

70 коп.



АКАДЕМИЯ НАУК СССР

---

АНАЛИЗ  
МЕТАЯЗЫКА  
СЛОВАРЯ  
С ПОМОЩЬЮ  
ЭВМ

ИЗДАТЕЛЬСТВО «НАУКА»

АКАДЕМИЯ НАУК СССР  
НАУЧНЫЙ СОВЕТ ПО ЛЕКСИКОЛОГИИ И ЛЕКСИКОГРАФИИ  
ИНСТИТУТ ЯЗЫКОЗНАНИЯ

АНАЛИЗ  
МЕТАЯЗЫКА  
СЛОВАРЯ  
С ПОМОЩЬЮ  
ЭВМ



ИЗДАТЕЛЬСТВО «НАУКА»  
Москва 1982

В книге рассматриваются возможности автоматической лексикографии и обработке и системном представлении различных аспектов языковой структуры с особым вниманием к семантическим параметрам языка. Обобщается опыт машинного построения «Русского семантического словаря» (тезауруса), характеризуется его лингвистическая и программная технология, вскрываются смысловые, системные и статистические закономерности в его структуре, отражающие «анатомию» русской лексической семантики.

Коллектив авторов:

*Ю. Н. Караулов, В. И. Молчанов, В. А. Афанасьев, Н. В. Михалев*

Ответственный редактор

доктор философских наук  
*Г. В. Осипов*

## ВВЕДЕНИЕ

Какие цели преследует анализ метаязыка толкового словаря? Авторы думают, что ни в коем случае не прикладные, не утилитарные только, т. е. не продиктованные лишь стремлением улучшить лексикографические толкования. Мы надеемся с помощью такого анализа получить какие-то новые представления (пусть даже субъективные) об особенностях русской лексической семантики. Ну, а если наряду с этим некоторые предложенные нами рекомендации по оптимизации структуры словарных дефиниций покажутся их творцам-лексикографам не совсем бесполезными, авторы испытают чувство исправно выполненного дела.

Мы не считаем при этом, что лексикографические толкования в наших словарях, хотя бы и отчасти, плохи или несовершенны. Отнюдь нет. И тем не менее они могли бы быть лучше. Мы не думаем также, что наши толковые словари не осуществляют всей полноты своих функций. Нет, они являются не только справочником, энциклопедией знаний о языке, учебником жизни, но и оружием. По свидетельству В. П. Павличенко, участника советско-американских переговоров по ОСВ-2, отправляясь на очередной раунд переговоров, члены нашей делегации каждый раз не забывали включить в свой багаж все 17 томов Большого академического словаря русского языка. При обсуждении формулировок договора часто вставала задача найти такой термин, такое слово или словосочетание, которые, отвечая по форме пожеланиям противоположной стороны, в то же время позволяли бы сохранить в неприкосновенности нашу принципиальную позицию по тому или иному вопросу. Но можно себе представить, как велики были трудности при поисках такого слова: ведь вход в БАС только один — от орфографического облика слова, а это как раз и есть в данном случае неизвестное, искомое. Насколько облегчилась бы задача, если бы поиск был автоматизирован и по желанию пользователя, задав тот или иной параметр, на телеэкран можно было бы вызвать любую словарную информацию — по семантическому подобию, стилистическому различию, контекстам употребления, грамматическим характеристикам, ассоциативным возможностям слова и т. п.

Иными словами, в настоящем их виде словари, конечно, расширяют наши возможности в изучении лексической семантики, дают новый материал для диахронической и современной лексикологии, однако эти возможности были бы многократно усилены, будь представление этой информации формализовано, а доступ к ней автоматизирован. Предпосылки к формализации заложены уже в самих особенностях лексикографической фиксации данных о языке. И под метаязыком словаря в широком смысле понимается совокупность лексикографических параметров, отражающая все сведения, передаваемые словарями о структуре данного языка, его истории, распространении, функционировании и изучении. Этому предмету посвящается I глава работы. Метаязыком же в узком смысле этого терми-

на мы называем семантический метаязык, т. е. язык описания значений в толковом или переводном словаре, язык дефиниций. Анализ языка словарных толкований в семантическом, системном и статистическом аспектах проводится в последующих главах. Причем этот анализ осуществляется в связи и по результатам автоматического получения идеографической классификации русской лексики, демонстрируя тем самым возможности автоматической лексикографии.

Итоги проведенных в ходе анализа экспериментов — как машинных, так и психологических, привели к формированию у авторов двух принципиально важных теоретических представлений. Во-первых, стало ясно, что дальнейшее развитие формализованных методов анализа и их приложение к автоматизированному построению «тезаурусной картины мира» возможно только на основе системно-целевого подхода, принципы которого были прояснены и частично реализованы уже на прошедшем этапе работы. Во-вторых, как показал опыт автоматического построения словаря, когда число параметров достаточно велико, оперирование ими как отдельными дифференциальными элементами при решении сложных конструктивных задач становится затруднительным и малоэффективным. Выход из этой трудности мы видим в выделении некоторых постоянных комбинаций, пучков параметров, своеобразных «лексикографических модулей», которые в данной работе были найдены эмпирическим путем, но, по нашему убеждению, должны стать объектом теоретического осмысления лексикографов и лексикологов.

## Глава I

### ОСНОВНЫЕ НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ ЭВМ В ЛЕКСИКОГРАФИИ

Скажите государю, что у англичан ружья кирпичом не чистят: пусть чтобы и у нас не чистили, а то, храни бог войны, они стрелять не годятся.

*Н. С. Лесков. Левша*

#### § 1. ЛЕКСИКОГРАФИЧЕСКАЯ ПАРАМЕТРИЗАЦИЯ ЯЗЫКА КАК ТЕНДЕНЦИЯ В СОВРЕМЕННОМ СЛОВАРОСТРОЕНИИ И КАК ОБЪЕКТИВНАЯ ПРЕДПОСЫЛКА ЕГО АВТОМАТИЗАЦИИ

Взаимоотношения науки о языке с электронно-вычислительной техникой исторически складывались таким образом, что в автоматизации информационных процессов первая выступала как средство, как носитель и поставщик содержательных элементов, но параллельно — и как один из объектов анализа, к которому мы подходим с этим мощным инструментом исследования. Таким образом, если потенциально мы соотносим некоторую абстрактную ЭВМ с искусственным разумом, то рассмотрение лингвистических проблем с ее помощью предстанет как своеобразная авторефлексия этого разума, а значит, вопрос об использовании ЭВМ в лингвистике приобретает не только технологический, но и философский аспект. Неизбежность философской оснащенности становится еще более очевидной, когда мы обращаемся к принципам устройства самих систем искусственного интеллекта, включающих в качестве обязательного компонента определенный комплекс «знаний о мире», специфическую «иерархию ценностей», которая задается исследователем, создателем, исходя из масштабов и целей системы. Даже при решении системой чисто классификационных задач такая философская ориентированность становится необходимой<sup>1</sup>, не говоря уже о системах более высокого ранга, использующих так называемые «предугадывающие схемы» (advanced preguessing circuits)<sup>2</sup>. «Предчувствуя» входной сигнал, подобная схема ускоряет его обработку, как бы моделируя и во много раз усиливая человеческую способность к опережающему отражению действительности.

Но чтобы установить какую-то иерархию, нужно прежде всего обладать набором упорядочиваемых элементов, т. е. словарем в широком смысле термина<sup>3</sup>, без которого не обходится ни одна база данных, будь то

<sup>1</sup> Виленская С. К., Смирнский В. Б. Проблемы разработки комплекса ИПЯ по общественным наукам. — В кн.: Проблемы автоматизированной обработки научно-технической информации. М., 1980, с. 186—187.

<sup>2</sup> Britain leads in intelligent circuits. — New Scientist, 1981, vol. 90, N 1247, p. 25.

<sup>3</sup> Левина Е. Л. Система СЛОВАРЬ как часть матобеспечения лингвистического процессора. — В кн.: Новые задачи информатики. Новосибирск, 1979.

информационно-поисковая система, система машинного перевода или искусственного интеллекта. И отношения между лексикографией и электронно-вычислительной техникой повторяют отношения последней с лингвистикой вообще: проблема «словарь для машины» — это авто-рефлексия по поводу проблемы «машина для словаря».

В этом смысле рассуждения, которые часто приходится в последнее время слышать, о том, что автоматический словарь полностью вытеснит словарь традиционный, дисплей везде заменит книгу, — это, дескать, лишь вопрос времени и средств<sup>4</sup>, на наш взгляд, несостоятельны. Взаимодействие и взаимозависимость автоматической и традиционной лексикографии не в том, что первая вытесняет вторую, а в том, что на современном уровне знаний одна не может развиваться без другой. Многие рутинные этапы работы над словарем-книгой переданы теперь машинам: составление словников и словоуказателей к текстам, подготовка картотек, подбор иллюстративных примеров и т. п. С другой стороны, само развитие традиционной лексикографии в настоящее время создает фундаментальные предпосылки к ее автоматизации.

Рассматривая особенности современного словаростроения и общее движение мировой лексикографии, можно охарактеризовать основную линию ее развития как «тенденцию к лексикографической параметризации языка»<sup>5</sup>. Суть ее в том, что в мировой лексикографической практике наблюдается отчетливое стремление закреплять в словарной форме результаты изучения всех уровней языковой структуры, представлять в виде лексикона все языковые единицы содержательного и формального плана и отношения между ними, экстраполировать в словарь результаты анализа самых разнообразных языковых явлений. Более того, можно сказать, что в описательном языкознании пафосом изучения любого уровня языка становится построение словаря соответствующих элементов, процедур, отношений. И дело здесь не просто в том, что словарь оказывается самой удобной формой фиксации наших знаний об изучаемом объекте. Скрытой, возможно, не всегда осознаваемой, но мощной пружиной отмеченной тенденции является то, что она закладывает базу для максимально формализованных описаний естественного языка. А в наши дни все более широкое распространение получает представление, что формализация языка скорее всего достижима на пути создания все-сторонних (т. е. разноаспектных) и исчерпывающих инвентарей его единиц, процессов, явлений, в числе которых предусматриваются инвентари семантические, лексико-семантические, морфосемантические, семантико-статистические и др.<sup>6</sup>

<sup>4</sup> Ср., например, подробные расчеты по годам снижения стоимости в связи с совершенствованием технологии карманных словарей-калькуляторов, разрабатываемых в учебных целях в университете Карнеги: Fox M. S., Bebel D. A., Parker A. C. The automated dictionary. — Computer, 1980, vol. 13, N 7.

<sup>5</sup> Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981, с. 34—74.

<sup>6</sup> Городецкий Б. Ю. Теоретические основы прикладной семантики. Автореф. докт. дис. М., 1978, с. 10, 12, 13, 21; Он же. Семантические проблемы построения автоматизированных систем обработки текстовой информации. — В кн.: Вычислительная лингвистика. М., 1976, с. 28 и сл.; Виск Ю. А. Классификаторная модель эстонской морфологии (автоматический синтез глагольных словоформ). Автореф. канд. дис. Таллин, 1978.

Принцип лексикографической редукции, лексикографической интерпретации отдельных языковых явлений, преломляясь в реальные словари различных языков, позволяет обнаружить две разнонаправленные линии расширения и роста словарей. Центростремительная линия характеризуется целевой установкой на создание универсального словаря, охватывающего в идеале все сведения о данном языке, в том числе грамматические, синтаксические и экстралингвистические. На практике такой идеал оказывается недостижимым, но почти для каждого языка с достаточно богатой лексикографической традицией можно назвать словарь, в той или иной мере приближающийся к универсальному<sup>7</sup>. Центробежная линия, напротив, связана с построением однопараметровых словарей, определяется стремлением лексикографировать каждое отдельное явление данного языка, каждое измерение данной языковой структуры. Здесь в качестве примера можно сослаться на имеющиеся почти в каждом письменном языке орфографические словари (параметр побуквенного состава слова, или орфографии), орфоэпические словари (параметр произношения), словообразовательные, синонимов, словоизменительные и др.

Таким образом, под параметром в самом общем виде понимается способ лексикографической интерпретации того или иного структурного элемента или функционального явления языка и их экстралингвистических соответствий. В прагматическом смысле можно определить лексикографический параметр как некоторый квант информации о языке, который в экстремальном случае может представлять для пользователя и самостоятельный интерес, но, как правило, выступает в сочетании с другими квантами (параметрами) и находит специфическое выражение в словарях. Не вдаваясь в подробный разбор статуса лексикографических параметров, их структуры и объема, отметим лишь некоторые их особенности, существенные для нашего дальнейшего изложения.

Прежде всего, «экстраполяция в словарь» того или иного явления может осуществляться не в том виде, как оно традиционно фиксируется при описании данной языковой структуры, и в этом отношении параметр не обязательно соотносится однозначно с языковой единицей или категорией. Главное отличие параметра от названных понятий — прежде всего в его глобальности, поскольку параметр всегда относится к слову в целом: это не слог, а слогоделение; не аффикс или суффикс, а морфемное членение слова; не отдельное словообразовательное значение, а комплекс словообразовательных отношений, т. е. гнездо; не фонема и не звук, а произношение, и т. д. Что касается его величины, то параметр, как правило, больше единицы, состоит из нескольких единиц (ср. произносительный параметр и отдельный звук, словообразовательное гнездо и набор отношений в нем между составляющими его словами, дефиниция и отдельные слова, из которых она строится); в ряде случаев он может и совпадать с некоторой единицей (ср. ядерная структура — и единица и параметр, фразеологизм — как единица и как параметр, и то же самое — словосо-

<sup>7</sup> Для примера назовем здесь несколько таких словарей, расположив их по убыванию степени приближения к универсальному типу: Trésor de la langue française. Dictionnaire alphabétique de la langue des XIX-e et XX-e siècle (1789—1960). T. 1—5. Paris, 1971—1976; Wahrig G. Das grosse deutsche Wörterbuch. Gutersloh, 1966; Webster's seventh new collegiate dictionary. Springfield, Mass., 1963; Словарь современного русского литературного языка. Т. 1—17. М.—Л., 1948—1965.

четание) или быть меньше нее, оказавшись тождественным грамматической категории (ср. род, переходность). В последнем случае как вполне оправданный должен расцениваться синкретический способ задания параметров, когда, например, в словарях русского языка указанные рода есть одновременно свидетельство принадлежности данного слова к разряду имен существительных (имя существительное — самостоятельный параметр), а обозначение переходности соответственно дает информацию о том, что мы имеем дело с глаголом (глагол — отдельный параметр). Но синкретизм лексикографической интерпретации свойствен не только параметрам, по величине меньшим той или иной языковой единицы, а распространяется и на ряд других. Ударение, например, как отдельный параметр (ср. словарь ударений) всегда привязано к орфографическому или произносительному (т. е. транскрипции), параметр длины слова (количество букв) невозможен без орфографического, а паронимический — без указания на точку возможной контаминации значений двух слов, т. е. без семантического параметра дефиниции. Вместе с тем, значительно число случаев, когда параметр вообще несонзимерим с языковой единицей: среди единиц языка ему ничто не соответствует. Таков частотный параметр, особенно в специфических условиях, когда, например, статистической обработке подвергается семантическая сторона слова<sup>8</sup>; таков параметр семантического пробела в двуязычном словаре лексических лакун<sup>9</sup>; таков, наконец, параметр интерференции в специальном интерференционном словаре для обучения неродному языку<sup>10</sup>.

В содержательном аспекте, с точки зрения передаваемой ими информации, все лексикографические параметры можно разделить на две группы. В одну из них войдут собственно языковые, т. е. структурогенные — ударения, орфографический, произносительный, словоизменительный, словообразовательный, категориальные, сочетаемости и др. Они дискретны по своей природе, возможность их варьирования распространяется только на способы их задания, фиксации в словарях. О таких параметрах мы говорим, что они обладают «значением»: так, параметр «род» может принимать три значения в русском (м., ж., ср.), два значения во французском (*m.*, *f.*) и полностью отсутствует в английских существительных; параметр «словоизменение существительных» предполагает приписывание имени одного из падежных аффиксов, а задан он может быть в русском языке не только как набор 12 служебных морфем, но и с помощью ключевых форм, на основе которых порождается полная парадигма (для большинства имен в русском языке такими ключевыми формами являются, помимо данной на входе во всякий словарь русского языка формы именительного падежа единственного числа, также род. п. ед. ч. и им. п. мн. ч.).

Другую группу составляют параметры, содержание которых включает по необходимости и экстралингвистический фактор — денотативный, историко-культурный, прагматический, т. е. моменты, вторичные по отношению к самому языку, связанные не только с языком, но и с его

изучением — языкознанием. Эти параметры недискретны, характеризуются изменяющейся глубиной раскрытия и отражают не собственно структурные отношения, а процессы — диахронические, психолингвистические, синтагматические, взаимодействия языков, интерференции и т. п. В противоположность собственно языковым, или структурогенным, мы называем их лингвистическими и относим к их числу параметр родства (этимологический), страноведческий, стилистический, библиографический (сведения об исследованиях по поводу данного слова, отраженные, например, в упомянутом в сноске 7 словаре французского языка), хронологический (дата первой письменной фиксации), частотный, иллюстративный и другие.

Параметры могут находиться в отношениях взаимной эквивалентности. Так, параметр дефиниции, помимо обычного толкования значения слова, может быть адекватно представлен синонимами (а синонимия — это один из самостоятельных системных параметров: ср., например, синонимический словарь), иллюстрациями употребления данного слова (ср. словарь-конкорданс, где значение задано перечнем контекстов), «неявным определением» в виде набора элементов соответствующего семантического поля (ср. идеографический словарь), переводом и наконец, в словарях сочетаемости мы находим еще один параметр, эквивалентный рассматриваемому, — указание синтаксической валентности слова. Содержание, передаваемое каждым из этих параметров по отношению к одному и тому же слову, не тождественно во всех этих случаях, но в значительной степени, очевидно, является общим, почему мы и имеем основания говорить об их эквивалентности.

Наконец, чтобы закончить краткую характеристику лексикографических параметров, стоит сказать несколько слов о совместимости их друг с другом в рамках одного словаря. Два аргумента говорят как будто в пользу того, что на сочетаемость параметров в словаре не накладывается никаких ограничений. Во-первых, последовательное сравнение словарей по мере исторического движения лексикографии позволяет констатировать, что ее развитие шло в направлении наращивания числа параметров в словаре. Так, в известном словаре Палласа (3-я четверть XVIII в.) мы находим лишь три параметра — приблизительное произношение, семантическое соответствие (переводной эквивалент на немецком языке) и название того или иного языка или диалекта. В первом русском толковом словаре (конец XVIII в.) уже 15 параметров, поскольку здесь впервые широкое отражение получила грамматическая информация о слове; в словаре же Даля их только 12. Далее число параметров опять растет, и в словаре Шахматова (начало XX в.) их 16, а в Большом академическом (середина нашего века) уже около 30.

Во-вторых, рассмотрение большого числа современных словарей разных языков, предпринятое в работе, на которую мы ссылались (см. сноску 5), не дает оснований говорить о наличии каких-то объективных препятствий на пути построения любого сочетания параметров. Значит, параметры могут совмещаться друг с другом в любых сочетаниях, а в экстремальном случае принципиально возможным представляется и соединение абсолютно всех известных лексикографических параметров в одном словаре. Таким образом, можно сформулировать предположение, что в словаростроении, в лингвистическом конструировании словарей

<sup>8</sup> Лексика английского языка. Семантический частотный словарь по автоматической обработке экономической информации. М., 1975.

<sup>9</sup> Муравьев В. Л. Лексические лакуны. (На материале лексики французского и русского языков). Владимир, 1975.

