4
П-71 11

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной
и прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

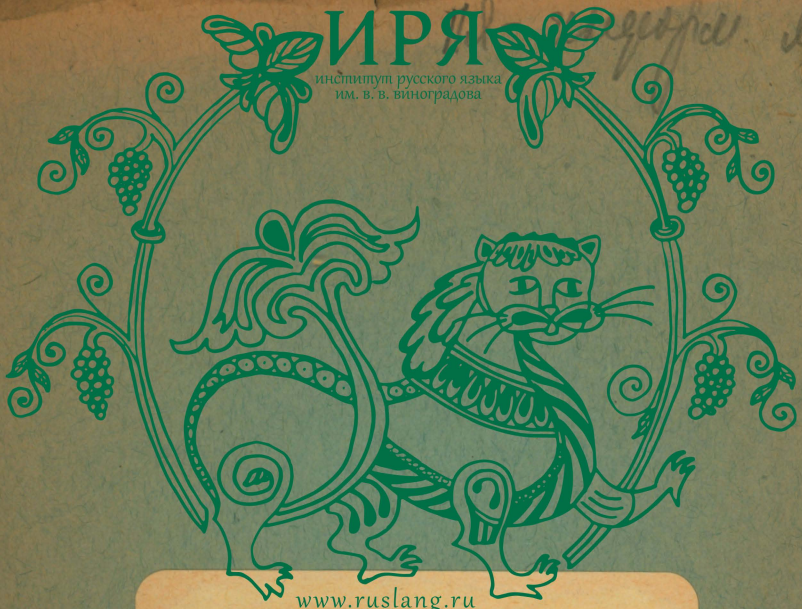
Выпуск 11

Н. Н. ЛЕОНТЬЕВА

О СОЗДАНИИ ИНФОРМАЦИОННОГО ЯЗЫКА НА БАЗЕ ПОЛНОГО СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Москва — 1970

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

65111

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа
по экспериментальной и прикладной
лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

Выпуск 11

Н.Н. Леонтьева

О СОЗДАНИИ ИНФОРМАЦИОННОГО ЯЗЫКА НА БАЗЕ ПОЛНОГО СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Москва - 1970

4
П-71
11
БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР
65111

1. Подавляющее большинство созданных и создаваемых в последнее время информационных языков имеют одно единственное назначение – обеспечивать документальный поиск информации. Их справедливо выделяют в класс поисковых языков, см. [1]. Среди поисковых языков наибольшее распространение имеют языки дескрипторного типа. Простота задачи, которую решают с помощью дескрипторных языков, находится в полном соответствии с простотой самого языка: классические дескрипторные языки (например, языки системы "Пусто-Непусто") имеют только лексику, состоящую из ограниченного для каждой определенной области набора дескрипторов, и небольшого количества базисных (постоянно существующих, словарных) отношений между дескрипторами, причем отношения используются только на стадии поиска. Синтагматических отношений поисковые языки, как правило, не имеют. Простота структуры дескрипторного языка позволяет осуществлять автоматическое индексирование документов. При том, что системы индексирования и документального поиска, основанные на дескрипторных языках, могут работать и давать удовлетворительные результаты, существует порог (в рамках задач только документального поиска), который непреодолим даже для самых развитых дескрипторных языков: попытка уменьшить потери информации приводит к большому информационному "шуму" и наоборот. Это и толкает некоторых исследователей на введение элементов грамматики (т.е. синтагматических отношений) в дескрипторные языки. Заметим, однако, что несистематическое использование элементов грамматики – плохой выход, так как он влечет за собой неалгоритмизуемость процедуры перевода с естественного языка на поисковый. Очевидно, имеет смысл только последовательное, систематическое использование грамматики.

Систематическое введение элементов грамматики в поисковые языки означает существенное усложнение его структуры и приводит к большим принципиальным трудностям. На основную из них указывают авторы работы [1] : "... поисковая система, которая улавливала бы

малейшие оттенки в содержании документов и запросов, могла бы оказаться неработоспособной, парализуя сама себя педантичной придирчивостью к каждому, даже малозначительному оттенку смысла и по всякому поводу отказываясь выдавать документы. Для того чтобы бороться с такой чрезмерной чувствительностью, поисковой системе потребовалась бы очень мощная логика". Встает вопрос, целесообразно ли обеспечивать мощной логикой систему, предназначенную только для документального поиска информации, не будет ли это непозволительной роскошью.

Нам представляется неразумным строить сложные языки с грамматикой, даже имеющие алгоритм перехода к ним от естественного языка, в расчете только на простые поисковые задачи: такие языки скорее всего не выдержат конкуренции с последовательно дескрипторными языками, так как слишком большой ценой дадут увеличение точности документального поиска. Сложность языка должна находиться в соответствии со сложностью решаемых задач. Так, едва ли не единственным недостатком системы СИНТОЛ является его ориентация на решение только документальных задач, тогда как стройность логики языка СИНТОЛ, богатство используемых сведений и наличие процедуры автоматического перевода с естественного языка дают основания ставить задачу хотя бы фактографического поиска информации.

2. Класс поисковых языков, не имеющих грамматики, не дает естественного выхода к решению задач бóльшей информационной сложности. Для того чтобы выйти за пределы задач чисто документального поиска, придется создавать совершенно новые искусственные языки. Трудно надеяться, что грамматику можно будет просто надстроить над множеством дескрипторов. Это означает, что каждая новая и более сложная задача будет вызывать к существованию новый язык: отдельный язык для фактографического поиска, отдельный для реферирования, отдельный для перевода и т.д.

Можно пойти по другому пути: по пути создания богатого языка, пригодного для решения наиболее сложных информационных задач. Вероятно, ИЯ, который может обслуживать такие сложные задачи, как реферирование, сможет использоваться и для более простых задач, например, фактографического и документального поиска.

Будем называть универсальными такие информационные языки, которые дают возможность решать широкий круг информационных задач: накопление и классификация информации, преобразование информации, сжатие информации, фактографический поиск, документальный поиск, автоматический перевод с одного языка на другой, в частности, на тот же самый. Заметим, что в это перечисление входит и задача автоматического реферирования: мы определяем его как перевод, сопровождающийся сжатием информации.

Универсальность языка понимается нами не только как способность языка решать разные задачи, но и как универсальность относительно характера описываемых текстов. Имеется в виду неизменность грамматики ИЯ при описании текстов разных областей знаний. Языки, которые имеют только лексику (напр., чисто дескрипторные языки) не являются и с этой точки зрения универсальными: лексический состав обязательно меняется от области к области. Не являются универсальными также и языки с так называемой "анкетной" грамматикой: роль грамматики в них выполняют анкеты, или таблицы со строго фиксированными вопросами. Ясно, что изменение лексического состава таких языков влечет за собой изменение состава таблиц. Грамматика, изменяющаяся вместе с лексикой, тем самым не является универсальной. Примером языка с анкетной грамматикой является ИПЯ для химии [7].

