой конструкции текста выделить по формальным признакам предложную группу;

2) перевести выделенную группу слов на русский язык с расстановкой правильных окончаний у прилагательных и существительных;

3) напечатать конструкцию, предложную группу и ее перевод на русский язык;

4) построить словарь, конструкция которого минимизировала бы время поиска в нем.

17.2. Математическая постановка задачи

Сформулированная выше лингвистическая задача может быть сведена к выполнению следующих заданий.

1) Каждая немецкая словоформа основного словаря с переводом и информацией преобразуется в стандартные машинные документы $Д_1$, каждый из которых размещается в 2 m_1 ячейках; в первых m_1 ячейках размещается код немецкой словоформы (верхняя часть документа), во вторых (m_1—1) ячейках — код русского перевода (нижняя часть документа); в последней ячейке документа $Д_1$ размещается информация к русской основе (цоколь документа). Все документы $Д_1$ размещаются в зоне $АС_1$ во внешнем накопителе машины.

2) Каждое сложное немецкое слово с переводом и информацией преобразуется в стандартные документы $Д_2$, каждый из которых размещается в 2 m_2 ячейках ($m_2 > m_1$). В первых m_2 ячейках размещается сложное немецкое слово, во вторых (m_2—1) ячейках — русский перевод; в последней ячейке документа $Д_2$ размещается информация к русскому переводу. Все документы $Д_2$ размещаются во внешнем накопителе машины в зоне $С_1С_2$.

3) Каждая частотная предложная группа с переводом преобразуется в стандартные машинные документы $Д_3$, каждый из которых содержит $2m_1$ ячеек МОЗУ. В первых m_1 ячейках размещается немецкая предложная группа, во вторых m_1 ячейках — ее перевод. Все документы $Д_3$ собираются в зоне $С_2С$ внешнего запоминающего устройства. Таким образом, весь автоматический словарь модели находится во внешнем накопителе машины в зоне АС:

$$АС = АС_1 + С_1С_2 + С_2С.$$

4) Каждый из 42-х немецких предлогов с их возможными переводами на русский язык преобразуются в стандартные машинные документы $Д_4$, содержащие $m_3+n\cdot m_3+n$ ячеек ($m_3 < m_1 < m_2$). В первых m_3 ячейках размещается немецкий предлог (верхняя часть документа); в каждом из n последующих групп по m_3 ячеек размещаются соответствующие переводы предлога на русский язык (нижняя часть документа); в последних n ячейках размещают информацию к русским переводам. Все документы $Д_4$ размещаются в зоне Ф оперативного накопителя.

5) Каждое словоупотребление немецкой конструкции приводится к стандартному виду путем записи его в m_1 ячейках оперативного накопителя.

6) Из общего текста выделяется очередная конструкция путем переноса каждого нового содержимого указанных в пункте 5 ячеек в определенную зону оперативного накопителя (зону К), содержащую n групп по m_1 ячеек. Признаком конца конструкции считается занесение в зону К кода точки.

7) Последовательным сравнением содержимого зоны К с содержимым зоны Ф находится и запоминается место группы таких ячеек m_1', что первые m_3 ячейки этой группы совпадают с m_3' ячейками из зоны Ф (определение места предлога в конструкции).

8) Из зоны К в зону ПГ оперативного накопителя k выделяются группы по m_1 ячеек, начиная с найденной группы m_1' и кончая группой, в первой ячейке которых два левых кода совпадают (выделение предложной группы).

9) Последовательным сравнением содержимого групп m_1 зоны ПГ с верхней частью документов $Д_1$ и $Д_2$ из зоны $АС_2$ внешнего накопителя выделяются в зону РПГ нижние и цокольные части этих последних.

10) Нижние части этих документов дополняются необходимыми кодами окончаний путем использования содержимого цокольных частей документов $Д_1$ и $Д_2$, расположенных в зоне РПГ, а также таблиц кодов окончаний прилагательных и существительных.

11) Расшифровывается содержимое зон К, ПГ и РПГ.

17.3. Подготовка начальных данных для ввода в машину

Немецкие словоформы и соответствующие русские основы кодировались в коде М-2 на одной и той же ленте в такой последовательности: пробел — признак латыни — немецкая словоформа — пробел — признак русского текста — русская основа — пробел — признак латыни — немецкая словоформа и т. д.

Если немецкое сложное слово переводилось на русский язык двумя или более основами, то эти основы соединялись между собой черточкой. Тогда машина воспринимала две или более русские основы как одну. Место присоединения окончания к основе обозначалось знаком ⨯. В русской части словаря к сложным немецким словам после этого знака ставилось несколько тире (черточек) — по числу букв в возможном окончании. Информация к основному и вспомогательному словарям кодировалась в коде машины «Минск-22» на отдельных перфолентах. Информация занимала одну неполную ячейку. Наличие или отсутствие двоичной единицы в определенных разрядах свидетельствовало о наличии или отсутствии некоторых признаков. Таблица соответствий между разрядами, цифрами в разрядах и сообщаемыми ими сведениями приведена ниже (см. табл. 9).

Например, пользуясь этой таблицей, информацию

0020 0206 0016

к основе «интегрировани⨯» можно расшифровать так: это основа существительного (цифра 2 на третьем месте слева), среднего рода (цифра 2 на шестом месте слева), мягкая основа единственного числа (цифра 6 на восьмом месте), тип формообразования — 16.

Таблица 9

Информация об основе слова и ее коды в ячейке машины «Минск-22»

Информация	Позиция в ячейке и код											
	1	2	3	4	5	6	7	8	9	10	11	12
Сложный синтез существительных			3									
Существительное			2									
Прилагательное			4									
Именительный падеж			1									
Родительный »				4								
Дательный »				2								
Винительный »				1								
Творительный »					4							
Предложный »					2							
Мужской род					1							
Женский »						4						
Средний »						2						
Твердая основа							1					
Мягкая »								4				
Единственное число								2				
Множественное »								1				
Тип формообразования существительных												1—34

Примечание. Если разная информация принадлежит одной и той же основе, то для получения общей информации об основе соответствующие коды частных информаций складываются.

В соответствии с табл. 9 кодировалась и информация к предлогам. Сами предлоги и их переводы заносились в упорядоченном коде У-1 непосредственно в программу анализа. Наиболее частотные предложные группы также кодировались на перфоленту. В отличие от слов концом каждой такой группы служила точка.

Часть словаря, подготовленного для кодирования, представлена в табл. 10.

Таблицы окончаний существительных и прилагательных переносились на перфоленты в коде печатающего устройства. Так как окончание могло содержать от одной до четырех букв, то применялась запись окончания с фиксированного разряда ячейки. Например, окончания *-у* и *-ами*, находящиеся в ячейках 6617 и 7277, записывались так:

у

6617) 000063000000

а м и

7277) 000040545000.

Рядом с каждым окончанием записывалась та суммарная информация, которая определяет выбор именно этого окончания.

Часть немецко-русского словаря с информацией

Немецкие словоформы	Русские основы	Информация
aabnahme	уменьшени⊠	0020 0206 0025
aabschnitt	отрезк⊠	0020 1002 0003
bbandabstand	зазор⊠ - - - полосы	0020 1002 0003
chemischen	химическ⊠	0040 0004 0000
dargestellte	изложени⊠	0040 0010 0000
ddezimmalwert	десятичн⊠ - - - числ⊠	0030 0202 0021
dort	там	0000 0000 0000
eeigenschaft	свойств⊠	0020 0202 0021
einfache	прост⊠	0040 0010 0000
eingehend	подробно	0000 0000 0000
ffrequenz	частот⊠	0020 0402 0026
ffrequenzbereich	диапазон⊠ - - - частоты	0020 1002 0003
ggegenkopplung	обратн⊠ - - - связ⊠	0030 0406 0013
gezeichneten	отмеченн⊠	0040 0010 0000
ggasart	вид⊠ - - - газ⊠	0020 1002 0003
ggrenze	границ⊠	0020 0402 0027
gut	хорошо	0000 0000 0000
kkanal	канал⊠	0020 0202 0021
pphase	фаз⊠	0020 0402 0026
kkern	ядр⊠	0020 0202 0021
sseite	сторон⊠	0020 0402 0026
ttheorie	теори⊠	0020 0402 0033
vverhalten	поведени⊠	0020 0202 0025

Она является суммой информаций предлога и существительного (для синтеза основы и окончания существительного) или же предлога, части информации существительного и информации прилагательного (для синтеза основы прилагательного и окончания). Эта информация подсчитывается заранее для каждого типа формообразования и для каждого значения предлога и переносится на перфоленту в коде машины.

Текст кодировался на перфолентах порциями по 50 конструкций в каждой. Конструкция содержит не более 12 слов.

17.4. Разработка метода решения задачи и принципиальный алгоритм ее решения

Прежде чем строить модель перевода на русский язык немецких предложных групп, мы представим, как бы это задание выполнил человек. Таким информантом для нас был преподаватель кафедры фонетики немецкого языка Минского ГПИИЯ А. Н. Шаранда. После этого нами был составлен принципиальный алгоритм решения этой задачи (рис. 30, см. Приложение), который в наиболее общем виде представляет весь ход решения.

17.5. Программирование алгоритма. Решение задачи на машине

При программировании алгоритмов для ЭВМ «Минск-22» были приняты следующие значения для переменных, введенных при математической постановке задачи:

$$1)\ m_1 = 4;$$
$$2)\ m_2 = 6;$$
$$3)\ m_3 = 2.$$

При реализации этой программы необходимо лишь 50 секунд на полную обработку машиной 50-ти конструкций. Сюда входило время на ввод порции, выбор из нее всех слов, относящихся к предложным группам, наполнение выборочного словаря, выделение предложных групп, их перевод и печать результатов.

Часть результатов, выданных машиной, приведена на рис. 15.

17.6. Анализ результатов работы модели

Все ошибки, выявившиеся в процессе перевода на ЭВМ более тысячи предложных групп из различных конструкций, можно разделить на две группы:

1) ошибки лингвистического плана;

2) ошибки технические.

В свою очередь в ошибках лингвистического плана можно выделить:

а) ошибки, связанные с тем, что при переводе не учитывалась идиоматичность немецкого языка по отношению к русскому; это выражается в отсутствии словоформ, управляющих предлогами, и в отсутствии деления существительных на семантические подклассы;

б) ошибки, связанные с неправильным заданием типа основы прилагательных;

в) ошибки, связанные с омонимией падежей;

г) ошибки, связанные с тем фактом, что в словарь включалось лишь по одному значению некоторых предлогов.

Рассмотрим подробнее каждый вид ошибок.

а) Наибольшее число ошибок вызвано отсутствием слов, управляющих предлогами. Например:

№ 39 handelt es sich um einen optischen uuebergang.

Предложная группа um einen optischen uuebergang была переведена : 'вокруг оптического перехода'.

С учетом управляющей предлогом um словоформы handelt перевод был бы таким: 'речь идет об оптическом переходе'.

Еще пример:


```
575  IN DER GLEICHEN SSPALTE STEHEN .
     IN DER GLEICHEN SSPALTE
     В ОДНОЙ КОЛОНКЕ

576  IST AM AAUSGANG NICHT MEHR > .
     AM AAUSGANG
     НА ВЫХОДЕ

577  WURDEN WEGEN DER SCHLECHTEN BBEARBEITBEIT VERMESSEN .
     WEGEN DES SCHLECHTEN BBEARBEITBEIT
     ИЗ ЗА ПЛОХОЙ ОБРАБОТЫВАЕМОСТИ

578  UNTER DEM INTEGRAL RECHTS MACHT ES UNMOEGLICH .
     UN ER DEM INTEGRAL
     ПРИ ЗНАКЕ    ИНТЕГРАЛА
     ПОД ЗНАКОМ   ИНТЕГРАЛА

579  FUEHRT JEDOCH UNTERHALB DER AANFANGSSPANNUNG .
     UNTERHALB DER AANFANGSSPANNUNG
     ПОД НАЧАЛЬНЫМ   НАПРЯЖЕНИЕМ

580  ETSTEHEN FUER DIE BEIDEN PPOLARISATIONSRICHTUNGEN.
     FUER DIE BEIDEN PPOLARISATIONSRICHTUNGEN
     ДЛЯ ОБОИХ НАПРАВЛЕНИЙ    ПОЛЯР.ЗАЦИИ

581  ABWEICHUNG VON DER STATISTISCHEN VVERTEILUNG .
     VON DER STATISTISCHEN VVERTEIL NG
     ОТ(ИЗ) СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ
     ПАССИВ СТАТИСТИЧЕСКИМ РАСПРЕДЕЛЕНИЕМ

582  WURDEN AN EINER FRISCH HERGESTELLTEN PPROBE AUFGENOMMEN .
     AN EINER FRISCH HERGESTELLTEN PPROBE
     НА(В) НЕДАВНО ИЗГОТОВЛЕННОЙ ПРОБЕ

583  INFOLGE DER LLAENGE LAESST SICH KLEIN HALTEN .
     INFOLGE DER LLAENGE
     ВСЛЕДСТВИЕ ДЛИНЫ

584  DIE IN DER EEINLEITUNG BESCHRIEBENE AABHAENGIGKEIT .
     IN DER EEINLEITUNG
     В ВВЕДЕНИИ

585  DURCH EEINBRINGEN IN DIE EENTL DUNG .
     IN DIE EENTLADUNG
     В РАЗРЯД

586  WEGEN DES GROSSEN EEINFANGQUERSCHNITTES KANN MAN DEUTEN .
     WEGEN DES GROSSEN EEINFANGQUERSCHNITTES
     ИЗ ЗА БОЛЬШОГО ЗАХВАТА    СЕЧЕНИЯ

587  WIRD IN ZZUSAMMENHANG MIT GROSSER PPHOTOLEITUNG DISKUTIERT .
     IM ZZUSAMMENHANG
     В СВЯЗИ

588  IST VON DER QQUELLE ABHAENGIG .
     VON DER QQUELLE
     ИЗ ИСТОЧНИКА
```

Рис. 15. Часть результатов работы первой модели перевода предложных групп на русский язык.

№ 743 nehmen wir fuer die aaustrattsarbeit.

Предложную группу fuer die aaustratsarbeit машина перевела: 'для работы выхода'.

Если бы для предлога fuer был задан управляющий им глагол nehmen, то машина перевела бы правильно: '(принимать) за работу выхода'.

Отсутствие деления существительных на семантические подклассы вызвало также ряд ошибок следующего вида. Например:

№ 643 stimmen bei den verschiedenen aautoren nicht ueberein.

Выделенную ею предложную группу bei den verschiedenen aautoren машина перевела: 'при различных авторах'.

Если бы были выделены семантические классы существительных, то существительное «авторы» было бы включено в класс одушевленных. Тогда правильный перевод предложной группы из конструкции № 643 был бы таким: 'у различных авторов'.

б) Как сказано, мы задавали для прилагательного лишь информацию, относящуюся к типу склонения (твердое или мягкое). Реализация модели показала, что такой минимум информации недостаточен для правильного определения окончания к прилагательному. Например:

№ 54 die fuer den piesoelektrische rresonator aabhaengigkeit.

Предложная группа fuer den piesoelektrischen rresonator была переведена: 'для пьезоэлектрическего резонатора' вместо правильного 'для пьезоэлектрического. . .'

Другой пример того же типа:

№ 818 schliesst mit einer kurzen bbetrachtung.

Выделенная машиной предложная группа mit einer kurzen bbetrachtung была переведена: '(с) краткым рассмотрением' вместо правильного: '(с) кратким рассмотрением'.

Ошибки в подобных примерах объясняются существованием смешанного типа образования прилагательных, т. е. тем, что в русском языке нет противопоставлений ($к : к'$), ($г : г'$). Поэтому часть прилагательных, основа которых оканчивается на $к, г$, дает образование одной группы форм по твердому, а другой — по мягкому окончанию.

в) Типичным примером ошибки в результате омонимии падежей является следующий:

№ 432 ist x ueber den vverstraerker triggerbar.

Выделенная в этом случае предложная группа была переведена: 'над усилителем' вместо правильного: 'через усилитель'.

Этой ошибки можно было избежать, если бы проверить существительное, следующее за артиклем den, на признак числа.

Известно, что артикль den употребляется в Akkusativ существительных мужского рода единственного числа и в существительных множественного числа. Поэтому указанная выше проверка на признак числа позволила бы выбрать правильный перевод для предлога.

г) В автоматическом словаре модели для каждой немецкой словоформы давался лишь один перевод, что приводило к неточности стиля, хотя семантика слова в целом передавалась. Например:

№ 159 ist aus einem weiteren ggrunde aufschlussreich.

Здесь предложная группа aus einem weiteren ggrunde была переведена: 'из дальнейшей причины' вместо правильного: 'по следующей причине'.

Еще пример:

№ 391 betrug bei den hier wiedergegebenen und ausgewerteten aaufnahmen.

Предложную группу bei den hier wiedergegebenen und ausgewerteten aaufnahmen машина перевела: 'при здесь вновь заданных и вычисленных потреблениях'. Более точным был бы перевод: 'при здесь вновь заданных и вычисленных поглощениях'.

К ошибкам технического характера можно отнести, например, ошибки, связанные с неверным кодированием информации к русским основам. Например:

№ 315 kkristalle aus der lloesung heraus genommen.

Выделенная машиной предложная группа aus der lloesung была переведена: 'из рассмотрений' вместо правильного 'из рассмотрения'.

Причина этой ошибки состоит в том, что у основы 'рассмотрени$\boxtimes$' ошибочно задана информация множественного числа 0020 0205 0025 (вместо 0020 0206 0025). При такого рода ошибках неверное окончание получает и словоформа, стоящая перед существительным.

Другим характерным примером технической ошибки является пропуск буквы при кодировании существительного. Как мы уже не раз говорили, начальная заглавная буква существительного в нашей системе кодируется удвоением соответствующей строчной буквы. В процессе подготовки словаря и текста иногда вместо двойной строчной оператор пробивает одну букву. Тогда возникает ошибка, подобная следующей:

№ 444 vverhaelthnisse fuer den fall.

Выделенная из этой конструкции предложная группа fuer den fall была переведена: 'для (не найдено) (не найдено)'. Или еще пример:

№ 479 war fuer den hier vorliegenden fall auch erforderlich.

Здесь в роли предложной группы машиной была выделена следующая группа слов: fuer den hier vorliegenden fall auch erforderlich, которая и переведена следующим образом: 'при здесь данн (не найдено) (не найдено) (не найдено) (не найдено)'.

В обоих этих случаях при кодировании текста существительное ffall было ошибочно закодировано как fall. Не найдя существительного ffall в конструкциях № 444 и № 479, машина выделяет в «предложную группу» весь остаток конструкции, от предлога до точки. После этого она начинает искать перевод для всех словоформ «предложной группы». В результате, не найдя в словаре словоформы fall (в словаре задано ffall) и всех остальных словоформ (ибо в конструкции № 479 словоформы за fall не входят в предложную группу, а значит и в выборочный словарь), машина печатает: «(не найдено)».

При исследовании случаев употребления для перевода наиболее частых предложных групп выяснилось, что наиболее целесообразно для этой цели использовать идиоматичные в данном подъязыке словосочетания типа:

auf dem ggebiet 'в области';
bei einer ffrequenz '(с) частотой';
durch die eentladung 'в ходе разряда';
durch die wwahl 'путем выбора'.

В остальных случаях трехсловное свободное сочетание не является достаточным микроконтекстом для определения контекстного значения предлога.

Анализ ошибок позволяет сделать следующие рекомендации по усовершенствованию модели.

1) Необходим возможно более полный учет идиоматичности немецкого языка по отношению к русскому. Это можно сделать двумя способами:

а) ввести в модель список наиболее частых словоформ, управляющих предлогами;

б) использовать деление класса существительных на семантические подклассы.

2) Необходимо разработать более строгие типы формообразования для прилагательных.

3) Нужно более детально разработать вопрос о точности определения падежа для предлогов, управляющих двумя и более падежами.

§ 18. Вторая модель перевода немецких предложно-именных групп

18. 1. Лингвистические аспекты задачи

Как было видно из рассмотрения предыдущей модели, контекстное значение предлога определяется значением управляемого им существительного и значения глагола, управляющего предлогом. Выяснилось также, что для грамматически правильного перевода прилагательных недостаточно той информации, которой сопровождалось каждое прилагательное в предыдущей модели. Нерациональным оказалось и использование для перевода наиболее частых предложных групп конкретного подъязыка. Наконец, не были учтены некоторые сведения, связанные с омонимией падежей.

Строя вторую модель перевода предложных групп, мы, с одной стороны, вводим в нее эту новую информацию, а с другой стороны, налагаем на нее более жесткие условия. Это выражается, в частности, в том, что машине для выбора задаются столько значений предлогов, сколько он их имеет в конкретном подъязыке.

Обобщая все вышесказанное, можно выделить следующие условия, которые мы налагаем на вторую модель:

1) между предлогом и следующим за ним именем существительным имеет место несвободная связь; значение предлога определяется семантическим подклассом управляемого предлогом существительного, с одной стороны, и значением глагола, управляющего предлогом, с другой;

2) каждый из предлогов получает в процессе автоматического анализа столько значений, сколько он их имеет в соответствующем подъязыке.

Рассмотрим отдельно каждый член расширенной предложной группы и ту информацию, которой он снабжается во второй модели.

П р е д л о г. Здесь используются 6 наиболее многозначных и частотных предлогов немецкого языка: über, nach, zu (zum, zur), um, bei (beim), vor. В таблице приведена зависимость значений этих предлогов от семантического подкласса существительных (табл. 11).

И м я с у щ е с т в и т е л ь н о е. В дополнение к информации, данной существительному в предыдущей модели, во второй модели вводится деление класса существительных на семантические подклассы. При этом выделяются следующие подклассы:

1) конкретное существительное;
2) существительное с временным значением;
3) существительное — имя собственное;
4) отвлеченное существительное;
5) одушевленное существительное;
6) прочие существительные.

Таблица 11

6 предлогов немецкого языка и их перевод на русский

Предлог	Класс существительных: конкретные существительные	существительные с временным значением	существительные — имена собственные	отвлеченные существительные	одушевленные существительные	прочие существительные	наиболее частый перевод
bei	'в' ('на')	—	'над',	—	'у'	'при',	'за',
nach	'по',	'после',	'в',	'после'	—	'по',	'к',
um	'вокруг',	'в' ('на')	'под'	—	—	'вокруг',	'о', 'на',
über	'над' (дат.),	'свыше'	—	—	—	'через'	'за',
	'через' (вин.)						'о', 'на', 'над'
vor	'перед'	'тому назад'	'над',	'от'	'перед'	'перед',	'от', 'перед',
zu	'к'	'в'	'в'	—	—	'для'	'к', 'в', 'на'

Таблица 12

Типы формообразования прилагательных во множественном числе

№№ типа	Падеж: им.	род.	дат.	вин.	твор.	предл.	Пример
1	ые	ых	ым	ые	ыми	ых	новые
2	ие	их	им	ие	ими	их	горячие

Имя прилагательное. Что касается информации к прилагательному, здесь не используются сведения, относящиеся к типу основы (твердая или мягкая). Используемые во второй модели типы формообразования прилагательных построены отдельно для единственного числа всех трех родов и для множественного числа (см. табл. 12—15).

