ISSN 0207-4435

TARTU RIIKLIKU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

658

KVANTITATIIVNE LINGVISTIKA JA STILISTIKA

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И СТИЛИСТИКА

Töid keelestatistika alalt
Труды по лингвостатистике

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 658 ВЫПУСК ОСНОВАНЫ В 1893.г.

KVANTITATIIVNE LINGVISTIKA JA STILISTIKA

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И СТИЛИСТИКА

Töid keelestatistika alalt
Труды по лингвостатистике

TARTU 1983

Kogumik "Töid keelestatistika alalt" ilmub TRÜ Toimetiste sarjas alates 1976.a. Käesolevas IX väljaandes on avaldatud rida uurimusi kvantitatiivlingvistika ja stilistika kokkupuutealadelt. Autoriteks on TRÜ rakenduslingvistika uurimisgrupi liikmed ja väliskaastöötajad ning külalisautor Prahast, filoloogiadoktor Marie Těšitelová.

Сборник "Труды по лингвостатистике" публикуется в серии Ученых записок Тартуского университета начиная с 1976 г. Настоящий 9-й выпуск содержит статьи по квантитативной лингвистике и стилистике. Авторами статей являются сотрудники Исследовательской группы по прикладной лингвистике при ТГУ и исследователи из других городов, в том числе д-р Марие Тешителова из Праги.

The collections "Papers on Linguo-Statistics" have been published serially as issues of ACUT (Acta et Commentationes Universitatis Tartuensis) since 1976. The present issue No. 9 contains investigations in the field of quantitive linguistics and stylistics. The authors of the articles are members of the Research Group of Applied Linguistics at Tartu State University and a guest author from abroad, Dr Marie Těšitelová of Prague, Czechoslovakia.

ПОЭТИЧЕСКИЙ ТЕКСТ И ПРОБЛЕМА ЕГО АВТОРСТВА

В.И. Батов, Ю.А. Сорокин

Установление авторства текста является одним из центральных вопросов, решаемых в рамках не только литературоведения, но и искусствоведения, психологии искусства, криминалистики, т.е. дисциплин, изучающих или процессы словесно-изобразительного творчества, или вопросы практической атрибуции продуктов творческой деятельности человека.

В.В. Виноградов выделял следующие принципы атрибуции в литературоведении: I. Субъективные - а) субъективно-коммерческие; б) субъективно-конъюнктурные; в) субъективно-эстетические; г) субъективно-психологические; д) субъективно-идеологические.

II. Объективные - а) документально-рукописные (археологические); б) исторические (биографии, сведения современников и пр.); в) историко-идеологические и сопоставительно-идеологические; г) историко-стилистические; д) художественно-стилистические; е) лингво-статистические (Виноградов В.В.,1960).

Нельзя сказать, что с помощью этих принципов проблема атрибуции в целом уже решена. Отмечая невысокую эффективность большинства подходов к установлению авторства текста, некоторые литературоведы следующим образом оценивают положение в этой сфере: "К сожалению, положение с атрибуциями в нашем литературоведении трудно считать нормальным" (Рейсер А., 1970, с. 234).

Одним из наиболее перспективных считается подход, базирующийся на статистическом анализе индивидуальных языковых/речевых признаков текста, - подход, в основе которого лежит метод "лингвистических спектров" Н.А. Морозова (1915). Работы в рамках этого направления, по мнению В.В. Виноградова, кроме собственно атрибуции, содействуют решению вопроса о правильности выбора того или иного варианта анонимного текста (при разборе черновых записей), обосновывают правомерность той или иной конъюктуры.

Хотя инвариантность признаков, характеризующая письменно-речевой навык некоторого автора, может быть выявлена

только на достаточно большом речевом массиве (порядка нескольких тысяч словоформ) (Ворончак Б., 1972), что, несомненно, является недостатком работ лингво-статистического направления, "так как неатрибутированных произведений достаточно большого объема мы почти не знаем" (Гришунин А. Л., 1960, с. 151), но если все же воспользоваться оценками субъективного отражения этих же характеристик в сознании носителя языка, то процесс выявления сходимости этих характеристик (к некоторой постоянной величине) может быть значительно сокращен, что позволяет идентифицировать тексты объемом уже в несколько сотен словоформ (Батов В.И., Сорокин Ю.А., 1977). Под этим углом зрения исследовались только прозаические тексты и, есть основания полагать, что языковые/речевые признаки субъективных образов, служащие идентификаторами авторства прозаического текста, окажутся непригодными для анализа поэтических текстов. Это заставляет искать признаки, релевантные для атрибуции и поэтических текстов.

Гипотеза

Основные требования, предъявляемые к этим признакам сводятся к следующим: во-первых, эти признаки должны быть таковы, чтобы их можно было "приписать" всем возможным формам поэтических текстов; во-вторых, они должны описывать структуру текста, но не его содержание; в-третьих, эти признаки должны быть формализуемы в такой степени, чтобы можно было использовать операциональные методы анализа художественных (поэтических) текстов.

В значительной степени этим условиям отвечают те признаки текста, которые сигнализируют о концептуальных схемах движения идеи произведения (Бахтин М., 1979).

В нашей работе эти схемы интерпретируются следующим образом.

Отношение авторского "Я" к чужому "Я". Это отношение двуполюсно: на одном полюсе находятся утверждения об отношении к "чужому Я" не как к объекту познания, а как к равноценному авторскому "Я" некоторому субъекту. Соответственно на другом полюсе находятся утверждения об отношении к "чужому Я" как к объекту познания.

Способы манифестации замысла произведения: монолог - однонаправленное движение идеи произведения из одного какого-либо локуса сообщения в некоторый другой; диалог - дву-

направленное причинно-следственная циркуляция идеи от некоторого икса к некоторому игреку; полифония - движение идеи из разных локусов к некоторым другим локусам.

Временная протяженность идеи произведения реализуется, как правило, в синхронической или диахронической "событийности" сообщения.

Форма конкретного речевого сообщения конструируется путем взаимодействия указанных гипотетических схем. Требуется установить, какие же формы репрезентации замысла/идеи произведения "порождаются" этим взаимодействием и какова значимость каждой из форм поэтического текста в реализации его замысла тем или иным автором.

МЕТОД ИССЛЕДОВАНИЯ. РЕЗУЛЬТАТЫ

Объектом настоящего исследования послужили 12 поэтических текстов (верлибр), принадлежащих трем авторам*. Каждый текст оценивался восемью испытуемыми, имеющими филологическое образование, с помощью ряда шкал, аналогичных шкалам метода "семантического дифференциала", представленных в таблице № I. Усредненные порядковые оценки (в форме медиан) текстов обрабатывались на ЭВМ по методу главных компонент факторного анализа с последующим вращением факторов. Цель такого анализа, как известно, состоит в выявлении факторов, связывающих шкалы между собой, т.е. в выявлении факторов, структурирующих семантическую массу поэтического текста. Результаты этого этапа исследования приведены в таблице № I.

Эти результаты свидетельствуют о том, что такие поэтические тексты, как верлибр характеризуются четырьмя различными формами презентации замысла автора с точки зрения читателей этих текстов.

Первая форма структурируется, в основном, с помощью четырех шкал (имеющих статистическую значимость и отмеченных в таблице № I звездочками). Эта форма репрезентации авторского замысла наиболее весома среди всех остальных используемых авторами форм верлибра (вес 35,28%). Характерная черта этой формы - противопоставление "диалога между персонажами" "диалогу с читателем", "полифонии" и "монологу". Такое

* Винокуров Е. Избранное из девяти книг. М., 1968; Сулейменов О. Глиняная книга. Алма-Ата, 1969; Луговской В. Середина века. Книга поэм. М., 1958.

Таблица I

ФОРМЫ КОММУНИКАЦИИ ВЕРЛИБРА

Семантическая шкала	Интракоммуникативный диалог	Интеркоммуникативный диалог	Альтернативность	Хронологичность
	35,28%	29,05%	I2,77%	II,02%
Субъектность идеи - Объектность идеи	0,80*	-0,I8	-0,39	0,08
Диалог между персонажами - Диалог с читателем	-0,73*	0,50	-0,I8	0,I0
Диалог между персонажами - Полифония	-0,80*	-0,53	0,08	0,I0
Диалог между персонажами - Монолог	-0,59*	0,05	0,20	0,39
Диалог с читателем - Монолог	0,05	-0,90*	0,33	0,05
Диалог с читателем - Полифония	0,23	-0,90*	0,0I	0,I0
Монолог - Полифония	0,46	0,34	0,8I*	-0,04
Синхрония - Диахрония	0,46	0,I5	0,I6	0,83*

противопоставление "диалога между персонажами" всем остальным шкалам позволяет, по-видимому, считать эту форму реализации замысла "интракоммуникативным диалогом", диалогом "внутри" произведения. Интракоммуникативный диалог такого типа сигнализирует об объективности позиции автора, т.е. о процессе познания им "чужого Я" как отчужденного от авторского "Я".

Вторая форма верлибра структурируется с помощью двух шкал, находящихся в обратной зависимости друг к другу. Эта форма репрезентации авторского замысла также достаточно широко распространена (вес 29,05%). Характерная ее особенность состоит в том, что "диалог с читателем" противопоставлен "монологу" и "полифонии". Такое противопоставление позволяет считать данную форму верлибра "интеркоммуникативным диалогом", диалогом, выходящим за рамки произведения.

Третья форма репрезентации авторского замысла (вес 12,77%) представлена одной семантической шкалой, а именно шкалой "монолог-полифония". Мы предлагаем называть эту форму верлибра "альтернативной".

Четвертая форма верлибра (вес 11,02%) представлена также одной шкалой, а именно шкалой "синхрония-диахрония". Эта форма реализации авторского замысла фиксирует временную развертку событий в произведении и поэтому может быть названа "хронологической".

Реальный поэтический текст (верлибр) представляет собой некоторую систему вышеприведенных форм репрезентации авторского замысла. В частности, эта система может быть организована как соподчинение форм. Выяснению предпочтения, оказываемого некоторым автором той или иной форме верлибра, и был посвящен второй этап нашего исследования.

На этом этапе, во-первых, с помощью регрессионного метода Томсона (см. Лоули Д., Максвелл А., 1967) определялись множественные линейные регрессии, связывающие значения гипотетических факторов (форм верлибра) с исходными оценками реципиентов (для того, чтобы определить значения отдельных форм верлибра, порядковые оценки, полученные от информантов, были переведены в интервальные).

Во-вторых, проводился дисперсионный анализ индикативной функции форм верлибра (на основании имеющихся значений выраженности этих форм) для экспликации авторства текста.

Результаты второго этапа нашего исследования представлены в таблицах № 2 и № 3.

Таблица 2

ВЛИЯНИЕ "ФОРМЫ" И "АВТОРСТВА" НА СЕМАНТИЧЕСКУЮ СТРУКТУРУ ВЕРЛИБРА

ГИПОТЕЗА	ФОРМА	АВТОРСТВО
	Фактическое значение критерия Фишера (табличное значение Фишера 4,8)	Фактическое значение критерия Фишера (табличное значение Фишера 5,I)
Раздельное влияние факторов на текстовую семантику	I3,37 (значимо)	0,02 (незначимо)
Взаимодействие факторов (табличный критерий Фишера 2,30)	I,39 (незначимо)	

В таблице № 2 приведены результаты двухфакторного дисперсионного анализа влияния факторов формы и авторства текста на семантическую структуру верлибра. Эти данные свидетельствуют о том, что фактор формы репрезентации верлибра статистически значимо влияет на семантическую организацию этих текстов, тогда как фактор авторства текста статистически не значим. По-видимому, полученные данные позволяют сделать следующие предположения.

У авторов проанализированных текстов существует предопределенная (например, личным опытом в художественно-словесной деятельности) тактика выбора строго определенных форм реализации идеи/замысла (произведения).

Не для всех авторов, тексты которых были проанализированы, справедливо первое предположение. Иными словами, среди наших авторов есть и такие, которые разновероятно используют ту или иную форму верлибра для реализации своего замысла.

Таблица 3

ВЛИЯНИЕ "ФОРМЫ" НА СЕМАНТИЧЕСКУЮ СТРУКТУРУ КОНКРЕТНЫХ ТЕКСТОВ

АВТОРЫ ТЕКСТОВ	Фактическое значение критерия Фишера (табличное значение Фишера 3,50)
Е. Винокуров	5,90 (значимо)
О. Сулейменов	53,24 (значимо)
В. Луговской	2,04 (незначимо)

Уточнить выводы нашего исследования помогают данные, приведенные в таблице № 3. Эти данные бесспорно свидетельствуют в пользу правильности второго предположения, а именно, о том, что такие авторы, как Винокуров и Сулейменов избирательно используют ту или иную форму верлибра. Иными словами, для каждого текста, принадлежащего этим авторам, характерна некоторая ведущая форма реализации авторского замысла, а в более широком смысле - характерны немногие, но структурно тождественные формы, независимые от содержательного плана текста. И поэтому следует с уверенностью полагать, что для этих авторов форма репрезентации поэтического текста является идентификационным признаком. Однако, этот вывод справедлив не для всех наших авторов.

Другое положение характерно для Луговского: по-видимому, выбор формы поэтического текста этим автором предопределяется целым рядом экстралингвистических (например, индивидуально-психологических) факторов, которые предстоит еще выяснить в планируемом исследовании.

Таким образом, может быть сделан следующий основной вывод из результатов нашего исследования.

Семантические признаки формы репрезентации авторской идеи в такого вида поэтических текстах, как верлибр могут служить объективной основой для установления авторства текста. Естественно, не ко всем текстам такого типа могут быть применены изложенные методические процедуры, но все же вышеизложенный метод анализа верлибра позволяет получить некоторые нетривиальные результаты в отношении этой поэтической формы художественного текста.

Л И Т Е Р А Т У Р А

Батов В.И., Сорокин Ю.А. Опыт построения методики для установления авторства текстов. - Известия АН СССР, серия литературы и языка, 1977, т. 36, № 4.

Бахтин М. Проблемы поэтики Достоевского. М., 1979.

Виноградов В.В. Теория литературных стилей и принципы атрибуции анонимных и псевдонимных произведений. - В кн.: О принципах определения авторства в связи с общими проблемами теории и истории литературы. Л., 1960.

Ворончак Б. Методы вычисления показателей лексического богатства текстов.-В кн.: Семиотика и искусствометрия.М., 1972.

Гришунин А.Л. Опыт обследования употребительности языковых дуплетов в целях атрибуции. - Вопросы текстологии.Вып. 2. М., 1960.

Лоули Д., Максвелл А. Факторный анализ как статистический метод. М., 1967.

Морозов Н.А. Лингвистические спектры. - ИОРЯС, т. 20, кн.4, 1915.

Рейсер С.А. Палеография и текстология нового времени. М., 1970.

THE POETICAL TEXT AND THE PROBLEM OF AUTHORSHIP

V.I. Batov and Yu.A. Sorokin

S u m m a r y

A new method of analyzing poetical texts is discussed. Twelve texts in vers-libre belonging to three authors were evaluated by eight readers with philological education using a series of scales analogous to those of the method of semantical differentials. The factor analysis based on the medium rank values of the texts revealed four main factors structuring the "semantic mass" of the poetical texts and pointing to the form of the representation of the artist's intention from the reader's point of view (Table 1). With the help of regression and variance analysis the authors' preference for one or another form of representation was revealed. It was established that two authors out of three used a stable form of representing the artist's intention irrespective of the subject-matter of the text. This makes it possible to a certain extent to use the method for determining the authorship of anonymous texts.

ОБ АВТОРСТВЕ ТЕКСТА "МОЯ СЕМЬЯ"

Г.В. Ермоленко

К настоящему времени уже вышло несколько томов академического издания произведений А.П. Чехова. Авторы вступительной статьи к примечаниям в т. 3 (Долотова Л.М., Соколова М.А., 1975, с. 540) отмечают, что в него не включены подписи под серией рисунков "Моя семья", (журнал "Зритель", 1883, № 19), так как принадлежность текста Чехову остается недоказанной. Между тем, в 68 томе "Литературного наследства", вышедшем в свет в 1960 году к 100-летнему юбилею великого писателя, Б.Д. Челышев высказывает предложение о возможной принадлежности этого текста Чехову. Свое предположение Б.Д. Челышев основывает на том, что в оглавлении против рассказа "Моя семья" стоит криптоним С,Б.Ч., который содержит те же буквы, что и подпись Чехова под рассказом "Патриот своего отечества" (но там буквы расположены в обратном порядке - Ч.Б.С.). Надо полагать, что это сокращение псевдонима Чехова "Человек без селезенки".

"Такое предположение, - заключает Б.Д. Челышев, - кажется тем более вероятным, что "Моя семья" по содержанию, стилю весьма близка к другим произведениям Чехова этих лет (ср., например, рассказ "Мои жены"); основу составляет перечисление характерных черт персонажей. Не противоречит предположению об авторстве Чехова и язык "Моей семьи" (Челышев Б.Д., 1960, с. 124).

Науке известно немало случаев ошибок при атрибуции текстов только по подписям. Случалось так, что иные подписи под каким-нибудь рассказом или заметкой появлялись по прихоти редактора журнала или газеты или по каким-либо другим причинам.

Установление авторства текста всегда является проблемой комплексной. Чем большее количество методов удается использовать в ходе исследования, тем глубже мы сможем проникнуть в суть проблемы. В данном случае мы имеем в виду некоторые статистические приемы исследования. В иных случаях они предоставляют исследователю ряд новых сведений об анализируемом тексте. Но не следует забывать, что статистический анализ текста проводится лишь с известной точностью и определенной

надежностью. Иными словами, он носит вероятностный характер. Но при большой точности и надежности, выводы, сделанные на основе результатов статистического анализа, могут максимально приблизиться к достоверным констатациям, полностью соответствующим языковой действительности.

Сравнение частот некоторых речевых параллелей и отдельных фразеосочетаний

Материал, выявленный нами для сопоставления, отобран из первых пяти томов ПССП, т.е. было просмотрено все написанное Чеховым за период с 1880 г. по 1886 г. Мы ограничили себя этими рамками не случайно. Во-первых, атрибутируемый текст относится к 1883 году, во-вторых, начиная с 1888 года Чехов уже не пишет юморесок. Он прекращает свое сотрудничество в юмористических журналах. Сужение периода до 1886 года необходимо, чтобы частично нейтрализовать результаты эволюции авторского стиля. Это важно при осмыслении и некоторых статистических сведений, полученных из просмотренных текстов. Так, например, в рассказах, написанных в период с 1887 года по 1900 год и отобранных Чеховым для собрания сочинений, он уже совсем не употребляет в авторской речи неусеченные формы отчеств мужчин (Иванович, Сидорович в противовес Иваныч, Сидорыч), а в период с 1880 по 1886 годы они составляют в среднем около третьей части отчеств мужчин.

В этот период встречаются случаи, когда в одном и том же рассказе автор употребляет как усеченную, так и неусеченную форму мужских отчеств. Ср.: 1. "Вызвал он, между прочим, дядю своего Клавдия Мироновича..." и несколькими строками ниже: "Огонь мелькал и еле освещал киот и большой портрет Клавдия Мироныча." (IУ/I2). 2. "Николай Андреевич Капитонов, нотариус, пообедал, выкурил сигару и отправился к себе в спальную отдыхать." и на этой же странице: "Николай Андреич встал с постели..." (IУ/I58).

В рассматриваемый период подобные случаи параллельного употребления обеих форм нередки. То же самое наблюдается в атрибутируемом тексте "Моя семья" (текст А). Автор пишет о первом брате своей жены: "Брат моей жены: Иван Емельянович ..." Останавливаясь на характеристике второго брата, Пантелея, тот же автор пишет: "Два раза судился за буйство и два раза по настоянию Ивана Емельяныча подавал апелляцию".

Частоты речевых параллелей и фразеосочетаний (Копыленн-

ко М.М., Попова З.Д., 1972, с. 32-33) только тогда могут быть "взвешены", когда исследователь располагает сведениями о них и из текстов какого-нибудь другого писателя, современного предполагаемому автору. Эти явления отличаются более частой встречаемостью. Естественно, что эти речевые параллели должны присутствовать в атрибутируемом тексте с частотой хотя бы равной единице.

Кроме рассматриваемой речевой параллели "Иваныч"/"Иванович" нами получены сведения и о других языковых явлениях. (См. таблицу I).

Частоты этих языковых явлений получены в результате анализа не равных по объему текстов. Чтобы сравнить их между собой, мы использовали критерий согласия "хи-квадрат"Усеченные формы отчества мужчин ("Иваныч"), существительное "дверь" (ед. ч.) и фразеосочетание "то и дело" ограничивают тексты Чехова от рассказов Куприна: расхождение частот этих явлений существенно. Расхождение частот остальных языковых явлений несущественно. При этом, фразеосочетание "несколько лет тому назад" оказалось общеречевым признаком (Ермоленко Г.В., 1977, с. 98), а данные о частотах "день и ночь" малодостоверны, так как число 22 меньше 33, которое принимается в качестве контрольного при величине надежности $\beta = 0,75$ и точности $\delta = 0,10$ (Ермоленко Г.В., 1977, с. 99). Итак, перечисленные I,3 и 7 явления надо считать признаками авторской принадлежности Чехову. Частоты их не различаются существенно от частот соответствующих явлений в тексте А и существенно отличаются от частот в рассказах Куприна (проверено по "хи-квадрат").

Следовательно, по использованию этих языковых явлений атрибутируемый текст примыкает к рассказам Чехова и не примыкает к рассказам Куприна.

О некоторых структурно-семантических особенностях текста А и чеховских рассказов (I-9)

Учитывая внутренние, чисто языковые особенности текста А, мы считаем, что его надо сопоставить с группой чеховских рассказов-характеристик (Чехов А.П., 1974, 1976, 1977). Это следующие рассказы и юморески (располагаем их в хронологическом порядке (n - длина текста в словоупотреблениях):

I. "По-американски", т. I (Стрекоза", 1880, № 49,7 декабря).

$n_I = 481$ - $Ч_I$

Абсолютные частоты речевых параллелей и фразеосочетаний

Таблица I

языковые явления / Авторы	"Иваныч"	"Иванович"	дверь (ед.ч.)	дверь (мн.ч.)	несколько лет тому назад	несколько лет назад	то и дело	день и ночь	Длина авторского текста (в словоупотреблениях)
	I	2	3	4	5	6	7	8	
Чехов т.т. I-5	486	I96	3I9	3I	50	4	78	22	28I000
Куприн т.т. I-2	5	I77	92	48	25	0	24	I	I93000
А	I	2	I	0	I	0	I	2	662

2. "Темпераменты", т. I ("Зритель", I88I, № 5, I7 сентября). n_2= I004 - $Ч_2$
3. "Свадебный сезон", т. 3 ("Зритель", I88I, № I8, ноябрь). n_3= 234 - $Ч_3$
4. "Он и она" (включаются только письма-характеристики его и ее), т. I ("Мирский толк", I882, № 26, 23 июля). n_4= I380 - $Ч_4$
5. "К характеристике народов", т. 3 ("Осколки", I884, № 46, I7 ноября). n_5= 487 - $Ч_5$
6. "Мои жены" т. 4 (начиная с характеристики жены № I)("Будильник", I885, № 24, июль). n_6= I408 - $Ч_6$
7. "К свадебному сезону" т. 4 ("Осколки", I885, № 30,28 сентября). n_7= 238 - $Ч_7$
8. "Руководство для желающих жениться", т. 4 ("Осколки", I885, № 44, 2 ноября). n_8= II04 - $Ч_8$
9. "Ряженые", т. 4 ("Петербургская газета", I886, № I,I января). n_9= 72I - $Ч_9$

Все эти рассказы и юморески содержат только авторскую речь, их легко можно было бы сопроводить рисунками. По своим композиционным и синтаксическим особенностям они схожы между собой. Кроме этих 9 текстов других рассказов подобного рода у Чехова нет.

Для установления значимости статистических сведений, которые были выявлены из рассказов А.П. Чехова, мы отобрали 4 одножанровых рассказа А.И. Куприна из серии "Киевские типы". Это тоже рассказы-характеристики. Все они были опубликованы осенью I895 года в газете "Киевское слово". Они помещены в т. I собрания сочинений Куприна (Куприн А.И.,I970, с. 388, 394, 399, 4I7).

I. "Певчий". n_I= 503 - $К_I$. 2. "Квартирная хозяйка". n_2=604- -$К_2$. 3. "Будущая Патти". n_3= 860 - $К_3$. 4. "Доктор"(начиная с I. "Доктор веселый"). n_4= 480 - $К_4$.

При статистическом анализе текстов реплики персонажей опускались.

Для выявления некоторых структурно-семантических особенностей текста А и чеховских рассказов был составлен час-

тотный список слов рассказа "Моя семья".

Относительные частоты слов в частотном списке текста А соответственно равны или почти равны относительным частотам этих слов в лексике русского языка (Частотный словарь русского языка, 1977), (ЧСРЯ - 77) (проверено по критерию "хи-квадрат"). Исключение составляют только 4 слова: частица НЕ и местоимения Я, МОЙ и ОНА. В тексте эти лексические единицы используются значительно чаще, чем в русском языке (расхождение частот существенно).

а/ частица НЕ в атрибутируемом тексте встречается в 2 раза чаще (ℓ = 0,038), чем в русском языке (ℓ_1 = 0,019), согласно ЧСРЯ - 77. Такая "концентрация" частицы НЕ в тексте А является одной из структурно-семантических особенностей этого текста. В рассказах Куприна того же жанра (см. $К_I$, $К_2$, $К_3$, $К_4$) эта частица имеет частоту употребления в 4 раза меньшую, чем в тексте А и в 3 раза меньшую относительной частоты в 9 рассказах Чехова. При этом, расхождение частот частицы НЕ в тексте А и в рассказах Чехова несущественно. Вот почему мы вправе утверждать, что отрицательные конструкции типичны для языка чеховских рассказов-характеристик. И действительно, в этих рассказах частица НЕ встречается много десятков раз (F = 196). Приведем несколько примеров. Ср.: 1. "Французы замечательны своим легкомыслием. Они читают нескромные романы, женятся без позволения родителей, не слушаются дворников, не уважают старших и даже не читают "Московских ведомостей". (Ш/113). 2. "Она не дышала, а задыхалась, не пила, а захлебывалась. (IУ/27). 3. "Когда она не спала и не ела, она играла, когда не играла, то пела". (IУ/28) и др.

Текст А тоже изобилует отрицательными конструкциями. Ср.: "Тесть: ... Не любит беспорядков и вечно читает мне нотации за нечистоплотность". 2. "Теща: ... Хвалится, что ее старичок не употребляет горячих напитков". 3. "Костька: ... Любит читать романы и выписывает газету Гатцука, которую никому не дает читать, боясь чтобы не запачкали". 4. "Жениться не хочет, потому что в женщине видит причину всех зол". 5. "Меня не любит за то, что я не позволяю ему ставить на мой стол бутылки с лекарствами" и др.

б) Относительная частота личного местоимения Я в тексте А (ℓ = 0,029) превышает общеязыковую в 2 раза (ℓ_1=0,014) Расхождение между этими частотами существенно. Этого и следовало ожидать, так как рассказ "Моя семья" написан от 1-го

лица. Если сопоставить абсолютную частоту местоимения Я в тексте А (F_1 = 19 при η_I= 662) с абсолютной частотой употребления ее в чеховских рассказах (F_2= 114 при η_2= 5377), написанных от I-го лица ($Ч_I$, $Ч_2$, $Ч_4$, $Ч_6$, и $Ч_8$), то оказывается, что расхождение между этими частотами несущественно. Это значит, что атрибутируемый текст и чеховские рассказы по использованию этой лексической единицы составляют одну совокупность. У Куприна мы не могли найти рассказы-характеристики, написанные от I-го лица, поэтому, в данном случае, мы лишены возможности контрольного сопоставления.

в) Концентрация местоимения МОЙ в тексте А очень высокая (ℓ = 0,0181). Она превышает общеязыковую в 11 раз (ℓ_I = = 0,0017). Расхождение между абсолютными частотами существенно. Но оно также существенно между абсолютными частотами в тексте А и в чеховских рассказах. Иными словами, по использованию местоимения МОЙ они не составляют одной совокупности. Это следствие того, что тема рассказа "Моя семья" не повторяется ни в одном из 9 чеховских рассказов. Между тем местоимение МОЙ в тексте А сочетается, в основном, со словами терминами родства. Ср.: брат моей жены, другой брат моей жены, третий брат моей жены, моя жена, моя мамаша, тетенька моя или женина, мои сыновья, моя дочь.

Здесь сталкиваемся с высокой частотой употребления местоимения МОЙ, в которой отражается тема рассказа "Моя семья".

г) В чеховских рассказах местоимение ОНА встречалось 129 раз (в текстах с общей длиной 7057 словоупотреблений). Почти все случаи употребления (кроме 5) этого местоимения относятся к лицам женского пола. Как в тексте А, так и в рассказах Чехова и Куприна это местоимение используется одинаковым образом. Расхождение частот местоимения ОНА в рассказах Чехова и Куприна несущественно, в чеховских рассказах и тексте А, в рассказах Куприна и тексте А тоже несущественно. Так как частота местоимения ОНА не отграничивает произведения Чехова от произведений Куприна, то можно считать эту лексическую единицу общеречевым признаком сопоставляемых одножанровых художественных текстов. Итак, из 4 выявленных служебных слов и местоимений только 2 (частица НЕ и местоимение Я) оказались релевантными при сопоставлении текста А с одножанровыми рассказами Чехова и Куприна.

О нейтральных словах в тексте А, чеховских рассказах и рассказах Куприна

Дальнейшее исследование семантических особенностей текста А в сопоставлении с рассказами Чехова и Куприна проведем по знаменательным словам, зафиксированным в зоне частотного списка рассказа "Моя семья" с частотой $F \geq 2$. Из совокупности знаменательных слов было выделено 9 слов (См.табл. 2), относящихся к нейтральному пласту лексики, частоты которых отличаются несущественно от частот этих слов в тексте А. Нейтральные единицы в нашем понимании - это слова, сконцентрированные в первой 1000 наиболее употребительных слов в ЧСРЯ - 77 (порог $F = 134$). Доля покрытия ими текста определяется по формуле: $Д = \frac{F^*}{N}$, где F^* - сумма частот выделенных слов в соответствующем тексте, N - длина текста в словоупотреблениях. В этой зоне из 75 слов 29 служебных слов и местоимений и 46 знаменательных слов. Присутствие знаменательных слов в этой зоне является следствием по крайней мере двух причин: 1) тематического фактора и 2) индивидуальноавторской наклонности употреблять каждое из данных слов в определенной речевой ситуации. Отграничить один фактор от другого не представляется возможным. Можно ослабить влияние тематического фактора, отобрав для сопоставления не один или два одножанровых рассказа, а несколько, и чем их будет больше, тем сильнее смогут проявить себя индивидуально-авторские особенности текста.

Мы уже отмечали, что было отобрано 9 рассказов-характеристик (других нет!), относящихся к раннему периоду творчества А.П. Чехова.

Доля покрытия текста А отобранными 9 словами в 3 раз превышает эту долю в рассказах Куприна и всего в 1,7 раза в рассказах Чехова. Расхождение между Д и $Д_1$ несущественно, а между Д и $Д_2$ существенно. Следовательно, текст А и чеховские рассказы составляют одну совокупность при использовании отобранных 9 слов. В то же время текст А и контрольные рассказы Куприна составляют по тем же признакам разные совокупности.

Эти данные указывают на некоторую тематическую близость чеховских рассказов и текста А. Мы чувствуем эту близость интуитивно.

Количественные данные о нейтральных словах в тексте А, рассказах А.П. Чехова и А.И. Куприна

Таблица 2

№№ п/п	Нейтральные слова	Частота в атрибутируемом тексте /А/ F	Частота в рассказах Чехова F_1	Частота в рассказах Куприна F_2	Расхождение между F и F_1 /по "хи-квадрату"/
I.	жена	4	I5	0	несуществ.
2.	любить	4	30	5	несуществ.
3.	деньги	3	I0	2	несуществ.
4.	спать	3	I2	0	несуществ.
5.	жениться	2	I3	0	несуществ.
6.	женщина	2	26	4	несуществ.
7.	иметь	2	I6	2	несуществ.
8.	писать	2	5	0	несуществ.
9.	умирать	2	I0	I	несуществ.
	Накопленная частота F^*	24	I37	I4	-
	Длина текста N	662	6585	2405	-
	Доля покрытия текста $\frac{F^*}{N}$	Д = 0,036 или 3,6%	$Д_1$= 0,02I или 2,I%	$Д_2$= 0,06 или 0,6%	несуществ.

Сопоставление фрагментов текста по содержанию

Текст А - это подписи зятя под I0 рисунками, изображавшими тестя, тещу, трех братьев жены, мать автора, тетку, двух сыновей и дочь. Эти подписи образуют I0 абзацев сплошного авторского текста. Они содержат 662 словоупотребления и являются характеристиками перечисленных персонажей.

В тексте А абзац I. а) "Тесть: Емельян Сидорович ... Мал ростом, худ, морщинист, но внушителен".[1] - Ср.: описание внешности губернского секретаря Ивана Саввича Кучкина: "Некрасив, ряб, гнусав, но весьма представителен".(IУ/I48). Наблюдается одна и та же синтаксическая конструкция, а по

[1] В этом и последующих примерах слова и словосочетания подчеркнуты нами.

содержанию - идентичное противопоставление отрицательных качеств персонажа его общему внешнему виду.

В этом абзаце читаем: б) "В день ангела получает визитную карточку от друга детства..." - Ср.: "Меланхолик. ... Ведет деятельную переписку с дяденьками, тетеньками, крестной мамашей и друзьями детства..." (I/83). Просторечные формы "тетенька" и "мамаша" характерны для языка ранних рассказов Чехова. (Они употребляются и в других местах текста А). Так, в т. I и 2 слово "мамаша" встречается 4I раз,а синонимы к нему ("маменька", "мамочка", "мама", "мать") - в единичных случаях.

в) "Не любит беспорядков и вечно читает мне нотации за нечистоплотность" - Ср.: "Она либила читать нотации".(I/I6I) "Битых два часа читал мне нотацию". (П/342).

г) "Искусства любит, но наук не признает". - Ср.:"Через неделю ко мне прибудет брат мой Иван (Маиор), человек хороший, но между нами сказать, Бурбон и наук не любит".(I/I5).

В тексте А абзац П. а) "Теща: Глафира Кузьминична ... Мяса не ест: обет дала". - Ср.: "Немного погодя она бросила курить и дала обет не есть мяса". (У/I64).

б) "Сваха, дает деньги на проценты..." - Ср.: "Живет отдачей денег под проценты". (IУ/I48).

в) "Хвастает, что ее старичок не употребляет горячих напитков". - Ср.: "...мой прапрадед... был погребен ... рядом с абатом католическим Иоакимом Шостаком, записки коего об умеренном климате и неумеренном употреблении горячих напитков хранятся еще доселе у брата моего Ивана (Маиора)".(I/I2). "Горячие напитки употребляете? да или нет?" (I/II6). "Угостила я их на славу, но, конечно, как и в те годы, горячих напитков - ни капли". (Ш/I72). "Картина масляными красками, изображающая "падение нравов" от употребления горячих налитков". (Ш/45I).

Когда речь идет об отношении к спиртному, Куприн поиному описывает это, совсем не употребляя словосочетания "горячие напитки". - Ср.: "Этот честный, но подверженный слабости к обильным возлияниям малый считал своим священным долгом напиваться каждый свободный вечер до состояния полного блаженства и горланить самые чувствительные песни".(I/II2). "Кучер Варфоломей был человек мрачный ... Пил ужасно, разговаривать ни с кем не любил ..." (I/278). "... пить водку вошло для меня в привычку именно на этой проклятой квартире..." (I/I2I).

Мы далеки от мысли считать, что только Чехову свойственно употребление словосочетания "горячие напитки". Он использовал те языковые средства, которые ему предоставлял русский литературный язык его времени. В данном случае, однако, мы можем утверждать, что это словосочетание более характерно для языка Чехова и нехарактерно для языка рассказов Куприна, Бунина и других писателей конца I9 века. Это словосочетание встречается в тексте А дважды.

г) "Я ли не даю ей денег на мазь Иванова, которой она по ночам натирает себе поясницу?" (А).

Мы встречаемся с упоминанием "мази Иванова" в трех рассказах Чехова: I. "Завещание старого, I883-го года" (I884). 2. "Жизнеописание достопримечательных современников" (I884) и 3. "Осколки московской жизни" (I884).

В середине 70-х годов А.И. Иванов приобрел известность в Москве как продавец целебных мазей. Эти мази были предметом карикатур и юмористических заметок. В первом рассказе автор завещает I884 году "Весь земной шар с его пятью частями света, океанами, ... мазью Иванова, Шестеркиным, обанкротившимися помещиками. ..." Во втором рассказе читаем: "Александр Иванович Иванов. Знаменитый изобретатель поседнокопытной, колесной и иных мазей". Рассказ заканчивается словами: "... новою мазью его я не только пользовался от прыщей, но также лечился ею от запоя и употреблял ее от клопов и прочих паразитов". (П/365-366).

В тексте А абзац Ш. а) "Брат моей жены: Иван Емельянович. Брандмейстер, изгнанный со службы за неумеренное употребление спиртных веществ". Ср.: с примерами в П/в

б) "Верит в спиритизм". Ср.: "Все трое не верили в спиритизм (П/276). "Я не верю в спиритизм..." (Ш/I39).

В тексте А абзац У: "Костька: третий брать моей жены".

а) "Жениться не хочет, потому что в женщине видит причины всех зол". - Ср.: "От нее (от женщины - Г.Е.) идет начало всех зол". (У/II4).

б) "Любит читать романы и выписывает газету Гатцука..." Фамилию Гатцука Чехов упоминает в рассказе "Масляничные правила дисциплины" (I885). Он пишет: "По мнению Гатцука, Суворина и других календаристов, она (Г.Ер.-масляница) начинается 28-го января..." (Ш/I62). Журналист А.Я. Гатцук издавал тогда "Крестный календарь".

В тексте А абзац УI. "Моя жена: Агаша. Маленькое, пришибленное, безгрудое ... существо". - Ср.: "маленькое, при-

шибленное, приплюснутое создание ..." (П/240). Состав определений почти тождественен, а порядок следования один и тот же.

В тексте А абзац УП. а) "Моя мамаша: Мавра Степановна..." О просторечии мамаша см. I-б.

б) "Горячие напитки отрицает:" О словосочетании горячие напитки см. П/в.

в) "День и ночь боится воров и то и дело ходит смотреть: заперта ли дверь?" Фразеосочетание "то и дело" типично для языка Чехова той поры (см. табл. I). Необходимо отметить также, что в повести "Драма на охоте" больной лесничий, отец Ольги, страдает той же манией, что и Мавра Степановна. Ср.: " - Митька, двери заперты? - услышали мы слабый тенор из соседней комнаты.

- Заперты-с, Николай Ефимыч! - прохрипел Митька и полетел опрометью в соседнюю комнату.

- То-то... Смотри, чтобы все заперты были ... - сказал тот же слабый голос. - На ключ, крепко-накрепко... Если воры будут лезть, то ты мне скажи ... Я их, мерзавцев, ружьем ... подлецов этаких ..." (Ш/269). (См. также Ш/271 и 273).

В тексте А абзац УШ. а) "Тетенька: (моя или женина;чья именно, не знаю)". См. о просторечии "тетенька" I/б.

В тексте А абзац IX. а) "Митя и Ваня: мои сыновья, близнецы... Усерднейшие потребители единиц и двоек. За поведение имеют тройку. Курят и на заборах пишут скверные слова". Ср. "Один злой мальчик имел дурную привычку писать на заборах неприличные слова". (П/281). Ср.: "Сангвиник... Грубит учителям, не стрижется... пачкает стены. Учится скверно, но курсы оканчивает" (I/80).

Сопоставленные фрагменты текстов свидетельствуют о том, что привычки и вкусы персонажей атрибутируемого текста схожи с привычками и вкусами некоторых героев чеховских рассказов раннего периода его творчества. Заметно, что характеристики персонажей оформляются, в общем, одними и теми же языковыми средствами.

Буквальные совпадения в тексте А и ранних произведениях А.П. Чехова

При сопоставлении текста А с текстами чеховских рассказов были обнаружены некоторые буквальные совпадения фрагментов анализируемых рассказов. Рассмотрим эти совпадения.

I) "Дрянь ты этакая!" (А) - "Ну ... тсс... Дрянь ты

этакая! Паршак! ("Раз в год", т. 2, с. 138).

2) "Родителей не почитает" (А) - "Родителей не почитает" ("Темпераменты", т. I, с. 80).

3) При перечислении автор текста А вместо числительного второй употребляет определительное местоимение другой в том же значении. "Брат моей жены: Иван Емельянович...","Пантелей: другой брат моей жены", "Костька: третий брат моей жены..." Ср.: а) "Случай первый", "Случай другой", "Случай третий" ("Исповедь, или Оля, Женя, Зоя," т. I, с. 133-136).

б) "Не успели застыть в воздухе волны от первого удара колокола, как послышался другой, за ним тот час же третий" ("Святою ночью", т. 5, с. 94).

в) "Первый настоящий страх, ..., имел своей причиной...", "Другой страх, пережитый мною ...", "Третий хороший страх мне пришлось испытать ..." ("Страх", т. 5, с. 186-190).

г) См. также "Мечты", т. 5, с. 395.

4) "Тесть: Емельян Сидорович ... В день ангела получает визитную карточку от друга детства, капитана 2-го ранга П.А. Дромадерова" (А), Ср.: а) "Однако! Что вижу! У вас цветут олеандры! - послышался внизу террасы женский голос,и через минуту на террасу входила княгиня Дромадерова, соседка по даче".

б) "Кнопка заговорил о быках и буйволах. Княгиня Дромадерова заявила, что все это скучно." в) "Дромадерова нашла, что все это очень скучно, и заговорила о сыне-поручике" ("Герой-барыня", т. 2, с. 150-152).

Рассказ "Герой-барыня" впервые появился в ж."Осколки", 1883, № 23, 4 июня, а текст "Моя семья" был напечатан в ж. "Зритель" 19 февраля того же года, т.е. тремя с половиной месяцами раньше.

Совпадение фамилий персонажей при атрибуции текстов имеет большое значение. Тем более, что фамилия "Дромадеров" отличается типичной для Чехова образностью и совершенно неожиданной этимологией, которая не могла бы прийти в голову двум разным лицам. Такое допущение совершенно невероятно. Мог ли бы Чехов заимствовать эту фамилию у автора А? Он не мог бы этого сделать (спустя 3,5 месяца!), так как его обвинили бы в плагиате. Чехов был тогда одним из самых активных сотрудников "Зрителя".

А.П. Чехов использует искаженную форму этой фамилии для обозначения воинской части в еще одном рассказе - "Делец", т. 4, с. 101: "Напомнить кстати о Пете Сивухине, желающем

поступить в Дримадерский полк". ("Осколки", 1885, № 33, 17 августа).

Итак, к этой фамилии Чехов вернулся снова спустя два с половиной года.

Кроме того, мы обнаружили у Чехова употребление слова "дромадер" (одногорбный верблюд) в прямом значении. Ср.: "Алеша поглядел вниз на землю и увидел штук десять котов. Вытянув хвосты, шипя и нежно ступая по травке, они дромадерами ходили вокруг хорошенькой кошечки, сидевшей на опрокинутой вверх дном лохани, и пели". ("Кот", т. 2, с. 132). Этот рассказ появился в тех же "Осколках" примерно за две недели до рассказа "Герой-барыня" и спустя два месяца после появления в печати рассказа "Моя семья".