3. Создание универсального ИЯ, неизбежно сложного по структуре, должно предполагать возможность автоматического перехода к нему от естественных текстов и обратно. Существование правил такого перехода не только избавляет нас от опасности неоднозначного представления на ИЯ одинаковых по смыслу высказываний, но и, по справедливому мнению авторов [1], увеличивает логические возможности системы: ... "неалгоритмический и неавтоматический характер процедуры перевода с естественного языка должен существенно ограничивать семантические возможности поисковых языков". (и не только поисковых, добавим мы). Поэтому представляется нецелесообразным создавать

сложные языки и сложные информационные системы, вход и выход из которых неалгоритмизуем.

Поскольку универсальный ИЯ предназначен для решения сложных задач автоматической обработки информации, заключенной в тексте, он должен располагать средствами, позволяющими передавать в принципе всю информацию, содержащуюся в тексте. Иначе говоря, ИЯ должен быть таким, чтобы любой текст мог быть переведен на него без потери информации. Это означает, что ИЯ должен иметь не только лексику, вводящую набор понятий, но и грамматику, отражающую синтагматические связи между понятиями.

Необходимость автоматизировать процесс перехода от естественного языка (ЕЯ) к ИЯ предъявляет свои требования к структуре ИЯ: она должна быть подобной, соизмеримой структуре ЕЯ. Только при этом условии возможно построить алгоритм, который в одно или несколько последовательных преобразований заменит текст на ЕЯ записью на ИЯ.

4. Насколько реально говорить о создании класса информационных языков, обладающих перечисленными выше свойствами - наличие грамматики, универсальность, алгоритмизуемость? Для того чтобы автоматической процедурой уметь извлекать из текстов нужную информацию, содержащуюся в них в явном или неявном виде, и переносить в запись на ИЯ, необходимо располагать сильным формальным аппаратом описания семантики. Имеется ли он? В последние годы область формального описания семантики естественных текстов развивается очень интенсивно. Создано и продолжает создаваться много конкретных семантических описаний, которые безусловно могут быть уже использованы в обсуждаемой области информатики; например, описание лексической сочетаемости слов, позволяющее снять при переходе к некоторому смысловому языку существенную часть вариативности выражения стандартных смыслов, регулярно повторяющихся в естественном тексте - работы [10], [11] и др.; представление смысла слов через набор элементарных понятий - [8], [9] и др.; синонимические преобразования предложений,

содержащих отрицание – [12], и др.; список работ по семантике естественного текста, которые необходимо учитывать при создании ИЯ, может быть продолжен. Однако просто перенести имеющиеся описания в некоторую создаваемую информационную систему нельзя: одно только согласование этих описаний между собой представляет собой большую работу. Но создание информационной системы, использующей сначала простые механизмы смысла и постепенно охватывающей все более и более сложные семантические явления, вполне реально; важно выбрать такую структуру ИЯ, которая позволит фиксировать любую информацию, извлекаемую из текста. К имеющимся сейчас результатам собственно лингвистических исследований можно добавить большой опыт, накопленный создателями информационных систем по составлению словарей дескрипторов и описанию базисных отношений между дескрипторами. Что касается описания текстуальных (синтагматических) смысловых отношений, являющихся, по нашему мнению, наиболее существенным элементом структуры ИЯ, то хотя они и используются в ряде ИПС – [4], [5], [6], мы не считаем лишним представить еще один вариант описания грамматики текстуальных смысловых отношений. Сам список смысловых отношений, а также определения отношений и их иерархию мы не считаем в настоящее время полными и окончательными.

5. Во второй части данной работы мы опишем конкретный семантический язык и покажем возможность его использования в качестве информационного языка охарактеризованного выше класса. Для удобства дальнейшего изложения дадим описываемому языку имя, например, ЭХО. (Единственный смысл имени – подчеркнуть зависимость свойств ИЯ от собственно лингвистических описаний.)

Язык ЭХО располагает средствами, позволяющими передавать в принципе всю информацию, содержащуюся в исходном тексте. Благодаря этому он может использоваться при переводе с одного естественного языка на другой, так как перевод должен осуществляться без потери информации.

Нас больше интересует другой аспект языка ЭХО, - и в этом мы видим его основное назначение, - ИЯ как средство сжатия смысловой информации исходного полного текста. Структура языка ЭХО такова, что она позволяет записывать на этом языке смысл исходного документа с различной степенью полноты, - от самого полного вида СЗ, являющегося пофразовым переводом текста, до самого краткого вида, являющегося формулировкой основных тем документа.

Все переходы внутри языка ЭХО - от любого более полного состояния к любому менее полному, а также внешние переходы - от естественного текста к его записи на ИЯ и от любого состояния ИЯ к его формулировке на естественном языке, - алгоритмизуемы. Следовательно, можно говорить о возможности автоматического сжатия смысла текста.

Разные состояния языка ЭХО (различающиеся степенью полноты представления смысла) могут использоваться на разных стадиях поиска информации: наиболее краткая СЗ, отражающая только темы документа - на стадии документального поиска информации, полная СЗ - на стадии фактографического поиска. Следовательно, поисковый образ (ПО) документа, нужный для документального поиска, создается автоматически как последняя ступень сжатия полного текста.

Перевод на естественный язык разных состояний записи на ИЯ ЭХО дает серию вторичных документов: автоматически отредактированный текст - реферат - резюме - аннотацию - формулировку темы. Схема получения разных состояний ИЯ из естественного текста, а также схема предполагаемого использования ИЯ помещены в конце статьи.

6. Структура семантической записи на языке ЭХО.

Любое высказывание может быть записано на языке ЭХО последовательностью строк вида R (A, B). Множество элементов, способных находиться только в позиции A или B, образуют лексику языка. Эти элементы можно называть термами, или, пользуясь устоявшейся в информатике терминологией, дескрипторами. Каждый дескриптор является представителем одного

или нескольких слов или грамматических категорий, имеющих один и тот же смысл.

В позиции R может находиться любой из элементов, входящих в список смысловых отношений.

В настоящем варианте языка ЭХО насчитывается 50 отношений.

Для наглядности дадим пример семантической записи (далее СЗ) на нашем ИЯ одного документа*).

Фраза: Просьба тов. Шитикова о направлении "Минска-22" Новосибирскому Ин-ту инженеров железнодорожного транспорта.

