АВТОМАТИЗИРОВАННЫЕ ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

Порядок функционирования автоматизированной информационно-поисковой системы

Автоматизированная информационно-поисковая система (АИПС) предназначена для ввода, обработки, хранения и поиска семантической информации. Поиск семантической информации предполагает сравнение смыслового содержания запроса со смысловым содержанием хранящихся в АИПС документов. Такая операция возможна, когда существует некоторый язык представления информации, позволяющий однозначно описывать смысловое содержание документов и запросов. Естественный язык для этой цели не подходит в силу своей многозначности и высокой сложности. При наличии такого языка, который носит название **информационно–поискового языка (ИПЯ)**,

1. перевод содержания документа и /или запроса с естественного языка на ИПЯ (**процесс индексирования текстов**). В результате индексирования полный текст документа (запроса) заменяется некоторой характеристикой, кратко отражающей его смысловое содержание. Эта характеристика носит название **поискового образа документа (ПОД)** и/или **поискового образа запроса (ПОЗ)**. Иногда ПОЗ называют **поисковым предписанием (ПП)**;

2. представление ПОДов и ПОЗов в машинных кодах (**кодирование**). Часто этот этап выполняется совместно с предыдущим. Организация массивов ПОДов и ПОЗов. Обработка элементов этих массивов и представление их в виде, наиболее удобном для

3. **поиск информации**, т.е. выделение из поискового массива тех документов, содержание которых соответствует поисковому предписанию. Эта операция осуществляется в соответствии с некоторым **критерием смыслового соответствия (КСС)** поискового образа документа поисковому образу запроса (критерий выдачи);

4. **выдача пользователю информации**, соответствующей отобранным ПОДам;

5. **корректировка запросов или ПП** и повторение предыдущих этапов. Эта операция выполняется в том случае, если потребитель не удовлетворен работой АИПС, и может производиться либо в пакетном режиме, либо в режиме диалога.

Выходной продукцией АИПС могут быть: оригиналы, копии или адреса документов; данные и факты, содержащиеся в документах в явном виде; факты, данные, сведения, которые в явном виде не содержатся во введенных документах.

В связи с этим различают следующие АИПС:

1. **документальные** (выдают оригиналы, копии документов или адреса введенных документов);
2. **фактографические** (выдают данные, факты, сведения, содержащиеся в явном виде во введенных документах);
3. **информационно-логические** (выдают данные, факты, сведения, которые в явном виде не вводились в АИПС, а получены в результате некоторого логического вывода).

Состав и структура АИПС

АИПС, также как и любая АИС является весьма сложной системой. Можно выделить несколько различных декомпозиций и, соответственно, представлений АИПС, каждая из которых описывает систему с определенной точки зрения и на различных уровнях детализации.

Наиболее необходимы для изучения АИПС следующие пять декомпозиций:

1. **функциональная декомпозиция**, т.е. разбиение АИПС на функциональные составляющие (подсистемы);

2. **покомпонентная декомпозиция**, т.е. разбиение АИПС, позволяющее выделить ее информационные, программные, технические и трудовые компоненты;

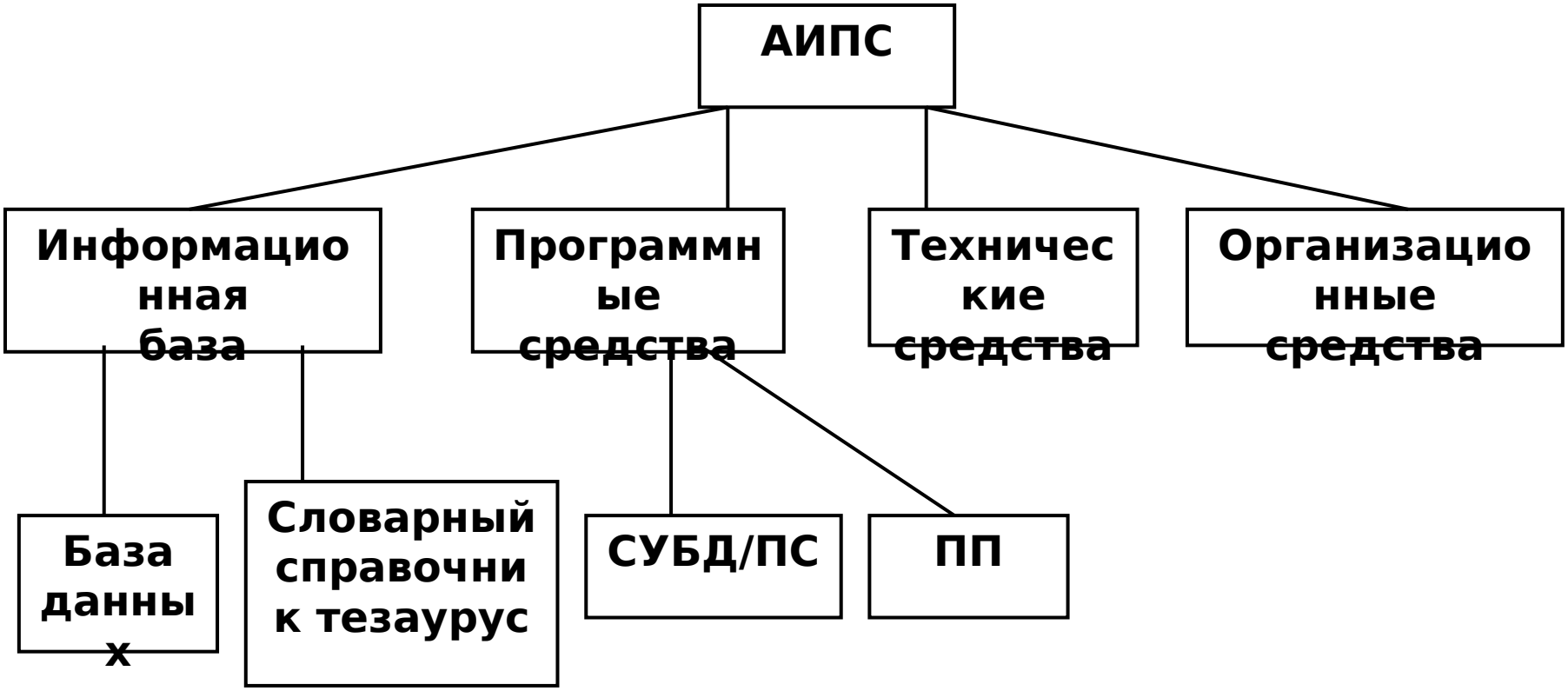
3. **декомпозиция на обеспечивающие составляющие**, т.е. разбиение АИПС на обеспечивающие подсистемы;

