Цена 33 коп.

В. В. Александров, А. В. Арсентьева, А. И. Семенков

СТРУКТУРНЫЙ АНАЛИЗ ДИАЛОГА

Ленинград

АКАДЕМИЯ НАУК СССР
ЛЕНИНГРАДСКИЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

Препринт № 80

В.В.Александров, А.В.Арсентьева, А.И.Семенков

СТРУКТУРНЫЙ АНАЛИЗ ДИАЛОГА

Ленинград
1983

УДК.681.142.4/519.24

Аннотация

В препринте обсуждаются вопросы эффективной организации диалогового взаимодействия с ЭВМ. Рассматривается проблема структурной организации проблемно-ориентированных языков. Путем сопоставительного анализа рангового распределения частот появления слов в различных текстах исследуются особенности построения текстов различного вида. Анализируются количественные характеристики словарного состава функционального ядра языка и проблемно-ориентированной терминологии. Предложен структурный метод анализа диалогов.

СЛНИЦ, 1983

Введение

Большинство достижений, связанных с развитием ЭВМ и расширением их прикладного назначения, значительно превзошли прогнозы. Современная технология обеспечивает быстродействие и объем памяти, а также экономические показатели значительно выше тех, о которых мечтали специалисты два десятилетия назад. Но одна проблема, связанная с вычислительной техникой (ВТ), которая считалась не самой сложной, не только не решена до сих пор, но более того, даже сейчас нет достаточно четкого понимания того, как ее следует решать. Речь идет о диалоге с ЭВМ удобным для пользователя образом. Не случайно в предыдущей фразе не было использовано выражение"на естественном языке". Это выражение, во-первых, довольно расплывчато, и , во-вторых, слишком часто используется в связи с ЭВМ без достаточных оснований.

Вообще говоря, анализ диалога довольно сложен из-за подвижности языка, его индивидуальности и изменчивости во времени. Кроме того, вероятно, не любой диалог можно считать эталоном для его воспроизведения на ЭВМ. Тем не менее мы имеем достаточное количество хороших образцов диалога для исследования. Человечество располагает огромным количеством прекрасно структурированных систем диалога, поскольку при определенных допущениях можно считать, что любой художественный текст вступает в диалог с читателем,как только он "погружается" в текст.

Первые шаги в исследовании свойств языка были сделаны совместными усилиями лингвистов и специалистов по ВТ. Большинство работ этого направления посвящено прежде всего формальному выявлению семантики языковых конструкций. Данный препринт посвящен несколько иному аспекту исследования языков в приложении к ЭВМ. Мы сделали попытку применить математические методы для исследования структурных особенностей ряда текстов с целью выявления закономерностей и принципиальных различий в построении диалогов,ориентированных на человека и на ЭВМ. Мы полагаем, что такое исследование поможет прояснить, какой же диалог с ЭВМ мы хотим иметь. Актуальность такого рода исследований подтверждается появлением в ВТ новых

терминов: friendly dialogue, soft methodology of dialogue , означающих комфортность и естественность работы специалиста в диалоге с ЭВМ.

Мы полагаем, что,выявив объективные количественные характеристики построения текстов, признанных наиболее эффективно воспринимаемыми собеседником, и проанализировав тенденции в изменении этих характеристик в зависимости от тех или иных функциональных назначений исследуемого текста, мы сможем учесть эти характеристики при построении критерия комфортности диалогового языка. Кроме того, анализ представительной выборки текстов позволит выявить необходимый объем словаря, место специальной терминологии в рамках такого словаря и некоторые другие характеристики диалогового языка, то есть мы будем исследовать, какими общими особенностями обладают текстуальные описания, как соотносится языковая форма описания с содержащимся в тексте смыслом.

Интересно, что ответы на эти вопросы ищутся специалистами в разных областях научных исследований от искусственного интеллекта до литературоведения, несмотря на то, что решается одна и та же задача: как организовать наиболее эффективный способ общения - коммуникационного информационного взаимодействия - между людьми и в человеко-машинных системах. С точки зрения применимости проводимого исследования для организации эффективного диалогового общения между человеком и ЭВМ важно выделить такие особенности построения проблемно-ориентированных текстов, как отношение общего числа слов к "фундаментальному словарю" [1].

Используемая нами методика проведения экспериментальных исследований текстов включает методы кластер-анализа, распознавания образов и принятия решений. В качестве параметров исследуемых текстов взяты коэффициенты функций, аппроксимирующих частотные распределения слов в текстах. Выбор типов текстов для исследования поясняется в разделе 4.Однако при выборе в рамках класса поэтических текстов мы руководствовались довольно субъективными соображениями.

Произведения А.С.Пушкина были взяты по двум причинам:

I. Словарь А.С.Пушкина - один из самых богатых словарей, и именно для него интересно было определить количественные характеристики той его части, которая позволила Пушкину создавать исключительно емкие по смыслу произведения.

2. 3I октября I983 года исполнилось I50 лет со дня создания Петербургской поэмы "Медный всадник".

Поэма "Медный всадник" является одной из самых загадочных поэм. Это произведение неоднократно анализировалось литературоведами разных лет, попытавшимися выявить семантическую особенность произведения. Мы же со своей стороны решились применить методы математики для выявления структурных особенностей этого гениального произведения.

В данном препринте проблемы организации диалога рассматриваются с различных сторон , но все они иллюстрируются общим набором экспериментальных данных.

Взяв в качестве исходного данного для эксперимента поэму-загадку "Медный всадник", мы, разумеется, не смогли избежать соблазна проанализировать ее и с точки зрения семантических особенностей.

"Следовать за мыслями великого человека есть наука самая занимательная"...

А.С.Пушкин

"Чем дальше, тем искусство становится более научным, а наука-более художественной; расставшись у основания, они встретятся когда-нибудь на вершине".

Г.Флобер

I. Структура, словарь и "естественность" языков.

При решении проблемы поиска комфортного языка общения с ЭВМ напрашивается очевидное решение - строить диалог на естественном языке. Оставим пока в стороне вопрос о целесообразности этого решения и посмотрим, что же мы понимаем под естественным языком.

Любой язык характеризуется двумя основными признаками: словарным составом и структурной организацией слов в текстах. Исследованию словарного состава текстов и законам развития знания посвящены многие работы Ципфа [2,3] . В этих работах показано, что рост общего знания, накопленного обществом, подчиняется определенному закону (огибающая на рис.I). На отдельных этапах накопления знания происходит его дробление на ряд проблемно-ориентированных областей[5].

В различных областях деятельности человека в рамках некоторого учреждения, организации, клуба и т.д. появляется ограниченный и приспособленный язык, так называемый профессиональный язык. Он развивается и используется людьми для профессиональных целей, т.е. для решения профессиональных задач в специфической области и для сообщения о них.

Характерными признаками проблемно-ориентированного знания служат, во-первых, специфические формы представления знаний своей предметной области (ПО), например, формулы в математике, экспедиционные дневники в геологии, таблицы с экспериментальными данными в физике, анкеты в социологии и т.п. и,

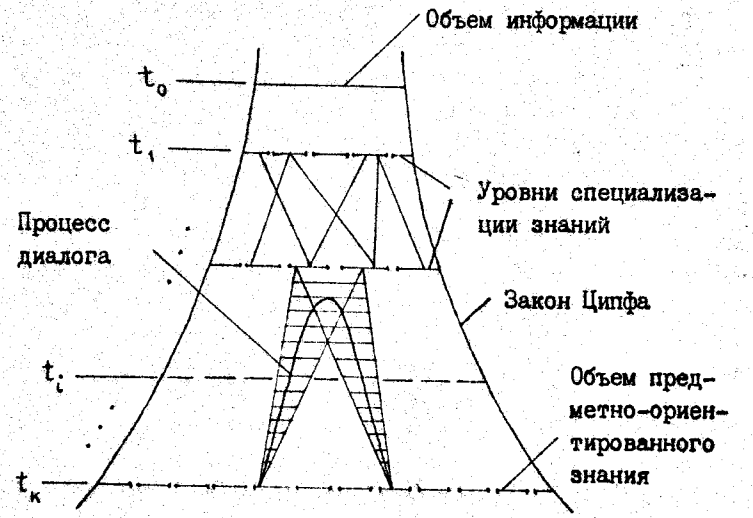


Рис.I. Процесс развития знаний.

во-вторых, собственный тезаурус. В процессе жизнедеятельности динамической системы знаний ее тезаурус может расти и сокращаться в объеме, а также перемещаться в информационном пространстве. Первый признак является следствием стремления применить наиболее эффективный для данной ПО формально-логический аппарат обработки знаний. Вторая - отражает необходимость семантической обработки и носит прагматический характер. Оба эти признака направлены на увеличение скорости обработки информации, т.е. на увеличение эффективности исследований и коммуникации в рамках данной ПО [5]. Таким образом, появление различных форм представления проблемно-ориентированных знаний и их тезаурусов есть следствие стремления к повышению скорости взаимодействия исследователей внутри своей ПО.

Приведенная иллюстрация закона Ципфа несколько упрощена. На самом деле элементы каждого яруса включаются в линейное

связи разной протяженности, обеспечивая тем самым контакты между различными иерархиями. Элементы высшего яруса по отношению к предыдущему низшему обладают интегративным характером. Принципы вертикальной и горизонтальной упорядоченности элементов взаимодействуют друг с другом.

Научная деятельность базируется прежде всего на развитии проблемно-ориентированного знания. В то же время при решении ряда задач требуются знания смежных дисциплин. В этом случае необходимо согласовать системы понятий, представленных в различных формах. Для согласования форм представления проблемно-ориентированных знаний и тезаурусов необходимо обратиться к предыдущему уровню представления знаний (см. рис.I), где понятия рассматриваемой и смежной ПО имеют единую семантику и, следовательно, возможно диалоговое сопоставление понятий специалистами разных ПО. Знания предыдущего уровня выполняют роль "транслятора" специфических форм представления знаний на более развитом уровне, а язык предыдущего уровня является "естественным языком" относительно "специализированных языков" следующего уровня, обеспечивающим согласование между ними. Следовательно, понятия "универсальной формы" представления знаний и "естественного языка" имеют относительный характер; роль общих знаний относительно проблемно-ориентированных состоит в обеспечении согласования различных форм представления знаний и тезаурусов проблемно-ориентированных ПО.

Знания каждой ПО, в свою очередь, развиваются и со временем становятся общими по отношению к своим подразделам. Рост проблемно-ориентированного знания также подчиняется закону Ципфа, при этом крутизна огибающей может быть различна для разных ПО и разных этапов его развития.

Таким образом, словарь "естественного языка" есть некая абстракция. В реальном мире мы пользуемся некоторым подмножеством словарного состава естественного языка.

Приведенные в [1] данные о структуре словаря "естественного языка" наглядно характеризуют так называемые "общую" и "узкоспециальные" его части.