Полная семантическая запись данной фразы выглядит так:

1. субъект (товарищ, просить)
2. имя (Шитиков, товарищ)
3. содержание (дать, просить)
4. объект (ЭЦВМ, дать)
5. имя (Минск-22, ЭЦВМ)
6. адресат (ВУЗ, дать)
7. локализация (Новосибирск, ВУЗ)
8. специализация (инженеры, ВУЗ)
9. специализация (транспорт, инженеры)
10. специализация (железнодорожный, транспорт)

(Каждая строка читается в порядке: 1-й член отношения - отношение - 2-й член отношения).

Чем отличается полная СЗ некоторого текста от самого текста?

Во-первых, наличием лишних, по сравнению с текстом, элементов - смысловых отношений. Они представляют собой по сути дела смысловую интерпретацию синтаксических связей, так что структура полной непреобразованной СЗ эквивалентна структуре дерева.

---

*) Материалом, с которым мы работаем, являются не первичные документы, а краткие резюме содержания экономических деловых бумаг. Резюме составляются людьми и, как правило, формулируются в одной назывной фразе.

Во-вторых, большей однозначностью лексики: разные слова ЕЯ, имеющие один смысл, представлены в СЗ одним дескриптором, и наоборот - одно слово ЕЯ, имеющее несколько значений, переходит в СЗ в виде одного из ее дескрипторов (так, слово Институт представлено в СЗ дескриптором ВУЗ; второй возможный дескриптор того же слова - НИИ).

В-третьих, большей эксплицитностью (вместо одного слова текста "Минск-22" в СЗ представлены два дескриптора "ЭЦВМ" и "Минск-22") и меньшей избыточностью.

В-четвертых, свободной членимостью высказывания: мысль, облеченная в языке в форму предложения, не делимого на простые отрезки, выражающие элементарные мысли, представлена в СЗ набором отдельных элементарных высказываний, которые можно располагать в любой последовательности, а также свободно извлекать из СЗ.

В-пятых, стандартностью и простотой грамматических средств: разнообразные синтаксические конструкции естественного языка представлены в СЗ одинаковыми тройками - отношение и два его члена. Состав смысловых отношений в принципе не меняется от области к области.

Перечисленные свойства семантической записи позволяют совершать те формальные операции преобразований смысла текста, которые неприменимы ни к самому тексту, ни к его синтаксическому дереву.

Однако преобразования семантической записи опираются не только на структуру самой записи и однозначность ее элементов, но и на все те словарные и грамматические соотношения, которые стоят за каждой семантической записью. Поэтому прежде чем описывать возможные в языке ЭХО преобразования, изложим кратко организацию лексики и систему грамматики языка.

7. Организация лексики.

К лексике ИЯ относится множество дескрипторов ИЯ

Организацию составляют два ряда отношений:

1) отношения между единицами ЕЯ (словами) и единицами ИЯ (дескрипторами); 2) отношения дескрипторов между собой - внутри ИЯ.

7 а. Единицы ЕЯ - единицы ИЯ. Не имея здесь возможности подробно останавливаться на сложном алгоритме перехода от единиц ЕЯ к единицам ИЯ (речь идет об автоматической процедуре такого перехода) и на разных промежуточных ступенях такого перехода (промежуточные дескрипторы, условные дескрипторы), проследим те окончательные соотношения, которые можно наблюдать, сравнивая некоторый текст с его полной семантической записью.

Единицы ЕЯ	Единицы ИЯ
слово напр., "Строительство"	дескриптор СТРОИТЬ
несколько слов напр., "Электронно-цифровая машина"	I дескриптор ЭЦВМ
пустое слово напр., "который"	дескриптор ПЛАН
слово напр., "непосредственно"	отсутствие дескриптора -
слово напр., "в"	не дескриптор, а смысловое отношение локализация (,)
слово напр., "вес"	дескриптор вместе с отношением параметр (ВЕС,)
отсутствие слова	дескриптор

Напр., одно слово "Минск - 22" будет представлено в СЗ двумя дескрипторами с отношением между ними: имя (МИНСК-22, ЭЦВМ). ЭЦВМ - тот самый дескриптор, которому не соответствует никакое слово текста.

одно слово	несколько разных дескрипторов
напр., "управление"	1) РУКОВОДИТЬ 2) УЧРЕЖДЕНИЕ
разные слова	один и тот же дескриптор
напр., 1) "верховный" 2) "головной" 3) "главный"	ГЛАВНЫЙ

Два последних соотношения (соответственно омонимия и синонимия) зафиксированы непосредственно в естественно-семантическом словаре, а в каждой конкретной СЗ реализуется одна из возможностей, т.е. имеет место соотношение "слово-дескриптор".

Сколь бы важным ни был вопрос о разрешении омонимии при переходе от ЕЯ к ИЯ, он решается достаточно естественно в рамках полного семантического анализа текста (ему мы посвятим отдельную статью). Гораздо более существенной является задача овладения синонимией, т.е. умение свести к одной семантической записи высказывания, имеющие разное выражение, но один смысл. Такое предпочтение подкрепляется высказывавшимся уже соображением, что синонимия представляет нормальное для языка явление, в то время как омонимия - патология [9]. Прибавим к этому, что большинство функций информационного языка связано с требованием - узнавать ту же информацию в разном.

При переходе от текста к СЗ снимается та часть вариативности выражения смысла, которая вызвана морфолого-синтаксическими различиями (так, словоформам "строит", "строящийся", "строительство", "построен" будет сопоставлен один дескриптор), лексическими конверсивами ("больше" - "меньше", "имеет" - "принадлежит"), обычными синонимами ("недостаток" - "нехватка" - "дефицит", "строить" - " возводить" - "сооружать"). Понятие синонима было существенно расширено в работах Апресяна - Жолковского - Мельчука о лексических функциях. Ими описаны большие группы слов, выражающих регулярно

повторяющиеся смыслы. Слова внутри групп приравниваются друг к другу не непосредственно, а сложным образом: A так относится к B , как C к D , как E к F и т.д. Слова A, C, E являются, так сказать, функциональными синонимами, т.е. в нашей системе могут быть заменены в принципе одним дескриптором.