<sup>10</sup> Чантуришвили Д. С. Основные проблемы сопоставительно-типологических исследований неродственных языков. — В кн.: Русский язык — язык межнационального общения народов СССР. М., 1976, с. 213—217.



действует такой же «закон свободы параметров», который установлен, например, в художественном конструировании<sup>11</sup>.

Тем не менее универсальный словарь пока не создан, а в разрабатываемых в последнее время проектах таких словарей набор параметров не покрывает даже половины их общего списка, приведенного в нашей работе (см. сноску 5). Хотя сам закон свободы параметров остается, как кажется, справедливым, на его действие наложены некоторые ограничения, которые и затрудняют решение вопроса об универсальном словаре. Первое из этих ограничений заключается в том, что, как показывает лексикографическая практика, возможно построить словари, содержащие исключительно структурогенные параметры, но невозможны словари из одних лингвистических: они обязательно должны дополняться и некоторыми структурогенными. Другое ограничение связано с первым и является как бы его уточнением: вход в словарь всегда базируется на одном из структурогенных параметров. Можно представить себе словарь со входом по лингвистическому параметру — скажем, частотный или хронологический (в последнем слова могли бы быть расположены в порядке приближения времени их фиксации в языке к настоящему моменту); можно вообразить этимологический словарь или словарь заимствований, в котором вход осуществлялся бы от языка-источника данного слова или группы слов. Но ни один из таких словарей не будет «работать», им нельзя эффективно пользоваться, если вход по лингвистическому параметру не будет дополнен входом по одному или нескольким структурогенным — орфографическим, т. е. алфавитным (прямым или обратным), смысловым (от понятия, как в тезаурусах), от длины слова, морфологическим (от корня, например) или каким-то иным. Более того, число входных параметров вообще оказывается довольно ограниченным, так как далеко не все даже из структурогенных могут реально использоваться на входе.

Наконец, третье ограничение касается сочетаемости: хотя сами параметры могут соединяться в любых комбинациях, не все способы их задания и степень глубины оказываются совместимыми друг с другом. Одна из объективных трудностей на пути построения универсального словаря лежит, таким образом, в отыскании оптимального соотношения между способами задания структурогенных и степенью глубины лингвистических параметров в таком словаре.

Будем надеяться, что эта беглая характеристика помогла читателю освоить понятия параметра языковой структуры и лексикографической параметризации языка. Вероятно, нет смысла подробно останавливаться на способах автоматической обработки и анализа каждого параметра. И не только потому, что разбор 68 пунктов (а именно такое число параметров включено в валовой, т. е. несистематизированный, их состав в уже неоднократно упоминавшейся выше работе) потребовал бы много места и отнял бы много времени у читателя. Дело в том, что сами приемы анализа в ряде случаев повторяются, а известная часть параметров и обрабатывается так же, как бывает задана, — синкретически. Поэтому поступим следующим образом: с определенной долей условности разобьем валовой состав параметров на три группы. В качестве основания разбиения примем три формальных критерия:

— степень препарированности параметров для машинной обработки и, соответственно, — степень сложности предмашинного анализа для выделения самого параметра и установления наилучшего способа его представления в машинных процессах;

— источник получения параметра — текст и его разновидности, словарь или специальные лингвистические исследования;

— цель автоматической обработки — построение словарей, получение новой информации о языковой структуре и синтезирование текстов.

В первую группу войдут параметры, требующие только широкой компиляции материала, домашняя обработка которых минимальна, а решение проблем их выделения и обобщения не превышает по своей сложности задач простой сортировки. К этому ряду следует отнести орфографический параметр, количественный (подсчет числа слов в словаре или вообще подсчет каких-то элементов словаря), хронологический, параметр длины слова, рифмы (скажем, словарь рифм поэта или вообще совпадающие финалы слов, а шире — обратный алфавитный порядок следования букв), аббревиатуры, иллюстративный параметр, стилистический и некот. др. Источником для их извлечения служат тексты, а целью их обработки является получение определенной базы данных. Назовем их компилятивными.

Вторую группу составят параметры, автоматическая обработка которых предполагает решение определенных аналитических задач, и объединяют они такие элементы информации о языковой структуре, которые воплощаются в категориях и формах. Сюда мы включаем ударение и связанные с ним произношение и слогоделение, часть речи, вид глагола, переходность, управление, залоги, причастие, будущее время, прошедшее время, морфологическое членение слова, многозначность — однозначность, транспозицию значения, ареальный параметр, фразеологический, стилистический и некот. др. Параметры этой группы выделяют на основе словарей (толковых, синонимических, идеологических, словообразовательных, грамматических), а также специальных языковедческих исследований, и результаты их автоматической обработки предназначены для анализа текста и построения новых типов словарей.

Границу между второй и третьей группами параметров провести труднее, но общий принцип здесь таков, что усложнение идет по линии комплексности параметра. Так, если в предыдущей группе каждый параметр воплощается в той или иной стандартной форме (причастие, вид глагола, морфологическое членение слова) или в модели, отражающей стандартный способ получения такой формы (управление глагола, будущее время, переходность), то параметры, относящиеся к третьей группе, предполагают развертывание уже в целый набор форм (или слов), представляют собой своего рода «парадигму», понимаемую в широком смысле этого термина, т. е. позволяют порождать любую форму или комплекс соответствующих форм. В этом смысле параметры третьей группы можно назвать «конструктивными», поскольку, будучи образованы по законам построения «новых лингвистических объектов»<sup>12</sup>, они способны не только обобщать известное, но и генерировать новую информацию о языке.

<sup>11</sup> Селезнев И. Ф. Дизайн. Проблемы материально-художественной культуры. Минск, 1978, с. 109—119.

<sup>12</sup> См.: Карацлов Ю. Н. Лингвистическое конструирование (О путях расширения практических приложений науки о языке). — Изв. АН СССР. Сер. лит. и яз., 1979, № 2.

(Ср. фиксацию в словоизменительном словаре таких форм, которые не засвидетельствованы в текстах, но образование которых не противоречит законам системы, и потому они как бы предсказаны соответствующей парадигмой: *блюдиш, волгнув, гния, ехав, жалев, кажась, могиш, сидев* <sup>13</sup>; ср. также выявление «пустых клеток», т. е. потенциальных слов в словообразовательной системе русского языка <sup>14</sup>.)

Одновременно с онтологическим разделением лексикографических параметров на три группы — компилятивные, аналитические и конструктивные, мы можем провести встречное, так сказать, деление в способах организации и целях функционирования разных групп параметров в автоматизированных лингвистических системах. Можно выделить три рода таких объектов, которые с известной долей условности отвечают соответствующим группам параметров. Во-первых, это банки лексикографических (и шире — вообще лингвистических) данных, которые ограничиваются собирательными и классификаторскими задачами и оперируют преимущественно параметрами первой группы — компилятивными. Более высокой степенью усложненности и, следовательно, использованием параметров более высокого ранга — аналитических, характеризуются действующие ныне системы машинного перевода, которые для своего функционирования требуют осуществления комплекса аналитических процедур — транскрипции, сегментации текста, морфологического и синтаксического анализа, лексического выбора при синтезе. И наконец, конструктивные параметры можно привязать к системам искусственного интеллекта, с их неизбежностью включения человеческого фактора, или «человеко-машинного принципа функционирования» <sup>15</sup>, следствием которого должен стать эвристический момент в решении той или иной задачи.

Условность предлагаемого деления очевидна, поскольку всякий банк данных, например, не существует сам по себе, а всегда имеет определенное предназначение, и ни одна система машинного перевода (МП) не может обойтись без некоторого банка информации. С другой стороны, можно полагать, что совершенствование МП направлено в отдаленной перспективе к оближению свойств соответствующих систем с принципами систем искусственного интеллекта, хотя реализованные в настоящее время МП пока далеки от таких параллелей. Тем не менее, при всей условности разделения трех групп лексикографических параметров по трем типам объектов, этот прием позволяет несколько упростить изложение и построить его таким образом, чтобы каждой из названных групп (компилятивным, аналитическим и конструктивным) и каждому типу объектов соответственно был посвящен отдельный параграф.

## § 2. АВТОМАТИЧЕСКАЯ КОМПИЛЯЦИЯ ПАРАМЕТРОВ И СОЗДАНИЕ БАНКОВ ЛЕКСИКОГРАФИЧЕСКИХ ДАННЫХ

Как было отмечено выше, домашняя обработка параметров этой группы минимальна и ограничивается некоторыми формальными процедурами, связанными с подготовкой материала к вводу в машину. Общее движение исследования в этом случае — от текста к словарю. Вероятно, одной из первых работ по словарю, подвергшихся автоматизации, было составление словника, т. е. перечня всех слов, встретившихся в тексте (текстах). Причем на заре автоматизации проблеме составляло даже получение простого алфавитного списка слов данного текста с подсчетом частоты их встречаемости, не говоря уже о составлении словоуказателей к текстам, включающим информацию об адресе данного слова — томе, названии произведения, странице, строке <sup>16</sup>. К настоящему времени автоматическое получение словников, которое является первым необходимым шагом в построении частотных, обратных словарей, конкордансов и в перманентном ведении, пополнении картотеки национального словаря, представляет собой простейшую задачу. Предварительная разметка текста при ее решении предполагает специальное обозначение конца строки, конца строфы (в поэзии), номера страницы, конца абзаца (в прозе) или монолога, реплики (в драме), выделение в текстах слов типа глава, действие, эпизод, сцена, предисловие а также названия произведения <sup>17</sup>.

Подобный словник, или словоуказатель, с одной стороны, хотя и содержит всего лишь перечень слов, с точки зрения параметризации не является элементарным параметром, а объединяет целый комплекс их: побуквенный состав слов (т. е. орфографический параметр), частотный, хронологический (поскольку обработке могут быть подвергнуты тексты, охватывающие только определенный исторический период), а также имплицитно присутствующие дистрибутивный и длины слова (количества букв в слове). Каждый из перечисленных параметров в зависимости от поставленной задачи может стать доминирующим, и тогда мы имеем дело либо с орфографическим <sup>18</sup>, либо с обратным <sup>19</sup> и т. п. словарем.

С другой стороны, словоуказатель — это вспомогательный для лексикографа материал, а собственно словарь начинается с картотеки. В картотеке слову соответствует от одной до нескольких десятков карточек, на каждой из которых зафиксирован тот или иной контекст этого слова. Частным случаем такой картотеки является конкорданс к произведению или совокупности произведений одного автора. Общие филологические словари составляются; естественно, не на материале одного автора, и лексикографы прибегают к выборкам. Выборки, сделанные с помощью конкордансов, способны более адекватно отражать и обобщать особенности словоупотребления той или иной эпохи, чем выборки, осуществляемые на основе опыта и интуиции лексикографа. Как считает Д. Кубурлис,

<sup>16</sup> Колин А. Ж. Т. Автоматическое составление словника. — В кн.: Автоматизация в лингвистике. М. — Л., 1966.

<sup>17</sup> Вертель В. А., Вертель Е. В., Рогожникова Р. П. К вопросу об автоматизации лексикографических работ. — ВЯ, 1978, № 2, с. 109.

<sup>18</sup> Viks U. Oigekeelsussõnaraamatü arvutivariant. Verbid. (Вариант орфографического словаря, обработанный на ЭВМ. Глаголы). — Keel ja kirjandus, 1974, № 10.

<sup>19</sup> Штиндлова И. Обратные словари. — В кн.: Автоматизация в лингвистике. М. — Л., 1966.

<sup>13</sup> Зализняк А. А. Грамматический словарь русского языка. М., 1977, с. 98—130.

<sup>14</sup> Тихонов А. Н. Школьный словообразовательный словарь русского языка. М., 1978.

<sup>15</sup> Пиотровский Р. Г. Лингвистические аспекты «искусственного разума». — ВЯ, 1981, № 3, с. 35—36.



составивший на ЭВМ несколько конкордансов, «если имеется конкорданс, исследователю уже не нужно ходить пешком, он способен к полету»<sup>20</sup>.

Автоматическое составление конкордансов сейчас более распространено применительно к поэтическим текстам, что связано с относительной простотой определения в них минимального контекста: он совпадает здесь со стихотворной строкой. Так, обработка текста при составлении словаря, упомянутого в сноске 20, включала следующие шаги: транслитерация с кириллического алфавита на латинский (что обусловлено особенностями машины, на которой проводилась работа по этому словарю) вместе со страницей, названием стихотворения (условно все стихи названы здесь по первой строке) и номером строки; шифровка на перфокартах с последующей автоматической и визуальной корректировкой перфорации, автоматическое извлечение титулов (названий стихотворений) и их сокращение. Затем номера, приданные каждому стихотворению, были автоматически заменены на сокращения соответствующих титулов, слова одной группы (т. е. словоформы одной лексемы) сведены вместе, проведена дифференциация омоформ и осуществлен обратный перевод с латинского на кириллический алфавит.

Полученный результат автор назвал «кластерным конкордансом» (cluster concordance), поскольку слова в нем собраны как бы «гроздьями» под «главным словом» (headword), т. е. словоформы — под своей лексемой, или слова — под гиперлексемой. Понятно, что такой «кластерный конкорданс» уже с некоторой долей условности может быть назван словарем, составленным из компилятивных параметров, поскольку по крайней мере две процедуры здесь выходят за рамки чистой компиляции и требуют проведения определенного анализа. Это, во-первых, разрешение трудности, которую Д. Кубурлис в другой своей работе назвал «диффузией словоформ» (word-form diffusion)<sup>21</sup> и суть которой заключается в том, что формы одного слова в обычном конкордансе, будучи упорядоченными по алфавиту, могут оказаться в разных местах словаря (ср. англ. see — saw, is — are — were, рус. идти — шел, звать — зову и т. п.). Во-вторых, аналитической работы требует «нечувствительность» машины к омоформам (homofonn insensitivity), из-за которой в одну единицу соединяются разные слова на основе их буквенного совпадения.

Компилятивный параметр представлен и в словаре рифм. Дж. Т. Шоу — составитель такого словаря на материале пушкинской поэзии — следующим образом характеризует этапы работы над словарем<sup>22</sup>. Были сделаны ксерокопии всех произведений поэта по академическому изданию. Каждая строка в этом тексте была помечена ударением для позиции конечного иктуса и добавлен код места строки, который позволяет легко связывать любое рифмующееся слово с его партнером. Затем каждая строка была перфорирована на отдельной карте. Одновременно всем конечным словам в строках была приписана информация, отражающая грамматический статус этого слова, его синтаксические особенности, форму рифмующегося сегмента, т. е. мужская ли это рифма

с ее разновидностями, или женская, или дактилическая и т. п. Эта информация была перфорирована на второй для каждой строки карте. Обе карты записывались на магнитную пленку для получения единой комбинированной информации о строке. Для обработки всего массива были составлены многочисленные программы, и результат был размножен в машинной печати.

Словарь пушкинских рифм состоит из трех частей.

I — лексикон всех конечных слов в строках (как с рифмой, так и без нее), например:

игр'ал...1 ем	2
игр'ал...2 ем	1
игр'ал...3 ем	3

Цифры в последней колонке означают здесь количество строк с данной рифмой (т. е. ее частоту), другие обозначения содержат грамматическую информацию — первое (или второе, третье) лицо, единственное число, мужской род.

II — конкорданс рифм — в обратном алфавитном порядке, сгруппированных под рифмующимися сегментами, например:

/ЕЧ/		
ПЛ'ЕЧ		
	р'ечь	ежИ-1
	до пл'еч	мсР12
2*С2.71.2.4.а*		
	м'еч	емИ-1
	с пл'еч	мсР41
2РЛ.2.141.42.6*		

Это означает, что под указанным сегментом слово *плеч* встретилось два раза; справа от слова дана грамматическая информация о нем, которая для последней, например, формы — *с плеч* — расшифровывается так: множественное число, средний род, родительный падеж, а цифры (41) означают тип предложного управления. Буквенно-цифровая запись под каждой рифмующейся парой сообщает ее адрес: в последнем случае это «Руслан и Людмила», глава 2, страница 141, строка 42; буква со звездочкой символизирует положение рифмы — в одном и том же предложении или в разных выступают партнеры, один и тот же ли субъект они характеризуют, к прямому или косвенному объекту относятся и т. п.

III — индекс стихотворений, с помощью которого можно расшифровать сокращенную запись в адресе.

Словари, составленные на материале поэзии Баратынского и Батюшкова<sup>23</sup>, представляют дальнейшее усложнение, увеличение числа параметров и объединяют конкорданс со словарем рифм. Принципы их построения не добавляют ничего нового к сказанному выше о каждом из типов словарей в отдельности. Здесь надо только подчеркнуть, что хотя рифма согласна нашим представлениям — это один отдельный параметр, в словаре

<sup>20</sup> A concordance to the poems of Osip Mandelstam. Ed. by D. J. Koubourlis. Ithaca — London, Cornell University press, 1974, p. XV.

<sup>21</sup> Koubourlis D. J. From a word-form concordance to a dictionary-form concordance. — In: Computers in the humanities. Minneapolis, 1975, p. 225.

<sup>22</sup> Pushkin's rhymes. A dictionary. By J. Th. Shaw. Madison, 1975, p. IX.

<sup>23</sup> Baratynskii. A dictionary of the rhymes & a concordance to the poetry. By J. Th. Shaw. The University of Wisconsin Press. Ann Arbor, 1975; Batjushkov. A Dictionary of the rhymes & a concordance to the poetry. By J. Th. Shaw. The University of Wisconsin Press. Madison, 1975.

рифм, как мы видели из приведенных примеров, выступают и многие добавочные параметры, в частности фонетическая, грамматическая и синтаксическая информация о конечном слове строки. Точно так же и в конкордансе — к элементарному иллюстративному параметру присоединяются орфографический, частотный, дистрибутивный. Более того, практика автоматического построения конкордансов приводит лингвистов к выводу, что просто выигрывать в скорости, по сравнению с ручным их составлением, уже не может считаться преимуществом: современные автоматические конкордансы должны включать максимально возможное число структурных параметров, оснащая исследователя исчерпывающей, быстро обобщаемой и легко доступной информацией о языке <sup>24</sup>.

Что касается сфер применимости конкордансов, то та нитка разочарования, которая прозвучала по их поводу в упомянутой статье Инграма, кажется нам преждевременной. Об этом свидетельствует, во-первых, сам размах ведущихся в данном направлении работ: конкордансы упорно составляются во всем мире — для прозы и поэзии, для отдельных произведений, отдельных авторов и целых периодов развития национальной словесности, вручную (как в Институте языка им. Р. Ачарьяна АН АрмССР — к произведениям всех классиков армянской литературы) и с помощью ЭВМ. Во-вторых, в ряде теоретических статей, скрыто или явно полемизирующих с точкой зрения Инграма, показано, что диапазон использования конкордансов на самом деле простирается от задач выборки и составления словоуказателей до превращения конкорданса в базу и инструмент подготовки толковых словарей <sup>25</sup>. В-третьих, наш собственный опыт свидетельствует о том, что при достаточно большом объеме исходного текста принципиально возможно предложить ряд формальных процедур, позволяющих алгоритмически строить лексикографическое определение значения слова из суммы контекстов его употребления, т. е. автоматически осуществлять переход «от конкорданса к дефиниции». Это становится особенно важным, если составители соответствующего толкового словаря хотят остаться максимально близкими к авторскому пониманию смысла каждой употребленной им лексемы, как, например, в случае работы над «Словарем языка произведений В. И. Ленина», где конкорданс мог бы стать основой корректного и точного построения лексикографических толкований из самих ленинских текстов. Наконец, говоря о возможностях, предоставляемых конкордансами, нельзя не отметить их роли в обучении иностранному языку. Именно такие задачи — препарирование морфологии русского языка в учебных целях — решает обратный конкорданс к «Евгению Онегину» <sup>26</sup>, например (форму представления текста в нем см. на с. 17—18).