По данным этимологического словаря Фасмера (Фасмер М., 1964, с. 538 и 541) интересующее нас слово имеет в русском языке три фонетических варианта "дремодар", "дромедар" и "дромадер". Последний вариант заимствован из французского языка. Именно от него и была образована фамилия капитана 2-го ранга П.А. Дромадерова.

Необходимо упомянуть об еще очень ценных сведениях, которые были выявлены А.П. Чудаковым из ранних произведений Чехова (Чудаков А.П., 1967, с. 166). Писатель иногда вводил в свои рассказы уже однажды использованные фамилии. А.П. Чудаков констатирует, что это Чехов делал в рассказах интересующего нас 1883 года. Ср.: фон Трамб ("Ушла", т. 2, с. 34) и барон Трамб ("Раз в год", т. 2, с. 135), Кулдаров ("Радость", т. 2, с. 12) и граф Кулдаров ("О том, как я в законный брак вступил", т. 2 с. 154).

Мы видим, что при вторичном использовании фамилий Чехов присоединял к ним разные титулы - "барон", "граф". В нашем случае наблюдается противопоставление аналогичного характера: П.А. Дромадеров - княгиня Дромадерова.

Итак, внешняя история текста А, анализ его содержания, языка и стиля, буквальные совпадения фрагментов атрибутируемого текста и рассказов Чехова, совпадение фамилий персонажей - все это приводит к одному единственному выводу: считать, что Б.Д. Челышев был прав, допустив возможную принадлежность текста "Моя семья" А.П. Чехову, и теперь признать, что рассказ этот действительно был написан великим писателем.

Л И Т Е Р А Т У Р А

Долотова Л.М., Соколова М.А. Вступительная статья к примечаниям в кн.: Чехов А.П. Полное собрание сочинений и писем в тридцати томах. Т. 3. - М.: Наука, 1975, с.540.

Ермоленко Г.В. Роль речевых параллелей при атрибуции художественных текстов. В кн.: Языковая норма и статистика. - М.: Наука, 1977, с. 98-99.

Копыленко М.М., Попова З.Д. Очерки по общей фразеологии. - Воронеж: Изд-во Ворон. ун-та, 1972, с. 32-33.

Куприн А.И. Собрание сочинений в девяти томах. Т. I. - М.: Худ. литература, 1970, с. 383, 394 и 417.

Фасмер М. Этимологический словарь русского языка. Т. I.- М.: Прогресс, 1964, с. 538 и 541.

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

Челышев Е.Д. Вступительная статья к тексту "Моя семья".В кн.: Литературное наследство. Т. 68. - М.: 1960, с.124.

Чехов А.П. Полное собрание сочинений и писем в тридцати томах. Т.т. I-4. - М.: Наука, 1974-1977. Издание продолжается.

Чудаков А.П. Неизвестные произведения раннего Чехова. - Вопросы литературы, 1967, № I, с. 166.

ON THE AUTHORSHIP OF THE TEXT "MY FAMILY"

Georgy Ermolenko

S u m m a r y

Adding to the already known ways of text attribution some new statistical methods of language and style analysis, the author proves that the text "My family" belongs to A.P. Chekhov. The new statistical methods are: 1) the method of distinguishing speech parallels; 2) the method of distinguishing the high frequency words typical for a certain text; 3) the method of comparing the parts of the text from the standpoint of the neutral words' usage. A number of literal coincidences in the text and the early writings by A. P. Chekhov, including the surnames' coincidence (Dromaderov), has been determined.

О ВОЗМОЖНОСТИ ФОРМАЛИЗАЦИИ ПУНКТУАЦИОННЫХ ПРАВИЛ

В.И. Критская

Лингвистическое обеспечение автоматизированной системы редакционно-издательских работ (АРИР) (Бондаренко В.В., 1982; Партыко З.В., 1981) включает подсистему пунктуационного контроля (ПК).

Знаки препинания в русских текстах регламентируются "Правилами русской орфографии и пунктуации" (1962, §§ 125-203), а также дополняющими их правилами, содержащимися в справочниках и пособиях (Розенталь Д.Э., 1978; Былинский К.И. Розенталь Д.Э. 1961; Былинский К.И., Никольский Н.Н., 1970). Для применения в автоматизированной системе правила должны быть описаны в удобной для ЭВМ форме. Возможно, что не все правила могут быть описаны достаточно формально. Задачей предлагаемой работы является анализ пунктуационных правил с целью определения возможности их формализации. Одновременно будет решаться и вторая задача - выявление требований к форме представления текста для работы подсистемы ПК АРИР.

Вся система правил пунктуации может быть разделена - в соответствии с Правилами (1962) - на три части: правила для авторской речи (точнее, общие правила), правила для прямой речи и правила сочетания знаков препинания*. Данное исследование касается только первой из названных частей. Формулировки этих правил в Правилах (1962) занимают §§ 125-194.

Рассмотрим общую структуру пунктуационного правила на примере правила 2 из § 153 раздела "Запятые при обстоятельственных оборотах". В связи с тем, что правила сгруппированы по признаку оформления обстоятельственных оборотов и выделена общая часть всех правил в шапку, мы восстановим полный текст правила, чтобы его смысл был понятен вне группы правил из § 153: "Запятыми выделяются существительные в косвенных падежах с предлогами и, реже, без предлогов, имеющие обстоятельственное (преимущественно причинное, условное

* Далее будем говорить о формулировках из Правил (1962) как основного собрания правил. В других пособиях формулировки строятся аналогичным образом.

и уступительное) значение, особенно если такие существительные имеют при себе пояснительные слова и стоят перед сказуемым".

Как видим, правило формулирует те условия, при которых ставятся парные запятые. Таким образом, процедура постановки парных запятых производится после проверки в тексте условий, предусмотренных правилом, а именно: I. Наличие существительного. 2. Существительное должно быть в форме косвенного падежа. 3. Существительное может иметь при себе предлог. Это условие необходимо проверять, так как при наличии предлога открывающая запятая ставится перед ним,а при отсутствии предлога - перед существительным. 4. Семантика существительного ограничена, но не строго, на что указывает слово "преимущественно". 5. Синтаксическая роль существительного - обстоятельство. 6. Существительное может иметь при себе пояснительные (т.е. зависимые) слова. 7. Существительное может стоять до и после сказуемого. Словом "особенно" подчеркивается тот факт, что условие 6 и одно значение условия 7 (позиция перед сказуемым) в комплексе делают постановку запятых более обязательной, чем при их невыполнении. В целом правило не имеет детерминистского характера, а указывает на возможность выделения таких обстоятельственных оборотов.

Кроме условий, правила включают утверждения о принятии решений по поводу постановки/непостановки знака препинания, его типа и месте в тексте.

Решение ставить/ не ставить пунктуационный знак записывается следующими способами: (знак) ставится; не ставится; не отделяется (запятой); выделяются (запятыми).

Разнообразно описание места постановки знаков препинания: конец предложения (§ I25); после (коротких предложений) - § I26; перед (союзами) - § I27; между (независимыми предложениями) - § I30; внутри (слова) - § I84.

Наибольшее разнообразие представляют описание условий.

I. Морфологические условия. К морфологическим относятся те контекстуальные признаки, которые обозначают либо грамматический класс слов, либо грамматическую категорию. Используются классы слов и категории: союзы (как их наличие, так и отсутствие - § I30); сочинительные союзы (§ I33); повторяющиеся союзы (§ I35); одиночные союзы (§ I39); сложные подчинительные союзы (§ I4I); подчинительный союз (§ I42); прилагательные (§ I43); существительное (§ I43); глаголы (§143);

причастия (§ I5I); личные местоимения (§ I5I);предлоги(§I5I); деепричастия (§ I53); формы глагола (§ I53); косвенные падежи существительного (§ I53); междометия (§ I57); частицы (§ I57); местоимения (§ I64).

Как видно из перечисленного, морфологическая информация занимает значительное место в условиях, формулируемых правилами, хотя основными принципами русской пунктуации считаются синтаксический, семантический и интонационный (Шапиро А.Б., I955). Вероятно, это результат практики редактирования, которая выработала "простые" признаки нахождения "опорных точек" в предложении, помогающих определять места пунктуационных знаков (наличие союзов) или находить конструктивные центры синтаксических единиц (причастные и деепричастные обороты). Необходимость использования морфологической информации предполагает автоматизацию морфологического анализа текста, что является в настоящее время реальной задачей.

2. Синтаксические условия. К синтаксическим относятся условия наличия в тексте определенных синтаксических единиц или отношений: законченное повествовательное предложение (§I25); независимые предложения (§ I30); сложное предложение (§I30); однородные члены предложения (§ I32); придаточные предложения, главное предложение (§ I33); второстепенный член предложения (§ I37); словосочетание (§ I43); именные части составных сказуемых (§ I50); пояснительные (зависимые слова) (§ I5I); определяемое (§ I5I); определяющее (§ I5I); определение (§ I5I); подлежащее (§ I5I); приложение (§I52);деепричастный оборот (§ I53); сказуемое (§ I53); вводные предложения (§ I55); вводные слова (§ I55); обращение (§ I56); вопросительные предложения (§ I80); восклицательные предложения (§ I82).

В настоящее время разрабатывается множество систем автоматического анализа, что делает реальной и задачу формализации синтаксических условий.

3. Логико-семантические условия. Сюда включены условия,опирающиеся на семантику единиц текста и логику, поскольку понятия логики положены в основу грамматики русского языка, используемой в пунктуационных правилах: связывать/не связывать в одно целое (§ I27); тесно связаны по смыслу (§ I38), в зависимости от смысла (§ I4I); тесно сливающиеся (§ I42); непосредственно относится (§ I43); образовывать единое смы-

словое целое (§ I42); с противоположными значениями (§I46); обозначения длительности действия (§ I49); употребляются не при сравнении (§ I50); тесно примыкают по смыслу (§ I5I); употребляемые в значении (§ I52); без изменения смысла (§ I52); по своему значению приближаются (§ I53); ограничивающие и уточняющие смысл (§ I54); определяется понятие (§ I64); пояснение и дополнение (§ I74); ослабить связь (§ I74); для указания (§ I77); для обозначения (§I77); высказывание (§ I85).

Автоматический семантический анализ текста пока не реализуем, поэтому и условия данного типа также не формализуемы.

4. Условия длины единиц текста. В пунктуационных правилах длина измеряется "синтаксически": короткие предложения(§ I26); достаточно развитые рубрики (§ I28); предложения значительно распространены (§ I30). Для выявления более точных оценок (измерения длины синтаксических структур в количестве слов или синтаксических связей, например) необходимо исследование текстов. От полученных результатов будет зависеть и определение возможности формализации данных условий.

5. Условия позиции единиц текста. Описания этих условий опираются на представление текста как линейной последовательности единиц: начинают собой самостоятельные предложения (§ I27); соединяют два предложения (§ I3I); между придаточными (§ I33); внутри имеются (§ I33); стоит после (§ I4I); начинающиеся союзом (§ I4I); стоит перед (§ I4I); сложный союз может распасться на две части: первая часть войдет в состав главного предложения, вторая будет выполнять роль союза (§ I4I); рядом стоящие союзы (§ I42);предшествующее прилагательное (§ I43); последующее словосочетание (§I43);следующие один за другим (§ I43); поставленные перед ним, но отделенные другими членами предложения (§ I5I); примыкают непосредственно к сказуемому (§ I53); предыдущих или следующих за ними (§ I54); которым заканчивается предложение (§ I59); предшествует (§ I59); находящимся в средине предложения (§ I60); вставляемые в предложение (§ I88).

Условия позиции связаны с морфологическими и синтаксическими условиями, и их формализация зависит от возможности формализации последних.

6. Условия присутствия других знаков препинания и зависимости от них. Проверка этих условий предполагает зависимость

знаков препинания между собой (в одной и той же позиции), а также зависимость постановки знака от наличия знаков препинания в единицах текста, входящих в большую единицу: внутри них уже есть какие-либо знаки препинания (§ I28);имеют внутри себя запятые (§ I30); если другими знаками препинания... это не может быть выражено (§ I73); когда выделение скобками... (§ I74); знак вопросительный в скобках (§ I8I); о тире при таких вставках см. § I74 (§ I83); которые без них (без союзов И, ДА) были бы разделены точкой (§ I3I).

Формализация этих условий возможна путем установления иерархии (последовательности применения) правил пунктуации. То есть необходимо выявить, какие правила в каждом конкретном случае или постоянно должны применяться первыми,какие - за ними, а какие - последними.

7. Условия трансформации единиц текста. К ним относятся немногочисленные условия: соединяют два предложения, которые без них (союзов) были бы соединены точкой (§ I3I);когда опущение придаточного предложения не требует перестройки главного предложения (§ I42); если перед ними можно без изменения смысла словосочетания вставить ТО ЕСТЬ, А ИМЕННО (§I52); относятся к отсутствующему в данном предложении,но подразумеваемому существительному или личному местоимению (§ I52).

Поскольку эта группа условий связана с логико-семантическими условиями, их формализация пока не возможна.

8. Стилистические условия. К этим условиям отнесены такие, которые опираются на экстралингвистическую информацию: точка ставится для придания изложению большей выразительности после коротких предложений, рисующих какую-нибудь единую картину или быструю смену событий (§ I26); точка ставится после предложения, вводящего в дальнейшее изложение, если последнее представляет собой развернутое повествование, описание или рассуждение (§ I29); придается большая самостоятельность, чем обычно (§ I3I).

Данная группа условий пока не формализуема, так как не разработаны способы формального выделения структурных частей произведения. Кроме того, невозможно в принципе предсказать авторские знаки препинания.

9. Условия наличия определенного слова. В этой группе собраны условия, в которых задаются либо множества слов, либо отдельные слова. Появление в тексте слова из заданного множества или отдельного слова и служит одним из условий при постановке знаков препинания: союзы И, А, НО, ОДНАКО и т.п. (§ 127); отрицание НЕ (§ 142).

При формализации правил эти слова должны задаваться списками.

10. Условия наличия сочетания слов. Сочетания слов также задаются либо множеством, либо отдельно: неразложимые выражения ВО ЧТО БЫ ТО НИ СТАЛО, ЧЕМ ПОПАЛО, КАК НИ В ЧЕМ НЕ БЫВАЛО, ЧЕРТ ЗНАЕТ ЧТО и т.п. (§ 141).

Такие сочетания слов также можно задавать списками.

Анализ формулировок правил пунктуации позволяет говорить о том, что каждое правило можно представить в виде независимого алгоритма, который, просматривая текст, проверяет выполнение определенных условий и вырабатывает решение о выборе и местоположении знака препинания. Рассмотрим общую схему алгоритма, построенного на основе приведенного выше правила из § 153.

Для работы этого алгоритма необходимо, чтобы в тексте была обозначена некоторая дополнительная информация. Каждому слову приписаны сведения о его принадлежности к части речи и грамматические характеристики, а также сведения о его формальных синтаксических связях и функции как члена предложения. Необходимо помнить и о соглашениях, принятых в системе правил пунктуации. Рассматриваемое правило действует только внутри предложения. Подразумевается также (как видно из иллюстративных примеров к правилу), что в случае расположения обстоятельственного оборота в абсолютном начале предложения оборот отделяется запятой, а не выделяется двумя запятыми – по общим правилам сочетания знаков препинания.

В анализируемом правиле можно выделить необходимые и достаточные условия (1, 2, 5) и дополнительные условия (см. выше). Дополнительные условия 3, 6 учитывают вариативность структуры обстоятельства в тексте, условие 4 – частотность обстоятельств с различной семантикой. О значении выполнения условия 7 было сказано ранее.

Предлагаем общий вид алгоритма правила постановки пунктуационных знаков при обстоятельственном обороте.

1. Если встречается существительное, переход ко 2; нет –

переход к другому правилу.

2. Если существительное в косвенном падеже, то переход к 3; нет - переход к другому правилу.
3. Если существительное - обстоятельство, то запоминается помета о возможности выделения запятыми и переход к 4; нет - переход к другому правилу.
4. Если существительное связано с предшествующим предлогом, то запоминается позиция перед предлогом и переход к 9; нет - переход к 6.
5. Если есть пояснительные слова после существительного, то запоминается позиция после последнего слова, запоминается помета о возможности запятых и переход к 9; нет - запоминается позиция после существительного и переход к 7.
6. Если есть пояснительные слова перед существительным, то запоминается позиция перед первым слева поясняющим словом, запоминается помета о возможности запятых и переход к 5; нет - запоминается позиция перед существительным и переход к 5.
7. Если обстоятельство - с указанной ограниченной семантикой, то запоминается помета о возможности запятых и переход к 8; нет - переход к 8.
8. Если обстоятельство стоит перед сказуемым, то запоминается помета о возможности запятых и переход к IO; нет - переход к IO; нет - переход к IO.
9. Если после предлога перед существительным есть пояснительные слова, то запоминается помета о возможности запятых и переход к 5; нет - переход к 5.
10. В соответствующие места текста ставятся запятые с особыми индексами: чем больше условий выполняется, тем вероятнее, что эти запятые останутся после окончательного редактирования текста.

АРИР не предполагает полного исключения человека-редактора из процесса редактирования ввиду принципиальной неформализуемости некоторых операций редактирования, что относится также и к подсистеме ПК: правка содержания и стиля текста - это творческая часть работы редактора (8). Однако большая доля ПК относится к техническому редактированию, что подтверждается и характером рассмотренных нами условий, которые нужно учитывать при постановке знаков препинания.

Возможность формализации пунктуационных правил, как было показано, в значительной степени зависит от наличия в тексте дополнительной информации. Эта информация может быть

введена в текст либо в результате автоматического анализа текста, либо редактором.

Таким образом, пунктуационное правило формализуемо, если формализуемы все условия, принятые в нем; частично формализуемо, если формализуемы некоторые условия; не формализуемо, если не формализуемы все условия. Целесообразность введения подсистемы ПК в АРИР должна проверяться на конкретных типах текстов: если в них преобладает употребление формализуемых и частично формализуемых правил, то подсистема ПК нужна; в противном случае введение такой подсистемы излишне.

ЛИТЕРАТУРА

Бондаренко В.В., Салко А.С. Задачи автоматизации редакционно-издательских работ. - В кн.: Переработка текста методами инженерной лингвистики. Тезисы докладов конференции, Минск, 1-2 февраля 1982 г. Минск, 1982.

Розенталь Д.Э. Справочник по правописанию и литературной правке. - М.: Книга, 1978.

Правила русской орфографии и пунктуации. - М.:Учпедгиз,1962.

Былинский К.И., Розенталь Д.Э. Трудные случаи пунктуации.М.: Искусство, 1961.

Былинский К.И., Никольский Н.Н. Справочник по орфографии и пунктуации для работников печати. - М.: МГУ, 1970.

Шапиро А.Б. Основы русской пунктуации. - М.: Изд. АН СССР, 1955.

Партыко З.В. О проблеме автоматизации процессов подготовки текстовой информации. - НТИ, сер. 1, 1981, № 3, с.8-13.

Феллер М.Д. Структура произведения. - М.: Книга, 1981.

ON THE FORMALIZATION POSSIBILITY OF FUNCTUATION RULES

Valentina Kritskaya

S u m m a r y

The article is devoted to the determination of conditions and possibilities of the Russian punctuation rules algorithmization. The rule formulas are analysed to find conditions and procedures. Classification of the conditions and text representation requirements are given. The results can be used in automatic editor systems.

СТАТИСТИЧЕСКИЙ АНАЛИЗ СЕМАНТИКИ ДЕЙСТВИЯ И СОСТОЯНИЯ В ПАТЕНТНО-ИНФОРМАЦИОННЫХ ПОТОКАХ

Г.Я.Мартыненко, Т.К.Чарская

I. Жанрово-стилевые аспекты двойственной природы патентной документации

Несмотря на то, что каждая десятая публикация в мировом потоке научно-технической информации относится к патентной литературе, функционально-стилистические черты этой разновидности письменной речи до недавнего времени находились на периферии лингвистического интереса, и лишь в самое последнее время появилось несколько работ, в которых рассматриваются наиболее общие особенности языка патентного документа (Хавкин, 1977; Могилевский, 1978; Мартыненко, 1979).

Жанрово-стилевое своеобразие патентной литературы определяется ее двойственной природой: патентный документ – это не только сообщение о новом техническом решении (информационно-технический аспект), но и средство фиксации объема прав изобретателя или патентовладельца (охранно-правовой аспект).

Выполнение патентным документом правовых функций обеспечивается совокупностью формальных требований, степень "деспотичности" которых зависит от того, на какую часть патентного документа они распространяются: на описание изобретения или на его формулу. Требования к описанию изобретения носят преимущественно номенклатурно-композиционный характер: указывается перечень обязательных тематических разделов и порядок их следования в тексте описания. В целом описание изобретения отличается от статьи с аналогичным техническим содержанием лишь более высокой степенью стандартизации изложения сущности технической идеи. О двойственной содержащейся в описании информации можно говорить лишь постольку, поскольку она неразрывно связана с формулой изобретения, в которой лаконично формулируется основная идея изобретения. Главная особенность формулы заключается в том, что в ней техническое содержание облекается в форму, имеющее правовое значение. Требования, предъявляемые к структуре формулы, мно-

гочисленны и каждое из них вносит свой вклад в формирование уникального облика того вида документа. Часть требований имеет универсальный характер: любая формула должна обладать такими свойствами, как лаконичность, широта (общность), полнота и определенность, а также отвечать требованию единства изобретения (Чихачев, 1965). Другие требования более конкретны: они указывают на то, каким образом достигаются перечисленные свойства. К таким требованиям относится, например, правило изложения каждого пункта формулы в виде одного предложения, имеющего стандартную структуру. Так, отечественная формула строится в виде дефиниции через ближайший род и видовое отличие, состоящей из двух частей: ограничительной части, в которой дается характеристика ранее достигнутого значения, и отличительной части, содержащей новую информацию. Ряд требований распространяется только на некоторые категории изобретений (устройства, способы, вещества). Примером такого требования является указание на необходимость описания устройства в статическом состоянии.

Для большинства нормативных требований характерно то, что они формируются в лингвистических терминах или могут быть истолкованы в таковых. Это не случайно, так как лексический состав текста формулы и ее семантико-синтаксическая организация по своей значимости соизмеримы с технической сущностью изобретения.

Это определяет исключительную важность изучения лингвистических последствий воздействия правовых аспектов на содержание технической задачи. Назовем некоторые проблемы, которые могут составить предмет лингвистического интереса:

I. Изучение закономерностей функционирования языка в экстремально-жестких условиях патентно-технологической коммуникации: компенсация бедности грамматики грамматикализацией семантических компонентов, деформация условно императивной языковой нормы категорически императивной юридической нормой, функционально-стилистические преобразования на разных уровнях языковой структуры и т.п.

2. Лингвистическое описание динамики и тенденций научно-технического прогресса, изучение структуры старого и нового технологического знания, а также лингвистических аспектов изобретательского творчества. Особый интерес представляет разработка "изобретательских словарей", которые могут выступать в роли центрального звена систем синтеза новых технических решений. (Методы поиска..., 1976), а в пер-

спективе – в роли лексикографического компонента автоматических моделей эвристических способностей человека (Караулов, 1982, 39).

3. Выявление перечня критериев, которыми руководствуются участники создания текста патентного документа при выборе способа воплощения технической идеи в реальные языковые формы.

Знание этих критериев позволит сформулировать рекомендации разной степени жесткости, на основании которых можно будет строить оптимальную в стилевом отношении формулу, не нарушая при этом требований, предъявляемых к ее структуре изобретательским правом. Важность разработки такого рода рекомендаций определяется тем, что "литературно-изобретательская" работа стала в настоящее время неотъемлемой частью деятельности каждого инженера и каждого ученого-естественника, а к качеству патентного документа предъявляются все более серьезные требования.

2. Объект и задачи исследования

Объект настоящего исследования – предикатные слова (в патентоведении их называют "связующими словами"), которые выступают в роли конструктивного центра словосочетаний, с помощью которых формулируются признаки изобретения на устройство.

Состав и форма связующих слов регулируются следующими нормативными указаниями: 1) фиксированным перечнем универсальных признаков, составляющих содержание устройства как конкретной категории изобретений (узлы и детали, из которых состоит устройство, пространственное расположение узлов и деталей, взаимосвязь между узлами и деталями, форма их конструктивного исполнения, материал, из которого изготовлены узлы и детали, соотношение их размеров); 2) необходимостью характеристики устройства в статическом состоянии; 3) необходимостью использования слов с предельно обобщенным значением.

Анализ "поведения" связующих слов разумно разбить на несколько этапов в порядке возрастания сложности подлежащих решению лингвистических задач. На первом этапе необходимо изучить морфолого-синтаксические характеристики связующих слов, на втором – их лексико-семантические характеристики, на третьем – семантико-синтаксические характеристики.

В данной работе отражены результаты первого этапа исследования.

Основной целью работы является выявление возможно более полного набора глагольных форм, обеспечивающих статальность характеристики конструктивного объекта в формуле, а также основных (допустимых и ошибочных) средств его характеристики действием.

Для достижения указанной цели были исследованы формулы конструктивных изобретений (за исключением формул на электрические схемы), опубликованные в 1982 г. в бюллетене "Открытия, изобретения, промышленные образцы, товарные знаки". При этом анализ проводился только на материале отличительной части, так как при составлении этого раздела формулы возникают наиболее серьезные редакционные трудности.

Кроме того, были исследованы некоторые закономерности количественного распределения глагольных слов со статальным значением и значением действия.

Основные итоговые данные (качественные и количественные) сведены в таблицы 1 и 2. К этим данным мы будем обращаться по ходу изложения результатов настоящей работы.

3. Правовое регулирование морфологических характеристик и синтаксических функций связующих слов

Прежде чем приступить к изложению результатов исследования, рассмотрим конкретные формулировки нормативных указаний, управляющих морфологическими и синтаксическими закономерностями употребления связующих слов.

3.1. Морфологические требования

Морфологические характеристики связующих слов регулируются жесткими правовыми требованиями, а также методическими указаниями, разъясняющими и уточняющими правовые положения.

Приведем несколько редакций этих требований:

1. В формуле изобретения не должно быть глаголов изъявительного наклонения (Патентоведение, 1977).

2. В формуле изобретения не должно быть глаголов изъявительного наклонения, выражающих незавершенное действие (Инструкция..., 1974).

3. Связи между элементами устройства характеризуются глаголами совершенного вида, прошедшего времени (Шепелев, 1967).

Цель этих требований – обеспечить выбор изобретателем или экспертом грамматической формы, соответствующей духу и букве патентного законодательства. Однако с лингвистической точки зрения эти требования сформулированы некорректно.

Если на практике составитель формулы будет слепо придерживаться первой рекомендации, то он должен будет взять на вооружение глаголы повелительного (!) и сослагательного (!) наклонений.

Вторая рекомендация призывает к аналогичным рискованным действиям, хотя этот призыв затемнен двусмысленностью формулировки благодаря введению дополнительного указания "выражающих незавершенное действие". В это указание вложен скорее причинный смысл (не должно быть глаголов изъявительного наклонения, ибо они выражают незавершенное действие), чем ограничительно-определительный (не должно быть таких глаголов изъявительного наклонения, которые выражают незавершенное действие). В пользу такого понимания говорит п.5.08 Инструкции по государственной научно-технической экспертизе изобретений (Инструкция..., 1974), в котором утверждается, что в отличие от конструктивных изобретений классического типа в формуле на электрические схемы допускаются глаголы изъявительного наклонения.

Из сказанного видно, что первые две формулировки терминологически некорректны. Однако это не влечет за собой никаких опасных последствий (никто, конечно, не составляет формулы в повелительном наклонении), так как некорректное использование лингвистических терминов нейтрализуется сильным семантическим требованием (устройство должно быть описано в статическом состоянии), а также примерами правильного и ошибочного использования глагольных форм: рекомендуется использовать слова типа установлен, снабжен, выполнен, но не устанавливается, устанавливают и т.п. Отметим, что как рекомендуемые формы, так и ошибочные имеют изъявительное наклонение.

В отличие от первых двух рекомендаций, привлекающих глагольную категорию наклонения, в третьей рекомендации используются категории вида и времени. Не возражая в принципе против этой рекомендации (она согласуется как с требованием статальности характеристики, так и с примерами правильного использования глагольных форм), отметим следующие ее недостатки. С одной стороны, под эту формулировку подпадают многие глагольные формы, которые на практике не используются

(чрезмерная общность формулировки в ущерб определенности), а с другой, - вне сферы ее действия оказывается значительный пласт глагольных слов, встречающихся в реальных формулах (недостаточная полнота формулировки).

3.2. Синтаксические требования

Выбор конкретной формы связующего слова в значительной степени зависит от выполняемой им синтаксической функции и очередности реализации в линейной цепи формулы.

Как было сказано выше, отечественная формула строится в виде формально-логического определения, состоящего из двух частей: ограничительной и отличительной. Первая часть играет роль логического субъекта (ближайший род), а вторая - роль логического предиката (видовое отличие).

С формально-грамматической точки зрения ограничительная часть представляет собой развернутое именное словосочетание, состоящее из названия изобретения, которое сопровождается цепочкой определений той или иной протяженности, вводимых с помощью предикатов включающий, содержащий, состоящий.

Что касается отличительной части, то она оформляется в виде предложения (или последовательности предложений), в котором в роли субъекта выступают слова-заместители, переобозначающие полностью или частично объем понятия, содержащийся в отличительной части, а в роли сказуемого - предикаты, с помощью которых достигается актуализация признаков, определяющих те или иные формы новизны конкретного объекта промышленной собственности. Часть признаков отличительной части, объединяясь в сочинительные и сочинительно-соподчинительные конструкции, образует предикативный каркас отличительной части.

Рассмотрим самые распространенные варианты формирования предикативного каркаса отличительной части.

I. Если усовершенствованию подвергается устройство в целом или один из его узлов, то отличительные признаки устройства или его узла вводятся посредством однородных сказуемых при общем субъекте (перечисленный ряд в этом случае замыкается союзом "и"):

Устройство для ..., отличающееся тем, что с целью ..., оно снабжено А, установлено в Б и выполнено в виде В;

Устройство для ..., содержащее А, отличающееся тем, что, с целью ..., А снабжен Б, установлен в В и выполнен в виде Г.

2. Если усовершенствованию подвергается два и более элемента, упомянутые в ограничительной части (одним из этих элементов может быть устройство в целом), то отличительные признаки этих элементов вводятся в виде последовательности предложений с собственными субъектами (перечислительный ряд в этом случае замыкается союзом "а");

Устройство для ..., содержащее А, Б, и В, отличающееся тем, что, с целью ..., А снабжен Г, Д установлен в Е, а выполнен в виде З.

Дополнительное расширение предикативного каркаса отличительной части осуществляется с помощью присоединительных конструкций, вводимых союзами причем, при этом, кроме того. Эти конструкции привлекаются преимущественно в двух случаях:

- если усовершенствованию подвергается элемент, упомянутый в ограничительной, но еще не упомянутый в отличительной части:

Устройство для ..., содержащее А, Б и В, отличающееся тем, что, с целью..., А снабжен Г, а Б установлен в Д, причем В взаимодействует с Г";

- если приводятся дополнительные сведения об элементе, уже упомянутым в отличительной части:

Устройство для ..., содержащее А и Б, отличающееся тем, что, с целью..., А снабжен В, а Б установлен в Г, причем В взаимодействует с Е.

Таким образом, отличительная часть достаточно большого объема распадается на несколько предикативных зон, последовательно присоединяемых друг к другу. Так, отличительная часть вида А снабжен Б и выполнен в виде В, а Г установлен в Д и взаимодействует с Г, причем Б несет Ж и контактирует с З состоит из трех зон: зоны признаков, в которые входит объект А, зоны признаков объекта Г и дополнительной зоны - зоны признаков объекта Б, уже упомянутого в отличительной части.

Каждая зона предикативного каркаса несет определенную смысловую нагрузку: наибольшую - зона, непосредственно примыкающая к слову отличающийся, наименьшую - зона, оформленная в виде присоединительной конструкции, которая имеет характер дополнительного сообщения.

Словесный объем каждой зоны предикативного каркаса наращивается с помощью определительных конструкций, которые уточняют признаки, определяющие форму новизны устройства, сообщают характеристике конструктивного объекта определен-

ность: А снабжен Б, содержащим В, а Г установлен в Д, связанным с Е, взаимодействующим с Ж, причем Б контактирует с З, примыкающим к Д".

4. Статистика связующих слов, имеющих значение состояния

Соответствие отличительной части нормативным указаниям реально проявляется в том, что в ней резко преобладают глаголы со статальным значением: около 87 % от употреблений всех глагольных форм. При том среди статальных слов господствующие позиции (79,4%) занимают глаголы совершенного вида в форме страдательных причастий прошедшего времени (см.табл. I). В функции сказуемого используются краткие формы этих причастий (установлен, снабжен, выполнен), а в функции определения – полные формы (установленный, снабженный, выполненный). Когда говорят о необходимости описания конструкции в статическом состоянии, имеют в виду именно эту, наиболее частотную форму. В лингвистической литературе ее обычно называют "статальным пассивом" или "пассивным состоянием". Преимущество этой формы заключается в том, что она выражает отнесенный к настоящему результат действия, совершенного в прошлом, т.е. с помощью слов этого типа достигается не только статальность характеристики конструктивного объекта, но подчеркивается, что этот объект является продуктом творческой деятельности, в данном случае изобретательской.

Однако кроме указанной, наиболее "популярной" глагольной формы в отличительной части достаточно широко (20,6 %) используются статальные глаголы несовершенного вида.

В функции сказуемого эти глаголы используются в форме 3-го лица настоящего времени (содержит), а в функции определения - в форме действительных причастий настоящего времени (содержащий).

Статальные глаголы несовершенного вида, как и глаголы типа установлен, привлекаются для формирования признаков, характеризующих устройства и составляющих его содержание:

I) узлы, механизмы, детали, образующие устройство: состоит из, включает, содержит, имеет и др.

2) пространственное расположение и взаимосвязь узлов, механизмов, деталей: несет, опирается, прилегает, примыкает, лежит, проходит, выступает из, контактирует, соприкасается, стыкуется, сопрягается, упирается, охватывает, объемлет и т.п.;

3) форма выполнения узлов, механизмов, деталей: представляет собой, имеет (вид, форму), является и т.п.;

4) соотношение размеров у отдельных узлов и деталей: диаметр узла "Л" превосходит, превышает диаметр узла "В", соотношение размеров составляет и т.п.

Следует отметить, что статальные глаголы этого типа - одновидовые глаголы совершенного вида и, следовательно, страдательные причастия могут быть образованы лишь от нескольких глаголов: сопрягается ⟶ сопряжен, сообщается ⟶ сообщен, опирается ⟶ оперт.

Статальные глаголы несовершенного вида обозначают состояние, не изменяющееся на любом отрезке своего протекания, но в отличие от глаголов типа "установлен" они не имеют результативного значения. Этим, по-видимому, объясняется предпочтение, отдаваемое глаголам совершенного вида в форме страдательных причастий.

Сознавая недостаточность семантики статальных глаголов несовершенного вида, составители формулы стремятся перевести их в план результативности. Эта операция осуществляется в основном двумя способами:

- путем сочетания глаголов несовершенного вида в форме действительных причастий со словом выполнен:

Экран выполнен состоящим из соединенных между собой полос;

Корпус выполнен объемлющим каретку;

- и (значительно реже) путем сочетания страдательных причастий с придаточными предложениями с союзами так что, таким образом что, в которых функцию сказуемого выполняет глагол несовершенного вида:

Траверсный путь для распределения заполненных контейнеров размещен так, что он пересекает крайние конвейеры-накопители для заполненных конвейеров и примыкает к контейнерам-накопителям для пустых контейнеров.

Как было указано выше, статальные глаголы совершенного вида реализуются примерно в четыре раза чаще, чем глаголы несовершенного вида. Однако это соотношение существенно нарушается при учете синтаксической функции, выполняемой этими глаголами (см. табл. I).

Если в предикативном каркасе среди статальных глаголов резко преобладают глаголы совершенного вида (86,9%), то в зависимой, атрибутивной позиции (причастные обороты и определительные придаточные предложения) это преобладание становится менее существенным (73,7%).

5. СТАТИСТИКА СВЯЗУЮЩИХ СЛОВ СО ЗНАЧЕНИЕМ ДЕЙСТВИЯ

Глагольные слова со значением действия в отличительной части используются достаточно часто (около 10% от употреблений глагольных слов в целом).

Все без исключения глаголы со значением действия - глаголы несовершенного вида. В функции сказуемого они, как и статальные глаголы второй группы, реализуются в форме 3-го лица настоящего времени (перемещается), а в функции определения - в форме действительных (перемещающий) или страдательных (перемещаемый) причастий настоящего времени.

Глаголы действия в зависимости от характера обозначаемого ими действия распадаются на две основные группы.

Первую группу образуют глаголы, указывающие на процесс создания конструкции или производимые ею операции. Так, вместо страдательных причастий прошедшего времени (установлен, снабжен, закреплен и т.п.) иногда используются глаголы несовершенного вида в форме 3-го лица настоящего времени:

в бункере на уровне материала устанавливается заземленная электропроводящая решетка; количество штырей устанавливают не менее четырех.

Хотя подобное использование связующих слов в формуле является грубым нарушением нормативных указаний, органы Государственной патентной экспертизы иногда проявляют беспечность, пропуская "незаконные" редакции.

Вторую группу слов со значением действия образуют слова, указывающие на кинематическую или динамическую связь узлов и деталей конструкции. Чаще всего эти глаголы указывают на двустороннее (взаимодействует) или одностороннее (воздействует) действие узлов и деталей устройства друг на друга, а также на вид движения, сообщаемый одним звеном кинематической цепи другому ("А вращает В", "В" вращается с помощью "В", А сообщает вращательное движение "В", "А" получает, заимствует вращательное движение от "В" и т.п.).

Если использование слов типа устанавливается, устанавливают в предикативном каркасе отличительной части является бесспорным нарушением нормативных указаний, то использование слов типа взаимодействует, перемещается, вращается очевидным нарушением не является.

Случаи употребления глаголов действия в предикативном

каркасе отличительной части, особенно в непосредственном контакте со словом отличающийся, встречаются крайне редко.

Так, среди 1000 формул, которые были привлечены для количественного анализа, встретилось лишь 10 таких случаев (см. табл. 1).

Глаголы действия концентрируются преимущественно в атрибутивной позиции, участвуя в формулировании признаков, развивающихся и уточняющих признаки, определяющие форму новизны устройства. В функции определения эти глаголы встречаются почти в каждой второй формуле.

Сознавая неопределенность характеристики, порождаемую использованием глагольных слов в предикативном каркасе формулы, составители формулы прибегают к комбинированным сочетаниям, с помощью которых характеристике устройства действием придается конструктивная окраска. "Иллюзия" конструктивности создается привлечением языковых средств, обеспечивающих перевод действия из актуального плана в план потенциальности или статальной результативности.

Для перевода действия из актуального плана в потенциальный используются слова или сочетания слов, обозначающих возможность: может, имеет возможность: "тоководы могут перемещаться один относительно другого", "одна из полос может зацепляться с приводным элементом" и т.п.

Более сильным средством устранения признака со значением действия является комбинированный прием: перевод действия из актуального плана в потенциальный в сочетании с переводом действия в план статальной результативности.

Рассмотрим основные случаи такого комбинированного перевода.

I. Наиболее употребительны словосочетания вида:

выполнен установлен закреплен	с возможностью	взаимодействия перемещения вращения

В комбинированных сочетаниях этого типа на значение действия наслаивается значение результативного состояния, создаваемое глаголами совершенного вида в форме страдательных причастий, и значение потенциальности, вносимое словом "возможность". При этом краткие формы страдательных причастий придают характеристике устройства действием некоторый оттенок конструктивности: например, слово выполнен в самом общем виде указывает на то, что движение, совершаемое каким-либо узлом устройства, является следствием формы его кон-

структивного выполнения, а слово установлен указывает на то, что движение узла является следствием места его расположения.

Иногда составители формулы при использовании этого приема проявляют "сверхстарание", привлекая для перевода действия в план результативного состояния даже два страдательных причастия. В качестве отрицательного примера можно привести такую фразу: "приспособление выполнено смонтированным с возможностью поворота вокруг своей оси".

2. Иногда для перевода действия в план потенциальности и статальной результативности используются словосочетания вида:

выполнен { перемещаемым / взаимодействующим / вращающимся }

Пример: "экран выполнен перемещаемым вдоль концов стержней".

В комбинированных сочетаниях этого типа значение результативного состояния, как и в сочетаниях, описанных выше, создается краткой формой страдательного причастия выполнен, а значение потенциальности заключено в причастиях настоящего времени, которые в таком употреблении близки по значению к прилагательным типа подвижный, поворотный, в которых действие представлено как свойство.

Использование комбинированных сочетаний является мощным средством "протаскивания" слов со значением действия в формулу (см. табл. I). В особенности это относится к предикативному каркасу отличительной части, в котором среди глаголов действия в 89% случаев реализуются комбинированные сочетания.

6. Зависимость частоты употребления связующих слов от очередности их реализации в отличительной части

Как было показано выше, интенсивность использования слов со значением состояния и действия находится в сильной зависимости от выполняемой ими синтаксической функции. Наблюдения показали, что частота использования этих слов в существенной мере зависит также и от очередности их реализации в предикативном каркасе отличительной части.

Из таблицы 2 и рис. I видно, что по мере удаления от слова отличающийся описание устройства в отличительной час-

ти становится все менее строгим. Это проявляется в том, что по мере приближения к "периферии" каждого пункта формулы постепенно падает результативность статальной характеристики конструктивного объекта и растет для признаков, сформулированных с участием слов со значением действия.

Если в позиции, находящейся на минимальном расстоянии от слова отличающийся (позиция I), 90,6 признаков формулируется при помощи слов со значением результативного состояния (установлен, снабжен, выполнен и т.д.), то уже в третьей позиции результативность статального описания падает до 32,2%, при этом доля статальных слов без результативного значения (содержит, несет, представляет собой и т.п.)достигает 25,5%, а доля слов и комбинированных сочетаний со значением действия (перемещается, установлен с возможностью перемещения и т.п.) достигает II,3%.

Интересен тот факт, что снижение доли результативности статального описания и увеличение доли признаков со значением действия происходит не беспредельно. В достаточно далекой позиции от слова отличающийся (см. рисунок) интенсивность использования слов со значением состояния и действия стабилизируется. Асимптотический уровень практически достигается в позиции 5. В этой позиции число глаголов совершенного и несовершенного вида будет примерно одинаковым,а среди последних глаголы со значением нерезультативного состояния будут составлять около 38%, а со значением действия - около I2%.

7. Заключение

Результаты данной работы говорят о том, что коллективный обычай употребления связующих слов (функционально-стилистический узус) в значительной степени отклоняется от жестких требований патентно-правовой нормы. Причина этого несоответствия видится в том, что функционально-процессуальное описание технической идеи в некоторых случаях является более эффективным средством защиты прав изобретения и патентовладельца. В пользу такого мнения говорит система патентования США, которая при описании устройств допускает активное использование функционально-процессуальных признаков. Это означает, что лингвистическая технология изобретательской деятельности может рассматриваться в более широком контексте - контексте лингвоправовой типологии.

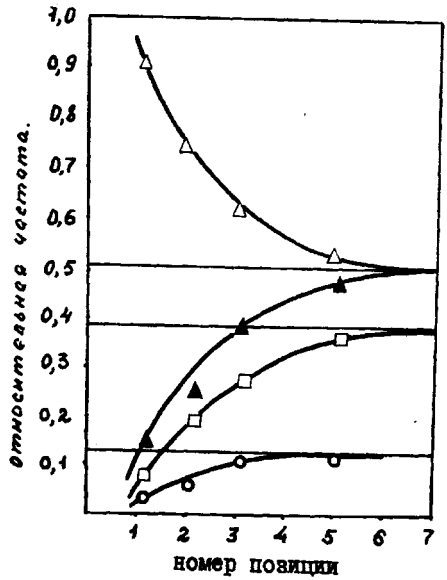


Рис. I. Зависимость частоты употребления связующих слов со значением действия и состояния от очередности их реализации в предикативном каркасе отличительной части.