Таблица 13

Типы формообразования прилагательных мужского рода в единственном числе

№№ типа	Падеж: им.	род.	дат.	вин.	твор.	предл.	Пример
1	ый	ого	ому	ый/ого	ым	ом	новый
2	ый	его	ему	ый/его	ым	ем	куцый
3	ий	его	ему	ий/его	им	ем	синий
4	ий	ого	ому	ий/ого	им	ом	короткий
5	ой	ого	ому	ой/ого	ым	ом	прямой
6	ой	ого	ому	ой/ого	им	ом	большой
7	ой	его	ему	й/его	им	ем	мой

Таблица 14

Типы формообразования прилагательных женского рода в единственном числе

№№ типа	Падеж: им.	род.	дат.	вин.	твор.	предл.	Пример
1	ая	ой	ой	ую	ой/ою	ой	новая
2	ая	ей	ей	ую	ей	ей	горячая
3	яя	ей	ей	юю	ей	ей	синяя

Таблица 15

Типы формообразования прилагательных среднего рода в единственном числе

№№ типа	Падеж: им.	род.	дат.	вин.	твор.	предл.	Пример
1	ое	ого	ому	ое	ым	ом	которое
2	ое	ого	ому	ое	им	ом	сухое
3	ее	его	ему	ее	им	ем	синее
4	ее	его	ему	ее	ым	ем	куцее

Таблица 16

Часть управляющих словоформ к предлогам

Левая управляющая словоформа			Предлог	Правая управляющая словоформа		
немецкая словоформа	перевод словоформы и предлога	требуемый русский падеж		немецкая словоформа	перевод словоформы и предлога	требуемый русский падеж
verfügten	'располагали'	твор.	über	siegen	'победить'	род.
sagt	'говорит о'	предл.		herfallen	'напасть на'	вин.
diskussion	'дискуссия о'	»		urteilen	'судить о'	предл.
ergebnissen	'результаты'	род.		llarm	'от шума'	—
kkapitel	'глава о'	предл.		pplan	'сверх плана'	—
sammeln	'собирать'	вин.	nach	suchen	'искать'	вин.
fragten	'спрашивали о'	предл.		gefragt	'спрашивают о'	предл.
steigt	'растет к'	дат.		hhause	'домой'	—
nnamen	'по имени'	—		aussen	'наружу'	—
gelangen	'приходят к'	дат.	zu	beitragen	'прибавлять к'	дат.
führen	'приводят к'	»		kommt	'приходит к'	»
proportional	'пропорционально'	»		bbeispiel	'например'	—
bbeitrag	'вклад в'	вин.		nnachteil	'в ущерб'	дат.
handelt	'речь идет о'	предл.	um	handelt	'речь идет о'	предл.
spielten	'играли на'	вин.		gestiegen	'вырос на'	род.
vermindert	'уменьшению на'	»		gefragt	'спросили'	»
schwankt	'колеблется на'	род.		zzeit	'в это время'	—
			bei	gefasst	'схвачен за'	вин.
				aarbeit	'быть занятым'	—
				kkräften	'быть в силах'	—
				mmehrzahl	'в большинстве'	—
warnte	'предостерегам от'	род.	vor	bewahrt	'сохраняет от'	род.
schützt	'защищает от'	»		fürchten	'бояться'	»
wie	'по-прежнему'	—		fliehen	'бежать от'	»
fürchteten	'боялись'	род.		kurzem	'недавно'	—
fürcht	'страх перед'	твор.		aanker	'стать на якорь'	—

Список глаголов, управляющих предлогами. Сюда включено 110 наиболее частотных немецких глагольных словоформ с их переводами (в виде русских словоформ). Для каждого предлога в отдельности даны в этом списке глаголы, которые могут стоять как в левой, так и в правой позиции по отношению к предлогу. Вполне естественно, что такой список не может содержать всех управляющих глаголов, которые возможны для взятых предлогов. Выбор их для нашей модели производился путем статистического анализа немецких текстов по радиоэлектронике. Наряду с информацией о значении управляющего глагола (а в некоторых случаях и значении предлога) в списке приводится информация о падеже, которым управляет соответствующий предлог. Часть списка управляющих глаголов приведена в табл. 16.

Автоматический словарь. При построении второй модели перевода предложных групп целиком использован автоматический словарь предыдущей модели. Разница заключается лишь в том, что в описываемой модели место наиболее частых предложных групп занимает список управляющих глаголов. Лингвистическое задание для построения второй модели перевода немецких предложно-именных групп на русский язык формулируется так же, как и для первой модели (см. стр. 376—387).

18.2. Математическая постановка задачи

Математические задания для построения второй модели совпадают с такими же заданиями для первой модели (см. стр. 388—389), за исключением следующих.

3) Каждый управляющий предлогом глагол с переводом и информацией преобразуется в стандартные машинные документы $Д_1$, из которых каждый размещается в $2m_1$ ячейках оперативного накопителя. В первых m_1 ячейках размещается код немецкой словоформы, во вторых ($m_1 - 1$) — русский перевод; в последней ячейке этой группы размещается информация о падеже, которым управляет предлог. Все документы $Д_1$ собираются в зоне $С_2С$ внешнего накопителя.

4) Каждый из девяти предлогов преобразуется к стандартному виду путем записи его в m_3 ячейках оперативного накопителя. Все коды предлогов собираются в зоне $Ф_1$ оперативного накопителя. Возможные переводы предлогов преобразуются к такому же стандартному виду, как и немецкие предлоги, и записываются в зоне $Ф_2$ МОЗУ. Соответствующая информация к русским значениям немецких предлогов собирается в зоне $Ф_3$ оперативного накопителя.

5) В формулировке п. 7 предыдущей модели зона Ф меняется на зону $Ф_1$.

18.3. Подготовка исходных данных для ввода в ЭВМ

Основной и вспомогательный словари, а также словарь управляющих глаголов кодировались на перфоленте так же, как это было изложено в пункте 17.3. предыдущей задачи см. стр. 389—391).

На отдельную перфоленту в коде машины перфорировалась информация к управляющим глаголам.

Информация о принадлежности существительного к тому или иному семантическому подклассу кодировалась цифрой, занимающей десятый восьмеричный разряд (считая слева направо) ячейки, несущей всю информацию к существительному. Различным значениям этой цифры соответствовали следующие семантические подклассы существительных:

Значение цифры	Семантический подкласс
1	— конкретное существительное
2	— существительное с временным значением
3	— существительное — имя собственное
4	— отвлеченное существительное
5	— одушевленное существительное

Каждое прилагательное также снабжается информацией о типе формообразования. Эти данные находятся в ячейке, несущей всю информацию для прилагательного, и занимают в ней следующие разряды:

Информация о типе формообразования	Разряд ячейки
во множественном числе —	9
в единственном числе среднего рода —	10
» » » женского » —	11
» » » мужского » —	12

Например, информация к основе *французск*▒ записывается в словаре так:

0040 0000 2214

Расшифровывается это число следующим образом: основа *французск*▒ является основой прилагательного (цифра 4 на третьем месте слева); при существительном во множественном числе эта основа получает окончание по 2-му типу формообразования (цифра 2 на девятом месте); при существительном среднего рода единственного числа основа получает окончание по 2-му типу формообразования (цифра 2 на десятом месте) и т. д. Вся остальная информация, необходимая для построения второй модели, подготавливается и кодируется так же, как и для первой модели (стр. 389—391).

Основной словарь содержит 970 немецких словоформ с их переводами; во вспомогательный словарь включено 89 сложных слов с их переводами.

Текст кодировался и вводился порциями по 40 конструкций в каждой.

18.4. Принципиальный алгоритм решения задачи

Принципиальный алгоритм решения этой задачи приведен на рис. 31 (см. Приложение). Как видно из сравнения этого алгоритма с принципиальным алгоритмом предыдущей задачи (рис. 30, см. Приложение), они имеют очень много общего. И это вполне естественно. По существу это две модели одного и того же языкового поведения человека при различных исходных данных.

18.5. Программирование алгоритма и решение задачи на машине

Программа по данному алгоритму составлена для ЭВМ «Минск-22».

При реализации на машине общее время обработки порции в 40 конструкций составило 1 мин. 50 сек. Часть результатов работы модели приведена на рис. 16.

18.6. Анализ результатов работы второй модели

Ошибки лингвистического характера можно подразделить на две группы:

1) связанные с неверным или недостаточно полном заданием лингвистической информации;

2) вызванные принципиальным отличием переводческой «работы» человека от тех логических операций, которые выполняет машина.

Рассмотрим первую группу ошибок.

Мы уже говорили, что в список управляющих словоформ были включены наиболее частые глаголы подъязыка радиоэлектроники. Вполне естественно, что они не могли охватить всех текстов этого подъязыка. Для некоторых конструкций управляющих глаголов в списке не оказалось. Тогда машина выдавала неверный перевод. Например:

№ 207 hat nach langen bbestrahlungszeiten hingewiesen.

Здесь предложная группа nach langen bbestrahlungszeiten была переведена: 'после длительных периодов облучения' вместо правильного: 'на длительные периоды облучения'.

Ошибка объясняется тем, что в списке управляющих слов к предлогу nach не была задана словоформа hingewiesen с жестким управлением: hinwiesen nach, указывающим перевод 'на'.

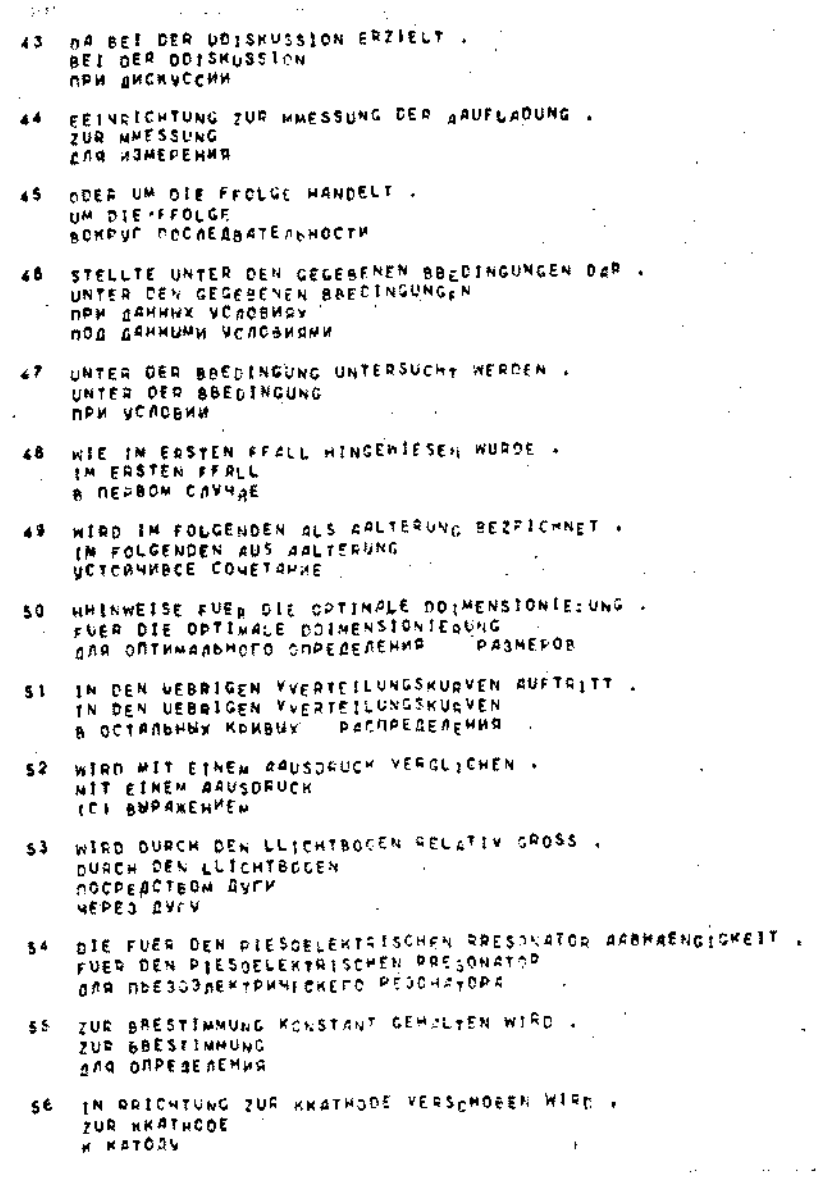


Рис. 16. Часть результатов работы второй модели перевода немецких предложных групп.

В первой нашей модели мы задавали список наиболее частых предложных групп. Их введение позволило нам избежать тех ошибок, которые связаны с идиоматичностью немецкого языка по отношению к русскому. Это выражается, в частности, в учете идиоматических выражений типа auf dem ggebiet (см. стр. 396). Во второй модели мы не задали машине ни списка частотных триад (задание всего списка нецелесообразно; ср. стр. 396), ни списка устойчивых идиоматических словосочетаний. Это привело к появлению ошибок, подобных следующим:

№ 574 zum gglueck schenkt man ein nnufeisen.

Выделенная машиной предложная группа zum gglueck была переведена: 'для счастья' вместо правильного 'на счастье'.

Другой пример:

№ 658 zur vvermeidung vor sstoerungen musste abgekuehlt werden.

Предложная группа zur vvermeidung была переведена неверно: 'для избежания'; правильный перевод: 'во избежание'.

В обоих приведенных примерах ошибки объясняются отсутствием специального списка устойчивых идиоматических сочетаний, куда бы вошли и словосочетания из конструкций № 574 и № 658: zum gglueck, zur vvermeidung и т. п. Наличие таких примеров при автоматическом переводе может служить базой для формального определения устойчивых сочетаний в том или ином языке: если нет словоформ, управляющих предлогами, и если, при детальном учете всех факторов, влияющих на перевод предлогов, машина все же выдает на выходном языке неправильный перевод предлога, то предложно-именная группа в таком случае может считаться устойчивым сочетанием.

Ряд ошибок связан с недостаточностью того деления существительных на семантические подклассы, которое было задано машине. Особенно это относится к таким подклассам, как «конкретное существительное» и «существительное с временным значением». Необходимо выделить в этих подклассах еще более частные смысловые группы и ввести их коды в машинные алгоритмы. Так, в подклассе «существительное с временным значением» можно было бы выделить такие подгруппы, как «существительные, выражающие временну́ю длительность», «существительные, выражающие предшествование», «существительные, выражающие перфективность», «существительные, выражающие предельность» и т. д. Например, за основу более детальной классификации семантических подклассов можно было бы принять 16 семантико-дистрибуционных разрядов существительных, предложенных в работе: Л. Н. Я р в. Проблема сочетаемости существительных с зависимыми словами в современном немецком языке. АКД, Л., 1967, стр. 12—13.

Ввиду отсутствия в нашей модели подобной семантической классификации мы вынуждены давать двойной перевод предлога. Например:

№ 16 und werden bei den anderen pproben gemessen.

Предложная группа bei den anderen pproben переводилась с двойным значением предлога: 'в (на) других пробах'.

Неверное отнесение существительного к тому или иному семантическому подклассу также приводило к ошибкам. Например:

№ 364 ist in ddiagrammen nach spaeteren zzeiten verschoben.

Выделенная машиной предложная группа nach spaeteren zzeiten была переведена ею: 'после более поздних времен' вместо несколько более правильного 'через более поздние времена'.

Анализ этой ошибки показал следующее: в информации к существительному zzeiten указано, что это — существительное с временным значением, в то время как в действительности это абстрактное существительное.

Рассмотрим лингвистические ошибки второй группы. Во второй модели поиск соловоформ, управляющих предлогом, проводился на один шаг влево от предлога и вправо — до точки. Машина, не обладая эвристическими способностями, точно выполняла эти условия. В результате происходили ошибки, подобные следующей:

№ 12 verschiebt sich von Z nach XX[158]

Из этой конструкции была выделена предложная группа nach XX и получен перевод: 'согласно XX' вместо: 'к XX'.

Здесь машина, не обнаружив в словоформе von перед предлогом управляющей словоформы, перешла к поиску правых управляющих словоформ и нашла среди них словоформу XX, которая входит к предложную группу и не является управляющей к предлогу в данной конструкции. Но эта словоформа есть в списке управляющих слов к предлогу nach со значением 'согласно'. Это значение было взято машиной и напечатано. На самом деле управляющей словоформой для предлога nach в рассматриваемой конструкции является словоформа von со значением 'к'. Но машина не нашла ее, так как анализ слева был ограничен лишь одной словоформой. Вероятно, человек в процессе анализа такой конструкции не ограничивается «одноразовым» просмотром слева и справа, а, проведя анализ справа, еще раз возвращается влево, чтобы убедиться в истинности своего вывода. Возможно, также, что подобный переход производится многократно.

[158] Через XX обозначалось в тексте любое имя собственное.

Ограниченность машинного «мышления» видна и на таком примере:

№ 245 der zu jeder zzeit sagen kann.

Предложная группа zu jeder zzeit была переведена: 'в настоящее время' вместо правильного: 'в любое время'.

Словоформа zzeit находится в списке правых управляющих словоформ со значением 'в настоящее время'. В таком значении она употребляется лишь непосредственно за предлогом zu (zur). В рассматриваемой конструкции zzeit является членом свободного словосочетания, но так как поиск управляющих слов производился вправо от предлога до точки, то оно попало в число управляющих, что и дало ошибочный результат. Ограничить же поиск управляющих слов справа также одним шагом (как поиск слева) нельзя, ибо существует достаточное число управляющих слов, которые могут стоять справа за предлогом более чем на один шаг.

Технические ошибки второй модели в основном совпадают с техническими ошибками первой модели. Новыми оказались ошибки, связанные с ошибочным кодированием семантического подкласса существительных, а также ошибки в информации к словоформам, управляющим предлогами. Первые дают неправильный выбор значения предлога, вторые влияют лишь на грамматическое оформление существительных и прилагательных окончанием.

Таким образом, процесс усовершенствования первой модели привел к вполне удовлетворительным результатам. Из общего числа в 1234 конструкции с предложными группами лишь в 60 случаях были обнаружены ошибки (и лингвистические и технические), т. е. точность работы второй модели составляет 95%. Результат реализации этой модели еще раз свидетельствует в пользу того, что перевод с использованием вероятностных (частотных) методов вполне возможен.

§ 19. От элементарных речевых моделей к языку лингвистического программирования

Как видно из программ, реализующих разработанные алгоритмы, они включают большое число команд конкретной машины. Такой переход от задачи, поставленной на естественном языке, к комбинациям команд требует значительного времени и знаний соответствующего «языка машины». Поэтому иногда необходимость такого ручного препарирования задач сводит на нет выгоды, получаемые от их быстрого решения на машине.[159] Одним из способов, позволяющих избежать ручного программирования, явились специальные алгоритмические языки (см. выше, стр. 289).

[159] Д. Ю. Панов. О взаимодействии человека и машины, стр. 41.

Алгоритмы, записанные на таком языке, переводятся в язык машинных команд самой машиной. Но и этот способ общения человека и машины остается достаточно сложным. До ввода в машину алгоритм задачи должен быть написан на соответствующем алгоритмическом языке, который требует специального изучения. Вполне естественно возникает вопрос: а нельзя разработать такую методику, которая позволяла бы вводить задачу в машину на естественном языке? Каким образом, не изучая всей системы команд и особенностей конкретной машины или специальных алгоритмических языков, можно на ЭВМ решить задачу, написанную на естественном языке? Ниже предлагается один из возможных подходов к решению этой проблемы.

При рассмотрении алгоритмов, представленных в главах I, II и III нашей статьи, нетрудно заметить, что некоторые отдельные подалгоритмы повторяются во всех моделях. Например, такие подалгоритмы, как «Ввод текста», «Чтение словоупотребления», «Формирование конструкций», «Сжатие конструкций», являются общими для всех моделей. Такие подалгоритмы, как «Выделение окончания», «Поиск единицы текста в неупорядоченном массиве единиц», «Синтез слова», являются общими для частей II и III.

Подалгоритмы, повторяющиеся в речевых моделях разного типа, назовем элементарными речевыми моделями (ЭЛРМ). При выполнении данной работы нами были разработаны следующие элементарные речевые модели:

1) Ввод текста в машину — ЭЛРМ-1
2) Чтение словоупотребления — ЭЛРМ-2
3) Выделение окончания словоформы — ЭЛРМ-3
4) Выбрасывание инфиксов — ЭЛРМ-4
5) Выделение основы слова — ЭЛРМ-5
6) Синтез основы слова и окончания — ЭЛРМ-6
7) Формирование конструкции — ЭЛРМ-7
8) Сжатие фазы — ЭЛРМ-8
9) Определение места присоединения окончания к основе — ЭЛРМ-9
10) Инверсия словоформы — ЭЛРМ-10
11) Организация машинного документа и свод документов в массив — ЭЛРМ-11
12) Запись массива документов на МЛ — ЭЛРМ-12
13) Поиск группы единиц текста в неупорядоченном массиве единиц — ЭЛРМ-13
14) Поиск словоформы в двуязычном словаре — ЭЛРМ-14
15) Упорядочение двуязычного словаря — ЭЛРМ-15
16) Перевод словоформ и сочетаний словоформ из кода машина в код печатающего устройства — ЭЛРМ-16
17) Вывод текстовой информации на устройство печати — ЭЛРМ-17

Как показал опыт работы группы «Статистика речи», такой набор элементарных речевых моделей является тем минимумом, который позволяет решать задачи по автоматической обработке лингвистической информации начиная от всевозможных статистических исследований и кончая автоматическим переводом.

Конечно, этот набор ЭЛРМ нельзя считать полным и окончательным. В процессе решения новых задач появится необходимость в разработке других ЭЛРМ, дополняющих имеющийся набор.

Каким же образом может лингвист использовать эти элементарные речевые модели? Говоря о знаниях, необходимых современному ученому, занимающемуся бионикой, известный американский нейрофизиолог У. Мак-Каллок писал: «Для того чтобы заниматься бионикой, необходимо самому достаточно хорошо ориентироваться и в технике, и в биологии, так как эти знания принесут гораздо больше пользы, если они сосредоточены в одной голове, а не просто в одной комнате».[160] То же самое можно сказать и об ученом, занимающемся проблемами инженерной лингвистики. Как нам кажется, представленный выше набор ЭЛРМ позволит в какой-то степени приблизить лингвиста к программисту.

Существует много различных методов записи алгоритмов при автоматической обработке лингвистической информации.[161] Однако все они предназначены в основном для специалистов, занимающихся проблемами автоматического перевода. Сейчас нужны такие правила, которые могли быть поняты любым лингвистом, ибо начавшееся широкое внедрение ЭВМ в различные области знаний непременно коснется и лингвистики. Машины помогут ускорению процесса научного лингвистического исследования. Они дадут возможность для объективной проверки выдвигаемых научных гипотез.

Мы предлагаем способ использования набора ЭЛРМ, основанный на знании лингвистом минимального числа простых команд конкретной машины. Любому человеку не трудно усвоить 5—10 команд любой машины. Тогда, зная эти команды и методику использования набора ЭЛРМ, сам лингвист может составлять программы для своих задач. При вводе в машину составленной лингвистом программы каждая ЭЛРМ в процессе работы автоматически «развертывается», выполняя то действие, для которого она предназначена.

---

[160] У. Мак-Каллок. Подражание одних форм жизни другим формам — биомиметис. В сб.: Проблемы бионики, М., 1965.

[161] См., например: И. А. Мельчук. О стандартных операторах для алгоритмов автоматического анализа русского научного текста. МП., 1961; О. С. Кулагина. Операторное описание алгоритмов перевода. В сб.: Машинный перевод и прикладная лингвистика, вып. 2 (9), М., 1959; В. А. Воронин. Операционная запись алгоритмов для МП. МП, 1961.

Таблица 17

Ряд команд ЭВМ «Минск-22»

Название команды	Код	Запись команды	Содержание команды
1) Сравнение	05	05 00 A_1 A_2	содержимое ячейки с адресом A_1 сравнивается с содержимым ячейки A_2, результат сравнения запоминается
2) Условная передача управления по нулю	—34	—34 00 A_1 A_2	если результат предыдущего действия равен нулю, то управление передается команде с адресом A_2, если не равен нулю, — то команде A_1
3) Условная передача управления по знаку	—32	—32 00 A_1 A_2	если результат предыдущего действия больше нуля, то управление передается команде A_1; если меньше нуля, — то команде A_2
4) Безусловная передача управления	—30	—30 00 A_1 A_2	передать управление команде с адресом A_1 и результат предыдущего действия записать по адресу A_2
5) Обращение к ЭЛРМ	—31	—31 00 A_1 A_2	обратиться к ЭЛРМ, начинающейся с адреса A_1 и заканчивающейся адресом A_2
6) Пересыл информации	—10	—10 00 A_1 A_2	содержимое адреса A_1 переслать в адрес A_2; если A_1=(000), то этой командой очищается ячейка с адресом A_2
7) Сложение	10	10 00 A_1 A_2	к содержимому ячейки с адресом A_2 прибавить содержимое ячейки с адресом A_1, результат оставить в ячейке A_2 и запомнить
8) Вычитание	20	20 00 A_1 A_2	из содержимого ячейки с адресом A_2 вычесть содержимое ячейки с адресом A_1, результат направить в ячейку с адресом A_2 и запомнить
9) Останов машины	—00	—00 0 0 0	останов машины по окончании работы программы

Для конкретной задачи, решаемой с использованием ЭЛРМ, задается таблица информации. В ней, в частности, должны быть указаны:

1) адрес A_1, начиная с которого в ЭВМ будет вводиться текст,

2) начальный адрес A_2 m ячеек, в которых будет читаться словоупотребление,

3) начальный адрес A_3 n групп по m ячеек, в которых будет записываться исследуемая конструкция,

4) адрес ячейки A_4 — признак, найдена ли словоформа при поиске;

5) адрес ячейки A_5 — счетчик словоформ (или конструкций) в порции;

6) ряд других адресов, связывающих между собой элементарные речевые модели.

Рассмотрим конкретную задачу и метод ее решения с использованием набора ЭЛРМ. Пусть задача формулируется следующим образом. Из английского текста, содержащего N словоупотреблений, выбрать конструкции, содержащие словоформы one и to. Конструкция должна содержать все предложение от точки до точки. Считаем, что лингвист знает следующие команды ЭВМ «Минск-22» (табл. 17). Задаем таблицу информации для данной задачи. Считаем, что максимальное число словоупотреблений в предложении равно 20 (n=20) и что при «чтении» словоупотребление записывается в 4-х ячейках оперативного накопителя (m=4). В этом случае таблицу информации представим следующим образом:

A_1=5000
A_2=4710
A_3=4500
A_4=4470
A_5=4452

Далее выделим еще такие адреса:

A_6=4455 — код единицы счета;
A_7=4456 — число словоупотреблений в порции;
A_8=4457 — » » во всем тексте.