Здесь цифры на левой стороне — порядковые номера словоупотреблений (в порядке обратного алфавита). Они упрощают учет особенно многочисленных вхождений: их количество можно получить путем вычита-

ния порядковых номеров. Цифры на правой стороне страницы — это точное место данного словоупотребления: первый столбец, всюду повторяющееся «01» — код «Онегина», второй столбец — номер главы, третий — номер строфы, четвертый — номер строки, пятый — номер слова в типографской строке.

Таким образом, в отличие от обычных для поэтических текстов KWOC-конкордансов (key word-out of-context), данный выполнен в форме KWIC-конкорданса (key word-in-context), причем роль key word выполняет здесь соответствующая буква в центре строки машинной распечатки. Помимо учебных целей такой конкорданс может заинтересовать, как пишет автор в предисловии, и исследователей русского языка, помогая решать такие, например, вопросы: какова функциональная нагрузка тех или иных концовок словосформ в текстах; каково окружение склоняемых и спрягаемых форм (падежи с предлогами и без предлогов, глагольные формы с личными местоимениями и без них) и т. п.

Заканчивая это «отступление» о конкордансах, следует напомнить, что они, представляя собой результат обработки главным образом компилятивных параметров, могут входить составной частью в базу (банк) данных, если эта база строится из текстов <sup>27</sup>, или сами выполнять функцию такой базы данных для последующих аналитических процедур над включенными в них параметрами, как это имеет место в некоторых вариантах эмпирического машинного перевода (см. § 3).

Создание узкооднопараметровых словарей преследует, как правило, какую-то вполне определенную, часто вспомогательную цель, как, например, в случае хронологического словаря английского языка, вход в который осуществляется именно от хронологического параметра — года первой письменной фиксации данного слова <sup>28</sup>.

Этот словарь был необходим для установления объема и состава английского лексикона времен Шекспира, чтобы изучение языка этого автора можно было осуществлять на фоне языка соответствующей эпохи. С этой целью Shorter Oxford English Dictionary (SOED), содержащий 80 096 словарных статей, был целиком введен в машину и проанализирован по ряду критериев. В частности, были исследованы омографы, число значений каждого слова, этимологии, упорядочены данные о письменной регистрации каждого слова, выверены характеристики по частям речи и т. п., что позволило выявить значительное число неточностей и ошибок в словарных статьях. Затем входные слова были переранжированы по времени их появления в письменном языке и получен Chronological English Dictionary (CED). Объем лексикона, ограниченный 1623 годом, — датой появления первого собрания сочинений Шекспира — и определил тот фон, на котором исследуется язык классика английской литературы <sup>29</sup>.

<sup>27</sup> Ср., например, конкорданс по 250 рукописям и янкунабулам в базе данных по старониспанскому языку (OSA — Old Spanish Archive), предназначенной для автоматического составления исторического словаря: Nitti J. J. Computers and the old Spanish dictionary. — Computers and the Humanities, 1978, vol. 12, N 1—2.

<sup>28</sup> Finkenstaedt Th., Leisi E., Wolff D. A chronological English dictionary listing 80,000 words in order of their earliest known occurrence. Heidelberg, 1970.

<sup>29</sup> Spevack M., Neuhaus H. J., Finkenstaedt Th. SHAD: a Shakespeare dictionary. — In: Computers in the humanities. Edinburg, 1974, p. 114; Neuhaus H. J., Spevack M. A Shakespeare dictionary (SHAD): some preliminaries for a semantic description. — Computers and the humanities, 1975, vol. 9, N 6.

<sup>24</sup> Ingram W. Concordances in the Seventies. — Computers and the Humanities, 1974, vol. 8, N 5—6.

<sup>25</sup> Герд А. С. и др. Автоматизация в лексикографии и словари-конкордансы. — ИДВШ. Филол. науки, 1981, № 1; Preston M. J., Coleman S. S. Some considerations concerning encoding and concordance texts. — Computers and the Humanities, 1978, vol. 12, N 1—2.

<sup>26</sup> Обратный конкорданс к «Онегину». Сост. М. Лацк. Ред. Ф. Папп. Дебрецен, 1980.

Форма представления текста в обратном конкордансе к «Евгению Онегину»

2164	сказалъ, он зналъ довольно по-латыни, чтоб
2165	глядѣл, один среди своихъ владѣний, чтоб
2166	m*onsieur l'a*bbé, французъ убогой, чтоб
2167	да, можетъ быть, боязни тайной, чтоб
.....	
2186	и тихо целить в блѣдный лолб
2187	д*ержавин насъ заметилъ и, в гроб

Результаты анализа представлены в полном и комплексном словаре шекспировских текстов (SHAD), который характеризует каждое слово по 32 (по нашему подсчету) параметрам, объединяющим информацию шекспировского конкорданса Дж. Бартлетта (1894 г.), автоматически составленного современного конкорданса (SHAC), а также упомянутых выше словарей — SOED и CED. Естественно, что далеко не все из этих параметров являются чисто компилятивными, и прежде всего это касается информации, связанной с лемматизацией, т. е. сведением текстовых форм к исходным, словарным формам. Сюда относятся также установление грамматической и синтаксической двусмысленности во фразе, определение принадлежности слова к той или иной специальной сфере, установление общей и более узкой этимологии и т. п.

Комплекс параметров по SHAD, охватывающий информацию и о структуре английского языка, и о его функционировании в Англии на рубеже XVI—XVII вв. в сопоставлении с современным английским (здесь, в частности, используется параметр, который я называю «словарным» и который указывает другие словари, в данном случае английского языка, содержащие сведения о соответствующем слове и различия — если они имеются — в его характеристике), и о специфике его употребления в шекспировских текстах, приближается по своей сложности, объему и задачам к так называемому банку данных по подязыку определенной области, в данном случае — по языку писателя.

Банки данных, как, впрочем, ясно из самого названия, призваны систематизировать и аккумулировать информацию об особой сфере знаний, техники, человеческого опыта, с тем, чтобы по запросу исследователя компоновать эту информацию заданным образом и выдавать в необходимых объемах для решения различных задач. Банки данных, по самой структуре своей предполагающие широкую компиляцию материала, в нашем случае — собрание языкового материала для его лексикографического представления, строятся в основном из компилятивных параметров и являются для этой группы, очевидно, «вершинным достижением», представляя собой максимальную возможность структурирования на основе простой сортировки.

Рассмотрим в качестве примера устройство машинного архива финских диалектов, созданного в университете г. Турку<sup>30</sup>. Архив подготавливался

<sup>30</sup> Описание дано на основе личного знакомства автора с указанной системой и бесед с ее создателями. Из публикаций о ней см., напр.: *Ikola O., Karjalainen Y.* Syntax archives of Finnish dialects for computer work. — In: *Computational and mathematical linguistics*, I. Frenze, 1977.

эпиграфы разбирать, потолковать об	01 1 06 04 1
только время проводить, сперва задумал	01 2 04 02 1
не измучилось дитя, учил его	01 1 03 10 1
муж или свет не угадал	01 8 35 02 1
.....	
на благородном расстоянии; но отослать	01 6 33 11 6
сходя, благословил, и я, в	01 8 02 04 3

в течение почти двух десятилетий. Записи диалектного материала, собранного финскими лингвистами и энтузиастами изучения родного языка с конца прошлого века, транскрибировались по единой системе (при необходимости переводились с магнитной пленки в обычный текст на бумагу), и текстовые формы размечались по определенной программе, предусматривающей как структурогенные — в нашей терминологии, т. е. морфологические, синтаксические, фонетические и прочие параметры, так и лингвистические — название диалекта и ареал его распространения, имя, возраст и другие данные об информанте, время записи и проч.

Отличительной особенностью этого архива (можно сказать даже «словаря», имея в виду, что последний тоже есть систематизированный комплекс параметров, вход в который осуществляется от одного или нескольких из них) является исключительно широкая представленность в нем синтаксических параметров — их 33 из общего числа около 100. Это и понятно: хотя так же, как в обычном словаре, перед вводом в машину каждое слово здесь было занесено на отдельную карточку, и ему была приписана вся связанная с ним морфологическая, синтаксическая и другая информация, а сами слова упорядочены как в прямом, так и в обратном алфавитном порядке, но не надо забывать, что данный архив имеет дело с текстом — текстом финских диалектных записей, отсюда и обилие, и особое значение в нем синтаксических параметров. Среди последних тоже можно выделить как структурогенные (напр., члены предложения; главное и подчиненное предложение; порядок следования главного и подчиненных; повествовательное, вопросительное или восклицательное предложение; прямая или косвенная речь), так и лингвистические (напр., число предложений в тексте, порядковый номер предложения в данном тексте, число слов в предложении).

При этом почти все названные параметры, за исключением, может быть, первого — «члены предложения», — получают определенные значения только в конкретном тексте, т. е. являются, иными словами, «текстовыми», и потому не могут быть включены в общий перечень лексикографических.

Аналогичный банк данных по другому подязыку (не диалектному) создан в Университете дружбы народов им. П. Лумумбы и межфакультетской лаборатории вычислительной лингвистики МГУ для обработки записей русской разговорной речи. Система, названная АЛЕКС, представляет собой лексикографическую информационно-поисковую систему, основной

структурной единицей которой является Лексикографический Банк Данных.

«Лексикографическим Банком Данных (ЛБД) называется система, состоящая из:

- 1) массива Модулей Лексикографических Данных (МЛД);
- 2) справочника Обращения к МЛД;
- 3) программы Управляющего Блока, осуществляющего функции накопления, коррекции, хранения, поиска и выдачи модулей по запросам.

Модулями лексикографических данных могут являться тексты (минимальные выборки), словарные статьи, иллюстрирующие контексты и другие лексикографические единства. Модули состоят из Записей Единиц Данных. Единицами данных являются словоформы, словарные формы, ключевые слова, указывающие место данной словоформы в структуре его словарной статьи (в парадигме) или в структуре предложения, синтаксические паспорта словоформ, указывающие их синтаксические функции, частоты употребления словоформ и словарных слов»<sup>31</sup>.

С позиций лексикографической параметризации языка, с которых мы ведем изложение, рассматриваемая лексикографическая ИПС включает главным образом компилятивные параметры, которые распределяются между четырьмя типами вспомогательных банков данных: аккумулятором, или банком входных данных, содержащим тексты стандартной длины с информацией об источнике, типе записи, типе речи и т. п.; словарем, или банком выходных данных, содержащим лексему с относящейся к ней информацией — часть речи, парадигма, указатель синтаксических функций, набор частот и др.; конкордансом, или банком резервных данных (предложений, контекстов), содержащим адреса хранения управляющих и управляемых слов; индексатором — специальным блоком для реализации диалогового режима.

Банк лексикографических данных, помимо того что он может использоваться для получения самой разнообразной информации о языке (как современном, так и любого исторического периода), представляет собой необходимую базу автоматизации работ в словаростроении. Так, корпус текстов на 5 млн. словоупотреблений, репрезентирующий функционирование английского языка в Америке наших дней<sup>32</sup>, послужил базой для автоматического построения частотного, ряда школьных словарей<sup>33</sup>, для уточнения при переиздании некоторых других известных словарей компании Мерриам — Уэбстер.

Аналогичное применение находят банки данных английского<sup>34</sup>, швед-

ского (The Swedish Logothèque: A Computer-Based Text and Word Bank)<sup>35</sup>, итальянского<sup>36</sup>, древнегреческого<sup>37</sup> и других языков.

Вне пределов собственно лингвистики особое значение приобретают банки данных по терминологии той или иной отрасли знаний, которые получили в последнее время широкое распространение во всем мире. Во время командировки в Канаду в 1979 г. нам удалось познакомиться с принципами организации и функционирования двух терминологических банков — в Оттаве и Квебеке. Оттавский терминологический банк является предприятием правительственного подчинения (он относится к ведению государственного секретариата — *Secretariat d'Etat*) и вместе с Бюро переводов (*Bureau de traductions*) и Генеральной дирекцией по терминологии и документации (*Direction générale de la terminologie et de la documentation*) составляет лингвистическую службу при государственном секретаре, столь необходимую в двуязычной стране, какой является Канада. Банк является универсальным, т. е. включает терминологию всех отраслей науки, техники, социальной жизни, культуры, спорта, предпринимательства и т. д., с той лишь разницей между этими областями, что термины из сферы чистой науки (теоретической физики, химии, биологии, например) включаются в банк лишь в той мере, в какой они встречаются в официальных документах. Таким образом, в его составе в основном общенаучные и общетехнические термины. Другая особенность этого банка — его двуязычность, дублирование информации в нем на двух языках — английском и французском. Банк содержит около полутора миллионов единиц хранения — слов и словосочетаний, из которых только около половины имеют статус стандартизованных терминов. На каждую единицу хранения заполняется сортировочная аналитическая карта (*fiche de dépouillement*), по структуре представляющая собой приспособленную для машинной обработки матрицу-досье на термин. Она состоит из четырех групп параметров, в основном компилятивного характера, причем собственно лексикографические составляют две первые группы.

I группа параметров (№ 1—19) дублируется на двух языках — английском и французском и включает название языка, орфографический облик термина, число слов в терминсочетании, указание отрасли науки, техники или отнесенность к социальной сфере, частотность, хронологические сведения, вариативность, аббревиатуру, синонимы, антонимы, рекомендательную оценку (правильно/неправильно), дефиницию, указание на источник, тип документации, указание на принадлежность к стандартизованным терминам, близкие ключевые термины.

II группа (№ 20—26) охватывает тезаурусную характеристику термина с указанием его вхождения в определенные фасеты, области и дескрипторы.

III и IV группы параметров (№ 27—32) относятся к внешним характеристикам системы и содержат полную информацию об источниках с точным адресом термина, шифры разработчиков-исполнителей, даты

<sup>31</sup> Андрищенко В. М. Некоторые проблемы автоматизации лексико-статистических работ. — В кн.: Вопросы лингвостатистического анализа русской разговорной речи. М., 1976, с. 69; см. также в том же сборнике статью: Ванин Ю. В. Лингвостатистический справочник русской разговорной речи и получение информации высших уровней.

<sup>32</sup> Paikeday T. M. The American heritage intermediate corpus. — In: Computational and mathematical linguistics, I. Firenze, 1977.

<sup>33</sup> См. рецензии на эти словари: Smith R. N. The American heritage school dictionary. N. Y., 1972; The American heritage word frequency book. N. Y., 1971. — In: Computers and the Humanities, 1974, vol. 8, N 5—6, p. 335—336.

<sup>34</sup> Leonard R. The computer archive of modern English texts. — In: Computational and mathematical linguistics, II. Firenze, 1977.

<sup>35</sup> Zettersten A. Current Scandinavian computer-assisted language and literature research. — Computers and the Humanities, 1976, vol. 10, N 5, p. 275.

<sup>36</sup> Spogli elettronici dell'italiano delle origini e del duecento. Utrecht—Bologna, 1974; Spogli elettronici dell'italiano letterario contemporaneo. Il Mulino, 1975.

<sup>37</sup> Bruner Th. F. El proyecto «Thesaurus linguae graecae». — In: Utilización de ordenadores en problemas de lingüística. Madrid, 1976 (Revista de la Universidad Complutense, vol. XXV, N 102).

первого включения термина в банк и последующих редакций и уточнений. а также указывают их авторов.

Вход в систему осуществляется от любого параметра, и пользователь может получить информацию, ориентированную в соответствии с интересующей его в данный момент проблемой. Решая в первую очередь утилитарные задачи — служить консультационной базой для принимаемых правительственных и административных решений в ситуации двуязычной страны, быть средством для создания аутентичных текстов государственных документов на обоих языках, банк оказывается одновременно мощным средством упорядочения и стандартизации научно-технической терминологии и может использоваться также в различных вспомогательных — прикладных и исследовательских — целях.

Опыт построения терминологических банков в разных странах и высокая оценка их эффективности приводят к мысли, что и в нашей стране задачам упорядочения, стандартизации терминологии и ее унификации на языках народов СССР наиболее полно отвечал бы всесоюзный автоматизированный многоязычный терминологический банк, сводящий в единую систему, с одной стороны, национальные двуязычные филиалы в каждой союзной республике, а с другой стороны, также и отраслевые терминологические банки, что обеспечило бы широту охвата и глубину представленности в нем специальной терминологии. Вопрос о создании такого банка в СССР назрел.

### § 3. МАШИННЫЙ ПЕРЕВОД И АНАЛИТИЧЕСКИЕ СЛОВАРНЫЕ ПАРАМЕТРЫ

Отход от простой констатации лингвистических данных, ее преодоление означает одновременно и углубление наших знаний о языке в связи с обращением к более сложным характеристикам его внутренней и внешней структуры. Для выделения параметров, названных нами аналитическими, их представления в удобной для машинных операций форме и автоматической их обработки уже недостаточно простого сравнения, на котором основывается опознание, идентификация компилятивных параметров, а требуются более сложные действия по разложению языковых единиц на составные части, обобщению элементов по новым признакам и выбору альтернативных решений. Направление исследования здесь противоположно тому, что мы фиксировали для компилятивных параметров, потому что, как правило, источником для аналитической обработки соответствующих лексикографических параметров служат толковые словари: обычный словарь перестраивается в автоматический, содержащий аналитические параметры. Последний позволяет решать и такие задачи, в которых, на первый взгляд, как бы генерируется новая информация о языке. Однако при ближайшем рассмотрении эти задачи нельзя считать эвристическими, и системы с аналитическими параметрами — так же, как и системы, базирующиеся на компилятивных параметрах, — обобщают и выдают («порождают») только то, что было заложено в них с самого начала. Впечатление оригинальности получаемой информации возникает в этих случаях за счет некоторой переклассификации материала, что само по себе, безусловно, важно и необходимо. Вместе с тем автоматический, а тем более машинный, словарь всегда проигрывает своему источнику

толковому словарю: при большей аналитичности и точности задания отдельных параметров число их всегда меньше, чем в толковом словаре, что обуславливается более строгой адресностью, более узкой целевой предназначенностью автоматического словаря. Отсюда можно сделать вывод, что автоматический анализ метаязыка словаря практически никогда не осуществляется в полном объеме, если понимать метаязык в широком смысле. Рассмотрим несколько систем такого рода.

В реализованной во Всесоюзном центре переводов системе англо-русского МП работает комплекс из семи машинных словарей<sup>38</sup>, которые параметризуют зоны языка-источника и языка-цели как по компилятивным, так и по усложненным аналитическим характеристикам. Английский словарь первичной обработки предназначен здесь для идентификации словоформ входного текста с основами, содержащимися в алгоритме морфологического анализа. Слова в словаре для первичной обработки расположены по убывающим длинам, а внутри групп одной длины — по алфавиту. Таким образом, важным компилятивным параметром является длина слова. Словарь оборотов с переводами служит для идентификации устойчивых словосочетаний в тексте. Затем в работу вступает словарь омонимов и алгоритм разрешения омографии. Однозначные и многозначные английские слова сгруппированы в два разных словаря: перевод однозначных осуществляется по таблице соответствий, при переводе многозначных слов используются специальные алгоритмы поиска контекстных признаков, поэтому последний словарь, объединяющий аналитические параметры, называют еще контекстологическим. Словари оборотов и контекстологический являются специфическими разновидностями синтаксического словаря. В тех случаях, когда алгоритм перевода многозначного слова не дает результата, в работу вступает шестой — контрольный словарь однозначного перевода многозначных слов. И наконец, словарь русского языка для синтеза, который уже никак не может обойтись одними компилятивными параметрами, воссоздает русскую словоформу. Аналитической работы и своеобразной категоризации языкового материала в этой системе требуют и особый способ трактовки и разрешения полисемии, и специфическое представление результатов морфологического анализа и классификации слов в выходном русском словаре, и т. п.