△ - глаголы совершенного вида;
▲ - глаголы несовершенного вида;
▫ - глаголы несовершенного вида со значением состояния;
○ - глаголы несовершенного вида со значением действия.

Таблица I

Типы связующих слов в отличительной части формулы на устройство и частота их употребления

(1000 формул, 4158 связующих слов)

Глаголы состояния					Глаголы действия				
3830					536				
Глаголы совершенного вида		Глаголы несовершенного вида			Все несовершенного вида				
2673		747							
В предикативной позиции	В атрибутивной позиции	В предикативной позиции		В атрибутивной позиции	В предикативной позиции			В атрибутивной позиции	
1416	1467	214		533	90			446	
		В одиночном употреблении	В комбинированном сочетании		В одиночном употреблении	В комбинированных сочетаниях		В одиночном употреблении	В комбинированных сочетаниях
установлен	установленный	состоит	выполнен состоящим	состоящий	перемещает	может перемещаться	выполнен перемещаемым, установлен в возможн. перемещения	перемещаемый, перемещающий	установленный с возможностью перемещения
1416	1467	212	2	533	10	8	72	402	44
Результативное состояние	Результативное состояние	Состояние	Состояние, переведенное в план результативности	Состояние	Действие	действие, переведенное в план потенциальности	действие, переведенное в план результативности	Действие, переведенное в план потенциальности	Действие, переведенное в план результативности

Таблица 2

Распределение связующих слов в отличительной части формулы на устройство в зависимости от очередности их реализации в предикативном каркасе формулы

Очередность употребления после слова "отличающийся"	Глаголы состояния сов. вида		Глаголы состояния несов. вида		Глаголы действия несов. вида	
	Количество употреблений	Частость	Количество употреблений	Частость	Количество употреблений	Частость
I	883	0,906	58	0,059	34	0,035
2	428	0,754	I04	0,I83	36	0,063
3	77	0,622	33	0,265	I4	0,II3
4-6	27	0,529	I8	0,353	6	0,II8
7 и более	I		I			
Сумма	I4I6		2I4		90	

Л И Т Е Р А Т У Р А

Инструкция по государственной научно-технической экспертизе изобретений. - Вопросы изобретательства, 1974, № 8, с. 29-35.

Караулов Ю.Н. Анализ метаязыка словаря с помощью ЭВМ. М., 1982.

Мартыненко Г.Я. Некоторые лингвистические аспекты двойственной природы патентной документации. - Научно-техническая информация. Серия 2. 1979, № 9, с. 10-16.

Методы поиска новых технических решений. /Под ред. А.И. Половинкина. Йошкар-Ола, 1976.

Могилевский В.М. Лингвистические особенности формулы изобретения на устройство. - Вопросы изобретательства, 1978, № 12, с. 25-32.

Патентоведение. /Под ред. В.А. Рясенцева. М., 1976.

Хавкин А.М. Некоторые актуальные вопросы языка патентной документации. - Вопросы изобретательства, 1977, № 8.

Лихачев Н.А. Формула изобретения. М., 1965.

Шепелев Н.П. Составление заявок на изобретение (практическое пособие). М., 1967.

STATISTICAL ANALYSIS OF ACTION AND STATE SEMANTICS IN PATENT INFORMATION FLOWS

Grigory Martynenko and Tamara Charskaya

S u m m a r y

The article deals with the regularities of language functioning under extremely strict conditions of patent technological communication. Predicates are the subject of the present investigation as they are the constructive centre of word-groups used for formulation of inventive features in patent claims. The statistical data presented in this paper show the correlation of verb forms and combined word-groups making it possible to describe the subject of invention in statics and dynamics. A special stress is also laid on the study of lingual mechanism of inventive step and lingual technology of patent and license activity.

ОПТИМАЛЬНОЕ СВЕРТЫВАНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА В ЗАДАЧАХ СТИЛИСТИЧЕСКОЙ ДИАГНОСТИКИ

М.А. Марусенко

Применение математических методов к решению различных прикладных задач как экономического, так и социального или лингвистического характера позволяет сформулировать некоторые методологические проблемы, свойственные, впрочем, всем молодым научным дисциплинам. Основная методологическая трудность заключается в избыточном разнообразии формулировок сходных по существу реальных задач (Айвазян, 1975). Анализ многообразия таких задач показывает, что многие проблемы, связанные с переработкой информации, могут формулироваться и решаться как задачи распознавания образов. В общем виде эта задача может быть описана следующим образом: пусть имеется множество M некоторых объектов, разделенное на n непересекающихся подмножеств, называемых классами или образами. Каждому объекту соответствует определенное описание x, которое можно рассматривать как многомерный вектор. Требуется построить оптимальный в некотором смысле алгоритм, который по данному описанию объекта указывал бы класс, к которому он принадлежит (Ковалевский, 1965). Как видно, в терминах распознавания образов могут формулироваться задачи весьма широкого круга. Помимо разработки вводных устройств различных типов для ЭВМ и систем обработки технико-экономической информации, к этому типу задач относится и целый класс лингвистических проблем, в первую очередь связанных с конструированием читающих автоматов, воспринимающих печатный или рукописный текст, а также устройств, воспринимающих устную речь. Можно показать, что к этому же классу задач относятся и статистические исследования функциональных и авторских стилей, а также атрибуция анонимных или псевдоанонимных текстов на основе лингвистических характеристик, уточнение или определение датировки и т.д. Все эти задачи объединяются в класс, удачно обозначенный термином "стилистическая диагностика". Как следует из этого названия, объектом анализа в подобных задачах являются индивидуальные или функциональные стили. Попытаемся дать операциональное определение этому широко употребительному понятию, позволяющее, в дальнейшем, перейти к решению задач стилистической диагностики с точки зрения распознавания образов. Для этого необходимо выявить основные концептуальные и методологические связи этого подхода с некоторыми смежными методами. Прежде всего необходимо отметить тесную связь с методами группировки и класси-

фикации многомерных совокупностей, кластер-анализом и информационной алгеброй. В рамках развиваемого здесь подхода термины "распознавание образов", "многомерная группировка и классификация" могут рассматриваться как синонимичные. В теории классификации понятие объекта определяется через пучок свойств, необходимых для отображения предметной области. Именно значения или кортежи значений свойств обозначают объект, становятся системным, формализованным именем объекта (Грейсух, 1982). Можно отметить, что используемые в стилистике определения стиля некоторым образом коррелируют с вышеуказанным определением. Так, например, А.И. Ефимов определяет стиль как исторически сложившуюся разновидность языка, отличающуюся специфическим составом, характером объединения и закономерностями употребления речевых средств (Ефимов, 1969). Таким образом, в терминах теории распознавания образов, стиль можно определить как набор свойств (параметров), характеризующих состав, способы объединения и закономерности употребления речевых средств, образующих данную разновидность языка.

Существенной особенностью теории распознавания образов является разграничение самого объекта ρ и математического объекта χ, который может быть описан заданием определенных значений конечного набора параметров. Каждый реальный объект становится математическим объектом только тогда, когда определен такой набор процедур измерений $\{M_1, \ldots, M_n\}$, что применение $\{M_1, \ldots, M_n\}$ к объекту ρ порождает набор n констант $(x_1, \ldots, x_n)$, представляющих значения признаков рассматриваемого объекта. Таким образом, полученный набор величин описывает математический объект χ, являющийся отображением реального объекта ρ (Заде, 1980). Поэтому, прежде чем приступить к процедуре собственно распознавания образов, необходимо осуществить преобразование реального объекта (стиль в сформулированном выше определении) в математический объект, являющийся отображением реального объекта в информационном пространстве, т.е. в набор значений признаков изучаемого объекта.

Если информационное пространство построено на признаках, то в нем может быть выделено пространство меньшей размерности, состоящее из необходимого и достаточного числа признаков для отнесения объекта к классу. В общем виде, каждый математический объект ρ может характеризоваться n-мерным вектором (n - число признаков). Следовательно, задача оптимального свертывания признакового пространства, являющаяся предметом данной статьи, может быть сформу-

лирована как задача нахождения m признаков ($m < n$), преобразующих реальный объект p в математический объект $M(p)=\{M_1(p),\ldots,M_m(p)\}$. При этом желательно, чтобы $M(p)$ определялся малым числом признаков и чтобы измерение этих признаков было относительно простым (Заде, указ. соч.). Снижение размерности исходного признакового пространства целесообразно и важно с нескольких точек зрения: во-первых, если объем исследуемой совокупности ненамного превосходит размерность признакового пространства, то среди признаков, по которым проводится распознавание, могут быть "шумовые", удаление которых лишь улучшит характеристики точности распознающего алгоритма, во-вторых, у исследователя появляется возможность значительно упростить математическую модель объекта, что, в свою очередь, позволяет лучше понять механизм изучаемого явления и дать содержательную интерпретацию полученным результатам, в-третьих, снижение размерности позволяет существенно снизить вычислительные трудности и затраты, неизбежно возникающие при машинной обработке значительных массивов информации большой размерности (Айвазян и др., 1974). Кроме того, существенным преимуществом этой операции является возможность уменьшить объем данных на этапе подготовки к вводу в ЭВМ и, тем самым, снизить количество ошибок, очень трудно поддающихся обнаружению и исправлению.

Прежде чем перейти к описанию эксперимента, необходимо несколько подробней остановиться на формулировке задачи оптимального свертывания признакового пространства. Число признаков, используемых в алгоритме распознавания, зависит от конкретных условий реализации каждого эксперимента. Известны эксперименты по определению авторства на основании одного параметра (Kjetsaa, 1976; Yule , 1939). Применив более развитые методы анализа, И.П.Севбо использовала в своих экспериментах 7 диагностических признаков, не объединяя их, однако, в рамках единого алгоритма (Севбо, 1981). В общем виде, минимальное необходимое и достаточное число признаков определяется требуемыми характеристиками точности распознавания и может быть установлено экспериментально. Поэтому задача оптимального свертывания признакового пространства может быть сведена к получению набора признаков, ранжированных по убыванию/возрастанию их индивидуальной информативности, под которой понимается способность не снижать достоверность классификации при переходе от исходного пространства к пространству сокращенной размерности. Из этого ранжированного ряда признаков исследователь, исходя из требований точности, возможностей и условий эксперимен-

та, может формировать математический объект, образованный n признаками ($n \geqslant 1$), такой, что информативность этих признаков будет превышать информативность любых других n признаков, взятых из того же исходного пространства.

При установлении исходного пространства признаков в описываемом ниже эксперименте учитывалось, что стиль - это прежде всего категория структурно-синтаксическая (Ефимов, указ.соч.) и что основной единицей синтаксиса является предложение. Поэтому признаки, составившие исходное пространство, взяты из обширного круга работ советских и зарубежных авторов, исследовавших структуру и состав предложения квантитативными методами. Учитывая требования простоты определения значения признака, в исходное пространство не включены признаки, определение значений которых требует промежуточных построений (деревьев зависимостей и т.п.). Анализ признаков показал, что их можно разделить на 2 группы: а) первичные признаки, значения которых определяются непосредственно на предложении и б) производные признаки, представляющие собой отношение двух или более первичных признаков (все вопросы, связанные с обоснованием, выделением и определением значений признаков, достаточно полно и подробно рассматриваются в работах В.Г.Адмони, И.А.Короленко, Г.А.Лесскиса, Л.В.Малаховского и др.). После унификации и приведения наименований было получено 56 признаков, отражающих структуру и состав предложения в различных индоевропейских языках (нумерация признаков произвольная): X01 - число слов в цельном предложении (далее ЦП); X02 - число графем в ЦП; X03 - число слов в простом самостоятельном предложении; X04 - число элементарных предложения (далее ЭП) в ЦП; X05 - число главных предложений; X06 - число сочиненных предложений; X07 - число сочиненных предложений без спрягаемой формы глагола; X08 - число подчиненных предложений; X09 - число подчиненных предложений I-й степени; X10 - число подчиненных предложений 2-й степени; X11 - число подчиненных предложений 3-й степени; X12 - число подчиненных предложений 4-й и высших степеней; X13 - число ЭП без номинативного подлежащего; X14 - число подчиненных предложений без спрягаемой формы глагола; X15 - число вставных предложений; X16 - число охватывающих предложений; X17 - число слов I-й группы (знаменательных); X18 - число слов 2-й группы (служебных); X19 - число существительных; X20 - число прилагательных; X21 - число местоимений; X22 - число спрягаемых форм гла-

гола; X23-число именных форм глагола; X24 - число наречий; X25 - число предлогов; X26 - число союзов; X27 - число подчинительных союзов; X28 - число сочинительных союзов; X29 - число предикативов; X30 - число слов в аккузативе; X31 - число слов в дативе; X32 - число подлежащих; X33 - число местоимений-подлежащих; X34 - число групп однородных членов; X35 - число членов однородных групп; X36 - число однородных сказуемых; X37 - число однородных групп дополнений; X38 - число причастных оборотов; X39 - число членов причастных оборотов; X40 - число распространенных причастных определений; X41 - число членов распространенных причастных определений; X42 - число согласованных определений; X43 - число причастий-согласованных определений; X44 - число несогласованных определений; X45 - число существительных-несогласованных определений; X46 - число обособленных членов; X47 - число членов в группах обособленных членов; X48 - число абсолютных оборотов; X49 - число членов абсолютных оборотов; X50 - число инфинитивных оборотов; X51 - число членов инфинитивных оборотов; X52 - число существительных без группы; X53 - число групп существительных (далее ГС); X54 - число членов ГС; X55 - число знаменательных слов в ГС; X56 - число служебных слов в ГС.

Эти первичные признаки входят в состав также 56 производных признаков (совпадение случайное). Способ вычисления производного признака указывается в скобках:
П01 - средний размер ЭП (X01/X04); П02 - средняя длина слова (X02/X01); П03 - средняя длина ЭП (X02/X04); П04 - коэффициент сложноподчиненности (X08/X04); П05 - коэффициент комплексности (X08/X05); П06 - коэффициент подчинения I-й степени (X09/X08); П07 - ранг ЦП (X05+X06+2xX09+3xX10+4xX11+5xX12); П08 - совмещенный коэффициент комплексности $[(X09+2xX10+3xX11+4xX12)/X05]$; П09 - коэффициент аффинитности (X14/X08); П10 - соотношение между средним размером ЭП и сложностью ЦП (П01/X04); П11 - доля слов I-й группы (X17/X01); П12 - доля слов 2-й группы (X18/X01); П13 - доля существительных (X19/X01); П14 - доля прилагательных (X20/X01); П15 - доля местоимений (X21/X01); П16 - доля спрягаемых форм глагола (X22/X01); П17 - доля именных форм глагола (X23/X01); П18 - доля наречий (X24/X01); П19 - доля предлогов (X25/X01); П20 - доля союзов (X26/X01); П21 - доля подчинительных союзов (X27/X01); П22 - доля сочинительных союзов (X28/X01); П23 - доля предикативов (X29/X01); П24 - доля аккузатива (X30/X01); П25 - доля датива (X31/X01); П26 - соотношение между числом су-

ществительных и спрягаемых форм глагола (X19/X22); П27 - коэффициент знаменательности (X17/X18); П28 - доля подлежащих (X32/X01); П29 - доля местоимений-подлежащих (X33/X01); П30 - доля однородных членов (X34/X01); П31 - доля однородных сказуемых (X36/X01); П32 - доля однородных групп дополнений (X37/X01); П33 - доля членов причастных оборотов (X39/X01); П34 - доля обособленных членов (X46/X01); П35 - доля членов инфинитивных оборотов (X51/X01); П36 - доля членов ГС (X54/X01); П37 - насыщенность ЭП однородными членами (X34/X04); П38 - насыщенность ЭП членами причастных оборотов (X39/X04); П39 - насыщенность ЭП компонентами обособленных групп (X46/X04); П40 - насыщенность ЭП существительными без группы (X52/X04); П41 - насыщенность ЭП группами существительных (X53/X04); П42 - доля членов ГС в ЭП (X54/X04); П43 - доля местоимений среди подлежащих (X33/X32); П44 - средний размер группы однородных членов (X35/X34); П45 - доля сказуемых среди однородных членов (X36/X34); П46 - средний размер причастного оборота (X39/X38); П47 - средний размер распространенного причастного определения (X41/X40); П48 - доля причастий среди согласованных определений (X43/X42); П49 - доля существительных среди несогласованных определений (X45/X44); П50 - средний размер обособленного члена (X47/X46); П51 - средний размер абсолютного оборота (X49/X48); П52 - средний размер инфинитивного оборота (X51/X50); П53 - средний размер ГС (X54/X53); П54 - доля знаменательных слов в ГС (X55/X54); П55 - доля служебных слов в ГС (X56/X54); П56 - соотношение между средними размерами ГС и средним объемом ЭП (П53/П01).

Таким образом, каждый объект, подвергаемый анализу в целях стилистической диагностики, может быть описан при помощи данного набора из 112 первичных и производных параметров. Необходимо учитывать, однако, что для каждого национального языка набор параметров может несколько сокращаться в силу особенностей структуры этого языка. В общем же виде, каждый диагностируемый объект может быть представлен в виде точки в 112-мерном пространстве или 112-мерного вектора. Поскольку априорных сведений об информативности каждого параметра не имеется, в ходе эксперимента должны быть решены следующие задачи: а) разработана методика уменьшения признакового пространства (уменьшение его размерности); б) построен ряд параметров, ранжированных по убыванию/возрастанию индивидуальной информативности; в) установлена обоснованность и целесообразность включения в признаковое пространство диагности-

руемого объекта производных параметров, так как они выводятся из первичных и могут оказаться избыточными.

Материалом для эксперимента послужила случайная выборка объемом 100 предложений авторской речи из художественных прозаических произведений 10 русских писателей конца XIX - XX века, на которой были установлены значения 103 параметров (X48, X49, X50, X51 и П35, П51, П52 в русском языке не существуют, XII и XI2 в выборке не зафиксированы).

Разработка алгоритмов свертывания признакового пространства ведется уже в течение нескольких десятилетий, однако теоретические достижения медленно входят в практику. До сих пор, даже в работах по теории классификации встречаются рекомендации выбирать из источников, описывающих некоторую предметную область, наборы параметров, отождествлять в разных наборах элементы, носящие разные имена у разных авторов (подобно тому, как это было сделанов описываемом эксперименте), а затем ... получать пересечение всех наборов: $M = M_1 \cap M_2 \cap \ldots \cap M_n$. Полученный набор предлагается считать "хорошим" набором (В.И.Драгуновский.Выступление на конференции "Теория классификации и анализ данных". Цит. по: Грейсух, 1982).

Гораздо более мощным методическим приемом является выделение признаков, показывающих наибольшую меру различия и, следовательно, наиболее значимых для диагностики речевых стилей, среди более или менее случайного набора самых разнообразных синтаксических характеристик речи, число которых может быть весьма большим (Лейкина и др., 1973). Такой подход позволяет отказаться от априорно задаваемых исследователем характеристик и построить формализованный алгоритм свертывания.

Эмпирические данные, полученные на разном материале, неоднократно давали основания говорить о существовании зависимостей между различными признаками. Так, например, Л.В.Малаховский установил взаимозависимость между объемом элементарного предложения и объемом цельного предложения, длиной цельного и элементарного предложений и т.д. (Малаховский, 1981). И.П.Севбо отмечала тесную связь между длиной пути в графе и количеством слов во фразе (Севбо, 1981). Много данных о взаимозависимостях различных параметров приводится в работах Я.А.Микка (Микк, 1975). Выявленные взаимозависимости являются лишь немногими компонентами системы "скрытой упорядоченности" в объеме синтаксических единиц, которая, будучи впервые выявленной Г.Аренсом на уровне слова и пред-

ложения, была затем подтверждена на более высоком уровне - на уровне предложения и абзаца. Здесь речь идет об общей тенденции к "изоквантности", то есть о взаимных связях тех синтаксических единиц равных уровней, из которых строится речь (Иванов, 1976).

Очевидно, что эти данные служат лишь потверждением одного из основных положений материалистической диалектики о том, что все явления действительности не изолированы друг от друга, а находятся в многообразных связях, отношениях. В.И.Ленин писал, что "всякое отдельное тысячами переходов связано с другого рода отдельными (вещами, явлениями, процессами) и т.д." В другом месте В.И.Ленин подчеркивает, что "отношения между вещами суть реальные отношения, вытекающие из природы вещей" (Ленин, ПСС). Поэтому методологически правильный алгоритм свертывания признакового пространства, учитывающий взаимозависимости между параметрами объекта, должен быть построен на учете всех связей между рассматриваемыми параметрами.

Анализ эмпирических данных о взаимозависимостях между различными параметрами позволил исследователям выдвинуть предположения о нецелесообразности измерения в тексте всех возможных параметров. В ряде случаев, при высокой степени совпадений, отмечались случаи функциональной зависимости, когда значение одного параметра полностью определяло значение другого. Поэтому смысл применяемой в данном эксперименте меры связи двух признаков не в том, что она оценивает степень отклонения их совместного распределения от независимости, а в том, что она дает возможность прогнозировать значение одного из них по значениям другого (Мирхин, 1980).

С этой целью на ЭВМ была получена корреляционная матрица связи параметров, элементами которой являются парные коэффициенты корреляции Пирсона, определяемые по формуле:

$$r=\frac{N\sum x_{i1}x_{i2}-(\sum x_{i1})(\sum x_{i2})}{[N\sum x_{i1}^2-(\sum x_{i1})^2]^{1/2}\cdot[N\sum x_{i2}^2-(\sum x_{i2})^2]^{1/2}},$$

где N - число пар значений, x_{i1}, x_{i2} - значения парных признаков. Таким образом была получена квадратная матрица порядка 103х х103 (по числу рассматриваемых признаков). Даже если принять во внимание, что такая матрица обладает свойством симметричности относительно главной диагонали и что диагональные ее элементы равны единице, все равно число оставшихся элементов (5.253) не позволяет содержательно интерпретировать выявленные связи между параметрами и требует использования специальных методов обработки данных.

В настоящее время известно несколько конкурирующих методов отбора информативных параметров на основе корреляционной матрицы. Одним из наиболее простых методов, допускающих к тому же ручную реализацию, является метод "корреляционных плеяд" (Терентьев, 1959; Терентьев, 1960; Выханду, 1964). В основе этого метода лежит предположение, что в основе сильной связи между параметрами (большие коэффициенты корреляции) лежит некоторый непосредственно ненаблюдаемый внутренний фактор, обусловливающий эту связь. Поэтому параметры группируются в корреляционные плеяды, каждая из которых объединяет близкие между собой параметры, а в разные плеяды попадают параметры, слабо связанные между собой. После разбиения параметров на плеяды можно оценить, сколько "представителей" от каждой группы нужно отобрать для последующего анализа. Однако при большом числе параметров количество связей настолько велико, что для выделения корреляционных плеяд приходится прибегать к формальным методам расщепления признакового пространства на уровни. Очевидно, целесообразно принимать во внимание только статистически существенные значения парных коэффициентов корреляции. Для проверки существенности парного коэффициента корреляции определяется величина t-критерия Стьюдента (Вайну, 1977):

$$t_r = r(N-2)^{1/2}/(1-r^2)^{1/2},$$

имеющая t- распределение с $N-2$ степенями свободы.

Преобразовав эту формулу, получим пороговое значение коэффициента корреляции:

$$r_{кр.} = t_r/[(N-2)+t_r^2]^{1/2}.$$

При доверительных уровнях 95% и 99% и объеме выборки $N=100$, табличные значения составят $t_{98;0,05}=1,659$ и $t_{98;0,01}=2,363$. Соответственно, пороговые значения коэффициентов корреляции будут составлять 0,165 и 0,232. Даже при учете одних только значимых связей картина остается весьма сложной. Аналогичный результат получается и при использовании других графических методов обработки.

Представляется, что гораздо более алгоритмизуемым является подход, основанный на критерии минимизации потерь информативности, поскольку все необходимые данные могут быть получены из самой совокупности полученных связей. Однако в задачах распознавания образов в целом, и в стилистической диагностике в частности, критерии информативности являются противоположными по сравнению

с традиционными задачами классификации, где главной целью ставится получение однородных групп признаков. Так, в методе корреляционных плеяд, где ставится задача выбора одного "представителя" из группы однородных признаков, наиболее информативными считаются признаки, имеющие максимальные суммы коэффициентов корреляции (Выханду, 1960). Напротив, для задач распознавания наибольший интерес, как уже указывалось выше, представляют признаки с минимальными значениями сумм коэффициентов корреляции, так как они показывают наибольшую меру различия и, следовательно, наиболее значимы для дифференциации данных речевых стилей.

Исходя из этих положений, были определены суммы коэффициентов корреляции для каждого признака, т.е. произведено суммирование по столбцам корреляционной матрицы (без учета диагональных элементов). Поскольку значения $\sum_{i=1}^{n} r_i$ варьируют в интервале (35,918; -18,331), было произведено расщепление на оптимальные интервалы по формуле Стерджесса (Венецкий, Венецкая, 1979):

$$\mathcal{Z} = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n},$$

где n - число единиц в совокупности, x_{max} и x_{min} - наибольший и наименьший варианты. При x_{max} =35,918, x_{min} =-18,331 и n =103, $\mathcal{Z}$ =7,057.

Таким образом, все признаковое пространство было разбито на 8 уровней, и значения признаков распределились следующим образом (отдельно по первичным и производным признакам):

Интервал расщепления	Первичные признаки	Производные признаки
(35,918; 28,861)	{X01, X02, XI9, XI8, X54, X55}	{0}
(28,861; 2I,804)	{X53, X35, X26, X42, X34, X56, X2I, X09, X23, X08, X27, X22, X20, X32}	{П07, П44}
(2I,804; I4,747)	{X30, X04, X52, X05, X45, X28, XI6, X44, XI7, X39 X24, X33, X46, X47, XI5}	{П05, П04, П03, П0I, П08, П42, П06, П53, П26, П46, П2I, П49, П34, ПI2}
(I4,747; 7,690)	{X38, XI0, X37, X4I, X36, X43, X3I, X40, XI4}	{П4I, П38, П55, П50, П47, П36, П37, П33, П09, П40, П45, П48, П30, ПI7}
(7,690; 0,633)	{X06, X07, X29}	{П20, П43, ПI9, П22, П56, П32, П54, ПI0,

(0,633;-6,424)	{X03}	П24, П13, П27, П15} {П25, П31, П14, П29 П23, П02}
(-6,424;-13,481)	{0}	{П18, П28}
(-13,481;-18,331)	{0}	{П11, П16}

Анализ таблицы группировки признаков позволяет сделать следующие выводы:

1. Многомерное признаковое пространство может быть алгоритмически преобразовано в одномерный ряд признаков, ранжированных по убыванию/возрастанию информативности.
2. Как показывает распределение признаков по интервалам, производные признаки в общем оказываются более информативными, чем первичные признаки.
3. При отборе необходимых и достаточных информативных параметров необходимо отдавать приоритет признакам с минимальным значением $\sum r_i$.

Алгоритм формирования минимальной необходимой и достаточной совокупности информативных параметров будет описан в последующих публикациях автора.

ЛИТЕРАТУРА

Ленин В.И. Полное собрание сочинений, т.29, с.318, 482.

Айвазян С.А. О некоторых направлениях применения многомерного статистического анализа в социально-экономических исследованиях.- В кн.: Алгоритмы многомерного статистического анализа и их применение. М.:ЦЭМИ, 1975.

Айвазян С.А., Бешаева З.И., Староверов С.В. Классификация многомерных наблюдений.- М.: Статистика, 1974.

Вайну Я.Я.-Ф. Корреляция рядов динамики.-М.:Статистика, 1977.

Венецкий И.Г., Венецкая В.И. Основные математико-статистические понятия и формулы в экономическом анализе. М.: Статистика, 1979.

Выханду Л.К. Об исследовании многопризнаковых биологических систем. - В кн.: Применение математических методов в биологии, вып.3. Л.: Изд-во ЛГУ, 1960.

Ефимов А.И. Стилистика русского языка.-М.: Просвещение, 1969.

Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе. - В кн.:Классификация и кластер. М.: Мир, 1980.

Иванов В.И. Соотношение размеров предложения и абзаца.- Вопросы языкознания , 1976, №1.
Ковалевский В.А. Задача распознавания образов с точки зрения математической статистики. - В кн.: Читающие автоматы и распознавание образов.-Киев: Наукова думка, 1965.
Лейкина Б.М., Откупщикова М.И., Случаевский Ф.И., Цейтин Г.С., Щербатов Б.А. Анализ некоторых числовых характеристик статистического исследования речи при патологии мышления. - В кн.: Лингвистические проблемы функционального моделирования речевой деятельности. Л.: Изд-во ЛГУ, вып.I, 1973.
Малаховский Л.В. Эволюция размеров слова и структуры предложения в английской научной прозе XVII-XX вв. - В кн.: Структура и объем предложения и словосочетания в индоевропейских языках. Л.: Наука, 1981.
Микк Я.А. Методика измерения трудности текста. - Вопросы психологии , 1975, №3.
Миркин В.Г. Анализ качественных признаков и структур.-М.: Статистика, 1980.
Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика.-Киев: Наукова думка, 1981.
Терентьев П.В. Метод корреляционных плеяд. - "Вестник ЛГУ", 1959, №9.
Терентьев П.В. Дальнейшее развитие метода корреляционных плеяд. - В кн.: Применение математических методов в биологии, вып.3. Л.: Изд-во ЛГУ, 1960.
Kjetsaa G. Storms on the Quiet Don: A pilot study. - Scando-Slavica, 1976, №22.
Yule G.U. On sentence-length as a statistical characteristics of style in prose: with application to two cases of disputed authorship. - Biometrika, vol.30, №3-4, 1939.

THE OPTIMAL REDUCTION OF PARAMETRICAL SPACE IN THE PROBLEMS OF STYLISTIC DIAGNOSTICS

Mikhail Marusenko

S u m m a r y

The problems of stylistic diagnostics are formulated in this article in the terms of the recognition of images. From this point of view the problem of the reduction of dimension of the initial parametrical space is formulated and the experiment on transformation of 112-dimension parametrical space into undimensional parametrical series rated according to their individual information capability is described.

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ПАТОЛОГИЧЕСКОГО ТЕКСТА

В.Э. Пашковский

Значение объективного анализа текстов душевнобольных осознавалось многими исследователями. Так, в текстах, составленных больными шизофренией, определялось отношение словарного запаса к объему текста (TTR), доли личных местоимений, существительных с абстрактными значениями (Fairbanks, H., 1944; Fliegel, H., 1965). Исследовалась также длина предложений, доли многосложных слов (Whiteborn, J., Zipf, G.K., 1943; Lorenz, M., 1953; Lorenz, M., Cobb, S., 1954), энтропийные показатели речи (Андреев М.П., Аминев Г.А., 1968). Эти работы продемонстрировали возможности лексикостатистического анализа, однако, несмотря на множество выделенных авторами лингвостатистических показателей, вопросы частотной структуры лексики в патологических текстах изучены недостаточно.

В последние годы в лингвостатистических исследованиях для получения поддающихся сравнению лингвостатистических показателей используются выборки, принадлежащие одному и тому же функциональному стилю (литературно-художественные тексты, научно-технические тексты и т.д.). Необходимым условием является также равенство объемов выборок, достигающих нескольких сот тысяч словоупотреблений. К сожалению, такой принцип не всегда возможно осуществить при исследовании патологических текстов. Объемы текстов, составленных психически больными малы - от нескольких сот до нескольких десятков тысяч словоупотреблений, а их жанровая принадлежность аморфна и неопределенна. Эта особенность патологических текстов создает необходимость поиска таких моделей, которые удовлетворяли бы следующим требованиям:

1) возможность сжатого описания статистической структуры текста;

2) относительная независимость сравниваемых текстов от объема и жанра;

3) возможность проверки тех или иных статистических гипотез с использованием данных исследуемых выборок.

Целью настоящей работы является изучение частотной структуры и относительного богатства лексики у больных параноид-

ной шизофренией с различной степенью выраженности патологических сдвигов в клинической картине заболевания.

Для получения текстового материала больным параноидной шизофренией предлагалось дать письменное определение словам: "береза, хлеб, жизнь, отпускать, лететь, сухой, каменный, быстро, рядом, нельзя". В исследовании участвовало 93 человека больных параноидной шизофренией и 112 здоровых людей. Исследуемый текстовой материал был разделен на четыре выборки: первая - тексты больных в начальной стадии заболевания, 19 человек; вторая - тексты больных с развернутой галлюцинаторно-параноидной симптоматикой, 20 человек; третья - тексты больных с выраженными клиническими расстройствами мышления и с более давними сроками заболевания, 49 человек; четвертая - тексты здоровых испытуемых, 112 человек. Все больные исследовались по минованию острой эффективной (страх, подозрительность, напряженность, тревожность) симптоматики.

Все исследуемые и контрольные тексты были преобразованы в частотно-алфавитные списки, где каждая лексическая единица записывалась в форме лексемы. Под лексемой мы понимаем слово, записанное в каноническом виде, например: существительное в именительном падеже, единственного числа и т.д.. Ниже приводятся выписки из протоколов здоровых и больных испытуемых. В скобках приводится абсолютная частота лексем по данным частотного словаря группы, к которой принадлежал испытуемый.

Испытуемый К., 4-я группа (здоровые). Результаты исследования: "Береза - дерево (48) с (47) белой (24) корой (10). Хлеб - продукт (28) питания (16). Жизнь - существование (33) органических (1) веществ (15)." У данного испытуемого, как и у других испытуемых группы здоровых, определения близки к словарным. Выявлена тенденция к выявлению родовых и видовых признаков определяемых понятий. Подсчитано соотношение одноразовых лексем ко всему объему текста дефиниций десяти определяемых слов в протоколе испытуемого. Это соотношение равно 0,111.

Больной Г., 1-я группа. Результаты исследования: "Береза - дерево (20). Хлеб - богатство (2). Жизнь - сильная (1). Лететь - на (8) самолете (4). Отпускать - в (12) срок (1). Сухой - чистый (1), честный (2)". Как видно из протокола, в одном случае определения логически корректны (береза - дерево), в других - используются слова, которые по смыс-

лу не сочетаются (сухой - чистый, честный; жизнь - сильная). Отношение одноразовых лексем к объему текста равно 0,428.

Больной В., 2-я группа. Результаты исследования: "Береза - дерево (20). Хлеб - пища (5). Жизнь - большая (I) птица (7), приносящая (I) человечеству (I) радость (I). Лететь - осознавать (I) себя (I) на (I5) месте (I) этой (I) птицы (7)". В данном случае определения разнородны. В одних случаях они носят обычный характер (береза - дерево, хлеб - пища). В других метафоричны, индивидуальны - (жизнь - большая птица). Соотношение одноразовых лексем к объему текста высокое и равно 0,607.

Больной М., 3-я группа. Результаты исследования: "Береза - лиственное (6) дерево (I3) растущее (5) в (33) средней (5) полосе (7), бумага (2), береста (I), тепло (I), уют (I), ценная (2) порода (5), сады (2), кислород (I), целлюлоза (2), импорт (I), экспорт (I), сок (2), выработка (I), прививок (I). Хлеб - пища (I2), пшеница (8), рожь (2), для (6) животных (3) - ячмень (I), просо (I), вырабатывается (I) жерновами (I), мельница (I), специальным (I) заводом (I), жизненная (3) необходимость (I), есть (I0) много (3) сортов (I), животный (I) мир (7), мир (7) людей (4), живой (4) человек (2I), природа (5) сложилась (I) издавна (I), это (34) все (I3) делает (I) обмен (I) веществ (I), желудок (I), кишечник (I), нервы (I) - все (I3) это (34) физическое (I), умственное (I), терапевтическое (I), государственное (I), патологическое (I), вразумительное (I) развитие (I)". В определениях отмечается разноплановость, определения представляют собой как бы автоматизированный поток ассоциаций, в результате чего создается впечатление о "течении мыслей по нескольким руслам одновременно". Соотношение одноразовых лексем к объему текста равно 0,500.

Из приведенных примеров видно, что по мере утяжеления патологического процесса нарушается дифференцировка между значением и смыслом определяемых слов. Определения становятся все более индивидуальными и непонятными для окружающих. Представляет интерес проследить, отражают ли лингвистические показатели глубину патологических сдвигов в клинической картине заболевания.

В таблице I приведены фрагменты частотных словарей исследуемых и контрольных выборок.

Таблица I.

I-я выборка N =490, $\hat{v}$ =233						2-я выборка N =930, $\hat{v}$ =442					
i	лексемы	F	F^*	P	P^*	i	лексемы	F	F^*	P	P^*
I	это	22	22	0.045	0.045	I	в	24	24	0.026	0.026
2	дерево	20	42	0.041	0.086	2	дерево	20	44	0.021	0.047
3	что	I3	55	0.026	0.112	3	не	20	64	0.021	0.068
4	в	I2	67	0.024	0.136	4	быть	I7	8I	0.018	0.086
5	либо	I2	79	0.024	0.160	5	это	I7	98	0.018	0.104
6	делать	9	88	0.018	0.178	6	и	I4	II2	0.015	0.119
7	не	9	97	0.018	0.196	7	человек	I4	I26	0.015	0.134
8	продукт	9	I06	0.018	0.214	8	который	I3	I39	0.014	0.148
9	на	8	II4	0.016	0.230	9	на	I3	I52	0.014	0.162
I0	друго*	7	I2I	0.014	0.244	I0	из	9	I6I	0.009	0.171
II	двигать	6	I27	0.012	0.256	II	отпускать	9	I70	0.009	0.180
I2	куда	6	I33	0.012	0.268	I2	предмет	9	I79	0.009	0.189
I3	питание	6	I39	0.012	0.280	I3	рядом	9	I88	0.009	0.198
I4	пища	6	I45	0.012	0.292	I4	жизнь	8	I96	0.008	0.206
I5	с	6	I5I	0.012	0.304	I5	сухой	8	204	0.008	0.214
I6	воздух	5	I56	0.010	0.314	I6	хлеб	8	2I2	0.008	0.222
I7	движение	5	I6I	0.010	0.324	I7	что	8	220	0.008	0.230
I8	из	5	I66	0.010	0.334	I8	воздух	7	227	0.007	0.237
I9	предмет	5	I7I	0.010	0.344	I9	значить	7	234	0.007	0.244

Продолжение таблицы I

i	лексемы	F	F*	p	p*	i	лексемы	F	F*	p	p*
20	твердый	5	176	0.010	0.354	20	камень	7	241	0.007	0.251
...						...					
...						...					
81	аппарат	I	333	0.002	0.684	I36	а	I	624	0.00I	0.693
82	бегом	I	334	0.002	0.686	I37	абстрактный	I	625	0.00I	0.694
83	бежать	I	335	0.002	0.688	I38	активность	I	626	0.00I	0.695
84	без	I	336	0.002	0.690	I39	аппарат	I	627	0.00I	0.696
85	безвлажный	I	337	0.002	0.692	I40	армия	I	628	0.00I	0.697
86	белый	I	338	0.002	0.694	I4I	басня	I	629	0.00I	0.698
87	бессердечный	I	339	0.002	0.696	I42	бегом	I	630	0.00I	0.699
88	близкий	I	340	0.002	0.698	I43	безвоздушный	I	63I	0.00I	0.700
89	близко	I	34I	0.002	0.700	I44	бездушный	I	632	0.00I	0.70I
90	больше	I	342	0.002	0.702	I45	безжизненный	I	633	0.00I	0.702
9I	большой	I	343	0.002	0.704	I46	белизна	I	634	0.00I	0.703
92	булка	I	344	0.002	0.706	I47	белковый	I	635	0.00I	0.704
93	важный	I	345	0.002	0.708	I48	береза	I	636	0.00I	0.705
94	вблизи	I	346	0.002	0.7I0	I49	березовый	I	637	0.00I	0.706
95	век	I	347	0.002	0.7I2	I50	бесчувственный	I	638	0.00I	0.707
96	весна	I	348	0.002	0.7I4	I5I	близкий	I	639	0.00I	0.708

Продолжение таблицы I

i	лексемы	F	F^x	P	P^x	i	лексемы	F	F^x	P	P^x
97	ветер	I	349	0.002	0.7I6	I52	близость	I	640	0.00I	0.709
98	вещество	I	350	0.002	0.7I8	I53	больница	I	64I	0.00I	0.7I0
99	вид	I	35I	0.002	0.720	I54	больной	I	642	0.00I	0.7II
I00	влажный	I	352	0.002	0.722	I55	большой	I	643	0.00I	0.7I2

	3-я выборка N =I758, $\hat{\nu}$ =654						4-я выборка N =2594, $\hat{\nu}$ =5I4				
i	лексемы	F	F^x	P	P^x	i	лексемы	F	F^x	P	P^x
I	не	4I	4I	0.023	0.023	I	в	74	74	0.028	0.028
2	это	34	75	0.0I9	0.042	2	из	74	I48	0.028	0.056
3	в	33	I03	0.0I8	0.060	3	дерево	48	I96	0.0I8	0.074
4	и	28	I36	0.0I5	0.075	4	с	47	243	0.0I8	0.092
5	что	27	I63	0.0I5	0.090	5	сделанный	45	288	0.0I7	0.I09
6	жизнь	23	I86	0.0I3	0.I03	6	не	44	332	0.0I6	0.I25
7	из	22	208	0.0I2	0.II5	7	передвигать	40	372	0.0I5	0.I40
8	на	22	230	0.0I2	0.I27	8	камень	36	408	0.0I3	0.I53
9	значить	2I	25I	0.0II	0.I38	9	на	36	444	0.0I3	0.I66
I0	человек	2I	272	0.0II	0.I49	I0	содержащий	35	479	0.0I3	0.I79
II	хлеб	I9	29I	0.0I0	0.I59	II	растущий	34	5I3	0.0I3	0.I92

Продолжение таблицы I

i	лексемы	F	F^x	P	P^x	i	лексемы	F	F^x	P	P^x
I2	сухой	I8	309	0.0I0	0.I69	I2	существование	33	546	0.0I2	0.204
I3	быстро	I7	326	0.009	0.I78	I3	освобождать	30	576	0.0II	0.2I5
I4	каменный	I7	343	0.009	0.I87	I4	время	29	605	0.0II	0.226
I5	нельзя	I5	358	0.008	0.I95	I5	действие	28	633	0.0I0	0.236
I6	отпускать	I5	373	0.008	0.203	I6	или	28	66I	0.0I0	0.246
I7	рядом	I5	388	0.008	0.2II	I7	продукт	28	689	0.0I0	0.256
I8	слово	I5	403	0.008	0.2I9	I8	происходить	27	7I6	0.0I0	0.266
I9	твердый	I5	4I8	0.008	0.227	I9	либо	26	742	0.0I0	0.276
20	быть	I3	43I	0.007	0.234	20	близко	25	767	0.009	0.285
...						...					
...						...					
253	авиация	I	I357	0.0005	0.7990	254	антоним	I	2334	0.0004	0.8956
254	актуальный	I	I358	0.0005	0.7995	255	атмосфера	I	2335	0.0004	0.8960
255	аппарат	I	I359	0.0005	0.8000	256	батон	I	2336	0.0004	0.8964
256	армия	I	I360	0.0005	0.8005	257	безводный	I	2337	0.0004	0.8968
257	ассоциировать	I	I36I	0.0005	0.80I0	258	безвоздушный	I	2338	0.0004	0.8972
258	безветренность	I	I362	0.0005	0.80I5	259	безжизненный	I	2339	0.0004	0.8976
259	безводный	I	I363	0.0005	0.8020	260	блюдо	I	2340	0.0004	0.8980
260	белизна	I	I364	0.0005	0.8025	26I	более	I	234I	0.0004	0.8984

Продолжение таблицы I

i	лексемы	F	F^x	P	Px	i	лексемы	F	F^x	P	Px
261	белковый	I	1365	0.0005	0.8030	262	больше	I	2342	0.0004	0.8988
262	белок	I	1366	0.0005	0.8035	263	булочная	I	2343	0.0004	0.8992
263	береста	I	1367	0.0005	0.8040	264	важный	I	2344	0.0004	0.8996
264	бесчевствен-ный	I	1368	0.0005	0.8045	265	витамин	I	2345	0.0004	0.9000
265	биться	I	1369	0.0005	0.8050	266	вкусный	I	2346	0.0004	0.9004
266	ближний	I	1370	0.0005	0.8055	267	вместе	I	2347	0.0004	0.9008
267	близкий	I	1371	0.0005	0.8060	268	внешний	I	2348	0.0004	0.9012
268	близость	I	1372	0.0005	0.8065	269	воля	I	2349	0.0004	0.9016
269	бог	I	1373	0.0005	0.8070	270	всякий	I	2350	0.0004	0.9020
270	бок	I	1374	0.0005	0.8075	271	второй	I	2351	0.0004	0.9024
271	болеть	I	1375	0.0005	0.8080	272	выводить	I	2352	0.0004	0.9028
272	больница	I	1376	0.0005	0.8085	273	вырабатывать	I	2353	0.0004	0.9032

Условные обозначения: N – объем текста; $\hat{v}$ – словарь текста; i – ранг; F – абсолютная частота лексической единицы; F^x – ее накопленная абсолютная частота; P – относительная частота лексической единицы; P^x – ее накопленная относительная частота.