Примечание. Соотношение между лексическими функциями (ЛФ) и единицами описываемого ИЯ не простое - нельзя сказать, что каждая ЛФ отображается в какой-либо дескриптор. Между ними могут быть все те же отношения, которые перечислены выше для единиц ЕЯ и дескрипторов. Так, например, ЛФ V, S, A, Adv, Able по возможности сводятся к значению одной основы, а дополнительный смысл выражается каким-то смысловым отношением либо отдельным дескриптором; ЛФ Incep , Cont, Fin - разные дескрипторы, общим у них является то, что каждый из них стоит на 1 месте смыслового отношения СТАДИЯ (,); ЛФ Caus, Real, Manif, Fact, Liqu - разные дескрипторы, имеющие семантический признак "действие" и валентности СУБЪЕКТ (,), ОБЪЕКТ (,) и др.; ЛФ Magn, Ver, Bon - разные дескрипторы, присоединяемые к определяемому дескриптору смысловым отношением ПРИЗНАК (,) или ОЦЕНКА (,), напр., ОЦЕНКА (Bon , КАЧЕСТВО); ЛФ Oper, Func (при i_0 = "действие") в окончательной СЗ не остаются, так как они характеризуют только синтаксические функции i_0, а к значению соответствующего действия ничего не добавляют; ЛФ Instr и Loc переходят в смысловые отношения С ПОМОЩЬЮ (,) и ЛОКАЛИЗАЦИЯ (,); всем ЛФ Syn от некоторого i_0 соответствует тот же дескриптор, что и i_0 ; ЛФ Anti, Plus, Minus - скорее не дескрипторы, а элементарные смыслы, являющиеся составной частью смысла многих дескрипторов.

Трудность перенесения ЛФ в наш ИЯ объясняется разными причинами, в том числе тем, что смысл слова ЕЯ передается в ИЯ не только и не столько лексически (выбо-

ром дескриптора), сколько его смысловыми связями (названиями смысловых отношений, являющихся частью смысла слова).

7 б. Словарные отношения между дескрипторами.

Хотя в полной СЗ текста устранена лексическая и синтаксическая синонимия, свойственная тексту (в СЗ нет синонимичных дескрипторов), остается синонимия на уровне высказываний. Так, в каждой паре следующих высказываний дважды выражена одна информация:

Новое горючее для самолетов	- Новое горючее в авиации
Доставка Новосибирску двух ЭВМ	- Получение Новосибирском двух ЭВМ
Использование "Минска-22" в Стройбанке	- Автоматизация операций в Стройбанке

Нужно иметь аппарат формальных преобразований СЗ, позволяющий обнаруживать одну и ту же информацию в разных СЗ. Цель эта достигается отчасти средствами грамматики ИЯ, отчасти лексическими средствами. К лексическим средствам мы относим словарь, фиксирующий смысловые сходства и различия между самими дескрипторами. Идея такого словаря, называемого словарем базисных (или парадигматических, или словарных) отношений, присутствует в большей или меньшей мере почти в любой информационной системе, являясь основой варьирования запросов при сопоставлении их с документами. В таких словарях используются обычно два вида отношений: более или менее определенные родо-видовые отношения и совсем туманные "ассоциативные" отношения. Информационная система, обладающая грамматикой, имеет средства конкретизировать ассоциативные отношения, разбив их на ряд точно определенных смысловых отношений. Будем называть множество отношений, характеризующих постоянные связи между дескрипторами, энциклопедическими отношениями (мы позволим себе пользоваться этим термином, пока не будет предложено более правильного его определения). Словарь, фиксирующий постоянные связи между дескрипторами, будем называть энциклопедическим словарем.

Итак, стоит задача определить круг и точные значения энциклопедических отношений и для каждого дескриптора заполнить по возможности все клетки этих значений. Источником многих из них является все тот же Толково-комбинаторный словарь, поскольку многие лексические функции являют собой по сути дела энциклопедические отношения (Gener, S_1, S_2, S_{loc}, Func и Oper от предметов, Result, Mult и др.). Использовав рамку толково-комбинаторного словаря, мы будем брать названия энциклопедических отношений в основном из нашего же списка синтагматических СО, добавляя к ним только слово "типичный". Так, отношение СУБЪЕКТ будет звучать в энциклопедическом словаре как "типичный субъект", ОБЪЕКТ - как "типичный объект", СПЕЦИАЛИЗАЦИЯ - как "типичная специализация" и т.д. По-видимому, целесообразно иметь отдельные энциклопедические словари для дескрипторов - "предметов" и дескрипторов - "действий", поскольку у них мало пересекающихся характеристик. Приведем для некоторых слов примеры сведений, которые войдут в энциклопедический словарь (не предопределяя его окончательную структуру).

Дескриптор	Родовое понятие	Типичная сфера	Типичная специализация	Актантная структура специализации	Типичные актанты	Типичный результат
ВУЗ	УЧЕБНОЕ ЗАВЕДЕНИЕ	ПРОСВЕЩЕНИЕ, ВЫСШЕЕ ОБРАЗОВАНИЕ	ОБУЧАТЬ	субъект (А, ОБУЧАТЬ) объект (В, ОБУЧАТЬ)	А=преподаватель В=студент	А=выпускник; А=молодой специалист; ПАРАМЕТР (высшее образование, А)

Дескриптор	Родовое понятие	Типичная сфера	Типичная специализация	Типичная локализация
СПЕЦИАЛИСТ	ПРОФЕССИЯ, КВАЛИФИКАЦИЯ, СПЕЦИАЛИЗАЦИЯ	КАДРЫ	РАБОТАТЬ	УЧРЕЖДЕНИЕ

Дескриптор	Родовое понятие	Типичная сфера	Типичное действие	Типичная специализация
САМОЛЕТ	ТРАНСПОРТ	АВИАЦИЯ	ЛЕТАТЬ	ПЕРЕВОЗИТЬ

Дескриптор	Родовое понятие	Типичная сфера	Типичная специализация	Типичное действие
МИНСК-22	ЭЦВМ	ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА	СЧИТАТЬ, ВЫЧИСЛЯТЬ, АВТОМАТИЗИРОВАТЬ	РАБОТАТЬ

Дескриптор	Родовое понятие	Актантная структура	Типичные актанты	Предпосылки	Следствия	Типичное средство	Типичная сфера
СЛАТЬ	ДАВАТЬ	субъект (A,) объект (B,) адресат (C,) исходн. пункт (D,) конечн. пункт (E,)	A=ОТПРАВИТЕЛЬ B=ПОСЫЛКА C=АДРЕСАТ E=МЕСТО НАЗНАЧЕНИЯ D=ОТПРАВНОЙ ПУНКТ	принадлежность (B,A) локализация (D,B) локализация (E,C)	принадлежность (B,C) локализация (E,B)	ПОЧТА, ТРАНСПОРТ	СРЕДСТВА СВЯЗИ

Сведения, помещенные в графах "Предпосылки" и "Следствия", чрезвычайно важны при поиске ответов на запросы, они дают возможность соотнести такие пары документ – запрос:

Передача Новосибирску двух ЭВМ – Наличие ВТ в Сибири;
О переслке IBM 4-70 в Тбилиси – Где имеются ЭВМ IBM 4-70?;
О переводе обувной фабрики из Горького в Казань в 1958 году – Сколько обувных фабрик было в Горьком в 1957 году? и др.