Если сопоставить с этой системой реализованную в Монреальском университете автоматизированную линию англо-французского перевода для субязыка метеорологии, то на первый взгляд принципиальных различий в цепочке перехода от текста на входном языке к тексту на выходном языке мы как будто не заметим. И в той, и в другой системе имеются словари, параметризующие данные соответствующего языка, правда, в ТАУМ-METEO<sup>39</sup> их не семь, а всего три; и в той и в другой системе используются похожие приемы морфологического и синтаксического анализа, строятся таблицы соответствий и т. п. Однако при ближайшем знакомстве с системой становится очевидно, что канадским ученым свой-

<sup>38</sup> Марчук Ю. Н., Тихомиров Б. Д., Щербинин В. И. Система машинного перевода с английского языка на русский. — В кн.: Машинный перевод и автоматизация информационных процессов. М., 1975, с. 19—24.

<sup>39</sup> Chevalier M. et al. TAUM-METEO: description du système. Groupe TAUM. Université de Montréal, 1978.

ственная иная стратегия, которую можно оценить как «организованный дескриптивизм». Исходя из тезиса о том, что 100% автоматического перевода получить нельзя в принципе, авторы отказываются и от построения теории МП, а видят выход в максимально полной компиляции языковых параметров данного субъязыка, их анализе, классификации и последующей лексикографической интерпретации, которая и позволяет производить переход от текста на одном языке к тексту на другом языке на основе, так сказать, «подбора прецедентов» в соответствующем машинном словаре. Перевод в системе TAUM-METEO, так же как в другой отлаживаемой здесь системе<sup>40</sup>, нельзя назвать «пословным» в строгом смысле, но осуществляется он путем пословного анализа каждой фразы. Организующими звеньями системы являются английский словарь анализа входного текста и французский синтезирующий словарь, которые строились из соответствующих KWIC-конкордансов, составленных по метеосводкам для восточных районов Канады, и охватывали примерно по 6000 словоупотреблений. Таким образом, каждый из словарей содержит практически всю аналитическую информацию (семантическую, синтаксическую, морфологическую, стилистическую) о каждом слове, которую можно было встретить в исходном массиве текстов. Синтаксический анализ, например, использует методику построения деревьев-предложений, которая реализуется в машине в скобочной записи и обеспечивает статистически обусловленный выбор, шаг за шагом сужающий каждый раз набор допустимых продолжений фразы, фиксированных, в частности, в KWIC-конкордансе и соответствующих словарях.

Вообще словарь занимает центральное место в любой автоматизированной системе переработки текста. Для анализа венгерских текстов, например, такой словарь создан на базе семитомного толкового, параметры которого были частично переосмыслены и переориентированы. В итоге в автоматическом варианте он включает 10 комплексных параметров, из которых только два последних являются компилятивными<sup>41</sup>:

- 1) сложное или простое слово (из скольких корней состоит лексема),
- 2) омонимия,
- 3) часть речи (в случае конверсии каждая отдельная форма квалифицируется как самостоятельная),
- 4) число значений,
- 5) стиль,
- 6) морфологические данные (для единиц, способных к словоизменению),
- 7) управление (только для глаголов),

<sup>40</sup> *Isabell P. et al. TAUM-AVIATION: description d'un système de traduction automatisée des manuels d'entretien en aéronautique. Groupe de recherche en traduction automatique (TAUM). Université de Montréal. COLING, août 1978.*

<sup>41</sup> *Papp F. Automatic analysis of Hungarian texts and linguistic data. — In: Computational and mathematical linguistics, I. Firenze, 1977, p. 287. — Как раз на этом примере удобно продемонстрировать сокращение, сжатие общего числа параметров в автоматическом варианте словаря по сравнению с исходным толковым. Словарная статья в «A Magyar nyelv értelmező szótára» (v. I—VII. Budapest, 1959—1962) включает помимо названных параметры частичного произношения, ареальный, арг., подробно разработанный словообразовательный параметр, историко-лингвистический, а также элементы страноведческой информации.*

8) этимология,

9) суффикс,

10) длина слова.

Переразложение параметров позволило выделить новые аспекты их сопоставления и обобщения и в результате предложить решение ряда неясных вопросов в грамматике венгерского языка. В частности, один из способов образования притяжательной формы 3-го лица единственного числа — с *j* или без него, оценивавшийся как нерегулярный (*lab* — *laba* 'его нога', но *comb* — *combja* 'его бедро'), обнаруживает закономерность, согласно которой способ с *j* используется как маркированный, для обозначения необычного, редко встречающегося или недавно появившегося в венгерском языке сочетания согласных в конце основы. Новые закономерности выявлены также в распределении правил сингармонизма между разными группами слов.

Переклассификация синонимических и антонимических отношений слов в английском языке с целью выявления сети взаимосвязанных семантических классов и получения точной и максимально полной картины структурирования английской лексики осуществлена с помощью ЭВМ на базе Webster's New Dictionary of Synonyms<sup>42</sup>. Эта работа выполняется на материале дефиниций, и одной из аналитических процедур, к которым прибегают авторы, является различие между словарной единицей (dictionary entry) и лексической единицей (lexical item). Так, если в Уэбстерском словаре синонимов и антонимов многозначное слово *simple* трактуется как одна отдельная словарная единица с пятью значениями, то в машинном варианте словаря предлагается каждое значение рассматривать как самостоятельную лексическую единицу, т. е. ввести *simple 1*, *simple 2* и т. д. Использование лексической единицы дает возможность представить синонимические и антонимические отношения как бинарные и переходить к лексико-семантическим классам как к группировкам лексических единиц.

Традиционный лексикографический параметр дефиниции может подвергаться другим разновидностям аналитических процедур с целью получения новой информации о семантико-синтаксических свойствах слов. При создании автоматического словаря итальянского языка дефиниция эксплуатируется как бы дважды — с одной стороны, как носитель характеристики означаемого, как семантический эквивалент слова, с другой — как модель синтаксической сочетаемости характеризуемого слова<sup>43</sup>. При этом автор словаря — Дж. Феррари — исходит из следующих предпосылок: во-первых, основой для конструирования словаря, предназначенного для анализа текста, остается лексическая и семантическая информация, тогда как синтаксическая используется в редуцированном виде и минимальном объеме; во-вторых, каждая словарная статья — прямо или косвенно, т. е. в снятом, нейтрализованном виде, — отражает один или несколько типов контекстов данной единицы; и в-третьих, дефиниция каждой словарной единицы принадлежит одновременно и языку, поскольку передает сведения о некоторой реальности — линг-

<sup>42</sup> *Edmundson H. P., Epstein M. N. Research on synonymy and antonymy: a model and its representation. — In: Papers in computational linguistics. Budapest, 1976.*

<sup>43</sup> *Ferrari G. Dictionnaire automatique et dictionnaire machine: une hypothèse. — In: Computational and mathematical linguistics, I. Firenze, 1977.*



вистической или внеязыковой, и метаязыку, поскольку каждый ее элемент может рассматриваться как символ определенного дифференциального признака. Так, из определения КРАСОТЫ как «качества того, что красиво», мы узнаем не только, что такое «красота», но и то, что это слово принадлежит к классу качеств, свойств и обладает набором признаков (семантических, логических и синтаксических — как заместитель определенных позиций во фразе), характеризующих этот класс. Точно так же определение СОБАКИ как «домашнего млекопитающего из рода плотоядных» дает нам сведения, что в классе животных собака имеет свойства млекопитающих, а значит в лингвистическом плане это слово связывается с группой глаголов, называющих действия: ACCOUCHER, POULINER, VELER, METTRE BAS etc. В итоге общая структура автоматического словаря предстает в виде трехуровневого дерева, первое ветвление которого отражает набор генерализованных категорий, второе устанавливает иерархические отношения между словами в вокабуляре и третье фиксирует отношения эквивалентности между ними.

В зависимости от задач, на которые ориентирован словарь, его масштабы и соответственно объем аналитических параметров в нем могут существенно колебаться. В проекте универсального макрословаря русского языка, строящегося как объединение словников основных русских филологических, энциклопедических и специальных словарей и рассчитанного на 1 млн. лексем (соответственно 10 млн. словоформ), главная аналитическая проблема связана с экономным представлением парадигм. Ее решение достигается за счет отказа от явного задания всех словоформ и изменения структуры словаря. «Структура словаря должна представлять собой сочетание словарей двух типов: словаря машинных основ и словаря словоформ, сжатых по методу Купера. В словаре основ все словоформы данной парадигмы порождаются от общей машинной основы, представляющей собой цепочку букв от начала слова, общую для всех словоформ одной парадигмы. Эта машинная основа не обязательно должна быть связана с традиционной основой. Порождение нужной словоформы осуществляется стандартной операцией апплицирования машинной флексии»<sup>44</sup>. Иначе решается та же задача в учебном словаре русского языка на 4 тыс. слов, созданном в университетском колледже г. Кардиффа. Здесь, по нашим подсчетам, использовано 26 параметров, все словоформы введены в машину в явном виде, и обучающийся может сформулировать свой запрос в нужной ему форме — от любого параметра: например, вывести на дисплей всю парадигму данного прилагательного или только его степени сравнения; найти форму прошедшего времени несовершенного вида от глагола купить (= покупал, -а, -о, -и); выяснить, какими падежами управляет предлог «под», или найти все соответствия английскому *stir* <sup>45</sup>.

С добавлением к категориальным, формально-грамматическим па-

<sup>44</sup> Беляева Л. Н., Кризевич В. С., Липницкий С. Ф., Пиотровский Р. Г. О многоцелевом автоматическом словаре русского языка (МАРС). — В кн.: Вопросы общей и прикладной лингвистики. Минск, 1975, с. 156—157.

<sup>45</sup> Shibayev V. The CARLEX computerized analytical Russian dictionary. The computerization of complete complex grammar structures into dictionaries of highly inflected languages — CARLEX results. — In: The computer in literary and linguistic studies. Cardiff, 1976.

раметрам семантических усложняются задачи их аналитической обработки. В английском словаре, предназначенном для систем, понимающих естественный язык, наряду с набором постоянных, ядерных признаков каждая часть речи характеризуется еще специфической серией параметров, и все они являются результатом специального анализа. В число постоянных, относящихся ко всем частям речи, включены такие:

- 1) орфографический, т. е. сама словарная единица,
- 2) часть речи,
- 3) семантическое поле,
- 4) словарная дефиниция,
- 5) нерегулярная словоизменительная форма,
- 6) словообразовательная морфология,
- 7) синонимы, включая синонимические отсылки,
- 8) антонимы,
- 9) примеры употребления для каждого случая, имеющего дефиницию,
- 10) время реакции информанта на предложение и усредненный тип ответа по всем информантам,
- 11) данные об информанте.

Имя существительное, например, характеризуется сверх того следующими параметрами, которые извлечены из дефиниции с помощью особых приемов анализа:

- 1) синтактико-семантические признаки: ± человек, ± одушевленное, ± исчисляемое, ± конкретное, ± мужское, ± женское,
- 2) падежные маркеры (в смысле Филлмора),
- 3) метафорические переносы значения,
- 4) социолингвистические ограничения употребительности <sup>46</sup>.

Аналогичной структурой обладает дрезденский машинный словарь, в котором в силу его специфики отсутствуют психолингвистические и социолингвистические параметры <sup>47</sup>.

Двуязычные и многоязычные словари, опирающиеся на такие же аналитики, принципиально не отличаются по своему устройству от рассмотренных выше. Не имея возможности останавливаться подробно на их характеристике, ограничимся здесь упоминанием лишь некоторых отечественных и зарубежных работ такого рода <sup>48</sup>.

<sup>46</sup> Maxwell E. R., Smith R. N. A computerized lexicon of English. — In: Computers in the humanities. Minneapolis, 1974; Smith R. N., Maxwell E. R. An English dictionary for computerized syntactic and semantic processing systems. — In: Computational and mathematical linguistics, I. Firenze, 1977, p. 310—311.

<sup>47</sup> Нейберг Г. О структуре и применении машинного словаря научно-технических подязыков. — В кн.: Использование математических моделей и электронных вычислительных машин в лингвистике. София, 1976.

<sup>48</sup> Шалапина З. М. Англо-русский многоаспектный автоматический словарь (АРМАС). — В сб.: Машинный перевод и прикладная лингвистика, вып. 17. М., 1974; Deweze A. The Trilingual computational dictionary «THESEE». The-saurus, compatible with the French, English and German indexing systems. — В кн.: Использование математических моделей и электронных вычислительных машин в лингвистике. София, 1976; Stachowitz A. Beyond the feasibility study: lexicographic progress. — In: Computational and mathematical linguistics, I. Firenze, 1977; Wilton M. T. Bilingual lexicography: computer-aided editing. — Ibid.; Зиман Ю. Л., Егорова Е. В. Использование ЦВМ для автоматизации составления и редактирования словарей. М., 1977. (Ин-т русского языка АН СССР. Проблемная группа по экспериментальной и прикладной лингвистике. Предварительные публикации, вып. 98).

Завершая рассмотрение автоматических словарей и систем с использованием большого числа аналитических параметров, остановимся еще на двух работах, которые представляются наиболее полными по охвату параметров и наиболее сложными по их структурированию из всех, отнесенных нами к данной группе. Первая из них — снимающий неоднозначность словарь (The disambiguation dictionary) для автоматического распознавания смысла высокочастотных английских слов <sup>49</sup>. Как известно, наиболее употребительные слова являются и самыми многозначными. Вместе с тем из 117 значений и употреблений глагола *take* или 91 значения существительного и конвертированного глагола *hand* (по словарю The Random House Dictionary of the English Language) далеко не все оказываются употребительными и даже реально встречающимися в современных английских текстах и устном общении. Авторы поставили перед собой задачу проанализировать, в каких значениях выступают 1815 самых частотных слов в текстах на полмиллиона словоупотреблений, относящихся к различным сферам общественной жизни, и разработать основы словаря, который позволял бы в ходе контент-анализа однозначно идентифицировать смысл, в котором употреблено данное многозначное слово в данном контексте. Результатом работы системы является приписывание каждой единице анализируемого текста номера значения этой единицы, зафиксированного в словарной статье автоматического снимающего неоднозначность словаря. Последний, таким образом, имеет два источника — корпус текстов и толковый словарь. Текст был обработан на уровне компилятивных параметров и представлен в виде KWIC-конкорданса (key word-in-context) для 1815 исходных слов — со всевозможными статистическими характеристиками. Параметры же толкового словаря подверглись аналитической обработке с неизбежной в подобных случаях редукцией некоторых из них (в частности, этимологического, исторического, словообразовательного, произносительного и некот. др.).

В числе аналитических процедур были следующие: создание анализатора словоизменительных форм, который давал возможность представить каждую словарную единицу как основу, к которой прибавляется — в случае регулярной парадигмы — определенный набор суффиксов (напр., *-s*, *-ed*, *-ing* для глагола); установление части речи, в функции которой выступает данная форма, что в значительной степени определяет синтаксические свойства ее ближайшего окружения; идентификация — с помощью толкового словаря — значения слова, реализованного в данном контексте, с выявлением неопределенных, двусмысленных или наоборот генерализованных его употреблений и т. п.

Отличительная особенность полученного словаря заключается в том, что в нем исключительно детальную характеристику приобретают синтаксические и семантические параметры, названные здесь «категориальными маркерами». Среди десяти групп синтаксических параметров:

— детерминативы, включающие помимо артиклей (*a*, *an*, *the*) и демонстративов (*this*, *these*; *that*, *those*), также показатели принадлежности («генитивы» — в терминологии авторов — *my*, *our*, *your*...), числительные (в двух субкатегориях — количественные и порядковые)

<sup>49</sup> Kelly E. F., Stone Ph. J. Computer recognition of English word senses. Amsterdam—Oxford, 1975.

и предартиклы (исчисляемые — *both*, *each*, *few*, *every*..., и смешанные — *all*, *any*, *some*, *half*...);

— указатели позиции слова во фразе (первое слово, последнее слово);

— специальные глаголы и глагольные элементы, содержащие девять субкатегорий, в том числе — глаголы-связки (*seem*, *look*, *feel*, *remain*...), глаголы превращения (*become*, *get*, *turn into*, *grow*...), модальные глаголы (*can*, *may*, *shall*...) и др.

Семантические параметры классифицируются в соответствии со стандартной практикой контент-анализа и представлены в данном случае семантическими полями, часть которых предполагает еще и дополнительное внутреннее подразделение: I одушевленное, II человек, III коллективное, IV абстрактное имя, V общественное место, VI часть тела, VII политическое понятие, VIII экономическое понятие, IX цвет, X коммуникация (в том числе — газета, книга, телефон, печать, карандаш...), XI эмоции, XII частота (редко, случайно, часто...), XIII оценочные прилагательные (плохой, хороший, прекрасный, трудный...), XIV прилагательные величины, XV прилагательные позиции (низкий, верхний, средний...), XVI наречия степени <sup>50</sup>.

В результате применения специальных правил анализа текста и его параметризации получен автоматический словарь, содержащий очень большой объем самой разнообразной информации и находящий применение не только при контент-анализе, но и при информационном поиске и некоторых других видах работ с текстами.

Наконец, последняя из рассматриваемых систем относится к машинному представлению всех лексикографических параметров Уэбстерского словаря — и компилятивных, и аналитических — с целью автоматического редактирования, переклассификации используемых данных, добавлений и изменений при перендании, а также для возможного применения его к задачам автоматического поиска информации. Система разработана на базе стандартного способа структурирования библиографических данных MARC (Machine Readable Catalog), имеющего распространение в Великобритании и США, и в применении к указанному словарю получила название WEBMARC <sup>51</sup>.

Приведение одной из словарных статей данного словаря в качестве примера потребовало бы слишком развернутого комментария из-за сложности используемой записи, поэтому мы ограничимся здесь самой общей характеристикой системы. Если исходить из понятия параметра, развиваемого в наших работах, то число таких параметров в системе WEBMARC не превышает 19. В классификации автора этой системы опорными лексикографическими категориями являются «поля», которых выделяется шесть: поле побуквенной записи (орфографическое), поле дефиниции, поле полного произношения, синонимическое поле, этимологическое поле, поле частичного произношения, — и которые подразделяются на субполя. Между полями и субполями, с одной стороны, и параметрами — с другой стороны, нет однозначного соответствия: параметр

<sup>50</sup> Kelly E. F., Stone Ph. J. Op. cit., p. 17—21.

<sup>51</sup> Sherman D. A common structure for lexicographic data. — In: Computers in the humanities. Minneapolis, 1974.

может совпадать и с полем (например, этимологический), и с субполем (например, омографический), некоторым субполям и даже полям (например, «частичного произношения») может не соответствовать ни один параметр, и наоборот, одно субполе может включать два параметра (например, субполе, названное «точки слогоделения» — hyphenation points — и кодированное знаком \$b в орфографическом поле, содержит информацию о числе слогов (один параметр) и количестве букв в каждом слоге (другой параметр), считая с конца, от точки, отделяющей последний слог слова; так, для слова *dic. tio. nary* этот код записывается \$b33).