Как видно из таблицы I, среди первых наиболее частых 20-и лексем в исследуемых и контрольных выборках расположены служебные слова, местоимения, значимые существительные, глаголы, прилагательные, наиболее тесно связанные с определяемыми словами. Обращает на себя внимание, что в выборках больных людей преобладают сами тестовые слова (рядом,жизнь, сухой, хлеб). Это говорит о тенденции больных к даче тавтологических ответов.

В таблице 2 представлены данные частотного спектра лексики исследуемых и контрольных выборок.

Таблица 2.

F	$\hat{V}m_1$ (абс)	$\hat{V}m_2$ (абс)	$\hat{V}m_3$ (абс)	$\hat{V}m_4$ (абс)	$\hat{V}m_1$ (%)	$\hat{V}m_2$ (%)	$\hat{V}m_3$ (%)	$\hat{V}m_4$ (%)
I	I58	307	402	26I	66,4	69,5	6I,5	50,8
2	35	47	I04	69	I4,7	I0,6	I5,9	I3,4
3	I5	23	37	33	6,3	5,2	5,7	6,4
4	9	20	I9	23	3,8	4,5	2,9	4,5
5	6	I4	24	2I	2,5	3,2	3,7	4,I
6	5	8	II	7	2,I	I,8	I,7	I,4
7	I	6	I0	9	0,4	I,4	I,5	I,7
8	I	4	8	8	0,4	0,9	I,2	I,6
9	3	4	5	7	I,3	0,9	0,8	I,4
I0	0	0	3	5	0	0	0,5	0,9
>I0	5	9	3I	7I	2,I	2,0	4,6	I3,8

Условные обозначения:

$\hat{V}m$ – количество лексем с указанной частотой (F) соответственно для I,2,3,4 выборок в абсолютных числах и в % по отношению к словарю.

Как видно из таблицы 2, процентное содержание одноразовых лексем у больных (I,2,3-я выборки) выше, а лексем с $F >$ I0 ниже, чем у здоровых (4-я выборка). Полученные статистические показатели хотя и дают представление об исследованных выборках, однако прямое сопоставление этих показателей затруднительно из-за различий в объемах выборок.

x x x

В связи с тем, что численные значения лингвостатистических показателей претерпевают систематические изменения в зависимости от объема текста или выборок,а исследовать тексты одинаковых объемов в клинической практике не представ-

ляется возможным, мы предприняли лексикостатистический анализ письменных текстов больных шизофренией с использованием модели частотной структуры лексики, разработанной Ю.К.Орловым (1976, 1978а), в основу которой положены некоторые результаты, полученные В.М. Калининым (1964, 1965). При этом используется метод теоретического пересчета словарного запаса выборки на стандартный объем (Orlov, J.K., 1982), контролируется степень адекватности используемой теоретической модели фактической частотной структуре выборок. В связи с тем, что прогноз словарного запаса на стандартный объем представляет собой математическое ожидание случайной величины, а фактическое ее значение может колебаться в каких-то пределах, вычислялись достоверные интервалы, которые позволили оценить степень достоверности наблюдаемых расхождений (Орлов Ю.К. 1978б).

В качестве основного лингвостатистического показателя нами выбран показатель относительного богатства лексики (ОБЛ), определяемый как число разных слов, приходящихся на стандартный объем текста. Этот показатель выбран не случайно. Общеизвестно, что многие психопатологические состояния характеризуются разной степенью речевой активности. Крайние степени речевых расстройств - это многословие, "скачка идей" при маниакальных состояниях и полное отсутствие речи при депрессии, кататоническом ступоре, истерии. Однако, делая заключение о богатстве или бедности словарного запаса больного, психиатры основываются на субъективном впечатлении. Поэтому вычисление показателя ОБЛ имело бы определенное значение как в дифференциальной диагностике психопатологических расстройств, так и для контроля за действием лекарственных препаратов.

В плане исходной гипотезы мы предположили, что у больных с различной степенью выраженности патологического процесса показатели ОБЛ должны достоверно различаться, при этом ОБЛ больных в начальной стадии заболевания будет более близок к ОБЛ здоровых, чем ОБЛ больных с выраженностью патологического процесса.

По методу, изложенному в работах Ю.К. Орлова (1978а, 1978б), производилось вычисление лингвостатистических показателей, включающее несколько этапов: 1) определение исходных производных статистических характеристик, 2) контроль соответствия теоретических и фактических показателей текста, 3) вычисление ОБЛ и его доверительных интервалов. Ниже при-

водится пример такого вычисления для первой выборки.
I) По исходным данным выборки N , $\hat{V}$ определяется значение ее объема Ципфа - показателя (z) - теоретического объема текста, на котором выполняются закономерности, найденные Ципфом, и значение производных параметров: K, B, ν (Z), определяющих ход частотной кривой:

$$Npi = \frac{NK}{B + i}; \quad (1)$$

$$i = 1, 2 \ldots\ldots V(z).$$

В нашем случае Z = 6137; K = 0,1779; Vz = 1091,84.
2) На втором этапе проверялось соответствие вычислений по формуле (I) фактическим абсолютным частотам (Fi) выборки. Данные вычислений приводятся в таблице 3.

Таблица 3.

i	Npi	Fi	i	Npi	Fi
I	22,05	22	I0	6,73	7
2	I7,59	20	I5	4,86	6
3	24,64	I3	20	3,79	5
4	I2,54	I2	25	3,I2	4
5	I0,96	I2	30	2,64	4
6	9,74	9	35	2,29	3
7	8,76	9	40	2,02	3
8	7,96	9	45	I,69	3
9	7,29	8	50	I,64	2

Соответствие теоретического количества m- разовых лексем фактическому производилось с использованием реккурентной формулы $Vm(N,Z) = \frac{Vm-1(NZ) - Vm-1(Z)}{1 - \frac{Z}{N}}$. (2)

При близком к единице, используется формула

$$\nu_m(N,z) = \nu(z) \sum_{j=m}^{\infty} \frac{(1-\frac{z}{N})^{j-m}}{j^{(j+1)}}. \quad (3)$$

Результаты подсчета даны в таблице 4.

Таблица 4.

m	Vm(NZ)	$\hat{V}m$	m	Vm(NZ)	$\hat{V}m$
I	I65,53	I58	6	2,79	5
2	32,93	35	7	2,0I	I
3	I2,93	I5	8	I,5I	I
4	6,77	9	9	I,I8	3
5	4,I4	6	I0	I,0	0

$V_m(N,Z)$ - теоретическое, $\hat{V}_m$ - фактическое количество m-разовых лексем. Убедившись в близости теоретических и фактических данных, что видно из таблиц 3 и 4, можно сделать вывод об адекватности используемой модели частотной структуры лексики для исследуемой выборки. Аналогичные результаты получились также при рассчетах второй, третьей и четвертой выборок.

3) Перейдем к определению условного показателя ОБЛ, который представляет собой теоретически определенный словарный запас на стандартном объеме $N=1000$.

$$ОБЛ = V(N,Z) = V_Z \frac{\ln \frac{Z}{N}}{\frac{Z}{N} - 1} . \qquad (4)$$

Данная величина (ОБЛ) представляет собой оценку, рассчитанную по фактическим наблюдениям, и естественно содержащую в себе некоторую случайную ошибку. Для учета этой ошибки использован метод построения доверительных интервалов для подобных прогнозов. Доверительные интервалы вычислены при уровне доверительной вероятности 0,95. Процедура их построения учитывает как ошибку исходного наблюдения, так и ошибку прогнозирования.

В нашем случае $ОБЛ = V(1000,Z) = V_Z \frac{\ln \frac{Z}{1000}}{\frac{Z}{1000} - 1}$.

В результате вычислений ОБЛ оказался равным 385,61 с доверительными интервалами 326,05 ÷ 445,17. Кроме вычислений ожидаемых словарных запасов на стандартных объемах по Ю.К. Орлову, дополнительно проводилось аналогичное вычисление с помощью формулы В.М. Калинина

$$V(N) = \hat{V}(N_0) - \sum_{j \geq 1} (1 - \frac{N}{N_0})^j \hat{V}_j(N_0), \qquad (5)$$

которая позволяет вычислить $V(N)$- ожидаемый словарный запас на объеме N, если известны фактический словарь $\hat{V}(N_0)$ и количество j - разовых лексем $\hat{V}_j(N_0)(j=1,2...)$ на выборке некоторого "базового" объема N_0. Статистические характеристики исследуемых и контрольных выборок представлены в таблице 5, а значения ОБЛ (по Орлову(и их доверительные интервалы продемонстрированы на графике (рис. I).

В таблице 5 представлены четыре показателя, отражающие лексическое богатство текста: TTR, ОБЛ по Калинину, показатель Z, ОБЛ по Орлову. С помощью трех из них (ОБЛ по Калинину, показателю Z, ОБЛ по Орлову) выявлен одинаковый порядок следования выборок по "богатству лексики": 2, 3, I,4.

Таблица 5.

№ п/п	Лингвостатистические характеристики	выборки			
		I	2	3	4
I	N	490	930	I758	2594
2	$\hat{V}$	238	442	654	5I4
3	$P_{i=1}$	0,045	0,026	0,023	0,028
4	$TTR = \frac{\hat{V}}{N}$	0,486	0,474	0,372	0,I98
5	ОБЛ по Калинину (N =I000)	378	463	457	3I4
6	Z	6I37	I0747	8049	I430
7	ОБЛ по Орлову (N =I000)	386	465	456	322
8	Доверительные интервалы ОБЛ по Орлову	326 445	4I8 5I2	4I9 493	297 348

Следует отметить, что в модели Ю.К. Орлова формула роста словарного запаса В.М. Калинина (формула 3) преобразована таким образом, что вместо фактического частотного спектра представлен некий идеальный спектр ("Ципфовский спектр") на объеме Z :

$$V_m(Z) = \frac{V(Z)}{m(m+1)} . \qquad (6)$$

Высокая степень совпадения прогнозов по Ю.К. Орлову и по В.М. Калинину свидетельствует о том, что положенная в основу модели Ю.К. Орлова идеализация частотного спектра близка к фактически наблюдаемому спектру.

Показатель TTR дает иной, чем ОБЛ по Калинину, показатель z , ОБЛ по Орлову порядок следования выборок: I, 2, 3, 4. Как видно, значения TTR определяются объемом выборок.

На **заключительном этапе** исследования произведена проверка степени статистической близости обследованных выборок. При этом допускалось, что если ОБЛ одной из выборок выходит за границы доверительного интервала другой выборки, то данную пару выборок следует считать взятыми из различных статистических совокупностей.

Пример:

выборка	доверительные интервалы	ОБЛ по Орлову предыдущей выборки
2	4I8 ÷ 5I2	386 (I)
3	4I9 ÷ 493	465 (2)

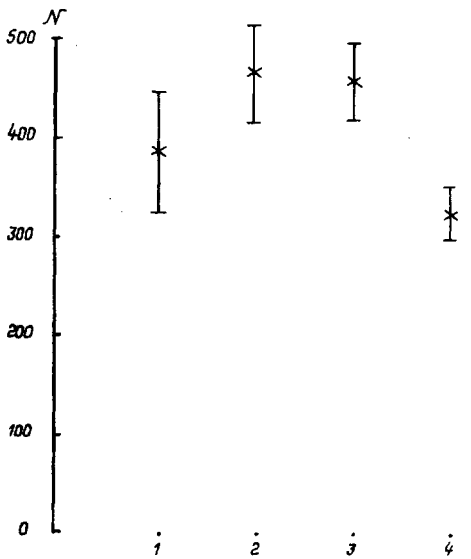


Рисунок I. Значения ОБЛ и их доверительных интервалов исследуемых и контрольных выборок.
N - объем выборки, I, 2, 3, 4 - номера выборок, × - значение ОБЛ по Орлову, $\mathbf{I}$ - доверительные интервалы ОБЛ по Орлову.

При сравнении I-й и 2-й выборок значение ОБЛ I-й выборки не вошло в границы доверительных интервалов ОБЛ 2-й выборки, а при сравнении 2-й и 3-й выборок показатель ОБЛ 2-й выборки вошел в границы доверительных интервалов 3-й выборки. Отсюда, расхождение значений ОБЛ между I-й и 2-й выборками следует считать статистически достоверными, а расхождения между ОБЛ 2-й и 3-й выборок считать статистически достоверными нельзя. В таблице 6 приводятся полученные данные. Минусом обозначено отсутствие, а плюсом наличие статистически достоверных различий между выборками.

Таблица 6.

выборка	I	2	3	4
I		+	+	+
2			-	+
3				+

Как видно из таблицы 6, обследованные пары выборок, кроме второй и третьей достоверно отличаются друг от друга.

x x x

Анализируя полученные данные, можно видеть, что показатель ОБЛ здоровых ниже, чем у больных шизофренией. Этот, казалось бы парадоксальный факт, объясняется тем, что в связи с нарушением дифференциации между значением и смыслом (в клинике этот феномен обозначается как патологический полисемантизм), дефиниции больных более разнородны, чем у здоровых. Иными словами, группа здоровых ведет себя более однообразно, чем группа больных. Операция определения у больных зачастую подменяется потоком ассоциаций, слабо связанных по смыслу с определяемым понятием. Рассматривая динамику изменения ОБЛ в текстах больных, подобранных в группы по принципу выраженности патологических сдвигов в клинической картине заболевания видно что показатели ОБЛ выборок больных в начальной стадии заболевания более близки аналогичным показателям выборок здоровых, чем выборки больных с более выраженной симптоматикой. В то же время значения показателей ОБЛ по Орлову в выборках больных с более или менее сходной симптоматикой (2 и 3 выборки) достоверно отличаются от выборок

здоровых, а различия их между собой недостоверны. Эти данные говорят о том, что показатель ОБЛ по Орлову отражает клиническую реальность.

Следует отметить, что, несмотря на статистическую неоднородность, исследуемые тексты подчиняются общим закономерностям, характеризующим объем Ципфа, то-есть редких слов много, есть небольшая группа частых слов и плавный переход между ними. Полученные с применением методов В.М. Калинина и Ю.К. Орлова данные о совпадении теоретических и фактических спектров в исследуемых выборках могут свидетельствовать об относительной сохранности системы статистической организации лингвистических единиц даже в патологических условиях (при исследуемом виде патологии).

Убедившись в возможности модели частотной структуры лексики давать уверенный прогноз и описание роста словаря на стандартном объеме, следует отметить, что в настоящем исследовании даны групповые статистические показатели патологических текстов. Для получения надежных данных для индивидуальной диагностики речевых расстройств и их психофизиологической и лингвистической интерпретации необходимо сосредоточить дальнейшие усилия на получении статистических характеристик индивидуальных текстов.

ЛИТЕРАТУРА

Андреев М.П., Аминев Г.А. Энтропийные показатели речи при шизофрении и органических заболеваниях мозга. - Журнал невропатологии и психиатрии, том 68, вып.3. М., 1968.

Калинин В.М. Некоторые статистические закона математической лингвистики. - Проблемы кибернетики, вып. П. М., 1964.

Калинин В.М. Функционалы, связанные с распределением Пуассона и статистическая структура текста. - Труды математического института им. Стеклова, LXXIX. М.-Л., 1965.

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

Орлов Ю.К. Статистическое моделирование речевых потоков. - Вопросы кибернетики, вып. 41. М.-Л., 1978а.

Орлов Ю.К. Модель частотной структуры лексики. - Исследования в области вычислительной лингвистики, вып.П, ч. I. М., 1978б.

Fairbanks,H. Studies in language behavior: II The quantitative differentiation of samples of spoken language. - In: Psychol. Monogr.vol. 56 1944.

Fliegel,H. Schizophasie in linguistischer Deutung.Berlin, 1965.

Lorenz,M. Language Behavior in Manic Patients. - In: Arch. Neurol. Psychiat. vol.69,1953.

Lorenz,M.,Cobb,S. Language patterns in psychotic and psychoneurotic subjects. - In: Arch. Neurol . Psychiat. vol.72, 1954.

Orlov,J.K. Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? - In: Sprache, Text, Kunst: Quantitative Analysen. - Bochum: Brockmayer,Studienverlag , 1982.

Whiteborn,I. Zipf,G.K.Schizophrenic Language. - In: Arch. Neurol. Psychiat. vol. 49,1943.

STATISTICAL CHARACTERISTICS OF PATHOLOGICAL TEXT

Vladimir Pashkovsky

S u m m a r y

By linguistic methods the written speech of 93 patients suffering from schizophrenia was studied. The control group was composed of 112 healthy persons. Since the sizes of the investigated samples were different, the text analysis was carried out with the help of a special model of the frequency structure of the vocabulary (Yu. K. Orlov). In so doing the degree of adequacy of the used theoretical model to the real frequency structure was controlled. The index of the "relative richness" (variety) of vocabulary - which proved to be lower in the texts of healthy persons than in those of sick ones - was calculated. The article is provided with frequency lists of words used by the healthy and sick persons and the corresponding tables of lexical spectra (word distribution).

МЕТОД КВАНТИФИКАЦИИ СВЯЗЕЙ МЕЖДУ ЭЛЕМЕНТАМИ ЯЗЫКОВОЙ СТРУКТУРЫ

М.С. Полинская

В настоящей работе представлен метод, позволяющий количественно оценивать связи между элементами какого-либо уровня языковой структуры. Метод был разработан прежде всего для измерения явлений верхних уровней языка, хотя в принципе может быть применен и не только к ним.

Статистические методы не раз использовались в грамматических и синтаксических исследованиях, см., например, обзор (Иванов, 1979), а также из недавних работ (Těšitelová, 1980). Однако, как правило, статистика в таких исследованиях описывает абсолютные показатели, являющиеся скорее "внешними проявлениями" языка (например, абсолютные показатели длины предложения, абсолютные частоты появления тех или иных структур в тексте и т.д.).

В связи с этим представляет интерес разработка метода, позволяющего получить количественную информацию о внутренней структуре, о связях между элементами исследуемого уровня языка.

§ 1. Исходные положения

Всякое предложение естественного языка есть некоторое соответствие экстралингвистической ситуации, в которой выделяется "событие" (предикат) и его "участники" (актанты). Структура предложения определяется - по крайней мере - соотношением единиц трех типов: семантических, синтаксических и референционных, или единиц коммуникативного плана. Кроме того, при анализе структурного типа предложения учитываются формальные средства кодирования, то есть поверхностная морфология в широком смысле, порядок слов (позиция в предложении), интонационное оформление.

При определении структурного типа языка происходит обобщение тех тенденций, которые обнаружены для индивидуальных конструкций в этом языке. Таким образом, тип языка определя-

ется количественным и качественным распределением в нем индивидуальных конструкций, соотношением этих конструкций*. Включая количественное распределение в число типологически и синтаксически значимых факторов, мы основываемся на убеждении, что существует положительная корреляция между частотностью конструкции в языке и ее центральностью и что пограничные (маргинальные) конструкции менее часты.

Теперь допустим, что в языке существуют какие-то количественные закономерности, характеризующие связи между элементами определенного уровня языка. По отношению к синтаксису это означает, что существуют определенные измеримые характеристики конструкции предложения. Эти характеристики можно исследовать, измерив в рамках предложения связи между актантами и предикатом (актантно-предикатные) и между актантами и актантами (межактантные). Как именно оценить эти связи?

Выдвигаются следующие предположения.

I. Каждая конструкция предложения характеризуется своей специфической количественной мерой (мерами). Эта мера, в дальнейшем называемая с т е п е н ь ю ж е с т к о с т и (СЖ), показывает, насколько жестко (строго) появление в предложении какой-либо формы одной составляющей детерминирует появление в том же предложении определенной формы другой его составляющей или других составляющих. Иначе говоря, мера СЖ показывает, насколько появление в предложении формы $\underline{x}$ из класса составляющих $\underline{P}$ ($x \in P$) требует появления//запрещает появление** в предложении формы $\underline{y}$ из класса составляющих $\underline{T}$ ($y \in T$).

2. Отношения жесткости, с которой составляющие задают/ /запрещают появление друг друга в рамках выбранного интервала, двунаправленны, но необязательно симметричны, то есть x задает y с СЖ = α, а y задает x с СЖ = β, где $\alpha \neq \beta$.

3. Каждый язык, поскольку он обладает своим набором и соотношением конструкций, может быть описан в показателях подобной жесткости.

---

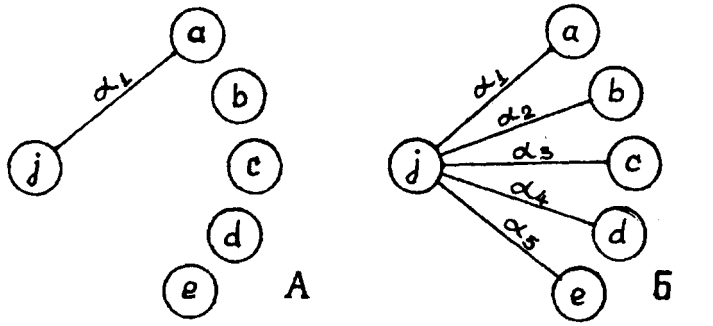
* Мы не касаемся здесь всего круга проблем, связанных с оценкой языка по тексту, т.е. по данным речи: отметим только, что на определенном этапе экстенсивного лингвистического исследования можно надеяться, что мы охватили большинство конструкций языка.

** Из этого видно, что вероятность осуществления события (E) и вероятность неосуществления того же события (¬E) не разграничены; мы сознаем необходимость такого разграничения, которое возможно при дальнейшем уточнении метода.

§ 2. Методика

Итак, у нас есть система (язык, точнее, некоторый уровень языковой структуры), в которой функционирует ряд множеств (классы, принятые за единицы анализа, см. ниже). В пределах определенного интервала элементы одного множества требуют появления//запрещают появление элементов другого множества (других множеств). Двунаправленные отношения детерминации между множествами количественно оцениваются по СЖ.

Пусть j – элемент множества X_I ($j \in X_I$), а a, b, c, d, e – элементы множества X_2 /X_2 {a, b, c, d, e}/. Если j связан с одним и только одним элементом из X_2, СЖ, с которой J задает появление этого элемента, т.е. СЖ от j,будет максимальной. Если j связан со всеми пятью элементами из X_2 равное число раз (то есть одинаковое число раз появляется совместно с каждым из этих элементов в пределах заданного интервала), СЖ, с которой j детерминирует появление каждого элемента, т.е. СЖ от j , будет минимальной. Все остальные случаи будут промежуточными. Предельные случаи представлены на рис. I.



$\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ $\quad$ $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$

СЖ = max. $\quad$ СЖ = min.

Рис. I.

Пусть теперь:

α_{ji} – число случаев совместной встречаемости элемента j из X_I с данным элементом i из X_2 в пределах заданного интервала,

n – число элементов в множестве X_I,

m – число элементов в множестве X_2,

X – любой класс составляющих предложения, принятый в анализе.

Тогда степень жесткости, с которой появление j в рамках заданного интервала требует появления//запрещает появление в нем элементов множества X_2, определяется по формуле:

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

Формула построена таким образом, что максимальная мера СЖ (случай "а", рис. I) равна I, минимальная (случай "б") равна 0. Подобный разброс величин удобен в подсчетах; он обеспечивается введением в формулу (I) соотношения $\frac{\sqrt{m}}{\sqrt{m}-1}$ (нормировочный член).

Пользуясь (I), получаем коэффициенты для каждого элемента во всех множествах (классах составляющих).

Множество репрезентировано своими элементами, следовательно, далее необходимо определить СЖ уже для всего множества составляющих. При этом необходимо принять во внимание следующее. Если элемент а из множеств X_2 встретился в пределах выборки 20 раз, а элемент b из того же множества встретился в выборке I000 раз, то а и b будут в разной мере влиять на общую СЖ множества X_2. Поэтому в следующей фазе расчетов должен быть учтен "вклад" каждого элемента из данного множества в совокупные отношения жесткости между множествами. ("Вклад", таким образом, коррелирует с абсолютной частотой появления элемента в выборке).

С этой целью для каждого элемента вычисляется средневзвешенный коэффициент сж, а затем вычисляется среднее арифметическое всех средне-взвешенных коэффициентов, полученных для элементов данного множества. Это и есть та искомая Сж, с которой какое-либо множество (класс составляющих) требует в рамках данного интервала появления* другого множества (другого класса составляющих, – например, СЖ, с которой X_I за-

* Здесь и далее для краткости опускается требование непоявления (см. об этом выше, с.I22).

дает X_2. Учет вклада элементов и определение общей СЖ проводятся по единой формуле:

$$СЖ_{X_I} = \frac{\sum_{j=I}^{n} сж_j \cdot g_j}{n} \qquad (2)$$

где

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

Помимо характеристики СЖ желательно также иметь обобщающую характеристику каждого множества (класса), принятого в анализе. Мы представляем эту характеристику как среднее от всех СЖ, полученных для данного множества; например $\overline{СЖ}_{X_I}$ – среднее тех СЖ, которые характеризуют отношение X_I к X_2, X_3, ... X_n соответственно.

Наконец, может быть введен еще один показатель, который, как выясняется из наших экспериментов, весьма существен. Это отношение $СЖ_{X_I} : СЖ_{X_2}$, то есть отношение меры и "контрмеры". По-видимому, наиболее актуальным этот парный коэффициент становится при сравнении нескольких выборок (нескольких языков, групп языков, стилей); в этих случаях предпочтителен его эксплицитный учет.

До сих пор метод обсуждался в применении к абстрактным множествам составляющих. Теперь необходимо выделить конкретные множества (классы), на основании которых возможен анализ.

§ 3. Подход

А. Какие элементы языковой структуры и какие классы этих элементов могут лечь в основу анализа? Это в большой степени зависит от цели исследования. Прежде всего условимся, что выбор подхода здесь будет ограничен уровнем предложения. За интервал принимается простое предложение*. Вторичная преди-

* Сказанное не отрицает возможности дальнейшего укрупнения и усложнения интервала.

кация не учитывается. Все придаточные, сочиненные, вставленные предложения рассматриваются как отдельные. Атрибутивные отношения учитываются только там, где на их основании удается реконструировать опущенные именные группы (NP).

Составляющие предложения могут быть описаны с точки зрения их семантических, синтаксических, референционных и формальных (кодировочных) свойств (см., например, Bechert, 1979). Однако не все перечисленные свойства одинаково легко поддаются количественной оценке. Пока неясно, как сравнивать языки на основании одних только категорий актуального членения. Затруднен и статистический учет синтаксических свойств: стало уже своего рода трюизмом говорить о нерелевантности категорий подлежащего, прямого дополнения и т.п. для целого ряда языков. Далеко не всегда просто приписать синтаксические статусы актантам в предложении подлежащного языка, - например, как трактовать актанты в русском предложении Мне хорошо видно дорогу?

Более перспективной представляется классификация элементов на основании двух других групп свойств - формальных, или свойств кодирования, и семантических. Выбор между ними диктуется, как уже было сказано, целями исследования и языковым материалом.

Рассмотрим два типа задач. Первый - сравнение ряда языков с точки зрения их типологического сходства и различия. Ясно, что в рамках такой задачи могут сравниваться языки, весьма далекие друг от друга, в частности, по морфологическому типу. А это затрудняет сравнение на основании формальных признаков.

При подобном сопоставлении наиболее удобен подход от семантики: он позволяет объединить понятийные признаки с планом выражения. В принципе опора на понятийные признаки расширяет основу для сравнения языков.

При семантическом анализе можно надеяться выяснить и ряд проблем, прямо не связанных с задачей типологического сравнения, а именно:

I. Существует ли прямая связь между семантическими свойствами (семантическими ролями актантов, семантикой предиката) и поверхностным представлением, одинаковы ли правила перехода от первого ко второму в разных языках?

2. Универсальны ли семантические роли (вернее, те теории семантических ролей, которыми мы пользуемся)?

Б. Итак, для межъязыкового анализа предлагается классификация элементов на основе их семантических свойств.

Методическая задача заключается в том, чтобы проиндексировать исследуемые предложения с точки зрения семантических свойств составляющих, то есть найти сетку взаимоотношений между множествами семантических ролей и предиката в пределах выбранного интервала.

Условимся, что мы умеем находить предикат в предложении и что нам известна поверхностная морфология данного языка, а следовательно, мы имеем право объединять алломорфы.

Прежде всего выделяется класс предикатов (Р). Это единый класс, элементами которого являются все финитные глагольные формы, формы экзистенциальных связок, прочие глагольные и глаголообразные формы и Ø (отсутствие предиката - явление поверхностной структуры). Таким образом, элементами класса Р являются все формы, которые кодируют предикат в данном языке.

При индексации предложений составляются рабочие матрицы встречаемости элементов класса Р с элементами других классов (попарно для классов). Вначале желательна максимальная дифференциация элементов каждого класса. Мы можем, как было сказано, объединять алломорфы, но не имеем права с самого начала объединять разноморфные формы в рамках порядка (ранга)*, даже если интуитивно это объединение кажется очевидным. Объединение такого рода допускается на следующей стадии анализа, если выясняется, что:

- элементы p_i и p_j различаются только одним морфом, имеющим одинаковую дистрибуцию в составе p_i и p_j;
- элементы p_i и p_j имеют одинаковое актантное окружение.

Единственный случай, когда объединение ряда форм Р возможно сразу, - моноперсональное согласование в рамках одной видовой, временной и т.п. группы. Так, при индексации нет смысла различать русские формы бежит - бегут или шел - шла - шли. Такое согласование прямо связывает предикат и актант и помечается в соответствующем классе актантов (то есть в классе данной роли выделяется как особая форма $R_{согл}$, например, Агенс - согл.).

Далее, при семантическом подходе учитывается семантика предиката. Знание глагольной семантики позволяет различать случаи, когда при предикате:

---

* О порядковом членении см.: (Ревзин, Лалашева, 1969); (Володин, Храковский, 1977).

I. актанта нет, но он может быть (сокращение, опущение актанта),

2. актанта нет и быть не может (у глагола отсутствует данная семантическая валентность).

Например, отсутствие Агенса при аффективном глаголе описывается случаем /2/, а при глаголе говорения - случаем /I/. Всякое отсутствие актанта в предложении, описываемое как (I), считается значимым и индексируется нулем (Ø). Таким образом, в ролевых классах непременно присутствует элемент Ø.

Если глагольная форма допускает двойное толкование (например, как пассив и как рефлексив) и контекст не разрешает многозначности, все предложение индексируется дважды - в соответствии с тем и с другим смыслом. Ясно, что при обработке связных текстов такие случаи будут достаточно редки.

Перейдем к обсуждению классов ролей. Отдельными элементами класса считаются те формы, которые кодируют данную семантическую роль в исследуемом языке. Знание поверхностной морфологии позволяет учесть конкретные способы падежного оформления, позиционные характеристики актантов и т.п. Ролевые классы, как следует из сказанного выше, содержат Ø; в языках с согласованием - согласованный элемент (элементы). Отдельными элементами во всех ролевых классах считаются местоимения, которые не объединяются с именами: это мотивировано распространенным специфическим поведением местоимений в различных языках.

При отборе ролей была сделана попытка примирить несколько лингвистических требований. С одной стороны, желательно было учесть данные различных падежных грамматик, а значит, сохранить в ролях максимум семантики при минимуме допущений. Это, естественно, предполагает большое число ролей. С другой стороны, в типологическом анализе приветствуется некоторое укрупнение таксонов (Успенский, 1977, 76). Укрупнения требует и подход к анализируемому материалу: ясно, что если мы учтем многообразные маргинальные случаи и раздробим токсоны, то для многих микроролей не наберется представительной статистики. Чтобы получить ее, пришлось бы обрабатывать необозримые выборки, что для ряда языков просто невозможно (язык представлен несколькими текстами).

Поэтому некоторые допущения были сделаны, рядом ролей пришлось пренебречь, а некоторые роли были сведены в гиперроли. Известно, что многие лингвистические объекты, и прежде всего семантические, наилучшим образом описываются как не-

предельные, имеющие иерархическую организацию. Таким образом, всякая семантическая роль может быть представлена как структура, обладающая ядром и периферией. Роли особенно легко будут перекрываться и пересекаться на периферии, в то время как ядро роли будет хорошо выделяться. Статистически это существенно, так как именно центральные значения будут частотнее, а значит, дадут больший вклад в искомые количественные показатели.

Выделяются следующие классы ролей: класс Агенсов (Аг), класс Пациенсов (Пац), класс Экзистантов (Экз), класс Адресатов (Адр), класс Каузаторов (Кауз), класс Инструментов (Инстр), класс Локативов (Лок).

Аг определяется в соответствии с традиционным подходом; при одноместных глаголах Аг обязательно появляется при глаголах движения, звукопроизводства, сложных глаголах с инкорпорированным именем объекта. Роль Аг приписывается к актанту при одноместных метеорологических глаголах (дождь идет). Это, конечно, допущение, но оно кажется удобным, поскольку, не будь его, пришлось бы вводить особую роль типа Сила или Стихия (Force),что статистически нежелательно.

Пац – нечто уничтожаемое//создаваемое//приводимое в данное производное состояние. Является целью действия, исходящего от Аг или агентивного Адр (таким образом, это фактически гиперроль, объединяющая традиционные Пац, Рец, Цель).

В трактовке Пац существенно также следующее. Если в языке нет формальных средств различения (а) состояния как такового вне связи с предшествующим действием и (б) производного состояния ("редуцированный пассив")*, – то есть, если конструкция многозначна, роль актанта не всегда может быть твердо определена. Тогда вводится двойное индексирование, уже упоминавшееся при описании класса Р. Конструкция индексируется как появившаяся дважды – как пассив с Пац и Аг, выраженным Ø, и как $р_{статив}$ с Экзистантом.

Термин Экз вводится потому, что основная масса случаев появления роли приходится на роль актанта при бытийном предикате. Кроме того, Экз может иметь негативное определение: роль Экз приписывается центральному актанту, который нельзя считать ни Аг, ни Пац, ни Адр и т.д. Таким образом, вводимая

* Ср. русск. дверь закрыта; японск. какару (Холодович, 1979, 156). Подобная многозначность распространена, но не универсальна. Ее нет, например, в ряде австронезийских языков, в некоторых ютоацтекских.

роль Экз частично пересекается с известными ролями Объектив, Комплетив, Дескрипт. Некоторые примеры: Он живет, Он вырос, Он умер, Валя умна, Антон – астроном, У меня$_{Адр}$ нет велосипеда$_{Экз}$.

Очевидно, при более точном теоретическом определении данной роли ее пришлось бы делить на подроли; это разумно семантически, но опять-таки нежелательно с точки зрения статистики.

Гиперролью является и Адр, однако и здесь количественные результаты будут смазаны, если мы произведем деление на роли типа Бенефактив, Экспериенцер и т.д. При индексировании посессивных конструкций, многие из которых подобны адресатным, желательно, насколько это возможно, различать посессоры с большей или меньшей агентивностью (ср. русск. У него есть собака, англ. He has a dog и более агентивное англ. He has much work to do). В ряде случаев необходимо двойное индексирование (как, например, трактовать русск. Он добился славы?). Вероятно и двойное индексирование типа Р Адр $\cup$ Р Лок (У них весело; Иван пошел к сестре).

Гиперролью может считаться и Лок. При первых испытаниях методики мы различали от-Лок, в-Лок и к-Лок (удаление, нахождение, приближение). Выясняется, что они сходны по дистрибуции и их разграничение не меняет общей картины показателей, а это дает право объединить их в гиперроль.

Для локативных актантов наблюдается интересная закономерность. В различных языках, по-видимому, независимо от типа, Лок охотно появляется и закрепляется там, где в поверхностной структуре более субъектные актанты не представлены (выражены нулями), то есть Р $\underset{Аг}{\emptyset}$ $\underset{Пац}{\emptyset}$ $\underset{Адр}{\emptyset}$ ... Лок. В целом по языкам появление Лок в таких структурах превышает ожидаемое в 2–4 раза. Это хорошая иллюстрация к шкале субъектных свойств: более периферийный актант гораздо легче тематизируется в отсутствии актантов других, более центральных ролей.

Инстр понимается вполне традиционно. Кауз в рамках данного подхода скорее гиперроль: под понятием Кауз объединяется ряд случаев, разделяемых при чисто семантическом подходе.

Кауз отличается от Аг меньшими ограничениями на агентивность//одушевленность, меньшей аффицированностью (непрямое участие, непрямой контроль ситуации, возможная ненамеренность), ср. Упавший стол$_{Кауз}$ придавил кошке хвост, Иван$_{Кауз}$

нечаянно придавил кошке хвост, Иван$_{Аг \cup Кауз}$ придавил кошке хвост. Большую часть класса Кауз составляют актанты при каузативных глаголах. В связи с каузативными глаголами в анализ вводится одна условность.

Дело в том, что актант "Каузируемый" является Пац при глаголе каузировать и одновременно - актантом некоторой роли при глаголе побуждаемого действия; во втором случае роль актанта определяется падежной рамкой глагола. Таким образом, получается следующее:

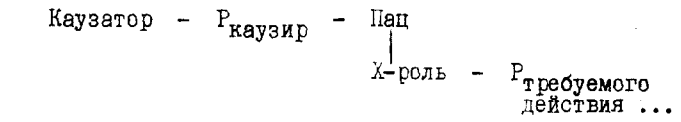


Рис. 2.

(ср. Nilsen, 1972).

В рамках простого предложения, взятого за интервал, рассматриваются - пусть с известной долей условности - однопредикатные структуры. Следовательно, все каузативные конструкции должны индексироваться как два предложения (соотв. первая и вторая ступени на рис. 2).

Не учитывается никакой роли с темпоральной семантикой: Темп - ситуативная роль, а значит, учтя ее, мы в большой степени привнесем в анализ свойства текста или жанра, что нежелательно. Кроме того, отчасти роль дублируется учетом разных временных форм Р.

В. Мы рассмотрели принципы введения классов на основе семантических свойств. Теперь обратимся к принципам определения формальных классов, на основании свойств кодирования элементов предложения.

Этот подход связан со вторым типом задач, обсуждаемым в настоящей работе, - с задачами внутриязыкового анализа.

При формальном подходе остаются те же "правила интервала", то же условие знания поверхностной морфологии и умения выделить предикат. Однако считается, что исследователю ничего не известно о семантических, синтаксических или референционных свойствах элементов предложения.

Выделяется класс Р, фиксируются все формы Р, встречающиеся в выборке. Остальные классы выделяются на основании следующих признаков: согласования с Р; грамматических показателей в широком смысле (падежное кодирование, местоименные показатели при глаголе, артикли, прочие служебные элементы);

позиции в предложении (учитывается абсолютная позиция в предложении и позиция относительно того элемента, который стоит в паре с данным); прономинализации в широком смысле.

Из перечня признаков ясно, что межъязыковое сравнение на основании связей между формальными классами на уровне предложения весьма трудно: одно и то же средство кодирования актантов может играть совершенно разную роль в двух разных языках, наличие средства кодирования К в одном языке и отсутствие его в другом препятствует сопоставлению формальных связей. Поэтому данный подход оправдан при описании структуры одного языка. В этом случае статистика может дать новые факты, требующие интерпретации, позволить сравнить разные периоды в истории языка (при условии, что его морфологический тип не изменился), наконец, может быть полезна в прикладном лингвистическом исследовании.

При формальной процедуре определения связей между классами возникает необходимость каким-то образом ранжировать признаки кодирования, чтобы найти среди них доминирующие, задающие классы. Для многих языков иерархия признаков кодирования интуитивно кажется ясной. Однако можно ввести простую процедуру ранжирования, позволяющую получать более объективные данные.

Пусть в языке есть набор различительных формальных признаков, один из которых – а. Для того чтобы оценить различительную силу а в данном языке (H_a), необходимо определить ее в каждом из N простых предложений N-мерной выборки и затем усреднить по всей выборке. Будем считать, что признак а обладает максимальной различительной силой (h_I) там, где этот признак один необходим и достаточен для различения актантов. Условно примем $h_I = I$. Если признака а недостаточно для различения актантов и он появляется в комбинации с еще одним признаком, будем считать, что его различительная сила (h_2) уменьшается вдвое. Аналогично, если признак встретился в комбинации из i признаков (то есть а и еще (i-I) признаков), h_i будет равняться $\frac{I}{i}$. Обозначим максимальное число признаков в комбинации друг с другом, встретившееся в данном языке, через М. Тогда

$$H_a = \frac{I}{N} \sum_{i=i}^{M} h_i \cdot n_i \qquad (3)$$

где h_i – различительная сила признака а в комбинации из i признаков,

n_i – количество появлений а в комбинации из i признаков.

Процедура ранжирования была применена к средствам кодирования в английском языке. Сравнивались: признак позиции, признак согласования актанта и предиката, признак наличия предложного показателя (признак падежа для местоимений мы сознательно не рассматривали). По подсчетам на выборке из 500 простых предложений выясняется, что наибольшей различительной силой обладает признак позиции, на втором месте признак наличия предложного показателя, на третьем - согласования, что согласуется с интуитивным представлением. Итак, для формального анализа английского желательно задавать признаки именных классов в указанной последовательности.

Представляется, что при внутриязыковом анализе интересные результаты можно получить, комбинируя два изложенных подхода, "семантический" и "формальный".

Наконец, рассмотрим требования, предъявляемые к выборке. Первое из них носит качественный, а не количественный характер; это требование включения в выборку связных текстов. Многие правила порождения предложений, многие свойства членов предложения никак нельзя выявить на отдельных внетекстовых примерах, как бы много их ни было. Интересные аргументы в пользу именно связных текстов как наилучших объектов лингвистического исследования приведены в (Heath, 1979).

Конечно, оценивая язык по текстам, мы всегда рискуем судить не о языке как таковом, а о речи. Однако трудно придумать какой-то иной способ получения количественной информации. Остается надеяться, что мы понижаем влияние индивидуального текста, обрабатывая большие выборки. Отсюда - два других требования к выборке.

Выборка должна быть представительной (в наших экспериментах обрабатывались тексты объемом не менее 20 стр.) и гетерогенной: сравниваются и соединяются тексты разных жанров и стилей. Здесь существенно то, что формулы (I) и (2) нечувствительны к длине предложения.