Словоформы one и to разместим в ячейках A_9=4450, A_{10}=4451.

Далее, пусть каждая ЭЛРМ занимает в оперативном накопителе определенное место:

ЭЛРМ-1	4210÷4400
ЭЛРМ-2	4000÷4200
ЭЛРМ-7	3700÷3777
ЭЛРМ-13	3500÷3660
ЭЛРМ-8	3300÷3470
ЭЛРМ-16	3200÷3270
ЭЛРМ-17	3100÷3170

Считаем, что весь массив в N словоупотреблений разделен на порции, каждая из которых содержит t словоупотреблений (это число хранится в ячейке 4456).

Тогда, начиная программу решения этой задачи с адреса 0100, запишем ее следующим образом:

```
0100) — 31 00 4210 4400 — ввести текст с адреса 5000;
   1) — 31 00 4000 4200 — прочесть очередное словоупотребление текста;
   2) — 31 00 3700 3777 — передать словоупотребление в конструкцию (формирование конструкции);
   3) — 31 00 3500 3660 — поиск в конструкции слов one и to;
   4)    05 00 4470 4455 — найдено?;
   5) — 34 00 0101 0106 — найдено → к 0106, нет → к 0101;
   6) — 31 00 3200 3270 — сжатие конструкции;
   7) — 31 00 3100 3170 — печать конструкции;
0110)    05 00 4452 4456 — порция закончилась?;
   1) — 34 00 0101 0112 — закончилась → к 0112, нет → к 0101;
   2)    05 00 4453 4457 — весь текст закончился?;
   3) — 34 00 0114 0116 — закончился → к 0116, нет → к 0114;
   4) — 10 00 0000 4452 — очистка счетчика словоупотреблений в порции;
   5) — 30 00 0100 0000 — ко вводу следующей порции текста;
   6) — 00 00 0000 0000 — конец работы.
```

Как видно из этого примера, в такой постановке задача не представит затруднения и для лингвиста, далекого от общей теории программирования.[162]

Имея достаточно полный набор ЭЛРМ, можно следующим образом организовать автоматический перевод научно-технических текстов с использованием вероятностных методов.

---

[162] Сейчас набор ЭЛРМ находится в процессе расширения и совершенствования. Сотрудниками группы «Статистика речи» разработано несколько новых ЭЛРМ. А. А. Пиотровская разработала подалгоритм «Синтез-2», позволяющий синтезировать в одну словоформу приставку, основу и окончание, а также подалгоритм «Выделение приставки», позволяющей выделить из словоформы, находящейся в машине в m ячейках, любое начальное количество букв. Другой сотрудник нашей группы А. Джубанов разрабатывает ЭЛРМ, связанную с выделением определенных признаков слова (признаков, характеризующих часть речи, род, число, падеж и т. д.). С использованием приведенного выше набора ЭЛРМ построен также ряд речевых моделей, не отраженных в настоящей работе. См.: А. Джубанов, А. В. Зубов. Автоматизация некоторых лингвистических процессов. Вестник АН Казахской ССР, вып. 9, Алма-Ата. 1968; А. В. Зубов, В. И. Киселева, А. А. Пинкевич, А. А. Пиотровская. Грамматический анализ испанского текста с помощью ЭВМ «Минск-22». В сб.: Научно-теоретическая конференция [Ленинградского ГПИ им. А. И. Герцена], посвященная 50-летию Великой Октябрьской социалистической революции», [Л.], 1967.

1) Автоматическая статистическая обработка текста. Сюда входит построение частотного словаря словоформ для данного текста (а при необходимости и частотного списка n-словных сочетаний). Каждый частотный список заносится на отдельную магнитную ленту.

2) Автоматическая комплектация частотных списков, предусматривающая нахождение русских эквивалентов для каждой словоформы иноязычного текста. Эти эквиваленты извлекаются из больших автоматических двуязычных словарей соответствующих подъязыков.

3) Анализ полученного двуязычного словаря к данному тексту. Само собой разумеется, что в большинстве случаев не для всех словоформ такого словаря будут найдены эквиваленты в большом словаре. В таких случаях, пользуясь «книжными» двуязычными словарями по соответствующей специальности, дежурящий у ЭВМ лингвист дополняет словари текста, давая наиболее вероятные в данном подъязыке переводы тем словоформам и выражениям, для которых не был найден эквивалент в автоматическом словаре. Одновременно оператор и лингвист сами или с помощью простейших программ, составленных с использованием набора ЭЛРМ, ведут статистику вновь появляющихся слов и выражений, не учтенных в автоматическом словаре. Наиболее часто повторяющиеся из них в дальнейшем включаются в автоматический словарь.

4) Первый автоматический перевод текста с использованием дополненных частотных словарей текста. ЭВМ в ходе первого перевода непременно допустит ряд ошибок. Для их устранения могут быть быстро построены, с использованием набора ЭЛРМ, программы корректировки. Эти последние могут легко меняться от одного текста к другому.

5) Окончательный бинарный вероятностный перевод с использованием программ корректировки.

Конечно, осуществление указанных этапов предполагает, что на магнитных лентах уже имеются автоматические словари по соответствующим подъязыкам, а также алгоритмы перевода.

Рассмотренная последовательность этапов перевода на первый взгляд кажется сложной. Но при наличии большого коллектива лингвистов и математиков-программистов, а также современных ЭВМ, обладающих достаточно большим объемом накопителей, весь процесс перевода может быть поставлен «на конвейер». «Побочным продуктом» процесса перевода могут быть многочисленные статистические данные о подъязыках и большое число автоматических речевых моделей.

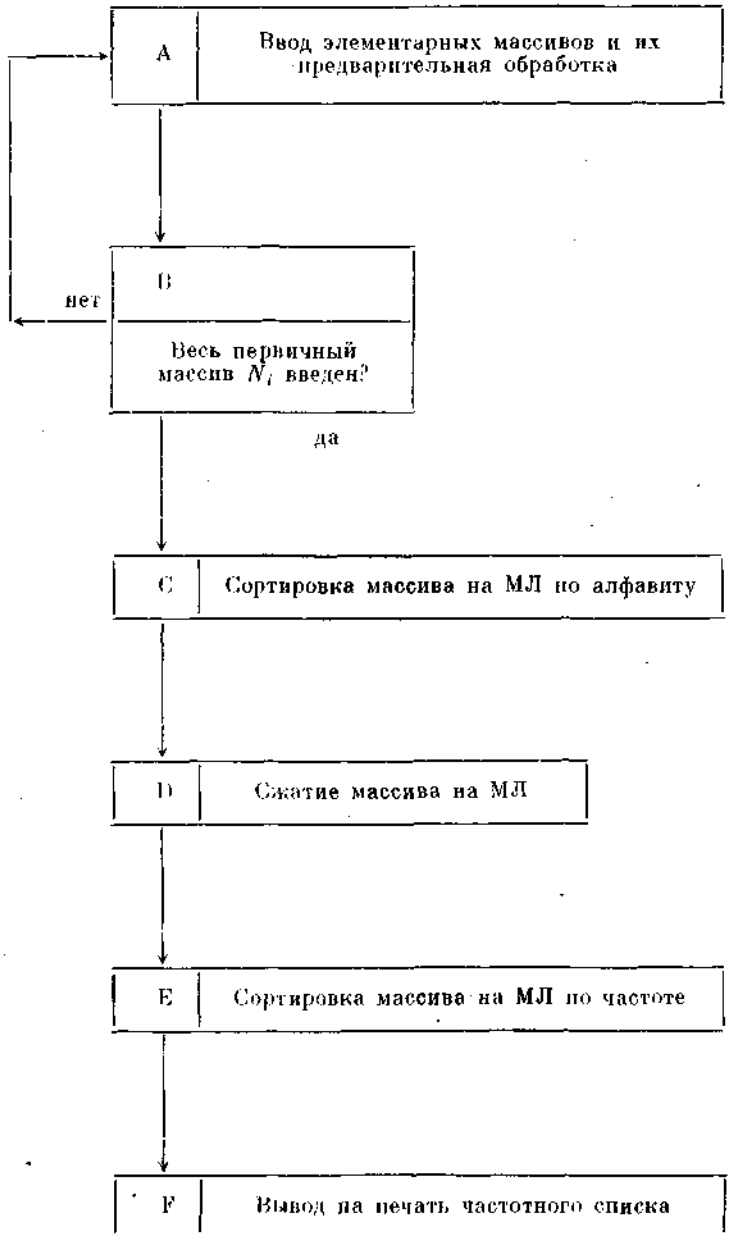


Рис. 17. Принципиальный алгоритм выделения наиболее частых n-буквенных сочетаний.

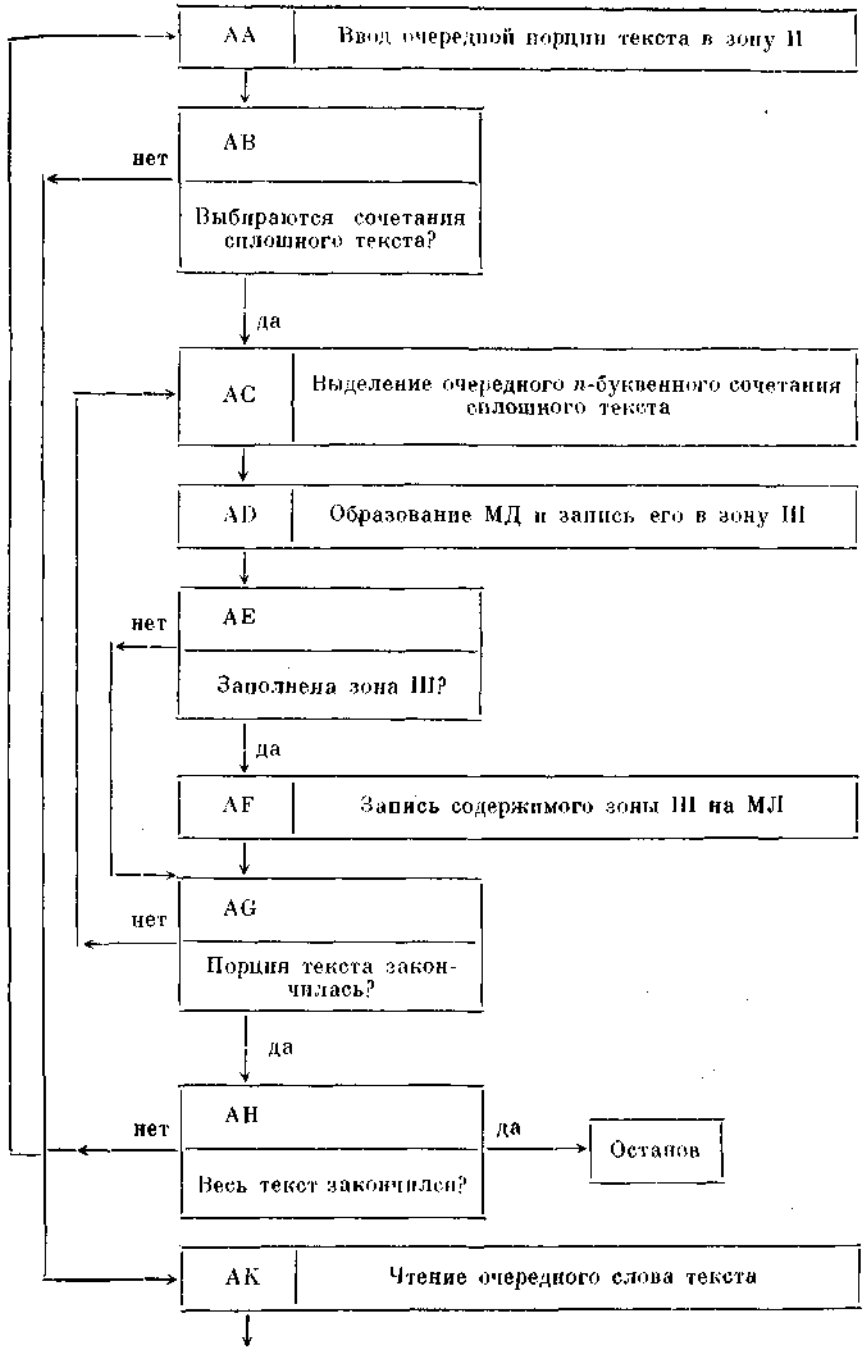


Рис. 18. Алгоритм A. Ввод элементарных массивов и их предварительная обработка (*см. продолжение*).

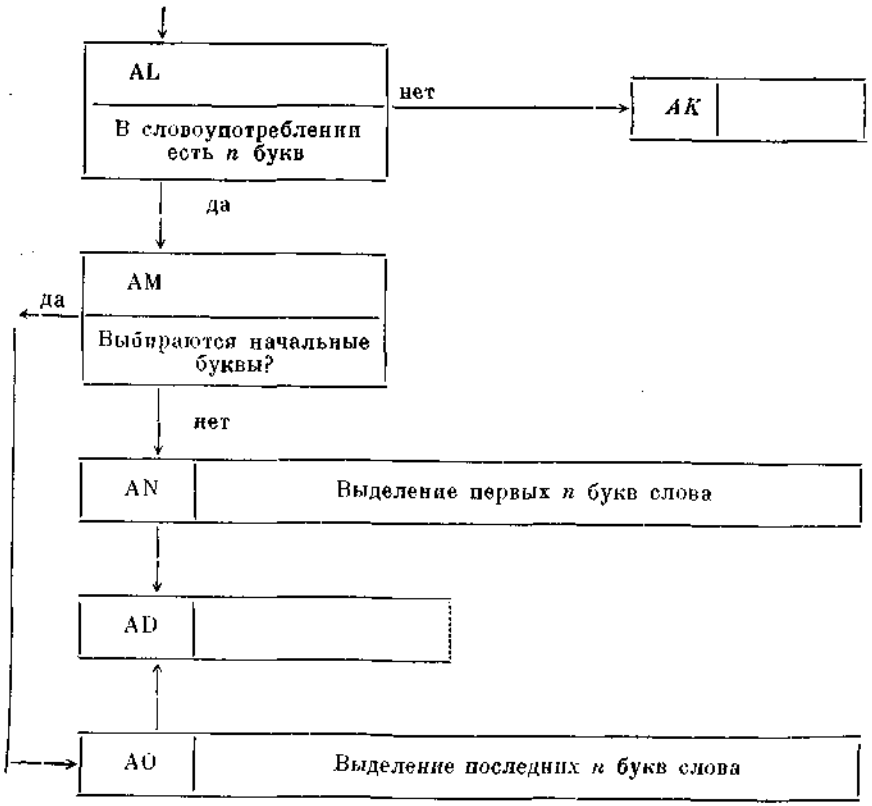


Рис. 18. (*окончание*).

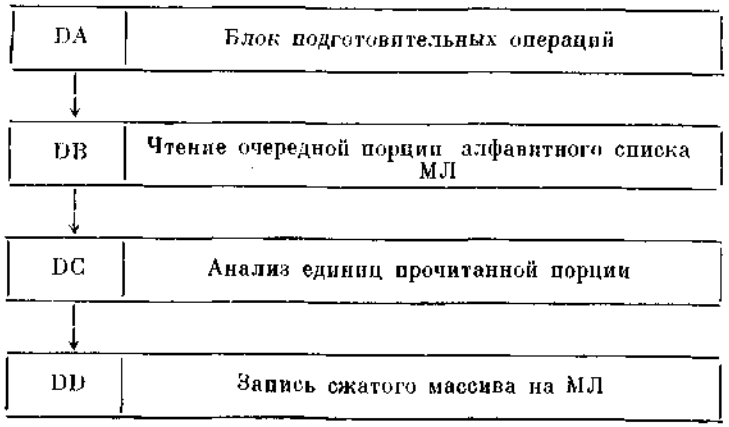


Рис. 19. Алгоритм Д. Сжатие массива информации на МЛ.

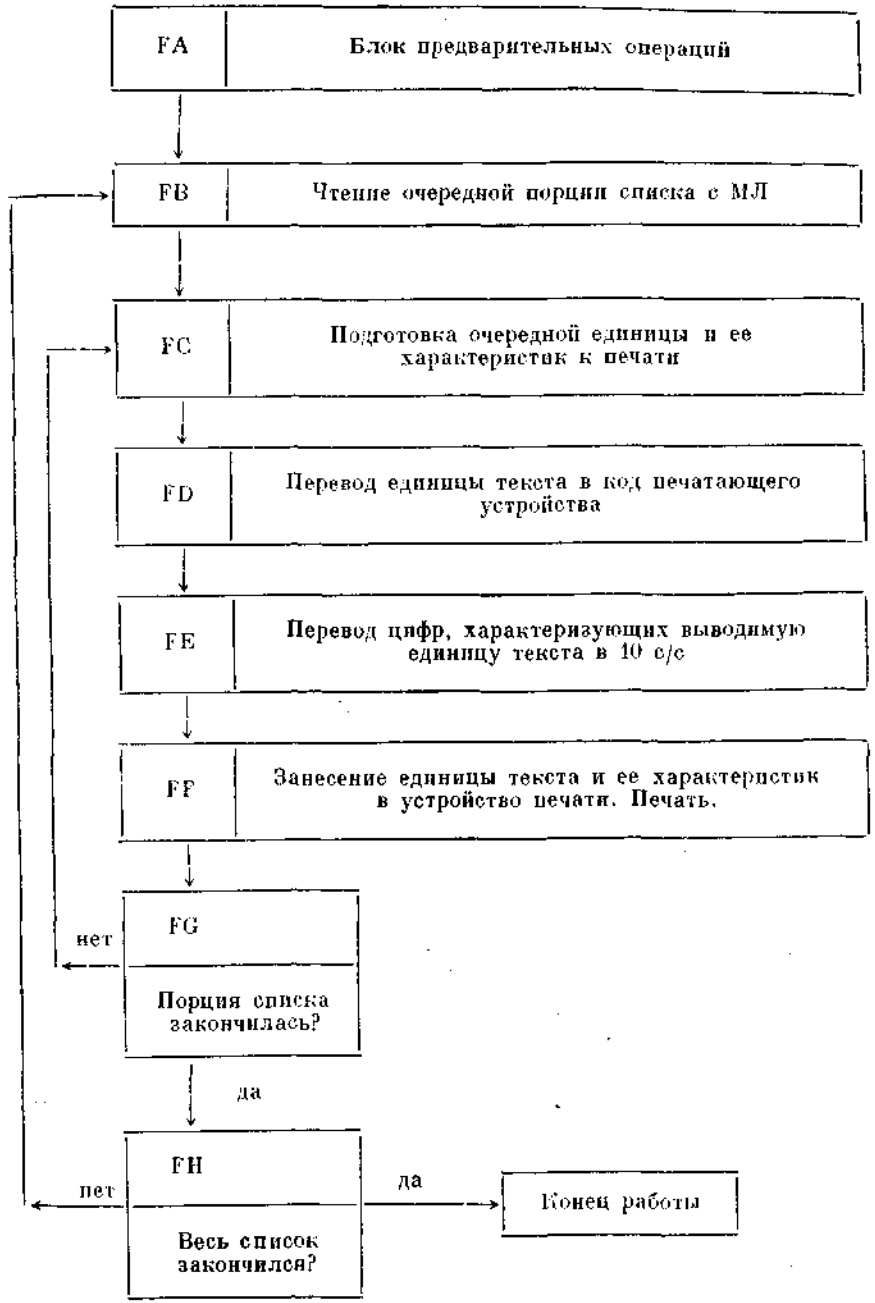


Рис. 20. Алгоритм F. Вывод на печать частотного списка.

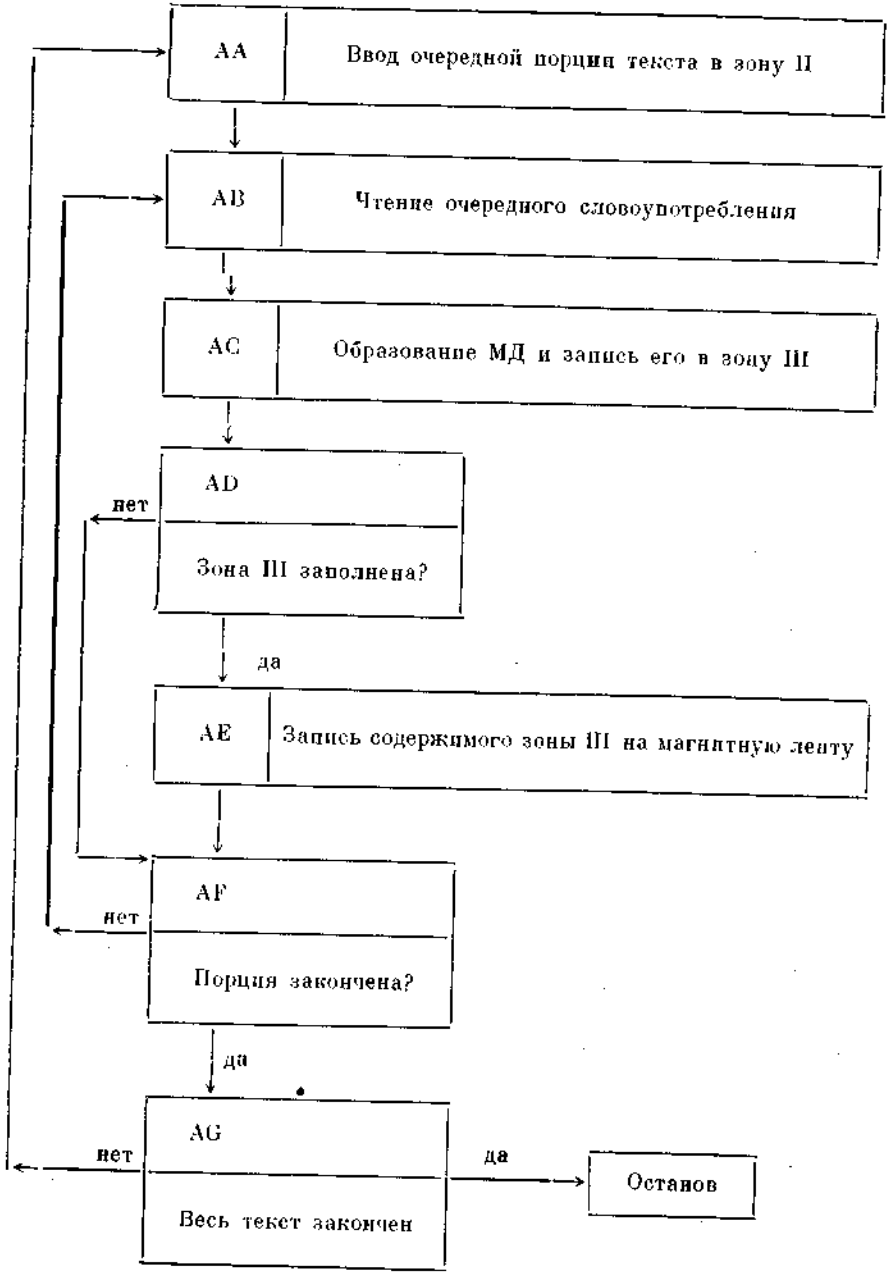


Рис. 21. Алгоритм А. Ввод элементарных массивов и их предварительная обработка.