Описанные таким образом два плана организации лексики будут использоваться на разных этапах работы информационной системы: словарь, сопоставляющий словам ЕЯ лексику ИЯ, – на этапе построения семантической записи для текста; энциклопедический словарь – для преобразований семантической записи, необходимых на этапах уточнения записи, сравнения документа с запросом, сжатия записи и других.

8. Организация системы смысловых отношений.

В настоящем варианте языка ЭХО используется 50 элементарных смысловых отношений (ЭСО). О значении ЭСО дают представление их названия, понимаемые достаточно естественно. Формально каждое ЭСО определяется с трех точек зрения: а) допустимые семантические признаки его членов; б) взаимоотношения с другими ЭСО; в) некоторые логические свойства. Рассмотрим их отдельно.

Семантические признаки членов отношений. В настоящее время в системе используется 16 семантических признаков, которые частично упорядочены между собой отношениями включения. Так, признак "документ" является, с одной стороны, частным случаем (т.е. включается в) признака "предмет", а с другой стороны, частным случаем признака "информативность":

"документ" $\subset$ "предмет"; $\subset$ "информативность".

Кроме того, в качестве семантического признака может использоваться любое ЭСО.

Каждое ЭСО, кроме одного, имеют какие-то ограничения на семантические признаки заполняющих дескрипторов. Самое общее ЭСО "связано с (А, В)" допускает в качестве своих членов лю-

65111

БИБЛИОТЕКА
ИНСТИТУТА РУССКОГО ЯЗЫКА
ИНСТИТУТА ЯЗЫКОЗНАНИЯ АН СССР

бые дескрипторы. Перечислим несколько ЭСО с указанием семантических признаков заполняющих слов.

Отношение	Член А	Член В
связано с (А,В)	любой	любой
принадлежность (А,В)	любой	предмет
субъект (А,В)	одушевленность	действие
адресат (А,В)	одушевленность	действие
причина (А,В)	действие или ЭСО	действие
параметр (А,В)	параметр (А,)	предмет
локализация (А,В)	пространственность	предмет или действие
модальность (А,В)	модальность (А,) (примеры А: возможно, необходимо, наличие, отсутствие, излишек и др.)	предмет или финансовость или действие
специализация (А,В)	действие	учреждение или устройство
в ведении (А,В)	учреждение	учреждение

Очевидно, что если в качестве заполняющего выступает какой-то крупный признак (напр., "предмет"), то на данное место в ЭСО могут встать не только дескрипторы с единственным собственным признаком "предмет", но и дескрипторы с любым признаком, восходящим к "предмету" ("документ", "учреждение", а также "предмет + информативность", "предмет + финансовость").

Иерархия смысловых отношений. Каждое ЭСО связано хотя бы с одним другим ЭСО отношением включения. Так, ЭСО "связано с (А,В)" является самым общим и включает в себя любое другое отношение, в том числе отношение "принадлежность (А,В)". Последнее тоже включает несколько отношений, в том числе "актант (А,В)", делящееся, в свою очередь, на четыре еще более конкретных ЭСО: субъект, объект, адресат, с помощью. Это последовательное деление можно представить схематически:

связано с (A, B) $\supset$ принадлежность (A,B) $\supset$

$\supset$ актант (B,A) $\supset$ { субъект (B,A); объект (B,A); адресат (B,A); с помощью (A,B) }

Последние четыре отношения в нашем языке никак не уточняются, т.е. не включают в себя ни одно другое отношение. Заметим, что каждое такое неделимое отношение в принципе могло быть разделено на более мелкие (напр., отношение "субъект (,)" можно разделить на отношения "деятель (,)", "испытывающий (,)", "автор (,)" и др.; отношение "с помощью (,)" - на "посредник (,)", "средство (,)" и др.; "объект (,)" - на "создаваемый объект (,)", "изменяемый объект (,)" и др.). Уровень дробности смысловых отношений определяется пока эмпирически, отношения корректируются в ходе работы над семантическим анализом текста; возможно, что детальная разработка других разделов создаваемой системы потребует изменения принятого уровня дробности смысловых отношений.

Наличие иерархии смысловых отношений позволяет нам уточнять слишком общие смысловые связи между двумя понятиями. Так, если в ЭСО "связано с (A,B)" член B будет иметь признак "предмет", само отношение можно уточнить как "принадлежность (A,B)". Далее, если член A отношения принадлежности имеет признак "действие", его можно уточнить как "актант (B,A)", причем этот переход сопровождается переменой мест: утверждение "действие A является принадлежностью предмета B" равнозначно утверждению "предмет B является участником (актантом) действия A". Поскольку отношение актанта включает в себя несколько отношений (см. выше), оно может быть уточнено одним из них при наличии необходимых семантических признаков; так, для перехода к отношениям "субъект (B,A)" или "адресат (B,A)" необходимо, чтобы предмет B имел еще признак "одушевленность".

Понятие иерархии смысловых отношений эксплуатируется на разных этапах работы системы: при семантическом анализе текста, при варьировании семантической записи для целей сопоставления документа с запросом. При этом оказывается необходимым

проделывать путь не только от общих к более частным отношениям, но и наоборот, от частных к более общим.

Некоторые логические свойства отношений.

Часть из списка ЭСО имеют свойство симметричности:

равно (A,B) = равно (B,A)

связано с (A,B) = связано с (B,A)

одновременно с (A,B) = одновременно с (B,A).

Все симметричные, а также некоторые другие отношения имеют свойство транзитивности:

локализация (A,B) и локализация (B,C) ⟶

⟶ локализация (A,C)

сфера (A,B) и сфера (B,C) ⟶

⟶ сфера (A,C)

принадлежность (A,B) и принадлежность (B,C) ⟶

⟶ принадлежность (A,C).

Указанные свойства служат дополнительным средством преобразований (вариаций) семантической записи в процессе ее построения и эксплуатации.

9. Смысловые выводы.

Смысловым выводом мы называем такую ситуацию, когда из некоторых утверждений на ИЯ следуют некоторые другие. Последовательность строк, образующих посылку, должна быть связной (т.е. в каждой строке посылки должен быть член, общий хотя бы с одной другой строкой), а в выводе обязательно должны быть те члены, которые упомянуты в посылке.