Рассмотрим подробнее субполевое структурирование орфографического поля в соотношении с нашими параметрами. Прежде всего, до побуквенной передачи слова, оно (поле) содержит указание на класс слов, к которому принадлежит данное, причем понятие «класс слов» оказывается несколько шире категориального лексикографического параметра «часть речи», поскольку включает частично морфологическую (PF — prefix, SF — suffix, VS — verb suffix...), частично словообразовательную информацию (CF — combining form, NC — noun combining form...). Таким образом, мы имеем основание записать:

класс слов = параметр «часть речи», параметр морфологического членения, параметр словообразования.

Далее следуют субполя:

\$a — побуквенная запись = собственно орфографический параметр.

\$b — точки слогоделения = параметр числа букв, параметр слогоделения.

\$c — показатель ударения = параметр ударения.

\$d — статус альтернативного класса слов (в случае конверсии) = ∅.

\$e — код альтернативного класса слов = ∅.

\$i — статус принадлежности данной единицы к тому или иному социальному срезу языка (dial, slang...) = ареальный параметр (указание региона), параметр арго (указание типа арго).

\$j — варьирование, типы вариантов = ∅.

\$s — статус вариантов (равноценный, предпочтительный, второстепенный) = параметр нормативности.

Итак, пример рассмотренной системы со всей наглядностью показывает, что широкая разработка и использование аналитических параметров всегда опирается на толковый словарь, т. е. источник получения их вторичен по отношению к тексту, тогда как результат их обработки и систематизации, как бы возвращается к тексту на новом уровне, поскольку предназначен для его автоматического анализа.

Что касается систем МП, с неизбежностью опирающихся в первую очередь на аналитические лексикографические параметры, то, как было показано выше (в частности, на системе TAUM-METEO), наряду с использованием толковых словарей для их выделения широко привлекаются и тексты соответствующего субъязыка. Практически действующие ныне системы МП не могут пока обходиться без достаточно объемных хранилищ текстовой информации (например в виде KWIC- или KWOC-конкордансов). Но тенденции развития исследований в области машинного перевода таковы, что главной проблемой на пути перехода от эмпири-

ческих систем к системам с теоретически разрешимыми задачами становится анализ семантической информации и ее систематическое кодирование для машинных словарей. Это связано уже с оперированием лексикографическими параметрами высшей степени сложности, названными нами конструктивными, к рассмотрению которых мы теперь и переходим.

#### § 4. ЛЕКСИКОГРАФИЧЕСКИЕ АСПЕКТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Вынося в заголовок раздела термин «искусственный интеллект», мы вовсе не претендуем на рассмотрение комплекса проблем, пусть даже только лингвистических, которые влечет за собой это понятие. Оно необходимо нам потому, что именно в этом направлении многие исследователи видят пути совершенствования машинного перевода, т. е. перехода от аналитических к конструктивным параметрам, и именно последние оказываются теми кирпичиками, из которых складывается собственно языковая часть искусственного разума и которые служат задачам моделирования мышления человека. «Необходимость в моделировании разумной и языковой деятельности человека возникает при создании интеллектуальных роботов третьего поколения, способных производить такие операции, как распознавание образов и смыслов, формирование понятий, принятие решений, соответствующих тем целям, которые возникают в ходе взаимодействия робота с внешней средой»<sup>52</sup>. Известный парадокс можно усмотреть в том, что для моделирования разумной деятельности прежде всего оказывается необходимой модель (опять-таки «модель»!) внешнего мира, отражающая «знания» искусственной интеллектуальной системы, воплощающая ее «картину мира» (ср. II главу). Важнейшей частью элементов для конструирования такой модели и являются конструктивные лексикографические параметры.

Под конструктивными, как мы помним, понимаются параметры, требующие для своей обработки построения специальных эвристик. Как правило, это параметры, включающие семантический компонент — словообразовательные, синтагматические (свободная и связанная сочетаемость), ассоциативные, лингвострановедческие, терминологические, параметры словоизменения (склонения, спряжения, степеней сравнения, наклонения) и др., и потому как бы подразумевающие присутствие человеческого фактора. Оперирование конструктивными параметрами с необходимостью предполагает и предварительную компиляцию соответствующих данных, и анализ определенных характеристик, но главной особенностью является связанное с ними решение эвристических задач, которое ведет к созданию новых лингвистических объектов. Новым лингвистическим объектом мы называем такой объект (это может быть словарь, лингвистический атлас, функционирующая модель фрагмента языковой структуры, целеориентированная грамматика и т. п.), который возникает не в результате описания некоторого языкового материала, а в результате эксперимента. Этот объект, будучи построенным на определенных теоретических предпосылках, сам дает новый материал для дальнейших наблюдений, исследований, выводов о данном языке. Таким образом, в случае

<sup>52</sup> Пиотровский Р. Г. Лингвистические аспекты «искусственного разума». — ВЯ, 1981, № 3.

работы с конструктивными параметрами одно их свойство как бы обуславливает другое: решение эвристических задач опирается на эксперимент, а возникающий в ходе эксперимента новый лингвистический объект позволяет генерировать новую информацию о языке, выявлять нетривиальные, иногда непредсказуемые его свойства.

В числе исследований подобного рода — работы разной степени сложности: в одних ставится задача построения, так сказать, однопараметровых объектов, другие направлены на объединение параметров, их синтез и универсализацию полученного объекта. Рассмотрим некоторые примеры.

Автоматический словарь эстонских глаголов<sup>53</sup> решает задачи опознания грамматического значения любой правильной формы (т. е. позволяет приписывать некоторой словоформе соответствующее время, лицо, залог, наклонение, число и т. п.) и задачи конструирования необходимой словоформы по заданному набору грамматических значений. Построение такого словаря для сильно флективного языка, каким является эстонский, потребовало представления каждого члена парадигмы в виде особой программы, которая предусматривает эвристическую задачу установления по характеру основы — типа спряжения, соответствующих ему правил дистрибуции основы и ее морфонологической перестройки в процессе глагольного словозменения. Сокращение программных процедур и их типизация в машинном эксперименте достигается с помощью оптимальной классификации самих основ, флексий и типов морфонологических чередований. Важной особенностью данного словаря является то, что он значительно увеличивает число «входов» в морфологию. Если в обычном описании вход осуществляется либо от словоформы (анализ), либо от слова-образца (т. е. через парадигму), то в автоматическом словаре при сохранении указанных возможностей вход осуществляется также и от перечня грамматических значений (как от каждого значения в отдельности, так и от различных их комбинаций), и от перечня форм (классический анализ), и от классификации. Последняя, являясь организующим стержнем этого нового лингвистического объекта, представляет собой предельно сжатое выражение структуры эстонской глагольной морфологии, т. е. сетку запретов и разрешений, которая как бы сдерживает «комбинаторный взрыв» логически возможных комбинаций грамматических значений и их фонологических репрезентаций. Этот словарь на ограниченном материале глагольного словозменения моделирует в известном смысле знание языка, поскольку с достаточной надежностью по типу основы и ее месту в классификации предсказывает ее морфонологическое и морфонологическое поведение, т. е. относит ее к классу определенных парадигм.

Машинный толковый терминологический словарь, относящийся к области ядерной физики и по своему типу приближающийся к традиционному энциклопедическому, разработан как составная часть системы, понимающей естественный язык<sup>54</sup>. Ограниченность предметной области дает здесь возможность рассматривать наименования ядерных реакций

как единый тематический ряд. План содержания каждой из подгрупп терминов, входящих в тематический ряд, описывается методом компонентного анализа, и все описания соединяются в одном классификаторе. Особенность построенного таким образом аналитико-смыслового словаря в том, что принцип порождения распространяется и на план содержания. Разработанный классификатор содержит три разбиения и дает возможность конструировать наборы семантических множителей (позиционные смысловые коды), которые однозначно соответствуют содержанию определенного научного понятия. Связи и отношения, свойственные описываемой предметной области, находят отражение в самой структуре классификатора.

При составлении позиционного кода из каждого разбиения берется по одному семантическому множителю, и смысл любого термина предстает в виде трехместного цифрового кода, внутренним свойством которого является значимый порядок следования позиций. Так, термину «упругое рассеяние нейтронов» соответствует смысловой код 2.2.1, где цифра первой позиции характеризует тип налетающих частиц (нейтроны), цифра второй позиции — тип вылетающих частиц (нейтроны), цифра третьей позиции характеризует состояние остаточного ядра (остаточное ядро равно начальному). Для порождения плана выражения соответствующего термина строятся матрицы перехода, отражающие связь между двумя планами. Использование матриц позволяет выявить такие закономерности в отношениях между планом содержания и планом выражения, которые в вербальной формулировке остаются скрытыми. Эксплицитное обозначение этих закономерностей служит основой для представления их в виде алгоритма. «Комбинаторный взрыв» и порождение бессмысленных кодов нейтрализуются здесь тоже специальными семантическими алгоритмами.

Таким образом, на основе алгоритмов, программ и исходных данных ЭВМ строит словарь, состоящий из эксплицитной записи смысла термина в соответствии с классификатором и формы выражения этого смысла в виде терминологического словосочетания или символа. Словарь хранится в ЭВМ в потенциальной форме, допускает расширение за счет введения новых разбиений в классификатор, новых дифференциальных признаков в каждое разбиение и т. д. и не требует при этом пересмотра всей структуры. Кроме того, данный словарь предусматривает использование его для лексикографического перевода терминов той же предметной области на английский и французский языки.

Рассмотренный лингвистический объект использует несколько эвристик, имитирующих знание языка: во-первых, умение отсеивать бессмысленные наборы семантических множителей, что требует «знания» о мире, знания данной предметной области; во-вторых, умение переходить от смысла к его выражению с учетом структурных особенностей соответствующего языка — возможностей комбинирования основ с флексиями и типов модификаций основ при образовании словоформ — для русского, или выбора формы артикля — для французского и т. д.

Присутствие человеческого фактора наиболее отчетливо проявляется при разработке такого параметра, как ассоциативный. Другой характерной чертой опытов автоматического построения ассоциативных словарей можно считать то, что в них наглядно подтверждается наличие системных

<sup>53</sup> Виск Ю. А. Классификаторная модель эстонской морфологии (автоматический синтез глагольных словоформ). Канд. дис. Таллин, 1978.

<sup>54</sup> Марюгина А. И. Лингвистические основы и структура аналитико-смыслового словаря для систем, понимающих естественный язык. Канд. дис. М., 1978.

связей между самими параметрами, обнаруживаются тесные взаимозависимости внутри отдельных их подгрупп. Установление же таких зависимостей, в свою очередь, несет новую информацию и о языковой структуре в целом, что дает основания отнести этого рода работы к рассматриваемой, третьей группе конструктивных параметров.

Что касается первой из названных черт — человеческого фактора, — то его присутствие совершенно очевидно, поскольку машинной обработке в этом случае подвергаются результаты тестовых ассоциативных экспериментов. Так, авторы ассоциативного тезауруса английского языка, отобрав по психолингвистическим (т. е. по принципу наиболее принятого использования данных слов в качестве стимулов в ассоциативных экспериментах), частотным и смысловым критериям 8400 слов-стимулов, проанализировали затем анкеты, полученные от 100 информантов, каждому из которых было предложено ответить на 100 стимулов<sup>55</sup>. Тезаурус представлен в виде сетей словесных ассоциаций и содержит 55 732 узла, которые характеризуются входящими связями. Исходящие связи тоже учитываются. В другой форме представления данный тезаурус выступает как перечень слов-стимулов с указанием к каждому из них всех входящих и исходящих реакций в порядке убывания их частотности. Наконец, для удобства пользования представлен и обратный перечень — всех входящих связей, какие только появились в ответах испытуемых. Ценной особенностью этого словаря является то, что в нем не просто зафиксированы, а затем проанализированы слова-реакции в их лемматизированной форме, а учтены словоизменительные разновидности, в которых эти слова выступали в ответах информантов, зафиксированы ошибки, варианты произношения и т. п. Статистическая обработка этих данных и их семантический анализ позволили выявить взаимозависимости между собственно ассоциативным параметром, частотным и рядом морфологических характеристик. Например, определение силы ассоциативной связи — как для входящих, так и для исходящих отношений — опирается на частотный критерий; обобщение данных по частоте словоизменительных форм в ответах приводит авторов к некоторым психологическим оценкам отдельных аспектов английской грамматики.

Указанные особенности выгодно отличают рассмотренный ассоциативный тезаурус от другого подобного ему словаря английского языка, в котором ассоциативный параметр выступает как бы в изолированном виде<sup>56</sup>. Здесь автор аксиоматически вводит 118 понятийных категорий, которые выступают как семантические примитивы и являются, по его мнению, достаточными для передачи значения любого слова. Они получают условные мнемонические обозначения и могут образовывать различные комбинации (AIM, BODY, EVER, GLAD, HEAR, LADY, MOTV, TIME, WRIT, NO...). Как справедливо отмечалось в рецензии на этот словарь, даже для выявления только сети ассоциативных связей «простые корреляции между словами оказываются неадекватными вне их соотношений с другими аспектами языковой структуры»<sup>57</sup>.

<sup>55</sup> Kiss G., Armstrong Ch., Milroy R., Piper J. An associative thesaurus of English and its computer analysis. — In: The computer in linguistic and literary studies. Edinburgh, 1973.

<sup>56</sup> Laffal J. A concept dictionary of English. Essex—N. Y., 1973.

<sup>57</sup> Huntman J. F. A concept dictionary of English, by Julius Laffal. — Computers and the Humanities, 1975, vol. 9, N 1, p. 47.

Необходимость использования — при построении новых лексикографических объектов — не только собственно конструктивных, но и параметров других рангов, а именно, компилятивных и аналитических, можно продемонстрировать на примере организации семантического архива английского языка в Объединении системных исследований Санта Моника (США)<sup>58</sup>. Само название этой системы («Архив») ориентирует как будто на восприятие ее как банка лексико-семантических данных, однако уже в проекте ее создания предусматривалось, что она будет функционировать одновременно и в роли информационного центра, и в качестве исследовательской базы — источника получения новых сведений о лексике английского языка. Исходные данные этой системы компилированы из четырех типов источников — Уэбстерского словаря (7-е издание), тезауруса Роже, «Брауновского корпуса» (Standart Corpus Present-Day Edited American English)<sup>59</sup> и большого круга лингвистических, философских, психологических, антропологических и другого рода гуманитарных исследований, в которых так или иначе анализируется значение и употребление конкретных слов, вошедших в общий лексикон по трем первым источникам.

Таким образом, в отличие от упоминавшегося французского Trésor, библиографический параметр в системе SOLAR не просто выполняет отсылочную функцию, а получает глубинную, содержательную интерпретацию, снабжая лексикографа и лексиколога сведениями о детальной разработке семантики отдельных слов. Суммарный объем используемого лексикона охватывает свыше 10 млн словоупотреблений. Весь материал систематизирован в 9 разрядах (группах) параметров — 7 основных и 2 вспомогательных, для каждой из которых разработан свой машинный формат. Первые три группы: 1) семантический анализ, 2) объяснение дескриптивных констант, используемых для семантического анализа, 3) концептуальный анализ понятий, связанных с дескриптивными константами — обрабатываются и подготавливаются к вводу в машину ручным способом. Остальные группы основных параметров — 4) дополнительные признаки, коррелирующие с отдельным значением слова, 5) семантические поля, 6) развертывание дефиниции, 7) контексты употребления, а также вспомогательные — индекс слов и библиография — обрабатываются автоматически.

Соотношение друг с другом трех первых групп параметров, в которых каждая последующая выступает как разъяснение существа предыдущей, позволяет оценить их как найденный авторами удачный, на наш взгляд, прием, останавливающий регресс в бесконечность при анализе лексической информации, обусловленный свойством ее «циркулярности» в словаре, где для объяснения значения слов употребляются те же самые слова. Так, семантический анализ строится на описании значений слов, содержащихся в специальных работах — главным образом лингвистов и филологов, и включает следующие параметры:

— установление предикатно-аргументных отношений, причем под предикатом понимается анализируемое значение, а дескриптивные

<sup>58</sup> Diller T., Olney J. SOLAR (A Semantically-Oriented Lexical Archive): current status and plans. — Computers and the Humanities, 1974, vol. 8, N 5—6.

<sup>59</sup> Kučera H., Francis W. N. Computational analysis of present-day American English. Brown University Press, Providence, 1967.

константы (т. е. семантические маркеры = семантические компоненты = семантические дифференциальные признаки = семантические примитивы), соответствующие этому значению, указывают на роль выражений, функционирующих в качестве аргументов;

- компонентный анализ значений;
- установление пресуппозиций отдельного значения;
- коннотации данного значения;
- условия его распространения, т. е. возможности его развития в предложении;
- допустимые дополнения;
- субкатегориальные признаки.

Объяснительные примечания к дескриптивным константам, используемым в семантическом анализе, содержат: толкование нестандартных символов; определение данной константы, извлеченное из оригинального текста автора, употребляющего и разрабатывающего эту константу; отсылки к другим словам, в анализе которых участвует данная константа. Наконец, концептуальный анализ, интегрируя итоги предыдущего, базируется на результатах аналитической философии и включает рассмотрение центральных понятий, связанных с дескриптивными константами, таких как «факт», «событие», «причина», «личность», «действие» и т. п.

Для иллюстрации процедуры обработки параметров первых трех групп приведем один пример. Семантический анализ слова *excuse* (извинять, прощать) выявил в качестве одной из дескриптивных констант *Defendant* (ответчик). Согласно Ч. Филлмору, использовавшему эту константу, понятием более высокого ранга, участвующим в ее определении, является *responsible* (ответственный). В соответствующем формате на это понятие критически рассматриваются различные точки зрения нескольких авторов, занимавшихся его анализом (в том числе и определение его в Уэбстерском словаре), и делается попытка сформулировать интегрированное представление о его содержании.

Что касается других групп параметров, то они в значительной степени традиционны.

Так, к дополнительным признакам значения авторы системы относят обычные стилистические пометы о сферах использования и ограничениях в употреблении, пояснения, даваемые к словарной дефиниции в скобках, иллюстрации употребления (которые одновременно представляют собой и часть общего конкорданса и потому дублируются 7-й группой параметров), некоторые сведения о сочетаемости, извлеченные из работ по «формальному семантическому анализу» рассматриваемой единицы.

Семантические поля (5-я группа) komponуются главным образом на основе словарей Уэбстера и Роже и строятся на двух типах отношений — дефиниционных и синонимических: некоторое слово *x* включается в поле данного *y*, если, во-первых, оно участвует в его дефиниции или, наоборот, данное слово *y* входит в дефиницию *x*; во-вторых, если находится с данным в одной из широко понимаемых синонимических связей (напр., *end* считается синонимом *last*). Тип взаимоотношений каждого слова в поле с заглавным указан, так же как указана и дополнительная информация — часть речи, число значений у слова помимо того, в котором оно включено в данное поле (так называемые «омографы»), переходность — непереходность у глагола и др. Распространение дефиниции (6-я группа парамет-

ров) осуществляется на двух уровнях — по общезыковому, т. е. филологическому словарю, а затем по специальному, в зависимости от той области, к которой относится данное слово (металлургия, горное дело, химия и т. д.): из дефиниции Уэбстерского словаря выбираются слова, которые имеются в словнике терминологического словаря соответствующей области (первый уровень), и из последнего выписываются их дефиниции (второй уровень). Наконец, последний — 7-й формат представляет собой обычный KWIC-конкорданс (key word-in-context), составленный по всем используемым в системе источникам.

Уже простое перечисление использованных в системе SOLAR параметров показывает, что их состав выходит за рамки собственно лексикографических (хотя, естественно, и не все из квалифицированных нами как лексикографические нашли отражение в этой системе: здесь отсутствуют, например, этимологический, словообразовательный, словоизменительный, производительный и многие другие параметры). Поэтому система способна не только служить базой при составлении различных словарей, но и давать ответы на широкий круг запросов, требующих и аналитической работы с исходными данными, и такой их переклассификации, которая выявляет неизвестные ранее свойства лексической семантики английского языка.