Тексты индексируются как набор простых предложений с элементами семантических или формальных классов. Для каждых двух классов составляются рабочие матрицы встречаемости элементов. Полученные данные обрабатываются, как описано в § 2. Для каждого языка получается 56 индексов семантических классов и некоторое число индексов формальных классов; индексы затем сводятся в парные показатели, см. § 4. Уменьшать число индексов, объединяя полученные показатели в более усредненные параметры, видимо, нет смысла, так как это обязательно

приведет к потере информации.

§ 4. Некоторые результаты

А. На основании семантического подхода сравнивались три выборки – латынь$_I$ (латынь Тита Ливия, конец I в. до н.э.), латынь$_2$ (латынь, на которой писал Николай Кузанский, XV в.) и современный французский. Результаты приведены в табл. I, см. там же списки выборок. Таблица построена следующим образом. Слева проставлены попарно названия классов, например, Аг/Адр. В правых столбцах приведены полученные показатели, левее знака дроби – показатель СЖ, с которой левый член пары задает "партнера", правее знака дроби – показатель СЖ от правого члена пары к левому.

Все три выборки демонстрируют существенное сходство показателей СЖ для классов предиката и наиболее центральных, субъектных ролей. Сходство уменьшается при переходе к периферийным ролям. Интересно, что в области периферийных ролей показатели лат.$_I$ и франц. также достаточно похожи. Наиболее заметные различия между лат.$_I$ и лат.$_2$ обнаруживаются в области СЖ Адр, Инстр, в паре Лок/Инстр. При расхождениях можно наблюдать одновременное нарастание или снижение показателей пары (по сравнению с другой выборкой), что согласуется с предположением о двунаправленности отношений детерминации.

Некоторые закономерности распределения СЖ можно проследить уже по этим трем группам показателей. В отношениях семантических классов можно выделить три типа: отношение "хозяина – слуги" (иерархичное), отношение равенства с взаимно низкими показателями СЖ и отношение равенства с взаимно высокими показателями СЖ. Несимметричность отношений характерна для ряда пар с Аг, где Аг явно выступает в роли "хозяина", например, в парах с более периферийными актантами. Подобные же отношения связывают Пац и Инстр; там, где в предложении есть Пац, он детерминирует появление Инстр, но не наоборот. Интересно, что то же отношение обнаруживается для Аг и Пац, что косвенно подтверждает наличие иерархий семантических ролей.

Далее, отношения субординации можно усмотреть в связях более центральных классов (предикат и более субъективные роли) и более периферийных; это указывает на принципиальную выделимость в предложении ядра и периферии.

Для более периферийных ролей обнаруживаются достаточно жесткие, примерно равные меры СЖ по отношению друг к другу.

Таблица I.

Семантический анализ, латынь$_I$, латынь$_2$, французский

Пара классов	латынь$_I$	латынь$_2$	франц.
Р/Аг	.07/.I9	.06/.I75	.08/.I8
Р/Экз	.09/.09	.08/.09	.09/.08
Р/Пац	.09/.06	.08/.06	.09/.06
Р/Адр	.II/.07	.I2/.09	.09/.08
Р/Кауз	.08/.I2	.07/.I2	.09/.II
Р/Лок	.I2/.07	.II/.05	.I0/.06
Р/Инстр	.08/.08	.I0/.09	.I2/.08
Аг/Экз	.I7/.I0	.I8/.I3	.2I/.II
Аг/Пац	.I75/.I2	.I7/.I0	.I9/.05
Аг/Адр	.2I/.08	.20/.I2	.20/.085
Аг/Кауз	.I9/.I5	.20/.I4	.22/.20
Аг/Лок	.I9/.II	.20/.I3	.I9/.09
Аг/Инстр	.20/.I0	.22/.II	.23/.I0
Пац/Экз	.I7/.20	.I6/.2I	.I55/.I9
Пац/Адр	.2I/.I8	.20/.I5	.I8/.I6
Пац/Кауз	.I5/.I7	.20/.22	.I4/.I8
Пац/Лок	.II/.06	.I5/.08	.I6/.09
Пац/Инстр	.26/.I4	.23/.I4	.2I/.I25
Адр/Экз	.I2/.25	.I6/.I9	.I4/.I9
Адр/Кауз	.I8/.2I	.I4/.20	.I6/.22
Адр/Лок	.I6/.27	.I5/.20	.I7/.235
Адр/Инстр	.I5/.22	.I6/.I9	.I5/.20
Кауз/Экз	.I0/.I2	.I4/.I4	.I6/.I4
Кауз/Лок	.I7/.09	.I9/.II5	.I5/.I0
Кауз/Инстр	.22/.I4	.20/.I5	.I85/.I7
Лок/Экз	.30/.22	.26/.22	.25/.20
Лок/Инстр	.26/.25	.2I/.23	.22/.I9
Инстр/Экз	.22/.I4	.I9/.I5	.20/.I5

Лат$_I$ – Titus Livius, Ab Urbe Condita, LL.XXI, XXII, exc. ca. 50 pp.

Лат$_2$ – Nicolai de Cusa opera omnia (Lipsiae, 1932), I, II, exc. 50 pp.

Франц. – Miscellanea, 50 pp.

По-видимому, в большом числе случаев это объясняется действием отрицательной (запретительной) детерминации: чем периферийнее составляющие, тем строже ограничения на их совместное появление в рамках простого предложения. То же конкурентное отношение между классами ответственно за высокие СЖ в паре Пац - Экз: эти роли нередко являются взаимоисключающими (см. § 3, Б). Жесткие связи в паре Адр-Пац, напротив, могут быть объяснены действием положительной детерминации; Адр и Пац сопряжены в надежной рамке целого ряда глаголов (глаголов непроизвольного действия, трехместных глаголов конверсивной семантики).

Третий тип отношений в парах наиболее существен для актантно-предикатных связей (за исключением пары Р - Аг). СЖ довольно низки, но примерно равны с обеих сторон, то есть отношение симметрично. По нулевой гипотезе, во всех парах должны получиться показатели семантической связи. Понижение СЖ в этой группе указывает на то, что семантическая связь перекрыта связью иного рода, прежде всего синтаксической. В паре Аг - предикат семантическая связь, по крайней мере, от Аг к Р, оказывается достаточно сильной.

Итак, судя по индексам актантно-предикатных пар, получаемые показатели, позволяющие соотносить содержательную структуру с поверхностным представлением, являются показателями комплексной связи: на них влияют разные структурные компоненты, в том или ином виде доходящие до поверхностного представления. Если так, ответ на вопрос о существовании одно-однозначного соответствия между семантическим компонентом и поверхностным представлением (с. 126) будет отрицательным.

Б. Формальный анализ был проведен на материале английского языка. Результаты и список выборки представлены в табл. 2. Сравнивались показатели общей выборки, повествовательного прозаического текста и научного текста. В таблице приведены СЖ для классов Р; класса -I позиции /A_I/; класса I позиции /A_2/ без предложных показателей; класса с показателем to позиции I-2 /A_3/ и класса с предложными показателями позиции I-2 /A_4/. Нетрудно видеть, что разбиение на классы, заданное в соответствии с первыми двумя ведущими признаками (алгоритмически), во многом совпадает с делением на синтаксические классы (S, DO, IO, Obl).

Полученные результаты позволяют сделать вывод о том, что стилистические различия не являются возмущающим фактором при анализе СЖ центральных классов, как и при семантическом ана-

Таблица 2

Формальный анализ, английский язык

Пара классов	Общая выборка	Выборка А	Выборка Б
P/A_I	.07/.II	.07/.075	.08/.09
P/A_2	.07/.05	.08/.05	.08/.07
P/A_3	.09/.I2	.09/.09	.09/.I3
P/A_4	.I0/.II	.I0/.I2	.II/.II
A_I/A_2	.I2/.06	.I3/.05	.I5/.05
A_I/A_3	.I3/.09	.II/.II	.I4/.II
A_I/A_4	.23/.04	.2I/.04	.25/.035
A_2/A_3	.I6/.22	.I5/.26	.I7/.28
A_2/A_4	.I9/.26	.I7/.22	.2I/.25
A_3/A_4	.3I/.27	.30/.28	.36/.32

Общая выборка – E. Hemingway, To Have and to Have Not; Trollope A., Barchester Towers, v. I; R. Aldington, Death of a Hero; "Language", 55, I (1979); "Observer", 19.II, 1978; "Morning Star", 6.10, 1980; B. Shaw, Arms and the Man; L. Hellman, The Autumn Garden; Sample B texts – Total ca. 100 pp.

Выборка А (повествовательная проза) – E. Hemingway, op. cit.; A. Trollope, op. cit., I. Murdoch, A Word Child; J. Cheever, Selected Short Stories – Total 50 pp.

Выборка Б (научная проза) – J. of the Amer. Chem. Soc.; Phys. Letters; J. of the Polynesian Soc.; Language – Total 50 pp.

лизе. Различия в СЖ растут при переходе к классам A_3 и A_4. Здесь обнаруживается ряд закономерностей. Во-первых, наблюдается та же запретительная детерминация, что и в отношениях классов периферийных семантических ролей: A_3 и A_4 вступают в отношение конкуренции. Некоторые показатели СЖ наиболее высоки в научном тексте, прежде всего в парах A_3/A_4, A_I/A_2 и A_I/A_3. Мы полагаем, что это относительное возрастание мер СЖ объясняется, во-первых, коммуникативными установками научного текста, а во-вторых, большей длиной предложения. В силу последнего актанты вынуждены жестче задавать друг друга: это как бы скрепляет конструкцию (это особенно существенно для A_I как доминантного класса; однако прирост СЖ весьма невелик). Итак, разные выборки дают весьма сходную картину, что подтверждает применимость метода при описании языка, а не стиля.

Выводы

Описанный метод количественной оценки связей между элементами языковой структуры применим к анализу связей в простом предложении; при этом классы элементов могут быть заданы как формально, так и на основании семантических признаков. Формальные признаки подлежат предварительному ранжированию.

Метод устойчиво работает при описании разных стилей одного языка (некоторые различия выявляются только для периферийных классов), следовательно, стилистические различия не являются возмущающим фактором. Таким образом, метод может быть использован для оценки актантно-предикатных и межактантных связей в предложениях данного языка. Полученные результаты позволяют предположить некоторые закономерности установления этих связей.

Поскольку метод дает сходные результаты для разных стилей и для родственных языков, он может быть применен и в типологическом анализе. При этом сравнение основывается на семантических классах. Языки могут быть сгруппированы по степени сходства показателей СЁ, особенно для центральных классов составляющих (предикат и субъектные актанты).

* *
*

Автор выражает глубокую признательность А.Я. Шайкевичу за обсуждение работы на всех ее этапах.

ЛИТЕРАТУРА

Володин А.П., Храковский В.С. Об основаниях выделения грамматических категорий (время и наклонение) – В кн.: Проблемы лингвистической типологии и структуры языка. Л., 1977.

Иванов В.С. Количественные методы в современном венгерском языкознании. – ВЯ, 1979, № 6.

Ревзин И.И., Юлдашева Г.Д. Грамматика порядков и ее использования. – ВЯ, 1969, № 1.

Успенский Б.А. К понятию диатезы. – В кн.: Проблемы лингвистической типологии и структуры языка. Л., 1977.

Холодович А.... Проблемы грамматической теории. Л., 1979.

Bechert J. Ergativity and the constitution of grammatical relations. - In: Ergativity, ed. F. Plank, N. Y.-L., 1979.

Heath J. Is Dyirbal ergative? - Linguistics, 17, 5/6 (219/220), 1979.

Nilsen D.L.F. Toward a semantic specification of deep case. The Hague, 1972.

Těšitelová M. Využití statistických metod v gramatice. Praha, 1980.

A METHOD OF STRUCTURAL RELATION QUANTIFICATION

Maria Polinskaya

S u m m a r y

A method has been proposed to quantify the relations between the elements relevant at a certain level of the linguistic structure. It has been shown possible to estimate the relations in question according to the rigidity with which the presence of an element determines the occurrence of the corresponding element within the chosen interval. A technique has been worked out to calculate the rigidity of determination (RD) relation between the two elements.

The method has been applied to the syntactic component, simple sentence taken as the interval. The elements involved have been classified according to their semantic and coding properties respectively. A procedure has been introduced to establish a Hierarchy of Coding Properties (HCP), which is language specific. The domains of the two approaches are stated. The method is shown to be indifferent to the style. The results obtained for English, French and Latin are provided (2 figs., 2 tables).

АББРЕВИАТУРЫ В ТЕРМИНОЛОГИИ КИБЕРНЕТИКИ /НА МАТЕРИАЛЕ СЛОВАРЯ ПО КИБЕРНЕТИКЕ/

Т.К. Пуздырева

Аббревиация - утвердившееся и успешно развивающееся явление в современных языках. Особенно важна ее роль в создании терминологии, так как именно в этой области в связи с бурным развитием науки и техники наиболее ощутима потребность в массовой номинации. Продуктивность аббревиации как способа образования специальной лексики отмечается многими исследователями (Алексеев Д.И., 1966; Борисов В.В., 1972; Лейчик В.М., 1981). Несмотря на то, что в последнее время проблемы аббревиации постоянно привлекают внимание лингвистов и опубликован ряд интересных работ, посвященных этим проблемам (см., например, Алексеев Д.И. 1966; Борисов В.В., 1972; Брагина А.А., 1978; Виноградов С.И., 1981;Уханов Г.П., 1962), многие вопросы, связанные с образованием и функционированием аббревиатур еще не решены. Например, до сих пор не установлено, каковы закономерности образования аббревиатур различных типов; не ясно, зависит ли структура аббревиатуры от структуры исходного словосочетания;чем объяснить, что одни аббревиатуры долгое время существуют в языке параллельно с исходными словосочетаниями, образуя с ними синонимические ряды, а другие вытесняют из употребления словосочетание и становятся самостоятельными лексическими единицами, приобретая несвойственные аббревиатерам грамматические качества (собственную категорию рода, способность склоняться, быть производящей основой), в какой мере этот процесс зависит от структурных особенностей аббревиатур. В настоящее время состояние изучения аббревиации как закономерного языкового явления таково: определен предмет и осознаны направления исследований. В этой связи нельзя не согласиться с мнением В.В. Борисова о том, что одна из важнейших задач в области создания общей теории аббревиации заключается в сборе и анализе достоверного и обширного фактического материала, на котором можно было бы строить более или менее глубокие обобщения (Борисов В.В., 1972, с. 22).

Описание структурных и семантических отношений между аббревиатурой и исходным словосочетанием имеет не только

теоретическое, но и практическое, прикладное значение. При машинной переработке текстовой информации, например, в системах автоматизированного реферирования, редактирования и информационного поиска неизбежно возникает проблема выявления взаимозаменяемых единиц текста. В круг вопросов, составляющих эту проблему, входит и соотнесение аббревиатур с исходными словосочетаниями.

В данной статье рассматриваются отношения между словами, образованными способом аббревиации, и мотивирующими их словосочетаниями, которые употребляются в языке кибернетики. Материалом для анализа послужил словник Словаря по кибернетике (1979), включающий термины, относящиеся к теоретической, технической, экономической, биологической и медицинской кибернетике и вычислительной технике. Анализ данных Словаря является первым и, как нам кажется, целесообразным этапом в исследовании использования аббревиатур в терминологии кибернетики, так как в словаре представлены наиболее употребительные, устоявшиеся термины.

Словарь содержит 2005 терминов и терминологических словосочетаний, длина которых колеблется от 2 до 8 слов (Табл. 1).

Аббревиатуры являются составляющими 151 термина. Список аббревиатур насчитывает 83 единицы, восемь из которых входят в состав более чем одного термина : ЦВМ - 33, АВМ - 22, перфокарта - 8, ЭВМ - 5, САУ - 2, ЕС - 2, ЗУ - 2, АН - 2.

Среди аббревиатур есть русские (62 единицы) и заимствованные (21 единица . Распределение аббревиатур по структурным типам представлено в табл. 2.

Русские аббревиатуры первых двух типов и мотивирующие их словосочетания синонимичны, и Словарь дает их как равноправные, взаимозаменяемые:

перфокарта, перфорационная карта;
оргтехника, организационная техника;
модем, модулятор-демодулятор.

Что же касается инициальных русских аббревиатур, то большинство из них в Словаре представлено тоже с указанием исходных словосочетаний. Это вполне соответствует традиционному пониманию аббревиатуры как второго сокращенного варианта существующего и параллельно функционирующего номинативного словосочетания (Брагина А.А., 1978; Виноградов С. И., 1981; Уханов Г.П., 1962). Но для трех аббревиатур - БЭСМ, ОГДА, АРКУС - Словарь расшифровки не дает. Возможно это сви-

детельсвует о том, что данные аббревиатуры в ходе употребления обособились от исходных словосочетаний и функционируют как самостоятельные номинативные единицы.

В тех случаях, когда сокращенное и полное названия используются в качестве синонимов, интересно сравнить структуру этих единиц. В инициальных аббревиатурах, зарегистрированных в Словаре, количество букв, в основном, совпадает с количеством слов в словосочетании. Несовпадение компонентов аббревиатуры и словосочетания бывает в двух случаях.

I. Если в исходное словосочетание входит сложное слово, то в аббревиатуре ему соответствует две буквы, например: СОЗУ - сверхоперативное запоминающее устройство, УСОПИА - универсальная система элементов промышленной пневмоавтоматики, АЦПУ - алфавитно-цифровое печатающее устройство.

2. Вследствие того, что не все слова словосочетания подвергаются усечению и вводятся в состав аббревиатуры. Могут быть пропущены предлоги и союзы. Например, при образовании аббревиатуры МИР в словосочетании машина для инженерных расчетов пропущен предлог.

В случае, если аббревиации подвергается многословное словосочетание, то аббревиатура составляется из начальных букв нескольких слов, достаточных для того, чтобы соотнести сокращенное и полное название. Таким образом создана аббревиатура ОГАС - общегосударственная автоматизированная система сбора и обработки информации. Данные Словаря показывают, что таких случаев несовпадения количества букв в аббревиатуре и слов в словосочетании сравнительно немного - 7 из 58.

Результаты анализа структурных особенностей соотнесенности аббревиатуры и исходного словосочетания необходимо учитывать при автоматизированном редактировании и реферировании научных текстов, так как оба эти процесса предусматривают идентификацию аббревиатур и соответствующих им словосочетаний, уже имеющихся в тексте, и образование новых аббревиатур как средство компрессии текста.

Как уже отмечалось, кроме русских, Словарь зафиксировал употребление и заимствованных аббревиатур в терминологии кибернетики. Большинство заимствованных аббревиатур (I9 из 2I) представляют собой транслитерации. В русском тексте они могут быть синонимичны с переводами иноязычных полных названий, поэтому соответствия между компонентами аббревиатуры и словосочетания-перевода в этих случаях нет. В Словаре есть три таких примера:

АИКА, Международная ассоциация по аналоговым вычислениям (Association Internationale pour le Calcul Analogique)

ИФИП, Международная федерация по обработке информации (International Federation for Information Processing)

ИФАК, Международная федерация по автоматическому управлению (International Federation of Automatic Control)

Остальные 16 аббревиатур-транслитераций переводов на русский язык не имеют. Все они являются названиями языков программирования, как например, ПЛ-I, АЛГОЛ, ФОРТРАН, КОБОЛ и т.п. По происхождению все они являются английскими аббревиатурами (ПЛ-I< PL - I, АЛГОЛ< ALGOL, ЛИСП< LISP, КОБОЛ< COBOL, ФОРТРАН< FORTRAN), причем за исключением двух (ПЛ-I и РПГ), принадлежащими к акронимам. Главная особенность акронимов состоит в том, что при их создании копируется или используется фонетическая структура слова. Акроним может получиться непреднамеренно, например, вуз, нэп, АСОР (Автоматизированная система организации работ), БИС (Большая интегральная схема) и пр. Но чаще создание акронима - процесс сознательный, в ходе которого аббревиатура и соответствующее ей словосочетание появляются одновременно и взаимно влияют на структуру друг друга (Борисов В.В., 1972, с. 169-225; Словообразование ..., 1968, с. 97-93). В этих случаях коррелятивной аббревиации назначение акронима отличается от назначения обычной аббревиатуры. Последняя создается с целью дать уже имеющемуся в языке описательному, иногда громоздкому, названию-словосочетанию сжатое синонимичное, более удобное в употреблении, название-слово. Акроним выполняет иную функцию: удовлетворяет возникшую потребность в образовании нового слова. Можно сказать, что мотивирующее словосочетание в процессе акронимии играет чисто вспомогательную роль. После того, как новое название создано, постепенно отпадает необходимость в соотнесении его со словосочетанием, и акроним начинает самостоятельную жизнь в языке. Например, уже немногие (даже специалисты) помнят, что МИР - это машина для инженерных расчетов и затрудняются в расшифровке БЭСМ (Большая электронная счетная машина или Быстродействующая электронная счетная машина). Если же разрыв связи между акронимом и исходным словосочетанием произошел уже в языке-источнике, то тем легче он осознается как немотивированная, неаб-

бревиатурная лексическая единица в заимствующем языке. Характерно, что из 14 заимствованных акронимов в Словаре только для двух указывается мотивирующее словосочетание: КОБОЛ (англ. COBOL сокр. от Common Business Oriented Language), АЛГОЛ (от англ. Algorithmic Language), а в текстах по кибернетике наиболее употребительные из них - ФОРТРАН, АЛГОЛ, КОБОЛ, ЛИСП - склоняются по образцу существительных 2-го склонения и графически часто оформляются не как аббревиатуры, а как собственные имена: Алгол, Фортран, Кобол, Лисп.

Акронимы, созданные на базе русского языка, в Словаре немногочисленны: БЭСМ, ЭГДА, АРКУС, МИР, БИС, АСОР, АСУП. * Эти названия в большей или меньшей мере утратили связь с мотивирующими словосочетаниями. Анализ текстов по кибернетике показывает, что аббревиатуры этой группы, особенно акронимы-амонимы (МИР, АМУР, ПАРИС), чаще встречаются без параллельного употребления соотносимых с ними полных названий.

В Словаре зафиксированы две аббревиатуры, представляющие собой заимствование с сохранением латинского алфавита: IBM - название семейства цифровых вычислительных машин, созданных в США, и UNCOL- название языка программирования. Оба слова русских эквивалентов не имеют.

Несмотря на то, что в Словаре представлены далеко не все аббревиатуры, употребляемые в текстах по кибернетике, очевиден тот факт, что имеются определенные трудности в соотнесении аббревиатуры с исходным словосочетанием, а иногда связь между ними рвется.

Результаты анализа даже того небольшого материала, который представлен в Словаре, дают возможность две группы аббревиатур, требующих различного подхода и различных алгоритмов при автоматической обработке текста. В первую входят аббревиатуры, генетически и функционально соотнесенные с исходными словосочетаниями, составляющие с ними синонимические пары. Вторая группа включает в себя аббревиатуры, связанные со словосочетанием только генетически.

Выявление закономерностей зависимости структуры аббревиатуры от структуры словосочетания, а также закономерностей параллельного использования аббревиатуры и исходного словосочетания, их взаимозаменяемости в тексте возможно только при обследовании больших массивов текста. Это задача дальнейших исследований.

* В кибернетических текстах встречаются и другие акронимы: АРИР, ПАРИС, ТАИР, АМУР, АЛИСА.

Табл. I.

Распределение терминов по количеству слов

Количество слов	Количество терминов	Количество словоупотреблений	Количество слов	Количество терминов	Количество словоупотреблений
I	33I	33I	6	8	48
2	I035	2070	7	2	I4
3	460	I380	8	I	8
4	I2I	484			
5	47	235	Всего	2005	4570

Табл. 2.

Распределение аббревиатур по структурным типам

№№	Структурный тип	Количество аббревиатур	
		русские	заимствованные
I	Сочетание начальной части слова с целым словом	2	
2	Слоговые	I	2
3	Инициальные	57	7
	в том числе:		
	двухбуквенные	I3	I
	3-х -"-	28	2
	4-х -"-	I4	4
	5-ти -"-	I	
	6-ти -"-	I	
4	Смешанный (слог и начальная буква)	2	I2
	Всего	62	2I

Л И Т Е Р А Т У Р А

Словарь по кибернетике. - Киев : Главная редакция Украинской советской энциклопедии, 1979.

Алексеев Д.И. Аббревиатуры как новый тип слов. - В кн.:Развитие словообразования современного русского языка.М.: Наука, 1966, с. 13-37.

Борисов В.В. Аббревиация и акронимия. Военные и научно-технические сокращения в иностранных языках. - М.: Воениздат, 1972.

Брагина А.А. Синонимический ряд: словосочетание - слово. - В кн. Новые слова и словари новых слов. Л.:Наука,1978, с. 81-94.

Виноградов С.И. Аббревиатуры как варианты обозначения в русском литературном языке 20-х - начала 30-х годов. - В кн. Литературная норма и вариантность. М.: Наука,1981, с. 148-181.

Лейчик В.М. Оптимальная длина и оптимальная структура термина. - Вопр. языкознания, 1981, № 2, с. 67-73.

Словообразование современного литературного русского языка. Русский язык и советское общество. - М.: Наука, 1968.

Сердюк М.Г. Абревіатури і співвідносні з ними слова та словосполучення. - Українське мовознавство, 1981. вып. 9, с. 69-76.

Уханов Г.П. Об отношении сложносокращенных слов к словосочетаниям с той же предметной отнесенностью. - Филолог. науки, 1962, № 1, с. 187-196.

ABBREVIATIONS IN THE TERMINOLOGY OF CYBERNETICS

/BASED ON THE MATERIALS OF THE DICTIONARY OF CYBERNETICS/

Tatiana Puzdyreva

S u m m a r y

The paper describes relations between abbreviations and their initial word-combinations used in the dictionary of cybernetics. Some difficulties in deciphering abbreviations for the needs of the automatic text analysis are discussed.

СТРУКТУРИРОВАННОСТЬ ДИАЛОГИЧЕСКОГО ТЕКСТА И ЕЕ ИЗМЕРЕНИЕ

Д.И. Сливняк

В этой работе рассматривается два вопроса, тесно связанные между собой. В самых общих чертах они сводятся к следующему.

Межфразовые связи в диалоге естественно делятся на диалогические - между репликами и монологические - внутри реплик. Первый вопрос - какой вид связей носит более обязательный характер.

В пределах одного предложения различные элементы его структуры в разной степени определяются контекстом. Второй вопрос - какие из этих элементов более тесно связаны с контекстом, а какие относительно независимы от него.

Разумеется, постановка обоих вопросов нуждается в уточнении, что и будет сделано ниже. Однако уже здесь выдвинем следующее предварительное соображение, которое послужит отправным пунктом при рассмотрении первого вопроса: диалогические связи, являющиеся плодом "коллективного творчества", должны быть подчинены более строгим законам, чем монологические, для которых наличие слушающего учитывается лишь косвенно*.

Условимся понимать под структурированностью степень обязательности межфразовых связей для тех или иных позиций в тексте. Тогда выдвинутое предположение можно выразить так: в диалогическом тексте позиции, примыкающие к границе реплики, структурированы сильнее, чем позиции внутри реплики.

В работе используется статистический подход к исследованию диалога, развиваемый в статьях (Сливняк, 1977, 1979, 1981). В основе его лежит понятие типа предложения. Выделяется некоторый набор категорий парадигмы предложения. Ниже

* К подобным выводам мы уже приходили в работе (Сливняк, 1981). Может сказаться полезной также следующая аналогия: фольклорные тексты, имеющие неограниченное множество "соавторов", строятся обычно по более строгим схемам, чем произведения индивидуального творчества.

используется три таких категории: лицо, время и целевая установка (оппозиция "утверждение-вопрос"). Фиксация одной или более из этого набора категорий выделяет некоторую совокупность предложений, называемую типом (например, совокупность вопросительных предложений). Исследуются статистические связи - как синтагматические, так и парадигматические, - существующие между типами предложения.

В рамках этого подхода поставленные выше вопросы сводятся к следующему: 1) в какой позиции (диалогической или монологической) выбор типа предложения подчинен более строгим законам; 2) какие из указанных трех категорий сильнее связаны с контекстом.

В основе всех подсчетов и выводов в данной работе лежат определяемые ниже меры λ, μ, характеризующие тесноту простейшего вида межфразовой связи - между соседними предложениями.

Рассмотрим некоторую пару типов x, y. Под цепочкой (x,y) будем понимать пару соседних предложений, из которых левое имеет тип x а правое - тип y. Пусть $p(x)$ - отношение числа предложений типа x к общему числу предложений в тексте; $p(y/x)$ - отношение числа цепочек (x,y) к общему числу цепочек вида $(x,\cdot)$, где на тип правого члена не накладывается никаких ограничений. Тогда по определению

$$\lambda(x,y)=\frac{p(y/x)}{p(y)}\ ,\ \mu(x,y)=p(y/x)-p(y). \qquad (1)$$

В случае независимости типов x, y - когда наличие предложения типа x в левом члене не влияет на шансы появления типа y в правом - величина $\lambda(x,y)$ равна 1. Неравенство $\lambda(x,y) > 1$ означает взаимное притяжение обоих типов, неравенство $\lambda(x,y) < 1$ - их взаимное отталкивание. Для меры μ аналогичную роль играет значение 0. Подробное обсуждение меры λ см. в статье (Сливняк, 1979).

Описанные ниже подсчеты проделаны в работе для обеих мер λ, μ. Одновременное рассмотрение этих мер вызвано их определенным несовершенством: пределы изменения каждой зависят от $p(x), p(y)$. Так, при $p(y) \geq p(x)$

$$0 \leq \lambda(x,y) \leq \frac{p(y)}{1-p(y)}\ .$$

Из-за этого значения λ, μ определяются не только теснотой связи между типами соседних предложений, но и распространенностью этих типов в тексте, что с точки зрения данной работы является привходящим фактором. Влияние его, однако, проявляется в мерах λ, μ по-разному, чем и оправдано их совместное рассмотрение - закономерности, выполняющиеся в терминах обеих мер, можно считать не связанными со специфическими особенностями каждой.

Казалось бы, естественно подвергнуть меру $\lambda(x,y)$ (или, что равносильно, μ) при фиксированных $p(x), p(y)$ линейному преобразованию так, чтобы нормированная мера изменялась от -I до +I, принимая значение 0 в случае независимости. Однако это существенно усложняет меру, причем, как показали расчеты, деформации, вызванные нормировкой, приводят к разрушению ряда содержательных закономерностей, выполняющихся для λ и μ .

Предположим, что для некоторого x вычислены значения $\lambda(x,y)$ и $\mu(x,y)$ при всех возможных y . Тогда перед нами - описание текстового поведения типа x , его "симпатий"и "антипатий".

Сравним такие характеристики для какой-либо пары типов a и b , то есть два ряда значений $\lambda(a,y)$ и $\lambda(b,y)$ (соответственно $\mu(a,y)$ и $\mu(b,y)$). В соответствии с принятой точкой зрения, тип y выступает здесь в роли контекста типов a и b . Выделим наиболее простые и важные для дальнейшего случаи.

I. Характеристики типов a, b практически совпадают - ряды $\lambda(a,y)$ и $\lambda(b,y)$ при изменении y ведут себя одинаково.

2. Характеристики этих типов максимально несхожи между собой - большим значениям λ в ряде $\lambda(a,y)$ отвечают малые значения в ряде $\lambda(b,y)$, и наоборот.

3. Характеристики типов a, b нейтральны друг по отношению к другу - ряд $\lambda(a,y)$ не содержит никакой информации о ряде $\lambda(b,y)$.

Случай I допускает одновременно два толкования:

- выбор между типами a и b свободен от контекста, они взаимозаменяемы*;

---

* В терминологии дескриптивной лингвистики здесь может идти речь как о свободном варьировании (взаимозаменяемости элементов без изменения смысла целого), так и о контрастной дистрибуции (взаимозаменяемости, связанной с изменением смысла). Для нас это различие несущественно.

– их дистрибуции жестко связаны, в отношении поведения в тексте типы α и β взаимно определяют друг друга.

Будем говорить, что это – случай максимальной свободы в синтагматике при максимальной связности в парадигматике.

В случае 2 дистрибуции обоих типов по-прежнему связаны жестко (хотя и с "обратным знаком"). Но и связь с контекстом здесь максимальна из-за взаимного избегания типов α, β. Таким образом, это случай максимальной связности и в парадигматике, и в синтагматике. Наконец, случай 3 характеризуется максимальной свободой (независимостью) в парадигматике. В синтагматике он занимает промежуточное положение между полной свободой от контекста и жесткой связью с ним.

Очевидно, второй из поставленных вопросов – о сравнительной силе связи с контекстом категорий лица, времени и целевой установки – требует измерения синтагматической свободы/несвободы соответствующих типов предложения. Первый вопрос – о сравнительной структурированности диалогических и монологических позиций – как будет показано, связан со степенью обязательности парадигматических связей в этих позициях.

В качестве меры сходства рядов значений $\lambda(\alpha,y)$, $\lambda(\beta,y)$ для пары типов α, β в работе применен хорошо известный в статистике коэффициент корреляции Пирсона–Бравэ $\rho_{\alpha,\beta}$ – см., например, (Головин, 1971, с. 162)*. Величина $\rho_{\alpha,\beta}$ изменяется от –1 до +1, причем значения $\rho_{\alpha,\beta} = 1$, $\rho_{\alpha,\beta} = -1$, $\rho_{\alpha,\beta} = 0$ соответствуют рассмотренным выше случаям 1, 2, 3. Таким образом, величина $\rho_{\alpha,\beta}$ выступает как мера синтагматической свободы выбора между типами α и β. С другой стороны, оба значения $\rho_{\alpha,\beta} = \pm 1$ характеризуют жесткую парадигматическую связь между типами α, β, а значение $\rho_{\alpha,\beta} = 0$ – отсутствие такой связи. Поэтому в качестве меры парадигматической связи между типами α и β выступает абсолютная величина $|\rho_{\alpha,\beta}|$. Заметим, что $\rho_{\alpha,\alpha} = 1$ и $\rho_{\beta,\alpha} = \rho_{\alpha,\beta}$. Благодаря этому при составлении таблиц можно ограничиться значениями ρ, расположенными выше главной диагонали – см. таблицы 2, 4, 5.

Введем два варианта меры λ – диалогический $/\lambda^{д}(x,y)/$ и монологический $/\lambda^{м}(x,y)/$, предназначенные для описания диа-

* Коэффициент корреляции как мера сходства между дистрибуциями слов в предложении применялся А.Я.Шайкевичем для выделения частей речи и семантических классов – см. (Шайкевич, 1976, 1980). При этом в качестве членов сравниваемых рядов использовались величины, сходные с мерами λ, μ.

логических и монологических **межфразовых связей.** Для $\lambda^g(x,y)$ в формуле (I) рассматриваются двучленные цепочки, расположенные по обе стороны от стыка реплик. При этом $p(y)$ - доля типа y среди правых членов таких цепочек; $p(y/x)$ - доля типа y среди правых членов тех из указанных цепочек, левый член которых имеет тип x . Величина $\lambda^m(x,y)$ определяется аналогично на множестве двучленных цепочек, расположенных внутри реплик.

Двум вариантам меры λ^g, λ^m отвечают и два варианта коэффициента корреляции - диалогический (ρ^g) и монологический (ρ^m). Аналогично определяются и μ^g , μ^m.

До сих пор, формируя ряд значений $\lambda(x,y)$ /или $\mu(x,y)$/ для некоторого x , мы рассматривали этот тип как фиксированный левый член пары (x,y) и для получения указанного ряда меняли правый ее член. Представляет интерес и симметричная конструкция, когда фиксируется правый член пары типов, а изменяется левый. Коэффициенты корреляции для получаемых при этом рядов λ , μ обозначим $\hat{\rho}$.

В результате для каждой меры λ , μ получаем четыре варианта коэффициента корреляции ρ : ρ^g, $\hat{\rho}^g$ (левый и правый диалогический) и ρ^m, $\hat{\rho}^m$ (левый и правый **монологический).**

Распространим эту терминологию и на позиции, связанные с соответствующими ρ . Именно, будем говорить, что предложение находится в левой диалогической позиции (ЛД-позиции), если оно является левым членом двучленной цепочки, расположенной по обе стороны от стыка реплик, и в правой диалогической позиции (ПД-позиции), если оно является правым членом такой цепочки. Аналогично определяются ЛМ- и ПМ-позиции. Напомним, что для типа предложения, находящегося в любой из этих четырех позиций, роль контекста играет тип предложения в другом члене рассматриваемой двучленной цепочки.

Перейдем к описанию эксперимента. Обработке подвергнуты 23 пьесы на пяти языках (русский, армянский, грузинский, немецкий, французский) и записи русской и армянской устной речи. Для получения устойчивых результатов весь массив разбит на 6 порций - по 4-6 тысяч предложений в каждой: устная речь, **русские** пьесы 20 в., пьесы I9 в., грузинские пьесы 20 в., немецкие и французские пьесы 20 в., пьесы I7-I8 веков. Порядок расположения материала в клетках таблиц 2, 4, 5 соответствует этому перечню.

Для каждой порции подсчитывается некоторый набор числовых характеристик - значений мер λ , μ и коэффициентов кор-

реляции ρ . Искомые закономерности, как правило, имеют вид неравенств между различными парами сравниваемых числовых характеристик. Статистически значимыми считаются неравенства, выполняющиеся на всех шести порциях для обеих мер λ , μ либо на шести порциях для одной из мер и на пяти - для другой. Последние подчеркнуты штриховой линией.

Как отмечено выше, рассматривается три категории парадигмы предложения: лицо, целевая установка и время. Категория лица принимает три значения (I, 2, 3), целевая установка - два значения (У, В). Категория времени имеет в рассматриваемых языках существенно различную структуру. Ввиду этого она подвергнута определенной унификации. Видо-временные формы, обозначающие действие, начинающееся до момента речи, объединены в "прошедшее" время, начинающееся после момента речи - в "будущее", а прочие формы - в "настоящее". Таким образом, категория времени принимает у нас три значения (П, Н, Б)*.

Уточним теперь понятие типа предложения. Под ним понимается класс предложений, задаваемый значениями трех, двух или одной категории. Примеры типов: ЗНВ, 2П, БУ, I. В данной работе рассматриваются типы, задаваемые двумя параметрами, так как фиксация всех трех параметров требует для получения устойчивых результатов существенно большего объема выборки, а задание одного параметра с точки зрения искомых закономерностей малоинформативно.

Итак, рассматривается три группы типов, задаваемых соответственно лицом и целевой установкой, временем и целевой установкой, лицом и временем. Третья группа содержит 9 типов, вторая и первая - по 6. Для каждой пары типов, входящих в одну и ту же группу, по всем шести порциям материала для обеих мер λ, μ были вычислены величины $\rho^{\vartheta}, \rho^{\mu}, \hat{\rho}^{\vartheta}$ и $\hat{\rho}^{\mu}$. Значения ρ^{ϑ} приведены в таблицах 2, 4, 5, клетки которых содержат результаты по каждой из шести порций. Для сокращения таблиц приведены только значения ρ^{ϑ} (умноженные на I00).

Значения $\rho^{\mu}, \hat{\rho}^{\mu}$ не приводятся ввиду их малой информативности, а $\hat{\rho}^{\vartheta}$ - так как они дублируют поведение ρ^{ϑ} . Кроме того, для каждой группы типов, каждой порции материала и каждой из мер λ, μ вычислены средние значения $\overline{\rho}$ величин $\rho^{\vartheta}, \rho^{\mu}$ и средние значения $\overline{|\rho|}$ их абсолютных значений $|\rho^{\vartheta}|, |\rho^{\mu}|$ (таблица I).

---

* Подробнее об используемых принципах обработки материала см. в наших статьях, указанных выше.

Перейдем к анализу полученных результатов. Начнем с вопроса об относительной структурированности диалогических и монологических позиций. Как видно из табл. I, для второй и первой групп значимо выполняется неравенство $|\rho^{\vartheta}| > |\rho^{\mu}|$. Для третьей же группы различие между $|\rho^{\vartheta}|$ и $|\rho^{\mu}|$ оказывается незначимым. Аналогичные соотношения имеют место и для величин $|\hat{\rho}^{\vartheta}|, |\hat{\rho}^{\mu}|$. Таким образом, в целом для парадигматических связей подтверждается высказанное выше предположение: в диалогических позициях они выражены сильнее, чем в монологических. Для синтагматических связей - в терминах величин ρ этот эффект проявляется значительно слабее. Более того, согласно той же таблице знак неравенства, связывающего ρ^{ϑ} и ρ^{μ}, оказывается различным для мер λ и μ. Тем самым подтверждается, что величины ρ и $|\rho|$ моделируют существенно различные аспекты межфразовых связей.

Полученный результат становится легко объяснимым, если исходить из следующего представления. Структурированность позиции тем больше, чем сильнее в ней противопоставление сходства и различия, подобия и контраста (здесь, как и ранее, речь идет о сходстве между наборами значений λ или μ, характеризующими "текстовое поведение" типов предложения). Но одновременно тем ближе для нее значения ρ к $\pm$I, то есть $|\rho|$ - к I. В этом смысле диалогические позиции оказываются сильно структурированными, в отличие от монологических, для которых и подобия, и контрасты выражены слабее.

Исключением, как уже говорилось, является группа типов "лицо-время". Причина этого, по-видимому, кроется в следующем. Как будет видно из дальнейшего, высокая структурированность диалогических позиций в группах типов "лицо-целевая установка" и "время-целевая установка" обусловлена наличием резко выраженного подобия для изменений целевой установки и контраста - для изменений лица и времени. При этом сила контраста для лица и времени примерно одинакова. Последнее обстоятельство оказывается решающим для группы "лицо-время": переходы за счет изменения как лица, так и времени близки по силе контраста, что и приводит к падению $|\rho|$.

Перейдем ко второму вопросу, поставленному в начале статьи, - о сравнительной связи с контекстом трех рассматриваемых категорий. Степень связи категории с контекстом назовем ее структурирующей силой. В соответствии с постановкой вопроса, инструментом анализа будет служить мера синтагматической свободы выбора между типами предложения - величина ρ (а не $|\rho|$, как выше).