Уже в словаре (энциклопедическая часть) для многих слов (обычно действий) содержится рамка для вывода. В качестве посылки в этом случае выступает набор строк, являющихся валентностями слова-дескриптора, в качестве вывода - новые отношения между дескрипторами, заполняющими валентность.

Примеры:

Дескриптор	Валентности	Выводы
ПРОСИТЬ	субъект (A, ПРОСИТЬ) адресат (B, ПРОСИТЬ) содержание (C, ПРОСИТЬ), где C - "действие".	субъект (B,C)

Просить товарища о помощи ⟶ Просить, чтобы товарищ помог

СЛАТЬ субъект (A, СЛАТЬ) ⟶ локализация (D, C)
объект (B, СЛАТЬ) ⟶ принадлежность (B, C)
адресат (C, СЛАТЬ) ⟶ локализация (D, B)
конечный пункт (D, СЛАТЬ)

из высказывания Послали ЭВМ Ин-ту в Новосибирск следует: 1) Ин-т находится в Новосибирске, 2) ЭВМ принадлежит Ин-ту, 3) ЭВМ находится в Новосибирске. Первое из этих утверждений представляет собой вывод в обратную сторону, или предпосылку.

Еще пример вывода:

КАУЗ объект (A, КАУЗ) ⟶ 1. специализация (B, A)
конечный пункт (B, КАУЗ), 2. объект (1., КАУЗ)
где A - "устройство",
B - "действие"

Ввод электростанции в действие ⟶ Каузирование такой ситуации, когда электростанция действует.

Кроме таких словарных выводов, заложенных в самом смысле слов (так как смысл слова описывается в нашей системе в основном набором валентностей), можно выделить группу выводов, строящихся на свойстве транзитивности некоторых отношений. Так, из высказывания "Иголка в яйце, яйцо в утке, утка в сундуке" следуют 3 вывода: "Иголка в утке", "Яйцо в сундуке", "Иголка в сундуке" (по транзитивности отношения локализации).

Третью группу выводов образуют так называемые уточняющие выводы, в которых и посылка, и вывод состоят из одной строки, причем вывод применим при соблюдении определенных условий.

субъект (A, B) ⟶ локализация (A, B)
при A = "административная единица":
Перевозки Горьковской области ⟶ Перевозки по (в, внутри, в пределах) Горьковской области";

объект (A, B) ⟶ сфера (A, B)
при A = "область деятельности":
Планирование сельского хозяйства ⟶ Планирование в области (в, в сфере) сельского хозяйства .

Наконец, существуют выводы, не зависящие от конкретных слов, а отражающие скорее законы действительности. Такие выводы никак пока не систематизированы, они накапливаются в ходе работы над материалом. Формулируются они также в терминах смысловых отношений.

специализация (A,B)
типичная специализация (C,B) } → сфера (A,C)

ВЦ планирует → Автоматизация в планировании .

содержание (A,B)
цель (C,A) } → содержание (C,B)

План исследований для создания... → План создания .

принадлежность (A,B)
локализация (C,B) } → локализация (C,A)

ЭВМ Хабаровского Ин-та" → "ЭВМ в Хабаровске".

объект (A,B)
специализация (B,C) } → специализация(A,C)

Автоматизация обработки информации → ВЦ для обработки информации .

локализация (A,B)
объект (C,B) } → локализация (A,C)

Прием сырья в Горьком → Сырье в Горьком .

Применяя операцию вывода к некоторой СЗ, мы получаем ее вариацию. Вариаций исходной СЗ посредством вывода может быть столько, сколько существует фрагментов записи, образующих посылку какого-либо вывода. К любой вариации СЗ снова может быть применена операция вывода, дающая новую вариацию; цепочка выводов может применяться до тех пор, пока получаемые вариации не станут повторять уже имеющиеся.

10. Смысловое сжатие.

После того как получена семантическая запись текста, можно подвергать ее сжатию - структура записи легко его допускает. В основе процедуры сжатия лежит: повторяемость членов отношений, грамматика смысловых отношений и наличие словарных энциклопедических сведений.

Смысловое сжатие может иметь разные пределы: а) устранение чистой избыточности, т.е. строк, повторяющихся внутри текста; б)устранение мало информативных (с какой-либо точки зрения) строк внутри текста; в) устранение части информации, содержащейся в других текстах на ту же тематику. Последний вид сжатия может, очевидно, совмещаться с избирательным распределением информации (ИРИ).

Повторяющиеся строки СЗ могут быть или следствием избыточности выражения в естественном тексте, или результатом многословия автора, или, наконец, результатом принятого способа семантического анализа, когда дважды (по информации, поступающей от главного и от зависимого слова) образуется одна и та же строка. В любом случае необходимо устранять повторяющиеся строки. Повторение может быть достаточно сложным, например, смысловые отношения двух строк могут не совпадать, а образовывать отношение включения; члены совпадающих отношений тоже могут сложным образом соответствовать друг другу (частично отсутствовать, быть местоимением, быть частным случаем один другого). Поэтому операции сжатия часто должно предшествовать доказательство того, что данные строки действительно повторяют друг друга.

Сжатием повторяющейся информации является также устранение из СЗ тех строк, которые содержатся и могут быть извлечены из энциклопедического словаря. Так, если в СЗ появился термин в развернутом виде (напр., "электронно-цифровая вычислительная машина", см. запись ниже), причем в словаре он представлен в более кратком виде (ЭЦВМ), в записи остается его краткая форма.

Если в правиле вывода посылка содержит больше информации, чем сам вывод, то применение вывода можно также считать операцией сжатия СЗ (см. раздел 9).

Кроме того, некоторые ЭСО по своему значению мало информативны и могут быть устранены, если ставится задача оставить только основные темы. Это отношения "часть (,)", "параметр (,)", "стадия (,)", "пункт (,)", "аспект (,)",

"единица измерения (,)", "количество (,)" и некоторые другие.

Задача автоматического определения основных тем документа по его семантической записи представляется достаточно самостоятельной (она опирается не только на сжатие ненужной информации, но и на нахождение главной части информации по структуре семантической записи и на основании шаблона, по которому построены анализируемые тексты) и здесь рассматриваться не будет.

II. Информационная запись документа как множество вариаций его семантической записи.

"От хорошей поисковой системы хотелось бы требовать, чтобы между двумя переводами текстов на ее язык формальное соответствие имело место тогда, и только тогда, когда между исходными текстами имеет место содержательное соответствие" [2]. Это требование достаточно очевидно выполняется для нашей записи. Действительно, поскольку запись отражает в принципе всю информацию исходного текста, в том числе все смысловые связи, одинаковые записи могут иметь только тексты, выражающие один и тот же смысл. Но этого недостаточно. Для того чтобы система могла давать точные ответы (документальные или содержательные), требуется, чтобы всякое смысловое соответствие приводилось к формальному соответствию.