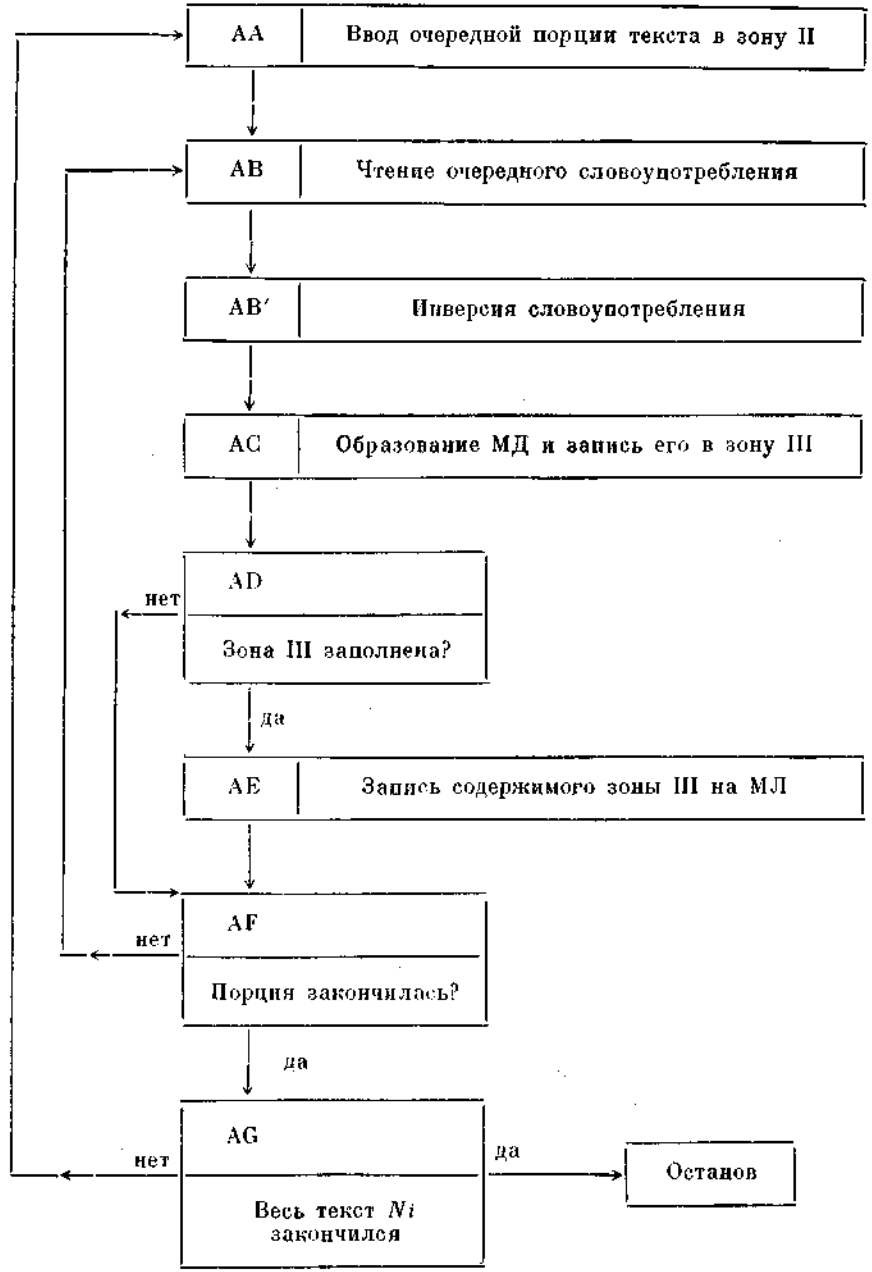


Рис. 22. Алгоритм А. Ввод элементарных массивов и их предварительная обработка (для обратного словаря).

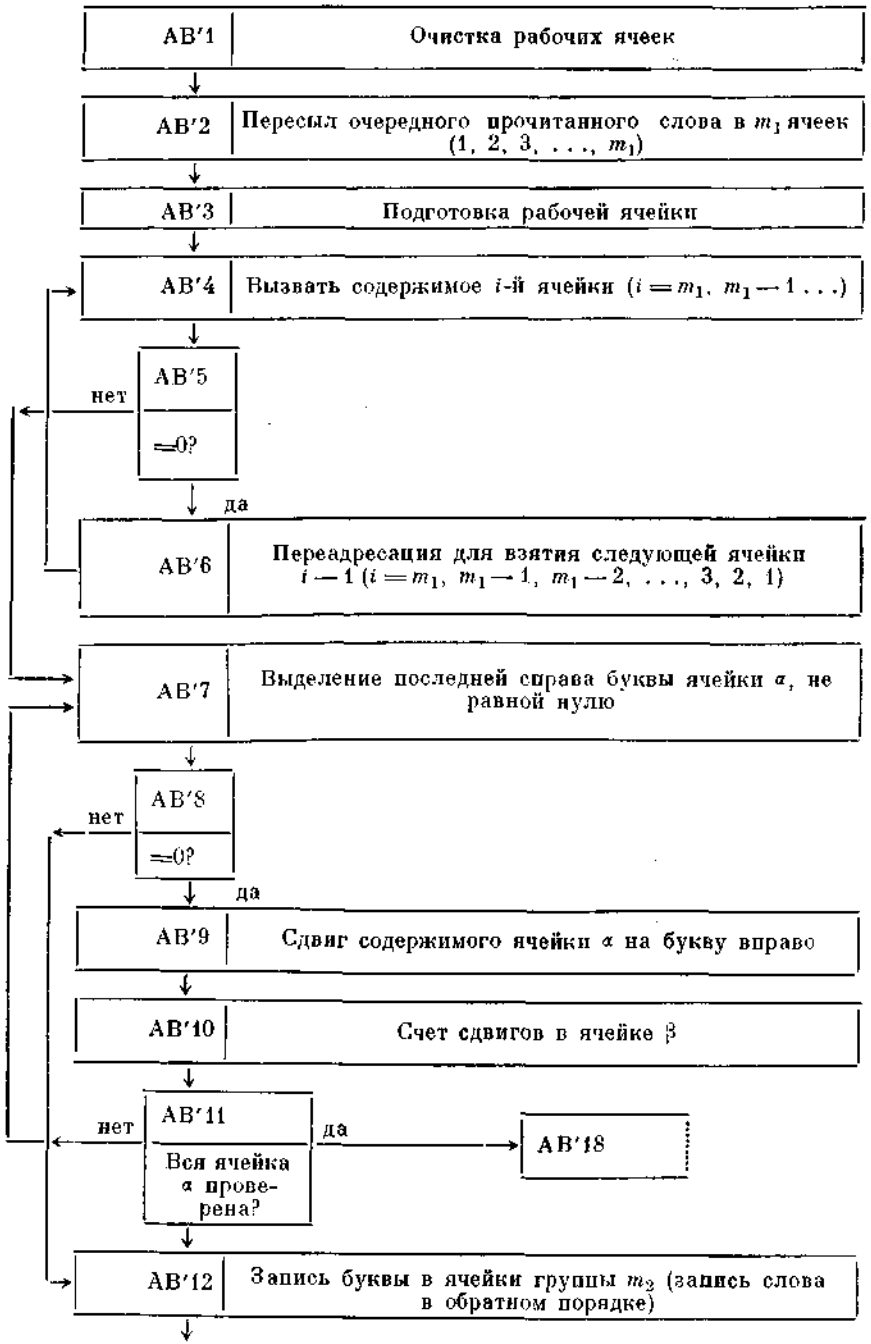


Рис. 23. Подалгоритм АВ′. Инверсия словоупотребления (*см. продолжение*).

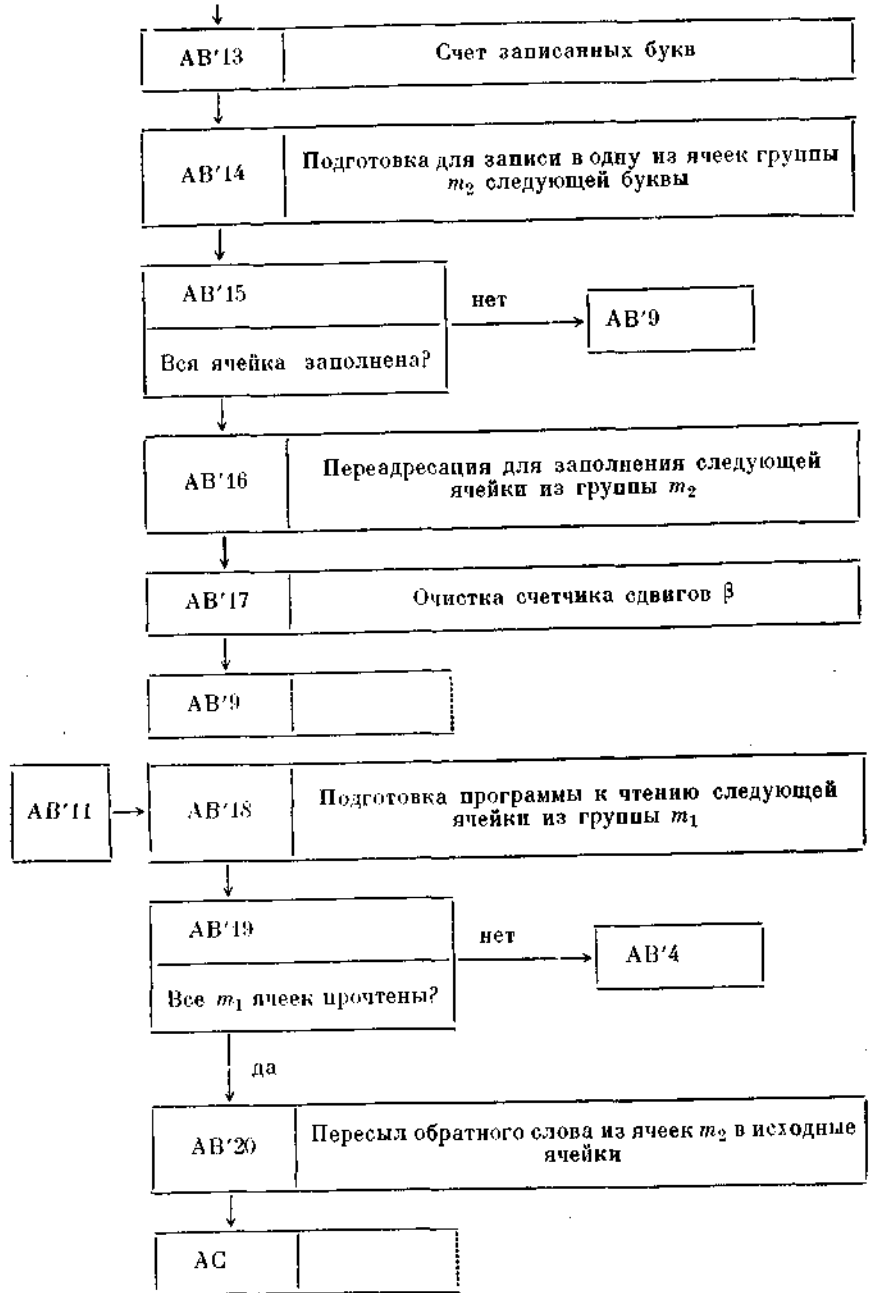


Рис. 23. (*окончание*).

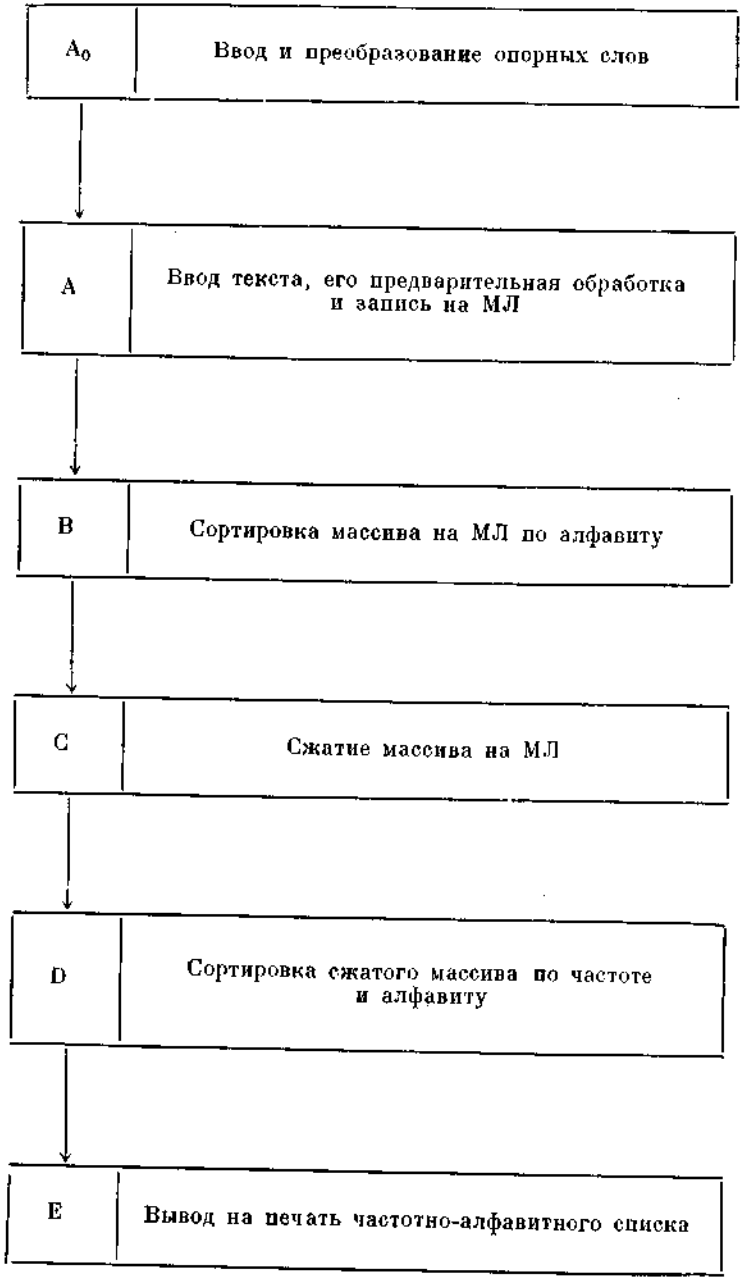


Рис. 24. Принципиальный алгоритм выделения наиболее частых n-словных сочетаний.

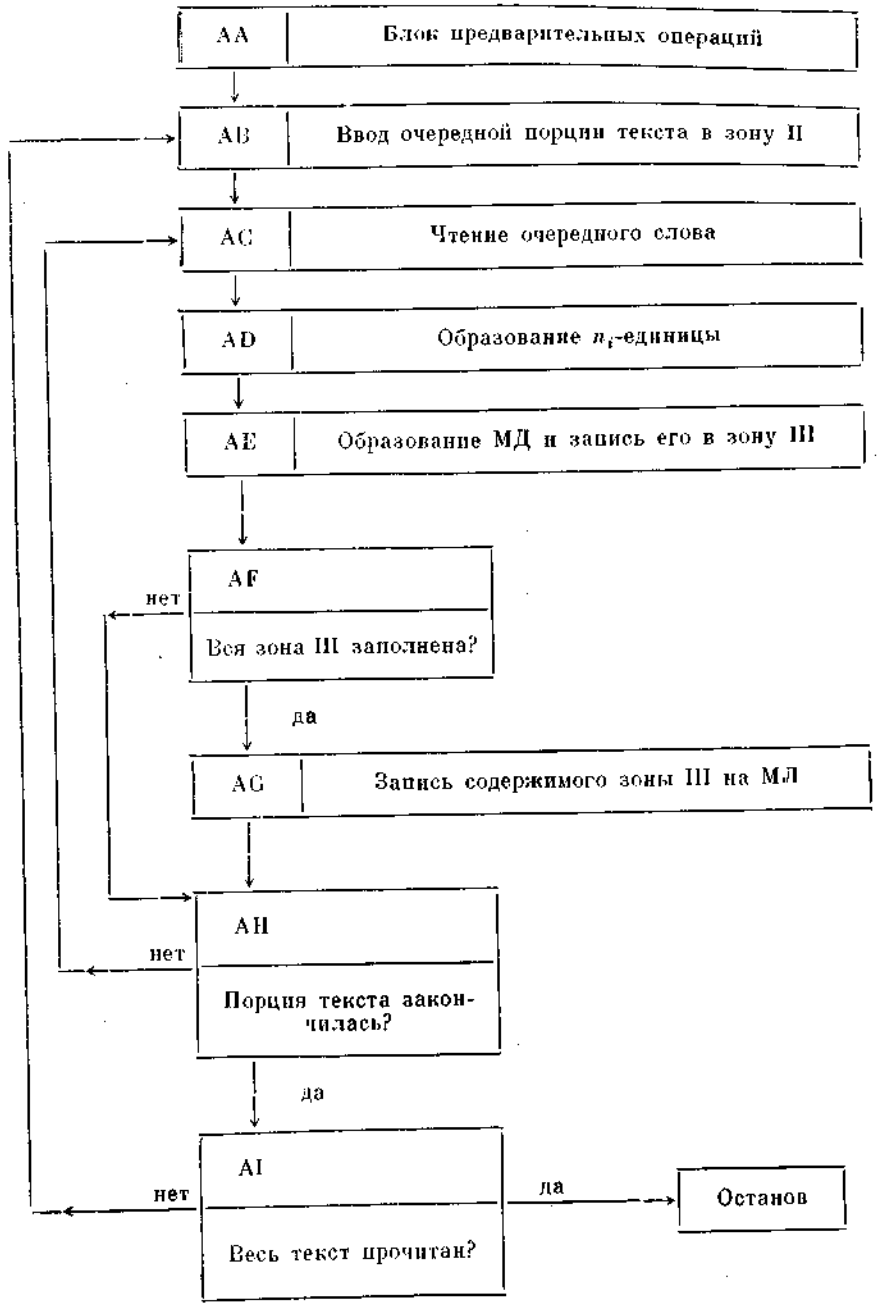


Рис. 25. Алгоритм А. Ввод текста и его предварительная обработка.

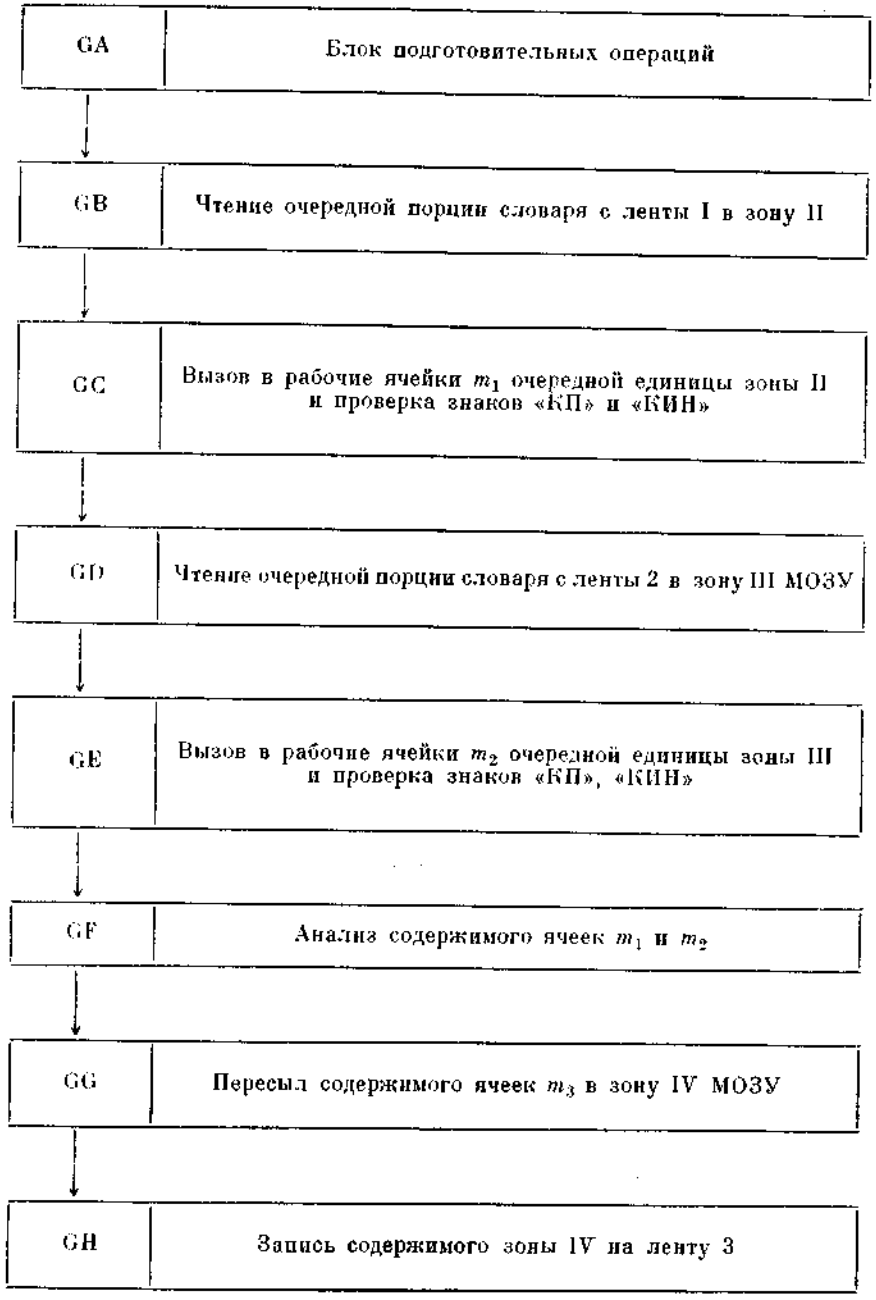


Рис. 26. Алгоритм G. Объединение первичных списков.

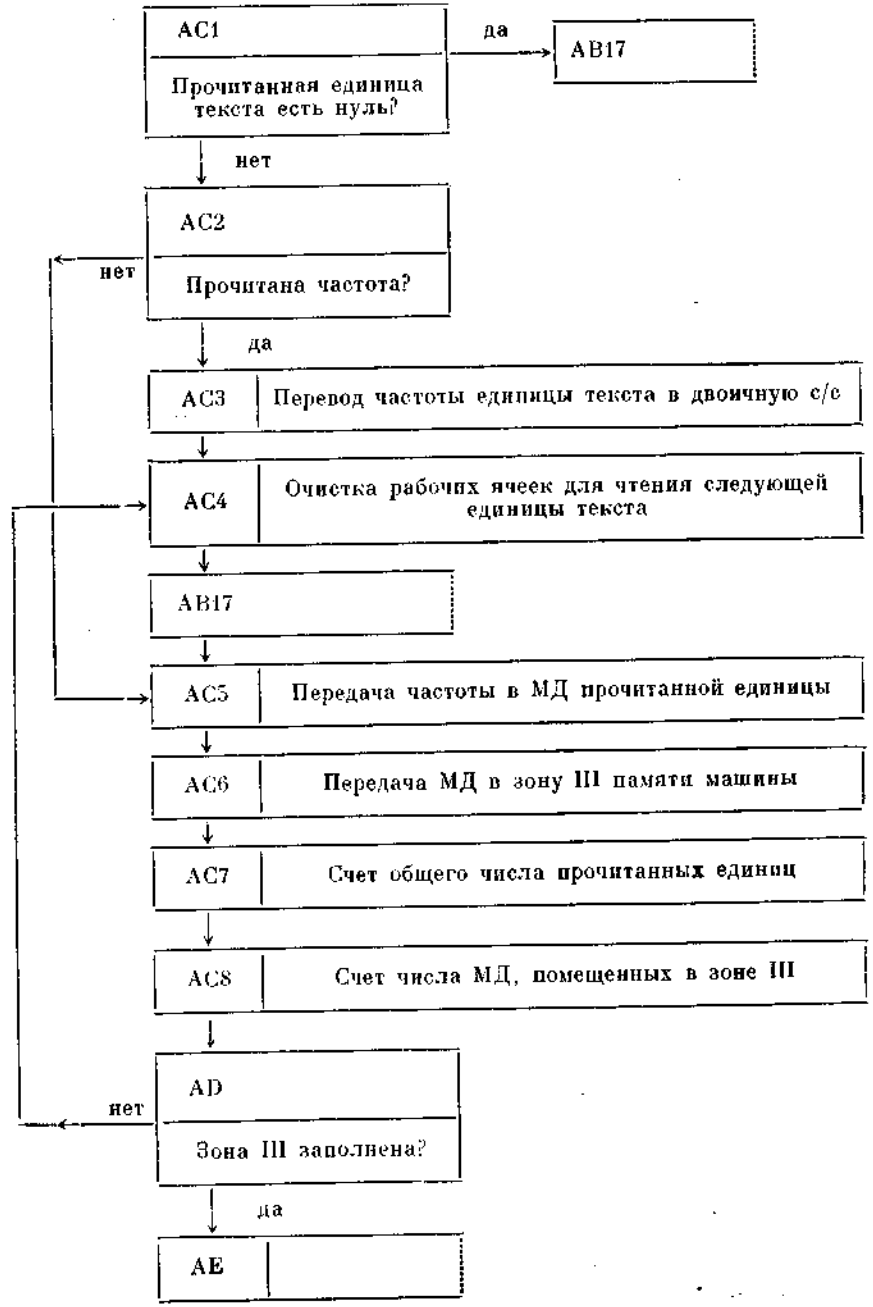


Рис. 27. Подалгоритм AC. Образование МД и запись его в зону III.