Таблица I

мера λ					мера μ			
	$\overline{\rho^{g}}$	$\overline{\rho^{\mu}}$	$\overline{\lvert\rho^{g}\rvert}$	$\overline{\lvert\rho^{\mu}\rvert}$	$\overline{\rho^{g}}$	$\overline{\rho^{\mu}}$	$\overline{\lvert\rho^{g}\rvert}$	$\overline{\lvert\rho^{\mu}\rvert}$
лицо – ц.уст.	-0,I2	-0,I7	0,52	0,40	-0,02	-0,02	0,64	0,5I
	-0,I6	-0,I7	0,4I	0,37	-0,I0	-0,07	0,57	0,43
	-0,I8	-0,I8	0,45	0,3I	-0,I4	0,00	0,50	0,45
	-0,I4	-0,I7	0,36	0,3I	-0,I4	-0,0I	0,59	0,4I
	-0,I7	-0,I8	0,39	0,23	-0,I5	-0,05	0,54	0,37
	-0,I6	-0,I9	0,4I	0,34	-0,I3	-0,05	0,52	0,38
время– ц.уст.	-0,I9	-0,I4	0,43	0,36	-0,I2	0,00	0,4I	0,39
	-0,I9	-0,I3	0,29	0,38	-0,I4	-0,02	0,35	0,42
	-0,I8	-0,I8	0,38	0,30	-0,I2	-0,I0	0,35	0,32
	-0,I4	-0,I5	0,33	0,3I	-0,I8	-0,0I	0,47	0,32
	-0,I8	-0,I8	0,42	0,32	-0,I7	-0,02	0,40	0,29
	-0,I8	-0,I7	0,30	0,33	-0,I5	-0,02	0,44	0,42

Анализ начнем с группы типов "время – целевая установка" в ЛД-позиции, когда предложение является левым членом двучленной цепочки, расположенной по обе стороны от стыка реплик. Покажем, что структурирующая сила времени значительно больше, чем целевой установки. Действительно, согласно таблице 2 выполняются сильно выраженные неравенства

$$\rho^{g}_{ПУ,НУ} < \rho^{g}_{ПУ,ПВ} > \rho^{g}_{ПУ,БУ}, \qquad (2)$$

$$\rho^{g}_{НУ,ПУ} < \rho^{g}_{НУ,НВ} > \rho^{g}_{НУ,БУ}, \qquad (3)$$

$$\rho^{g}_{БУ,НУ} < \rho^{g}_{БУ,БВ} > \rho^{g}_{БУ,ПУ}, \qquad (4)$$

и аналогичные соотношения, которые получаются, если поменять местами параметры У и Ь. Таким образом, утверждение и вопрос при фиксированном времени находятся значительно ближе к слу-

Значения 100 ρ^3

Таблица 2

	мера λ							мера μ						
		НУ	БУ	ПВ	НВ	БВ			НУ	БУ	ПВ	НВ	БВ	
	-40	-51	52	-58	-38		-44	-42	-52	0	68	-85	14	-64	11
	-40	-34	17	-40	-22		17	-28	-28	31	34	-74	20	-28	-10
НУ	-53	-50	23	-43	-49	ПВ	-12	-42	НУ -46	- 4	49	-73	-24	ПВ -67	-23
	-54	-21	9	-57	-40		-53	-22	-78	- 8	60	-78	- 3	-20	-12
	-75	-37	37	-51	-37		-27	-42	-79	25	78	-83	- 8	-70	-24
	-74	-32	28	-18	-60		26	-15	-81	-43	72	-71	2	-49	17
		-36	-42	46	-47			-12		-22	-80	-30	-27		-41
		- 7	-49	9	-27			-12		- 7	-60	- 5	-31		-44
НУ		9	-70	36	13	НВ		-36	НУ	- 7	-79	31	- 3	НВ	4
		-56	6	89	7			-48		-42	-55	85	-27		-85
		- 7	-62	62	13			-33		-11	-85	49	6		-14
		22	-80	11	40			-20		39	-92	40	-11		-44
			-43	-13	85						- 2	-24	95		
			-49	-49	29						-27	-57	77		
БУ			-38	-22	70				БУ		- 4	-27	85		
			-62	-57	8						-36	-37	68		
			-32	-43	67						8	-54	87		
			-37	-71	9						-41	- 5	40		

чаю взаимозаменяемости, чем времена при фиксированной целевой установке. Иначе говоря, в ДД-позиции **категория** времени существенно сильнее связана с контекстом, чем целевая установка, то есть его структурирующая сила по определению больше.

Механизм этого явления можно уточнить, обратившись к рядам значений величин λ, μ, для которых вычислялись обсуждаемые коэффициенты корреляции. Указанные ряды, подсчитанные для меры λ по сводному массиву, объединяющему все шесть порций материала, приведены в табл. 3, где они выступают в качестве строк. Как видим, высокие значения $\rho^{\vartheta}_{\text{ПУ,ПВ}}, \rho^{\vartheta}_{\text{НУ,НВ}}, \rho^{\vartheta}_{\text{БУ,БВ}}$ в неравенствах (2)-(4), выражающие сходство соответственно I-й и 2-й, 3-й и 4-й, 5-й и 6-й строк таблицы, обусловлены относительно высокими значениями λ в группах клеток, обведенных жирными линиями. Последнее означает, что определяющим в относительно высокой структурирующей силе времени является сохранение его при переходе через границу реплики, то есть согласованность по этой категории предложений, расположенных по обе стороны от стыка реплик. Для меры μ результаты аналогичны.

Проделаем аналогичный анализ для категорий лицо-целевая установка. Можно предположить, что структурирующая сила целевой установки остается по-прежнему малой. И действительно, согласно табл. 4 значимо выполняются неравенства

$$\rho^{\vartheta}_{\text{3У,2У}} < \rho^{\vartheta}_{\text{3У,3В}} > \rho^{\vartheta}_{\text{3У,1У}}, \qquad (5)$$

$$\rho^{\vartheta}_{\text{2У,3У}} < \rho^{\vartheta}_{\text{2У,2В}} > \rho^{\vartheta}_{\text{2У,1У}}, \quad \rho^{\vartheta}_{\text{1У,3У}} < \rho^{\vartheta}_{\text{1У,1В}}, \qquad (6)$$

а также соотношения, полученные в результате замены местами параметров У и В. Однако неравенства

$$\rho^{\vartheta}_{\text{1У,1В}} > \rho^{\vartheta}_{\text{1У,2У}}, \quad \rho^{\vartheta}_{\text{1В,1У}} > \rho^{\vartheta}_{\text{1В,2В}}$$

уже не выполняются. Как видим, структурирующая сила целевой установки действительно в целом невелика. Вместе с тем в рамках оппозиции "I лицо-2 лицо" четкой картины получить не удается, что, скорее всего, связано с малой структурирующей

Таблица 3

Мера λ

справа / слева	ПУ	ПВ	НУ	НВ	БУ	БВ
ПУ	I,43	2,7I	0,7I	0,83	0,76	0,95
ПВ	3,03	I,I9	0,6I	0,44	0,45	0,42
НУ	0,7I	0,7I	I,05	I,3I	0,87	0,84
НВ	0,65	0,42	I,38	0,8I	0,56	0,45
БУ	0,69	0,75	0,83	0,84	2,7I	3,42
БВ	0,60	0,38	0,72	0,57	4,44	2,7I

силой противопоставления первого лица второму.

О том же говорит относительно высокое значение $\rho^{\partial}_{2У,1В}$, связывающее типы предложений, различающиеся даже по обоим параметрам. Наблюдаемая картина согласуется с распространенной в лингвистике точкой зрения, согласно которой в системе лица наиболее важно противопоставление 3-го лица объединенному (I+2)-му, внутри которого уже первое лицо противопоставляется второму - см. (Бенвенист, I974, с. 266, Якобсон, I972). Как видим, и с точки зрения структурирующей силы последняя оппозиция менее существенна.

Остается сравнить структурирующую силу лица и времени. Согласно табл. 5 значимо выполняются соотношения

$$\rho^{\partial}_{3П,3Б} < \begin{pmatrix} \rho^{\partial}_{3П,2П} \\ \rho^{\partial}_{3П,1П} \end{pmatrix} > \rho^{\partial}_{3П,3Н}, \quad \rho^{\partial}_{3Б,2Б} > \rho^{\partial}_{3Б,3П}. \qquad (7)$$

Таким образом, структурирующая сила времени несколько больше, чем лица. Из той же таблицы следуют неравенства, являющиеся дополнительным свидетельством относительно малой структурирующей силы противопоставления первого лица второму:

$$\rho^{\partial}_{1Н,3Н} < \rho^{\partial}_{1Н,2Н} > \rho^{\partial}_{3Н,2Н}, \quad \rho^{\partial}_{1Б,3Б} < \rho^{\partial}_{1Б,2Б}. \qquad (8)$$

Значения $100\,\rho^3$ Таблица 4

	мера λ							мера μ						
	2У	3У	1В	2В	3В		2В	3В	2У	3У	1В	2В	3В		2В	3В
	45	-45	12	-12	-60		54	-34	74	-34	84	59	-77		71	-85
	-36	-32	50	-42	-38		16	-38	23	-28	69	10	-75		58	-74
1У	-39	-11	48	-29	-45	1В	-40	-15	1У 14	-39	65	29	-71	1В	7	-50
	50	-71	5	34	-46		-17	-32	60	-75	4	55	-81		9	-10
	17	-48	48	-22	-44		- 5	-11	8	-22	45	-17	-47		31	-48
	-36	-42	-25	7	-14		20	-59	18	-50	11	42	-77		30	-32
		-94	77	77	-59			-42		-70	78	96	-69			-71
		-41	32	73	-49			-30		-60	72	91	-72			-55
2У		-78	-29	91	-50	2В		-57	2У	-76	13	95	-68	2В		-76
		-93	29	48	1			-38		-93	34	87	-84			-87
		-70	2	51	-78			-34		-85	57	89	-81			-62
		-52	77	67	-57			-29		-89	53	92	-48			-64
			-74	-63	33						-28	-60	- 2			
			-68	-57	19						-67	-77	30			
3У			5	-84	61				3У		-17	-83	51			
			-16	-62	10						-30	-89	85			
			-43	-65	53						-64	-86	60			
			-30	-82	17						-40	-92	51			

Значения $100\,\rho^3$ для меры λ Таблица 5

	1Н	1Б	2П	2Н	2Б	3П	3Н	3Б		2Б	3П	3Н	3Б
	2	-1	19	-30	-24	3	-18	-45		13	-27	-5	0
	24	30	-14	-33	-12	24	-37	-31		-7	-22	-1	-15
1П	23	-16	-15	-24	-13	-2	-28	-31	2Н	4	-31	-36	-30
	-15	-48	51	-30	-34	-24	-9	-34		32	-10	-45	-31
	-18	-27	7	-31	-17	8	-34	4		-6	-23	-1	-27
	-39	-14	-11	-44	-21	13	2	-27		-21	-29	5	-39
		50	20	5	-42	-47	-11	-29			-28	-24	34
		27	-9	-5	1	-48	-35	-29			-31	-16	-10
1Н		33	-26	22	-18	-22	-32	-36	2Б		-9	-13	-20
		-22	-33	80	23	-21	-22	-14			-43	-71	34
		12	-24	19	-11	-31	12	-37			-35	-6	53
		40	-22	20	-30	-43	-27	56			-22	26	-9
			53	51	80	-1	-63	44				-10	-13
			-41	-49	39	-35	-17	1				7	-12
1Б			-27	3	36	-23	-27	11	3П			-15	-25
			-13	9	24	6	-33	-11				27	-10
			-28	-31	79	-30	1	37				-32	-26
			-18	31	0	-49	-19	-22				-16	-9
				-21	-16	-9	-30	-27					-12
				11	-13	7	-36	-24					-5
2П				5	-22	-15	-29	-14	3Н				63
				-19	16	-55	-23	-32					7
				-13	-37	13	-52	-37					9
				5	-27	19	-59	2					-28

Значения $100\,\rho^3$ для меры μ

	1Н	1Б	2П	2Н	2Б	3П	3Н	3Б		2Б	3П	3Н	3Б
	25	55	68	41	43	26	-70	-2		60	8	-67	7
	53	54	33	-4	36	35	-76	10		8	-36	-33	-3
1П	35	43	0	-8	7	7	-36	-25	2Н	43	-26	-60	-64
	-31	20	66	-10	23	17	-52	-29		56	-9	-58	-18
	-6	15	39	-18	20	41	-43	22		54	-11	-57	22
	-12	-14	16	-47	4	25	-4	-11		19	-31	-58	-49
		52	28	41	19	-44	-15	10			-4	-55	46
		37	23	18	31	-30	-68	13			-8	-55	15
1Н		66	6	40	18	-28	-46	-43	2Б		-4	-52	-40

Значения $100\,\rho^{э}$ для меры μ Окончание таблицы 5

	1Н	1Б	2П	2Н	2Б	3П	3Н	3Б
1Н		-26	-41	80	27	-41	-18	- 2
		2	-15	19	22	-20	-31	42
		52	26	54	40	-33	-65	-43
			- 7	-39	7	-44	-34	-32
			4	-38	66	3	-44	34
1Б			34	33	63	- 4	-70	-34
			32	1	46	20	-54	6
			19	-22	61	4	-14	53
			36	82	32	- 6	-79	-48
				32	40	17	-63	6
				6	17	10	-58	1
2П				30	11	- 3	-52	-28
				-16	35	-16	-35	-24
				- 5	12	35	-52	7
				13	20	25	-68	-38

	2Б	3П	3Н	3Б
2Б		- 6	-84	21
		15	-68	72
		- 5	-44	-34
			-65	-26
			-12	-15
3П			-37	-22
			-28	-34
			-52	8
			-28	-34
				- 5
				-31
3Н				73
				8
				-52
				70

Здесь, в отличие от предыдущего, сопоставление производится в рамках одного параметра - лица при фиксированном втором параметре - времени. По неизвестным причинам, аналогичные соотношения для прошедшего времени не имеют места.

Итак, для ЛД-позиции имеет место следующая картина: наибольшей структурирующей силой обладает категория времени, далее идет оппозиция "1+2 лицо-3 лицо", а последнее место делят оппозиция "1 лицо-2" и целевая установка.

Эти результаты в основном сохраняются и для ЛД-позиции (то есть для $\hat{\rho}^{э}$), хотя и в несколько "размытом" виде. Подобное ослабление закономерностей, связанное с переходом от ЛД- к ЛД-позиции, уже наблюдалось нами при изучении тематических границ в диалоге) (Сливняк, 1981). Совсем размытая картина получается в монологических позициях (величины $\rho^{м}$, $\hat{\rho}^{м}$), что соответствует установленной выше их слабой структурированности. На общем хаотическом фоне выделяются относительно высокие значения $\rho^{м}_{1п,3п}$, которые, видимо, отражают беспорядочное чередование 1-го и 3-го лица в устноречевом повествовании, ведущемся в основном в прошедшем времени.

Подведем итоги. К основным результатам работы можно отнести следующие. 1) Диалогические позиции структурированы

сильнее, чем монологические, причем это проявляется через связь типов предложения в парадигматике. 2) Для диалогических позиций в синтагматическом аспекте грамматические категории расположены в порядке убывания их структурирующей силы следующим образом: время, оппозиция "I+2 лицо-3 лицо", оппозиция "I лицо-2 лицо" и целевая установка. 3) Для правой диалогической позиции получена такая же картина, как и для левой, но в более размытом виде.

Как отмечалось по ходу изложения, эти результаты согласуются с имевшимися ранее представлениями о структуре диалогического текста и в определенной мере дополняют их.

Автор приносит благодарность А.Я. Шайкевичу за внимание к работе и ценные советы.

Л И Т Е Р А Т У Р А

Бенвенист Э. Общая лингвистика. – М.: Прогресс, 1974.

Головин Б.Н. Язык и статистика. – М.: Просвещение, 1971.

Сливняк Д.И. О количественной связи двух характеристик предложения. – Вестник общественных наук АН Армянской ССР, 1977, № 1, с. 80-87.

Сливняк Д.И. Об одном способе статистического анализа диалога. – Вестник общественных наук АН Армянской ССР, 1979, № 7, с. 58-66.

Сливняк Д.И. Статистические характеристики межсегментных границ в диалоге. – В кн.: Актуальные проблемы квантитативной лингвистики и автоматического анализа текстов. Труды по лингвостатистике УII. Тарту, 1981, с.120-135 (Учен. зап. Тартуского ун-та, вып. 591).

Шайкевич А.Я. Выделение классов слов и парадигм посредством дистрибутивно-статистического метода.-Прикладная лингвистика, вып. 18 (Труды МГПИИЯ), 1976.

Шайкевич А.Я. Гипотезы о естественных классах и возможность количественной таксономии в лингвистике. – В кн.: Гипотеза в современной лингвистике. – М.: Наука, 1980,с.319–357.

Якобсон Р.О. Шифтеры, глагольные категории и русский глагол. – В кн.: Принципы типологического анализа языков различного строя. М., 1972.

THE STRUCTURATEDNESS OF THE DIALOGICAL TEXT AND ITS MEASUREMENT

Dmitrii Slivnyak

S u m m a r y

The subject-matter of the article is the obligatoriness degree of inter-phrase links in different positions of the dialogical text, called structuratedness of the position. A statistical procedure for measuring structuratedness is proposed. Another characteristic measured is the relative context-dependence degree of a sentence grammar category (called its structurating force). It is established that the positions situated near the queue boundaries are more structurated than those in the interior of the queues. A hierarchy of sentence grammar categories according to their structurating force is obtained.

ЭМПИРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ЧАСТОТНОСТИ ФОНЕМ В ОРОЧСКОМ ЯЗЫКЕ

Ю.А.Тамбовцев

В настоящее время ведется интенсивное исследование языков Сибири и Дальнего Востока фоностатистическими методами. Мансийский (Тамбовцев, 1977, 1979, 1980, 1981), хантыйский (Тамбовцев, в печати), кетский (Вернер, Тамбовцев, 1979), хакасский (Тамбовцев, в печати) уже были обработаны при помощи методов фоностатистики. Целью данной статьи является анализ эмпирического распределения фонем в орочском языке.

Орочский язык относится к тунгусо-маньчжурским языкам, а точнее к южной подгруппе этой языковой общности. Эта небольшая народность проживает в Советском Союзе двумя неравными изолированными группами. В данной работе рассматривается язык бо́льшей группы, проживающей в районе Советской Гавани (по рекам, впадающим в Татарский пролив).

Материалом для фоностатистической обработки послужили орочские тексты и словарь, записанные В.А.Авроpиным и Е.П.Лебедевой во время экспедиции 1959 года в район Советской Гавани (Аврорин, Лебедева, 1978). Тексты на орочском языке содержат различного рода устные рассказы о прошлой жизни, охоте и рыболовстве, религиозных взглядах и суевериях и т.д.

Статистическая обработка текста, объем которого составил 123 761 фонемы, производилась на ВЦ НГУ при помощи ЭВМ ЕС-1033. Выборка такой величины представляется достаточно надежной (Тамбовцев, Утев 1981; Ludvíková, Königové, 1967).

В орочском языке, также как и предыдущие исследователи (Аврорин, Лебедева 1958, Аврорин 1978), мы выявляем 28 гласных и 19 согласных фонем. Гласные: и, æ, э, о, а, у, ў, й, э̄, а̄, ӯ, ў̄, иэ, иа, иу, æа, æу, эи, эу, аи, уи, уэ, ўи, ўа, ои, о̄, а̄ў.

Согласные: п, б,в,м,т,д,с,н,л,р,ч,j, ʒ ,нʲ к,г,х,ɣ,ŋ .

Фоностатистически обработав данные тексты получаем следующие частотности фонем, которые представим в виде таблицы (таб.I). Цифровые данные этой таблицы показывают, что тремя самыми частотными фонемами являются гласные: и (I4,I0%), а (8,98%) и э (7,I6%). Отметим, что по данным Они ши в японском языке самыми распространенными также являются а(I5,96%), о(I2,00%), и(I0,59%), е(5,76%) (Onishi, 1937).

Выделим медиану частотного функционирования фонем в орочской речи. Для того, чтобы найти медиану, мы должны разделить I00% на количество фонем в языке, в данном случае на 46. Следовательно, медиана равна 2,I7%. Практически медианой в системе частотного функционирования орочских фонем является фонема (ŋ) с частотностью 2,I0%. Теперь весь упорядоченный ряд фонем разобъем на три подмножества. В верхнее подмножество войдут самые характерные для орочской речи фонемы, в нижнее подмножество войдут довольно редко встречающиеся и, следовательно, менее характерные для орочской речи фонемы, а в среднее подмножество войдут соседние с медианой (ŋ) фонемы, т.е. (j) и (л). Все фонемы, частотность которых больше, чем 2,40% войдут в верхнее подмножество, а фонемы с частотностью менее 2,00% войдут в нижнее подмножество.

Рассмотрим отношение, которое назовем коэффициентом медиальности. Вычислим его следующим образом. В инвентаре данного языка 46 фонем, следовательно медиальными фонемами упорядоченного ряда являются фонемы под номерами 23 и 24 (ẏ , ӯ) т.е. фонемы, со средней арифметической частотностью употребления в потоке речи I,265%. Как мы видим она не совпадает со средней теоретической частотностью фонемы, которая составила 2,I.%. Коэффициентом медиальности будет отношение частотности медиальной фонемы в упорядоченном ряду к средней теоретической частотности фонемного функционирования в данном языке. Здесь этот коэффициент равен I,26: 2,I7 = 0,58I. Покажем, чему равен этот коэффициент в некоторых других языках. Для большего контраста возьмем языки разных семей и расположим их в соответствии с этим коэффициентом в порядке убывания:

Английский (британский)	(Denes, 1963)	I,056
Английский (американский)	(Voelker, 1934)	0,998

Мансийский (по нашим данным)	0,876
Коми-зырянский (по нашим данным)	0,812
Кетский (Вернер, Тамбовцев, 1979)	0,801
Польский (Сегал, 1972)	0,782
Русский (Елкина, Одина, 1964)	0,687
Чешский (Ludvíková, Königová, 1967)	0,624
Украинский (Перебейнос, 1964)	0,618
Орочский (по нашим данным)	0,583
Японский (Onishi, 1937)	0,316

Хорошо известно, какое большое значение уделяется величине отношение в языке согласных и гласных. Ещё в прошлом веке известный лингвист Б.Бурдон предложил разделить все языки в соответствии с величиной содержания в них инвентаре гласных и согласных на три типа: вокалические, консонантные и смешанные (Bourdon, 1892, p.64). Но такая классификация не отражала частотное функционирование фонем в потоке речи, поэтому совершенно правомерным представляется дополнение, сделанное чешским фонологом Й.Крамски (Krámský , 1948, 1959, 1978).

Й.Крамски ввел новый критерий, учитывающий величину частотности классов гласных и согласных фонем в инвентаре и речи. Вслед за Й.Крамским рассмотрим, какие классы согласных превалируют в речи по сравнению с инвентарем. В качестве классов согласных возьмём классификацию согласных в зависимости от места артикуляции (т.е. активного органа, который вовлекается в производство данного звука). Выделим следующие классы: губные, переднеязычные, среднеязычные и заднеязычные. Й.Крамски выделяет 4 типа языков: 1) языки с повышенным употреблением лабиальных и альвеолярных; 2) языки с повышенным употреблением альвеолярных; 3) языки с повышенным употреблением альвеолярных и палатальных; 4) языки с повышенным употреблением альвеолярных и велярных (Крамски, 1959, p.84). Нами выделяются также 4 класса: 1) языки с повышенным употреблением в тексте по сравнению с употреблением в инвентаре губных; 2) языки с повышенным употреблением переднеязычных; 3) языки с повышенным употреблением среднеязычных; 4) языки с повышенным употреблением заднеязычных. После выявления количественных величин употребления указанных классов согласных в орочском языке,

сравним эти величины в других языках, на которых говорят в Сибири и родственных им языках. Это прежде всего довольно близко территориально расположенный якутский язык, также такие языки Сибири как мансийский, хантыйский, селькупский и родственные им венгерский, коми-зырянский и карельский. Следует отметить объем выборок, на которых были получены данные по этим языкам, так как из статистики хорошо известно, что о значимых результатах можно говорить только тогда, когда выборка взята достаточно большой (Тамбовцев, 1980, Тамбовцев, Утев, 1981). В данном случае языки имеют следующие выборки: мансийский - 276 418 фонем, коми-зырянский - 80 168 фонем, хантыйский 74 762 фонемы (по данным автора), венгерский 551 828 фонем (Jékel, Papp, 1974), селькупский - 10 000 (Морев, 1973), карельский - 62 360 фонем (Баранцев, 1975), якутский - 236 245. Кроме указанных языков представляется интересным сравнить орочский язык также и с японским языком, который, как известно, ещё не получил достаточное обоснование своих родственных связей (генетических связей) с другими языками мира (Dörfer, 1961; Chew, 1981).

На основании фоностатистических данных построим таблицы и рисунок, в которых покажем численные характеристики функционирования в инвентаре и в речи выделенных четырех классов согласных (Таблицы 1,2,3,4 и рисунок 1).

Применяя критерий Й.Крамски находим, что к первому классу языков, т.е. к языкам с повышенным употреблением в тексте губных согласных относятся из данных языков хантыйский, мансийский, селькупский, в то время как такие языки, как венгерский, коми-зырянский и орочский употребляют в речи такой же процент губных согласных, как и в словаре. К языкам с пониженным употреблением губных относятся карельский, якутский и японский. Ко второму классу языков, т.е. к языкам с повышенным употреблением переднеязычных согласных относится большинство данных языков: мансийский, якутский, селькупский, карельский, коми-зырянский и хантыйский. Орочский также больше тяготеет к этой группе языков, хотя стоит между ними и языками, где процент употребления переднеязычных в тексте и инвентаре совпадает: венгерским и японским. В третий класс языков из данных не выходит ни один. Даже орочский язык не вошёл в этот класс, хотя из обсуждаемых языков он более всех приближен к нему. В орочском языке опять наблюда-

ется равновесие в динамике употребления среднеязычных в речи и в инвентаре. У самого близкого в этом отношении орочскому языку японского языка, это соотношение несколько ниже (0,7), что показывает, что в японской речи среднеязычные в объеме занимают меньший процент, чем в инвентаре. Для того, чтобы отметить, что величина этого отношения, близкая к единице, показывает на высокую степень палатализации речи, приведем величину этого отношения в таком языке с широким употреблением палатальных согласных как русский. В русском языке это отношение равно 0,68 (частотность фонем взята из работы по фоностатистике русского языка (Ёлкина, Юдина, 1964). К четвертому классу, т.е. к классу с повышенным употреблением заднеязычных согласных относятся только два из данных языков: карельский и венгерский. Японский с коэффициентом отношения 1,1 тоже входит в эту группу, но стоит как бы на границе. Опять очень интересным образом проявляет себя орочский язык: объём заднеязычных согласных в инвентаре и в речи равен. Это очень интересное явление из всех девяти языков в большей степени присуще только одной орочскому языку. Особенно ясно это видно по графику: по всем четырем параметрам соотношение согласных равно или очень близко единице. В этом отношении подобная тенденция прослеживается в японском языке. В других приводимых языках объемы, занимаемые данным классом фонем в инвентаре и в речи сильно различаются.

Из всего сказанного сделаем следующие выводы:

1. С помощью фоностатистических методов было проанализировано девять языков. Шесть из них (мансийский, хантыйский, селькупский, орочский, якутский и японский) принадлежит к языкам Сибири и Дальнего Востока. Остальные три - венгерский, коми-зырянский, карельский - являются родственными мансийскому, хантыйскому и селькупскому. Данные по мансийскому, хантыйскому, якутскому и орочскому языкам получены автором. В данной статье основное внимание уделяется анализу фоностатистических данных по орочскому языку, которые получены с помощью ЭВМ и на сравнительно большой выборке впервые.

2. Объём орочских текстов составил 123 761 фонему. Полученные данные по частотности фонем представлены в таблицах (Табл. 1,2,3,4) и на рисунке (Рис.1), из которых следует, что соотношение гласных и согласных в орочском языке равно 1:1,7, а точ-

нее на 1000 гласных приходится 1680 согласных, если исходить из данных инвентаря. В потоке речи соотношение гласных и согласных совсем другое, оно равно 1:1 или точнее на 1000 гласных приходится 1040 согласных. Орочский язык можно назвать очень мелодичным языком, так как сумма гласных и сонорных в нем составляет 64,1%, что является довольно высокой величиной по сравнению с другими языками (выше, чем в русском и чешском, но ниже, чем в мансийском и хантыйском, величины которых соответственно равны 61,3%; 53,4%;73,7%; 71,3%). Общая сумма частотностей долгих гласных составила 8,62%, а кратких 40,28%, следовательно, долгие и краткие соотносятся как 1:5. Глаеные заднего, переднего и среднего ряда относятся как 1:1,6 :2,1. Глаеные нижнего, среднего и верхнего подъёма относятся как 1:1,2:1,7. Губные, среднеязычные, заднеязычные и переднеязычные согласные соотносятся как 1:1,2: 1,6, а шумные щелевые, сонорные и шумные смычные как 1:1,5:2,4.

3. В статье предлагается анализировать различные языки по коэффициенту медиальности. Одиннадцать языков различных семей располагаются в упорядоченный ряд. Интереено отметить, что языки генетически родственные, по этому коэффициенту в упорядоченном ряду занимают соседние места.

4. Учитывая отношения величин частотностей фонем в инвентаре и в речи выделяются четые тина языков (классы согласных в соответствии с местом их артикуляции): 1) языки с повышенным употреблением губных; 2) языки с повышенным употреблением переднеязычных; 3) языки с повышенным употреблением среднеязычных; 4) языки с повышенным употреблением заднеязычных. Очень интересным образом проявляет себя орочский язык: по всем четырём параметрам он имеет коэффициент близкий к единице, таким образом его нельзя причислить ни к одному из выделенных классов. Следовательно, необходимо выделить особый тип языков, у которых распределение объёмов этих четырёх классов согласных в инвентаре и в речи примерно одинаково.

17

Таблица I.

Абсолютная частота орочских фонем ($\sum$ - I23 76I)

№	фонема	%	№	фонема	%	№	фонема	%	№	фонема	%
I.	и	I4,I0	I3.	б	2,90	25.	п	I,2I	37.	эу	0,I4
2.	а	8,98	I4.	у	2,75	26.	ō	I,I7	38.	уи	0,I4
3.	э	7,I6	I5.	в	2,66	27.	ӣ	I,I0	39.	ɣ̆	0,II
4.	т	5,55	I6.	с	2,5I	28.	ē	I,I0	40.	æу	0,07
5.	н'	4,I4	I7.	н	2,44	29.	эи	I,03	4I.	ӯа	0,05
6.	к	4,03	I8.	ӡ	2,40	30.	æ	0,84	42.	ӯи	0,04
7.	дь	4,00	I9.	л	2,40	3I.	ӯ	0,68	43.	иэ	0,03
8.	м	3,70	20.	ң	2,I0	32.	аи	0,52	44.	уэ	0,03
9.	г	3,40	2I.	j	2,00	33.	р	0,40	45.	иу	0,02
I0.	х	3,I7	22.	ҷ	I,97	34.	иа	0,I9	46.	æа	0,0I
II.	о	2,97	23.	ӱ	I,27	35.	ои	0,I7	$\sum$	= I00,00	
I2.	ā	2,97	24.	ӯ	I,26	36.	аӱ	0,I5			

Таблица 2.

Группы согласных в инвентаре.

№		Губные	Передне-язычные	Средне-язычные	Задне-язычные
I.	Орочский	2I,0	3I,6	2I,0	26,3
2.	Японский	I9,0	57,I	4,8	I9,0
3.	Кетский	2I,7	34,8	2I,6	I7,4
4.	Якутский	I5,0	30,0	25,0	30,0
5.	Казахский	20,0	48,0	4,0	28,0
6.	Узбекский	20,8	45,8	4,2	29,2
7.	Туркменский	23,8	52,4	4,8	I9,0
8.	Мансийский	I7,6	29,4	29,4	23,5
9.	Хантыйский	I6,7	38,9	27,8	I6,7
I0.	Венгерский	20,8	50,0	I6,7	I2,5
II.	Коми-зырянский	I7,2	4I,4	3I,0	I0,3
I2.	Селькупский	I9,6	28,3	28,3	23,9

Таблица 3.

Группы согласных в речи, в %.

№	Язык	Губные	Передне-язычные	Средне-язычные	Задне-язычные
I.	Орочский	20,5	33,9	20,6	25,0
2.	Японский	I6,8	58,8	3,2	2I,2
3.	Кетский	I4,8	49,I	II,8	24,2
4.	Якутский	I0,6	57,I	9,I	23,I
5.	Казахский	I4,8	60,6	5,3	I9,3
6.	Узбекский	I6,2	59,I	4,9	I9,8
7.	Туркменский	I4,7	59,5	7,I	I8,8
8.	Мансийский	22,2	49,3	II,I	I7,4
9.	Хантыйский	2I,I	5I,7	I2,8	I4,4
I0.	Венгерский	I8,4	57,3	7,0	I7,2
II.	Коми-зырянский	I7,5	52,9	I9,4	I0,2
I2.	Селькупский	22,0	5I,6	7,6	I8,8

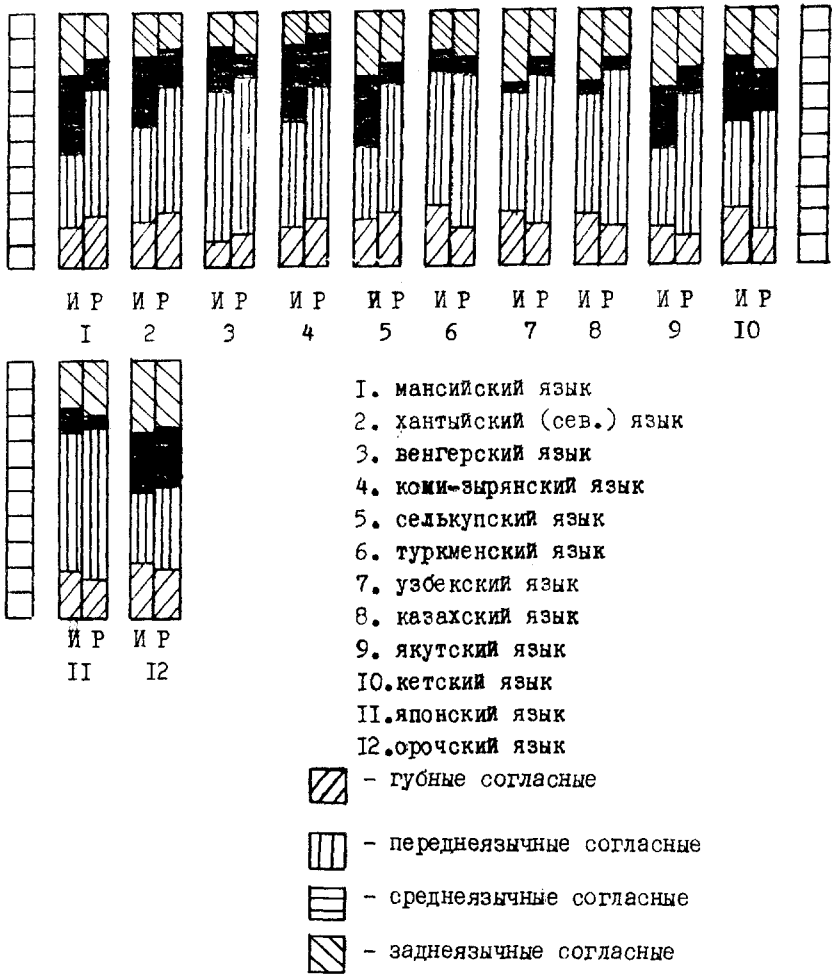
Таблица 4.

Отношение и разность между группами согласных в речи и в инвентаре ($P_р$ – суммарная частотность согласных определенной группы в речи в %, $P_и$ – суммарный процент этой же группы согласных в инвентаре, $R = P_р/P_и$, $D = P_р - P_и$)

Группы согласных	Губные		Передне-язычные		Средне-язычные		Задне-язычные	
№ Язык	R	D	R	D	R	D	R	D
I. Орочский	I,0	-0,6	I,I	+2,3	I,0	-0,5	I,0	-I,2
2. Японский	0,9	-2,3	I,0	+I,7	0,7	-I,6	I,I	+2,2
3. Кетский	0,7	-6,9	I,4	+I4,3	0,4	-I4,3	I,4	+6,9
4. Якутский	0,7	-4,4	I,9	+27,I	0,4	-I5,9	0,8	-6,7
5. Казахский	0,7	-5,2	I,3	+I2,6	I,3	+I,3	0,7	-8,7
6. Узбекский	0,8	-4,6	I,3	+I3,3	I,2	+0,7	0,7	-9,4
7. Туркменский	0,6	-9,I	I,I	+7,I	I,5	+2,3	0,9	-0,2
8. Мансийский	I,2	+4,5	2,5	+I9,9	0,4	-I8,3	0,7	-6,I
9. Хантыйский	I,3	+4,4	I,3	+I2,8	0,4	-I5,0	0,9	-2,2
I0. Венгерский	I,0	+0,4	I,0	+4,5	0,4	-9,7	I,4	+4,7
II. Коми-зырянский	I,0	+0,3	I,3	+II,5	0,6	-II,6	I,0	-0,2
I2. Селькупский	I,I	+2,5	I,8	+23,3	0,3	-20,7	0,8	-5,I

Рисунок I

Графическое представление соотношения суммарной частотности согласных, выделенных по активному органу (месту образования), в инвентаре (И) и в речи (Р)



И- в инвентаре; Р - в речи

ЛИТЕРАТУРА

Аврорин В.А. Фонетика орочского языка. – В кн.: Изучение языков Сибири. Новосибирск, 1978, с.3-52.

Аврорин В.А., Лебедева Е.П. Орочский язык. – В кн.: Языки народов СССР, том 5, Монгольские, тунгусо-маньчжурские и палеоазиатские языки. Л., Наука, 1968, с.191-209.

Аврорин В.А., Лебедева Е.П. Орочские тексты и словарь. Л.: Наука, 1978, с.264.

Вернер Г.К., Тамбовцев Ю.А. Некоторые результаты фоностатистического анализа звукового состава югского языка. – В кн.: Теоретические вопросы фонетики и грамматики языков народов СССР. Новосибирск, 1979, с.47-52.

Баранцев А.П. Фонологические средства людиковской речи. Л.: Наука, 1975, с.178-182.

Ёлкина В.Н., Юдина Л.С. Статистика слогов русской речи. – Вычислительные системы. вып.10. Новосибирск, 1964,с.61.

Морев Ю.А. Звуковой строй среднеобского (ласкинского) говора селькупского языка. АКД, Томск, 1973, с.4-12.

Перебейнос В.И. Частота и сочетаемость фонем современного украинского языка. Киев, 1965, с.25-30.

Тамбовцев Ю.А. Некоторые характеристики распределения фонем мансийского языка. – Советское финно-угроведение, XIII, № 3, 1977, с.195-198.

Тамбовцев Ю.А. Распределение гласных фонем в мансийской поэзии. Советское финно-угроведение, XV, № 3, 1979, с.164-167.

Тамбовцев Ю.А. Закономерности частотного функционирования долгих и кратких гласных в ударных и неударных слогах мансийского слова. – Советское финно-угроведение, XVII, № 2, 1981, с.105-109.

Тамбовцев Ю.А. Некоторые закономерности распределения фонем в хакасском языке. В печати.

Тамбовцев Ю.А. Эмпирическое распределение фонем в казымском диалекте хантыйского языка. В печати.

Тамбовцев Ю.А., Утев С.А. Зависимость величины частот мансийских гласных I-го слога от величины объема выборки.– В кн.: Теоретические вопросы фонетики и грамматики языков народов СССР. Новосибирск, 1979, с.97-105.

Bourdon B. L'expression des emotions et des tendences dans le language. Paris,I892.

Chew J.J. The Relationship between Japanese,Korean,and the Altaic Languages.In what Sence Genetic.-In:The Bulletin of the International Institute for Linguistic Sciences.Kyoto Sangyo University.Vol. II,4,I98I,p.7-38.

Denes P.B. On the Statistics of Spoken English.-In:Journal Acoustic Soc.Amer. 30,1963,p.892-904.

Doerfer G. The Conditions for Proving the Genetic Relationship of Languages.-In:The Bulletin of the Inter.Inst. for Linguistic Sciences. Kyoto Sangyo University.Vol.II, 4.1981,p.39-58.

Jékel P.,Papp F. Ady Endre Őszes Költői Műveinek Fonémastatisztikája.-Budapest,1974,p.3.

Krámský J. Fonologické využitísamohlàskových fonemat.-In:Linguistica Slovaca.IV-VI,1946-1948,N° 39.

"------" A Quantitative Typology of Languages.-In:Language and Speech. Vol.2,part 2,1959,p.72-85.

"------" Quantitative Analysis of Near-identical Phonological Systems.-In:Prague Studies in Mathematical Linguistics,6.Prague,1978,p.9-38.

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

Onishi M. Articulation Bases of Ten Major Languages. A Study from the Viewpoint of Frequency of Occurrence of Sounds.-In:Study of Sounds.Vol.6,1937,p.153-181.

Tambovcev Y.A. Zur Vorkommenshaufigkeit Langer und Kurzer Vokale in der Zweiten Silbe des Vogulischen.-In: Finnisch-Ugrische Mitteilungen,4,1980,p.73-74.

Voelker C.H. Count of Phonic Frequency in Formal American Speech. -In:Journ.Accoust.Soc.Amer. 5,1934,p.242.

EMPIRICAL DISTRIBUTION OF THE PHONEMES IN THE OROCH LANGUAGE

Yuri Tambovtsev

S u m m a r y

For the last few years, some Siberian languages have been intensively studied by the methods of phonostatistics. The article presents the results of computing and an analysis of the empirical distribution of the phonemes in Oroch, one of the southern groups of the Tungus-Manchurian languages. The data obtained are compared with the phonostatistical results for eleven other languages of Siberia, the Far East and the North. The article considers three sub-series of the Oroch series of phonemes. It has been found that Oroch is close to Japanese in some respects which may help to more solidly establish the relations of the Tungus-Manchurian languages and Japanese.

О ТАК НАЗЫВАЕМОМ ДЕЛОВОМ, НЕХУДОЖЕСТВЕННОМ СТИЛЕ С КВАНТИТАТИВНОЙ ТОЧКИ ЗРЕНИЯ

Марие Тешителова

(Прага)

Понятие функционального стиля, или же функциональное использование языковых средств, берет свое начало в чехословацкой лингвистике от тезисов Пражской лингвистической школы, заложенной в 1929 г. (These, 1929). В дальнейшем отмечается, между прочим, расширение количества отдельных функциональных стилей. Это, конечно, имеет связь с развитием литературного языка и с новыми задачами, которых требует от языка развитие нашего общества. В настоящей статье рассматривается дифференциация главных функциональных стилей в современном литературном чешском языке с квантитативной точки зрения. В течение развития литературного чешского языка установились наряду с функциональным стилем разговорного языка и художественной литературы следующие стили: профессионально-технический, или же научный (Б. Гавранек выделяет его в своих трудах как "рабочий" и научный; Havránek, 1963), публицистический (самостоятельное его существование долгое время служило объектом дискуссий) и относительно "наиболее молодой" административный стиль (Mistrík, 1970, 151), стиль законов, объявлений, циркуляров, отчетов и т.п. В развитом социалистическом обществе перечисленные функциональные стили имеют основное значение. Кроме того, можно отметить, что развитие научного и технического прогресса, публичные отношения в административе, в юридической сфере и т.д. содействуют стабилизации специфических черт соответствующих функциональных стилей.

В настоящей статье делается попытка посредством статистических данных подтвердить существование так называемого делового, нехудожественного стиля в рамках его составных частей, стиля научного, публицистического и административного, противопоставив стилю художественной литературы, который, как правило, включает стиль художественной прозы, поэзии и драмы. Художественному стилю будет уделяться внимание только тогда, когда это потребуется для определения нехудо-

жественного стиля. Статистические данные, на которые мы опираемся в настоящей статье, взяты как из чешского частотного словаря (сокр. ЧСЧ, Jelínek - Bečka - Těšitelová, 1961), так и из корпуса 540 000 слов (180 000 слов из публицистических, 60 000 слов из административных, 300 000 слов из научных текстов, в том числе 75 % из письменных работ и 25 % из устных выступлений), который был собран под руководством автора данной статьи в отделении математической лингвистики и фонетики Института чешского языка Чехословацкой Академии наук в Праге в 1972-1980 гг. Научные работники и специалисты данного отделения подготовили для обработки на вычислительной машине лексический и грамматический (морфологический и синтаксический) анализ всего корпуса и обсудили результаты, полученные с помощью вычислительной машины (ср. напр. Тешителова и кол., Частотный словарь современной чешской публицистики, 1980, сокр. ЧСП, Linguistica II, Частотный словарь современной чешской административы, 1980, сокр. ЧСА, PSML 8). Весь материал, внесенный с помощью перфокарт на магнитную ленту вычислительной машины, разрабатывается дальше в вычислительных центрах ЧАН в Праге, особенно на вычислительной машине ТЕСЛА 200.