Что касается структурной связи слов в текстах на некото-

ром языке, то эта область значительно менее исследована. Традиционные описания в виде грамматических правил при обращении к ЭВМ показали свою явную несостоятельность. В лучшем случае они удовлетворительно используются при анализе существующего текста. На наличие в законченных текстах структуры, которая может характеризоваться количественными оценками, обратили внимание Эсту (J.Estoup) и Ципф (G.Zipf) [4]. Анализируя частотные распределения слов в законченных текстах, они нашли функции, аппроксимирующие эти распределения. Их работы интересны именно тем, что позволяют объективными количественными параметрами характеризовать структуру законченного текста и отличать его от фрагментов или суммы текстов, для которых найденные аппроксимирующие функции не годятся. Таким образом, наше интуитивное восприятие гармонии, "законченности" текста, комфортности его восприятия имеет в своей основе некоторый объективный закон, который в полной мере еще не ясен, но проявляется через вполне детерминированные частотные распределения слов. Поскольку любой разработчик диалоговой информационной системы (ИС) является творцом языка, то перед ним естественно возникают вопросы выбора словаря и структурной организации диалогов на базе выбранного словаря.

ИС, в основе которых используются ЭВМ, представляют собой новое направление с еще не сложившимися традициями и методологией. Однако уже наметилась устойчивая тенденция "гуманизации" технических систем, т.е. такой их организации, при которой человек является активным действующим лицом, вступающим в информационное взаимодействие с другими людьми посредством ИС. В этом случае ИС должны представлять собой удобный, а главное, "понятный" инструмент общения. "Понятный" в данном контексте означает, что человек, работая со своей информацией, полностью отдает себе отчет в том, какие изменения она претерпевает при использовании ЭВМ. Поэтому целью научной методологии создания и использования ИС является достижение взаимопонимания и эффективной коммуникации между людьми посредством ИС. При этом следует учитывать, что человек во всех своих действиях стремится к некоторой цели, а сама ИС сущест-

вует в рамках предметной области и неразрывно связана с формальным представлением ПО.

Участник любого диалога производит активное реконструирование сообщения. Интерпретация зиждется на предварительных установках: ожидании, концепциях, значениях, знании языка и т.д. Коммуникация нацелена на взаимопонимание субъектов. Однако она невозможна, если до акта коммуникации между субъектами не было понимания. Взаимопонимание должно основываться на знании как языковых правил, так и соответствующей области деятельности. Поэтому при использовании ЭВМ в качестве посредника в диалоге язык общения вынужденно дополняется правилами и некоторым множеством слов, "понятных" ЭВМ. Мы, таким образом, имеем, с одной стороны, словарь и правила порождения текстов на его основе для ЭВМ, а с другой стороны, словарь и правила порождения текстов предметной области вне зависимости от ЭВМ.

Каким образом следует соединить эти компоненты, чтобы максимально адаптировать диалог к пользователю, т.е. чтобы использование ЭВМ в качестве посредника в диалоге минимально искажало привычный способ общения между специалистами данной ПО?

Те знания, которыми располагают в настоящее время лингвисты и специалисты по ВТ, не позволяют дать однозначного ответа на этот вопрос. Большинство существующих на сегодняшний день диалоговых систем имеют в своей основе интуитивное стремление либо к минимизации словаря и жесткой структуре порождаемых текстов, либо, напротив, накоплению огромного словаря, который за счет использования синонимов создает иллюзию "естественности".

Данная работа, разумеется, также не содержит окончательного ответа на этот вопрос. Однако проведенные эксперименты позволяют составить представления о некоторых количественных закономерностях в текстах на "естественном языке" и сопоставить их с аналогичными характеристиками для "искусственных языков". Термин "искусственный язык", вероятно, не совсем удачен, мы его используем за неимением устойчивой терминологии как антоним к "естественному языку" и включаем в класс "искусственных языков" как формализованные проблемно-ориентированные языки, так и языки программирования, т.е. те языки, для которых правила порождения текстов из слов заданы вполне детерминированно человеком. Примером текстов на "искусственных языках" могут служить формальные математические выводы, описание химических реакций, программы для ЭВМ и т.д.

В экспериментальных исследованиях текстов разного характера мы попытались выявить закономерности в словарном и структурном решении диалогов на разных языках: естественных языках поэзии и прозы, языках программирования и формализованных языках предметной области. Во всех случаях были взяты примеры "хороших" текстов, т.е. текстов, удовлетворяющих комфортному восприятию адресатом.

2. Лингвистика, семантика и принятие решений.

Используемые при анализе текстов художественных произведений методы, как правило, допускают большую долю субъективности при выборе той или иной гипотезы. В ряде случаев это просто связано с отсутствием удобного математического (или какого-либо иного формального) аппарата, пригодного для анализа художественных произведений. Так, например, при трактовке смысла художественного произведения зачастую отдается предпочтение одной версии перед другой на основании произвольно выбранных разрозненных фактов, что серьезно противоречит методам теории принятия решений.

В теории принятия решений при выборе одной из нескольких гипотез все основано на следующих двух тривиальных утверждениях:

- достоверный вывод может быть сделан только на достоверных фактах;
- если достоверные факты не противоречат сразу нескольким гипотезам, то выбор одной из них неправомочен: имеют право на существование все гипотезы.

На примере поэмы-загадки А.С.Пушкина "Медный Всадник" перечислим основные гипотезы, альтернативный выбор из которых обычно пытаются обосновать.

I. "Медный Всадник" является поэмой, форма, содержание и внутренний смысл которой составляют единое целое произведение, определяющее отношение поэта к Петру, Петербургу и описание наводнения 1824 года и связанную с этим наводнением трагическую судьбу Евгения.

2. Поэма имеет второй смысловой ряд, который вложен логически внутрь текста и определяет основной смысл поэмы как памфлет против Николая I и напоминания о необходимости смягчения участи участников декабрьского восстания. В этом случае форма и содержание выступают как оболочка относительно внутреннего смысла поэмы.

3. Форма, содержание и внутренний смысл поэмы отражают сложное психологическое состояние поэта в 1833 году и длительный период работы над историческими тенденциями развития России. Параллельность работы над "Историей восстания Пугачева" и "Историей Петра I" вызвала суггестивный образ двойственного отношения к монархии, определяющий как судьбу государства, так и судьбы потомственных дворян.

Чтобы провести объективный (насколько это возможно) анализ этих альтернативных гипотез, необходимо их проанализировать на одном и том же наборе фактов. Выбор набора фактов, разумеется, тоже может привнести субъективность, но она будет тем меньше, чем больший объем фактов, относящийся к произведению, будет рассмотрен и чем ближе эти факты в пространственно-временном отношении к периоду работы над ним. В следующем разделе приведен семиотический подход к анализу художественного произведения с целью выбора достоверной гипотезы о замысле поэта при создании поэмы-загадки.

Однако существующие на сегодняшний день математические методы выявления структурного построения произведения на основании только объективных количественных оценок могут также быть полезны при его анализе.

Вторая гипотеза - наличие второго смыслового ряда в поэме - требует наличия в поэме двух самостоятельных текстов, отражающих двойной смысл. А это значит, что частотное распределение слов внутри законченного в смысловом отношении текста должно отличаться от распределения в составленном из двух отдельных в смысловом отношении текстов, сложная композиция из которых представлена в виде одного текста. Более того, как правило, основные используемые словари отдельных текстов имеют зону пересечения, т.е. вероятность того, что словари двух отдельных в смысловом отношении законченных текстов полностью включатся друг в друга,очень мала.

Существование одновременно первой и второй гипотез не требует никаких доказательств, т.к. их появление и определяется спонтанно-интуитивной работой поэта.

Исследование же зависимостей частного распределения слов в тексте при сознательной длительной работе над художественным произведением или же при интуитивно-суггестивном спонтанном творчестве находится за рамками данной работы.

3. Семиотические методы в анализе художественного текста.

Развитие семиотического подхода [6] и приложение его в гуманитарных науках позволяет взглянуть на традиционные объекты исследований в области лингвистики как на сложно организованные системы.

Кроме того, тексты, приближающиеся по своим характеристикам к художественным, стали не только средством общения между людьми, но и важным элементом человеко-машинной коммуникации, где они выступают в качестве одного из объектов обработки. Включение таких текстов как единиц обработки в человеко-машинную коммуникацию делает необходимым разработку формальных методов их анализа и способов различного рода их преобразований.

Анализ существующих методов показывает, что формализация текстов осуществляется на уровне языкового и логического компонентов текста. В то же время задачи человеко-машинной коммуникации необходимо решать с учетом семантического компонента текста [7].

В настоящее время специалисты в различных областях рассматривают культуру как сложную иерархическую систему, которая обеспечивает:

I. Обмен сообщениями двух видов:
   - внутренний между членами данного социума;
   - внешний между социумом и Природой;
2. Хранение сообщений - Память.
3. Выработку новых организованных форм порождения сообщений [8].

Под сообщением понимается подмножество множества всех конечных последовательностей элементов словаря или словарей, если сообщение является композицией последовательностей на различных языках, с определенной на нем ассоциативной и некоммутативной бинарной операцией конкатенации.

Назовем текстом сообщение любого вида, которое возникает в системе информационного взаимодействия между членами данного социума, хранится в ее памяти и (или) активно используется в коммуникационных актах.

С точки зрения семиотики художественное произведение может рассматриваться как текст или знак. В наиболее простом случае под текстом понимается последовательность знаков, а под знаком - структура, включающая означаемую и означающую стороны знака. Для семиотического исследования существенно то, что текст может выступать как единое целое, не членящееся на отдельные самостоятельные знаки, хотя при этом не членимый на знаки текст может разлагаться на отдельные компоненты - блоки [9].

К первому разделу культуры как иерархической системы относится все связанное с описанием языков культуры как элементов информационного взаимодействия, системы циркуляции текстов внутри данной культуры, разных систем интерпретирования, осмысления текстов в пределах той или иной системы информационного взаимодействия.

Ко второму разделу - проблема "Культура как память данного социума", типы организации этой коллективной памяти, ее эволюция и самооптимизация [10].

Третий раздел рассматривает сложный феномен постоянного семиотического самообновления информационной модели внешнего мира (ИМВМ), которая формируется на основе глобальной историко-социальной сферы знаний, выработки ею новых языков и механизмов возникновения новых организационных форм порождения текстов.

Нас интересует первый раздел, а именно - тезаурусы, языки и интерпретации текстов, циркулирующих внутри рассматриваемой ИМВМ. При таком подходе нас не может не заинтересовать одна особенность ИМВМ - наличие огромного числа текстов - произведений искусства, участвующих во множестве коммуникационных актов.

Художественный текст как знак представляет собой сложную структуру, которая, например, в естественном языке описывается как состоящая из нескольких уровней. Наибольший интерес представляет соотношение между различными уровнями внутри знаковой структуры. Эти уровни, по которым могут быть расклассифицированы различные художественные средства, распо-

лагаются между замыслом, представляющим собой высший уровень или означаемую сторону художественного текста, и его конечным воплощением в последовательность сигналов, воспринимаемых органами чувств - означающей стороной или низшим уровнем художественного текста.