4. **организационная декомпозиция** - декомпозиция АИПС на организационные составляющие;

Функциональная декомпозиция - декомпозиция на функциональные подсистемы. При такой декомпозиции наиболее рационально выделять следующие функциональные подсистемы АИПС:

- отбора информации из внешней среды;
- предмашинной обработки и ввода информации;
- обработки и хранения информации;
- поиска и выдачи информации;
- информационного обслуживания потребителей информации.

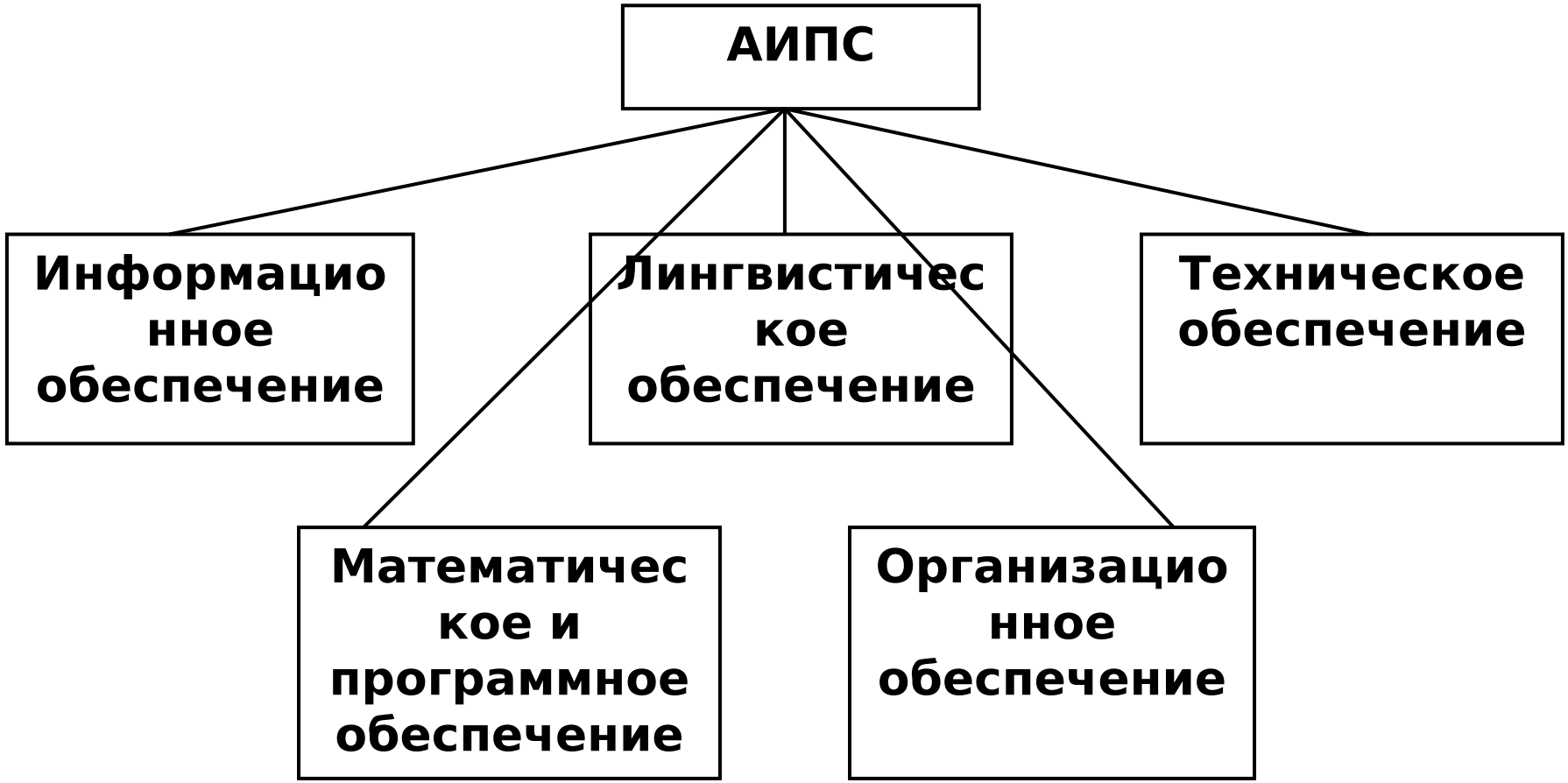
Покомпонентная декомпозиция. Такая декомпозиция вызвана необходимостью самостоятельного рассмотрения информационной, программной и технической среды АИПС. С этих позиций в составе АИПС целесообразно выделить: информационную базу (базу данных, словари, справочники и т.д.), программные средства (СУБД/ ПС, пользовательские программы - software АИПС); технические средства (hardware АИПС), организационные средства (рисунок). Большинство функций предыдущей (функциональной) декомпозиции реализуются соответствующими техническими программными и информационными средствами покомпонентной декомпозиции. При этом почти все функциональные подсистемы (кроме подсистемы отбора) используют соответствующие программные и технические средства. Обе


АИПС
Информационная база
Программные средства
Технические средства
Организационные средства
База данных
Словарный справочник тезаурус
СУБД/ПС
ПП

Декомпозиция на обеспечивающие составляющие. Обеспечивающими составляющими или подсистемами АИПС называют элементы, которые обеспечивают реализацию заданных функций АИПС.