Подобные исследования находятся уже на грани между лексикографическим объектом и моделью языка в целом, поскольку вплотную подводят к процессам «понимания» сообщения, к уровню принятия решений.

Охарактеризованная система SOLAR, например, в принципе обладает теми же «знаниями» (только в иных пропорциях), что и системы, «понимающие естественный язык», а именно — семантикой, синтаксисом, логикой и представлением о мире (т. е. знанием определенной предметной области), отличаясь от таких систем меньшим динамизмом, но зато превосходя их масштабами<sup>60</sup>.

Определенную аналогию с разумным поведением (искусственного интеллекта) можно усмотреть в автоматическом ведении словаря в процессе изменения массива текстов, для которых он первоначально был составлен.

Точно так же как тезаурус личности по мере изменения и роста опыта человека, постоянно находясь во взаимодействии со средой, поддерживает с ней основанные на обратных связях отношения динамического информационного равновесия, могут быть организованы и искусственные системы. Так, можно представить себе систему (например, ИПС), в которой удаление некоторой старой документации и соответствующих терминов и введение новой документации с новыми терминами вызывает пополнение словаря, перекомплектование групп терминов и реорганизацию базовых отношений между ними. О практической реализации в Королевском университете г. Кингстона (Канада) такой системы, в которой все этапы ее функционирования — построение словаря, индексирование, классификация и поиск — осуществляются автоматически, рассказывает Р. Крауфорд<sup>61</sup>. В обычной (статистической) ИПС термино-

<sup>60</sup> Ср., напр.: Виноград Т. Программа, понимающая естественный язык. М., 1976.

<sup>61</sup> Crawford R. G. Dynamic dictionary updating. — Information processing & management, 1977, vol. 13, N 4.



логический файл со словарной информацией содержит следующие элементы:

a) *Dictionary*

<i>Term</i>	<i>Concept Number</i>
Bacillus	1031
Bacteria	172
Basophilic	3019
Behavior	462
Bening	781

b) *Document Vector*

document number 1291
concept/weight pairs
172/1, 191/2, 367/1, 491/1
567/3, 3410/1, 4096/2

Предположим, — говорит автор далее, — что некоторый термин «ocular» не используется в этой системе для индексирования совокупности документов по офтальмологии, потому что он встречается в большей части этих документов и не может выполнять функцию различителя. Однако по прошествии некоторого периода времени состав документов расширился за счет добавления их из более общих областей медицины. Одновременно ряд специфических документов по офтальмологии был исключен из базы данных. В итоге в получившемся множестве термин «ocular» стал менее частотным, превратился в эффективный различитель и должен быть включен в число индексирующих терминов. Каким образом можно обеспечить автоматическую перестройку системы с учетом изменения статуса термина?

Для обеспечения динамического поведения этой системы предлагается в терминологический файл, наряду с указанными выше параметрами — отнесенностью термина к определенным дескрипторам и областям, числом документов и представленностью термина в каждом документе, — ввести дополнительные характеристики статуса каждого термина (*Term Status Map*), включая определение его индексирующих возможностей, вычисляемых автоматически по специальным формулам, в которых учитываются статистические зависимости между соответствующими показателями на разных массивах, характеристики «компактности» самих массивов и различительная сила термина. Предложенный в статье способ решения задачи оказывается принципиально близким к тем приемам, которые были использованы нами при автоматическом получении идеографической классификации русской лексики (см. главу II).

Завершая наш обзор, представляется необходимым остановиться на характеристике еще одной системы, которая по существу не относится к лексикографии прямо, но по своему устройству и функционированию аналогична словарю. Речь идет о фонде (банке данных) эвристических приемов решения конструкторско-изобретательских задач, и понятно, что единичей в этом своеобразном словаре уже становится не слово, а эвристический прием, т. е. краткое предписание (правило), как преобразовать прототип, чтобы получить новое техническое решение, или как синтезировать новое решение<sup>62</sup>. Источником выявления и систематизации эвристических приемов при составлении фонда послужила обширная мировая литература, посвященная методам поиска новых технических решений. В республиканской научно-исследовательской лаборатории ма-

тематических методов оптимального проектирования при Марийском политехническом институте им. М. Горького проанализированы методы: морфологического ящика (США), матрицы открытия (Франция), организуемых понятий (ГДР), алгоритма решения изобретательских задач (СССР), систематической эвристики (ГДР), ступенчатого подхода к решению задачи (США), функционального изобретательства (Англия), интегральный метод открытия (Франция), десятичных матриц поиска (СССР), мозгового штурма (США) и мн. др. Всего собрано около 800 не повторяющих друг друга приемов, которые классифицированы в несколько групп и распределены по подфондам, представляющим собой массивы информации, используемые на различных этапах работы обобщенного эвристического алгоритма поиска новых технических решений. Помимо того, что созданная машинная библиотека («словарь» — в нашем смысле) эвристических приемов и поисковых процедур может использоваться как обычный банк данных, т. е. как вспомогательное информационное средство при решении научных, технических и организационных проблем, этот «изобретательский словарь» служит центральным узлом реализованной в Марийском политехническом институте автоматической системы синтеза новых технических решений. Система прошла апробирование, и с ее помощью найдены принципиально новые пути конструирования запоминающих устройств, элементов аналоговых вычислительных машин, сооружений из мягких материалов и др.

Эта работа является пионерным исследованием в области изучения и формализации приемов научного творчества: она вплотную подводит к созданию автоматической модели эвристических способностей человека, т. е. фактически создает новый научный объект, обладающий в том числе и лингвистическими свойствами, что в условиях отсутствия общей теории творчества<sup>63</sup> является важным шагом по пути ее построения. Понятно, что в этих своих качествах данная система далеко выходит за рамки лексикографии, вторгаясь в сферу исследования проблем искусственного интеллекта. Вместе с тем она демонстрирует основополагающую роль собственно словаря в решении целого ряда эвристических задач.

Подводя итоги сделанному обзору, следует констатировать, что в ходе автоматического построения различных лексикографических объектов с необходимостью осуществляется анализ метаязыка словаря (в широком смысле), причем в зависимости от характера этого объекта акцент делается на той или иной группе параметров, каждая из которых для своей обработки требует специфического подхода, использования особых приемов и процедур. Естественно, что при обработке параметров высшего уровня нельзя миновать предыдущие, поэтому работа, например, с аналитическими параметрами предполагает предварительную компиляцию, а оперирование конструктивными не может обойтись без предшествующего ему анализа. Как показывает опыт составления тезауруса

<sup>62</sup> Методы поиска новых технических решений. Под ред. А. И. Половинкина. Йошкар-Ола, 1976.

<sup>63</sup> Как констатируют многие исследователи и в нашей стране, и за рубежом, вопрос о создании подобной теории по-настоящему еще не поставлен на научную основу. Ср., напр.: Бун Г. Я. Методологические основы научного управления изобретательством. Рига, 1974; Bender A. D. Creativity and productivity programs — how widely they used? — Research Management, 1975, vol. 18, N 5; Mc Pherson J., Andrews W. Collected insights for enlightened managers/administrators. — Journal of Creative Behavior, Buffalo (N. Y.), 1971, vol. 5, N 3.

русского литературного языка (см. II главу), все эти этапы являются неотъемлемой частью работы над ним, однако в силу особенностей самого этого объекта, для которого определяющими являются семантические параметры (т. е. параметры конструктивного уровня), центральные проблемы на пути достижения поставленной цели связаны с анализом прежде всего семантической информации. Характеристике этих проблем и способов их решения и посвящается следующая глава.

## Глава II

### АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ РУССКОГО ТЕЗАУРУСА КАК СПОСОБ СЕМАНТИЧЕСКОГО АНАЛИЗА МЕТАЯЗЫКА ТОЛКОВОГО СЛОВАРЯ

#### § 1. ЛИНГВИСТИЧЕСКАЯ ТЕХНОЛОГИЯ КОНСТРУИРОВАНИЯ ТЕЗАУРУСА И ОБЩАЯ ХАРАКТЕРИСТИКА РЕЗУЛЬТАТА

Внешние параметры словаря, построенного авторами с помощью ЭВМ, таковы: он содержит 1600 дескрипторов (или основных понятий, концептов, идей, охватывающих и организующих всю русскую лексику) и около 9000 слов, которые распределены машиной по этим дескрипторам на основе семантических связей. Тезаурус в печатном виде состоит из двух частей<sup>1</sup>: первая — «от понятия к слову» и вторая — «от слова к понятию». Таким образом, раскрыв первую часть тезауруса, пользователь может получить информацию о том, например, какие слова (из исходного списка в 9000 слов, и только из него!) относятся к, или раскрывают смысл, или семантизируют и т. п., понятие БЕДА:

беда	искупление	печальный	спасение
бедный	испытание	плакать	страдать
выход	катастрофа	плач	суровый
гибель	кризис	плачевный	темный
гибнуть	мрачный	плохой	тонуть
голод	наводнение	пожалеть	трагедия
горе	наказание	постичь	траур
горький	напасть	пострадать	трудность
грустный	несчастный	потерянный	трудный
жалеть	несчастье	прикорбный	тяжелый
жертва	опасный	пронести	утешение
затруднение	печаль	сожаление	
зло	печаль	сострадание	

Аналогично представлены следующие, например, понятия:  
**БЕРЕЖЛИВОСТЬ**

бережливость	осторожный	хозяйственный	экономный
беречь	расчет	хранить	
жалеть	скупой	экономить	ДОСКА
жалко	соблюсти	экономия	бревно

<sup>1</sup> О третьей части тезауруса, существующей в машинном исполнении и представляющей собой банк семантических множителей, см. ниже.

доска	МАНЕРЫ	нравственность	претензия
замок	вид	облик	привить
лапа	вкус	образ	привиться
лист	владеть	обхождение	привыкнуть
основа	внешний	общество	привычка
пила	воспитание	обычай	привычный
планка	врезаться	окружать	придать
пластинка	выражение	окружающий	придерживаться
плита	выразительный	перерождение	приличие
плитка	герой	печать	примазаться
плот	держат	повадиться	принять
подоконник	держаться	повадка	прислушаться
полено	думать	поведение	присмотреться
продольный	красивый	поддерживать	приучить
просвет	круг	подобие	свойственный
распилить	линия	подхватить	склад
рубить	люди	политика	совесть
столб	манера	правило	среда
ступень	мир	практиковаться	туалет
сук	мода	предрассудок	улица
толщина	наклонность	представиться	форма
шкаф	наружный	преобразить	характерный
щель	носить	преобразиться	ходить

И наоборот, обратившись ко второй части, мы найдем ответ на вопрос, по каким понятиям распределяется, в смысловые группы каких дескрипторов входит или к каким семантическим полям относится слово, например, *башня*. Это такие понятия, или дескрипторы: КОЛОННА, АРКА, АРХИТЕКТУРА, БАШНЯ, ПАМЯТНИК, МАЯК. Аналогичным образом для слов:

*переговоры*: БЕСЕДА, ЯЗЫК, ПРОЩАНИЕ, ОБМЕН, ДОГОВОР, ПОСРЕДНИЧЕСТВО, ТЕМА, СЛОВО;

*полоскать*: ВОДА, ДЕЗИНФЕКЦИЯ, ОЧИЩАТЬ, КОСМЕТИКА, ЩЕЛОК;

*предрассудок*: РЕЛИГИЯ, ПРЕДРАССУДОК, МНЕНИЕ, УБЕЖДЕНИЕ, ПРОПОВЕДЬ, СУЕВЕРИЕ, ПРЕДУБЕЖДЕНИЕ, ПРИВЫЧКА, МАНЕРЫ;

*рифма*: МУЗЫКА, ПОЭЗИЯ;

*смущение*: СМУЩЕНИЕ, ЗАТРУДНЕНИЕ, СТЫД, СКРОМНОСТЬ, СОВЕСТЬ, ГРЕХ.

Дескрипторы в первой части и слова во второй части тезауруса расположены по алфавиту. Внутри тезаурусных статей также принят алфавитный порядок расположения.

Полученный словарь можно расценивать как «Основы русского тезауруса», поскольку, как считают авторы-составители, для охвата всей лексики литературного русского языка и ее представления в тезаурусном виде этот словарь уже не меняет свою принципиальную структуру, общий состав понятий, хотя отдельные из них могут варьироваться, а до-

пускает лишь экстенсивный рост своего лексического наполнения. В пользу такой его оценки говорят следующие соображения.

Обычный, стандартный тезаурус имеет, как правило, четыре входа, соответствующие четырем размерностям семантического представления лексики.

1)  $K \rightarrow K$ , вход «от концепта к концепту», от одного понятия к другому, воплощенный в схеме вертикально и горизонтально (т. е. иерархически и корреляционно) связанных друг с другом понятий.

2)  $K \rightarrow Z$ , вход, представленный дескрипторной частью тезауруса, отражающий переход «от концепта к знаку», ср. подзаголовки первой части рассматриваемого словаря — «от понятия к слову».

3)  $Z \rightarrow K$ , вход «от знака к концепту», соответствующий алфавитному указателю адресов слов в понятийных полях дескрипторов и представленный в нашем словаре второй его частью «от слова к понятию».

4)  $Z \rightarrow Z$ , вход «от знака к знаку», или пермутационный указатель, который имеет смысл для тезаурусов, включающих многословные единицы, т. е. словосочетания, и не нужен в нашем случае, так как в этом варианте тезауруса его единицы унитарны, однословны, что исключает возможность поисков в нем в направлении «от слова к слову».

Из этих четырех размерностей семантического пространства, в которых полный тезаурус представляет лексику языка или некоторого ограниченного подязыка, в нашем словаре наличествуют, таким образом, две —  $K \rightarrow Z$  и  $Z \rightarrow K$ , но размерности или входы, определяющие данный тип словаря.

Что касается первого из охарактеризованных выше входов в тезаурус, входа  $K \rightarrow K$ , т. е. общей схемы взаимосвязи понятий или, если угодно, тезаурусной картины мира, тезаурусных знаний о мире, то на первом этапе работы над словарем (этапе, который теперь уже можно считать завершенным), вопрос о ее построении — для упрощения и без того сложной задачи — сознательно исключался из рассмотрения. Во-первых потому, что его решение требует непереносимого выхода за рамки собственно языкознания и учета достижений не только философской науки, но, по сути дела, всего комплекса естественнотехнических и гуманитарных дисциплин<sup>2</sup>. Во-вторых же, по той причине, что современный уровень наших знаний об устройстве лексической семантики и о механизме формирования и функционирования лексикона человека не дает ясного представления ни о степени психолингвистической реальности подобных схем, ни о конкретном влиянии такой схемы на окончательный выбор слова в процессе речепроизводства или ее роли в речевосприятии<sup>3</sup>. Тем не менее сама логика работы над тезаурусом требовала от авторов учесть тем или иным способом и эту размерность лексико-семантического пространства, что было сделано с использованием синоптической схемы идео-

<sup>2</sup> Обзор различных возможностей устройства «тезаурусных знаний о мире» или разных подходов к построению «картины мира» на основе систематизации лексики см. в кн.: Карацлов Ю. Н. Общая и русская идеография. М., 1976, с. 246—259.

<sup>3</sup> В другой нашей работе (Карацлов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981, с. 197—218) показана роль концептуализации в процессах речепорождения и речевосприятия, но анализ ограничивается размерностями  $Z \rightarrow K$  и  $K \rightarrow Z$ , не выходя на уровень идеологического обобщения «знаний о мире».

логического словаря Х. Касареса. Таким образом, в структуре нашего тезауруса была как бы оставлена «пустая клетка», которая временно заполнялась соответствующей информацией из словаря Касареса. Это выразилось в приписывании каждому дескриптору номера той дескрипторной области, того семантического «квадрата», к которому данный дескриптор принадлежит в указанном словаре. Однако эта информация носила чисто условный характер, поскольку она никак не учитывалась при решении нашей основной задачи — распределения слов по дескрипторам.

Кроме того, синоптическая схема Касареса не может считаться удовлетворительной ни с точки зрения ее соответствия современному состоянию наук о природе, обществе и человеке, ни в плане ее мировоззренческого статуса, определяемого совсем иной чем наша иерархией ценностей. Исходя из необходимости соблюдения двух этих методологических требований, а также отводя системе отношений между концептами (= дескрипторами) совершенно определенную методическую (техническую) роль как средству коррекции полученного распределения слов по дескрипторам, авторы разработали свое представление «тезаурусных знаний о мире». В основе «картины мира» (см. с. 45) лежит классификация наук, базирующаяся на марксистско-ленинской методологии науки и охватывающая все современные отрасли знаний<sup>4</sup>.

Термин «научная картина мира» прочно вошел в современную литературу по методологии и истории науки, и соответствующее ему понятие занимает определенное место в структуре научного знания. Исходя из этого, В. Ф. Черноволенко определяет содержание этого понятия как «горизонт систематизации знания, где как раз происходит теоретический синтез результатов исследования конкретных наук со знанием мировоззренческого характера, представляющим собой целостное обобщение совокупного практического и познавательного опыта человечества»<sup>5</sup>.

Будем исходить из того, что понятийный аппарат, используемый для представления научной картины мира, может послужить основой для иерархизации дескрипторов на входе  $K \rightarrow K$  при построении тезауруса русского литературного языка. На пути решения этой задачи возникает ряд новых проблем, которые усложняют задачу построения этого входа тезауруса. Среди них необходимо отметить по крайней мере две. Во-первых, в настоящее время еще не разработана полностью общенаучная картина мира. Ее построение предлагает синтез научных картин мира, разработанных в отдельных областях знания: в физике, биологии, обществоведении, астрономии и т. д. Однако их обобщению в единую и целостную картину мира препятствует не только нерешенная проблема синтеза частных картин мира, но и неравномерность развития отдельных отраслей научного знания. Так, если в физике произошла уже смена трех картин мира — аристотелевской, механической, электродинамической, — а господствующей в настоящее время стала четвертая — квантово-полевая картина, то в биологии, например, имела место только одна смена: первая картина мира (дарвиновская) была в последнее время заменена молеку-

<sup>4</sup> Кедров Б. М. О современной классификации наук. (Основные тенденции в ее эволюции). — Вопросы философии, 1980, № 10.

<sup>5</sup> Черноволенко В. Ф. Мировоззрение и научное познание. Киев, 1970, с. 122.

лярно-генетической, в которой используются уже достижения физики, химии и других наук.

Однако взаимосвязь между такими картинами мира, как квантово-полевая и молекулярно-генетическая, можно проследить только по отдельным аспектам. Для создания же единой картины мира живой и неживой природы, микро- и макромира, биологического и социального и т. д. в науке нет пока достаточных оснований.

Если же использовать для нашей цели частные картины мира — физическую, биологическую, астрономическую и т. д., то при этом возникает вторая проблема — их сводимости и соотношения понятийного аппарата, используемого в общенаучной картине мира, с понятийным аппаратом, применяемым в отдельных теориях, теоретических схемах, моделях. Более того, в каждой частной картине сами эти компоненты, т. е. понятия соответствующей теории или модели, построены по иерархическому принципу. Так, в физической картине мира в верхнем ярусе иерархии понятий содержатся представления «о фундаментальных (элементарных) объектах, взаимодействие которых порождает все остальные, исследуемые физикой объекты и процессы; об общих особенностях такого взаимодействия (характер причинности и закономерности физических процессов); о пространственных и временных характеристиках физического мира»<sup>6</sup>. Далее, совершенно не изучена взаимосвязь не только между понятийным аппаратом частных научных картин мира и теориями, концентрирующимися в общую картину мира, но и между общезыковыми понятиями и понятийным аппаратом научной картины мира<sup>7</sup>.