Для установления черт нехудожественного стиля и его составных частей изучалась частота некоторых (I.) лексических и (2.) грамматических средств.

(I.) Лексические средства

Общеизвестно, что к самым частотным лексическим средствам в языках, в которых отсутствует артикль, принадлежат союзы, предлоги, местоимения, глаголы-связки, включая модальные глаголы, т.е. слова грамматические, формальные. Эти слова применяются в отдельных функциональных стилях дифференцированно (Těšitelová, 1974). Об индивидуальном применении данных средств можно говорить даже в нехудожественном стиле и в его составных частях.

I.I. Начнем, например, с самого частого слова в чешском языке - с союза a (и). Этот союз выступает как самое частое слово в публицистическом, административном и научном стилях, т.е. занимает первое место в нехудожественном стиле. Частота данного союза обусловлена письменной формой большинства анализированных текстов (ср. ЧСЧ). В устных текстах, напр., публицистических, союз a является самым частым союзом, но не самым частым словом. Таким словом служит место-

имение ten - указательное и анафорическое (ср. ниже).

Явные различия в распределении частот в отдельных анализированных функциональных стилях мы замечаем при союзе že (что). В нехудожественном стиле он является однозначно вторым наиболее частым союзом и используется для усложнения структуры предложения. В административном стиле союз že выразительно менее частый; в порядке слов по убывающей частоте мы находим его между двадцатым и тридцатым словом. Из этого вытекает, что структура предложения в административном стиле намного проще.

Третий самый частый союз в нехудожественном стиле - союз i (и, даже). Третье место он занимает как в публицистическом (где находим его между первым и десятым словом), так и в научном стиле. В административном стиле союз i является вторым наиболее частым союзом (он предшествует союзу že - ср. выше). О союзе i, который соединяет слова, тесно связанные по своему значению, можно сказать что он типичен для нехудожественного стиля; этот факт подтверждает также сравнение с ЧСЧ (в текстах нехудожественного стиля его частота 56,85 %, в текстах художественного стиля - 43,15 %).

I.2. То же, самое можно сказать о предлоге v/e (в), который является самым частым предлогом в публицистическом, административном и научном стилях; предлог v можно классифицировать как наиболее частый предлог в нехудожественном стиле. В ЧСЧ его частота достигает 50,56 % в нехудожественном стиле и 49,44 % в стиле художественной литературы. Можно отметить, что аналогичная, однако, обратная пропорция наблюдается при предлоге na (на). По данным ЧСЧ, предлог na в художественном стиле встречается гораздо чаще (66,05 %), чем в нехудожественном (33,95 %). Предлог na, как можно предполагать, выступает вторым наиболее частым предлогом в нехудожественном стиле; в публицистике и в административе он встречается относительно чаще, чем в научном стиле. Относительно более высокая частота предлога z/e (из) по сравнению с предлогом s/e (с) характеризует разницу между нехудожественным и художественным стилями. Хотя оба предлога, по данным ЧСЧ, мы находим довольно часто в художественном стиле (s/e 68,66 % в художественном стиле, 31,34 % в нехудожественном стиле; z/e 61,11 % в художественном стиле, 38,89% в нехудожественном стиле), предлог z встречается в нехудожественном стиле относительно чаще, чем предлог s ; это имеет связь с высокой частотой родительного падежа у существительных (ср. ниже).

I.3. К самым частым грамматическим словам принадлежат в флективных языках (напр. в чешском, русском и др.) также местоимения, прежде всего указательное или же анафорическое местоимение **ten** (этот, тот), которого мы уже коснулись выше. Правда, местоимение **ten** относится однозначно к самым частым языковым средствам в публицистическом и научном стилях, но не в административном стиле. Там преобладает, как уже указывалось, письменная форма соответствующих жанров стиля, поэтому местоимение **ten** заменяется его сложной формой **tento**. Эта сложная форма обозначает кроме основного указательного значения также подчеркивание. Местоимение **tento** можно квалифицировать как типичное, хотя не наиболее частое местоимение не только административного стиля, но и нехудожественного (по данным ЧСЧ, его частота 73,22 %, в художественном стиле - 26,78 %). Специфика сложного местоимения **tento** для нехудожественного стиля обнаруживается при сравнении с частотой местоимения **ten**, которое можно считать специфичным для художественного стиля (по данным ЧСЧ, его находим в 73,44 % в художественном стиле, против 26,56 % в нехудожественном. Наряду с местоимением **tento** вторым характерным лексическим средством нехудожественного стиля служит относительное и вопросительное местоимение **který** (который), хотя не преобладает численно (по данным ЧСЧ, 52,3I % в нехудожественном стиле, 47,69 % в художественном). Это местоимение присоединяет наиболее часто придаточные предложения относительные, а именно атрибутивные (**Těšitelová a kol.**, I982, 70). В административном стиле местоимение **který** является даже самым частым местоимением вообще, более частым, чем местоимения **tento** и **ten**. Для нехудожественного стиля характерен следующий порядок самых частых местоимений: **ten** , **který**, **tento** (между первыми десятью наиболее частыми словами); для художественного стиля типичен порядок: **ten**, **on**, **který** (по данным ЧСЧ). Местоимение третьего лица **on** (он) находим в нехудожественном стиле лишь между десятым и двадцатым словом по убывающей частоте.

I.4. Глагол-связка **být** (быть) и глагол **mit** (иметь), который имеет в чешском языке в значительной степени аналогичный характер, принадлежат к относительно самым частым словам в чешском языке вообще, но в нехудожественном стиле их частота ниже, чем в художественном стиле (ср. в ЧСЧ **být** 62,68 % в художественном стиле, 37,42 % в нехудожественном; **mit** 67,5I % в художественном стиле, 32,49 % в нехудожест-

венном. Аналогичную ситуацию можно видеть у модальных глаголов, которые мы часто находим в художественном стиле (в ЧСЧ **moci** (мочь) 59,28 %, **muset** (быть должен) 61,58 %, **chtít** (хотеть) 81,24 %) по сравнению с нехудожественным стилем (**moci** 40,72 %, **muset** 38,42 %, **chtít** 18,76 %). Из этого вытекает, что характерной чертой нехудожественного стиля с точки зрения лексикальных средств является относительно более низкое количество употребления глагола-связки **být**, глагола **mít** и даже модальных глаголов **moci, muset, chtít** (**Těšitelová**, 1974, 135-136).

1.5. Полнозначные слова в нехудожественном стиле не входят в большинстве случаев в группу самых частых слов. Это касается также первых двадцати наиболее частых слов в публицистическом и научном стилях. Исключение составляет административный стиль, в котором уже во второй половине первых двадцати наиболее частых слов (ранги 11.-20.) находим существительные (в чешском языке, напр. **pracovník** (работник), **úkol** (задача), **práce** (работа, труд)) и прилагательное **pracovní** (рабочий). Это, видимо, отглагольные слова, которые вопреки именной форме сохраняют значение действия и увеличивают характерные черты действия в нехудожественном стиле (ср. ниже распределение частей речи). По сравнению с этим в художественном стиле к самым частым словам относятся прежде всего собственные имена главных персонажей в тексте, имена местные, существительные, обусловленные темой, и т.п.

1.5.1. С у щ е с т в и т е л ь н ы е. Даже среди самых частых существительных встречаются такие, которые типичны для нехудожественного стиля. Хотя они принадлежат в значительной мере к самым частым существительным в чешском словарном запасе вообще (ср. ЧСЧ), их распределение показывает, что главное применение они находят в нехудожественном стиле или же в некоторой из его составных частей. В нашем материале это касается, напр. существительного **práce** (по данным ЧСЧ, 71,24 % в нехудожественном стиле, 28,76 % в художественном) и **doba** (время) (69,41 % в нехудожественном стиле, 30,59 % в художественном). Другие относительно частые существительные, которые мы находим среди первых десяти наиболее частых существительных, более выразительно относятся к лексическим средствам нехудожественного стиля. Это, напр., существительные **rok** (год) (по данным ЧСЧ 67,19 % в нехудожественном стиле, 32,81 % в художественном), **úkol**

(82,07 % в нехудожественном стиле, 17,93 % в художественном). С убывающей частотой растет число существительных, которые относятся главным образом к языковым средствам нехудожественного стиля или же к его определенной составной части. Ср., напр., существительные pracovník (по данным ЧСЧ 97,60 % относится к нехудожественному стилю, 2,40 % к художественному; в административном стиле это слово - даже самое частое существительное вообще), oblast (область) (в нехудожественном стиле 96,39 %, в художественном 3,61 %); в научном стиле слово oblast занимает по частоте третье место после существительных práce и doba (их распределение ср. выше). Некоторые из самых частых существительных принадлежат однозначно к отдельным частям нехудожественного стиля. Например, в публицистическом стиле это касается существительного strana (сторона, партия) (по данным ЧСЧ 63,05 % в нехудожественном стиле (из этого 70 % в публицистике), 36,95 % в художественном стиле), organizace (организация)(в нехудожественном стиле 92,13 % (из этого 61,39 % в публицистике), в художественном стиле 7,87 %) и т.п. Для административного стиля таким существительным является, напр., существительное závod (завод) (в нехудожественном стиле 91,61 %, в художественном 8,39 %), podnik (предприятие) (в нехудожественном стиле 85,79 %, в художественном 14,21 %) и т.п. Из лексических средств научного стиля приведем, напр., существительное teorie (теория) (99,04 % в нехудожественном стиле (из этого в научном стиле 93,20 %), 0,96 % в художественном стиле) и т.д. Приведенные примеры существительных как лексических средств нехудожественного стиля и его отдельных частей показывают, - на наш взгляд, - специфику нехудожественного стиля и его частей. Учитывая значение существительных в тексте (ср. ниже), данные признаки нехудожественного стиля и его частей становятся более важными и существенным образом способствуют образованию нехудожественного стиля и его частей.

1.5.2. Имена прилагательные. Так как прилагательные соединяются по своему значению особенно с существительными, можно предполагать, что они отражают характерные черты нехудожественного стиля только ограниченно. Это касается, напр., некоторых определенных прилагательных, частота которых относительно высока; такими прилагательными являются, напр., nový (новый)(в нехудожественном стиле 54,49%, в художественном 45,51 %), jiný (другой, иной) (в нехудоже-

ственном стиле 52,83 %, в художественном 47,17 %), celý (весь) (в нехудожественном стиле только 35,45 %, в художественном даже 64,55 %) и т.д. Хотя прилагательные nový и jiný мы находим в нехудожественном стиле и его частях среди первых наиболее частых прилагательных, прилагательное celý показывает особенно низкую частоту в административном стиле. Типичным прилагательным для нехудожественного стиля является další (следующий, дальнейший), имеющее значение как места, так времени. Даже в ЧСЧ мы находим прилагательное další, составляющее 78,32 % в нехудожественном стиле (причем научный стиль несколько преобладает над публицистическим (55 %: 45 %)), в художественном 21,18 %. Другие прилагательные более или менее характерны для некоторой из определенных частей нехудожественного стиля. Напр., прилагательное pracovní (в нехудожественном стиле 97,27 % (в том числе 67 % в научном стиле; в административном стиле (ЧСА) выступает самым частым прилагательным), в художественном стиле 2,73 %). Для публицистического стиля характерна, напр., высокая частота прилагательного sovětský (советский)(в нехудожественном стиле 97,09 % (в том числе 75 % в публицистическом стиле, в ЧСП это второе наиболее частое прилагательное), в художественном 2,91 %). Для научного стиля характерно, напр., прилагательное společenský (общественный)(в нехудожественном стиле 89,62 % (в том числе в научном стиле 90 %), в художественном 10,38 %) и т.п. Из сказанного вытекает, что прилагательные даже в ограниченных условиях их присутствия в текстах разных функциональных стилей проявляют характерные черты нехудожественного стиля и его частей.

I.5.3. Г л а г о л ы. Если мы пропустим глаголы-связки и модальные глаголы (ср. выше), то полнозначные глаголы можно разделить в нехудожественном стиле на две группы: первая из них обусловлена спецификой лексических средств данного функционального стиля, вторая - спецификой тематической. В первую группу входит, напр., глагол znamenat (значить) (в 39,01 % случаев мы находим его в нехудожественном стиле (в том числе 61,54 % в научном стиле, 38,46 % в публицистике), 30,99 % в художественном стиле), uvést (ввести, привести) (82,55 % в нехудожественном стиле (в том числе 74 % в науке), 17,45 % в художественном стиле), používat (использовать, пользоваться) (87,13 % в нехудожественном стиле (в том числе в науке 81,82 %), в художественном стиле 12,87 %) и т.п. Во вторую группу входят глаголы, типичные

для нехудожественного стиля, напр., глагол pracovat (работать) в 68,69 % случаев в нехудожественном стиле (в том числе 51,61 % в публицистике, 48,39 % в науке), 31,31 % в художественном) и т.д. Другие глаголы типичны для отдельных частей нехудожественного стиля. Особенно ярко это проявляется в административном стиле (ср. ЧСА). Это, напр., глаголы zajistit (обеспечить), (83,13 % в нехудожественном стиле, 12,87 % в художественном), provést (произвести, провести) (55,47 % в нехудожественном стиле, 44,53 % в художественном) и т.д. Для публицистики характерен, напр., глагол patřit (принадлежать) (по данным ЧСЧ, его частота в художественном стиле выше (64,25 %), чем в научном стиле (35,75%, но 53 % покрывает публицистический стиль). Для научного стиля характерна, напр., частота глагола existovat (существовать) (по данным ЧСЧ, мы находим его в 86,44 % случаев в нехудожественном стиле, в том числе в 73 % в науке, в 13,56% в художественном стиле). Можно, конечно, продолжить число примеров, но из вышеприведенного явствует, что даже полнозначные глаголы способствуют образованию нехудожественного стиля и его частей.

2. Грамматические средства

1. Из грамматических категорий, в частоте которых проявляются характерные черты нехудожественного стиля и его частей, рассмотрим прежде всего части речи. В соответствии с ЧСЧ в настоящей статье проведено деление с квантитативной точки зрения на две группы: (1.) именную и (2.) глагольную.

1.1. К именной группе, по данным ЧСЧ, принадлежат существительные, прилагательные и предлоги. В нехудожественном стиле на существительные приходится около 34% всех слов (по данным ЧСЧ, в художественном стиле на существительные приходится 25 %). Что касается существительных в отдельных частях нехудожественного стиля, то относительно большое число их мы находим в административном стиле(41,69%); в публицистике и науке данная пропорция почти одинакова: 33,51 % и 33,13 %. В предложениях нехудожественного стиля существительные более часто употребляются как несогласованный атрибут (38,78 %), частота других синтаксических функций намного ниже. Это касается обстоятельства (17,99 %), объекта (17,35 %) и субъекта (17,14 %) (L. Uhlířová, в печати). Почти идентичное разделение синтаксических функций существи-

тельных мы находим также в текстах публицистических и научных. Этим можно объяснить, за исключением существительных мужского рода одушевленных и среднего рода в единственном числе, относительно наиболее высокую частоту родительного падежа ед. и множ. числа. Численное преимущество вин.пад.ед. ч. мужского рода неодушевленного, вин. пад. мн. ч. женского рода и вин. пад. ед. и мн. числа среднего рода над предл. пад. ед. и мн. числа указывает на взаимоотношение синтаксической функции обстоятельства и объекта. В мужском роде в ед. числе преобладает однозначно имен. пад. (54,99 %), т.е.преимущественно синтаксическая функция субъекта, в среднем роде с незначительным преимуществом (в имен. пад. ед. ч. в 29,31 % случаев против 27,09 % в род. пад. ед. ч.).В отдельных частях нехудожественного стиля также преобладает род. пад. ед. ч. (за исключением мужского рода одушевленного), т.е. частота несогласованного атрибута или же объекта. Таким образом, существительные как категория частей речи и их категории морфологические (падеж, число, род) и синтаксические составляют своим присутствием характерные черты нехудожественного стиля и его отдельных частей.

Нехудожественный стиль отличается также более высоким числом прилагательных (19 %); в художественном стиле их меньше (8,45 %). В составных частях нехудожественного стиля прилагательное достигает самой высокой частоты в научном стиле (около 20 %), относительно низкой частоты в публицистике (около 18 %), частота прилагательных в административе совпадает в основном с частотой в нехудожественном стиле вообще. Относительно высокая частота употребления прилагательных в нехудожественном стиле и его частях имеет связь особенно с большой частотой терминов, которые состоят часто из устойчивых словосочетаний, из существительного и атрибута и выражаются прилагательным в синтаксической функции согласованного атрибута или существительным в функции несогласованного атрибута (ср. выше). Поэтому частота несогласованного атрибута достигает 87,49 % в нехудожественном стиле, в административных текстах даже 91,92 %. Это является также причиной наиболее высокой частоты родительного падежа прилагательных, хотя не в таком преимуществе, как это бывает при существительных. Действительно, прилагательные связаны в предложении с существительными, но не последовательно.Иначе говоря, прилагательное сопровождает существительное в различных падежах разным способом. Итак, кроме род.пад.(не-

взирая на категорию именного рода) с частотой, равной 29,80%, т.е. более низкой, чем род. пад. существительных, относительно часто встречается именительный падеж ед. и мн. ч. прилагательных (26,98 % в ед. ч., 23,02 % во мн. ч.), следует винительный падеж (с почти идентичной частотой в ед. и мн.ч. 18,49 % : 18,37 %), творительный (в ед. ч. 11,06 %, во мн.ч. 8,32 %) и предложный падеж (в ед. ч. 10,84 %, во мн.ч. 7,32%). Значит, в нехудожественном стиле прилагательное сопровождает существительное наиболее часто в синтаксической функции субъекта, объекта и обстоятельства (кроме несогласованного атрибута). В науке мы встречаем именительный падеж чаще, чем родительный, т.е. прилагательное, сопровождающее существительное, выступает в синтаксической функции субъекта чаще, чем в функции несогласованного атрибута. Частота встречаемости творительного падежа выше, чем предложного. Это значит, что применение прилагательного у существительных мужского рода приводит к тому, что существительное в синтаксической функции объекта или обстоятельства в творительном падеже встречается чаще, чем в синтаксической функции обстоятельства или объекта в предложном падеже и т.п. В нехудожественном стиле и его частях можно наблюдать, хотя и с некоторыми вариациями, следующий порядок самых частых падежей: родительный, именительный, винительный, творительный и предложный, т.е. в основном соотношение синтаксической функции несогласованного атрибута, субъекта, объекта, и/или объекта. Такова еще одна дополнительная грамматическая черта нехудожественного стиля и его частей. Этот факт подтверждается также сравнением с частотой падежей и их синтаксических функций у прилагательных в художественном стиле: именительный, родительный, винительный и творительный падежи. Следовательно, в художественном стиле применение прилагательного приводит к более частому употреблению в предложении существительных в синтаксической функции субъекта, несогласованного атрибута, объекта или именной формы сказуемого (Těšitelová, 1980, 72).

Частота существительных и прилагательных как категорий частей речи и частота их морфологических и синтаксических категорий, которые мы в настоящей статье в отборе показали, свидетельствуют о существовании нехудожественного стиля и его самостоятельных частей.

1.2. Частота глагола в нехудожественном стиле, как основного представителя г л а г о л ь н о й группы наряду

с местоимениями, наречиями и союзами), относительно ниже (I3 %), чем в художественном стиле (приблизительно 2I %). Это имеет связь со стилевыми методами в нехудожественном стиле, к которым относятся прежде всего объяснение и описание и которые больше опираются на наименование субстанций и их свойств, чем на точное выражение деятельности, состояния и т.п. Кроме того, для наименования данных явлений служат также существительные и прилагательные отглагольного происхождения, напр., отглагольные существительные, отглагольные прилагательные и т.п., которые в нехудожественном стиле и его частях значительно увеличивают глагольную группу. Например, в корпусе нехудожественного стиля в чешском языке (ср. выше) мы находим 8,08 % существительных отглагольного происхождения; примерно столько же мы находим их в научной речи (8,37 %), больше всего в административе (I0,45 %), меньше в публицистике (6,63 %). Этим можно объяснить динамичный характер нехудожественного стиля и его частей, напр., публицистики.

Разумеется, что глагол выступает также в нехудожественном стиле наиболее часто в синтаксической функции предиката (приблизительно 86 %). Это касается и отдельных частей нехудожественного стиля. Если обратить внимание на тип сказуемого, то видим, что в нехудожественном стиле предикат, состоящий из полнозначного глагола, встречается реже (80,49 %), чем в публицистике и административе (85,68 %), причем реже всего в научной речи (77,95 %). Самую высокую частоту в научной речи проявляет именное сказуемое со связкой (20,4I %); в нехудожественном стиле встречается реже (I7,83 %). В публицистике частота его встречаемости достигает только I5,53 %, в административе - лишь II,5I %. Отношение сказуемого, состоящего из полнозначных глаголов, к именному сказуемому со связкой является еще одной характерной чертой для нехудожественного стиля и его частей.

При полнозначных глаголах мы можем изучать частоту их морфологических категорий. Для нехудожественного стиля типично, напр., преобладание настоящего времени, в художественном стиле доминирует однозначно прошедшее время. Также категория глагольного рода дифференцирует своей частотой нехудожественный стиль и его части и художественный стиль. Между тем как в художественном стиле преобладают формы действительного залога, в нехудожественном стиле более частыми являются формы страдательного залога, и даже в отдельных

частях нехудожественного стиля. Что касается категорий лица и числа, для нехудожественного стиля типична пропорция 3-го л. ед. и мн. ч. (которая имеет наиболее высокую частоту как в нехудожественном, так и в художественном стилях) I-го л. мн. ч.; в художественном стиле это соотносится с I-ым л.ед. ч.

С менее низким числом глаголов в нехудожественном стиле связано менее низкое количество местоимений в рамках отглагольной группы. То же самое относится к менее низкому числу наречий в нехудожественном стиле. Более высоким числом наречий выделяются публицистика (8,76 %) и научная терминология (I2,80 %), самое низкое число приходится на административу (5,05 %). Количество наречий преобладает,по данным ЧСЧ, в художественном стиле (66 %) по сравнению с нехудожественным (34 %).

Союзы несколько расходятся в нехудожественном стиле с другими частями речи из отглагольной группы. В нехудожественном стиле на союзы приходится приблизительно 8 %, то же самое касается публицистики и науки. Чрезвычайно низкое количество союзов мы находим в административе (6,36 %). Это имеет связь, по-видимому, с очень простой структурой предложений: 66,28 % простых предложений против 33,72 % сложных предложений. В нехудожественном стиле число сложных предложений (53,20 %) преобладает над числом простых предложений (46,80 %). В общем ту же самую пропорцию мы находим в публицистике : 53,99 % сложных предложений, 46,0I % простых предложений. Самой сложной структурой предложения выделяется наука: 58,03 % сложных предложений и 4I,97 % простых предложений. В художественном стиле количество простых предложений достигает 55 %, количество сложных предложений - 45%. Это значит, что пропорциональность структуры предложения также определяет границу между нехудожественным и художественным стилями.

З а к л ю ч е н и е. На основе квантитативной оценки некоторых лексических и грамматических средств нехудожественный, деловой стиль ясно выделяется при противопоставлении с художественным стилем. В рамках делового, нехудожественного стиля индивидуальными чертами обладают его составные части: наука, публицистика и административ. Деловой,нехудожественный стиль представляет собой важный функциональный стиль, который типичен для развитого социалистического общества.

ЛИТЕРАТУРА

Havránek B. Studie o spisovném jazyce. Praha, 1963.

Jedlička A. a kol. Základy české stylistiky. Praha, 1970.

Jelínek J., Bečka J. V., Těšitelová M. Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha, 1961.

Mistrík J. Štylistika slovenského jazyka. Bratislava, 1970.

Těšitelová M. Otázky lexikální statistiky. Praha, 1974.

Těšitelová M. Využití statistických metod v gramatice. Praha, 1980.

Těšitelová M. a kol. Frekvenční slovník současné české publicistiky. Praha, 1980.

Těšitelová M. a kol. Frekvenční slovník současné české administrativy. Praha, 1980.

Těšitelová M. a kol. Kvantitativní charakteristiky současné české publicistiky - In: Linguistica II. Praha, 1982.

These o poznámky k diskusi. Propositions. Sekce II, Praha, 1929.

Uhlířová, L. Simple sentence structure from the quantitative point of view (based on present-day Czech texts of non-fiction). - In: PSML, 8 (в печати)

ON THE SO-CALLED NON-FICTION STYLE FROM A QUANTITATIVE POINT OF VIEW

Marie Těšitelová

S u m m a r y

On the basis of statistical data on some (1.) lexical and (2.) grammatical language means the author identifies characteristic features of the functional non-fiction style in contradistinction to the style of fiction. Within the frame of the non-fiction style three components are noted for their specific features: the scientific style, the publicist style and the administrative style. The degree of specificity of the styles corresponds to the quantitative representation of certain lexical and grammatical phenomena; the administrative and publicist styles appear to take extreme positions on the scale, with the publicist style rather approaching the style of fiction. The non-fiction style in itself represents an important complex functional style characteristic of the period of the developed socialist society.

СОЦИАЛЬНАЯ ДИФФЕРЕНЦИАЦИЯ ЛЕКСИКИ ЭСТОНСКОГО ЯЗЫКА С КВАНТИТАТИВНОЙ ТОЧКИ ЗРЕНИЯ

Ю.А. Тулдава

Словарный состав языка дифференцируется по разным признакам, в том числе по сферам употребления, имеющим внелингвистическое, социальное основание. Сферы употребления лексики тесно связаны с функционально-языковыми (функционально-стилистическими) характеристиками слов. На этой основе в данной статье рассматриваются проблемы выделения основных разновидностей лексики, исходя из установок квантитативно-системного анализа языка. Исследование проводится на материале эстонского языка с привлечением сравнительных данных из других языков. Статья является частью более обширной работы по изучению общих квантитативно-типологических свойств лексики эстонского языка (предыдущую статью в этой серии о генетическом составе эстонской лексики см. Тулдава Ю., 1982).

Сферы употребления лексики. Современный эстонский язык представляет собой сложную систему ("архисистему"), объединяющую различные формы его существования: язык общего употребления, или "общий" язык (эст. üldkeel), территориальные и социальные диалекты (murdekeel ja argoo), языки (подъязыки) науки и техники (oskuskeel).* Соответственно лексика современного эстонского языка, т.е. весь его словарный состав распадается по сферам употребления (распространения) на три основные группы, или разновидности: общеупотребительная, или

* Язык народа как совокупность всех своих вариантов именуется "этноязыком" (термин А.И.Горшкова; см. Березин Ф.М., Головин Б.Н., 1979, с. 59). В таком же всеобъемлющем значении употребляется часто термин "национальный язык" для обозначения языка эпохи нации (см. Филин Ф.П., 1975). Однако некоторые исследователи оставляют территориальные диалекты за пределами национального языка (например, Гроссе Р., 1970, в отношении немецкого языка). Такая более узкая трактовка понятия национального языка (эст. rahvuskeel) является традиционной и в эстонском языкознании, причем особо выделяются такие черты развитого национального языка, как общность для всего народа и возможность общаться с его помощью во всех сферах жизни (см. Erelt T., 1982, с. 178).

общая лексика (эст. üldsõnavara), диалектная лексика (murdesõnavara) и специальная, или терминологическая лексика (oskussõnavara). Последние две группы - диалектная и специальная лексики - можно объединить под общим названием "лексика ограничительного употребления", причем некоторые исследователи выделяют в ней еще жаргонную лексику (эст. argoosõnavara) как самостоятельную подгруппу. В данной работе мы относим жаргоны (и арго) к социальным диалектам и рассматриваем жаргонную лексику как особую подгруппу диалектной лексики (в широком смысле слова). В то же время следует рассматривать т.н. профессионализмы (полуофициальные, преимущественно разговорные разновидности терминов) как подгруппу специальной лексики.

Основные разновидности лексики существуют как реальные данности, но будучи подсистемами в составе общей системы лексики данного языка, они характеризуются всеми свойствами подсистем, т.е. они системно связаны, взаимопроникают друг в друга и т.п. Всю систему лексики можно для наглядности представить в виде пересекающихся множеств, представляющих основные компоненты словарного состава языка (см. рис. I). Особое место в этой системе занимает лексика литературного языка, которая в рассматриваемой модели занимает промежуточное положение.

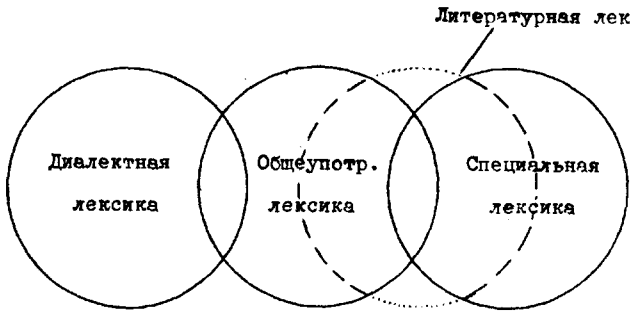


Рис. I. Основные компоненты лексики современного языка в синхронном разрезе (деление по признаку сферы употребления).

Понятие о б щ е у п о т р е б и т е л ь н о й л е к с и к и является в известной мере условным. К общеупотребительной лексике мы относим слова, понимание и употребление которых "не ограничено какой-то местностью, профессией или родом деятельности" (см. Калинин А.В., 1978, с.119, который в том же значении употребляет термин "общенародная лексика"). Подобная лексика составляет устойчивую основу языка, хотя по своему составу она не является однородной. В ней различаются пласты лексики устной разговорной речи (в том числе просторечия) и лексики письменной речи (в том числе "книжной" лексики). С точки зрения исторической перспективы в ней выделяются различные генетические пласты лексики, неологизмы и архаизмы. На основе статистических критериев можно выделить ядерную лексику (которая в основном совпадает с так называемым основным словарным фондом языка) и периферийную лексику. Во всех случаях дифференциации основным признаком общеупотребительной лексики остается то, что, по определению, ее употребление не связано с местом жительства или родом деятельности.

Нет четкой демаркационной линии между общеупотребительной лексикой и другими разновидностями лексики (см.рис. I). Общеупотребительная лексика обогащается постоянно за счет диалектных слов, которые проникают в общее употребление в результате сближения города и деревни или в результате сознательного внедрения диалектизмов языковедами, особенно в периоды становления и формирования литературного языка.В то же время имеет место обратный процесс - влияние общеупотребительной лексики на диалекты, вследствие чего происходит сближение диалектов с общенародным языком. Общепонятными и общеупотребительными становятся со временем также многие слова социальных диалектов (жаргонные слова и арготизмы),которые употребляются в устной или письменной речи в целях эмфаза и экспрессии.

Нет четких границ и между общеупотребительной и специальной лексикой. В наше время наблюдается усиленный переход терминов в общее употребление, в результате чего в языке постоянно увеличивается слой общепонятных терминов. Это является одним из следствий всеобщей "интеллектуализации" современных языков. С другой стороны, многие общеупотребительные слова приобретают наряду со своим общим значением еще специальное значение (например, эст. alus 'основание' в общем языке и как термин в подъязыке химии - в значении 'ос-

нование' и в подъязыке лингвистики – в значении 'подлежащее') или полностью переходят в разряд специальных слов, не говоря уже о том, что большое количество терминов образуется на базе общеупотребительной лексики (aastarõngas 'годичное кольцо' в ботанике, juursõna 'корневое слово' в лингвистике и др.).

Таким образом приходится констатировать, что слова отдельных сфер употребления не имеют точно фиксированных, неподвижных разграничений, и наблюдаются постепенные, "непрерывные" переходы между лексическими группами. Объем и границы лексических групп отдельных сфер употребления могут объективно определяться с помощью вероятностно-статистических критериев (частотность, употребительность, распределение частот слов и др.). Известны, например, критерии разграничения общей и специальной лексики на основе дифференцированных данных о степени употребительности слов в различных подъязыках (Андреев Н.Д., 1967) и критерии автоматического выделения терминологической лексики на основе распределения частот слов ("эффект Бектаева", см. Пиотровский Р.Г., 1980). Интерес представляют и эмпирические наблюдения в работах, посвященных исследованию специальных текстов, например, констатация того, что на словаре, составленном на основе специального текста объемом 100 тыс. словоупотреблений, узкоспециальные слова (относящиеся к главной тематике отрасли) концентрируются в "среднечастотной" зоне: $50 > F > 15$, где F – частота слова в тексте (Негуляев Г.А. и др., 1973).

Литературная лексика. На фоне основных групп лексики, образовавшихся в результате расчленения общей системы лексики по признаку сфер употребления (см. рис. 1), особо выделяется лексика литературного языка, или литературная лексика как своеобразный промежуточный слой и в то же время важнейший компонент лексической системы современного языка. Литературный язык (эст. kirjakeel), к которому, естественно, относится литературная лексика, занимает ведущее положение в составе современного языка как главное средство общения нации. Эстонский литературный язык охватывает в наши дни не только язык письменности, но и общепринятый у образованных людей язык устного общения. Этот литературно-разговорный язык отличается от обиходно-бытового языка (эст. argikeel), главным образом, большей нормированностью грамматики, но имеются и различия в выборе

лексики (в обиходно-бытовом языке, например, отсутствуют книжные слова, но зато чаще встречаются просторечные или жаргонные слова). Все же разница между литературно-разговорной и обиходно-бытовой разновидностями современного эстонского языка не такая большая, как в некоторых других языках, например, в близкородственном финском языке (Rätsep H. 1981, с. 203) или, в особенности, в чешском языке, где существуют две заметно отличающихся друг от друга формы устной речи (см. Филин Ф.П., 1975, с. 5-6).

Литературный язык эпохи нации продолжает оставаться в тесных взаимосвязях с другими разновидностями языка, видоизменяется под их влиянием и сам воздействует на них. Естественно, что и лексика литературного языка находится в непрерывном взаимодействии с другими разновидностями лексики; она возникла на их основе и продолжает впитывать в себя элементы лексик разных сфер употребления. В то же время литературная лексика обладает известной самостоятельностью и устойчивостью, благодаря тому, что самому литературному языку в современном понимании свойственны стабильность, нормативность и общеобязательность для всех членов языкового коллектива. Если попытаться более точно разграничить литературную лексику от других разновидностей лексики языка, то целесообразно воспользоваться методом противопоставлений (так как "всякое понятие лучше всего выясняется из противоположений" - Щерба Л.В., 1957, с. 115).

Прежде всего надо отметить, что хотя литературный язык рассматривается в наши дни как главное средство общения нации, все же лексика литературного языка полностью не совпадает с "общеупотребительной" лексикой в принятом нами значении. От общеупотребительной лексики она отличается в первую очередь четкой нормированностью (нормы закреплены соответствующими правилами и словарями литературного языка). В состав литературной лексики не входят просторечные или грубые слова, хотя элементы просторечия и т.д. могут быть применены, например, в художественной литературе в стилистических целях. Далее, в состав литературной лексики не входят диалектные или жаргонные слова, хотя и они могут встречаться в литературных текстах как "внелитературные" вкрапления. Наконец, из литературной лексики исключаются некоторые слои специальной лексики, в частности слова профессионального просторечия. На вопрос о включении узкоспециальной терминологии в состав литературной лексики решается по-раз-

ному. Можно, по-видимому, говорить о литературной лексике в широком и узком смысле слова; в последнем случае выделяется общеупотребительная литературная лексика, или о б щ е л и т е р а т у р н а я л е к с и к а (эст. kirjakeele üldsõnavara), которая составляет основную, базисную часть литературной лексики современного языка. Общелитературная лексика совпадает с литературной лексикой в широком смысле в части общеупотребительных слов, но отличается от последней тем, что в нее не включена узкоспециальная терминология.

Таким образом, понятие литературной лексики, как особой подсистемы лексики языка, раскрывается в общих чертах через противопоставление "литературной" и "нелитературной" лексики; к последней следует отнести диалектные, жаргонные и просторечные слова, а также просторечные профессионализмы. Притом допускается применение нелитературных слов в литературном языке в стилистических целях, т.е. эти слова приобретают в литературном языке особую эмоционально-экспрессивную окраску (особую лингвостилистическую функцию) и их можно отнести к "временным" (порой окказиональным) элементам в составе литературной лексики. Более точное и объективное разграничение литературной и нелитературной лексик может осуществиться, как и во многих других задачах разграничения лексических групп, с помощью вероятностно-статистических критериев в сочетании с качественным анализом.

Следует отметить, что предлагаемая выше схема социальной дифференциации языка и лексики (ср. также рис. I) отнюдь не исчерпывает всех возможностей взаимосвязи и взаимодействия основных социальных разновидностей лексики, и могут быть и другие решения проблемы. Наша схема дает лишь наиболее общую картину взаимосвязей в синхронном плане, достаточную для решения некоторых задач общего значения в рамках квантитативно-системного подхода к изучению лексики.

Функциональные стили и лексика. Известно, что литературный язык не является каким-то однородным целым, а разветвляется на многие жанры и стили. На основе общественных функций языка (общение, сообщение, воздействие) и учитывая сферу общения (эстетическая, научная, официально-деловая и др.), в литературном языке различаются т.наз. ф у н к ц и о н а л ь н ы е с т и л и.* Обычно выделяют четыре основных,

* Если упор делается на тематику (предметную область действительности), то говорят также о п о д ъ я з ы к а х, которые по объему могут совпадать или не совпадать с функциональными стилями (подробнее см. Тулдава Ю., 1981, с. 120-121).

"базовых" функциональных стиля литературного языка: художественный[**], публицистический, научный, официально-деловой. Эти литературные "книжные" стили противопоставляются разговорному (разговорно-бытовому) стилю, характеризующемуся преимущественно функцией общения в сочетании со спонтанностью и непринужденностью. Базовые функциональные стили распадаются в свою очередь на частные разновидности в зависимости от проявления конкретных задач и ситуаций общения, вплоть до выражения функционально-стилистических особенностей индивидуального характера (см. Кожина М.Н., 1977, с. 157).

Из сказанного следует, что принципы формирования функционального стиля выводятся из его внеязыковой, экстралингвистической основы (функции, цели и задачи коммуникации, сферы общения и др.). Но возникнув на внеязыковой основе, будучи тесно связанными с предметом высказывания, функциональные стили (и их подвиды) различаются между собой внутриязыковыми признаками, а именно особенностями отбора и употребления языковых средств. Что касается л е к с и к и функциональных стилей, то очевидными представляются два момента. Во-первых, основную часть словаря любого функционального стиля составляет в стилевом отношении нейтральная общеупотребительная лексика. Во-вторых, существует еще специфичная для каждого стиля лексика, которая характерна только для данного стиля (или подъязыка). Однако на практике приходится более дифференцированно подходить к вопросу о разделении лексики на группы. При сравнительном исследовании текстов, принадлежащих к разным функциональным стилям (подъязыкам), выделяется особый, ядерный слой общеупотребительной лексики, который оказывается в буквальном смысле общим для всех текстов даже при сравнительно небольших объемах выборок. Кроме того, выделяются слои лексики, общие для определенного количества текстов (групп текстов) и т.д. вплоть до специфичной лексики одной узкой группы текстов или даже одного индивидуального текста. При таком подходе можно говорить о противопоставлении "общей" и "специфичной" лексик, но при

[**] О правомерности выделения художественного (художественно-беллетристического) стиля среди других функциональных стилей см. Кожина М.Н., 1977, с. 58 и след. Несмотря на то, что в художественной речи могут быть использованы элементы других стилей, важно, что эти последние используются в художественной речи в эстетической функции, а не в той, которую они выполняют в тех стилях, откуда они взяты.

разных уровнях общности/специфичности, выбор которых зависит от целей и задач исследования. Во всех случаях противопоставления общей и специфичной лексик надо различать уровни словаря и текста, т.е. проводить различие между инвентарем и употреблением. Приведем пример. Частотный словарь русского языка (1977) составлен на основе четырех функциональных стилей (художественная проза, драматургия, газеты-журналы, научная литература), которые представлены одинаковым объемом текстов (по 250 тыс. словоупотреблений). Общая совпадающая часть лексики (т.е. слова, встречающиеся во всех четырех видах текстов) составляет 6440 единиц (лексем), или 16,4 % при объеме сводного словаря 39 268 единиц. Но эти 6440 слов покрывают 82,2 % всех текстов, на основе которых составлен частотный словарь (868 577 словоупотреблений из общего числа - 1 056 382). Таким образом, выявляется большое различие между квантитативными характеристиками словаря и текста. Это различие обусловлено тем, что небольшая часть общеупотребительной лексики - в строгом смысле "общая", или "ядерная" лексика - имеет большую частотность и в количественном отношении доминирует в актуальной речи (в текстах) всех сфер общения. В этой связи важно отметить, что общеупотребительная лексика ведет себя по-разному в разных сферах общения, т.е. наблюдается различие между частотными характеристиками общеупотребительных слов в разных видах текстов. Это обстоятельство позволяет (наряду с учетом количества и состава специфичной лексики) дифференцировать стили (и подстили) на основе объективных вероятностно-статистических критериев. Например, по данным упомянутого частотного словаря некоторые строевые и знаменательные слова распределяются в функциональных стилях следующим образом:

	худож. проза	драма-тургия	газеты журналы	научно-технич.
а (союз)	2909	5203	1355	1252
который	673	514	1381	1600
год	346	286	1080	555
большой	297	708	571	490
сказать	1278	978	359	294

Безусловно, важное значение имеет еще тот факт, что единицы общеупотребительной лексики могут полностью не совпадать в разных функциональных стилях не только по своим количественным, но и по качественным свойствам (значение, се-

Таблица I

Десять наиболее частотных существительных в разных функциональных стилях эстонского, финского и русского языков (Kaasik Ü. et al., 1977; Valge J., 1972; Saukkonen P. et al., 1979; Частотный словарь русского языка, 1977)

Эстонский язык				Финский язык				Русский язык	
Худож.проза		Газеты		Худож. проза		Газеты		Худож.пр.	Газеты
mees	муж(чина)	aasta	год	aika	время	vuosi	год	человек	правительство
aeg	время	valitsus	правительство	mies	муж(чина)	aika	время	рука	страна
silm	глаз	riik	государство	päivä	день	osa	часть	глаз	год
inimene	человек	rahvas	народ	käsi	рука	maa	страна;земля	дело	партия
käsi	рука	partei	партия	poika	сын;мальчик	asia	дело	жизнь	борьба
naine	жен(щин)а	poliitika	политика	ihminen	человек	ihminen	человек	голова	день
päev	день	suhe	отношение	lapsi	ребенок	työ	работа	лицо	газета
asi	дело;вещь	aeg	время	asia	дело	kysymys	вопрос	день	государство
pea	голова	osa	часть	maa	страна;земля	päivä	день	мать	сила
nägu	лицо	õigus	право	pää	голова	määrä	степень	время	вопрос

мантический объем, экспрессивность и т.д.).

В результате различий в частотности слов выясняются и различия в распределениях, в покрываемости текста и других квантитативно-типологических характеристиках текстов, принадлежащих к разным сферам общения и тем самым к разным функциональным стилям (или подъязыкам). Если сравнивать, например, списки наиболее частотных существительных в разных функциональных стилях и языках (см. табл. I), то выявляется следующее. Большое различие наблюдается в списках слов разных стилей, например, среди десяти наиболее частотных существительных в художественной прозе и в газетном тексте в эстонском языке общим является только одно слово (aeg 'время'), в русском языке - также только одно слово (день). В то же время можно констатировать близость в распределении частотных существительных в одном и том же стиле, в частности в художественной прозе разных языков. Например, эстонский и финский языки имеют здесь семь общих (в данном случае генетически родственных) слов из десяти: (в переводе) мужчина, человек, рука, голова, время, день, дело. Эстонский и русский языки имеют восемь "общих", т.е. взаимопереводимых слов: человек, рука, глаз, голова, лицо, время, день, дело. Обращает на себя внимание, что среди общих частотных слов художественной прозы разных языков на видном месте находятся слова, обозначающие человека (человек, мужчина, женщина) и части его тела (рука, глаз, голова, лицо). Меньше сходств в газетном тексте разных языков. Например, в эстонском и финском языках среди десяти наиболее частотных существительных-понятий общими являются только три (год, время, часть), а в эстонском и русском языках - четыре (год, правительство, партия, государство). Здесь сказывается различие в тематике, частично обусловленное социально-политическими условиями, а также тем, что рассматриваемые тексты принадлежат к разным подвидам публицистического стиля (в эстонском языке рассматриваются внешнеполитические, а в финском и русском языках - смешанные газетные тексты).