Знаковый характер текста предполагает наличие не меньше чем двух таких уровней, обязательно не совпадающих друг с другом. В простейших случаях означаемая сторона художественного текста может быть отождествлена с концептом (смыслом), означающая сторона - с денотатом (конкретным текстом). Кроме того, художественный текст является таким объектом в коммуникационном акте, который, вступая в обратные связи с различной аудиторией, изменяет каждый раз возможную свою интерпретацию в соответствии с данной индивидуальной ситуацией. Этому соответствует не только исключительно емкий и компактный объем памяти художественного текста, но и способность бесконечно варьировать сообщение на выходе в зависимости от запросов адресата.

В основе изучения коммуникационных процессов в иерархической системе культуры чаще всего лежит классическая модель Якобсона [11]:

АДРЕСАНТ — СООБЩЕНИЯ / ТЕКСТ — АДРЕСАТ

Согласно этой модели коммуникационная ситуация, тезаурусы, интерпретационные механизмы адресанта и адресата одинаковы, т.е. коммуникационный канал в этой простейшей модели симметричен. Следствием из этого положения является то, что в идеальных условиях объем сообщения и его интерпретация (осмысление) одинаковы для обеих сторон коммуникационного канала. Изучение реальных коммуникационных процессов приводит к убеждению, что эта модель явно неполна. Одним из частных, но показательных случаев является коммуникационный канал при неодинаковых словарях C_1 у адресанта и C_2 у адресата. Даже при одинаковых грамматиках, порождающих язык общения в процессе коммуникации, несимметрия канала коммуникации при-

водит к сдвигу семантики в процессе интерпретации текста адресатом или же к изменению семантики. При сдвиге семантики образуются новые, не предусмотренные адресантом значения текста, поскольку адресат интерпретирует текст с помощью только общей части словаря адресата и адресанта. Модель коммуникационного процесса имеет в этом случае вид:

АДРЕСАНТ - ТЕКСТ I - $C_1 \diamond C_2$ - ТЕКСТ 2 - АДРЕСАТ

Так как в ходе передачи по каналу ТЕКСТ I трансформируется в ТЕКСТ 2 из-за несимметрии коммуникационного канала, то нас интересует процесс этой трансформации - самого массового и, одновременно, самого сложного явления в системе коммуникации.

Трансформация - это отображение текста в текст, отображение не однозначное, но позволяющее провести преобразование ТЕКСТ → СМЫСЛ.

Под смыслом в данном случае следует понимать ту информацию, которая извлекается из текста при его восприятии, причем "информация" понимается как некоторое знание. Определение смысла текста в целом требует выхода за рамки самого текста и обращения к знанию ИМВМ, образующего в мышлении как бы "внутренний" текст. При конкретизации понятия смысла следует подчеркнуть, что смысл необходимо отождествлять с новым знанием. Так как всякий текст содержит новое знание только в контексте конкретной ИМВМ, через комбинацию элементов которой это новое знание только и может быть выражено, то соотношение между ИМВМ адресанта и адресата становится наиболее важным компонентом для понимания трансформации текста.

В рассматриваемой ситуации механика трансформации текста сводится к тому, что текст, подлежащий передаче, синтезируется и интерпретируется последовательно с помощью двух различных, но пересекающихся словарей C_1 и C_2, причем на уровне семантики пересечение словарей должно давать тождественность смысла в плане содержательного восприятия. В реальности, конечно, все сложнее - каждый текст синтезируется и интерпре-

тируется (осмысляется) на основе существующей в каждый момент времени, пространства и конкретного коммуникационного акта ИМВМ, т.е. не одного, а целого ряда словарей и грамматик языков, причем число их может быть значительно. ИМВМ адресата может иметь ряд пересечений с пространством ИМВМ адресанта, и только в том случае, когда текст и синтезирован и интерпретирован на базе одной и той же области пересечения ИМВМ адресанта и ИМВМ адресата, возможно правильное преобразование ТЕКСТ → СМЫСЛ. Именно из-за того, что ИМВМ динамически изменяется и у каждого адресанта и адресата есть своя ИМВМ, преобразования ТЕКСТ → СМЫСЛ может быть выполнено полностью, приблизительно, или вообще невыполнимо.

Нас интересует влияние на процесс преобразования ТЕКСТ → СМЫСЛ различных факторов и особенности этого преобразования в ситуациях межличностной коммуникации и диалога человек-ЭВМ.

Анализ реально существующих в культуре видов коммуникаций и текстов позволяет выделить две группы ситуаций:

-Первая группа - это ситуация, когда целью коммуникационного акта является передача конкретных данных. В этих случаях ценность всей системы определяется тем, в какой мере текст - без потерь и искажений - передается от адресанта к адресату. В этом случае всякое несовпадение между словарями адресанта и адресата рассматривается как помеха, а текст является пассивным носителем вложенного в него смысла. Трансформации, которым может подвергаться текст в процессе коммуникационного взаимодействия, в этом случае делятся на допустимые и недопустимые. Первые совершаются в соответствии с заложенными в структуре системы коммуникации алгоритмами, и, следовательно, являются обратимыми. К недопустимым трансформациям относятся описки, ошибки, искажения, являющиеся шумом в системе коммуникации, а также все виды искажения смысла при преобразованиях ТЕКСТ → СМЫСЛ. Индивидуальность ИМВМ адресанта и адресата, затрудняющая преобразование ТЕКСТ → СМЫСЛ, также рассматривается как помеха, для уменьшения которой должна быть мобилизована структура языка. Идеальным видом такой коммуникации является передача текста на искусственном языке.

Под искусственным языком здесь понимается такой формализованный язык, который задается как произвольная совокупность специализированных языковых средств с более или менее точно фиксированными правилами образования выражений и приписывания этим выражениям определенного смысла. Все тексты на естественных языках, и особенно на языках искусства, будут в этом случае с точки зрения взаимнооднозначности отображения ТЕКСТ → СМЫСЛ, проводимого в рамках конкретной ИМВМ, выглядеть как "неэффективные".

Вторая группа - это ситуации, когда целью коммуникационного акта является выработка нового знания, т.е. новых текстов. В этом случае ценность системы определяется сдвигом смысла текста в процессе его движения от адресанта к адресату. Нетривиальным сдвигом смысла назовем такой семантический сдвиг, который однозначно не предсказуем и не задан определенным алгоритмом трансформации текста. Текст, который получается в результате такого сдвига, мы и будем называть новым. Возможность образования новых текстов определяется структурой и особенностями ИМВМ адресанта и адресата. В связи с неоднозначностью различий в ИМВМ адресатов один и тот же текст интерпретируется ими различным образом. Проследим это различие на примере интерпретации текста поэмы А.С.Пушкина "Медный Всадник" различными исследователями.

Прежде всего выделим "тривиальную" интерпретацию текста "Медного Всадника": "печальный рассказ" об "ужасной поре" - определение Петербургской повести А.С.Пушкиным. (Рис.2.Блок № I).

Все попытки выделения смысла поэмы, т.е. преобразования ТЕКСТ → СМЫСЛ, легко разделить на две группы [12]. Первая группа основана на выделении смысла из прямых, ничем не прикрытых данных поэмы, и сводится к "тривиальной" интерпретации текста. Назовем это преобразование ТЕКСТ → СМЫСЛ I. Вторая группа учитывает наличие в тексте неявных элементов смысла, затрудняющих понимание его в целом. Так как все исследователи поэмы одинаково производят преобразование ТЕКСТ → СМЫСЛ I, основанное на непосредственных данных произведения, то сосредо-

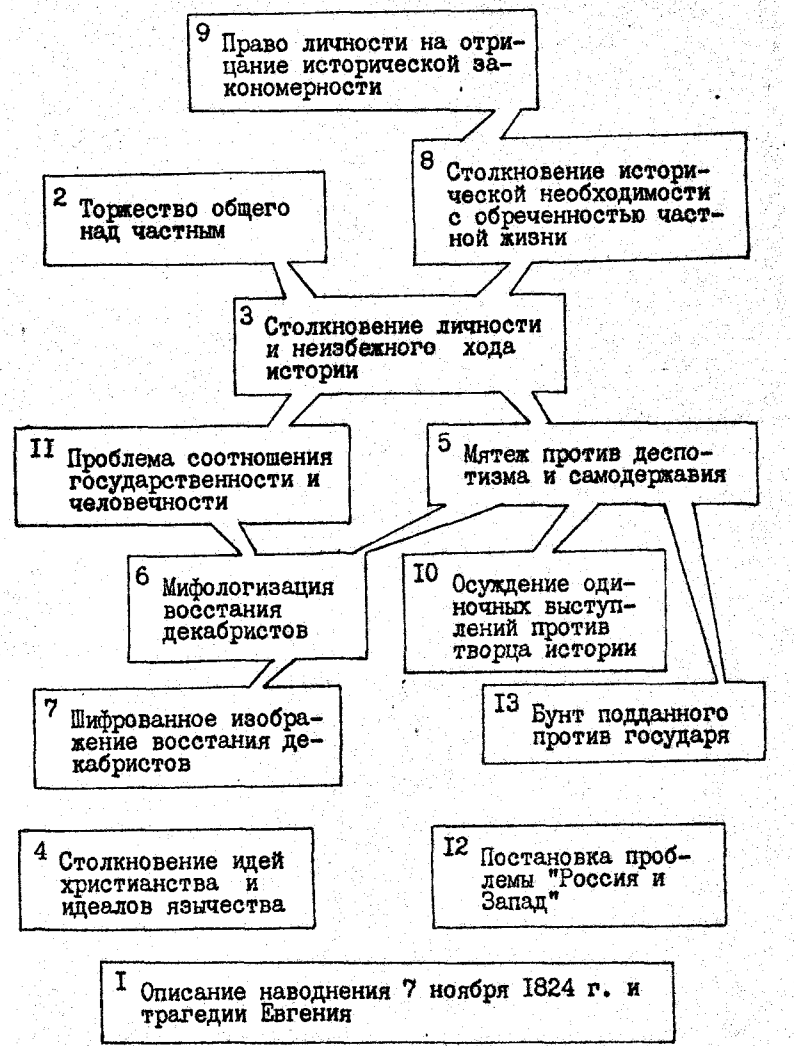


Рис.2. Блок-схема связей интерпретаций исследуемого текста.

точим внимание на преобразованиях ТЕКСТ→СМЫСЛ i, $i = 2,3,\ldots,n$, которые проводятся исследователями независимо от первого преобразования.