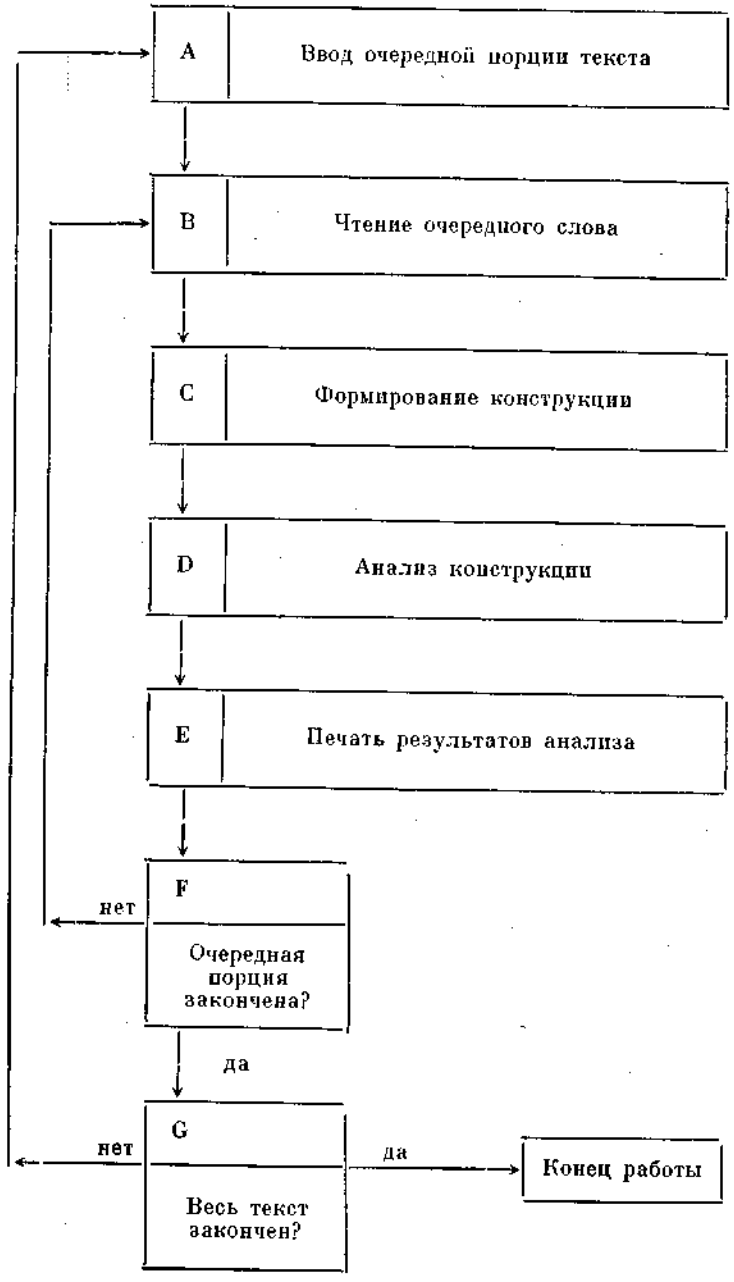


Рис. 28. Принципиальный алгоритм задачи различения лексико-грамматической омонимии слова zu.

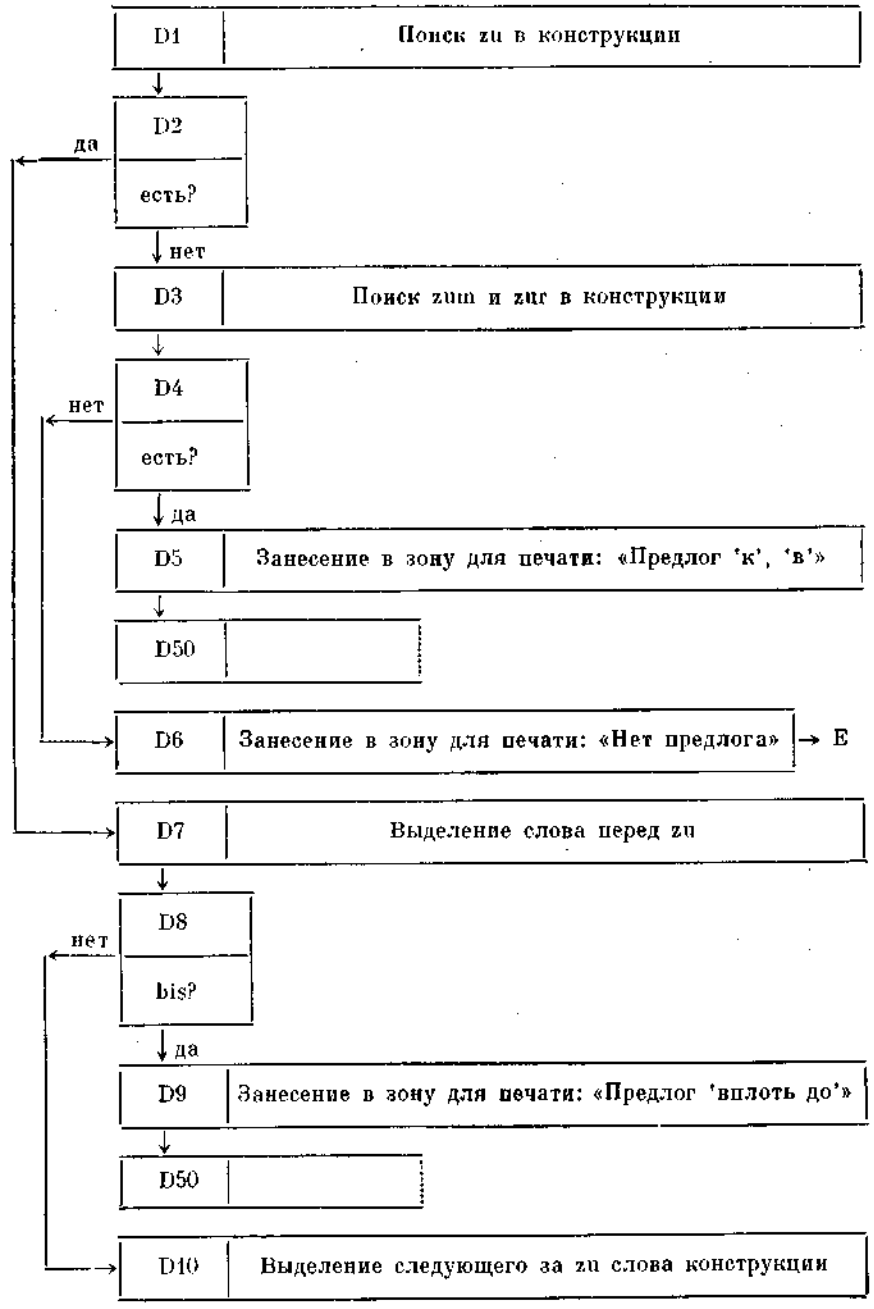


Рис. 29. Алгоритм D. Анализ конструкции с zu.

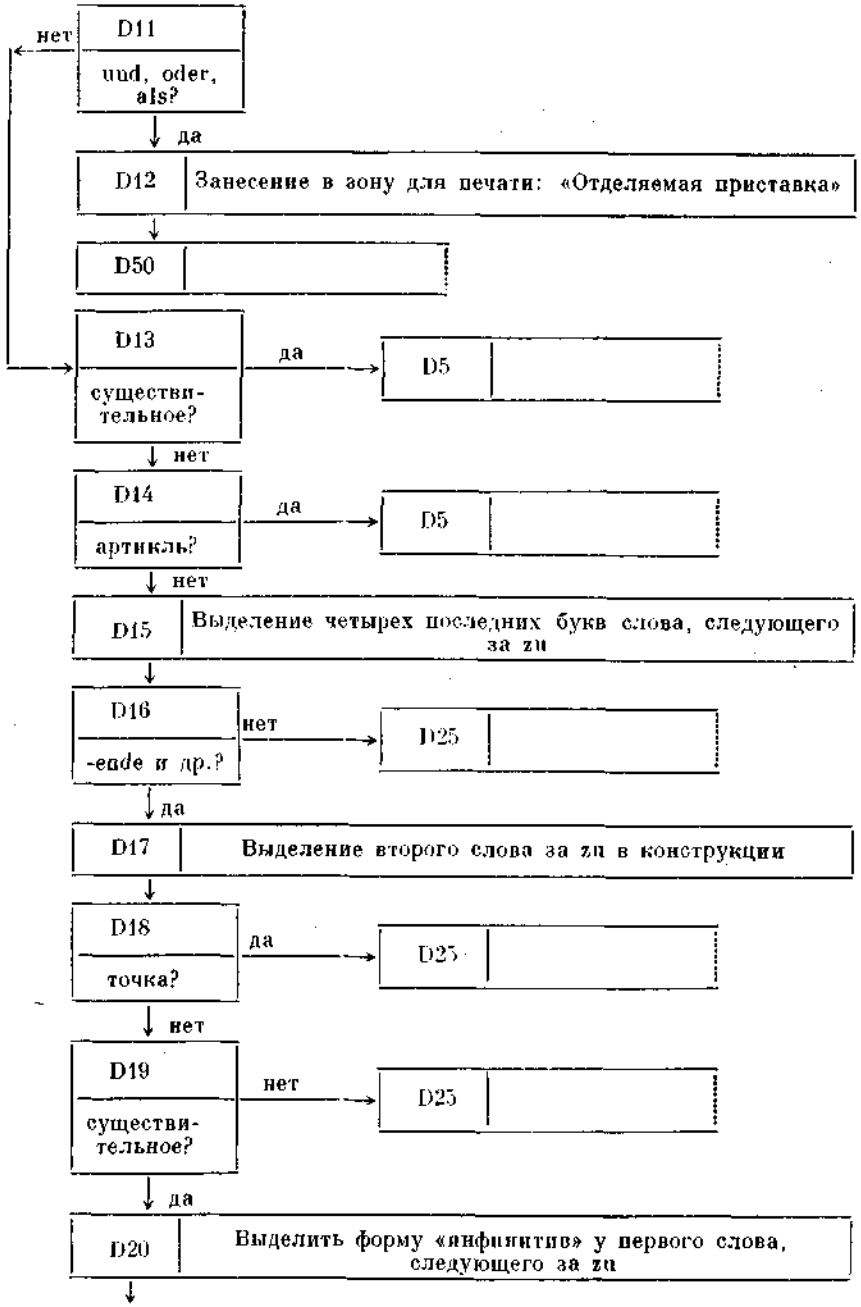


Рис. 29 (*продолжение*).

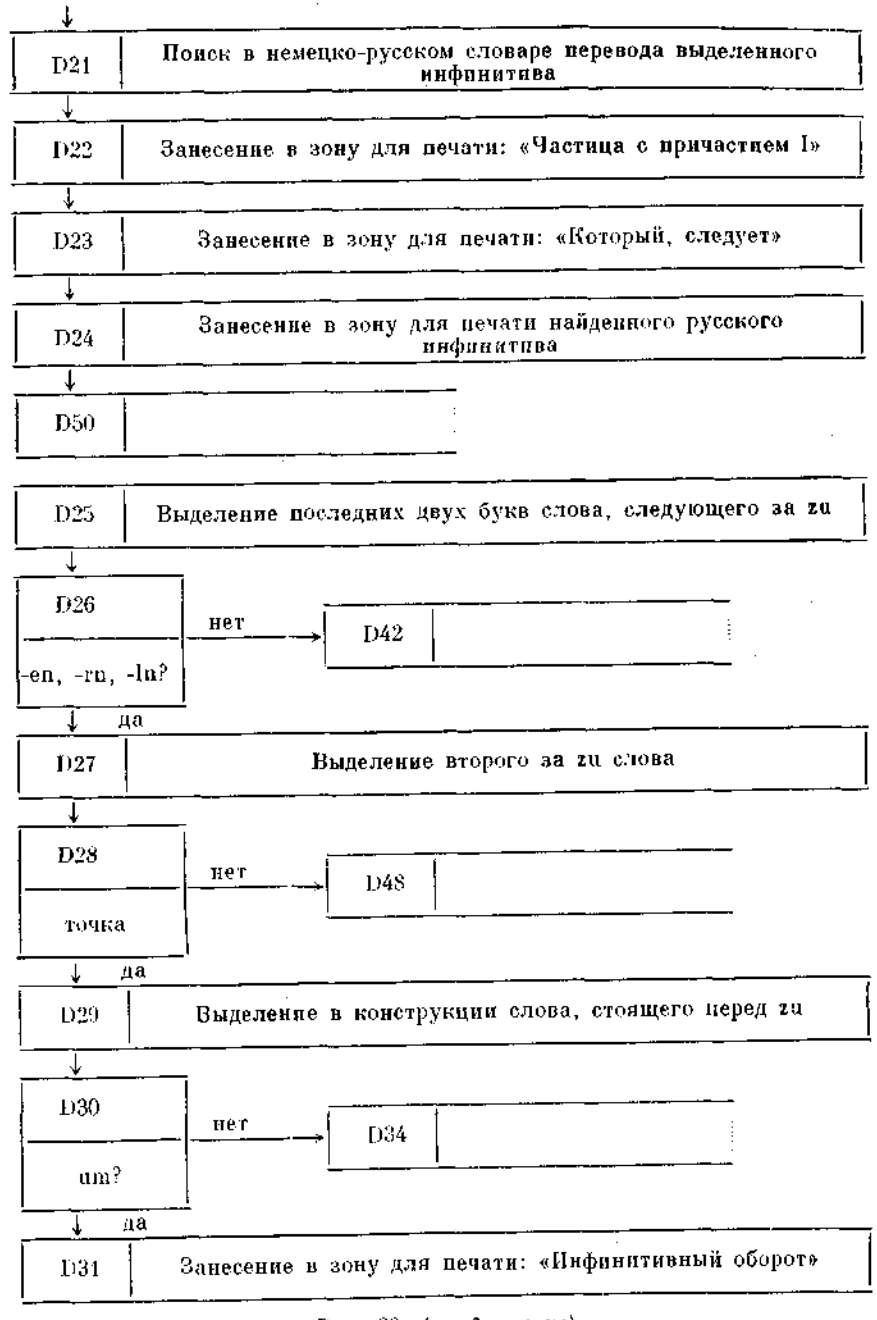


Рис. 29 (*продолжение*).

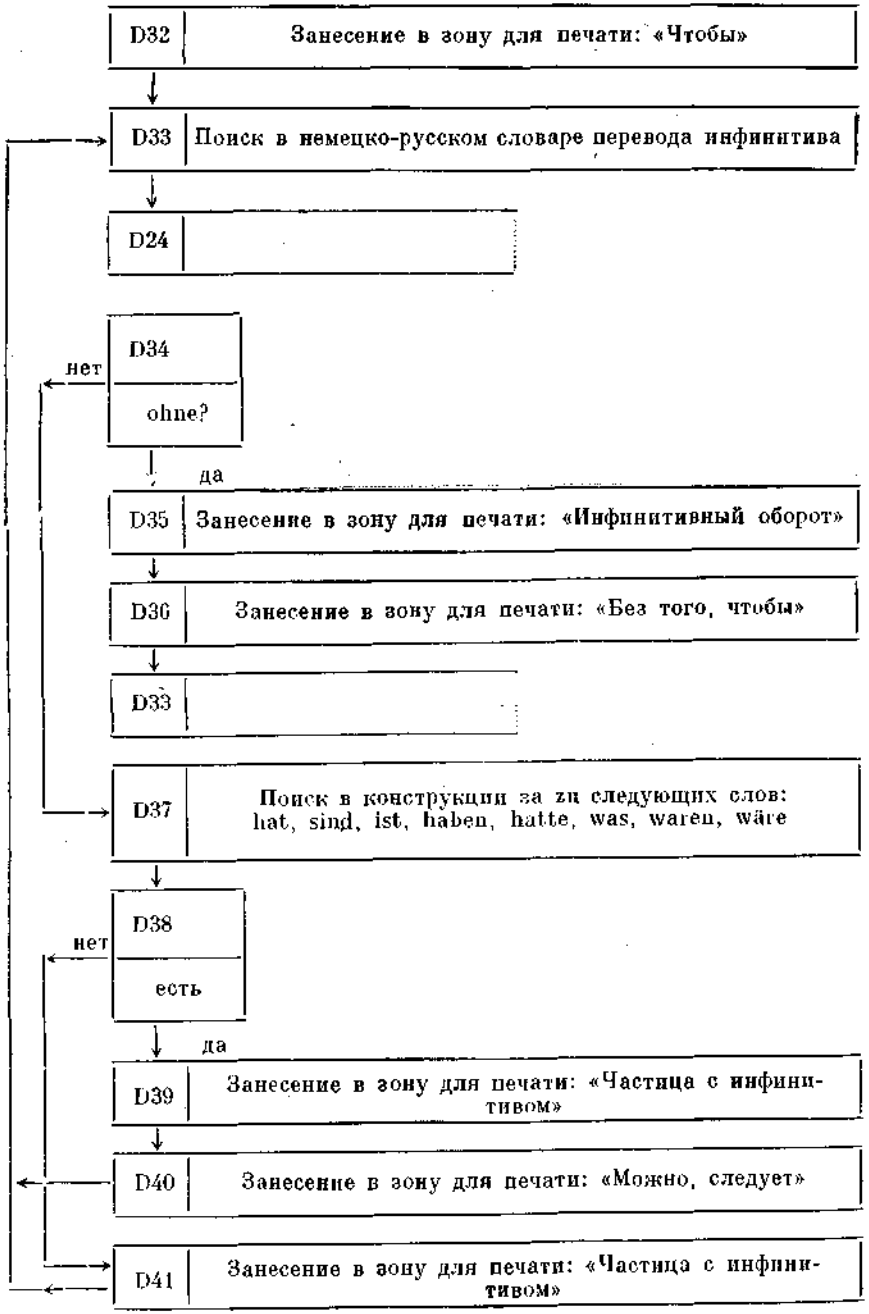


Рис. 29 (*продолжение*).

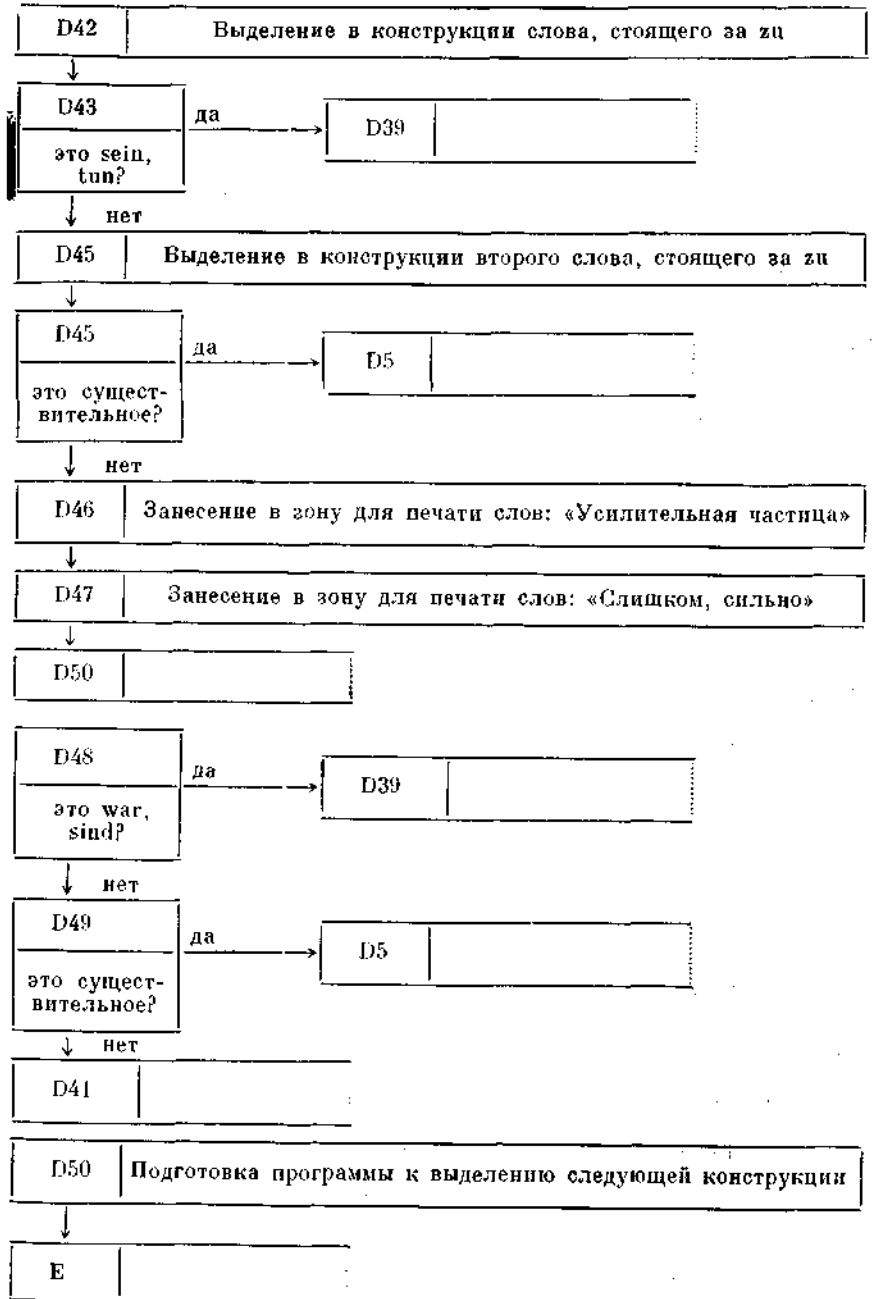


Рис. 29 (*окончание*).

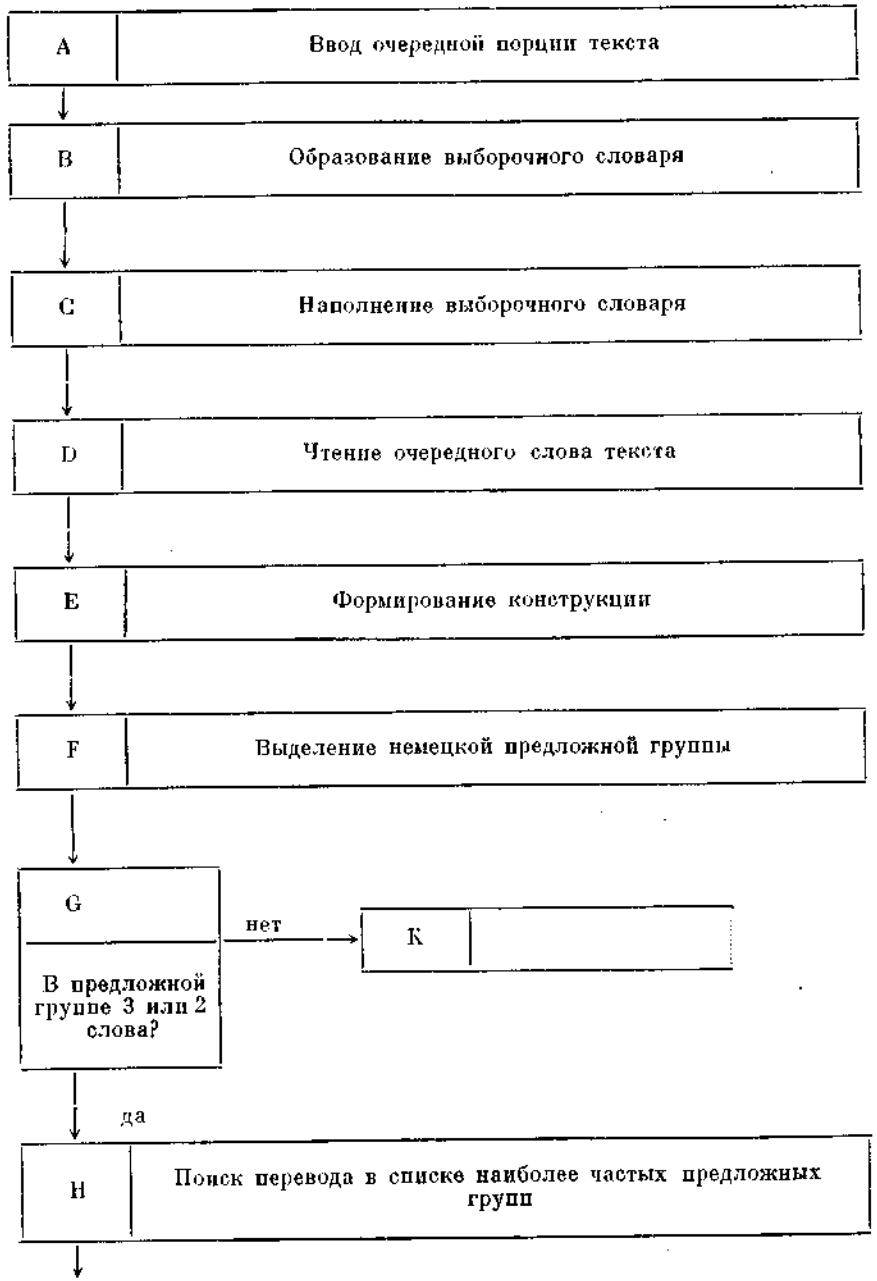


Рис. 30. Принципиальный алгоритм задачи перевода немецких предложно-именных групп (первая модель).