В АИПС обычно выделяют следующие обеспечивающие подсистемы (рисунок):
- информационного обеспечения;
- лингвистического обеспечения;
- математического и программного обеспечения;
- технического обеспечения;
- организационного обеспечения.

Подсистема **лингвистического обеспечения** включает совокупность словарей, справочников, положений и инструкций предмашинной и машинной обработки и поиска информации.


АИПС
Информационное обеспечение
Лингвистическое обеспечение
Техническое обеспечение
Математическое и программное обеспечение
Организационное обеспечение

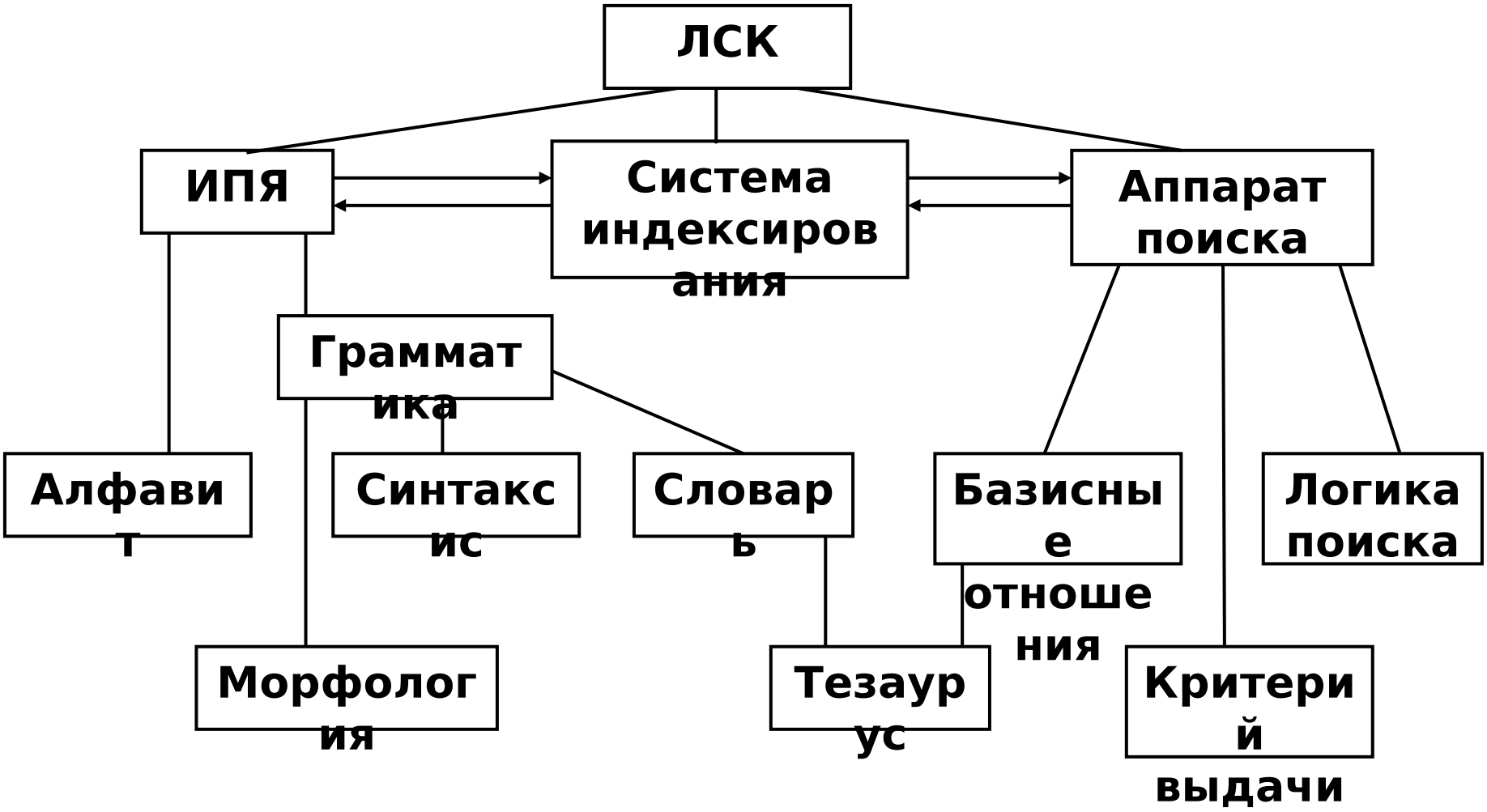
Организационная декомпозиция АИПС. Такая декомпозиция соответствует организационной структуре информационного института, центра или иной организации, в структуру которой входит АИПС.

Среди элементов организационной декомпозиции могут быть: вычислительный центр, отделы или лаборатории.

Декомпозиция на обеспечивающие подсистемы в чем-то перекрываясь с покомпонентной декомпозицией, тем не менее представляет новую точку зрения на состав и структуру АИПС.

Логико-семантический комплекс АИПС. Логико-семантический комплекс (ЛСК) - комплекс языковых, логических, и математических средств формализованного представления семантической информации с целью ее автоматизированной обработки и поиска (рисунок).

ЛСК представляет собой теоретическую и практическую базу создания и функционирования как каждой составляющей всех ранее рассмотренных декомпозиций АИПС, так и АИПС в целом.