Учитывая, что новые тексты, а значит и новые смыслы текста появляются из-за различий ИМВМ адресатов, которые приводят к различным трансформациям одного и того же текста, можно надеяться найти действительный "первоначальный" смысл преобразования СМЫСЛ→ТЕКСТ адресанта на наибольшем пересечении результатов многочисленных преобразований ТЕКСТ→СМЫСЛ. Приведем примеры различных преобразований ТЕКСТ→СМЫСЛ. Результат преобразования ТЕКСТ→СМЫСЛ 2 (рис.2, Блок 2) Белинским был сформулирован в статье [13]: "... и смиренным сердцем признаем мы торжество общего над частным", т.е. "невозможность соединить счастье индивидуальностей и обеспечить участь народа и государства" [14]. Продолжатели Белинского формулируют близкий к СМЫСЛУ 2 СМЫСЛ 3 [14]: "сопоставление коллективной воли ... и воли единичной, ... столкновение личности и неизбежного хода истории" (Блок 3). Мережковский и согласные с ним исследователи выделяют СМЫСЛ 4 [14]: "восстание христианства на идеалы язычества". В.Брюсов интерпретирует текст следующим образом [14]: "мятеж против деспотизма и самодержавия" (Блок 5). Интерпретация Д.Д.Благого в 20-е годы приведена в [15]: "мифологизированное" изображение восстания I4 декабря (Блок 6). А.Белый в 20-е годы поддержал вывод Д.Д.Благого, но интерпретировал текст "Медного Всадника" как "зашифрованное изображение восстания I4 декабря I825 г." [16] (Блок 7). С.М.Бонди считает, что основная мысль поэмы состоит в "столкновении" исторической необходимости с обреченностью частной личной жизни" [17] (Блок 8). И.Л.Бродский утверждает, что "Пушкин в своей поэме признал в диалектике социальной действительности наряду с исторической закономерностью существующего и право на его отрицание" [18] (Блок 9). Л.П.Гроссман интерпретирует в [19] смысл "Медного Всадника" следующим образом: " поэт осуждает все одиночные, не связанные с народом, и, значит, безнадежные политические выступления" (Блок I0). Одна из главных интерпретаций Н.В.Измайлова состоит в

следующем [14]: основная мысль поэмы - "постановка вопроса о государственности и человечности. Этот вопрос требовал и требует решения в пользу человечности, но в пушкинское время государственность настолько подавляла все человеческое, что Пушкин на поставленный в его поэме вопрос не мог найти ответа" (Блок II). М.И.Гиппельсон дает интерпретацию поэмы Пушкина как "постановку проблемы "Россия и Запад", а именно выражение этой мысли в том, что "Петр I скорее поднял Россию на дыбы, чем погнал ее вперед" [20] (Блок I2). Прямых данных о преобразовании ТЕКСТ → СМЫСЛ одного из первых адресатов Пушкина - Николая I-нет, но интерпретация смысла может быть определена косвенно из направленности пометок на цензурном автографе поэмы. Смысл может быть определен следующим образом [14] : "восстание "ничтожного героя" против виновника его трагедии", что представляет собой бунт подданного против государя (Блок I3). На основе этих интерпретаций была составлена общая схема возможных связей результатов преобразований ТЕКСТ → СМЫСЛ i , $i = \overline{1,13}$.

Рассматриваемый текст участвовал в неперечислимом множестве коммуникационных актов и создал множество интерпретаций. Некоторые из них совпадают, некоторые в каком-то смысле пересекаются, а некоторые не имеют пересечений. В связи с тем, что осмысление текста производится адресатами на содержательном уровне, где интеллект человека производит перестройку предметной соотнесенности лексических уровней значений слов и на основе своей ИМВМ находит денотаты, соответствующие конкретному сочетанию слов текста, то для построения связей результатов преобразований ТЕКСТ → СМЫСЛ i, $i = 2,3,\ldots,n$ была построена денотатная структура текста - денотатный граф на базе квазиалгоритма, предложенного А.И.Новиковым [7] . Структура денотатного графа не приводится в работе из-за размеров графа и объемов таблиц связи денотатов и отношений.

Из установленных связей всевозможных интерпретаций исследуемого текста следует, что к интерпретации текста, наиболее близкой к замыслу адресанта, можно отнести преобразование ТЕКСТ → СМЫСЛ 3 - **"Столкновение личности и неизбежного хода истории"**.

Необходимо отметить тот факт, что отсутствует однозначное соответствие между элементами языка и сферой обозначаемого. Это связано с тем, что в основе способности языка выражать новое знание лежит возможность комбинирования знаков, имеющих расплывчатость лексических значений, однако для формализованных языков и текстов на них в определенной предметной области денотатная структура в сжатом виде сохраняет всю содержательную часть текста и может быть использована для повышения адаптируемости диалога человек - ЭВМ к обоим компонентам этой системы.

4. Выбор текстов для анализа.

Выбор текстов определялся следующими соображениями. Во-первых, в набор должны были войти художественные и нехудожественные тексты на "естественном языке", а также тексты на искусственных языках. Перечень текстов с их основными характеристиками приведен в табл. I.

Таблица I.

Нумерация и параметры исследуемых текстов.

Текст	Объем выборки, N	Объем словаря, V
А.С.Пушкин:		
I."Моя родословная";	304	208
2."Езерский";	755	500
3."Домик в Коломне";	I286	772
4."Медный Всадник";	I676	994
5."Кавказский пленник".	2683	I224
6.М.Ю.Лермонтов."Мцыри".	27I0	II80
7."Вступление" в тексте № 4.	379	275
8.Текст № 4 без "Вступления".	I303	720
9.А.Блок."Двенадцать".	88I	5I8
I0.В.В.Маяковский."Флейта-позвоночник".	885	569
II.Частотный словарь русского языка.Под ред. Л.Н.Засориной.	I056382	39268
I2.Текст программы OCENKI-2 (АЛГОЛ)	I978	20I
I3.Текст ИС "ДИАНА" (БЭЙСИК).	I4659	I992
I4.Текст ИС "ДИАНА" без комментариев (БЭЙСИК).	I0247	924
I5.Текст ИС "SITO" (АЛГОЛ).	347I	I48
I6.Текст процедуры MAPON" (ФОРТРАН)	576	77
I7.Диалог с ИС "ДИАНА"	I039	3II

Текст "Медного Всадника" выбран по причинам, указанным во введении. Остальные Пушкинские произведения выбраны так, чтобы анализируемый набор включал те произведения, в которых некоторые литературоведы видят связь с "Медным Всадником" (таковы "Моя родословная", "Езерский"), а также те, которые не имеют этой связи, но написаны в тот же период. Произведения М.Ю. Лермонтова, А.А.Блока, В.В.Маяковского взяты, поскольку эти поэты по общему признанию являются выразителями своих эпох. Частотный словарь русского языка представляет собой пример нехудожественного текста на "естественном языке". Кроме того, этот текст не является целостным в смысловом плане.

В набор текстов на "искусственных языках" входят четыре программы на языках программирования. Один из текстов программ проанализирован с включением комментариев и без них. Тексты программ различаются по объему, назначению (для вычислений и обработки информации) и по языку программирования.

В качестве примера текста на проблемно-ориентированном формализованном языке (иногда такого рода языки называют "ограниченными естественными", на наш взгляд это не совсем удачный термин, поскольку любой человек или любая область знаний используют именно ограниченный естественный язык) взят текст конкретного диалога, проведенного специалистом - медиком с системой диагностики заболеваний органов брюшной полости (ДИАНА). Текст программы этой системы также включен в исследуемый набор.

Художественные тексты на "естественном языке" и диалог на проблемно-ориентированном языке сравнивались как по словарному составу, так и по структурной организации, выраженной через количественные характеристики функций, аппроксимирующих частотные распределения слов.

Анализ текстов на языках программирования имел двоякую цель. С одной стороны, - выявить характеристики "комфортности" трех языков программирования для программиста (в том числе - роль комментариев для восприятия текста программы). С другой стороны, представлял интерес вопрос о наличии корреляции

между структурной организацией текста программы и порождаемым ею диалогом. Словари языка программирования и языка диалога имеют пересечение (рис. 3). При этом та часть словаря ПО, которая составляет его функциональное ядро (т.е. встречается в основном в любом диалоге на языке этой ПО), является, с

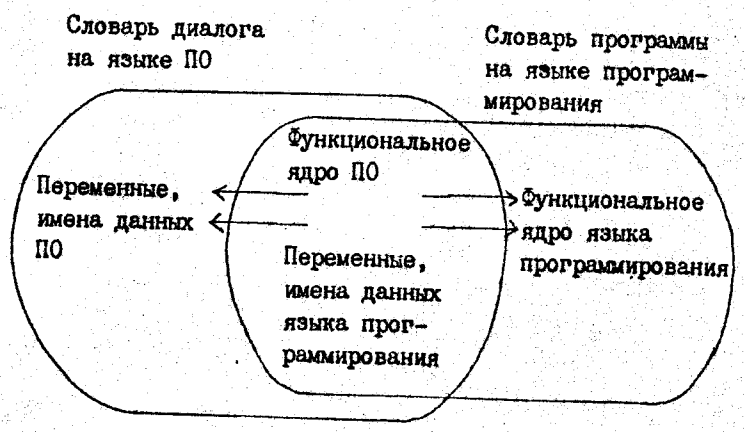


Рис.3. Корреляция между словарем программы и словарем диалога.

другой стороны, переменной частью для словаря программ, написанных на языке программирования. При использовании диалоговой системы на базе ЭВМ в диалог вовлекается несколько лиц. В действительности, опытный специалист своей ПО, назовем его A , разработавший формальный язык ПО, задает все необходимые с его точки зрения связи между словами, включенными им в словарь ПО. Его коллега (B), желающий воспользоваться его знаниями, в идеале хотел бы иметь диалог непосредственно с A . Однако он вынужден иметь посредника в виде ЭВМ. В нашем исследуемом примере речь идет о системе медицинской диагностики. Формальное описание связи симптомов, синдромов, анамнеза с решением о тактике лечения сделано на медицинском языке (подмножестве естественного), затем формально преобразовано к языку предикатов (искусственному). Возникает вопрос, какой язык программирования может с минимальными искажениями обеспечить диалог между A и B . Зависит ли вообще это от языка программирования?

На первый взгляд явной корреляции между языком диалога и языком программирования нет. Более существенное влияние будет оказывать, например, наличие в языке программирования возможности манипулирования со строками. Однако вопрос не столь тривиален. Сейчас явно наметилась тенденция организации программирования без программистов, т.е. создание языков очень высокого уровня, в которых функциональное ядро не будет присутствовать явно. Пока имеются лишь попытки создать такого рода языки, примером может служить APL, функциональное ядро языка уйдет в именованные процедуры, которыми легко сможет манипулировать специалист ПО. В идеальном случае, если имена процедур будут совпадать с терминологией его ПО, то пользователь будет иметь иллюзию, что работает на естественном (разумеется, для данной ПО) языке. Такая ситуация аналогично тому, как читатель, читая художественное произведение, каждое слово ассоциирует (замещает) мысленно его определением. Причем этот процесс может иметь большую глубину рекурсии.

При использовании такого рода языков программирования практически порождающая диалог программа и порожденный диалог будут иметь общий алфавит, т.е. корреляция будет 100%.