Несовпадение окончательных СЗ документа и запроса еще не свидетельствует о том, что ⟵⟶ документ-запрос несоотносимы. Так, документ О направлении СО АН СССР "Минска-22" и запрос Наличие ЭЦВМ в Новосибирске или Средства вычислительной техники (ВТ) в Сибири безусловно соотносимы, хотя имеет место полное несовпадение их лексического состава и семантических записей. Мы уже говорили выше, что за каждой СЗ стоят словарные энциклопедические сведения и грамматика смысловых отношений, которые позволяют создавать различные вариации исходной СЗ. Варьирование может заключаться и в изменении лексического состава (напр., "Минск-22" ⟶ ЭЦВМ ⟶ ВТ), и в применении вывода (слать ⟶ иметь), и в ис-

пользовании правил грамматики:

локализация (Сибирь, СО АН СССР) ⟶
принадлежность (Минск-22, СО АН СССР) ⟶

⟶ локализация (Минск-22,Сибирь), или, что то же самое, ⟶ принадлежность (ВТ,Сибирь). Это последняя строка уже полностью совпадает с СЗ запроса.

Итак, чтобы доказать, что документ-запрос несоотносимы, нужно образовать все вариации СЗ и документа, и запроса, и убедиться, что ни при каких вариациях между этими СЗ не существует формального совпадения.

Если исходить из содержательного определения информационной записи документа как записи, в которой отражен и весь смысл, непосредственно передаваемый в документе, и все вытекающие из него утверждения, то в нашей системе эту роль выполняет семантическая запись документа плюс все ее вариации. При этом неважно, существует ли информационная запись документа в явном виде (т.е. для каждого документа строится его СЗ и сразу все ее вариации, безотносительно к запросу) или она существует как возможность, заложенная в грамматике смысловых отношений и словаре энциклопедических сведений, и при поступлении каждого конкретного запроса строятся только некоторые вариации СЗ, диктуемые данным запросом.

12. Система поиска информации с помощью языка ЭХО.

Начиная с приведенной СЗ, любое состояние записи может рассматриваться в качестве поискового образа (ПО), служащего для документального или фактографического поиска информации. На первой стадии поиска в качестве ПО можно использовать самый краткий вид записи - перечисление тем. Для этой стадии выгодно инвертированное устройство памяти, существенно сокращающее время поиска. В этой части работа системы ничем не будет отличаться от работы систем дескрипторного типа, имеющей целью определение множества потенциально релевантных документов. Но то, что является выходом в дескрипторной системе, будет у нас входом в следующую стадию поиска. На второй стадии поиска из множества потенциально

релевантных документов должно быть выделено множество действительно релевантных. Цель этой стадии - выяснение, в тех ли смысловых отношениях, которые интересуют потребителя, находятся темы в документе. Здесь будет осуществлен сплошной просмотр всех ПО документов, выделенных на первой стадии, причем в качестве ПО будет использована самая полная СЗ.

В качестве ответа могут выдаваться и номера документов, оказавшихся действительно релевантными (документальная выдача), и переводы СЗ (или частей СЗ, соответствующих запросу) этих документов на ЕЯ (фактографическая выдача). При этом может ставиться задача суммирования информации, содержащейся в семантических записях всех релевантных документов.

13. Заключение.

Предлагая в качестве ИЯ язык семантической записи, сошлемся на доклад [3], в котором подчеркивается необходимость планомерного применения в информатике глубоких семантических методов. "При этом для информатики как прикладной дисциплины, для которой фундаментальным понятием является пертинентность, необходимая степень раскрытия семантических свойств сообщений более глубока, чем для других практических целей. В частности, эта степень глубже, чем для "автоматического перевода" (где в принципе можно ограничиться выявлением смысловой эквивалентности сообщений, не вдаваясь в выяснение более сложных смысловых отношений между разными сообщениями" [3].

Описанный семантический язык применялся пока к отдельным фразам, представляющим собой краткие резюме экономических документов. Следующий принципиальный шаг в разработке системы - переход к семантическому анализу целых текстов.

Описанная работа ведется в Лаборатории машинного перевода МГПИИЯ. Создан и сдан для программирования алгоритм морфологического анализа фраз (авторы - канд. филол. наук В.З. Санников, С.А. Кодзасов), создан и проходит ручную проверку алгоритм естественно-семантического анализа, составляется словарь, на который опирается семантический

анализ.

Структура принятого семантического языка и вопросы семантического анализа текста отражены в [13] и [14]. Возможное применение семантического языка и метода анализа к текстам номенклатурных наименований описано в [15]

Литература.

1. Э.С.Бернштейн, Д.Г.Лахути, В.С.Чернявский. О критериях оценки поисковых систем. НТИ, 1964, № 3.

2. Д.Г.Лахути. Семантическая поисковая система "Пусто-Непусто-2". Сб. "Системы управления непромышленными предприятиями". М., 1969.

3. Г.Э.Влэдуц, В.А.Успенский, Ю.А.Шрейдер. Семантика и научная информация (семантические аспекты информатики). Доклад на Международном Симпозиуме стран-членов СЭВ. М., ВИНИТИ, 1970.

4. Perry J.W., Kent A. Tools for Machine Literature Searching. N.-Y., 1958.

5. R.C.Cros, J.C.Garden, F.Levy. L'automatisation des recherches documentaire, un modèle général: le SYNTOL.Paris,1964.

6. Информационно-поисковая система "БИТ". Киев, 1968.

7. Н.И.Иванова, В.Б.Маргаритов. Введение элементов грамматики в дескрипторный информационно-поисковый язык по химии и химической промышленности. НТИ, 1968, серия 2, № 10.

8. Машинный перевод, вып. 2. Изд. ИТМ и ВТ АН СССР, М., 1960.

9. Машинный перевод и прикладная лингвистика, вып. 8.

М., 1964.

10. А.К.Жолковский, И.А.Мельчук. О семантическом синтезе. Проблемы кибернетики, вып. 19. М., 1967.

11. Ю.Д.Апресян, А.К.Жолковский, И.А.Мельчук. К построению действующей модели языка "смысл - текст". Машинный перевод и прикладная лингвистика, вып. II. М., 1969.

12. Е.В.Падучева. Семантический анализ отрицательных предложений при переводе на язык исчисления предикатов. Тезисы "Третьей Всесоюзной конференции по информационно-поисковым системам и автоматизированной обработке научно-технической информации". М., 1969.

13. Н.Н.Леонтьева. Об одном способе представления смысла текста. Сб. "Информационно-поисковые системы и автоматизированная обработка научно-технической информации", т. 2. М., 1967.