ЛСК
ИПЯ
Система индексирования
Аппарат поиска
Грамматика
Алфавит
Синтаксис
Словарь
Базисные отношения
Логика поиска
Морфология
Тезаурус
Критерий выдачи

Информационно-поисковые языки (ИПЯ)

В последние годы создаются самые разнообразные искусственные языки, ориентированные на определенный аспект решаемых задач.

Это языки описания данных, информационно-поисковые языки, языки моделирования, управления заданиями, автоматизации проектирования, языки манипулирования данными и т.д.

Описать все разнообразие существующих языков или тем более дать их исчерпывающую классификацию не представляется возможным. Среди множества классов искусственных языков нас интересуют только **информационно-**

Основные элементы ИПЯ

Для определения роли и места ИПЯ рассмотрим основные понятия языков, которые тесно связанны с информационно-поисковыми языками и некоторые из описаны в стандарте ГОСТ 7.74–96.
Язык - это знаковая система любой физической природы, выполняющая познавательную и коммуникативную функции и процессе человеческой деятельности. Естественный язык (ЕЯ) есть особого рода преобразователь заданных смыслов в тексты, и наоборот.
Информационный язык - формальная семантическая система, включающая алфавит, правила образования конструкций, их преобразования и интерпретации и предназначенная для описания, обработки,

Информационно-поисковый язык - специализированный искусственный язык, предназначенный для описания основного содержания (центральной темы) и формальных характеристик документов с целью информационного поиска.

Алгоритмический язык - язык, предназначенный для записи информации и алгоритмов ее обработки в форме, воспринимаемой ЭВМ. Каждый из названных языков предназначен для описания языковых объектов и, следовательно, в той или иной мере обладает смысло-выразительной способностью, т.е. способностью выражать смысловое содержание текстов. Указанная способность зависит от того, на каких уровнях представляются языковые объекты средствами данного языка.

Различают следующие **уровни представления** языковых объектов.

Семантика - основные закономерности строения внутренней (смысловой) стороны языковых объектов. Семантический уровень представления языковых объектов позволяет отобразить их смысловое содержание, выразить связь смыслов отдельных знаков со смыслом текста (связь смысла языковых объектов между собой и со смыслом образуемого ими более сложного языкового объекта).

Синтаксис - основные закономерности, определяющие отношения между единицами языка в пределах конкретных текстов. Синтаксический уровень представления языковых объектов позволяет выразить их структуру.

Морфология - основные закономерности построения слов языка, т.е. система грамматических категорий и способов их выражения.

Правописание - система правил, устанавливающая единообразные способы передачи речи на письме.

Фонетика - основные закономерности поведения речевого аппарата и способы их использования.

Указанные уровни представления языковых объектов позволяют описать преобразование: звук - фонема - морфема - слово - текст – смысл. ИПЯ представляют языковые объекты на 1, 2, 3, 4 уровнях. Однако арсенал средств ИПЯ для представления языковых объектов на семантическом уровне менее развит по сравнению

Основными **элементами ИПЯ** являются: **алфавит**, **лексика** и **грамматика**.

Алфавит ИПЯ - система знаков, используемых для записи слов и выражений ИПЯ. Это могут быть буквы русского и/или английского языка, знаки препинания, арабские цифры, любые иные символы.

Лексика, или словарный состав ИПЯ, - совокупность слов, словосочетаний и выражений, используемых для построения текстов ИПЯ.

В качестве лексических единиц ИПЯ могут быть использованы:

• слова, фрагменты слов, словосочетания и выражения любого естественного языка;

• коды и шифры (цифровые, буквенные, буквенно-цифровые) словосочетаний, слов и выражений, выступающие в роли имен соответствующих классов;

• шифры и коды в сочетании со словами, словосочетаниями и выражениями.

Существуют различные способы задания словарного состава ИПЯ, в том числе:

1. перечисление всех лексических единиц ИПЯ;
2. перечисление части лексических единиц и задание правил формирования из них других лексических единиц;
3. задание правил построения лексических

Первый способ задания лексики требует больших интеллектуальных усилий. Процесс построения лексики нельзя автоматизировать. Лексика ИПЯ оказывается жестко фиксированной и в ряде случаев не позволяет достаточно точно выразить смысловое содержание текстов.

Третий способ поддается полной автоматизации, хотя и требует больших интеллектуальных затрат на определение правил формирования лексики. Однако научный подход к формированию словарного состава делает его более совершенным, обеспечивает единообразие и уменьшает субъективизм при построении лексики.

Второй способ занимает промежуточное положение и в отношении интеллектуальных усилий, и в отношении автоматизации процессов.

Грамматика ИПЯ - совокупность средств и способов построения, изменения и сочетания лексических единиц.

Грамматика включает **морфологию** и **синтаксис**.

Морфология - совокупность средств и способов построения и изменения слов.

Синтаксис - совокупность средств и способов соединения слов в выражения и фразы.

Требования к ИПЯ

К информационно-поисковым языкам, его конструкциям и элементам могут быть предъявлены следующие **требования**:

1. **ИПЯ должен располагать лексико-грамматическими средствами** для точного выражения основного содержания (центральной темы или предмета) текста.
2. **ИПЯ не должен быть двусмысленным**. Любое выражение ИПЯ должно пониматься вполне однозначно в силу того, что приемником текстов ИПЯ является программно-техническая система, а не человек.
3. **ИПЯ не должен содержать элементы, отображающие волевое побуждение, эмоци**и и т.д. Выражение ИПЯ, его значение, смысл не должны зависеть от «настроения» приемника

Типы отношений между словами ИПЯ

Построение выражений ИПЯ требует решения, по крайней мере, двух проблем.