Разумеется, реально эта предельная ситуация не была бы эффективной, да и вряд ли возможна вообще, поскольку избыточное укрупнение процедур и передача их в стандартные библиотеки привела бы к чрезмерно большому их количеству, и тогда сам выбор нужной процедуры был бы затруднен. Вероятно, должно быть некоторое промежуточное наиболее удобное соотношение между словарями A и B . Используемая в эксперименте методология может и в этом вопросе дать некоторые количественные оценки.

5. Количественные методы анализа текстов.

Многие исследователи ставили вопрос, какую роль играет частота использования различных слов языка в текстах. Филология рассматривает текст как языковую форму, в которой с помощью приемов, специфичных для данной эпохи и культуры, автор выражает определенную, в какой-то степени целостную систему образов, представлений и взглядов. Телефонная книга, словарь, случайная выборка - не являются текстами в законченном смысловом отношении. Как противоположность перечисленным выше объектам мы выбрали для анализа поэтические тексты, которые сильно насыщены семантическим содержанием.

Текст может быть расчленен на единицы различных уровней: абзацы, предложения, слова, морфемы, фонемы и т.д. В данном исследовании мы проводим членение текста на слова.

Одно из основных наблюдений лингвистики состоит в том, что выделенные в результате членения единицы текста имеют резко различные частоты употребления. Широко известен способ описания разнообразия частот в тексте ранговым распределением, когда частота слова F_n рассматривается как функция от ранга τ -числа слов текста, которые и имеют частоту F_n.

Эмпирически найденные формулы аппроксимируют реальные ранговые распределения. При переходе от одного текста к другому форма таких зависимостей сохраняется с точностью до значения параметров, но частоты отдельных слов могут заметным образом меняться. Основной тезис лингвистики состоит в том, что целостный текст обладает специфической структурой, выражающейся в том, что для законченного, целостного, единого в смысловом отношении текста и только для него выполняется закон распределения частот появления слов текста. Этот эмпирический закон - закон Ципфа - в наиболее общем виде определяется следующим образом.

Пусть имеется набор N элементов. Каждый из элементов снабжен меткой, выбираемой из некоторого множества. Пусть $n(x, N)$ - число различных меток, каждая из которых встречается ровно x раз в выборке из N элементов. Тогда для достаточно больших N имеем следующую эмпирическую

зависимость [21]:

$$n(x,N) = \frac{A}{x^{\gamma}}, \quad \gamma = 1+\alpha, \qquad (1)$$

где A - константа, определяемая, вообще говоря, объемом выборки N; $\gamma = 1+\alpha$ - показатель закона Ципфа; α -характеристический показатель (const.).

Если все метки расположены в ряд в порядке убывания их встречаемости, то величина τ, представляющая положение в этом ряду метки, встречающейся x раз, называется рангом. Ранговое представление закона Ципфа имеет вид:

$$x_\tau = \frac{C}{\tau^{\beta}}, \qquad (2)$$

где $\beta = 1/\alpha$, $C = \alpha B^{1/\alpha}$.

Часто для аппроксимации реальных частотных распределений используется формула, предложенная Б.Мандельбротом

$$x_\tau = \frac{C}{(B+\tau)^{\beta}}, \qquad (3)$$

где C, B и β — const.

Закон Ципфа проверен на самом различном эмпирическом материале, связанном с человеческой деятельностью. Огромное количество самых различных подтверждений закона Ципфа, полученных известными статистиками (Юлом, Кенделлом, Лоткой и др.), приведено как в книге Ципфа [2], так и в обширной литературе, посвященной этому закону.

Исследование структурных особенностей словарных множеств рассматриваемого набора текстов, использующее анализ коэффициентов функций, аппроксимирующих расчетные ранговые распределения частот появления слов в тексте (РРЧС), приводит к необходимости поиска функций такого вида, для которых отклонения теоретической кривой от расчетных РРЧС были бы минимальны. В то же время эти функции должны быть достаточно простыми и быть сравнимыми с формулами закона Ципфа и Ципфа-Мандельброта.

Характерной особенностью набора поэтических текстов является то, что аппроксимация расчетных РРЧС законами Ципфа и Ципфа-Мандельброта не удовлетворяет условиям наилучшей аппроксимации по всем точкам РРЧС.

Так как дальнейшее исследование проводится на основе данных о множестве численных значений коэффициентов аппроксимирующих функций, то естественным и логичным является выбор в качестве аппроксимирующих функций простейших функций типа равносторонней гиперболы $xy = K$ или параболы $y = Kx^2$ Для приведения кривых аппроксимирующих функций в область расположения точек расчетных РРЧС, задаваемых в логарифмическом масштабе, необходимо определить коэффициенты функций, соответствующие наложению кривых этих функций на множество точек РРЧС, с учетом соответствия (совпадения) по характерным точкам. Коэффициент K выбирается из условия наилучшей аппроксимации.

В работе в качестве аппроксимирующих функций рассматриваются обе ветви равносторонней гиперболы $(x-\alpha)(y-\beta) = K$ Коэффициент K взят равным 50 для максимального приближения кривой аппроксимирующей функции к множеству точек расчетных РРЧС корпуса поэтических текстов.

На графиках расчетных РРЧС можно выделить две характерные точки: (см. рис.4).

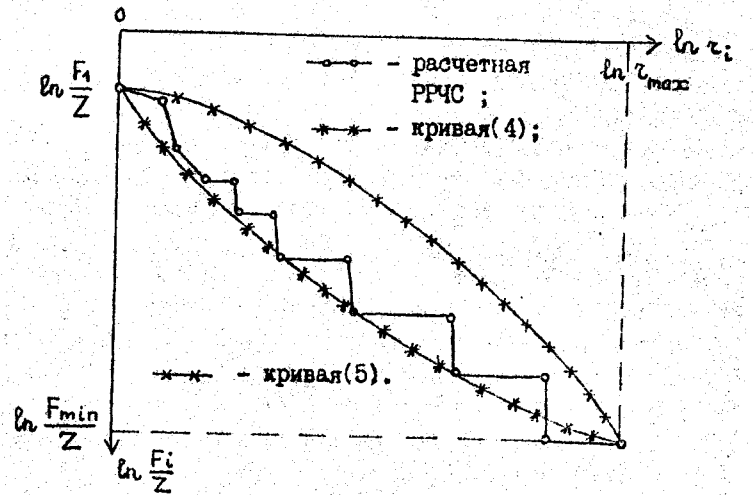


Рис.4. Пример расчетной РРЧС и аппроксимирующих кривых (4) и (5).

I. При ранге $r = 1$ (на оси абсцисс $\ln r_i = 0$) на оси ординат $\ln F_i / Z = \ln F_1 / Z$;

2. При ранге $r = r_{max}$ (на оси абсцисс $\ln r_i = \ln r_{max}$) на оси ординат $\ln F_i / Z = \ln 1 / Z$, так как $F_{min} = 1$.

Коэффициенты аппроксимирующих кривых

$$\ln \frac{F_i}{Z} = \alpha + \frac{50}{\ln r_i - \beta} \qquad (4)$$

и

$$\ln \frac{F_i}{Z} = \frac{50}{\ln r_i + \beta} - \alpha \qquad (5)$$

определяются таким образом, чтобы аппроксимирующие кривые проходили через характерные точки расчетных РРЧС. Для рассматриваемой аппроксимирующей кривой в координатах x, y имеем

$$(y-\alpha)(x-\beta) = 50 \qquad (6)$$

и две точки (x_1, y_1), $x_1 = 0$, $y_1 < 0$, (x_2, y_2), $x_2 > 0$, $y_2 < 0$. Тогда

$$\beta_{1,2} = \frac{x_2}{2} \pm \sqrt{\frac{x_2^2}{4} + \frac{100 x_2}{y_1 - y_2}} \qquad (7)$$

и

$$\alpha_{1,2} = \frac{50}{\beta_{1,2}} + y_1 \qquad (8)$$

Все графики РРЧС строятся в логарифмическом масштабе. По оси абсцисс откладывается $\ln r$ - натуральный логарифм ранга слова, а по оси ординат откладывается натуральный логарифм относительной частоты встречаемости слова в тексте F / Z , где Z - объем текста.

В таблице I(см. приложение I) представлены исходные данные для проведения кластер-анализа (автоматической классификации с помощью системы обработки данных SITO), которые состоят из коэффициентов, рассчитанных для каждого из текстов по аппроксимирующим функциям вида (2)-(5), и данных об объеме каждого текста и соответствующем объеме словаря.

6. Структурный метод анализа текстов.

Рассмотренные в предыдущем разделе приведенные функции (2-5) в той или иной степени аппроксимируют реальное распределение слов в текстах. На рис. 5 приведены все четыре вида аппроксимации текста № 4.

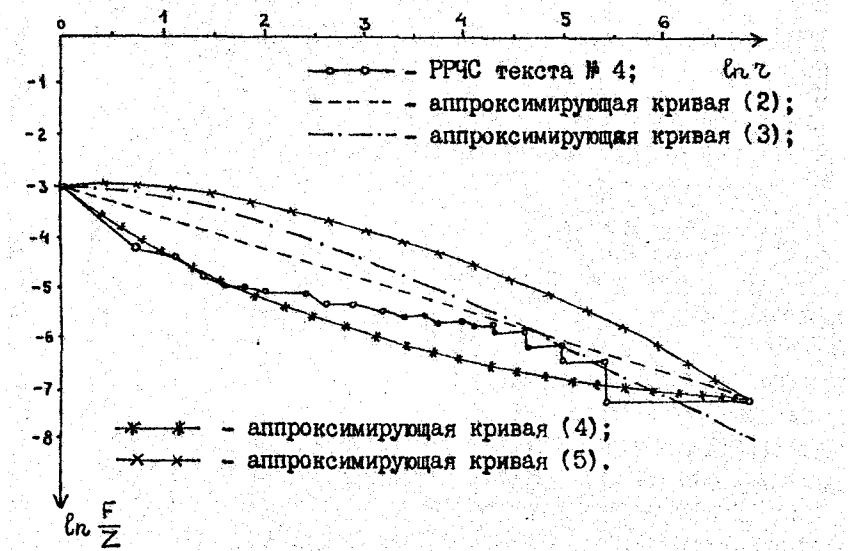


Рис.5. Расчетное ранговое распределение частот появления слов текста № 4 ("Медный Всадник") и графики аппроксимирующих функций.

Поставим следующий вопрос. Будет ли устойчивой группировка текстов или она будет различаться при использовании параметров разных аппроксимирующих функций? Поскольку в параметрах аппроксимирующих функций в той или иной форме присутствуют основные характеристики текста: объем алфавита, общее количество слов, максимальная частота появления слова, то устойчивость группировки свидетельствует о достаточности этих характеристик для

классификации.

Так как любой текст (более подробно, что понимается под текстом в литературоведении, см. [22], а в вычислительной технике-[23]) есть конечная выборка, а исследования связаны с анализом различных частотных распределений слов и их совокупностей в текстах, то оценку достоверности получаемых результатов мы будем основывать не на подходе авторов [24,25], основанном на модификации закона Ципфа и различных поправках к объему выборки, а на основе структурных методов обработки данных [26].