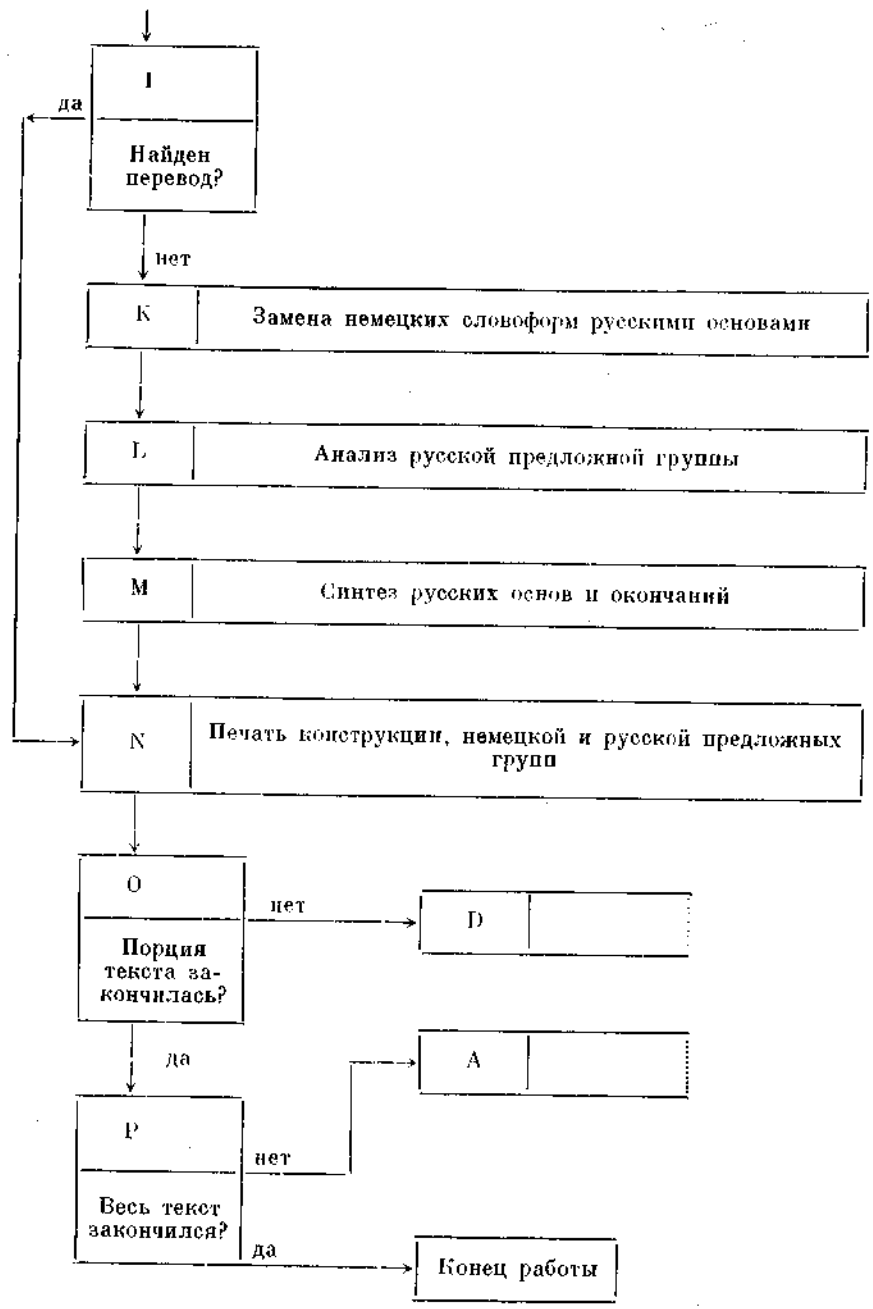


Рис. 30 (*окончание*).

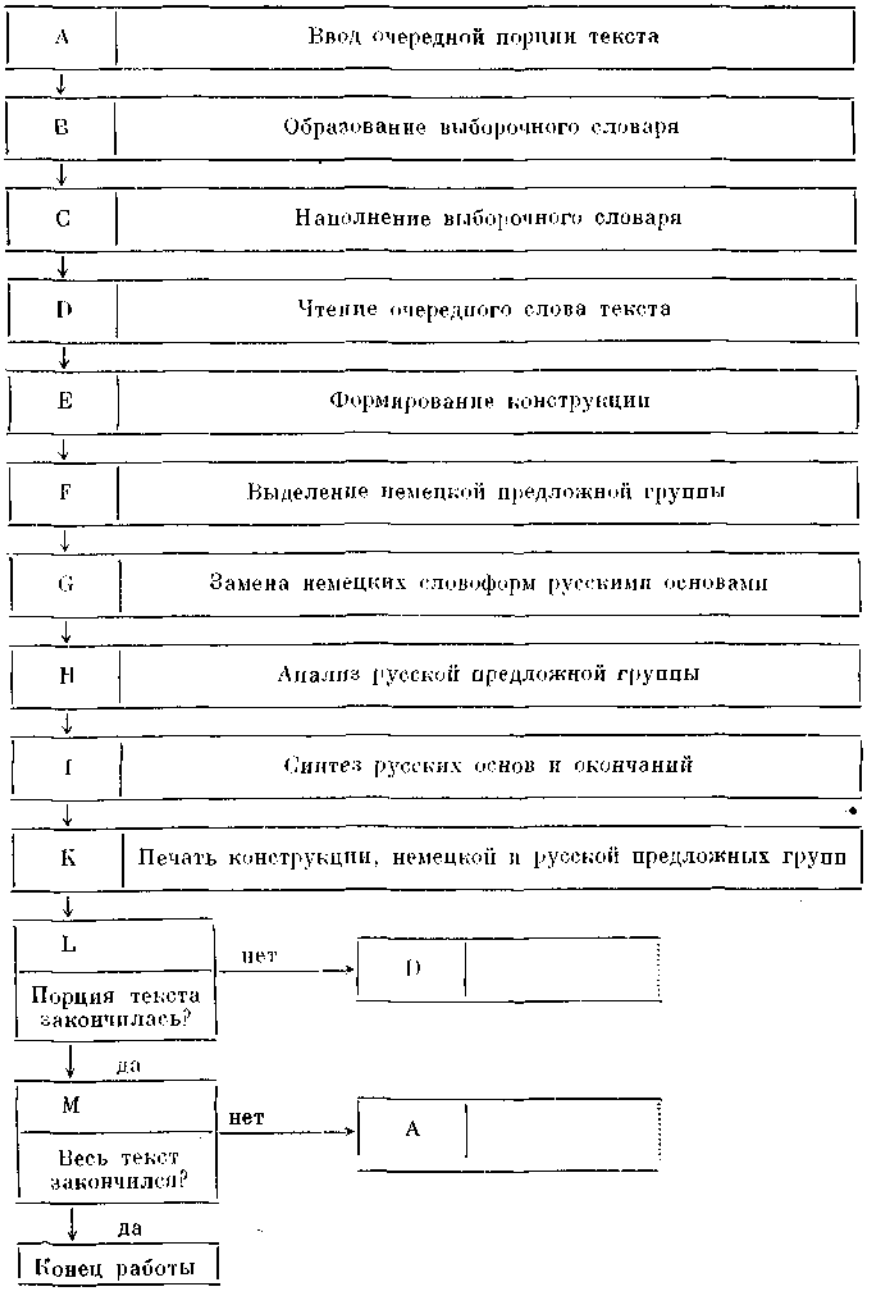


Рис. 31. Принципиальный алгоритм задачи перевода немецких предложно-именных групп (вторая модель).

В. Н. Исаков и Н. Г. Лоскутов

О ВЫБОРЕ АЛГОРИТМИЧЕСКОГО ЯЗЫКА ДЛЯ ПРОГРАММИРОВАНИЯ ЗАДАЧ ОБРАБОТКИ СМЫСЛОВОЙ ИНФОРМАЦИИ

Автоматизация того или иного процесса с помощью электронно-вычислительной машины включает в себя этап формализации процесса, результатом которого является алгоритм решения задачи, и этап программирования.

Программа является формой записи алгоритма на языке, на котором алгоритм может быть введен в машину и «воспринят» ею. Специфичность и большая трудоемкость процесса программирования требуют специальной профессиональной подготовки и значительных затрат времени на его выполнение. Поэтому прогресс в использовании ЭВМ тесно связан с автоматизацией программирования. Наибольшие результаты в области автоматизации программирования достигнуты на пути введения алгоритмических языков и трансляторов к ним. Как известно, существо этого пути состоит в следующем.

Все трудности программирования вытекают из большого различия естественного языка, языка математических формул, на котором описывается алгоритм, и языка цифрового кодирования, присущего машине. Для облегчения перехода с одного уровня языка для описания алгоритма на другой уровень вводится еще один, промежуточный уровень языка — алгоритмический язык.

Перевод алгоритма с промежуточного языка на машинный осуществляется с помощью специальной программы-транслятора на самой ЭВМ. Внедрение алгоритмических языков и трансляторов в практику использования ЭВМ дает много преимуществ, так как позволяет резко сократить сроки подготовки задачи к решению, улучшить использование машин за счет сокращения времени на отладку программ и облегчить обмен программами.

В настоящее время в различных странах разработано и используется большое количество алгоритмических языков.[1] Одной

[1] См. сб.: Алгоритмы и алгоритмические языки, М., 1968.

из причин такого их многообразия является трудность создания одного языка, одинаково хорошо приспособленного к задачам (проблемам) различного типа и к вычислительным машинам различного класса. Поэтому одни языки привязаны к задачам определенного типа (проблемно-ориентированные языки), другие — к конкретным машинам(машинно-ориентированные языки). Использование при программировании задачи не свойственного для данного типа задач алгоритмического языка сопряжено с большими неудобствами: велики затраты труда и времени, сами описания ненаглядны, оттранслированные программы велики по объему и требуют для решения больших затрат машинного времени.

В данной статье рассматриваются вопросы выбора алгоритмических языков при программировании задач, связанных с обработкой смысловой информации. Проблема обработки неформализованной информации возникает при создании автоматических информационных систем (АИС). Работы по созданию таких систем ведутся в основном в трех областях:

1) системы оперативного управления отдельных предприятий и целых отраслей производства, обеспечивающие автоматический сбор, накопление и обработку экономической информации;

2) автоматический поиск документов в больших массивах информации и их выдача по запросу, с осуществлением, при необходимости, автоматического составления рефератов;

3) автоматический перевод текстов с одного языка на другой.

Среди ряда специфических задач, решаемых в каждой из этих областей в отдельности, можно отметить наиболее важную и присущую всем областям задачу анализа смыслового содержания сообщений, которая включает в себя машинное представление смысловых связей между понятиями, автоматический перевод сообщений с естественного языка на информационный и обратно, автоматическое реферирование и оптимальное кодирование произвольных речевых сообщений для записи в память ЭВМ или передачи по каналам связи и т. п.

Подобные задачи принято относить к классу информационно-логических, которые, как правило, связаны с обработкой и анализом значительных объемов информации, а программы их решения характеризуются большим объемом и разветвленностью.[2]

Рассмотрение возможности использования одного из типов алгоритмических языков для описания алгоритмов информационно-логических задач в данной статье осуществляется на примере конкретной задачи, имеющей самостоятельное практическое значение для анализа смыслового содержания текста (сообщений).

[2] См.: А. И. Китов. Программирование информационно-логических задач. М., 1967.

В качестве таковой взята задача анализа текста на ЭВМ с целью составления частотного списка наиболее употребительных слов и словосочетаний русских текстов по радиоэлектронике.

Наш пример может рассматриваться как один из путей решения задачи автоматического реферирования текстов.

Задача автоматического реферирования сводится к получению на основе исходного текста некоторого «вторичного» текста, меньшего по объему, но отражающего основное содержание оригинала.

Известно два пути решения этой задачи:

1) извлечение из текста отдельных фрагментов (слов, словосочетаний, фраз, предложений), набор которых и является рефератом;

2) семантический анализ текста, в результате которого можно выделить основное содержание текста и представить его в виде новых фраз и предложений.

Второй путь в настоящее время рассматривается только как перспективный и не поддается машинной реализации.

Задача будет решаться первым способом, при этом необходимо:

1) выбрать критерии выделения словосочетаний текста с таким расчетом, чтобы охватить и перенести в частотный список оптимальное число синтаксических связей;

2) разработать алгоритм и программу (с последующей проверкой на ЭВМ) составления частотных словарей слов и словосочетаний современных русских текстов по радиоэлектронике.

Процедура выделения слов и словосочетаний из текста

Для выборки из текста наиболее употребительных слов используются известные методы составления частотных словарей.[3] При этом нижняя граница (по F_i^*) выборки слова задается в зависимости от объема анализируемого текста и может быть оценена по формуле:

$$F_{i\,\min} = \frac{Z_p^2}{\delta_{i\,\max}},$$

где Z_p — коэффициент, являющийся функцией ρ — заданного доверительно интервала; $\delta_{i_{max}}$ — заданная максимальная ошибка.

В качестве границ словосочетаний текста выбраны:

1) левая граница — наиболее употребительные предлоги, встречающиеся в текстах по радиоэлектронике; предлоги взяты из частотного словаря русского подъязыка по электронике Е. А. Калининой;[4]

[3] См. СР.

[4] Там же.

2) правая граница включает все предлоги, являющиеся левой границей, и дополнительно — знаки препинания (,/;/./:/?/(), два союза (*и*, *или*) и признак перехода с латинского регистра на русский.

Дополнительное ограничение: любое словосочетание имеет в своем составе не более четырех слов.

Для ускорения анализа текста все признаки границ разбиты на группы в зависимости от числа символов в признаке. Внутри каждой группы анализ производится методом последовательного перебора.

Общее число таких групп равно 6 (1+2 символа; 3+4 символа; 5 символов; 6 символов; 9 символов (1 предлог); 11 символов (1 предлог)).

Укрупненная блок-схема алгоритма и ее описание

Блок-схема алгоритма представлена на рис. 1.

Алгоритм работает следующим образом.

Исходный текст участками объемом не более 5000 знаков (максимальный объем страницы печатного текста) вводится с помощью блока 1 и размещается в массиве № 1.

В блоке 2 производится анализ очередного слова массива № 1, который состоит в определении конца слова (наличие пробела) и подсчете числа символов слова.

В блоке 3 производится анализ каждого выделенного слова на окончание всего текста. В качестве такого признака избрана группа вида: К К К К К —. Если текст не закончен, осуществляется переход к блоку 3 при анализе каждого слова текста (блок 3^{I}—3^{III}). Проверяется признак конца участка. В качестве его избрана комбинация вида: У У У У У ⌴.

Если участок текста закончен, то проверяется наличие свободного массива ячеек памяти для размещения новых слов и словосочетаний с целью анализа очередного участка текста. При этом предполагается, что необходимо иметь память в объеме до 30% от исходного текста. Если памяти не хватает, то производится сброс частотных словарей во внешнюю память. В дальнейшем информация на магнитной ленте рассматривается как исходная для работы алгоритма. Смысл такого рассмотрения состоит в учете одинаковых слов и словосочетаний, полученных при анализе до сброса на МЛ и после него.

Если участок текста не закончен, то осуществляется переход к блоку 6. Назначение его ясно из названия.

При составлении частотного словаря все слова, имеющие один символ, исключаются из рассмотрения.

Назначением переключателя № 1 (блок 7) является изменение направления анализа словосочетания.

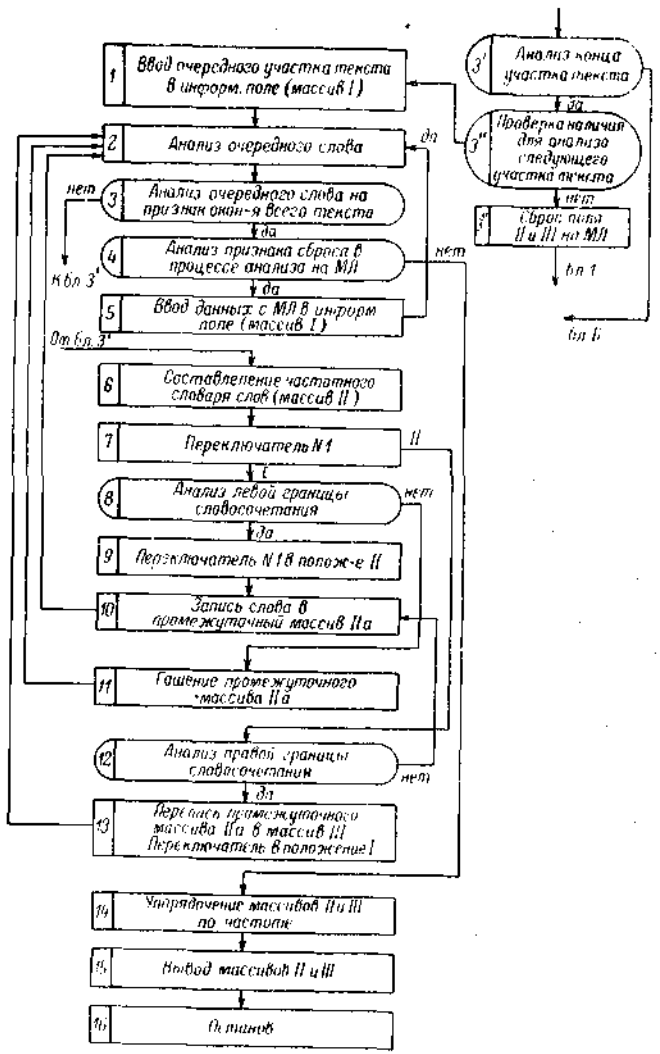


Рис. 1. Укрупненная блок-схема алгоритма.

В положении I анализируется левая граница словосочетания. После ее нахождения (блок 8) переключатель ставится в положение II, в результате чего каждое очередное слово после левой границы анализируется на признак правой границы (блок 12). После нахождения правой границы переключатель вновь устанавливается в положение I, а выделенное словосочетание записывается в выходной массив III (блок 13). Кроме того, в блоке 13 определяется, не встречалось ли подобное словосочетание раньше. Если да, то запись словосочетания не производится, а происходит добавление 1 в счетчик F_i^* данного словосочетания.

После анализа последнего участка текста с учетом, если необходимо, информации на магнитной ленте производится упорядочение частотных словарей слов (массив II) и словосочетаний (массив III) по частоте (блок 14) и выдача словарей на печать (блок 15).

Таким образом, результатом работы алгоритма является упорядоченные по частоте списки наиболее употребительных словосочетаний и слов исходного текста.

В настоящее время в СССР наиболее распространены алгоритмические языки, построенные на основе международного алгоритмического языка АЛГОЛ-60.

Так, для вычислительных машин класса «Минск», широко используемых для решения лингвистических задач, разработаны трансляторы с языков АЛГАМС [5] и АЛГЭМ.[6]

Алгоритмический язык АЛГАМС является версией АЛГОЛ-60 и приспособлен к решению задач вычислительного характера.

На языке АЛГАМС описываются задачи со сложной разветвленной логикой, что характерно для лингвистических задач. И однако этот язык не может быть применен в данном случае по следующим причинам:

1) он не позволяет работать с частями машинного слова, так как в нем не определено понятие двоичного кода;

2) здесь недостаточно конкретизуется, по сравнению, например, с языком АЛГЭМ, понятие оператора процедуры-кода, что затрудняет использование стандартных подпрограмм, записанных на языке конкретной машины.

Язык АЛГЭМ приспособлен к программированию экономических задач и построен путем наращивания языка АЛГОЛ средствами, необходимыми для обработки больших массивов информации, последовательно размещаемых в памяти машины.

По своему характеру экономические задачи ближе к лингвистическим, чем задачи вычислительные. Так, и в экономических и в лингвистических задачах необходима обработка больших массивов информации, имеющей полностью детерминированную струк-

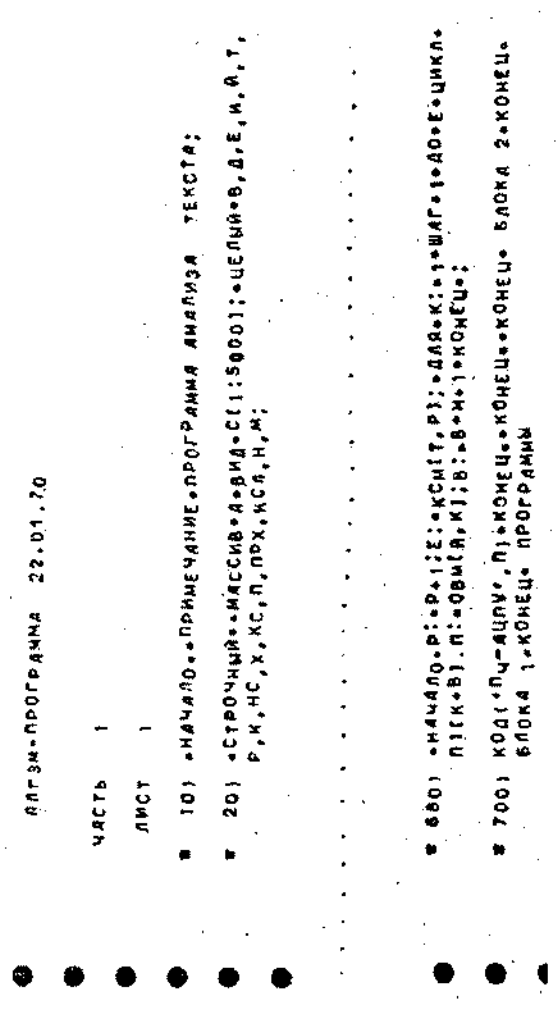
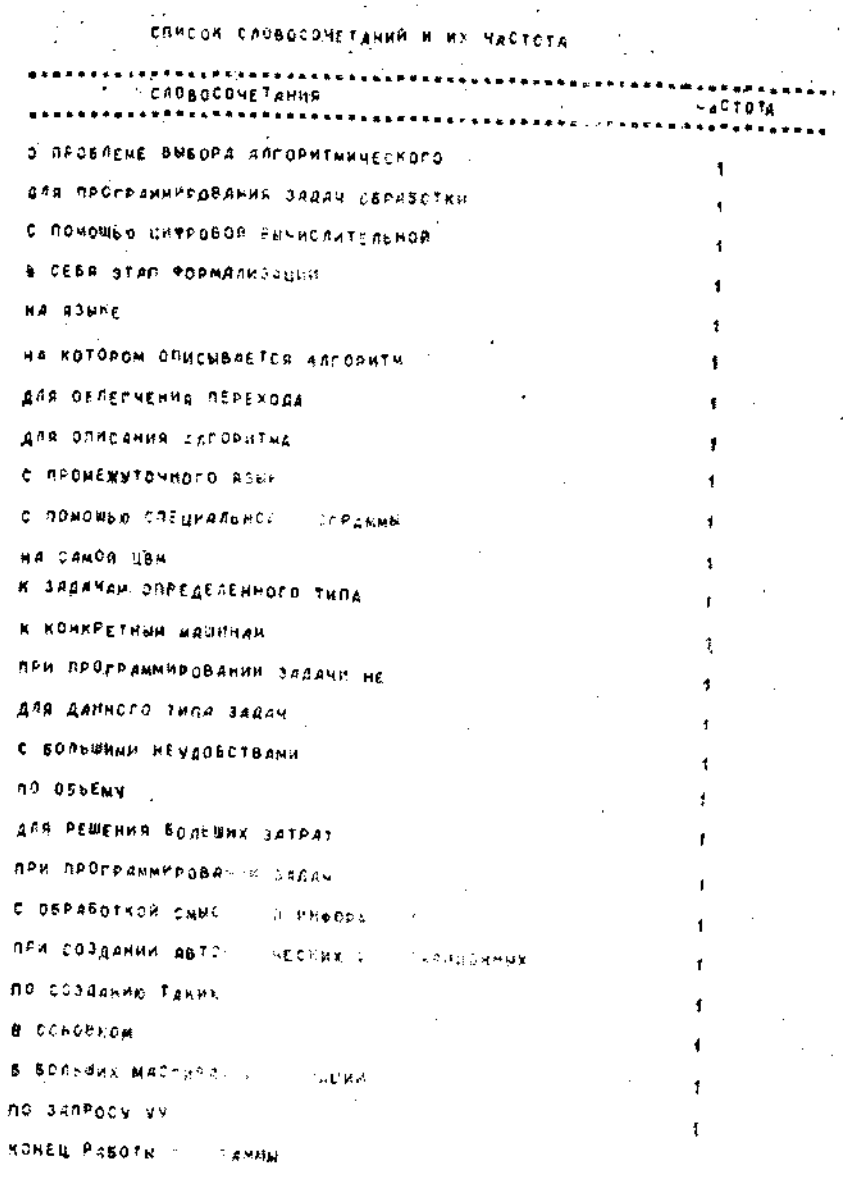


Рис. 2. Фрагмент АЛГЭМ-программы.

[5] См. сб.: Алгоритмы и алгоритмические языки.
[6] См.: А. И. Китов, ук. соч.