Таким образом, в настоящее время уровень методологического знания недостаточно развит для того, чтобы результаты существующих классификаций могли быть непосредственно приложимы к задаче последовательной иерархизации дескрипторов на входе К — К в нашем тезаурусе. Существующие картины мира еще не являются такими феноменами, которые позволяли бы применять их без принципиальных теоретических реконструкций к исследованию лексики национального, в частности русского, языка. В связи с этим нами применительно к нашим целям осуществлен синтез современных знаний о структуре лексики и некоторых представлений о научной картине мира и предложена предварительная схема основных компонентов (составляющих) знания об окружающем нас мире, которая может, на наш взгляд, лечь в основу синоптической схемы русского тезауруса. При ее разработке мы исходили из учета двух взаимосвязанных функций, которые эта схема должна выполнять в ходе построения тезауруса и семантического анализа метаязыка словаря.

1. Данная схема является системой, в определенной мере отражающей основные фрагменты объективного мира, знания о котором не только зафиксированы в подязыке той или иной научной дисциплины, но и составили часть корпуса общенаучной терминологии литературного русского языка. Поэтому схема является необходимым компонентом при разработке семантических кодов для представления связей между дескрипторами в литературном языке. Прежде всего она нацелена на систематическое

Схема «тезаурусных знаний о мире»

1. Формы существования материи	Биологические формы существования материи	Микромир	1. Время 2. Пространство 3. Движение 4. Развитие 5. Свойства 6. Отношения 7. Энергия, сила 8. Состояние 9. Форма, строение	II. Физический микромир	1. Номенклатура 2. Свойства 3. Процессы, явления
			III. Биологический микромир	1. Номенклатура 2. Свойства 3. Процессы, явления	
			IV. Флора	1. Номенклатура 2. Общность 3. Структура 4. Свойства, форма строения 5. Процессы, явления	
			V. Фауна	1. Номенклатура 2. Свойства 3. Процессы, условия жизни	
			Человек	VI. Человек как живое существо	1. Питание 2. Анатомия 3. Физиология 4. Медицина
		VII. Человек как разумное существо		1. Потребности 2. Психология 3. Мышление 4. Сознание	
		VIII. Человек как общественное существо		1. Индивид 2. Общественные отношения 3. Семья 4. Народ, государство, типы расселения 5. Социальные институты	
		IX. Человек как существо социальное		1. Труд 2. Образование 3. Обычай 4. Право 5. Язык 6. Искусство, культура 7. Наука 8. Техника 9. Промышленность 10. Сельское хозяйство 11. Прочая деятельность 12. Условия жизнедеятельности (быт, жилье, одежда) 13. Религия, псевдонаука	
		Макромир	X. Земля как объект	1. Номенклатура 2. Процессы, явления 3. Строение	
	XI. Солнечная система		1. Номенклатура 2. Свойства 3. Процессы, явления		
	XII. Макромир		1. Номенклатура 2. Свойства 3. Процессы, явления		

<sup>6</sup> Степин В. С. Структура и эволюция теоретических знаний. — В кн.: Природа научного познания. Минск, 1979, с. 187.

<sup>7</sup> Дышлевый П. С. Естественнонаучная картина мира как форма синтеза знания. — В кн.: Синтез современного научного знания. М., 1973, с. 98.

выявление элементарных понятий этого семантического кода, в результате логического умножения которых можно образовать все остальные понятия. Кроме того, она позволяет определить суммарное число всех парадигматических отношений, служащих для выражения логических связей между данным понятием и семантическими множителями, на которые разложено это понятие<sup>8</sup>.

2. Схема является методическим средством, позволяющим осуществлять процедуру верификации и коррекции семантического кодирования по квазиосновам, которое производится на базе дефиниций толкового словаря.

Предлагаемая схема состоит из 12 блоков:

I. Формы существования материи (время, пространство, движение, развитие, свойства, отношения и т. д.).

II. Физический микромир.

III. Биологический микромир.

IV. Флора.

V. Фауна.

VI. Человек как живое существо.

VII. Человек как разумное существо.

VIII. Человек как общественное существо.

IX. Человек как существо созидющее.

X. Земля как объект.

XI. Солнечная система.

XII. Макромир.

В свою очередь эти блоки состоят из отдельных фасетов, которые представлены основными компонентами, процессами, свойствами и отношениями, специфическими для данного блока. Фасеты состоят из 55 областей, которые намечены с таким расчетом, чтобы каждое понятие одной области можно было определить только через понятия той же самой области. Насколько реально выдержать такое условие — должен показать опыт применения схемы в работе.

Приводимая схема не может расцениваться как готовый вход (K — K) в наш тезаурус. Данная система понятий отражает структуру такого входа только в первом приближении, потому что она не прошла испытаний на адекватность полученному распределению слов по дескрипторам. Не вполне ясным остается также способ распределения дескрипторов по областям этой схемы. Тем не менее наш экскурс в проблематику научной картины мира и ее редукции в схему тезаурусных знаний о мире был необходим не только для характеристики соответствующего входа (K — K) в тезаурус, но и для выяснения перспектив совершенствования другого входа (K — З), т. е. результатов уже полученного распределения слов по дескрипторам. Дело в том, что известная доля неточностей в этом распределении может быть объяснена именно отсутствием ориентации на схему иерархии дескрипторов. В соответствии с принятым принципом распределения все дескрипторы выступали как концепты одного уровня, одного ранга, тогда как в действительности им должна быть свойственна определенная субординация в зависимости от их принадлежности к тому или иному блоку, фасету, области. Иными словами, имеет

место принципиальное несоответствие между одноуровневостью распределения и разноуровневостью самих дескрипторов. Этими предварительными соображениями мы пока и вынуждены ограничиться в рассмотрении тезаурусного входа K — K.

Особенность построенного на ЭВМ словаря на понятийной основе в том, что он обладает добавочным входом, которого лишены обычные идеографические словари и терминологические тезаурусы. Речь идет о входе «от семантического множителя к слову» (СМ — З).

Чтобы пояснить существо этой размерности семантического пространства тезауруса, необходимо вернуться к тому, как он был сделан. Общая идея возникла 15 лет назад, и в то время самому автору она казалась довольно смелой и оригинальной, хотя теперь она, естественно, превратилась в общее место и звучит как банальность. Заключалась идея в том, что распределение слов по понятиям должно осуществляться на основе общности семантических элементов у того и другого, причем в качестве конструктивного аналога записи смысла слова (и понятия) набором семантических множителей была взята его лексикографическая дефиниция. Иначе говоря, перечень слов, из которых строится определение в толковом словаре, конструктивно приравнивался набору семантических множителей, задающих смысл определяемого слова. Тогда наличие одинаковых слов в толкованиях двух сопоставляемых единиц должно свидетельствовать и о наличии семантической связи между этими единицами. Для достижения механической сравнимости разных словоформ одного и того же слова, употребленных в различных дефинициях, была введена сокращенная запись состава дефиниции по квазиосновам. Квазиоснова — это набор букв слева от начала графического слова до некоторой буквы корня или суффикса, который позволяет однозначно идентифицировать группу словоформ и слов-derivатов, представляющих одну гиперлексеми. Квазиоснова — это своеобразный представитель, заместитель, если угодно, знак гиперлексеми. Например, квазиоснова *поощ* представляет группу слов: *поощрять, поощрить, поощрение, поощрительный*, составляющих одну гиперлексеми, а квазиоснова *пья* — *пьяный, пьяница, пьянствовать, пьянство* и т. д.

Затем дефиниции 1600 дескрипторов и 9000 слов переписываются по квазиосновам и вводятся в машину. Выглядят они при этом так (для дескриптора): БОГОСЛУЖЕНИЕ богослужение цер хрис религ обряд культ почит бог служ свят святы поклон икон ризн престол храм праз утва. Примерно то же для слов (в неполной записи): *крест*...симв хрис культ...; *молитва*...вер бог религ...; *монастырь*... религ цер...; *набожность*...религ обряд....

Сопоставление этих записей делает очевидным пересечение слов с дескриптором по одному или нескольким семантическим множителям, что делает возможным автоматическое распределение. В итоге, открыв тезаурус на с. 27, где дан дескриптор БОГОСЛУЖЕНИЕ, мы обнаружим под ним список слов, которые программа отобрала из исходного словника по этим самым признакам:

ангел	болеть	крест	обряд
атеизм	вера	молитва	отец
бог	верование	монастырь	панихида
богослужение	ересь	набожность	пасха

<sup>8</sup> Ср. анализ этих отношений и перечень их типов в кн.: Карацлов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка, с. 305—312.

плач	почитать	провидение	священник
поклонник	православие	пророк	служба
поклоняться	православный	пророческий	таинство
посаженный	праздник	просвира	теология
посуда	престол	протестанство	хоронить
похоронить	причащение	религия	храм
похороны	причитание	свадьба	церковь
почтитель	причт	святой	

Следовательно, на этапе распределения слов по дескрипторам (или знаков по концептам, если вернуться к схеме размерностей семантического пространства) квазиосновы функционируют как семантические множители.

Естественно, что для будущих более глубоких экскурсов в анализ русской лексической семантики было бы целесообразнее снабдить на этом входе (К — З) информацию о связи понятия с раскрывающими его содержание словами также и перечнем всех семантических множителей, послуживших основанием для отнесения каждого слова к соответствующему концепту. Получение такой информации из машинной базы данных по нашему тезаурусу не составит трудности, и ее отсутствие в полученном печатном формате объясняется в данном случае только стремлением несколько сократить объем первой части тезауруса. Эту информацию, как и данные по принципиально новому, неканоническому входу в тезаурус — «от семантического множителя к знаку» (СМ — З), можно при необходимости получать по запросу потребителей.

Для иллюстрации структуры дополнительного входа (СМ → З) приведем несколько примеров. Так, семантический множитель сторон содержат следующие слова:

анализ	обмазка	боковой	подкуп
признак	взаимный	справа	посредничество
остров	передний	слева	надпись
разветвление	четырёхугольник	лезвие	барабан
поверхность	цилиндрический	острие	встреча
ограничение	угол	направляться	
сердце	отклонение	прыгать	
подкладка	направление	заворачивать	

Множитель вод объединяет такую группу слов:

бассейн	берег	роса	мыло
замерзание	остров	туман	бетон
резервуар	поток	родник	мост
раствор	дождь	водоросль	плыть
течение	море	корневище	котел
гидравлика	горизонт	тростник	пловец
жидкость	залив	раки	судно
влажность	непроточный	рыба	пруд
вода	овраг	утка	руль
плотина	болото	каша	плот
канал	пещера	питье	трубопровод
лужа	водопад	грязь	

В ряду рассматриваемых отношений между тремя массивами данных — дескрипторами (К), словами (З) и семантическими множителями (СМ) — логически допустимым кажется постулировать отношение СМ → К и говорить о возможности извлечения из тезауруса информации о связи семантических множителей непосредственно и с понятиями, поскольку принципы кодирования, т. е. приписывания семантической информации, для дескрипторов (понятий) были теми же, что и для слов. Однако отношение СМ → К<sup>9</sup> оказывается принципиально невозможным теоретически и нереализуемым конструктивно в силу специфического устройства данного словаря. Теоретическая его невозможность объясняется тем, что члены этого отношения принадлежат не к одному и тому же и не к смежным, непосредственно контактирующим друг с другом уровням в семантической иерархии, но разделены еще одним уровнем, а именно, уровнем слов<sup>10</sup>, между тем как все рассмотренные выше размерности семантического пространства представляют собой отношения между элементами одного и того же либо смежных уровней. Конструктивно же переход «от семантического множителя к понятию» нельзя осуществить потому, что в этой размерности понятия просто неотличимы от слов: согласно принципам комплектования массивов исходных данных, элементы высшего уровня обязательно включались во множество элементов ближайшего низшего уровня, но не наоборот<sup>11</sup>. Таким образом, в число слов, т. е. в подлежащий распределению исходный словник, включены все дескрипторы нашего тезауруса, а в число семантических множителей по тому же принципу вошли все слова этого исходного словника. Значит, получив из машины информацию о том, например, какие конкретно слова характеризуются множителем жалос или вник, единственно, что мы можем утверждать на этом этапе с достоверностью, так это то, что два из пяти слов для первого множителя и один из четырех — для второго<sup>12</sup> являются одновременно и дескрипторами. Далее можно идентифицировать эти дескрипторы по их алфавитному списку, однако все эти процедуры не дают нам оснований для того, чтобы провести принципиальное различие семантических множителей дескрипторов и семантических множителей слов.

Если обратиться теперь ко второй части тезауруса «от слова к понятию», то в ней уже можно почерпнуть информацию не только об отношениях З — К, но и о тех множителях, которые послужили основой для отнесения данного слова к тому или иному дескриптору. Например: враждебный

НЕНАВИСТЬ	8 вра	3 ненав	
ВРАЖДА	8 вра	3 ненав	8 неприя
ПРОКЛЯТИЕ	3 ненав		

Это означает, что слово *враждебный* включено в семантические поля перечисленных дескрипторов на основании общности указанных семанти-

<sup>9</sup> Реверсивные отношения возможны для обеих пар, но оказываются тривиальными, поскольку оба они — и З → СМ, и К → СМ — представляют собой исходную информацию при построении данного тезауруса и восходят к дефинициям толковых словарей, на базе которых он строился.

<sup>10</sup> См.: Караулов Ю. Н. Общая и русская идеография, с. 187.

<sup>11</sup> Караулов Ю. Н. Частотный словарь семантических множителей русского языка. М., 1980, с. 11, 12.

<sup>12</sup> Караулов Ю. Н. Частотный словарь..., с. 105, 97.



ческих множителей (цифра перед каждым семантическим множителем фиксирует его частоту на массиве дескрипторов).

Аналогично для слов:

*пожениться*

ЖЕНСКИЙ	7 бра	20 жен	
СУПРУЖЕСТВО	7 бра	20 жен	10 муж
СЕКСУАЛЬНЫЙ	20 жен	10 муж	
ОПЛОДОТВОРЕНИЕ	20 жен	10 муж	
ХОЛОСТЯК	20 жен	10 муж	
СЕМЬЯ	20 жен	10 муж	
ВДОВЫЙ	20 жен	10 муж	

*пожилой*

СТАРСТЬ	2 немол	6 стар
ХОЛОСТЯК	2 немол	
ВОЗРАСТ	6 стар	
ДРЕВНИЙ	6 стар	
ПОВТОРЕНИЕ	6 стар	
ПЕНСИЯ	6 стар	

*поза*

ПРИТВОРСТВО	2 неиск	21 повед	6 притв
ИСКУССТВЕННОСТЬ	2 неиск	6 притв	
ХАНЖЕСТВО	21 повед	6 притв	
ПОЗА	1 поз	24 положен	46 телес
ДВИЖЕНИЕ	24 положен	46 телес	
ЛОЖЬ	6 притв		
МАСКИРОВКА	6 притв		
АКТЕР	6 притв		

*робкий*

ТРУСОСТЬ	4 боя		
РЕВНОСТЬ	4 боя		
РОБОСТЬ	4 боя	1 несмел	1 опасл 3 роб
ОПАСНЫЙ	4 боя		
ЗАЯЦ	3 роб		

Практически этот вход в словарь дает некоторый максимальный набор концептов, с помощью которого раскрывается смысл входного слова. Эти концепты могут служить ключевыми при построении лексикографической дефиниции данного слова в толковом словаре. Ср.:

*враждебный* 1. Крайне неприязненный, полный вражды, ненависти. 2. Вражеский, неприятельский.

## § 2. СЕМАНТИЧЕСКАЯ СВЯЗЬ И СОДЕРЖАТЕЛЬНЫЙ СМЫСЛ СЕЛЕКТИВНЫХ КРИТЕРИЕВ

Остановимся на характеристике той важной роли, которую сыграло понятие семантического множителя в этой работе. Оказалось, что безусловное применение начального принципа «если есть общий семантический множитель у двух слов, то между ними существует семантическая связь» — не работает, не дает разумного результата. Для первичного распределения, сделанного только на основе этого принципа, был ха-

рактерен сильный «шум», мы получали очень нечистый результат, где надежно согласующиеся с интуитивными представлениями, т. е. безусловно верные с точки зрения среднего носителя языка, связи занимали всего 15—20% в полученных списках. Иными словами, мы получали какие-то безбрежные по составу группы, в которых более или менее обоснованные и максимально ожидаемые, с точки зрения экспертов, семантические связи просто терялись на фоне громадных списков единиц, отнесенных программой к каждому дескриптору (К — З), или наоборот, обширных списков дескрипторов, отнесенных к каждому слову (З — К). Приведем в качестве примера начало нашего словника — образцы статей для входа З — К, полученных при первичном распределении (курсивом напечатано слово, прописными буквами — дескрипторы, в состав которых должно входить это слово, справа от дескриптора указаны общие для него и входного слова, т. е. совпадающие, семантические множители; если данное слово является одновременно и дескриптором, как, например, в случае *абсолютный*, то число таких совпадений, естественно, велико; минимально необходимым в соответствии с феноменологической моделью является совпадение в одном семантическом множителе).