В этом случае в качестве данных будут использоваться количественные параметры различных функциональных зависимостей, аппроксимирующих ранговые распределения частот появления в текстах (РРЧС), приведенных выше.

Последовательность шагов эксперимента приведена в приложении 2 для одного вида аппроксимирующей функции (4). Результаты иерархической кластер-процедуры приведены на рис.7. Из рис. 7 видно, что в зависимости от "подробности" классификации, т.е. от шага классификации возможны более "тонкие" или "грубые" объединения текстов в классы, что может трактоваться как некоторый индивидуальный почерк автора. Достоверность и объективность проведенной классификации следует из того, что различные параметры различных функций, аппроксимирующих РРЧС, при классификации дают сходные результаты при разных уровнях этой классификации. Например, текст № 11 – Частотный словарь – не имеет отношения к поэтическому тексту ни при одной аппроксимации, ни при одном уровне классификации не образует устойчивой группы с другими текстами. Это обстоятельство позволяет продемонстрировать, что выбор различных аппроксимирующих функций и алгоритмов группировки позволяет усилить достоверность результатов анализа без увеличения представительности выборки при их совместном использовании. Следовательно, полученные группировки текстов имеют объективный характер и могут служить достоверным фактом для анализа гипотез о семантике текстов, а также о сходстве и различии в текстах разных типов.

Исходным материалом для анализа являлись ранговые распределения частот появления слов в тексте, часть из которых изображена в виде графиков на рис.6.

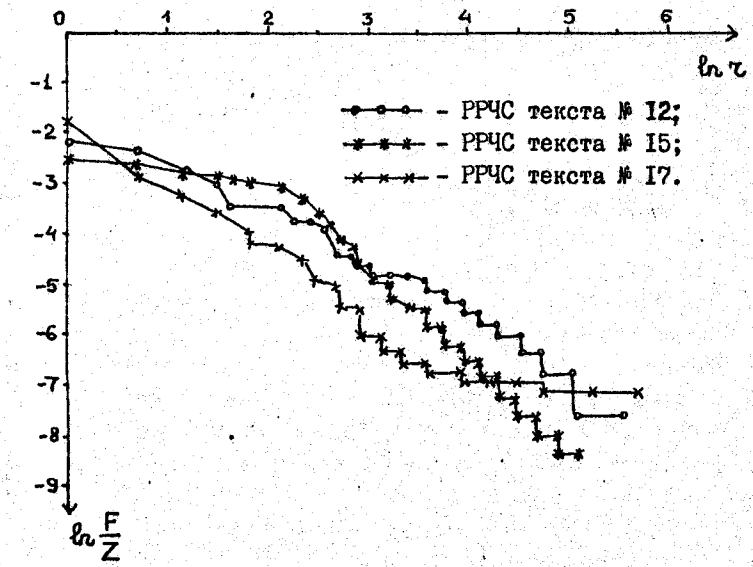


Рис.6. Ранговые распределения частот появления слов в текстах № I2, № I5, № I7.

Для использования структурного метода анализа из частотного распределения слов в тексте необходимо выделить основные параметры, с помощью которых они могут быть аппроксимированы. Исходная таблица данных для классификации приведена в приложении I.

"Номер объекта" соответствует номеру текста в таблице 1. Номера вертикальных столбцов таблицы 1 соответствуют следующим коэффициентам аппроксимирующих функций: I - нормированному коэффициенту A_1 функции (4); 2 - нормированному коэффициенту B_1 этой же функции; 3 - нормированному коэффициенту A_2 функции (5); 4 - нормированному коэффициенту

ВАРИАНТ БЛИЖНЕГО СОСЕДА РАССТОЯНИЕ ЕВКЛИДА

ШАГ 6
КЛАСС 1 1
КЛАСС 2 2 9
КЛАСС 3 3 4 5
КЛАСС 4 6
КЛАСС 5 7
КЛАСС 6 8 10
КЛАСС 7 11
КЛАСС 8 12
КЛАСС 9 13 14
КЛАСС 10 15 17
КЛАСС 11 16

ШАГ 8
КЛАСС 1 1
КЛАСС 2 2 9
КЛАСС 3 3 4 5 6 8 10
КЛАСС 4 7
КЛАСС 5 11
КЛАСС 6 12
КЛАСС 7 13 14
КЛАСС 8 15 17
КЛАСС 9 16

ШАГ 11
КЛАСС 1 1 7
КЛАСС 2 2 3 4 5 6 8 9 10
КЛАСС 3 11
КЛАСС 4 12
КЛАСС 5 13 14
КЛАСС 6 15 16 17

ШАГ 12
КЛАСС 1 1 2 3 4 5 6 7 8 9 10
КЛАСС 2 11
КЛАСС 3 12
КЛАСС 4 13 14
КЛАСС 5 15 16 17

ШАГ 13
КЛАСС 1 1 2 3 4 5 6 7 8 9 10
КЛАСС 2 11
КЛАСС 3 12
КЛАСС 4 13 14 15 16 17

Рис.7. Результаты работы иерархической кластер-процедуры.

этой же функции;5 $-\ln Z$; 6 - нормированному коэффициенту C функции (2); 7 - нормированному коэффициенту β этой же функции; 8 - величина F_1 (частота слова с рангом I); 9 - нормированному коэффициенту K функции (3); I0 - нормированному коэффициенту B этой же функции.

Численные данные, приведенные в приложении I, и явились теми данными, которые обрабатывались на основе ППП SITO. ППП SITO представляет собой систему обработки разнотипных данных разнообразными методами. В данном эксперименте проведены следующие виды обработки исходных данных:

I. Формирование локального банка данных, введение таблицы экспериментальных данных (ТЭД);

2. Корректировка ТЭД с клавиатуры терминала;

3. Нормировка чисел в ТЭД стандартными отклонениями;

4. Определение одномерных распределений, выделенных для анализа признаков по всей совокупности объектов;

5. Определение и вывод на печать проекций всей совокупности объектов на плоскость, образованную выделенными признаками;

6. Определение главных компонент для анализируемых признаков по всей совокупности объектов;

7. Расчет группировки признаков с использованием в качестве меры близости взаимной корреляции анализируемых данных;

8. Использование иерархической кластер-процедуры для пошагового определения числа и состава классов;

9. Вычисление функции приращения межклассового расстояния.

7. Анализ результатов эксперимента.

Выделим четыре вопроса, связанные с исследуемыми текстами.

I. Сходство и различие в словарном составе текстов на естественных и искусственных языках.

2. Структурные различия текстов на естественных, проблемно-ориентированных формализованных языках и языках программирования.

3. Возможность использования результатов классификации для подтверждения гипотез о семантических особенностях художественных текстов.

4. Использование результатов классификации для выработки единого критерия "комфортности" диалогового языка для пользователя.

I. Ранговые распределения частот появления слов в разных произведениях, отражающие внутреннюю структуру построения произведения, имеют общие закономерности. Для любого произведения слева на частотной гистограмме (большие частоты) группируются слова функционального ядра языка, характерные для данного автора, а справа (малые частоты) - слова, харктерные только для данного произведения. Аналогичную картину, только еще более ярко выраженную, мы находим и для текстов на искусственных языках: слева - служебные слова, справа - имена данных, процедур, констант и меток, присущие только данной программе. (Несколько особо ведут себя лишь две величины: единица и индекс, наиболее предпочитаемый программистом. Эти величины могут попасть и в левую область, на рис. 6 особо указано их положение. Эта ситуация вполне объяснима: действительно, эти два объекта имеют особую функцию, близкую к служебным словам).

Таким образом, если известен объем словаря некоторой ПО, то можно вполне определенно сказать, каков должен быть объем функционального ядра языка. И наоборот, если имеется язык программирования с данным функциональным ядром, можно приближенно определить, годится ли он для управления словарем данной ПО.

2. В результате классификации мы получили устойчивую группу текстов программ, написанных на языках программирования. Все тексты этой группы характеризуются и особым видом частотного распределения (см. рис. 6.). Если при использовании различных аппроксимирующих функций художественные тексты группировались несколько разным образом, что соответствует учету какого-то одного из параметров с большим весом, то для текстов программ все коэффициенты имеют существенное отклонение от коэффициентов для прочих текстов. То же самое можно сказать и о частотном словаре, который не вошел ни в одну из групп художественных произведений, либо примыкает к текстам программ, либо является единственным представителем группы. Это обстоятельство характеризует его промежуточное положение по признаку "естественности" между программами и художественными произведениями. Важно отметить, что характерные группы наилучшим образом аппроксимируются лишь одной из функций. Это означает, что для разных типов языков можно применять различные аппроксимирующие функции: для художественных текстов - функция (4);
для текстов на языках программирования - функция (5);
для текстов на формализованных ПО языках - функция (3).

С точки зрения удобного для человека соотношения "универсальности" и "специальности" языка можно сделать следующие выводы. Прежде всего стремление к чрезвычайной универсализации искусственных языков по словарному составу не имеет достаточного основания, поскольку использование слов в каждом конкретном акте общения (программе, художественном произведении) характеризуется большим количеством повторов. Это связано с законом адекватного восприятия текстов (сообщений).

Аналогично тому, как в любом обществе коллективный научный интеллект является основой для научной деятельности, но каждый индивид обладает собственным стилем восприятия идей и их выражения, так и в системах человек-ЭВМ необходимо выдержать грань между "универсальностью" и "специальностью". "Универсальность" должна давать некоторый запас гибкости для возможности дальнейшего усложнения понятий и развития системы общения. "Специальность" должна обеспечить комфортность работы в данный момент с гибким упорядочением информации сообразно индивидуальным потребностям пользователя данной ПО.

3. Иерархическая кластер-процедура дает последовательную группировку текстов, которая может служить существенным аргументом при рассмотрении семантических вопросов художественных произведений.

Так, например, существующая гипотеза о тесной связи произведений А.С.Пушкина "Моя родословная", "Езерский" и "Медный Всадник" не подтверждается анализом поведения экспериментальных РРЧС и их аппроксимаций по (2) и (3). Имеется значительно большее расхождение между экспериментальной РРЧС суммарного текста и ее аппроксимации по (2) и (3), чем для отдельно взятых текстов. Кроме того, кластер-анализ этих же отдельно взятых текстов не привел к образованию устойчивой группы.

В то же время "Медный Всадник" (текст № 4) при различных аппроксимациях группируется с текстами № 3,5,6, что говорит о целостности и неразрывности семантического содержания поэмы-загадки "Медный Всадник". Характерно, что отдельно текст № 7 или № 8 не образуют устойчивых групп, что дополнительно свидетельствует о семантической целостности и неразрывности текста № 4.

Интерпретации полученных группировок других текстов - дело специалистов-филологов.

4. С точки зрения ответов на поставленные вопросы полученные результаты позволяют выделить следующие особенности текстов, адаптированных для диалогового взаимодействия с человеком:

I. Стабильность отношения общего словаря к частному;

2. Закон поведения РРЧС показывает границы комфортно-воспринимаемого при диалоге текста.