СПИСОК СЛОВОСОЧЕТАНИЙ И ИХ ЧАСТОТА

СЛОВОСОЧЕТАНИЯ	ЧАСТОТА
О ПРОБЛЕМЕ ВЫБОРА АЛГОРИТМИЧЕСКОГО	1
ДЛЯ ПРОГРАММИРОВАНИЯ ЗАДАЧ ОБРАБОТКИ	1
С ПОМОЩЬЮ ЦИФРОВОЙ ВЫЧИСЛИТЕЛЬНОЙ	1
В СЕБЯ ЭТАП ФОРМАЛИЗАЦИИ	1
НА ЯЗЫКЕ	1
НА КОТОРОМ ОПИСЫВАЕТСЯ АЛГОРИТМ	1
ДЛЯ ОБЛЕГЧЕНИЯ ПЕРЕХОДА	1
ДЛЯ ОПИСАНИЯ АЛГОРИТМА	1
С ПРОМЕЖУТОЧНОГО ЯЗЫ	1
С ПОМОЩЬЮ СПЕЦИАЛЬНОЙ ПРОГРАММЫ	1
НА САМОЙ ЦВМ	1
К ЗАДАЧАМ ОПРЕДЕЛЕННОГО ТИПА	1
К КОНКРЕТНЫМ МАШИНАМ	1
ПРИ ПРОГРАММИРОВАНИИ ЗАДАЧИ НЕ	1
ДЛЯ ДАННОГО ТИПА ЗАДАЧ	1
С БОЛЬШИМИ НЕУДОБСТВАМИ	1
ПО ОБЪЕМУ	1
ДЛЯ РЕШЕНИЯ БОЛЬШИХ ЗАТРАТ	1
ПРИ ПРОГРАММИРОВАНИИ ЗАДАЧ	1
С ОБРАБОТКОЙ СМЫС ИНФОРМ	1
ПРИ СОЗДАНИИ АВТО ЧЕСКИХ	1
ПО СОЗДАНИЮ ТАКИХ	1
В ОСНОВНОМ	1
В БОЛЬШИХ МАСШТАБ ЗАДАЧИ	1
ПО ЗАПРОСУ УУ	1
КОНЕЦ РАБОТЫ ПРОГРАММЫ	1

Рис. 3. Машинный результат.

туру; специфичным для них является также большой удельный вес логических операций (проверка различных признаков, выделение отдельных частей слова и работа над ними и т. д.). Введение в языке понятия строчной переменной, восьмеричного и двоичного кода, нумерации разрядов слова и оператора процедуры-кода отвечает указанным особенностям задач. Поэтому из имеющихся у нас алгоритмических языков именно язык АЛГЭМ наиболее приспособлен к лингвистическим задачам.

Особо важными достоинствами языка АЛГЭМ, облегчающими программирование лингвистических задач, являются:

1) допущение во входной информации к задаче не только цифр и букв, но и знаков математических операций (↑, [,], >, ', ×, / и др.), введение повторителей при набивке одинаковых символов облегчает и ускоряет процесс подготовки информации;

2) возможность сегментации программ позволяет разрабатывать по частям большие по объему программы;

3) легкость включения в АЛГЭМ-программу готовых участков программ;

4) способность транслятора СТ-3 выдавать детальную информацию об ошибках в программе, обеспечить простоту внесения исправлений в нее, что значительно упрощает и сокращает процесс отладки программы.

Неудобством языка является исключение в нем первичных логических операций, что затрудняет описание задач с разветвленной логикой и удлиняет алгоритм.

Однако опыт использования АЛГЭМ для сформулированной выше задачи показывает, что последнее обстоятельство не является решающим и применение АЛГЭМа к данному классу задач дает большие преимущества перед другими методами программирования.

Фрагмент программы и результаты решения сформулированной задачи, записанной на язык АЛГЭМ, приведены на рис. 2 и 3. Общий объем программы в машинных кодах — 1800 команд, что на 20—30% больше объема программы, которую мог бы составить опытный программист. Время трансляции программы с транслятором СТ-3=30 мин. На отладку программы затрачено около 4-х часов машинного времени.

Р. Г. Пиотровский и *В. А. Чижаковский*

О ДВУЯЗЫЧНОЙ СИТУАЦИИ

В ходе перевода с одного языка на другой (будем называть такой перевод бинарным в отличие от перевода через язык-посредник) возникает особая система лингвистических отношений, которую принято называть д в у я з ы ч н о й с и т у а ц и е й.

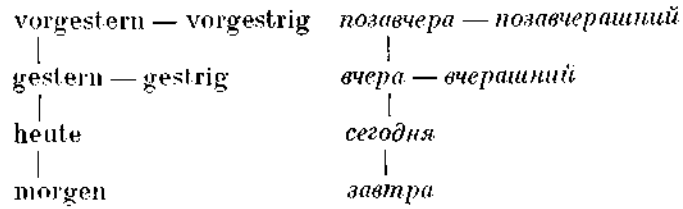
Сущность двуязычной ситуации состоит в следующем. Каждая единица языка и каждая связь между этими единицами характеризуются набором определенных парадигматических и синтагматических отношений. Напоминаем, что эти отношения характеризуют не только грамматические формы, фонемы, слова и словосочетания, но и значения отдельных морфем, слов и словосочетаний, а также семантико-синтаксические связи.

В ходе перевода (ручного или машинного) мы ставим элементы и связи входного языка в некоторое соответствие элементам и связям выходного языка. Поскольку каждая лингвистическая единица имеет свою сетку структурных отношений, постольку соотнесение лингвистических единиц предусматривает соотнесение и их «структурных» сеток. Механизм этого соотнесения представляет собой функционирование двуязычной ситуации. Иными словами, двуязычная ситуация — это суперструктура, объединяющая структуры как входного, так и выходного языка.

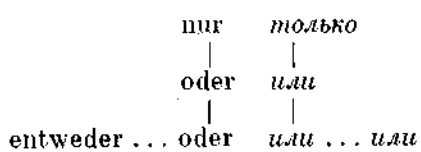
В рамках двуязычной ситуации обнаруживаются три типа соотнесений сегментов входного текста соответствующим сегментам выходного текста: полная калькируемость, квазикалькируемость и некалькируемость.

1. П о л н а я к а л ь к и р у е м о с т ь (ПК) предусматривает взаимно-однозначную изоморфность каждого элемента входного сегмента соответствующему элементу выходного сегмента. Это значит, что первое слово входного сегмента имеет парадигматическую и синтагматическую сетку отношений, совершенно аналогичную сетке первого слова выходного словосочетания. При этом значения входного и выходного слова полностью совпадают, они одинаково соотнесены с другими словами семантических и морфологических парадигм и имеют идентичную валентность. В аналогичных отношениях взаимно-однозначного изоморфизма находятся первое, второе, третье и т. д. слова входного и выходного сегментов. Естественно, порядок слов выходного словосочетания точно соответствует порядку слов входного сегмента.

Рассмотрим в качестве примера фрагмент двуязычной ситуации, представляющей собой взаимоотношение немецкого словосочетания gestern oder heute [1] и его русского эквивалента *вчера или сегодня*.[2] Парадигматические отношения значений немецких наречий gestern и heute и русских наречий *вчера* и *сегодня* можно представить в виде следующих изоморфных графов:



Упрощенную семантическую (антонимическо-синонимическую) парадигматику немецкого союза oder и его русского эквивалента *или* представим также в виде двух изоморфных графов:[3]



Синтагматические отношения наречий gestern и heute и их русских эквивалентов *сегодня* и *завтра* также идентичны: и немецкие и русские наречия употребляются с глаголами любых значений и могут стоять как после них, так и перед ними; вместе с наречиями nachmitag и abend, соответствующими русским *после обеда* и *вечером*, они образуют сочетания heute nachmitag и heute abend, соответствующие русским *сегодня после обеда* и *сегодня вечером*. Аналогичным образом совпадает синтагматика союзов oder и *или*.

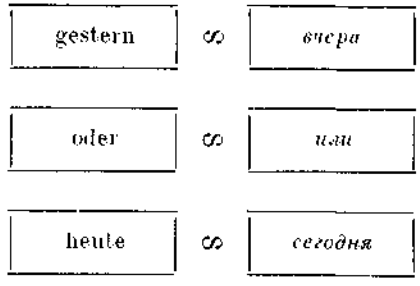
Что же происходит при переводе немецкого выражения gestern oder heute на русский язык? Как уже указывалось, каждому немецкому слову ставится в соответствие его русский эквивалент, причем одновременно сравниваются сетки их парадигматических и семантических отношений. Изображая сумму парадигмати-

[1] См.: Немецко-русский словарь, под ред. А. А. Лепина и Н. П. Смирновой. М., 1964.

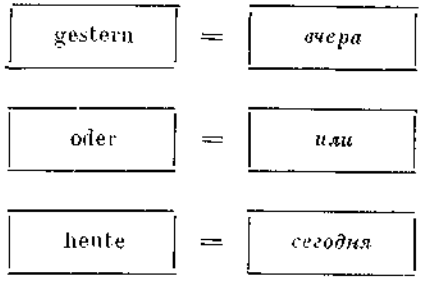
[2] См.: Русско-немецкий словарь, под ред. А. Б. Лохтвица. М., 1956.

[3] Упрощение состоит здесь в том, что мы не учитываем существование русского союза *либо*.

ческих и синтагматических отношений в виде прямоугольника, а операцию сравнения обозначая знаком ∽, мы сможем изобразить описанную ситуацию в виде следующей схемы:



Выше уже было показано, что между сравниваемыми немецкими словами и их сетками, с одной стороны, и их русскими эквивалентами и их сетками, с другой стороны, имеется полное взаимозначное соответствие. Иными словами:



Отсюда следует, что перевод немецкого gestern oder heute может быть осуществлен как вручную, так и на машине путем пословной замены (перекодировки) каждого немецкого слова его русским эквивалентом: вместо gestern — *вчера*, вместо oder — *или*, вместо heute — *сегодня*; в итоге имеем: gestern oder heute= *вчера или сегодня*. Перевод сегмента (выражения) входного языка осуществляется здесь путем пословной перекодировки на выходной язык, без какого бы то ни было нарушения парадигматических и синтагматических отношений. Примером полной калькируемости может служить также пара англ., фр. «Fig. 2» — русск. «Рис. 2» (в дальнейшем мы будем оперировать только немецкими примерами).

2. К в а з и к а л ь к и р у е м о с т ь (КК) наблюдается в тех случаях, когда между единицами входного и выходного сегментов имеют место (или искусственно устанавливаются) однозначные соотношения: каждому элементу (и связи) входного словосочетания соответствует единственный выходной элемент (но не наоборот). Ср., например, такое соотношение немецких и русских лексем:

немецк.	русск.
Instrukteur	*инструктор*
Unterweiser	*инструктор*
Anlerner	*инструктор* [4]

Структурная КК имеет место и в тех случаях, когда между словоформами, образующими данный входной сегмент, и соответствующими словами выходного сегмента наблюдается полная лексико-грамматическая эквивалентность, однако их парадигматические (соотв. синтагматические) отношения не изоморфны: нескольким связям входного языка соответствует одна связь выходного; ср. нем. sing, Willy! — русск. *пой*, *Вилли*! Хотя каждый элемент немецкого сегмента однозначно перекодируется в русский эквивалент, однако их морфологические парадигмы будут совершенно различны. Нем. Willy имеет форму родительного падежа Willys, в то время как в русском языке морфологические противопоставления прямых падежей родительному нейтрализованы. Аналогичным образом парадигма нем. sing не изоморфна парадигме русск. *пой*. В условиях МП к КК можно также отнести и те случаи, когда одной словоформе входного языка однозначно соотнесен сегмент выходного языка, состоящий из двух и более словоформ; ср. нем. einerseits — русск. *с одной стороны*, или нем. anderseits — русск. *с другой стороны*.

Иными словами, в случае КК допустим пословный перевод каждой словоформы входного сегмента, однако изоморфного соотнесения морфологических, синтаксических и семантических отношений каждого элемента входного сегмента сетке соответствующего выходного элемента нет. Все эти виды неизоморфности могут, разумеется, вступать в комбинации. В связи с этим КК распадается на семь подтипов.

а) КК, связанная с морфологической неизоморфностью элементов входного и выходного сегментов; ср. нем. sing, Willy! — русск. *пой*, *Вилли*! (комментарий см. выше).

б) КК, определяемая синтаксической неизоморфностью элементов входного и выходного сегментов. Практически чистых случаев синтаксической квазикалькируемости обнаружить не удается, поскольку эта КК, как правило, бывает осложнена морфолого-семантической квазикалькируемостью. Она может иметь место тогда, когда потенциально имеются два входных синтаксических синонима, которым соответствует один выходной эквивалент.

в) КК, связанная с семантической неизоморфностью элементов входного и выходного сегментов; ср. нем. sogar du? — русск.

[4] При определении русских эквивалентов для немецких слов и выражений мы пользуемся словарями, указанными в прим. 1 и 2 (стр. 445).

даже ты? Семантическая парадигма русск. *даже* не соответствует парадигме нем. sogar, которое имеет синоним selbst, также переводящийся русск. *даже* (кроме того selbst может переводиться как *сам, сама, само, сами. . .*).

г) КК, обусловленная морфолого-синтаксической неизоморфностью. Если сравнить входной русский сегмент *спишь ты?* и его немецкий эквивалент schläfst du?, то обнаруживается, во-первых, несоответствие морфологических парадигм русск. *спишь* и нем. schläfst, а во-вторых, синтагматические отношения *спишь ты?* — schläfst du? неизоморфны, поскольку свободному порядку слов во входном сегменте, допускающему варианты *спишь ты?* и *ты спишь?*, соответствует жесткий синтаксис выходного эквивалента schläfst du?

д) КК, имеющая морфолого-семантическую неизоморфность элементов входного и выходного сегментов; ср. нем. Radio Oslo и русск. *радио Осло*. Здесь имеет место неизоморфность морфологических парадигм первых словоформ; ср. нем. Radio — Radios и русск. *радио* (см. пример sing, Willy!). Кроме того, значение, выраженное одним русским словом *радио*, передается в немецком языке парой синонимов: Radio — Funk.

е) КК, связанная с семантико-синтаксической неизоморфностью; ср. нем. einerseits und anderseits — русск. *с одной стороны и с другой стороны*. Здесь мы обнаруживаем, во-первых, несоответствие семантики немецкого und и русск. *и*: немецкому und может соответствовать не только русск. *и*, но и русск. союз *а*; во-вторых, значения, покрываемые русскими сегментами *с одной стороны* и *с другой стороны*, передаются в немецком языке каждое парой синонимов соотв. einerseits — einesteils, anderseits — andersteils. Синтаксическая неизоморфность einerseits — *с одной стороны* и anderseits — *с другой стороны* была описана выше (см. стр. 447).

ж) КК, определяемая морфолого-семантико-синтаксической неизоморфностью элементов выходного и входного сегментов; ср. русск. *лечишь ты Буби?* — и нем. kurierst du Buby? Морфологическая неизоморфность русск. *Буби* и нем. Buby аналогична неизоморфности русск. *Вилли* и нем. Willy (см. стр. 447); русск. *лечишь* и нем. kurierst обнаруживают семантическую неизоморфность, поскольку у русск. *лечить* синонимы практически отсутствуют, а нем. kurieren имеет синоним behandeln (см. стр. 449). Иными словами, семантическое поле русск. *лечишь* шире по сравнению с нем. kurieren. В свою очередь нем. behandeln имеет дополнительные русские эквиваленты: *обращаться, обходиться, обрабатывать, излагать, обслуживать, практиковать, трактовать, разрабатывать. . .* Синтаксическая неизоморфность обнаруживается в существовании нескольких вариантов порядка слов русского входного сегмента: *лечишь ты Буби?*; *ты лечишь Буби?*; *Буби ты лечишь?*; *ты Буби лечишь?*; *лечишь Буби ты?*, которым

соответствует жесткий порядок слов выходного эквивалента: kurierst du Buby?

3. Некалькируемость (НК) входного сегмента возникает в связи с морфологической, синтаксической или семантической неоднозначностью одного или нескольких элементов входного и выходного словосочетаний. Здесь различаются следующие случаи.

а) НК, связанная с морфологической неадекватностью элементов входного и выходного сегментов.

Рассмотрим фрагмент немецко-русской двуязычной ситуации, связанной с соотнесением сегментов gelbe Banane — *желтый банан*. И синтаксис сегментов, и значение составляющих его элементов вполне адекватны; зато с морфологической точки зрения пары gelbe — *желтый*, Banane — *банан* неадекватны. Действительно, входной словоформе Banane соответствуют эквиваленты: *банан, банану, банана, (о) банане, бананом*; нем. gelbe будет переводиться русск. *желтый, желтая, желтое, желтую* . . .. Но и это еще не все, ибо входное и выходное существительные имеют разные родовые характеристики, что затрудняет выбор правильного русского морфологического эквивалента для немецкого прилагательного gelbe. Такая морфологическая некалькируемость обоих сегментов, представляющая значительную трудность при МП (напоминаем, что ЭВМ переводит текст формально-алгоритмическим путем), может быть снята только путем введения дополнительных программ морфологического анализа и синтеза.

б) НК, определяемая синтаксической неадекватностью элементов входного и выходного сегментов; ср. нем. двухчленный сегмент hau ruck! и его русский одночленный эквивалент *взяли!*

в) НК, вызванная семантической неадекватностью входного и выходного сегментов. Рассмотрим пары: нем. heute früh и русск. *сегодня утром*. Хотя каждая из немецких словоформ, взятая порознь, имеет однозначный перевод: heute — *сегодня*, früh — *рано*, входной сегмент не может быть калькирован с помощью пословного перевода. Дело в том, что früh, находясь рядом с heute, развивает новое, «связанное», значение — *утром*. Другими словами, для правильного перевода сегмента heute früh машине нужно задать специальную программу, учитывающую эту семантическую особенность нем. früh.

г) НК, обусловленная морфолого-синтаксической неадекватностью элементов входного и выходного сегментов; ср. нем. Ottos Vater — русск. *отец Отто*. Морфологическая неадекватность немецкой словоформы ее русскому переводу очевидна: ведь немецкому Vater соответствует не только *отец*, но и *отцу, отца, отцом, (об) отце* и т. д. Что же касается синтаксической неадекватности, то она обнаруживается в разном порядке слов немецкого и русского сегментов.

д) НК, имеющая морфолого-семантическую неадекватность элементов входного и выходного сегментов; ср. нем. Deutsche

Demokratische Republik — русск. *Германская Демократическая Республика*. Морфологическая неадекватность пар Demokratische — *демократическая*, Republik — *республика* аналогична морфологической неадекватности пар gelbe — *желтый*, Banane — *банан* (см. выше). Семантически неадекватными оказываются члены пары Deutsche — *Германская*, поскольку нем. deutsche имеет второе русское соответствие — *немецкий* (*немецкая*, *немецкое* и т. д.). В связи с этим правильный автоматический перевод сегмента Deutsche Demokratische Republik может быть получен только в результате сообщения машине дополнительных программ, устраняющих морфологическую и семантическую неадекватность.

е) НК, связанная с семантико-синтаксической неадекватностью элементов входного и выходного сегментов, может быть проиллюстрирована парой um so mehr — *тем более*. Нем. mehr семантически неадекватно русск. *более* (ср. другой его возможный перевод: *больше*); um so только в данном сегменте имеет соответствие *тем*, в свободных же употреблениях um и so имеют другие русские переводы (um — *вокруг*, *в*, *около*, *на*, *за* . . .; so — *так*, *таким образом*, *следовательно*. . .).

Трехчленному построению немецкого сегмента um so mehr соответствует двухчленное построение русского выражения *тем более*. В этом синтаксическая неадекватность входного и выходного словосочетаний.

ж) НК, проистекающую из морфолого-семантико-синтаксической неадекватности, рассмотрим на примере пары сегментов Westberliner über Attentat beunruhigt. . . («Neues Deutschland» от 4 июля 1969 г.) [5] — *жители Западного Берлина обеспокоены покушением*. . . Морфологическая неадекватность легко обнаруживается при сравнении пар: (West)berliner — *жители* (*Западного*) *Берлина*, Attentat — *покушением*, beunruhigt — *обеспокоены*.

Семантическая неадекватность обнаруживается практически у всех словоформ входного и выходного сегментов. Ср.: Westberliner — *жители Западного Берлина*, но и *западные берлинцы*. Аналогичным образом семантическое поле нем. Attentat покрывают два русских синонима: *покушение* и *посягательство*, нем. beunruhigen имеет два русских эквивалента: *беспокоить* и *тревожить*. Предлогу über соответствует большое количество русских эквивалентов: *над*, *через*, *по*, *более*, *свыше*, *о* и т. д.

Семантическую неадекватность легко обнаружить в построении сегментов Westberliner — *жители Западного Берлина*, в позициях причастных сказуемых beunruhigt — *обеспокоены*, а также в том, как оформляется управление зависящих от них существительных: beunruhigt + über + Akk. — *обеспокоены* + (твор. п).[6]

Уже знакомство с примерами, использованными в только что приведенной классификации, говорит о том, что полное калькирование сегмента входного языка сегментом выходного языка встречается чрезвычайно редко. Достаточно редки и случаи квазикалькирования, которые в условиях построенного на индукции вероятностного бинарного перевода мы можем разрешать путем пословного перекодирования (в условиях перевода через искусственный метаязык-посредник КК-обороты анализируются так же, как и НК-словосочетания). Основную же массу составляют некалькирующиеся сегменты. Центральная проблема МП поэтому состоит в том, чтобы определить наиболее экономные и эффективные приемы автоматического устранения некалькируемости входных сегментов на выходной язык.

[5] Данный сегмент представляет собой «усеченный» заголовок статьи. Соответствующая ему полная форма выглядела бы так: Die Westberliner sind über das Attentat beunruhigt. Этот сегмент также не калькируется русским переводом.

[6] Само собой разумеется, что все рассмотренные случаи неадекватности входного и выходного сегментов могут либо сопровождаться, либо не сопровождаться неизоморфностью парадигм адекватных элементов этих сегментов. В связи с этим приведенная классификация может быть значительно разветвлена и расширена. Однако, учитывая нужды бинарного вероятностного перевода, такая расширенная классификация не дает никаких видимых преимуществ с точки зрения решения интересующих нас вопросов.

Л. В. Малаховский

О СПОСОБАХ ОБРАБОТКИ ЗНАКОВ ПРЕПИНАНИЯ ПРИ АВТОМАТИЧЕСКОМ ПЕРЕВОДЕ С АНГЛИЙСКОГО ЯЗЫКА НА РУССКИЙ

В исследованиях по автоматическому анализу текста вопрос о знаках препинания обычно игнорируется. Однако информация, передаваемая в тексте знаками препинания, может оказаться для целей автоматического анализа весьма полезной. Исследователи, занимавшиеся этим вопросом,[1] получили результаты, которые могут быть использованы при составлении алгоритмов автоматического перевода (АП) с разных языков.

Здесь рассматривается вопрос об использовании информации, передаваемой знаками препинания в английском тексте, и о способах обработки знаков препинания при бинарном АП с английского языка на русский.

Знаки препинания могут нести в тексте разнообразную информацию. В английском языке, так же как и в русском, знаки препинания несут в основном информацию грамматическую — о конце полного предложения, о границе между простыми предложениями, о границе между однородными членами, об обособлении членов предложения и об иных синтаксических отношениях между ними и, наконец, информацию о пропуске слова.[2] Однако наряду с грамматической они могут нести также и другие виды информации,

[1] См., например: G. S a l t o n. A Method for Using Punctuation Patterns in the Machine Translation of Languages. In: Papers Presented at the Seminar in Mathematical Linguistics (Harvard university). 2, Cambridge (Mass.), 1956, pp. 1—55; Т. М. Н и к о л а е в а. Анализ знаков препинания при машинном переводе с русского языка. Сб. статей по машинному переводу, М., 1958, стр. 33—46; L. H i r s c h b e r g. Ponctuations et Analyse Syntaxique Automatique. Bruxelles, 1962.

[2] Некоторые знаки (точка, удлиненное тире, реже многоточие) могут передавать также информацию о пропуске части слова. В этом случае они выступают уже не как знаки препинания, а как графемы. В качестве графемы может выступать и косая черта. Подробнее об этом см.: Л. В. М а л а х о в с к и й. Об информации, содержащейся в небуквенных графемах английского языка. В кн.: Статистика текста. Сборник статей. Т. II. Автоматическая переработка текста. Минск, 1970, стр. 173—188.

например информацию семантическую (об употреблении слова в особом значении, об отношении автора к высказыванию, об авторской принадлежности речи) и стилистическую (о стилевой принадлежности слова или выражения, об эмоциональной окраске речи).[3]

Для разработки эффективной системы автоматического перевода важно знать, является ли информация, передаваемая знаками препинания, необходимой, существенной. Иначе говоря, не дублируют ли знаки препинания информацию, уже выраженную в тексте другими средствами?

В большинстве случаев дело обстоит именно так. Действительно, если оставить пока в стороне вопрос о семантической и стилистической информации, разнообразную грамматическую информацию мы можем извлечь из текста посредством синтаксического анализа, не прибегая к помощи знаков препинания. Это доказывается уже тем, что большинство предложений можно понять и без знаков препинания. Но это не значит, что знаки препинания в таких предложениях не нужны. Читать текст, игнорируя знаки препинания, вообще-то возможно, так же как можно ехать по шоссе, не обращая внимания на дорожные знаки и указатели. Однако понимание текста, лишенного знаков препинания, будет затруднено, а иногда и совсем невозможно.[4]

Дело здесь в том, что даже когда знаки препинания только дублируют информацию, передаваемую в тексте другими средствами (морфологическими и синтаксическими), они выражают ее более наглядно, эксплицитно, и при анализе текста это часто заставляет предпочитать их другим носителям информации. Убедительным примером этого является информация о конце предложения. При рассмотрении текста, не разделенного на предложения точками, в принципе почти всегда возможно определить границы между предложениями на основе только синтаксического анализа. Однако никто этого обычно не делает, так как гораздо проще и быстрее опираться на пунктуацию и считать границей предложения точку.