14. Н.Н.Леонтьева. Семантический анализ и смысловая неполнота текста. Автореферат. М., 1968.

15. Н.Н.Леонтьева, Ю.С.Мартемьянов, В.Ю.Розенцвейг. О выявлении и представлении смысловой структуры текстов экономических документов (в печати).

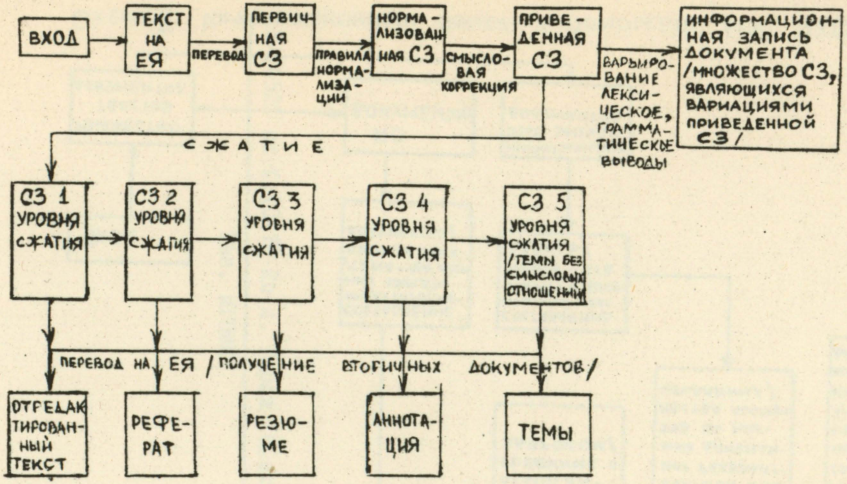


Схема I. Построение и преобразования ИЯ ЭХО.

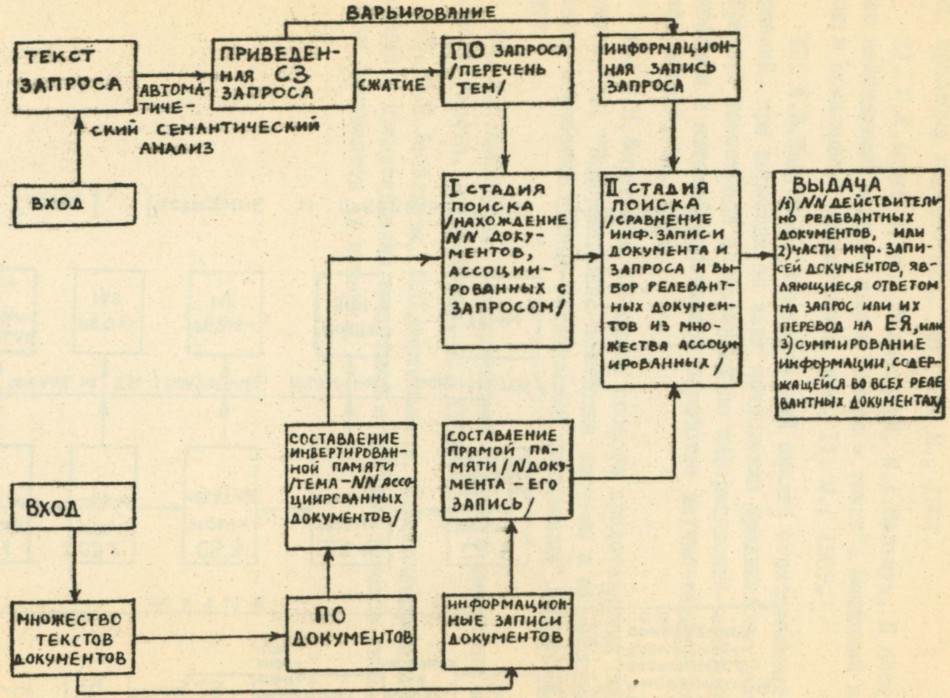
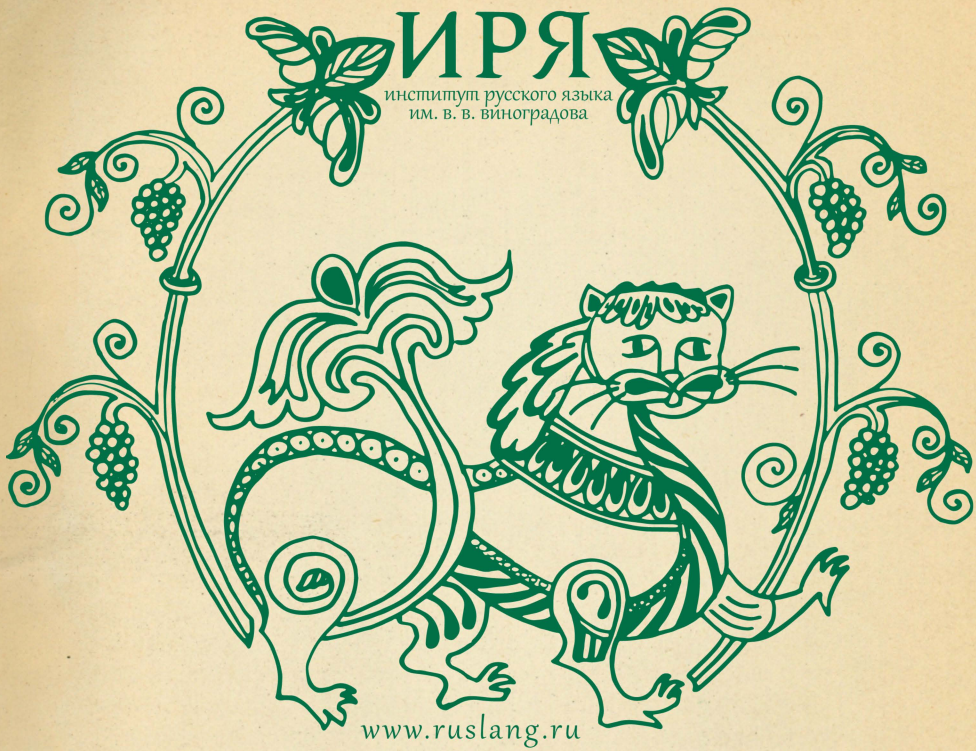


Схема II. Информационный поиск с помощью ИЯ ЭХО

БИБЛИОТЕКА ИН-ТА РУССКОГО ЯЗЫКА АН СССР



Т-14521 от 1.10.70 2п.л. Тираж 100экз. Зак.766

Тип. МГПИИЯ