Отмеченность/неотмеченность лексики. Лексические нормы эстонского литературного языка закреплены в "Ортологическом словаре" (Õigekeelsussõnaraamat, 1976) и в готовящемся к печати толковом словаре эстонского литературного языка. В эти словари включена в первую очередь общелитературная лексика в принятом нами значении. Многие слова, которые не являются общелитературными в строгом смысле слова,

хотя и включены в словарь, снабжены соответствующими стилистическими пометами. В словарях применена система помет, указывающих на сферу употребления (диалектная и специальная лексика, разговорные слова общеупотребительной лексики), историческую перспективу (устарелые слова) и экспрессивность (фигуральные, шутливые, пренебрежительные слова и т.п.). Отличительной чертой эстонских нормативных словарей является то, что в них преобладают пометы различных подвидов специальной лексики. Например, в ортологическом словаре общее число помет - 61, из них 52 относятся к специальной лексике.

Рассмотрим подробнее квантитативное распределение отмеченных и неотмеченных слов в ортологическом словаре эстонского языка. Слварь охватывает 115 000 заглавных слов, причем общая встречаемость помет - 42 033. Чтобы правильно определить доли отмеченной и неотмеченной лексики в ортологическом словаре, следует учесть тот факт, что многие слова в словаре снабжены двумя или более пометами, например, reaskülv 'рядовой посев' (имеет пометы "сельское хозяйство" и "лесоводство"), põhikude 'основная ткань' (с пометами "ботаника", "зоология", и "текстильное дело"). В некоторых случаях встречаются совместно общее и специальные значения слов; например, заглавное слово laager, имеющее общее значение 'лагерь', отмечено еще пометами, указывающими на принадлежность к отраслям "военное дело" и "техника" (в последнем случае в значении 'подшипник'). В результате проверки (на основе выборки) было установлено, что около 6 % отмеченных слов имеют более одной пометы (из них 5,5 % слов с 2 пометами, 0,4 % - с 3 пометами и 0,1 % - с 4 пометами). Общее соотношение количества отмеченных слов и количества помет составляет приблизительно 0,9 : 1. Следовательно, при суммарной частоте помет - 42 083 - общее число слов, отмеченных одной или более пометами, составляет около 38 000 (=0,9· 42 083), т.е. 33 % всего состава ортологического словаря. С другой стороны, неотмеченных слов оказалось в словаре около 77 000, или 67 %. Если к ним добавить слова, не относящиеся ни к диалектной, ни к специальной лексике, но имеющие в словаре другие, например, экспрессивно-оценочные пометы, то можно приблизительно считать, что количество о б щ е у п о т р е б и т е л ь н ы х с л о в в ортологическом словаре современного эстонского языка достигает 80 000, т.е. около

70 % словаря.* Остальные 30 % состава словаря распределяются между стилистическим пластом диалектизмов и попадающей в сферу общеупотребительной лексики группой специальных терминов. В последнем случае имеется в виду, что приведенные в словаре термины могут встречаться не только в узкоспециальной литературе, но и достаточно часто в прессе, в научно-популярных статьях, в учебниках общеобразовательных школ и т.п.

Более точное разделение лексики на отмеченную и неотмеченную возможно на основе толковых словарей, в которых различаются отдельные значения слов. Здесь можно подсчитывать т. наз. стилистические позиции (т.е. отдельные значения и нюансы значений, фразеологические обороты и т.д.; см. Филин Ф.П., 1973, с. 7) и на их основе выяснить долю отмеченных единиц. Нами была сделана выборка на основе пробной тетради толкового словаря эстонского языка (Eesti kirjakeele sõnaraamat, 1969), где приводятся фрагменты словаря на буквы А, Е, Н.** В пробной тетради на 1510 заглавных слов оказалось 2211 позиций, из них отмеченных 648, т.е. 29,2% всех позиций.

Для сравнения можно привести данные толковых словарей русского языка. В 7-м томе "Словаря современного русского литературного языка" (1948-1965) из общего числа позиций - 15 530 - были отмечены пометами 3925, т.е. 25,3 % (Филин Ф.П., 1973, с. 7-8). В "Словаре русского языка" С.И.Ожегова (1963), насытывающем 51 533 слова, пометы употреблены 17 005 раза (Денисов П.Н., Костомаров В.Г., 1970, с. 72). По приблизительным подсчетам в словаре должно быть около 83 000 позиций, т.е. отмеченные единицы составляют примерно 20 %.

---

* Ориентировочно можно считать, что количество общеупотребительных слов в ортологическом словаре (80 000) соответствует объему „общелитературной лексики" современного эстонского языка. Однако нельзя забывать о богатых возможностях окказионального словообразования (особенно словосложения), которое является источником постоянного пополнения эстонского словаря новыми общеупотребительными словами. Анализ показывает, что около 20 % слов, зарегистрированных в малочастотных зонах (при частотах 1 и 2 в тексте объемом 100 тыс. словоупотреблений) в частотном словаре современной художественной прозы эстонского языка (Kaasik Ü. et al., 1977), отсутствуют в ортологическом словаре 1976 года.

** Данные пробной тетради следует считать лишь предварительными, так как в отредактированном (еще не опубликованном) варианте словаря могут быть различные поправки и изменения (см. Raiet E., 1983).

Таким образом, доля отмеченных единиц в рассмотренных нормативных словарях литературного языка колеблется между 20 и 30 процентами, причем в эстонских словарях этот процент достигает максимума. Притом следует иметь в виду, что распределение отмеченных единиц по типам стилистических характеристик (функционально-стилистических, экспрессивно-оценочных и т.д.) может в разных словарях существенно различаться (см. ниже).

Распределение отмеченной лексики. Распределение помет по частотности в ортологическом словаре эстонского языка приводится в таблице 2.* Различия в частотности помет ясно указывают на различную степень значимости отдельных групп слов, составляющих стилистически отмеченный слой в общей системе литературной лексики. Обращает на себя внимание плавное, монотонное убывание частот примененных помет, из которых большинство относится к специальной лексике. Такая регулярность в ранговом распределении частот напоминает о скрытом действии некоторых общих законов, вроде закона Ципфа, в коммуникативных и других сложных системах. Если (при соблюдении принципа однородности) взять только пометы специальной лексики (52 пометы из 61), то выяснится, что эмпирическое ранговое распределение их частот (и тем самым - частот отдельных тематических групп специальной лексики) в целом хорошо подчиняется логарифмическому закону (см. рис. 2) **, причем отклонение от общей тенденции наблюдается в "ядре", т.е. среди наиболее частотных отмеченных групп. Надо полагать, что ранговое распределение по логарифму представляет собой особый тип упорядоченности объектов в сложных социальных системах. Этот тип распределения встречается и в некоторых областях лингвистики (действие логарифмического закона обнаруживается, например, в ранговом распределении частот букв в тексте (см. Маскау А., 1965; Tuldava J., 1980).

Как уже было упомянуто, основная масса отмеченных слов в эстонских нормативных словарях относится к специальной лексике, т.е. терминологии. Например, в ортологическом сло-

* Автор благодарит научного сотрудника сектора вычислительной лингвистики ИЯЛ АН ЭССР к.ф.н. Ю. Викс за предоставление численных данных о распределении стилистических помет в ортологическом словаре эстонского языка.

** Линейная зависимость между частотой (F) и логарифмом ранга (ln i) по формуле $F = a + b \ln i$, где a и b - константы. В данном слечае $a \approx 3200$ (теоретически максимальная частота), $b \approx -800$ (указывает на темп убывания частот).

Таблица 2

Распределение стилистических помет в Ортологическом словаре эстонского языка. Объем словаря: 115000 слов; число помет: 61; общая встречаемость помет в словаре: 42083.

3433	tehn.	(техника)	484	farm.	(фармация)
2585	bot.	(ботаника)	435	kunst	(изобр.искусство)
2475	zool.	(зоология)	423	mets.	(лесоводство)
2437	med.	(медицина)	413	vet.	(ветеринария)
1991	põll.	(сельское хоз.)	406	etn.	(этнография)
1701	maj.	(экономика)	383	kirj.	(литература)
1589	sport	(спорт,физкульт.)	372	trük.	(типогр. дело)
1465	keem.	(химия)	353	füsiol.	(физиология)
1442	aj.	(история)	343	min.	(минералогия)
1269	el.	(электричество)	308	kirikl.	(церковное)
1191	ehit.	(строительство)	286	astr.	(астрономия)
1117	anat.	(анатомия)	272	folkl.	(фольклор)
1093	sõj.	(военное дело)	262	fot.	(фотография)
1024	lgv.	(лингвистика)	227	mäend.	(горное дело)
990	füüs.	(физика)	194	teatr.	(театр)
805	kõnek.	(разговорное)	192	meteor.	(метеорология)
799	van.	(устарелое)	182	kal.	(рыбоводство)
744	mat.	(математика)	173	filos.	(философия)
672	piltl.	(фигуральное)	134	ped.	(педагогика)
659	biol.	(биология)	132	arheol.	(археология)
634	geol.	(геология)	129	psühh.	(психология)
625	tekst.	(текстильное дело)	122	bibl.	(библиография)
623	muus.	(музыка)	92	loog.	(логика)
611	murd.	(диалектное)	74	paleont.	(палеонтология)
607	aiand.	(садоводство)	63	müt.	(мифология)
605	geogr.	(география)	57	vulg.	(вульгарное)
595	pol.	(политика)	49	antr.	(антропология)
575	jur.	(юриспруденция)	49	lastek.	(детское)
546	kok.	(кулинария)	36	nalj.	(шутливое)
507	mer.	(морское дело)	21	luulek.	(поэтическое)
			8	halv.	(пренебрежительное)

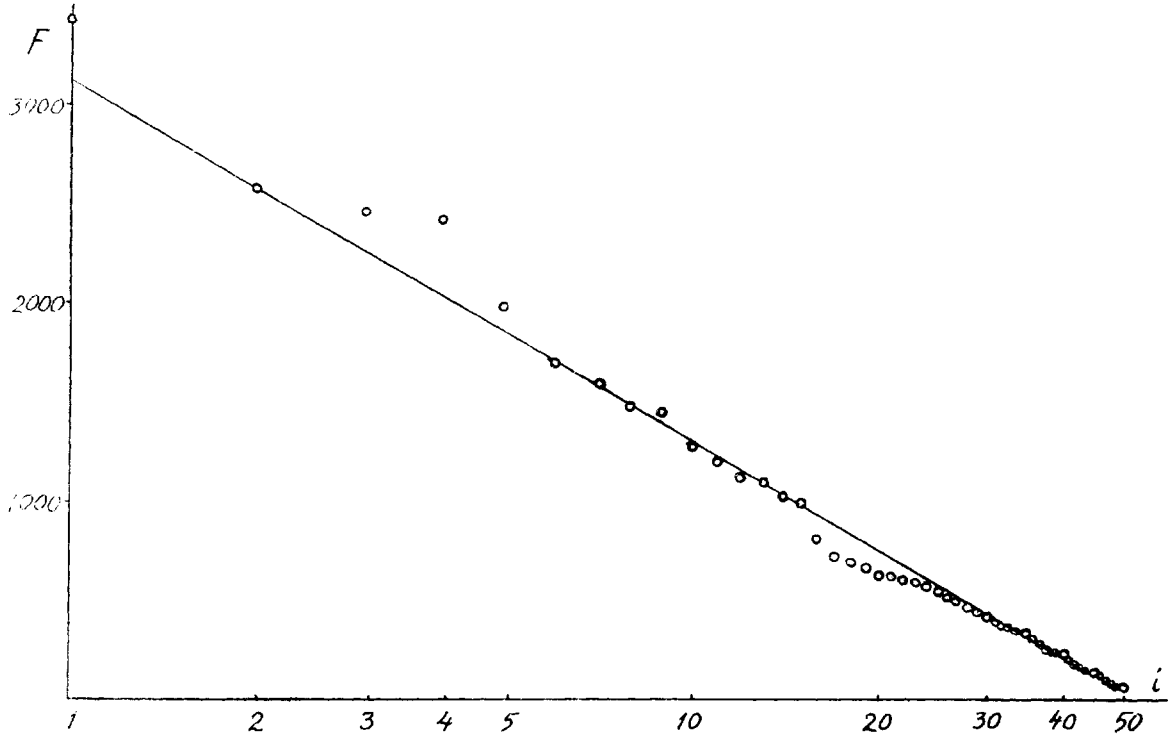


Рис. 2. Ранговое распределение групп специальной лексики в Ортологическом словаре эстонского языка (F - частота пометы, i - ранг). По оси абсцисс логарифмический масштаб.

варе встречаемость помет специальной лексики (39 025) составляет 92,7 % общей частотности помет (42 083), а во всем словаре специальная лексика занимает около 30 % (с учетом омонимии). Специальная лексика обильно представлена и в пробной тетради толкового словаря эстонского языка, но имеется некоторое отличие в распределении диалектизмов, архаизмов и разговорной лексики по сравнению с ортологическим словарем. В табл. 3 приводятся обзорные данные о распределении отмеченной лексики в эстонских словарях сравнительно со словарями русского языка.*

Таблица 3

Распределение стилистически отмеченных лексических единиц в словарях эстонского и русского языков (в%)

Словарь / Лексика	Ортол.сл. эст.яз.	Толк.сл. эст.яз.	Сл.совр.рус. лит. яз.	Сл.рус.яз. Ожегова
Специальная л.	92,7	72,8	?	17,0
Диалектизмы	1,5	2,9	3,7	1,8
Разговорная л.	1,9	6,8	38,4	33,9
Просторечие	-	-	24,6	9,3
Архаизмы	1,9	6,2	?	13,5
Прочие	2,0	11,3	?	24,5
Всего (%)	100,0	100,0	100,0	100,0
Общее число помет	42083	646 (выборка)	3925 (выборка)	17003

Рассмотрим основные группы отмеченной лексики подробнее.

Среди слов специальной лексики наибольшую частоту в ортологическом словаре имеют термины с пометой "техника", которые встречаются в словаре 3433 раза (8,8 % общей встречаемости терминов). К этой группе относятся слова, обозначающие вещества, материалы, оборудование и т.п., которые имеют применение в различных отраслях современной промышленности, например, masuut 'мазут', valukoda 'литейная', valtsmasin 'вальцовый станок'; сюда же относятся

* О распределении отмеченной лексики в "Словаре современного русского литературного языка" см. Филин Ф.П., 1970, и в "Словаре русского языка" С.И. Ожегова см. Денисов П.Н., Костомаров В.Г., 1970.

наименования различных технологических процессов, например, kuumtöötlemine 'горячая обработка'. Некоторые из этих терминов можно, по-видимому, отнести к слою "общетехнической" терминологии (agregaat 'агрегат', reaktiivne 'реактивный' и т.п.).* В ортологическом словаре многочисленно представлены еще термины с пометами "ботаника" (6,6 %), "зоология" (6,3 %), "медицина" (6,2 %), "сельское хозяйство" (5,I %), "экономика" (4,4 %) и др. (см. табл. 2).Концентрация частых терминов большая: десять наиболее частотных групп покрывают 52,2 % отмеченной специальной лексики (48,5 % всех отмеченных слов).

Всю совокупность терминов в ортологическом словаре можно условно сгруппировать по трем основным разделам науки (см. классификацию наук в БСЭ, т. I7, с. 330). См. табл. 4.

Таблица 4

Распределение специальной лексики в ортологическом словаре эстонского языка по основным разделам науки

Раздел	Число помет	Частота терминов число	%
Технические науки	I8	I7 65I	45,2
Естественные науки	I4	I2 497	32,0
Общественные науки	20	8 877	22,8
Всего	52	39 025	I00,0

Согласно принятой классификации к терминам технических (практических) наук мы относим слова с пометами "техника" (8,8 % всей совокупности терминов), "электричество,электротехника" (3,3 %), "строительство" (3,I %), "военное дело" (2,8 %), а также "медицина" (6,2 %), "сельское хозяйство" (5,I %), "ветеринария" (I,I %), "спорт" (4,I %); "садоводство", "лесоводство","морское дело" и др. Термины этой группы составляют вместе взятыми почти половину всех отмеченных слов специальной лексики.

К терминам естественных наук относятся слова с пометами "ботаника" (6,6 %), "зоология" (6,3 %), "химия" (3,8 %), "анатомия" (2,9 %), "физика" (2,6 %), "биология" (I,7 %);

* О статистическом методе разделения терминов на общетехнические (или общенаучные) и специальные см. Пиотровский Р.Г., Истребова С.В., I969.

"геология", "география", "физиология", "минералогия", "астрономия", "метеорология", "антропология". К этой же группе причисляется "математика" (I,9 %).

Раздел общественных наук охватывает историю, экономику, политику, филологию, искусство, а также философию. Наиболее частотными в этой группе являются термины с пометами "экономика" (4,4 %), "история" (3,7 %), "лингвистика" (2,6 %), музыка (I,6 %) и политика (I,5 %).

Встает вопрос, в какой мере представление специальной лексики в ортологическом словаре отражает социальный вес и актуальность соответствующих разделов науки в наши дни. Представляется, что актуальность наук в определенной степени можно измерить по объему публикаций по соответствующим разделам в определенный период времени. Для этого была просмотрена "Статистика печати Эстонской ССР за 1980 год" и составлена соответствующая таблица (см. табл. 5).

Таблица 5

Выпуск книг в Эстонской ССР в 1980 г.
(исключая художественную и справочную литературу)

Раздел	Книг на эст. яз. число	%	Печатн. листов число	%
Техника и промышленность	462	42,9	4130	37,0
Естествознание	154	14,3	1620	14,5
Общественные науки	460	42,8	5420	48,5
В с е г о	1076	100,0	11170	100,0

Сравнивая таблицы 4 и 5 приходится констатировать, что главным различием является выдвижение публикаций в области общественных наук, словарные пометы которых по частотности были на последнем месте. Противоречие оказывается все же мнимым, так как надо учесть, что специальная лексика общественных наук имеет большое распространение (она встречается ежедневно в прессе, радиопередачах, телевидении и т.д.), в результате чего многие термины общественных наук в наши дни перешли в разряд общеупотребительных (общелитературных) слов и в словаре уже не отмечаются специальной пометой. Здесь сказывается принцип связи между частотностью и известностью слов.

Доля специальной лексики в толковом словаре эстонского языка (по данным пробной тетради) несколько меньше, чем в

ортологическом, но все же значительная. Из всех отмеченных позиций 72,8 % (470 из 646) относятся к терминологии, а во всей выборке они составляют 21,3 %. Эти числа значительно больше, чем соответствующие данные словарей русского языка (см. табл. 3); например, в "Словаре русского языка" С.И.Ожегова специальная лексика занимает 17,0 % всех отмеченных позиций и лишь около 3,5 % всех позиций в словаре. Обилие специальной лексики в эстонских словарях является, конечно, рефлексией языковой ситуации наших дней, но можно все же поставить вопрос о рациональном объеме терминологической лексики в составе общих нормативных словарей, учитывая, что уже имеются десятки специальных терминологических словарей по разным отраслям науки и техники.* На будущее планируется издание ортологического словаря нового типа, где объем терминологической лексики значительно сократится и будет расширена лингвистическая информация (см. Ahven E., 1976, с. 37).

Особый слой специальной лексики составляют разговорно-терминологические и разговорно-профессиональные слова, которые ввиду своей "ненормативности" и ограниченности сферы употребления выходят за пределы литературного языка. Если они представлены в нормативном словаре, то они обычно снабжены пометой "разговорное". К специальной лексике относятся также номенклатуры, т.е. системные совокупности наименований профессий (Erelt T., 1982, с. 116). Наиболее употребительные наименования профессий приводятся в ортологическом словаре на общих основаниях с другими терминами (например, eetööline 'забойщик' - с пометой "горное дело"). На периферии специальной лексики находятся т.наз. товарные знаки - "прагмонимы", которые в эстонском языке (как и в других языках) во многих случаях представляют собой семантически мотивированные производные или сложные слова (см. Komolova Z., Tuldava J., 1974). Прагмонимы в нормативных словарях не приводятся.

Диалектная лексика в словарях эстонского литературного языка представлено довольно скудно. Безусловно, это отражает факт резкого уменьшения роли диалектов в наше время. Еще в середине прошлого века в большом

* Только за период 1955-1980 гг. в Эстонии опубликовано более 20 (в основном двуязычных) терминологических словарей с общим числом терминов около 200 тыс. (Erelt T., 1982, с. 145-146). В Институте языка и литературы АН ЭССР ведется интенсивное научное исследование вопросов теории и практики терминологии (см., например, Kull R., Saari H., 1975; Саари Х., 1981).

эстонско-немецком словаре Ф. Видеманна (Wiedemann F. J., 1869), объемом 50 тыс. слов, доля отмеченных эстонских диалектных слов составила 25 %. Хотя прямое сравнение словаря Ф. Видеманна с ортологическим словарем невозможно, т.к. первый был задуман как словарь всего этноязыка (т.е. всех социальных разновидностей языка), все же здесь в известной степени нашел отражение тот факт, что в период формирования национального литературного языка диалекты играют еще значительную роль в общеязыковой ситуации. Конечно, и в наши дни диалекты полностью не утратили своего значения как средства общения некоторых слоев населения, но в период зрелости литературного языка значение диалектов уменьшается, и сами диалектные различия постепенно сглаживаются (см. Rätsep H., 1976, с. 123). В настоящее время отбор диалектных слов в словари литературного языка решается критерием минимально-достаточной употребительности: за основу берется встречаемость слова у 2–3 разных писателей (см. Eesti kirjakeele sõnaraamat, 1969, с. 5). Надобность включения диалектизмов в нормативные словари уменьшается и потому, что уже существуют или готовятся к печати специальные словари диалектов.*

В ортологическом словаре диалектная лексика отмечена соответствующей пометой 611 раз, а это составляет лишь 1,5 % общей частотности отмеченных слов. Во всем словаре диалектизмы покрывают около 0,5 %. В пробной тетради толкового словаря эстонского языка соответствующие величины – 2,9 и 0,9% (в переводе на "позиции"). Сравнение со словарями русского языка показывает, что и там диалектно окрашенные слова относительно мало представлены (от 1,8 до 3,7 % всех отмеченных слов, см. табл. 3, и от 0,4 до 1 % во всей выборке). ** Для сравнения можно привести данные о толковом словаре грузинского языка, где лексические диалектизмы составляют более 7 % всех слов (Партенадзе М.Х., 1980, с. 91).

Среди диалектизмов, приведенных в нормативных словарях эстонского языка, преобладают имена существительные, прилагательные и глаголы. Например, в ортологическом словаре (по

* В институте языка и литературы АН ЭССР готовится к печати шеститомный словарь эстонских диалектов (на основе картотеки, охватывающей более 2 миллионов слов). Опубликован "Краткий диалектный словарь" (Väike murdesõnastik, 1982–1983), насчитывающий около 50 тыс. заглавных слов.

** По последним, более точным данным, во всем "Словаре современного русского языка" из 120 480 [illegible] 2435 являются диалектно окрашенными (2,0 %), причем "полными" диалектизмами являются 1453 единицы, или 1,2 % (см. Партенадзе М.Х., 1980).

данным выборки, состоящей из I4 200 слов на I00 страницах, выбранных с помощи таблицы случайных чисел) существительные составляют 60 % (всех диалектизмов), прилагательные - 20 %, глаголы - I5 %, остальные - 5 %. Тематически диалектные слова - это в основном наименования предметов и явлений сельского быта (около 35 % всех диалектизмов), а также наименования растений и животных (I5 %). Встречаются и слова, обозначающие разного рода действия или свойства. В большинстве случаев к диалектным словам можно подобрать соответствия (синонимы) из общего языка, например, диал. labu (= общеупотр. heinasaad 'копна сена'), taosed (= rangid 'хомут'); lember (= remmelgas 'ракита'), ätse (= karikakar 'пупавник'); letu (= jänes 'заяц'), küü (= vaskuss 'медяница'); hõdras (= habras 'хрупкий'). Но встречаются и диалектные слова (около 20 %), которые выражают какие-то специфические явления, не имеющие однословного обозначения в общем языке, например, soonetis '(родниковое) трудно проходимое место', soop 'весной и осенью затопляемая земля', õtsk 'болотистый берег'; õsuma 'жать серпом'.

Диалектные слова попадают в состав лексики литературного языка через сферу общего языка, т.е. в зависимости от степени проникновения в слой общеупотребительной лексики (принцип употребительности). Можно добавить, что во многих случаях переходным этапом является т. наз. полудиалект - речь, переходная от местных диалектов к общему языку и, далее, к правильному литературному языку.

Что касается лексики социальных диалектов (разговорно-жаргонные слова и арготизмы), то эта разновидность словарного состава эстонского языка изучена мало. В нормативных словарях она специально не отмечается, но отдельные слова, имеющие помету "разговорное", можно, по существу, отнести к пласту жаргонизмов (например, inter (= internaat 'общежитие', tekkel 'студенческая фуражка' - в студенческом жаргоне).

Понятие разговорной лексики требует особого рассмотрения. В прямом смысле слова разговорная лексика означает всю совокупность слов, которые фактически используются в повседневном бытовом и трудовом общении, т.е. в разговорной речи. Однако более распространенной является трактовка "разговорной лексики" как особого пласта слов, характерных (специфичных) только при преимущественно для разговорной речи. Можно добавить, что разго-

ворная речь обнаруживается не только в устной форме, но она может встречаться и в письменном виде (в частных письмах, записках и т.д.)*. Кроме того, надо учесть, что разговорная речь по своему распространению охватывает все основные сферы употребления языка (общий язык, диалекты, подъязыки науки и техники) и поэтому необходимо различать разновидности общеязыковой, диалектной и профессиональной разговорной речи. В данном случае остановимся на некоторых проблемах общеязыковой разговорной речи и соответствующей лексики в эстонском языке.

В рамках общего языка разговорную речь (эст. kõnekeel) можно подразделить на следующие основные подвиды: литературно-разговорную речь, обиходно-бытовую речь и просторечие.** Соответственно различаются и отдельные слои специфичной разговорной лексики. Но так как литературно-разговорная лексика по существу относится к нормативному литературному языку, то собственно "разговорными" считаются лишь специфичные слова обиходно-бытовой речи или просторечия. В ортологическом яловаре эстонского языка все такие слова отмечены стилистической пометой "разговорное", причем не делается разницы между обиходно-разговорными и просторечными словами (хотя некоторые грубые просторечные слова встречаются с пометой "вульгарное"). Как видно из сопоставления данных разных словарей (см. табл. 3), в эстонских нормативных словарях отмеченная разговорная лексика занимает весьма незначительное место: в ортологическом словаре 1,9 % из всех отмеченных слов (0,7 % всех слов словаря), в толковом словаре (в переводе на позиции) - 6,8 % (2,0 %). На фоне этих данных особенно выделяются показатели Словаря современного русского литературного языка: лексика с пометой "разговорное" составляет 38,4 % всех отмеченных позиций, с пометой "просторечное" - 24,6 % (по отношению ко всем позициям в словаре,

---

* О лингвостатистических характеристиках русской эпистолярной речи (лексики частных писем) см. Григорьева А. С., 1981.

** Если принять тезис о различении стилистики языка и стилистики речи (причем под языком будем понимать потенцию, а под речью - реализацию; см. Тулдава Ю., 1981), то при более строгом подходе следовало бы различать и понятия "литературно-разговорный язык" - "литературно-разговорная речь", "обиходно-бытовой язык" - "обиходно-бытовая речь", а также "просторечие" на уровне языка - на уровне речи. Для эстонского языка предлагаем следующую стилистическую градацию на двух уровнях: normkeel - argikeel - madalkeel; normkõne - argikõne - madalkõne.

соответственно, 9,7 и 6,2 %). Доля разговорных и просторечных слов значительна также в словаре С. И. Ожегова. Большая разница между данными словарей эстонского и русского языков свидетельствует, по-видимому, не только о различной трактовке некоторых проблем определения границ разговорной лексики, но и о том, что в эстонском языке вне конкретного текста социально-культурная градация стилистических характеристик слов (принадлежность к высокому, нейтральному, низкому стилю) и различие между обиходно-бытовой речью и литературной речью не всегда столь отчетливое, как в русском языке. Некоторые трудности в разграничении стилистических пластов лексики в лексикографической работе объясняются также недостаточной разработанностью проблем стилистики эстонского языка.

С точки зрения стилистической оценки можно отметить двойственную природу "разговорных" слов: они составляют стилистический пласт экспрессивных ("сниженных", "юмористических" и т.п.) элементов в рамках литературного языка, но в своей собственной среде - в бытовом общении - они могут быть вполне "нейтральными". Судя по данным выборки, сделанной на основе отмеченных разговорных слов в ортологическом словаре, приблизительно 50 % слов являются именно такими нейтральными или мало экспрессивными, в частности народные названия растений, веществ и т.д. (põdrasammal 'олений мох', hundisavi 'синяя глина' и др.) и фонетические или морфологические варианты слов, которые в той или другой степени отличаются от литературной нормы (misukene - сокращение от missugune 'какой'; ärme = литер. ärgem 'пусть не'; takka = литер. tagant 'сзади' и др.). Обращает на себя внимание то, что среди отмеченных разговорных слов много заимствований, главным образом из немецкого[x] и русского языков (klaar 'ясный', krell 'яркий', krempel 'хлам', vanderdama 'путешествовать'; junts 'юнец; молокосос', kruusa 'грузовик', kruttima 'крутить' и т.д.). Все же надо отметить, что среди немецких заимствований многие слова находятся в стадии вымирания (krihv 'рукоятка', krehvtine 'острый, пряный' и др.).

Около 35 % специфичных разговорных слов отличаются ярко выраженной экспрессивностью, проявляющейся в мотивации (внутренней форме) сложных и производных слов. Например, valevorst 'враль, врун' (буквально: "лже-колбаса"), sopajoodik 'пьянчуга' ("грязь-пьяница"), näolapp 'мордочка' ("лицо-

[x] Результат тесных контактов с прибалтийскими немцами (до 1940 г.).

лоскут"), **inimesetükk** 'человечишко' ("человек-кусок"), **ahnepäits** 'жадюга' ("жадная голова"), **jõmmkärakas** 'толстяк, тумба' ("толстяк-хлопок"), **kirvenägu** 'уродина'("топорное лицо"). Сюда же относятся производные слова с некоторыми характерными суффиксами и концовками, например, **kõlakas** 'слушок', **ohmu** 'простофиля', и сокращения типа **kriminull** 'детектив (детективный роман)'. Подобные сложные,производные и сокращенные слова ("грубовато-экспрессивная лексика") можно считать одним из наиболее типичных видов разговорного (особенно окказионального) словообразования.

Немало (15 %) среди разговорных слов ономатопоэтических или дескриптивных образований, например, **krõnks** (сухочарный старый человек), **krõbi** (дряхлый старый человек), **jõrris** 'злюдий', **jukerdama** или **jupsima** 'возиться'.

Доля разговорной лексики в толковом словаре эстонского языка оказывается несколько больше, чем в ортологическом словаре, главным образом, по той причине, что всегда учитываются случаи, когда у литературного слова может быть особое значение в разговорной речи. По данным пробной тетради толкового словаря, около половины всех позиций, снабженных пометой "разговорное", оказывается только отдельными значениями какого-нибудь слова. Например, слово **esmaabi** 'первая (медицинская) помощь' употребляется в разговорной речи также в значении "скорая помощь"; слово **habe** 'борода' употребляется в значении 'бородатый человек'; одно из значений глагола **hakkama** 'начинать' в разговорной речи - 'подходить,быть впору' (например: **sinine värv hakkab talle hästi**'синий цвет идет ей хорошо'). Надо отметить, что в ортологическом словаре в отдельных случаях отмечается двойное использование слова, например, в отношении слова **punt** 'связка; букет' отмечается, что в разговорной речи оно может иметь значение **salk** ('ватага'). Все это свидетельствует о том, что разговорной речи свойственна "повышенная вариативность лексики" (Филин Ф.П., 1973, с. 9), в частности семантические и стилистические сдвиги в значениях слов.

По сравнению с разговорной лексикой в эстонских нормативных словарях в количественном отношении примерно в таком же объеме представлены а р х а и з м ы - слова с пометой "устарелое" (см. табл. 3). Здесь имеются в виду слова, которые встречаются в более старом литературном языке, в частности в произведениях эстонских классиков, хотя в настоящее время эти слова уже вышли из употребления. Например,

teal (= siin'здесь'), **esipuhku** (= **esialgu** 'сначала'), **jumalavits** (в значении: **nuhtlus**; **õnnetus** 'кара; несчастье'). Среди отмеченных в эстонских словарях архаизмов встречается множество устарелых терминов, названий профессий, учреждений и др. Можно отметить интересный факт, что именно в терминологических названиях во многих случаях изначальные эстонские образования оказались устарелыми и они заменены интернациональными терминами, например, **inimeseteadus** (=**antropoloogia** 'антропология'), **vaimuteadus** (= **humanitaarteadus** 'гуманитарная наука'), **õppetool** (= **kateeder** 'кафедра'), **ergukava** (= **närvisüsteem** 'нервная система'), **kirjatoimetaja** (= **sekretär** 'секретарь'), **halastejaõde** (= **meditsiiniõde**'медсестра'). Архаизмами считаются также (по данным словарей) старинные эстонский названия месяцев: **küünlakuu** (= **veebruar** 'февраль'), **heinakuu** (= **juuli** 'июль') и др. С другой стороны, отдельные иностранные слова-термины оказались устарелыми, например, **dežuur** (= **valvekord** 'дежурство'), **gümnastika** (= **võimlemine** 'гимнастика'), **ökonoom** (= **majapidaja**'эконом'). В числе устарелых слов довольно много старых заимствований в области ежедневной жизни, в частности заимствований из немецкого языка, например, **rehkendama** (= **arvutama** 'вычислять, считать'), **valsk** (= **võlts**, **valelik** 'фальшивый, лживый'). Встречаются и "грамматические" архаизмы, особенно словообразовательные варианты: **andeline** (= **andekas** 'даровитый'), **esiotsalt** (= **esiotsa** 'сначала'), **järgemisi** (= **järgemööda**'по очереди'), **haigemaja** (= **haigla** 'больница') и др., а также отдельные семантические архаизмы, например, слово **halb** 'плохой' в одном из значений 'дешевый' (по данным толкового словаря). Многие из отмеченных устарелых слов примыкают к "историзмам", т.е. терминам, которые употребляются преимущественно в исторических исследованиях или в художественных произведениях для изображения исторического колорита прежних времен, например, **junkur** в значении 'управляющий имением'; **konka** 'конка, конно-железная дорога'. Собственно историзмы приводятся, как правило, с пометой "историческое" и относятся к специальной лексике.

В лексической системе эстонского языка свое место занимают и н е о л о г и з м ы , но они в нормативных словарях специально не отмечаются. Вопрос о неологизмах требует специального исследования. В особом рассмотрении нуждается и проблема дифференциации лексики по экспрессивно-эмоциональному признаку.

Заключение. На фоне общественных сфер употребления языка выделяются основные социально-функциональные группы эстонской лексики : неограниченная местностью или родом деятельности общеупотребительная лексика и группы слов ограниченного употребления - диалектная и специальная лексики. Особое положение занимает литературная лексика, которая покрывает определенную часть общеупотребительной и специальной лексик. Констатируется "размытость" границ между отдельными группами - компонентами общей системы лексики - и необходимость определения границ и объемов групп лексики с помощью вероятностно-статистических критериев в сочетании с качественным анализом. С точки зрения квантитативно-системного подхода к изучению лексики особый интерес представляют взаимные отношения между отдельными группами и подгруппами, а также статистические распределения единиц и классов единиц. В данной работе анализ проводится в основном на материале репрезентативных словарей эстонского языка. Таким образом, лексика рассматривается только на уровне языка (лексикона, инвентаря). Анализ функционирования лексики в речи, т.е. в текстах, требует особого рассмотрения и применения других методов.

Л И Т Е Р А Т У Р А

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

Березин Ф.М., Головин Б.Н. Общее языкознание. - М.: Просвещение, 1979.

БСЭ = Большая советская энциклопедия. 3-е изд. Том 17. - М.: Советская энциклопедия, 1974.

Григорьева А.С. Статистическая структура русского эпистолярного текста (лексика частных писем). АКД. Л., 1981.

Гроссе Р. О соотношении языка и нации. - Иностранные языки в школе, 1970, № 3, с. 2-10.

Денисов П.Н., Костомаров В.Г. Стилистическая дифференциация лексики и проблема разговорной речи. - В кн.: Русская разговорная речь. - Саратов: Изд-во Саратовского ун-та, 1970, с. 69-75.

Калинин А.В. Лексика русского языка. - М.: Изд-во МГУ, 1978.

Кожина М.Н. Стилистика русского языка. - М.: Просвещение, 1977.

Негуляев Г.А., Покрас Ю.Л., Колесников Л.И. Автоматизированный отбор лексики для информационно-поисковых тезаурусов. - Научно-техническая информация. Серия 2. М., 1973, № 2, с. 16-24.

Ожегов С.И. Словарь русского языка. Изд. 5-е. М., 1963.

Партенадзе М.Х. Об основаниях отбора диалектных слов для словарей литературных языков. - Вопросы языкознания, 1980, № 2, с. 87-98.

Пиотровский Р.Г. Лингвостатистический эффект Бектаева. - Материалы семинара "Статистическая оптимизация преподавания языков и инженерная лингвистика". Чимкент, 1980, с. 25-26.

Пиотровский Р.Г., Истребова С.В. Статистическое опознание термина. - В кн.: Статистика текста. Том I. - Минск: Изд-во БГУ, 1969, с. 249-259.

Саари Х.М. Анализ принципов эстонской терминологии. АКД. Тарту, 1981.

Словарь современного русского литературного языка. Т. I-17. М.-Л., 1948-1965.

Статистика печати Эстонской ССР 1980. - Таллин: Ээсти Раамат, 1981.

Тулдава Ю. О теоретико-методологических основах квантитативно-системного анализа лексики (2): лингвистические аспекты исследования. - Учен. зап. Тартуск. ун-та, вып. 585. **Linguistica XIV.** Тарту, 1981, с. 114-133.

Тулдава Ю. Квантитативное исследование генетического состава лексики эстонского языка. - Учен. зап. Тартуск. ун-та, вып. 628. Труды по лингвостатистике. Тарту, 1982, с. 136-166.

Филин Ф.П. О структуре современного русского литературного языка. - Вопросы языкознания, 1973, № 2, с. 5-12.

Филин Ф.П. О свойствах и границах литературного языка. - Вопросы языкознания, 1975, № 5, с. 3-12.

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

Щерба Л.В. Избранные работы по русскому языку. М., 1957.

Ahven E. Keele ja Kirjanduse Instituudi keelepool. - Rmt. : Keel, mida me harime. - Tallinn: Valgus, 1976, lk.33-44.

Eesti kirjakeele sõnaraamat. Makett. - Tallinn: Valgus, 1969.

Erelt T. Eesti oskuskeel. - Tallinn: Valgus, 1982.

Kaasik U., Tuldava J., Villup A., Ääremaa K. Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik. TRÜ Toimetised, vihik 413. Töid keelestatistika alalt II. Tartu, 1977, lk. 5-140.

Komolova Z., Tuldava J. Estonian Brand Names. - In: Linguistica V. Tartu, 1974, pp. 74-90.

Kull R., Saari H. Die Entwicklung des estnischen terminologischen Gedankens und die Förderung der Fachsprachen in den letzten Jahren. - In: Congressus tertius internationalis Fenno-ugristarum. Tallinn, 1975, S. 245-249.

Mackay A. On the Type-count of the Phaistos Disc. - Statistical Methods in Linguistics, No 4. Stockholm: Skriptor, 1965, pp. 15-25.

Raiet E. "Eesti kirjakeele sõnaraamat" ilmumisjärgus. - Keel ja Kirjandus, 1983, nr. 3, lk. 134-136.

Rätsep H. Murdekeel ja kirjakeel. - Rmt: Keel, mida me harime. -Tallinn: Valgus, 1976, lk. 121-124.

Rätsep H. Some Tendencies in the Development of Estonian. - Soviet Finno-Ugric Studies, 1981, No 3, pp. 202-211.

Saukkonen P., Haipus M., Niemikorpi A., Sulkala H. Suomen kielen taajuussanasto. A Frequency Dictionary of Finnish. Porvoo - Helsinki - Juva, 1979.

Tuldava J. Eesti keele sõnavara foneetilis-grafeemilised mõõted.- TRÜ Toimetised, vihik 518. Töid keelestatistika alalt V. Tartu, 1980, lk. 51-98.

Valge J. Ajalehekeele sõnavara statistiline analüüs. TRÜ diplomitöö. Juhendaja H. Rätsep. Tartu, 1972.

Väike murdesõnastik I-II. Toimetanud V. Pall. - Tallinn: Valgus, 1982-1983.

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

Õigekeelsussõnaraamat./ Toimetanud R. Kull ja E. Raiet. -Tallinn: Valgus, 1976.

THE SOCIAL DIFFERENTIATION OF THE LEXIS OF THE ESTONIAN LANGUAGE FROM A QUANTITATIVE POINT OF VIEW

Juhan Tuldava

S u m m a r y

In this article an attempt has been made to provide a quantitative analysis of the Estonian lexis according to the main spheres of the social usage of language: the common (overall) language, territorial and social dialects, and the language of science and technology. The conception of the literary language as an intermediate phenomenon (cf. Fig. 1) has been elaborated. Futher, the so-called functional styles are discussed. On the basis of the normative Estonian dictionaries an analysis of the distribution of terminology, dialectal words, colloquial lexis, and archaisms in the language has been made. Of the total volume of 115,000 words in the " Orthological Dictionary " (1976) about 30% are provided with stylistic labels mainly pointing to various fields of terminology (Tables 2 and 4). About 70% or 80,000 words may be considered as the bulk of the literary lexis of the contemporary Estonian language (excluding terminology and nonce word-building).

The article belongs to a series written by the author dealing with some general quantitative-typological characteristics of the Estonian language. The previous article - on the genetic structure of the Estonian vocabulary - was published in ACUT vol. 623, 1982.

СОДЕРЖАНИЕ

SUMMARIES

CORRIGENDA

В посвящение моей статьи в предыдущем выпуске "Трудов по лингвостатистике" (Ученые записки ТГУ, вып. 628. Тарту, 1982) вкралась досадная опечатка - ошибочно указано отчество Игоря Владимировича Рахманова. Приношу редколлегии и читателям мои извинения за этот недосмотр и прошу обладателей книги внести соответствующее исправление в текст посвящения.

С.И. Гиндин

Ученые записки Тартуского государственного университета.
Выпуск 658.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И СТИЛИСТИКА.
Труды по лингвостатистике.
На русском языке.
Резюме на английском языке.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, I8.
Ответственный редактор М. Соонтак.
Подписано к печати I.II.1983.
МВ I0537.
Формат 60x90/I6.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов I0,8I.
Печатных листов II,25.
Тираж 500.
Заказ № II60.
Цена I руб. 60 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Пялсона, I4.