Таким образом, зная словарь некоторой предметной области, можно подобрать некоторое частотное распределение функционального ядра. Любой искусственный язык с данным функциональным ядром порождает текст, отвечающий требованиям "комфортности" диалога человека с ЭВМ. Разумеется, речь не идет об автоматическом синтезе языка, а лишь о критерии, формальной процедуре

проверки разрабатываемых языков на эффективность диалога. Опыт разработки диалоговых информационных систем показал полезность такой оценки при проектировании диалоговых языков.

Литература

I. Моль А. Социодинамика культуры. М.: Прогресс, I973

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

3. Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. - В кн.: Вычислительная лингвистика. М.: Наука, I976, с.I79-202.

4. Яблонский А.И. Стохастические модели научной деятельности. - В кн.: Системные исследования. Ежегодник, I975, М.: Наука, I976.

5. Ponomaryov V.M., Alexandrov V.V. Constructive approach to knowledge representation. "Comput. Ling. and Comput. Lang., 1980, 14, pp. 245-259.

6. Лотман Ю.М. Феномен культуры. - В кн.: Труды по знаковым системам, т.IO. Семиотика культуры. Уч. записки Тартуского гос. ун-та. Вып. 463. Тарту, I978.

7. Новиков А.И. Семантика текста и ее формализация. М.: Наука, I983.

8. Лотман Ю.М. Мозг - текст - культура - искусственный интеллект. - В кн.: Семиотика и информатика, Вып.I7. - М.;I98I.

9. Иванов В.В. Очерки по истории семиотики в СССР. - М.: Наука, I976.

IO. Арсентьева А.В., Поднозова И.П. Особенности построения баз данных культурных ценностей. - В кн.: Алгоритмические модели в автоматизации исследований. М.: Наука, I980, с. 24I-246.

II. Jacobson R. Linguistics and Poetics. "Style in Language", Mass., Second printing, 1964.

I2. "Вечерний Ленинград", I974, от 2I.02, с.2, "Новое о "Медном Всаднике".

I3. Белинский В.Г., ПСС, т.УII, М., I955, с.542-548.

I4. Пушкин А.С. Медный Всадник. Л.: Наука, I978.

I5. Благой Д.Д. Миф Пушкина о декабристах. Социологическая интерпретация "Медного Всадника". М., I927.

I6. Белый А. Ритм как диалектика и "Медный Всадник". М., I929.

I7. Бонди С.М. История заполнения "Альбома I833 - I835 годов". - В кн.: Рукописи А.С.Пушкина. Фототипическое издание. Альбом I833-I835 г.г. М., I939.

I8. Бродский И.Л. А.С.Пушкин. Биография. - М., I937, с.773-788.

I9. Гроссман Л.П. Пушкин. Изд-е 2-е. М., I958.

20. Гилельсон М.И. От Арзамасского братства к пушкинскому кругу писателей. - Л.: Наука, I977.

2I. Арапов М.В. Классификация и распределения в лингвистике.- В кн.: Семиотика и информатика. Вып. I7. М., I98I.

22. Лихачев Д.С. Текстология. М.-Л., I964.

23. Хомский Н., Миллер Дж. Конечные модели использования языка. - В кн.: Кибернетический сборник, нов. сер. № 4, I967.

24. Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текстов. - В кн.: Проблемы передачи информации, т.3, № I, I967.

25. Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. - В кн.: Семиотика и информатика, вып. IO, I978, с. 74-95.

26. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. Л.: Наука, I983.

27. Александров В.В., Алексеев А.И., Горский Н.Д., Никифоров А.М., Система обработки разнотипных данных . 3. Интерактивный вариант (материалы по математическому обеспечению). Л.: ЛНИВЦ АН СССР, I982.

Приложение I.

Таблица I.

НОМЕР ОБ'ЕКТА И СООТВЕТСТВУЮЩИЙ НОМЕР ОБУЧАЮЩЕЙ ВЫБОРКИ	Аппроксимация вида: $x_2 = A_1 + \frac{50}{z - B_1}$	Аппроксимация вида: $x_2 = \frac{50}{B_2 - z} - A_2$	Объем выборки	Закон Ципфа $x_2 = \frac{C}{z^\beta}$	Частоты слов	Закон Ципфа-Мандельброта $x_2 = \frac{K}{B+z}$
	1 A_1 2 B_1	3 A_2 4 B_2	5 $\ln Z$	6 C 7 β	8 F_1	9 K 10 B
1	9.800 6.900	1.600 12.200	5.700	7.400 6.000	25.000	.310 2.780
2	10.200 7.600	-.100 13.800	6.600	4.400 4.700	19.000	.340 12.490
3	10.900 6.500	.500 13.200	7.100	4.600 5.800	48.000	.260 5.920
4	12.700 5.100	1.200 12.000	7.400	4.800 6.400	82.000	.230 3.640
5	11.700 5.900	.600 12.900	7.900	4.300 6.500	104.000	.220 4.550
6	13.300 4.600	1.700 11.700	7.900	5.600 7.600	217.000	.190 1.320
7	9.900 7.100	1.100 12.700	5.900	6.400 5.600	23.000	.310 2.780
8	14.900 4.100	1.900 10.900	7.200	5.500 6.700	92.000	.230 2.610
9	10.400 7.800	-.400 14.000	6.800	4.100 4.500	17.000	.350 17.210
10	10.800 6.300	1.100 12.700	6.800	5.400 6.200	50.000	.260 3.520
11	15.200 4.200	.600 13.300	13.900	2.900 10.000	42454	.100 7.010
12	12.360 5.020	2.400 10.500	7.600	7.900 9.800	181.000	.200 1.090
13	14.643 4.117	2.810 9.420	9.327	6.837 .03300	932.000	.146 1.286
14	15.790 3.875	3.668 8.256	9.235	6.848 .00300	942.000	.146 .588
15	16.118 3.693	5.390 6.332	8.222	5.642 .121	782.000	.177 1.368
16	13.699 4.324	5.206 6.809	6.356	4.220 .02900	69.000	.237 1.007
17	11.798 5.027	2.999 10.306	7.212	5.094 .158	163.000	.196 .251

Приложение 2.

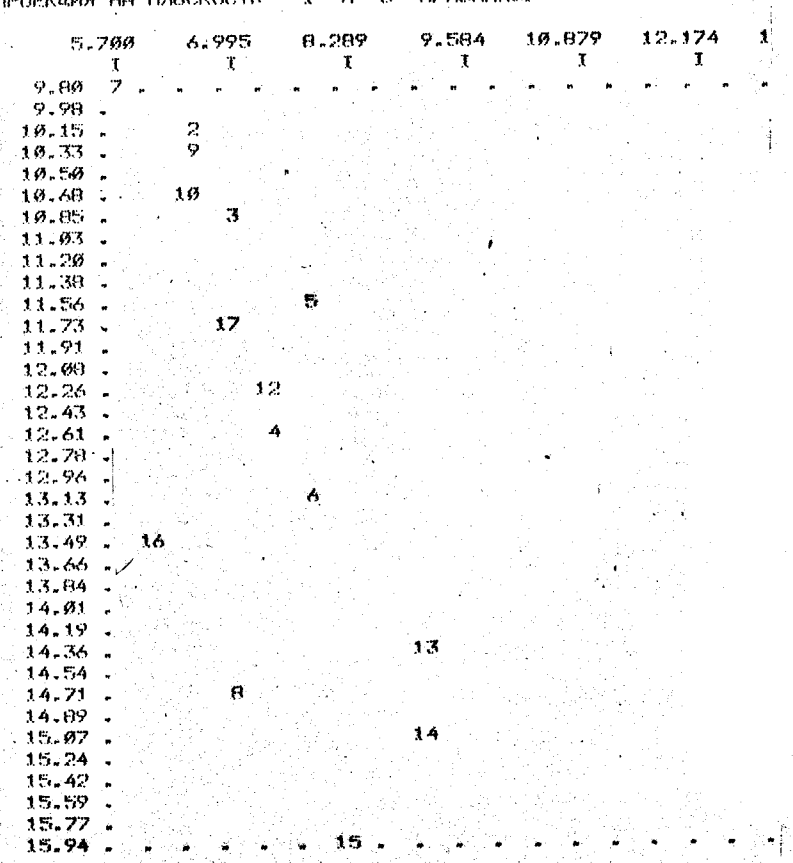
```
>> ИНТЕРАКТИВНАЯ СИСТЕМА ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ
SITO. НУЖНЫ ЛИ ПОЯСНЕНИЯ ?
      0-НЕТ
      1-ДА
? 0
  >> БЫЛ ЛИ ПРЕДВАРИТЕЛЬНО
  СФОРМИРОВАН ЛОКАЛЬНЫЙ БАНК ДАННЫХ ?
      0-НЕТ ИЛИ БЫЛ НО НЕОБХОДИМО ФОРМИРОВАТЬ НОВЫЙ
      1-БЫЛ И ИМ НАДО ПОЛЬЗОВАТЬСЯ ПОЛНОСТЬЮ
      2-БЫЛ, НО ИСПОЛЬЗОВАТЬ ТОЛЬКО ТЭД
? 0
  >> КОЛИЧЕСТВО ОБ'ЕКТОВ ?
? 17
  >> КОЛИЧЕСТВО ПРИЗНАКОВ ?
? 10
  >> ОТКУДА БУДЕТ ВВОДИТСЯ ТЭД ?
      1-ИЗ ФАЙЛА ДАТА, СОЗДАННОГО ВАМИ
      2-С КЛАВИАТУРЫ ТЕРМИНАЛА
? 1
  >> КАК ВВОДИТСЯ ТЭД ?
      1-ПО СТРОКАМ
      2-ПО СТОЛБЦАМ
      3-РАЗ'ЯСНЕНИЕ
? 1
  >> ТИП РЕШАЕМОЙ ЗАДАЧИ ?
      0-АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ
      1-РАСПОЗНАВАНИЕ ОБРАЗОВ
      2-РЕГРЕССИОННЫЙ АНАЛИЗ
? 0
  >> ТИПЫ ПРИЗНАКОВ В ТЭД ?
      0-РАЗНОТИПНЫЕ
      1-КЛАССИФИКАЦИОННЫЕ
      2-КАЧЕСТВЕННЫЕ
      3-КОЛИЧЕСТВЕННЫЕ
      4-РАЗ'ЯСНЕНИЕ
? 3
  >> ЕСТЬ ЛИ ИНФОРМАЦИЯ О РАЗЛИЧНОЙ ВАЖНОСТИ ПРИЗНАКОВ ?
      0-НЕТ
      1-ДА
      2-РАЗ'ЯСНЕНИЕ
? 0

  >> ДЕЛАТЬ РАСПЕЧАТКУ ТЭД ?
      0-НЕТ
      1-НА ЭКРАН
      2-В ФАЙЛ HISTORY
      3-НА ЭКРАН И В ФАЙЛ HISTORY
? 1

ЧИСЛО ОБ'ЕКТОВ=  17
ЧИСЛО ПРИЗНАКОВ=  10
```

```
 >> НУЖНО ЛИ КОРРЕКТИРОВАТЬ ТЭД ?
       0-НЕТ
       1-ДА (ЗАМЕНА ЗНАЧЕНИЙ В ТЭД)
       2-ДА (ИСКЛЮЧЕНИЕ ОБ'ЕКТОВ И/ИЛИ ПРИЗНАКОВ)
       3-РАЗ'ЯСНЕНИЕ
? 2
 >> ОТКУДА БУДЕТ ВВОДИТЬСЯ ИНФОРМАЦИЯ О РЕДАКТИРОВАНИИ ?
       1-ИЗ ФАЙЛА ДАТА
       2-С КЛАВИАТУРЫ ТЕРМИНАЛА
? 2
  >> ЧИСЛО ИСКЛЮЧАЕМЫХ ПРИЗНАКОВ
     ЧИСЛО ИСКЛЮЧАЕМЫХ ОБ'ЕКТОВ ?
? 7,0
  >> ВВЕДИТЕ НОМЕРА ИСКЛЮЧАЕМЫХ ПРИЗНАКОВ
? 3,4,6,7,8,9,10

 >> ДЕЛАТЬ РАСПЕЧАТКУ ТЭД ?
       0-НЕТ
       1-НА ЭКРАН
       2-В ФАЙЛ HISTORY
       3-НА ЭКРАН И В ФАЙЛ HISTORY
? 1

ЧИСЛО ОБ'ЕКТОВ-  17
ЧИСЛО ПРИЗНАКОВ-  3
```