В этом случае (как, впрочем, и во многих других) трудно сказать, дублирует ли пунктуация синтаксическую информацию,

[3] См.: Н. А. К о б р и н а и Л. В. М а л а х о в с к и й. Английская пунктуация. 2-е изд., М., 1961.

[4] Всем известны примеры вроде: *казнить нельзя помиловать*, смысл которых зависит от того, как расставить знаки препинания. Такие примеры встречаются не очень часто, но и не так редко, как можно было бы думать. Ср. русск. *Он, — сказал я, — хотел ей помочь*, или: *Экспедиция отправляется завтра* (ср.: *Он сказал: — Я хотел ей помочь*; *Экспедиция отправляется завтра?*) или следующие английские предложения: They call their own profound boredom and incapacity to understand the beauty of life, virtue (Sk); It is this sense that gives to this forlorn uncomely land, power to speak to the spirit (Sk). Без запятой понимание этих последних предложений было бы сильно затруднено.

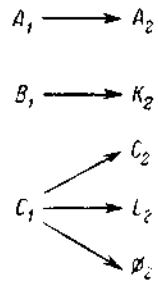
или наоборот, синтаксис дублирует пунктуацию: при наличии точки всякая другая информация о конце предложения является избыточной. Правильнее всего было бы считать, что информация о структуре текста передается разными средствами, в том числе синтаксисом, морфологией и пунктуацией, которые иногда дополняют, а иногда и перекрывают друг друга, и при разработке системы АП важно в каждом конкретном случае использовать то средство, которое позволяет осуществлять анализ текста наиболее просто и надежно.

Таким образом, следует ставить вопрос не об избыточности или неизбыточности пунктуации как носителя грамматической информации, а о ее эффективности как средства анализа текста в каждом конкретном случае ее применения. При грамматическом анализе эффективность пунктуации в определенных случаях может быть очень высока.[5] Тем более это относится к семантической и стилистической информации: пунктуация часто оказывается единственным носителем информации об авторской принадлежности речи, об употреблении слов в особом значении, их эмоциональной окраске и т. д.

Однако возможность использования пунктуации при автоматической обработке текста определяется не только тем, какую информацию она несет и какова ее эффективность как средства грамматического анализа. При бинарном АП подход к пунктуации зависит еще и от того, как соотносятся между собой пунктуационные системы рассматриваемых языков.

Здесь возможны по крайней мере три различных случая. Во-первых, одному знаку входного языка может всегда соответствовать тот же знак в выходном языке. В этом случае знаки могут рассматриваться как «цитаты» и переноситься в неизменном виде из входного языка в выходной. Во-вторых, может оказаться, что какому-либо знаку входного языка соответствует другой — но всегда один и тот же — знак выходного языка. Тогда вводится простое правило, предписывающее заменять данный знак на ему соответствующий.[6] Наконец — наиболее сложный случай — когда постоянного, прямого соответствия между знаками нет: данный знак входного языка может передаваться в выходном языке в одних условиях тем же знаком, в других — другим, а в третьих, вообще опускаться. В этом случае должны быть определены эти условия и составлен алгоритм перевода данного знака. Может оказаться, что определить полностью эти условия не удастся. Тогда данный знак при анализе придется игнорировать, а на выходе расстановка недостающих знаков препинания будет осуществляться постредактором.[7]

Рассмотренные три случая можно представить схематически следующим образом:



Здесь A_1, B_1, C_I — знаки препинания во входном языке, A_2, K_2 и т. д. — знаки препинания в выходном языке ($\varnothing$ — отсутствие знака). Случай, когда отсутствию знака во входном языке соответствуют какие-либо знаки в выходном, на схеме не представлен, так как он может быть сведен к одному из трех указанных случаев, если рассматривать отсутствие знака препинания как особый «нулевой» знак (см. ниже).

Сопоставление систем английской и русской пунктуации начнем с уточнения состава знаков препинания в этих языках.

Большинство знаков препинания являются для обоих языков общими. Это точка, вопросительный знак, восклицательный знак, точка с запятой, запятая, двоеточие, тире, многоточие, скобки (круглые и квадратные) и кавычки (двойные). Среди знаков, общих для обоих языков, следует назвать и «абзац».[8] В английском языке имеется, кроме того, три знака, которых нет в русском: кавычки одинарные (single quotation marks), удлиненное тире (two-em dash) и косая черта (virgule). В русском языке дополнительных знаков препинания нет.

---

[5] См. предыдущ. примеры Ср. также предложения: He tried every possible means, but in vain и I had no room for thoughts but two», где наличие запятой указывает, что слово является противительным союзом ('но'), а отсутствие запятой — что слово but является предлогом ('кроме'). В предложении: The signal is applied to the amplifier, which secures the desired result запятая указывает на отсутствие синтаксической связи между словами amplifier и which; отсюда и правильный перевод: Сигнал подается на усилитель, что и обеспечивает желаемый результат (при отсутствии запятой было бы: . . . на усилитель, который обеспечивает . . .).

[6] Это же правило может быть применено и в отношении первого случая, поскольку там для каждого знака также имеется постоянный эквивалент в другом языке. Таким образом, при составлении алгоритма обработки знаков препинания первый случай может быть сведен ко второму.

[7] Именно это и делается обычно в отношении 100% случаев употребления знаков препинания. Цель настоящей работы как раз и состоит в том, чтобы найти способы, позволяющие свести процент необрабатываемых случаев употребления пунктуационных знаков до минимума.

[8] На необходимость отнесения «абзаца» к числу знаков препинания указывают и другие исследователи (см., например: L. Hirschberg, ук. соч.; А. А. Реформатский. О кодировании и трансформации коммуникативных систем. В сб.: Исследования по структурной типологии, М., 1963, стр. 208—215).

Следует выделить еще один, общий для обоих языков, знак препинания. Это «нулевой» знак (∅), т. е. пробел, не заполненный каким-либо другим пунктуационным знаком. При исследовании системы пунктуации в пределах одного конкретного языка необходимость выделения «нулевого» знака препинания, возможно, и не ощущается. Однако при сопоставлении пунктуационных систем разных языков, когда выясняется, что тому или иному знаку одного языка соответствует отсутствие знака в другом, введение нулевого знака сильно упрощает процедуру исследования. Введение этого знака оправдывается и чисто теоретическими соображениями, поскольку функцию членения текста на слова нельзя не отнести к числу функций, выполняемых знаками препинания.[9] Между прочим, тот факт, что в некоторых древних памятниках функцию членения текста на слова выполняет особый знак (косая черта), показывает, что вводимый нами нулевой знак не всегда был «нулевым».

Существуют незначительные различия между языками в начертании отдельных знаков. Английское многоточие в отличие от русского печатается не единым наборным знаком, а тремя отдельно стоящими точками, причем в конце предложения оно ставится не взамен точки (как в русском языке), а в дополнение к ней (подробнее об этом см. ниже). Английские кавычки, как двойные, так и одинарные, ставятся, в отличие от русских только по верхнему срезу строки и повернуты к выделяемому ими материалу не выпуклой, а вогнутой стороной: "..." или '...'. Кавычки другого рисунка (« ») в английской издательской практике не применяются.

Таким образом, уже при рассмотрении состава знаков препинания обнаруживается частичное несовпадение систем английской и русской пунктуации. Имеются расхождения в употреблении знаков, общих для обоих языков.

Сравнение функций конкретных знаков препинания в английском и русском языках позволяет разделить английские пунктуационные знаки на три группы — в соответствии с описанными выше тремя возможными способами взаимодействия знаков препинания входного и выходного языков.

1. К п е р в о й г р у п п е относятся знаки, которые не нуждаются в переводе, знаки-«цитаты»: «абзац», точка,[10] вопросительный знак, восклицательный знак, тире, круглые скобки и квадратные скобки. Во всех случаях, когда эти знаки встречаются в английском тексте, их можно переносить без изменения в русский текст.[11] При этом могут возникать некоторые неточности, но очень незначительные. Так, в английском языке существует правило, по которому в конце предложения, содержащего вопросительную конструкцию, но выражающего не вопрос, а вежливую просьбу, ставится точка (а не вопросительный знак, как в русском языке), например: Will you kindly fill out and return this questionnaire (Webst.) 'Не будете ли вы так любезны заполнить и вернуть эту анкету?'. Однако, не говоря о том, что эти случаи очень редки, подобные предложения могут переводиться на русский язык и без вопросительной конструкции ('Будьте любезны заполнить и вернуть эту анкету'). К тому же и в английском языке иногда в таких случаях ставится не точка, а вопросительный знак, например: David, will you please sit down? (Mast.).

Некоторые неточности могут возникать и при переносе из английского текста в русский точки с запятой, так как этот знак может, хотя и очень редко, употребляться и для разделения однородных членов предложения, например: Think of the laws appertaining to real property; to the bequest and devise of real property; to the mortgage and redemption of real property; to leasehold, freehold, and copyhold estate. . . (Dick.).

В русском языке в подобных случаях принято ставить запятую, однако употребление точки с запятой является допустимым.

Что касается тире, то функции его в английском языке более многообразны, чем в русском. В частности, оно может использоваться (как и удлиненное тире) для обозначения незаконченности высказывания, длительных пауз или заминок в устной речи, а также для передачи эмоциональной авторской речи. В русском языке в таких случаях используется многоточие или восклицательный знак. Однако все эти случаи принадлежат стилю художественной литературы и при обсуждении проблем АП могут не рассматриваться. Все остальные функции английского тире — выделение вводного члена, вводного предложения, обособленных членов, приложения, а также отделение обобщающего слова от следующего за ним перечисления — совпадают с функциями тире в русском языке. Это и позволяет отнести тире к первой группе знаков препинания.

Остальные знаки первой группы могут во всех случаях переноситься в русский текст без изменений.

2. В т о р а я г р у п п а включает знаки, для которых в русском языке имеется постоянный переводной эквивалент. Сюда относятся одинарные кавычки, удлиненное тире и косая черта.

[9] Это не исключает возможности рассматривать пробел на графемном уровне как особую нулевую графему. Необходимо лишь учитывать, что нулевой графемой является любой пробел в тексте, тогда как нулевым знаком препинания — только «пустой» пробел, т. е. такой, который не заполнен другими пунктуационными знаками; см. также: Р. В. Макарова. Основные вопросы графики современного русского языка (алфавит и пунктуация). АКД. М., 1969, стр. 8.

[10] Употребление точки и других знаков в составе чисел (например, в десятичных дробях) в настоящей работе не рассматривается.

[11] Это не означает, что можно делать то же самое и при переводе с русского языка на английский. Например, английскому многоточию всегда соответствует в русском языке тоже многоточие, но функции русского многоточия чаще передаются в английском языке при помощи тире.

а) Одинарные кавычки обычно применяются, когда необходимо выделить слова, которые входят в состав текста, уже заключенного в двойные кавычки, а также слова, относящиеся к так называемой внутренней прямой речи.[12] Одинарным кавычкам соответствуют в русском языке кавычки другого рисунка („. . .“). Они могут использоваться и при выделении слов внутри отрывка, помещенного в другие кавычки («. . .»), и при выделении внутренней прямой речи.

б) Удлиненное тире употребляется (наряду с обычным тире) для обозначения незаконченности высказывания или длительных пауз в устной речи.[13] В текстах научно-технического характера этот знак не встречается. В русском языке удлиненному тире во всех случаях соответствует многоточие.

в) Косая черта — знак, который в руководствах по английской пунктуации обычно не упоминается, но который в последнее время устойчиво употребляется в научно-технических и публицистических текстах. Косая черта, стоящая между двумя словами (чаще всего между словами and и or), означает, что «в истолковании смысла [предложения] может быть использовано как то, так и другое слово».[14] Например: This voltage is applied to a fused switch and/or a circuit breaker. 'Это напряжение подается на плавкий предохранитель или на прерыватель цепи или на то и другое одновременно'.

Косую черту удобно передавать в русском языке при помощи скобок. Тогда данный пример можно перевести следующим образом: 'Это напряжение подается на плавкий предохранитель и (или) на прерыватель цепи.[15]

3. К третьей группе относятся знаки, которые передаются на русский язык по-разному, в зависимости от условий. В эту группу входят все остальные знаки: запятая, нулевой знак (∅), двоеточие, двойные кавычки и многоточие.

Рассмотрим сперва более простые соответствия, связанные с переводом трех последних знаков.

а) Английское двоеточие чаще всего передается в русском языке двоеточием же. Имеется всего несколько случаев, когда оно должно передаваться в русском языке другим знаком.

[12] См.: Н. А. Кобрина и Л. В. Малаховский, ук. соч., стр. 87—93.

[13] Об использовании удлиненного тире в качестве графемы см. прим. 2 (стр. 452).

[14] Webster's Collegiate Dictionary. 5th ed., Springfield (Mass.), 1939, pp. 1122; см. также прим. 2 (стр. 452).

[15] Косая черта иногда прямо переносится переводчиками в русский научно-технический текст (например: и/или), однако такое употребление нельзя признать правильным хотя бы потому, что оно непонятно большинству читателей.

Во-первых, двоеточие ставится иногда после обращения в деловом письме или приветственном адресе: Dear Mr. Saxl: It was a pleasure to speak to you . . . (Mast.). В этом случае двоеточие следует заменять на восклицательный знак. Сигналом для такой замены могут служить слова: Dear Mr. . . ., Dear Sir, Dear Dr. . . . и т. п.

Во-вторых, двоеточие может использоваться для разделения частей сложносочиненного предложения: I muttered something: then I enquired how many people knew (Ch. Snow). В русском языке в таких случаях ставится точка с запятой или запятая; однако формальных признаков, сигнализирующих о необходимости такой замены, вероятно, нет.[16] Поэтому двоеточие здесь придется сохранять, что почти не увеличит процент ошибок, поскольку случаи эти встречаются крайне редко, а в текстах научно-технических и публицистических, по-видимому, вообще не встречаются.

И наконец, двоеточие ставится перед приложением, стоящим в конце предложения: The only work available for his people is seasonal: blueberry-picking in June and potato-picking in autumn (W). При переводе на русский язык двоеточие следовало бы заменить здесь на тире; однако сохранение двоеточия не искажает смысла предложения и не затрудняет его понимания.

Двоеточие, кроме того, может употребляться в сочетании с тире после слов, которые вводят прямую речь, начинающуюся с абзаца. В этом случае сочетание знаков «двоеточие+тире+абзац+кавычки» должно заменяться в русском тексте на сочетание «двоеточие+абзац+тире».

б) Английским двойным кавычкам соответствуют в русском языке тоже кавычки (обычно рисунка «») — во всех случаях, кроме прямой речи. Кавычки при прямой речи могут сохраняться, если прямая речь в русском тексте дается в строку. Если же она начинается с абзаца, то открывающие кавычки заменяются на тире, а закрывающие — снимаются. Можно выработать систему довольно простых правил, по которым английские кавычки будут преобразовываться в тот или иной пунктуационный знак русского текста — в зависимости от того, с каким знаком в английском тексте они сочетаются. Формальным признаком прямой речи являются следующие сочетания знаков: «двоеточие+кавычки (открывающие)», «запятая+кавычки», «двоеточие+абзац+кавычки» или (реже) «двоеточие+тире+абзац+кавычки». При этом и закрывающие кавычки должны обязательно сочетаться с запятой или с точкой (или со знаком, заменяющим точку, — вопросительным, восклицательным, многоточием или

[16] Дело в том, что формально эти предложения не отличаются от предложений, в которых двоеточие должно сохраняться, например: We never saw that lumber: all of it was resold (W).

тире). При одновременном наличии указанных признаков текст, помещенный между открывающими и закрывающими кавычками, может рассматриваться как прямая речь, а кавычки (и сочетающиеся с ними знаки) — заменяться на те знаки, которыми оформляется прямая речь в русском языке. Например, сочетание |„текст“| преобразуется в сочетание |:А — текст| (где буква А обозначает абзац) и т. п. Во всех остальных случаях, как уже было сказано, никаких преобразований не производится и кавычки просто переносятся из английского текста в русский (разумеется, с изменением их рисунка).

в) Многоточие отнесено к третьей группе по чисто формальным соображениям. Английскому многоточию всегда соответствует в русском языке тоже многоточие. Однако в некоторых случаях, а именно когда многоточие следует после точки, вопросительного или восклицательного знака, количество точек, из которых оно состоит, уменьшается на одну. Таким образом, английские сочетания |....|, |?...| и |!...| передаются в русском тексте знаками |...|, |?..| и |!..|.

Значительно сложнее обстоит дело с запятой и нулевым знаком. Создание алгоритма обработки этих знаков представляет собой тему для самостоятельного исследования. Здесь же делается попытка лишь наметить пути к решению этой задачи.

Исходным пунктом при разработке такого алгоритма может служить выяснение того, какие вообще пунктуационные знаки русского текста могут служить в качестве эквивалентов английской запятой и нулевого знака.

Сопоставление правил английской и русской пунктуации показывает, что английская запятая может передаваться в русском языке либо запятой же, либо нулевым знаком. Имеется всего два случая, когда запятая передается другими знаками.

Во-первых, в английском языке запятая ставится в эллиптических предложениях, при пропуске глагола-связки или иного члена предложения, например: Paris is the capital of France; Madrid, of Spain (Sk.). В этом случае запятая должна при переводе заменяться на тире.

Во-вторых, запятая ставится в начале письма после обращения (в деловых письмах чаще ставят двоеточие; см. выше). Здесь при переводе требуется заменять запятую на восклицательный знак.

Формальные признаки, требующие перехода от запятой к восклицательному знаку или, как в предыдущем случае, к тире, выявить, по-видимому, сравнительно легко. Что касается двух основных возможностей (сохранение запятой или переход к нулевому знаку), то в большинстве случаев требуется запятую сохранять. Можно отметить всего несколько случаев, когда английская запятая передается в русском тексте нулевым знаком:

1) когда запятая разделяет соподчиненные придаточные предложения, соединенные союзом and: The Foreign office was satisfied that the terms of the treaty were being compiled with, and the files were then passed to the Home Office (MS);

2) когда запятая стоит перед последним из нескольких однородных членов, присоединенным к предшествующему ряду при помощи союза and или or: She lifted her eyebrows, smiled, turned her nead on its long neck, *and did not care about Julia* (Less.);

3) когда запятая отделяет обособленное дополнение, стоящее в начале предложения: *For materialism*, thought is a product of matter and a reflection of matter (Cornf.);

4) когда запятая отделяет ряд однородных членов от следующего за ними члена предложения; например, ряд однородных подлежащих отделяется запятой от общего для них сказуемого: *Hard times, and hard weather, and hard work*, make it trying now and then (Dick.). В английском языке существует правило, согласно которому последняя из нескольких однородных частей аналитического сказуемого отделяется запятой от общей для них части сказуемого: Anyone who *cannot, does not, or will not*, fight must leave the field (Sk.). Точно так же отделяются запятой несколько глаголов от общего для них дополнения. Подобным же образом, если к одному и тому же существительному относится два (или более) предлога, то после каждого из них ставится запятая, в том числе и после последнего, который, таким образом, отделяется от этого существительного запятой: [It is] . . . a unique centre for treatment *of*, and research *into*, mental deficiency (DW).

Для случаев, перечисленных в пункте 4, по-видимому, возможна выработка достаточно простых формальных правил перехода от запятой к нулевому знаку (например, если в предложении имеется два или более предлога, после каждого из которых стоит запятая, то при переводе запятые отбрасываются, а все предлоги соотносятся с тем существительным, которое идет за последним из них). В случаях же, перечисленных в первых трех пунктах, замена запятой на нулевой знак (т. е. устранение запятой) может производиться при синтезе на основании нескольких формальных правил (запятая между однородными членами, соединенными союзом 'и', устраняется; также снимается запятая, отделяющая сказуемое от дополнения, и т. п.).

Нулевому знаку препинания в подавляющем большинстве случаев соответствует в русском тексте также нулевой знак. В более редких случаях нулевой знак должен заменяться на запятую и в одном случае — на кавычки. Этот единственный случай относится к названиям книг, журналов, газет, кораблей и т. п. Такие названия в английском языке обычно в кавычки не берутся, а выделяются курсивом или прописными буквами. Не говоря уже об относительной маловажности этого случая, на-

личие шрифтового выделения, вероятно, может быть использовано для выработки правил замены нулевого знака на кавычки.[17]

Случаи, когда английскому нулевому знаку соответствует в русском языке запятая, связаны со следующими особенностями английской пунктуации:

1) между частями сложносочиненного предложения, соединяемыми при помощи союзов and, or или but, может не ставиться никакого знака: The snowstorms came down in a yellow darkness and every lull was followed by a fresh fury of snow (Linds.);

2) придаточное предложение во многих случаях не отделяется от главного: There is not the slightest sign in this speech that the U. S. authorities have drawn any lessons from the defeats they have suffered in Vietnam (MS);

3) причастные обороты также не всегда выделяются запятыми:[18] There have been several outbreaks of fire in the multiple diesel units introduced some years ago on this line (MS);

4) обычно не разделяются запятой однородные члены, соединенные союзами but, though, as well as или парными союзами both . . . and, either . . . or, neither . . . nor: All Stilleveld . . . had also been the birthplace of new people. A people who were neither white nor black; neither Europeans nor Africans but a blending of the two that was at once different from both white and black . . . (Abr.).

Выработка формальных правил замены нулевого знака на запятую для первых трех случаев принципиальных трудностей не вызывает; в последнем же случае она вообще очень проста, поскольку в качестве сигнала такой замены выступают (при синтезе) сами союзы: но, хотя и т. п.

Подведем некоторые итоги.

1) Знаки препинания являются одним из источников информации о грамматической структуре текста. Они обычно лишь дублируют информацию, передаваемую в тексте другими средствами, но благодаря присущей им эксплицитности оказываются в ряде случаев более удобными для использования в АП.

2) Традиционные списки пунктуационных знаков английского и русского языков необходимо пополнить такими знаками, как «абзац» и «нулевой знак». Это позволит дать более точное описание пунктуационных систем и значительно облегчит обработку знаков препинания при АП.

3) Английские знаки препинания можно разбить, в целях бинарного АП, на три группы: знаки, которые могут прямо переноситься из английского языка в русский в неизменном виде; знаки, которые хотя и не переносятся без изменений, но имеют постоянный переводной эквивалент в русском языке; знаки, которые имеют по нескольку переводных эквивалентов, выбор которых зависит от словесного и пунктуационного окружения. Знаки первой группы могут рассматриваться и обрабатываться при АП как «цитаты» (т. е. как математические или иные символы или формулы), знаки второй группы — как однозначные слова, а знаки третьей группы — как слова многозначные.

4) Обработка при АП знаков-«цитат» и однозначных знаков препинания с английского языка на русский не вызывает трудностей; для многозначных же знаков в большинстве случаев могут быть определены формальные признаки, по которым выбирается нужный знак при переводе. Наиболее сложным является вопрос об определении таких признаков для запятой и нулевого знака. В нашей статье намечены пути для решения этого вопроса.

5) Некоторые знаки препинания находятся в отношениях омонимии с совпадающими с ними по начертанию небуквенными графемами. Вопрос о формальных способах преодоления этой омонимии должен решаться особо.

[17] Следует, по-видимому, поставить вопрос об отнесении шрифтового выделения и особенно прописных букв к системе пунктуационных средств языка.

[18] О знаках препинания при придаточных предложениях и причастных оборотах см.: Н. А. Кобрина и Л. В. Малаховский, ук. соч., стр. 24—64.

Сокращения

Abr. — P. Abrahams
Cornf. — M. Cornforth
Dick. — Ch. Dickens
Less. — D. Lessing
Mast. — D. Masters
DW — The Daily Worker, London.
MS — The Morning Star, London
SK — R. Skelton. Modern English Punctuation. 2nd ed., London, 1949
W — The Worker, New York
Webst. — Webster's Collegiate Dictionary. 5th ed., Springfield (Mass.), 1939

СОДЕРЖАНИЕ

ИСПРАВЛЕНИЯ И ОПЕЧАТКИ

Страница	*Строка*	*Напечатано*	*Должно быть*
41	Табл. 19, правый стлб., 1 сверху	foreig,*	foreign,*
45	5—6 сверху	вероятности $P(\chi^2 \geqslant \chi_0^2) : P(\chi^2 \geqslant \chi_0^2) = \int_{\chi_0^2}^{\infty} P(\chi^2)\, d\chi^2$	вероятности $P(\chi^2 \geqslant \chi_0^2) = \int_{\chi_0^2}^{\infty} P(\chi^2)\, d\chi^2$
164	19 сверху	67581	67851
189	12 »	1349	1350
251	24 »	Einssachchleppern	Einachsschleppern
256	13 снизу	acide	acide,
309	3 »	артоматический	автоматический
348	5 »	Ambignity	Ambiguity
454	11 »	слово является	слово but является