Первая из них связана с выбором слов (лексических единиц) из множества лексических единиц ИПЯ, необходимых для построения выражений. Здесь решается вопрос, какие использовать слова по принципу «или-или» (или то слово - или иное слово). Выбор слов определяется их смысловыми значениями, обусловленными отношениями между предметами и явлениями, которые они определяют. Такие отношения называются **парадигматическими**.

Вторая проблема построения фраз ИПЯ связана с определением последовательности употребления или написания выбранных слов (словосочетаний), поскольку в каждый данный момент может быть использовано только одно слово (словосочетание), лексические единицы могут следовать одна за другой, но не одновременно.

Отношения, устанавливаемые при соединении слов в словосочетания и фразы, носят название **синтагматических** отношений.

Парадигматические отношения - это отношения, обусловленные наличием не языковых, а логических связей между предметами и явлениями, обозначенными данными словами. Наиболее важны следующие парадигматические отношения:

1. **«вид-род»**, например, «шкаф-мебель». В данном случае понятие «шкаф» является видовым но отношению к понятию «мебель» - понятие мебель» является родовым по отношению к понятию «шкаф». Родовое понятие всегда включает в себя видовое;
2. **«часть-целое»**, например «лезвие-нож». Лезвие является частью ножа;
3. **«причина-следствие»**, например «лампа-свет»;
4. **«функциональное сходство»**, например «лопата-экскаватор», "телега-автомобиль».

Естественный язык обладает высокой многозначностью. Это создает богатство его форм и содержания. При написании текстов (особенно художественных) стремятся использовать эту многозначность для придания тексту элегантности, литературности. В ИПЯ недопустима многозначность. Поэтому здесь необходимо учитывать отношения **синонимии** и **омонимии** слов ЕЯ, используемых в ИПЯ.

Синтагматические отношения - это совокупность всех отношений, реализуемых синтаксисом ИПЯ. С этой точки зрения синтаксис представляет собой совокупность способов и средств выражения синтагматических отношений. Простейшим видом синтагматических отношений является отношение вхождения нескольких лексических единиц ИПЯ в один и тот же текст, фрагмент текста, фразу и т.д., т.е. отношение координации.

Парадигматика и синтагматика - это два различных аспекта ИПЯ, первый связан с его лексикой, второй - с грамматикой.

Многообразие используемых в ИПЯ парадигматических и синтагматических отношений определяет смысловыразительную способность или **семантическую силу ИПЯ**.

Классификация ИПЯ

По характеру использования грамматических средств различают **прекоординированные** и **посткоординированные** ИПЯ.

Прекоординированные ИПЯ - это ИПЯ, словарный состав которых жестко связан грамматическими средствами в единую структуру. Лексика и грамматика такого языка, а также синтаксис, морфология, все парадигматические и синтагматические отношения самостоятельно не существуют, а образуют единую жесткую связанную структуру.

Индексирование текстов (перевод текстов на ИПЯ) выполняется только с использованием элементов такой жесткой структуры. По сути дела, каждый ИПЯ этого типа представляет собой некоторую систему классификации.

На практике используются два типа **классификационных** ИПЯ.

Первый тип - перечислительные классификации. В таких классификациях жесткая структура понятий языка определена заданием одного или нескольких отношений, устанавливающих взаимосвязь понятий и обеспечивающих попадание любого объекта в один единственный класс. Различают два вида перечислительных классификаций - **иерархические** и **алфавитно-предметные (рубрикационные**) классификации.

Иерархические классификации получают заданием отношений древесного порядка (или совокупности таких отношений, пересечение или объединение которых позволяет получить удовлетворительную классификацию). Примерами таких классификаций являются: Десятичная классификация Дьюи, Классификация Библиотеки

Алфавитно-предметные классификации получают заданием отношений алфавитного порядка на семействе множеств лексических единиц ИПЯ, определяемых предметными классами понятий. Примерами таких классификаций являются структуры различного рода каталогов и указателей.

Второй тип систем классификаций - **фасетные классификации.**
В этих классификациях существуют несколько жестких структур (фасетов), каждая из которых отображает один аспект отношений между словами ИПЯ.

Построению фасетных классификаций предшествует фасетный анализ, в результате которого вся лексика ИПЯ разбивается на

Посткоординированные ИПЯ - ИПЯ, словарный состав которых не связан грамматикой заранее и такая связь осуществляется в процессе индексирования и/или поиска.

Выделяют три типа таких ИПЯ:

1. **дескрипторные** ИПЯ;
2. **семантические коды** (RX-коды, семантический код Перри-Кента) и
3. **синтагматические** ИПЯ (например, СИНТОЛ).

Иногда выделяют ИПЯ:

1. **классификационные** (к ним относят иерархические и фасетные классификации);
2. **рубрикационные** (алфавитно-предметные классификации);
3. **дескрипторные** (все посткоординировапиые ИПЯ).

По способу образования словарного состава различают:

4. ИПЯ с жестким словарем, задаваемым перечислением всех лексических единиц языка;
5. ИПЯ со свободным словарем, задаваемым перечислением определенной части лексических единиц и правилами образования новых лексических единиц,
6. ИПЯ без словаря, в котором лексические единицы заранее не фиксируются, а задаются

По характеру словаря различают ИПЯ:

1. со словарем ключевых слов;
2. со словарем словосочетаний;
3. со словарем дескрипторов;
4. с тезаурусом.

По наличию парадигматических отношений:

5. ИПЯ с базисными отношениями;
6. ИПЯ баз базисных отношений.

По учету синтагматических отношений различают ИПЯ:

7. со слабой синтагматикой (вхождение слова в текст);
8. с неразвитой синтагматикой (учитываются указатели роли и связи);
9. с развитой синтагматикой (указатели роли и связи, позиционная грамматика и т.д.).