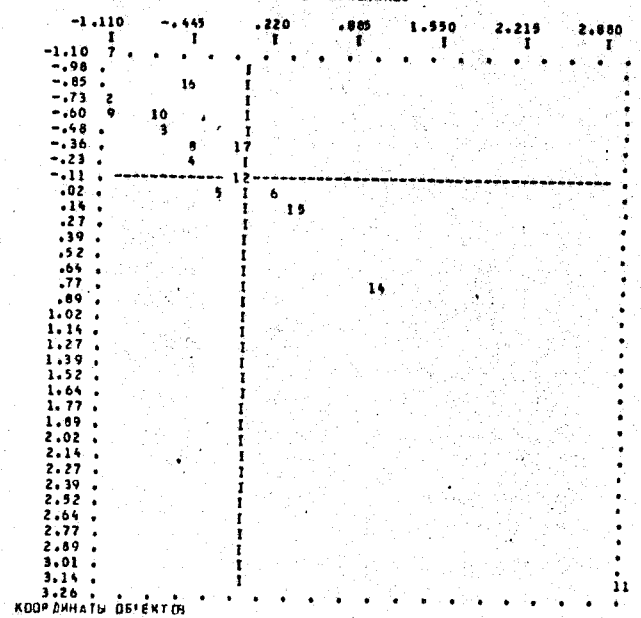
НОМЕР ОБ'ЕКТА И СООТВЕТСТВУЮЩИЙ НОМЕР ОБУЧАЮЩЕЙ ВЫБОРКИ	1 КОЛИЧЕСТВЕННЫЙ ПРИЗНАК 1.000	2 КОЛИЧЕСТВЕННЫЙ ПРИЗНАК 1.000	5 КОЛИЧЕСТВЕННЫЙ ПРИЗНАК 1.000
1	9.800	6.900	5.700
2	10.200	7.600	6.600
3	10.900	6.500	7.100
4	12.700	5.100	7.400
5	11.700	5.900	7.900
6	13.300	4.600	7.900
7	9.900	7.100	5.900
8	14.900	4.100	7.200
9	10.400	7.800	6.800
10	10.800	6.300	6.800
11	15.200	4.200	13.900
12	12.360	5.020	7.600
13	14.643	4.117	9.327
14	15.290	3.875	9.235
15	16.118	3.693	8.222
16	13.699	4.324	6.356
17	11.798	5.827	7.212


```
 >> ПЕЧАТАТЬ ПРОЕКЦИИ ОБ'ЕКТОВ
 НА ПЛОСКОСТЬ, ОБРАЗОВАННУЮ ПРИЗНАКАМИ,
КОТОРЫЕ ВЫ УКАЖИТЕ ? .

       0-НЕТ(ЗАКОНЧИТЬ РАБОТУ)
       1-ДА (НА ЭКРАН ДИСПЛЕЯ)
       2-ДА (В ФАЙЛ HISTORY)
       3-ДА (НА ЭКРАН И В ФАЙЛ HISTORY)
? 1
  >> ВВЕДИТЕ НОМЕРА ПРИЗНАКОВ (ПАРУ),
  КОТОРЫЕ ОБРАЗУЮТ ПЛОСКОСТЬ ПРОЕКЦИИ.
? 1,5
```

ПРОЕКЦИЯ НА ПЛОСКОСТЬ 1 И 3 ПРИЗНАКОВ



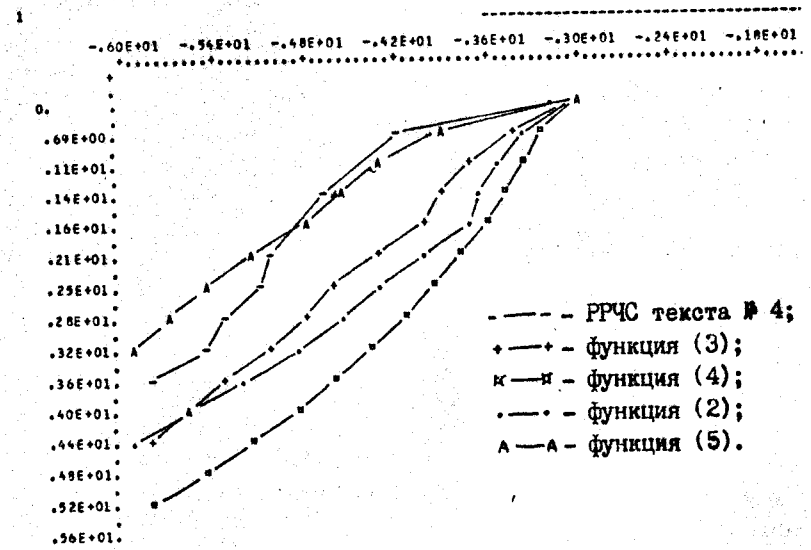
ПРОЕКЦИЯ НА ПЛОСКОСТЬ 5 И 8 ПРИЗНАКОВ



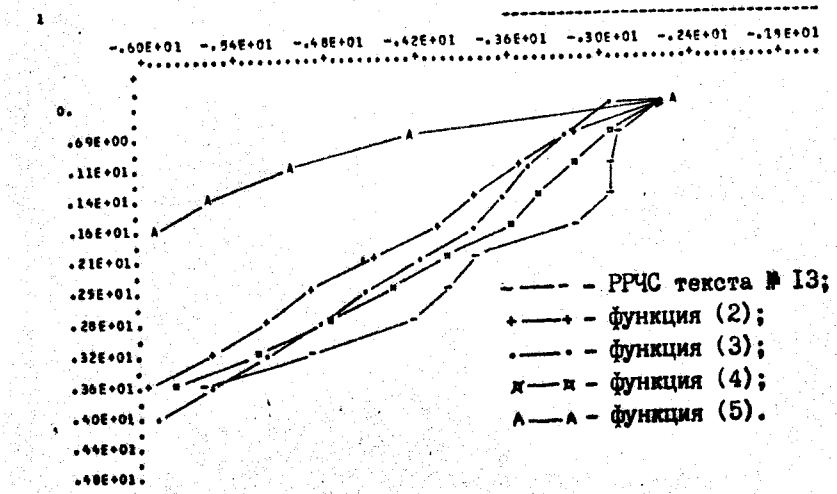
КООРДИНАТЫ ОБЪЕКТОВ

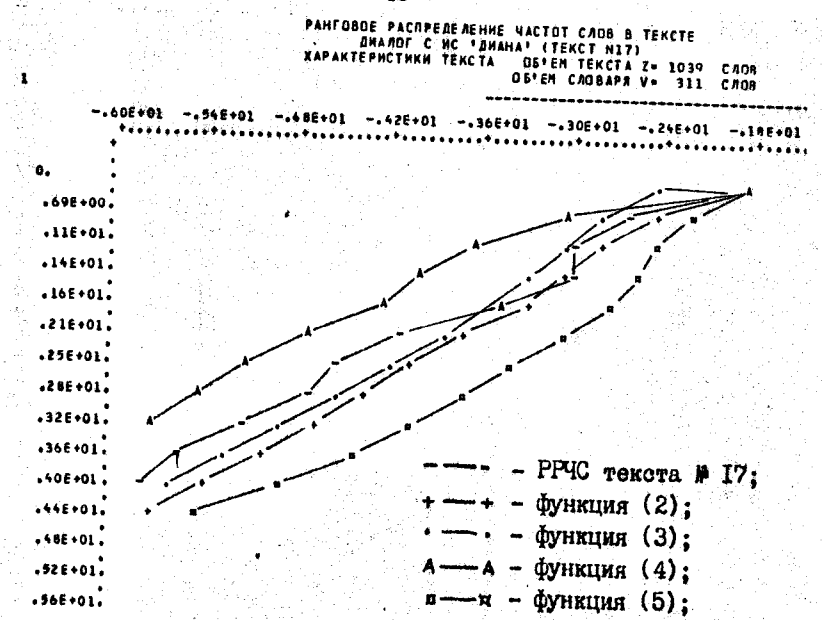
*	1	-1.134	-.903	***	2	-.611	-1.050	***	3	-.337
	6	.101	.259	***	7	-.994	-.948	***	8	-.282
	11	3.389	3.102	***	12	-.063	.162	***	13	.883
	16	-.745	-.365	***	17	-.276	.106	**		

РАНГОВОЕ РАСПРЕДЕЛЕНИЕ ЧАСТОТ СЛОВ В ТЕКСТЕ
А.С.ПУШКИН. МЕДНЫЙ ВСАДНИК (ТЕКСТ N4)
ХАРАКТЕРИСТИКИ ТЕКСТА ОБ'ЕМ ТЕКСТА Z= 1659 СЛОВ
ОБ'ЕМ СЛОВАРЯ V= 994 СЛОВ



РАНГОВОЕ РАСПРЕДЕЛЕНИЕ ЧАСТОТ СЛОВ В ТЕКСТЕ
ИС 'ДИАНА' (ТЕКСТ N13)
ХАРАКТЕРИСТИКИ ТЕКСТА ОБ'ЕМ ТЕКСТА Z=11329 СЛОВ
ОБ'ЕМ СЛОВАРЯ V= 1182 СЛОВ





Оглавление.

Редактор Орехов Ц.И.

ЛИЯФ, зак. 996 , тир.I60, уч.-изд.л. 2,23, 08. 12.83, М-24754
Цена 33 коп.