Дескрипторные ИПЯ

В основе построения дескрипторных ИПЯ лежит **принцип координатного индексирования**, который предполагает, что основное смысловое содержание документа может быть выражено списком ключевых слов, т.е. списком наиболее существенных для понимания текста назывных полнозначных слов. **Полнозначные слова** - существительные, прилагательные, глаголы, наречия, числительные, местоимения. Неполнозначные слова - предлоги, союзы, связки, частицы.

Принцип чистого координатного индексирования и поиска состоит в индексировании документов и запросов списками ключевых слов, являющихся ПОДами

Метод координатного индексирования и поиска

Пусть задано универсальное множество ключевых слов $A = \{a_1, ..., a_n\}$ и некоторое множество документов $T = \{d_1, ..., d_m\}$.

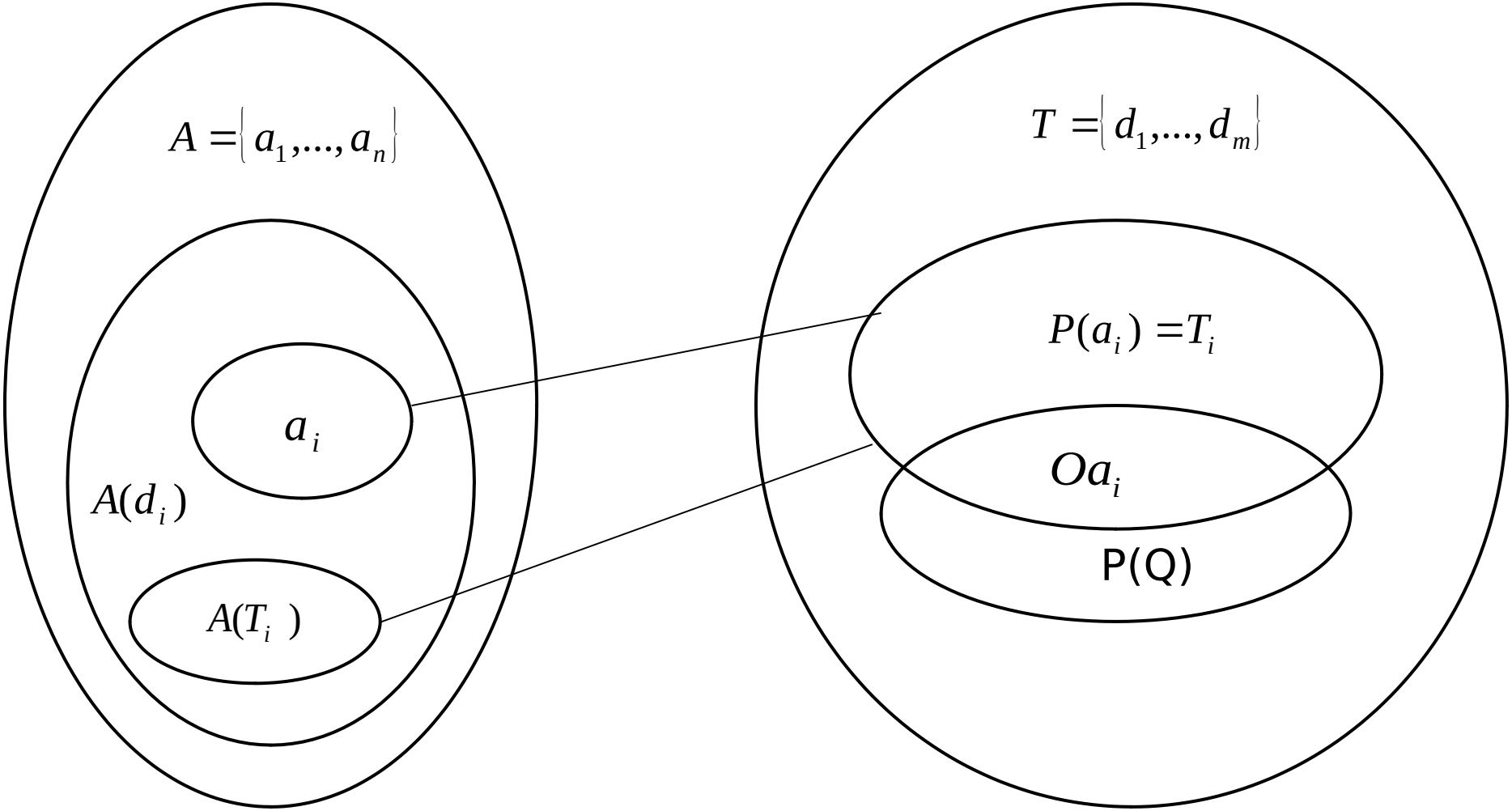
$\overline{a_i} = \{d_{i1}, ..., d_{i\alpha}\} = T_i = P(a_i)$ или $P : A \to 2^T$.

Пусть далее 2^T есть $\forall a_i \subset A$, поисковый образ документа $P(a_i) = \{d_{i1}, ..., d_{i\alpha}\} = T_i$, т.е. существует такое отображение множества документов А в множество $A(T_i)$, что T_i, $P[A(T_i)] = \{d_{i1}, ..., d_{i\alpha}\} = T_i$, $A(T_i) \subset A$.

Пусть $A(d_i)$ - подмножество документов с ПОДом, равным d_i, $T_i = \{d_{i1}, ..., d_{i\alpha}\}$, т.е. d_{ik} $A(d_{ik})$.

$A(T_i) = \cap A(d_{ik}), A(d_{ik}) \subset A$

Обозначим подмножество документов, содержащих через


$A = \{a_1, \ldots, a_n\}$
$T = \{d_1, \ldots, d_m\}$
a_i
$A(d_i)$
$A(T_i)$
$P(a_i) = T_i$
Oa_i
P(Q)

Рассмотрим запрос Q, поисковый образ которого $P(Q) = \{ d_{i1}, ..., d_{i\beta} \}, d_{ik} \in T_i$

Документ a_i отвечает на запрос Q (релевантен Q), если $|P(Q) \cap P(a_i)| \geq K.$

$$\left| P(Q) \cap P\left[A(T_i) \right] \right| \geq K$$

Подмножество $A(T_i)$ релевантно запросу Q, если (рисунок).

В соответствии с принципом чистой координации документ a_i выдается на запрос Q в том случае, если их поисковые образы имеют не менее K общих ключевых слов.

При использовании чистой координации при поиске могут возникнуть следующие **нежелательные ситуации**:

1. **ложная координации** (в массиве, выданном на запрос, может быть документ, которые не отвечает запросу);
2. **неполная координация** (выдача документа, несоответствующего запросу);
3. **синонимия ключевых слов** (выдача отсутствует, хотя необходимо было выдать документ, содержащие синоним искомого термина);
4. **полисемия** (выдача ненужных документов);
5. **необозначенность родо-видовых** (парадигматических) **связей** (выдача отсутствует, хотя необходимо было выдать документ, содержащие родо-видовую связь с искомым термином);

Для ликвидации указанных недостатков необходимы:

1. устранение синонимии, полисемии, омонимии;
2. учет парадигматических связей;
3. учет синтагматических связей.

Состав и структура дескрипторных ИПЯ

Основными **элементами ДИПЯ** являются:

1. **Словарь лексических единиц (ЛЕ)**, обеспечивающий выделение определенных частей текста и их замену на коды лексических единиц.
2. **Правила применения ИПЯ** (грамматика), определяющие процедуру перевода текстов документов и запросов (слов и словосочетаний - морфология; фраз, текстов в целом - синтаксис) с естественного языка на ИПЯ.
3. **Правила построения и ведения ИПЯ**, определяющие процедуру изменения и совершенствования ИПЯ, т.е. его словаря и правил применения.

Анализ информации и построение словарей

Задача построения словарей состоит в следующем: по заданному классу текстов необходимо выбрать попарно-различимые лексические единицы (словоформы, основы слов, КС. дескрипторы и т.д.), определить их морфологические, синтаксические и семантические характеристики и расположить в заранее обусловленном порядке.

Существуют три способа построения словарей: **априорный**, **апостериорный**, **динамический**.

Априорный. Лексические единицы выделяются из различных терминологических источников (справочников, энциклопедий, словарей, классификаторов и т.д.) по заданной тематике. После отбора лексики проводят ее семантическую обработку и строят словари.

Апостериорный. Лексика формируется из представительной выборки будущего фонда документов. Далее проводят ее семантическую обработку и строят словари.

Динамический способ. Процессы накопления лексики, ее семантическая обработка и построение словарей совмещены с процессом эксплуатации ИПС.

Принципы отбора лексических единиц

В настоящее время не существует методов построения оптимальных словарей. Наука и практика располагает лишь определенными принципами построения более или менее хороших словарей.

Эти принципы базируются на свойствах слов и текстов естественного языка, таких как информативность слов, омонимия, синонимия и полисемия слов и фраз; синтаксическая эквивалентность фраз; отношения между словами; изменение со временем значений слов; ненормализованность слов и т.д.

При построении словарей приходится решать **три основные проблемы:**

1. Какие слова включать в словарь?

2. Какие учесть типы отношений?

Решение первой проблемы в основном базируется на учете синонимии, омонимии, полисемии, а также информативности слов, косвенным показателем которой является частота их встречаемости в текстах.

С учетом этого **принципы отбора слов** при решении первой проблемы состоят в следующем:

1. не включать в словари редких терминов;
2. исключать общие понятия с высокой частотой встречаемости;
3. в каждый класс понятий вводить слова только с одинаковой частотой встречаемости;
4. использовать только устойчивые слова и словосочетания;
5. исключать незначащие (в пределах данных текстов) слова, тщательно их проанализировав;
6. неоднозначные термины применять в том значении, которое они имеют в данном массиве.

Типы парадигматических и синтагматических отношений, используемых в ИПЯ, определяют его смысловыразительную способность, которая возрастает с увеличением числа и усложнением типов учитываемых отношений.

Основные принципы, которыми необходимо руководствоваться при выборе таких отношений:

1. затраты на разработку, ведение и использование словарей не должны превышать эффекта от их применения;
2. выбор типов отношений зависит от предполагаемых целей и областей использования ИПЯ и определяется необходимой полнотой и точностью поиска информации;
3. перед переходом к учету синтагматических отношений необходимо исчерпать возможности

Степень детализации словаря определяет полноту и точность поиска. Широко употребляемые термины дают большую полноту, но низкую точность поиска.
При выборе степени детализации словарей **необходимо учитывать заданные ограничения на желаемую полноту и точность поиска**, а также иметь иерархию словарей и использовать их различные уровни при поиске информации по разным запросам.

Одной из актуальных задач информационно-поисковых систем является поиск аналогов. Сложность этой проблемы заключается в том, что по поисковому образу запроса, выраженному в терминах одной области знаний или отрасли техники, необходимо найти документ-аналог, поисковый образ которого выражен в терминах

Один из путей преодоления такого барьера состоит в фасетном принципе организации словарей, т.е. в построении одноименных фасет в словарях всех областей знаний и метафасет или трансляторов для перевода терминов одной области знаний в термины другой области знаний в пределах заданного фасета.

Другой путь решения той же проблемы состоит в построении иерархического комплекса словарей, охватывающего все области знаний.

Количественные характеристики словарей

Эффективность информационного поиска в значительной мере определяется уровнем качества словарей информационно-поискового языка ДИПС. **Качество словарей можно характеризовать различными показателями**. Наиболее часто для этой цели используются следующие:

- количество типов словарей;
- число лексических единиц словарей;
- полнота словаря;
- коэффициент отображения лексики поискового массива;
- коэффициент динамики роста словаря;
- средняя длина лексической единицы словаря;
- среднее число символов в лексической единице словаря;

Полнота словаря

Рассмотрим ИПЯ конкретной АИПС, обслуживающей заданную предметную область. N_1 - общее число понятий данной предметной области, которые могут быть построены из лексических единиц ИПЯ ($N_{ИПЯ}$) по правилам их образования в данном ИПЯ. Тогда **коэффициент полноты словаря**

$$П_C = \frac{N_C}{N_{ИПЯ}}, 0 \le П_C \le 1$$

На практике используют:

$$П_C^1 = \frac{N_C}{N_O}, 0 \le П_C^1 \le 1$$

где N_O - количество лексических единиц словаря, по которым должен проводиться поиск (определяется по общему количеству несовпадающих лексических единиц массива запросов),

Коэффициент отображения лексики

Коэффициент отображения лексики поискового массива определяется как:

$$K_M = \frac{N_D}{N_C}$$

где N_D - количество дескрипторов в словаре.

Распределение лексических единиц по длине словосочетаний

Средняя длина словосочетаний, используемых в ИПЯ в качестве лексических единиц, характеризует степень прекоординации ИПЯ, и, тем самым, является важной характеристикой смысловыразительной способности ИПЯ.

Для характеристики ИПЯ с этой точки зрения используют **распределение длин словосочетаний**

$$\Gamma_g = \Gamma(g_1, g_2, \ldots, g_m), \quad g_l = \frac{K_l}{m}$$

K_l

где - количество лексических единиц, содержащих l слов;

m - максимальная длина словосочетания в ИПЯ (в числе слов).

Средняя длина ЛЕ

$$g_{cp} = \frac{\sum_{l=1}^{m} l \cdot g_l}{\sum_{l=1}^{m} g_l} = \frac{\sum_{l=1}^{m} l \cdot K_l}{\sum_{l=1}^{m} K_l}$$

Распределение ЛЕ по количеству символов

Длину лексических единиц ИПЯ можно характеризовать распределением

$$F_n = F(C_1, C_2, ..., C_i, ..., C_n), \quad C_i = \frac{B_i}{n}$$

где B_i - количество ЛЕ, содержащих i символов,

n - максимальное число символов в ЛЕ.

Среднее число символов в лексической единице

$$C_{cp} = \sum_{i=1}^{n} iB_i \Big/ \sum_{i=1}^{n} B_i$$

Динамика роста словаря

Динамика роста словаря характеризуется коэффициентом

$$K_d = \frac{S_d}{D}$$

где S_d - количество ЛЕ, введенных в словарь в процессе обработки ***D*** документов.

Ранговое распределение слов

Пусть $V = \{x\}$ - словарь ИПС. Обозначим F(x) - частоту встречаемости слова ***x*** во всех текстах массива.

Перенумеруем словарь так, чтобы частота слова, $V = \{x_1, x_2 \ldots, x_{m1}\}$, F(x) была невозрастающей функцией его номера, $F(x_1) \geq F(x_2) \geq \ldots \geq F(x_m)$.

т.е. если $\Phi(n) = F(x_n)$

то $\Phi(n) = \{F(x_1), \ldots, F(x_n)\} = \{f_1, \ldots, f_n\}$.

Назовем функцию f_n ранговым распределением слов

Показано, что частота слова n-го ранга $f_n = \frac{f_1}{n^\gamma}$ связана с частотой слова 1-го ранга следующей зависимостью:

где n - ранг слова,

- число, определяемое экспериментально.

Поисковый аппарат АИПС

Технология функционирования АИПС состоит **в переводе сообщений** (документов, текстов) и **информационных запросов на ИПЯ** (формировании поисковых образов документов и запросов), **сравнение ПОЗов и ПОДов и выдачи пользователям АИПС сообщений, отвечающих их информационным потребностям.**

При переводе сообщений на ИПЯ возможны **различные подходы**:

1. полный перевод сообщения на ИПЯ;
2. частичный перевод сообщения на ИПЯ (перевод па ИПЯ только отдельного элемента сообщения, например, его названия или реферата);
3. полный отказ от перевода на ИПЯ и использование в процессе поиска оригинального сообщения или его составляющих (текста,

Перевод запросов на ИПЯ тоже может быть выполнен **в различных вариантах**:

1. перевод всего информационного запроса на ИПЯ и формирование единого ПОЗа;
2. перевод отдельных составляющих на ИПЯ и формирование поисковых образов подзапросов.

Поисковое предписание (ПП), т. е. задание АИПС на поиск информации тоже **может быть сформулировано по-разному**:

3. формулировка единого ПП, соответствующего единому ПОЗу;
4. формулировка нескольких ПП, соответствующих подзапросам.

Процедура сравнения ПОЗов (или ПП) и ПОДов и принятия решений о выдаче или невыдаче пользователям АИПС тех или иных сообщений тоже характеризуются большим многообразием. Такое многообразие определятся многими факторами и, прежде всего, возможностями использования при формировании ПП логических операций И, ИЛИ, НЕ и различных критериев выдачи.

Организация и используемые методы и средства реализации процессов индексирования документов и запросов и проведения собственно поиска оказывают основополагающее влияние на эффективность поиска и, соответственно, эффективность АИПС.

Совокупность методов и средств реализации процесса поиска информации в автоматизированных ИПС назовем **аппаратом поиска или поисковым аппаратом.**

Поисковый аппарат АИПС включает:

1. **математический аппарат формализованного представления и поиска информации;**

2. **методы и средства структурирования информационных запросов;**

3. **критерии выдачи (смыслового соответствия) информации;**

4. **стратегии поиска и организации массивов.**