*абсолютный*

АБСОЛЮТНЫЙ	абсол	КОЛЕСО	дви
	безотно		маш
	безусл		механ
	внесрав	ЗВУК	дви
	независ	СИЛА	дви
	неотно		механ
	несрав	ЭЛЕКТРИЧЕСТВО	дви
	полн	ШУМ	дви
	соверше	ТЕПЛОТА	дви
СЛАВА	безусл	ТЕЧЕНИЕ	дви
ОБЯЗАТЕЛЬНЫЙ	безусл	ГИДРАВЛИКА	дви
ПРОКЛЯТИЕ	безусл	КОЛЕБАТЕЛЬНЫЙ	дви
	полн	ФИЗИКА	дви
ОСОБЕННОСТЬ	независ		механ
СУДЬБА	независ	ПОТОК	дви
НЕЗАВИСИМЫЙ	независ	ВЕТЕР	дви
СВОБОДА	независ	ТУЧА	дви
ПОГРУЖЕНИЕ	полн	ЖИВОТНОЕ	дви
НАПОЛНЯТЬ	полн	КРЫЛО	дви
ИЗОБИЛИЕ	полн	НЕРВ	дви
ТОЧНОСТЬ	полн	ПУЛЬС	дви
СОВЕРШЕНСТВО	полн	ЖИВОЕ	дви
	соверше	НАПРАВЛЕНИЕ	дви
ЗАВЕРШЕННОСТЬ	полн	ЛИНИЯ	дви
	соверше	БЕЖАТЬ	дви
РОСТ	соверше	ВРАЩЕНИЕ	дви
<i>авария</i>		ДВИЖЕНИЕ	дви
ОСЬ	дви		механ
	маш		работ
	механ		

ДРОЖЬ	дви	СЛЕСАРНАЯ	работ	КРЫЛО	аппа	ДОРОГА	передви
КОЛЕБАНИЕ	дви	ОРУДИЯ	работ		возд	ПЛЫТЬ	передви
НАПРАВЛЯТЬСЯ	дви	МОЛОТОК	работ		лет	ПУТЕШЕСТВИЕ	передви
ОСТАНОВКА	дви	ЩЕБЕНЬ	работ	ФОТОГРАФИЯ	аппа	СКОЛЬЗИТЬ	передви
ОПЕРЕЖАТЬ	дви	ИЗОБРЕТЕНИЕ	работ	ВЕЛОСИПЕД	аппа	БРОДЯЖНИЧАТЬ	передви
ОТСТАВАТЬ	дви	ДЕЙСТВИЕ	работ	САМОЛЕТ	аппа	ПЛОВЕЦ	передви
ПЛЫТЬ	дви	НЕБРЕЖНОСТЬ	работ		возд	БЕГ	передви
ПОДЪЕМ	дви	УСПЕХ	работ		лет	ШАШКИ	передви
ПРЫГАТЬ	дви	РАБОТАТЬ	работ	ДЫМ	возд	СУДНО	передви
СКОЛЬЗИТЬ	дви	ЗАНЯТИЕ	работ		лет		средс
ТАНЦЕВАТЬ	дви	СЛУЖБА	работ	ЗВУК	возд	ЖЕЛЕЗНОДОРОЖНЫЙ	передви
ТОЛЧОК	дви	ПОМОЩЬ	работ	НАСОС	возд	ГИДРАВЛИКА	прак
ТРЕНИЕ	дви	ПРОИЗВОДСТВО	работ	ВОЗДУХ	возд	ГИГИЕНА	прак
ШАГАТЬ	дви	ПОДГОТОВКА	работ	ХОЛОД	возд	ОПЫТНОСТЬ	прак
ЧАСТОТА	дви	ИНСТРУМЕНТ	работ	ПРОЗРАЧНОСТЬ	возд	ПРОБЛЕМА	прак
РИТМ	дви	ПРОФСОЮЗ	работ	АТМОСФЕРА	возд		теор
ЖЕСТ	дви	СОРЕВНОВАНИЕ	работ	БЕЗВЕТРИЕ	возд	ИСТИНА	прак
ПРЕПЯТСТВИЕ	дви	ЗАРАБОТОК	работ	ВЕТЕР	возд	ПРОЗА	прак
ИСТОРИЯ	дви	УЧРЕЖДЕНИЕ	работ	РОСА	возд	БУХГАЛТЕРИЯ	прак
ТАНЦЕВАТЬ	дви	август		ТУМАН	возд		теор
КИНО	дви	ГИТАРА	восъм	ГЛАЗУРЬ	возд	ВЕТЕРИНАРИЯ	прак
БЕГ	дви	ЮНОСТЬ	год	ЛИСТВА	возд	КОСМЕТИКА	средс
АВТОМОБИЛЬ	маш	ДЕТСТВО	год	БРОНХИ	возд	ОРУДИЯ	средс
	механ	ВЕСНА	год	ДЫХАНИЕ	возд	ДЕЗИНФЕКЦИЯ	средс
		ВРЕМЯ	год		возд	ЛЕЧИТЬ	средс
ОБМЕН	дви	ГОД	год	ОКНО	возд	ВОЗМОЖНОСТЬ	средс
ПАРУС	дви		календ	ПОДУШКА	возд	ОБОРОНА	средс
ТРОЛЛЕЙБУС	маш	ЗИМА	год	БАНЯ	возд	ПРОИЗВОДСТВО	средс
ЖЕЛЕЗНОДОРОЖНЫЙ	дви	КАЛЕНДАРЬ	год	СВИСТ	лет	ЯЗЫК	средс
ГРЕСТИ	дви		календ	ЛЕТЕТЬ	передви	ИНСТРУМЕНТ	средс
НАСОС	маш		меся		возд	БУРЖУАЗНЫЙ	средс
	механ	ЛЕТО	год	ПРОГУЛКА	возд	ФЕОДАЛЬНЫЙ	средс
СБОРКА	маш	МЕСЯЦ	год	ПРЫГАТЬ	возд	АВТОМОБИЛЬ	средс
РАЗБОРКА	маш		меся	ГУЛЯНИЕ	возд	ВЫГОДА	средс
	механ	ОСЕНЬ	год	РУЛЬ	возд	ВКЛАД	средс
КОВЕР	маш	ДАТА	календ	ТРУБОПРОВОД	лет	ДЕНЬГИ	средс
	работ		меся	МУХА	лет	НИЩЕНСТВО	средс
ВЕЛОСИПЕД	маш	СРОК	календ	ПЧЕЛА	лет	ВООРУЖЕНИЕ	средс
	механ	ЛУНА	меся	БРОСАТЬ	лет	КООПЕРАЦИЯ	средс
ЖАТВА	маш	МЕНСТРУАЦИЯ	меся	ПАДЕНИЕ	лет	ПОЧТА	средс
РАВНОВЕСИЕ	механ	ДЕНЬ	меся	РАКЕТНЫЙ	лет	ПОВОЗКА	средс
СМЕСЬ	механ	авиация		КОЛЕСО	передви	ТРАНСПОРТ	средс
ЧАСЫ	механ	АВИАЦИЯ	авиа	ЭКРАНИРОВАТЬ	передви		флот
ЧАСТЬ	механ		аппа	ЖИВОТНОЕ	передви	ФЛОТ	флот
РУЛЬ	механ		возд	ПРЕСМЫКАЮЩИЕСЯ	передви	КОРАБЛЬ	флот
РАНА	повре		лет	ЧЕРЕПАХА	передви		
УВЕЧЬЕ	повре		передви	БЕЖАТЬ	передви	автобус	
ЦАРАПАТЬ	повре		средс	БРОДИТЬ	передви	ОСТАНОВКА	автоб
РАЗБИВАТЬ	повре		флот				пассажир

ТРОЛЛЕЙБУС	автоб	АВТОБУС	автомо
	многомес		многомес
АВТОМОБИЛЬ	автомо		пассажир
	пассажир	ТАКСИ	автомо
РУЛЬ	автомо	ТРАНСПОРТ	пассажир

Полученные распределения для первых 200 единиц словника и их анализ сделали очевидным тот факт, что существо семантической связи многообразнее и сложнее, чем это позволяет представить принятый принцип распределения. Модель, конструктивно воплощающая понимание семантической связи на поверхностном, так сказать, уровне, на уровне явлений (и соответственно названная в других наших работах феноменологической), как жесткой и однозначно определяемой зависимости между двумя единицами, не отражала, очевидно, всей сложности самого феномена, требовала определенной коррекции и дальнейшей аппроксимации, адаптации к нашей задаче.

Вместе с тем у нас не было и оснований безоговорочно утверждать, что полученные в соответствии с феноменологической моделью распределения или группы, отвечающие, по мнению экспертов, канонам тезаурусных статей лишь на 15—20%, в остальном бессмысленны. Нет, между входящими в них единицами и заглавным словом (для входа 3 — К) или дескриптором (для другого входа К→3) практически во всех случаях констатируются те или иные связи, но связи эти характеризуются большей или меньшей силой. Об этом, в частности, свидетельствуют результаты проведенных нами ассоциативных экспериментов, целью которых было выявить некоторое безусловно принимаемое всеми носителями языка ядро тезаурусной статьи, наиболее тесно связанное с заглавным словом. Забегая вперед, скажем сразу, что эксперименты указанной цели не достигли: во всех трех группах испытуемых при сложении их результатов отмеченными оказались все дескрипторы (образцы статей, с которыми работали информанты, см. выше), отнесенные машиной к каждому из двухсот слов словника, фигурировавших в эксперименте, только одна часть из них получила большее, а другая — меньшее количество «отметок».

**Ассоциативный эксперимент.** Материал эксперимента составили 200 слов, под каждым из которых перечислены все дескрипторы, обнаружившие с заглавным словом хотя бы один общий семантический множитель (вход 3 — К), аналогично образцам, приведенным на с. 51. Испытуемыми были три группы людей: I — студенты-лингвисты 5-го курса, полностью информированные о принципах автоматического построения тезауруса, о возникших при этом проблемах и о целях проводимого ассоциативного эксперимента. Эти информанты были знакомы также с существующими в современной науке представлениями о существе и типах семантических связей в лексике, о взаимоотношениях знака и концепта, или слова и дескриптора (14 человек); II — филологи, работающие в разных областях литературоведения или языкознания, но не являющиеся специалистами по лексикологии или лексикографии, которым были сообщены только условия данного эксперимента (7 человек); III — нефилологи с высшим математическим или философским образованием, специалисты по автоматической обработке информации, осведомленные о принципах данного тезауруса и задачах эксперимента (4 человека).

Условия эксперимента. Испытуемому предъявлялось не число распределений и ставилась задача отметить те из дескрипторов (концептов), которые по его представлениям каким-то образом связаны по смыслу с заглавным словом.

Результаты эксперимента. Поразительным и, может быть, несколько обескураживающим оказалось то единодушие, которое проявили представители разных групп носителей русского языка в выполнении поставленной задачи. Во-первых, каждый из них стремился к тому, чтобы отметить максимум дескрипторов, т. е. обнаружить семантическую зависимость между словом и концептом во что бы то ни стало. Отсюда всем спискам оказалось свойственно очень широкое понимание семантической связи. Во-вторых, при наложении одних и тех же распределений с отметками разных информантов друг на друга, т. е. при их суммировании, оказалось, что не отмеченных концептов вообще нет, причем разброс «отметок» был такой, который не позволял, как планировалось вначале, очертить определенное ядро в той или иной группе дескрипторов, поскольку единичных связей было зафиксировано очень мало и они не совпадали друг с другом.

Интерпретация результатов. Условия нашего эксперимента отличались от обычных в ассоциативном эксперименте и как бы заранее создавали благоприятные условия для положительного ответа на поставленный вопрос. В самом деле, если стоит задача — «найти» нечто, то человек, проявляя зачастую всю изобретательность, на какую он способен, обязательно «находит»<sup>13</sup>, тем более, что на информанта своеобразное гипнотическое действие оказывает сама машинная распечатка распределений, создавая в некотором роде «культ спецов» — «уж они-то знали, что делали...». И все же этот суммарный итог — выявление семантической связи в каждой сопоставляемой паре слово—понятие — нельзя объяснить исключительно субъективными предпосылками в организации нашего эксперимента, причина лежит глубже — в объективных условиях существования в языке семантических связей и проявления у носителей языка механизма семантических взаимодействий. Что касается этого механизма, то при всей его общечеловеческой универсальности на конкретных способах его проявления всегда лежит печать определенной человеческой личности. Это особенно отчетливо видно в свободном ассоциативном эксперименте, когда легкость ассоциирования испытуемого в том или ином направлении выявляет доминирующий у него либо кустовой, либо цепной тип ассоциаций<sup>14</sup>. В нашем же эксперименте, когда даны две единицы и испытуемый должен открыть основание их

<sup>13</sup> Сама формулировка задания испытуемым могла варьироваться и звучать, например, так: «В соответствии с программой машина выбрала из исходного списка слова, которые по некоторым формальным критериям считаются связанными с заглавным словом. Однако машина не обладает пониманием, подобным человеческому, программа несовершенна, а формальные критерии не вполне адекватны реальным смысловым зависимостям. Вам предстоит исправить машинный результат». Но даже при этой формулировке информанты видели свою задачу в том, чтобы раскрыть, обнаружить, сделать явной скрытую семантическую связь. Т. е. субъективное ощущение их было таким, что связь есть всегда, надо ее лишь выявить и сформулировать.

<sup>14</sup> Levin I. Creativity and two modes of associative fluency: chains and stars. — Journal of Personality, Dirham, 1978, vol. 46, N 3.

общности, сами условия оказываются сдерживающим фактором для предпочтительного проявления только ассоциативной доминанты данной личности и вводят в действие, актуализируют все другие, обычно не используемые ею возможности ассоциирования, заставляя испытуемого комбинировать разные способы семантического согласования в поисках этой общности. Таким образом, спектр выбора оснований для семантической связи двух единиц существенно расширяется.

Надо сказать, что с подобной ситуацией не раз сталкивались и другие исследователи. Дж. Энглин, например, развивая предикативную гипотезу для объяснения закономерностей ассоциативной памяти, в серии экспериментов (сортировка, свободное воспроизведение, ассоциативный тест) показал, что признаки, положенные испытуемыми в основу ассоциативной общности в той или иной паре слов, могут быть настолько дробными и малозначительными или наоборот чрезвычайно обобщенными, что число их тем самым потенциально увеличивается до бесконечности, и ассоциативная связь почти всегда оказывается возможной. Так, слова *мальчики* и *лошади* в его экспериментах объединялись испытуемыми на основе свойственных им одних и тех же предикатов — «едят, ходят, имеют ноги, являются теплокровными», т. е. потому, что обозначают они определенные виды из класса «животные»<sup>15</sup>.

Другая объективная причина получения в нашем эксперименте охарактеризованного выше результата заключается в фундаментальной неопределенности самого понятия «семантическая связь». Например, в психолингвистике устойчивым является представление, что семантическая связь имеет единственную реальность в языке — это зафиксированные в экспериментах ассоциации. На совершенно иных позициях стоят лексикологи, которые видят семантическую связь в общности сем, тогда как грамматисты усматривают ее в возможности парадигматических преобразований данной единицы или в условиях ее сочетаемости с другими на синтагматической оси. Если теперь вспомнить эксперименты В. С. Старинца, в которых было показано, что при направленном ассоциировании для установления цепочки связей между двумя произвольно выбранными словами в экстремальном случае достаточно шести переходов, шести ассоциативных преобразований<sup>16</sup>, а также выведенное нами правило — «правило шести шагов», согласно которому между двумя любыми лексемами в словаре можно установить связь по общности элементов, используемых в их дефинициях, с помощью не более чем 5—6 промежуточных слов и их определений<sup>17</sup>, то утверждение, что «все в словаре связано со всем», уже не покажется абсурдным и лишенным всяких конструктивных следствий. Таким образом, главный итог проведенного нами эксперимента состоит в том, что он заставляет переосмыслить само существо понятия «семантическая связь» и, с учетом бытующих в науке разных представлений о нем, сделать следующее заключение.

Семантическая связь — понятие неоднородное и не абсолютное.

<sup>15</sup> Anglin J. M. The growth of word meaning. Cambridge, Mass., The MIT Press, 1970.

<sup>16</sup> Старинец В. С., Агабабян К. Г., Недялкова Г. И. Экспериментальные исследования семантической организации ассоциативных сетей. — В кн.: Моделирование в биологии и медицине, вып. 3. Киев, 1968.

<sup>17</sup> Караулов Ю. Н. Общая и русская идеография, с. 77—79.

Она обладает многоступенчатой или спектральной структурой и характеризуется изменяющейся интенсивностью, а связанные между собой единицы различаются как интенсивностью, так и количеством своих связей.

Положение о том, что в лексике нет ни одной единицы, не имеющей связей с другими, семантически изолированной, представляется интуитивно ясным. Но обратное утверждение, которое вытекает из размышлений над результатами проведенного нами эксперимента, а именно, что семантическая связь между двумя произвольно выбранными единицами существует всегда, но только может быть при этом более сильной либо более слабой, приближающейся в своем значении к нулю, такое утверждение не кажется само собой разумеющимся и требует доказательств. По аналогии с явлением всемирного тяготения представим себе силы семантического притяжения в виде повсеместно существующего, разлитого в языке поля, в которое помещены тела — лексические единицы языка. Разные единицы в этом поле взаимодействуют между собой так же, как атомы, молекулы, макротела, планеты и космические объекты — и на одном уровне, т. е. с однородными единицами, и межуровнево. Эти взаимодействия могут быть как сильными (например, внутриядерные), так и слабыми. Ассоциативная связь — это в общем случае пример слабого взаимодействия, тогда как общность сем есть показатель, как правило, сильного семантического взаимодействия. Оговорки, использованные при характеристике этих разновидностей семантических связей — «в общем случае» и «как правило», — продиктованы известной неоднозначностью предлагаемых критериев определения типа (сильное или слабое) семантического взаимодействия. Дело в том, что в число реакций-ассоциатов попадают и такие, которые могут иметь одинаковые семы со стимулом. И наоборот, совпадение далеко не всяких сем у двух слов оказывается достаточным для установления между ними сильного взаимодействия. Так, обратившись к «Ассоциативному словарю белорусского языка», мы увидим на стимул БЕЛЫ<sup>18</sup> следующие реакции (единичные ответы опускаются): 417 снег, 69 хлеб, 58 черны, 21 цукар, 19 мядзведзь, 17 твар, 15 дом, 13 колер, 11 дзень. Среди них только снег, черны, и колер имеют общие семы с белы, в то время как остальные, в том числе около ста единичных ассоциаций, таких сем лишены независимо от того, опираются ли они на парадигматические или синтагматические отношения в паре стимул—реакция. Аналогично для стимула БЛАГІ — 208 чалавек, 68 дрэнны, 61 учынак, 48 намер, 42 добры, 37 сябар, 34 дзень, 30 настрой, 28 мат, хлопец, 18 выпадак, 16 харошы, 14 таварыш, 12 характар, 10 кепскі..., где общие семы обнаруживают дрэнны, добры, харошы и кепскі, образующие с ним две антонимические пары. С другой стороны, ходовые примеры типовых ассоциаций, которыми часто оперируют психолингвисты, такие, скажем, как мрачный и ночь, не позволяют установить общие семы у ассоциатов даже с третьим, промежуточным словом, связь через которое согласно теории медиации должна объяснять само наличие такой ассоциации. Ср. цепочку определений, в которой еще не появляется общий элемент: мрачный — темный, погруженный во мрак; ночь — часть суток от вечера до утра; темный — лишенный света, погруженный во

<sup>18</sup> Цітова А. І. Асыцыятыўны слоўнік беларускай мовы. Мінск, 1981, с. 22.

тьму; *мрак* — отсутствие света, тьма; *сутки* — продолжительность времени, равная 24 часам; *вечер* — часть суток перед наступлением ночи, следующая после окончания дня, и т. д. До установления одинаковых сем в этой паре нужно сделать не менее 5—6 шагов в глубину по дефинициям словаря, а ассоциативная связь, т. е. слабое взаимодействие, между ними тем не менее существует, что подтверждается множеством экспериментов.

Для иллюстрации противоположного нетипичного случая, когда наличие показателя сильного взаимодействия, или «равносемность», так сказать, двух слов не подтверждается другими свидетельствами существования между ними не только сильного, но как будто и слабого, ассоциативного взаимодействия, можно обратиться к примерам на с. 48, ср. *анализ* и *остров*, *разветвление* и *подкладка* и т. п. пары, которые объединяются множителем сторон. Тем не менее подобные крайние случаи не должны разрушать представления о наличии сильного и слабого взаимодействия и их формальных показателей — равносемности, с одной стороны, и ассоциативной повторяемости — с другой. Прочие способы выявления семантических связей с конструктивной точки зрения обладают меньшей эффективностью и могут расцениваться как варианты двух охарактеризованных приемов. Например, возможность преобразования (одношагового), или, в иной терминологии, лексическая функция, полностью покрывается критерием равносемности: *купить* + *Сопв.* → *продать*; *реакционер* + *Магп.* → *махровый*; *грустный* + *Апг.* → *веселый* и т. п. И в этом смысле лексическая функция есть показатель сильного семантического взаимодействия. Другой используемый иногда показатель связи — возможность установить и сформулировать некоторое типовое отношение для данной пары слов — по сфере его приложимости оказывается значительно шире, и потому он соизмерим, на наш взгляд, с критерием ассоциативной общности, т. е. выступает как необходимое условие слабого семантического взаимодействия: выше, в примере из Дж. Энглина — отношение в паре *мальчики* — *лошади* «тепловкровные, относящиеся к классу животные». Набор таких отношений в лексике, по-видимому, конечен, хотя и не описан полностью, и установить отношение, обладая известной фантазией, воображением, практически можно всегда. Так, когда при обсуждении содержания статьи ЗАПАХ нашего тезауруса было высказано недоумение по поводу того, на каком основании включено в статью слово *гвоздика*, участвовавший в обсуждении Е. Л. Гинзбург предложил остроумное и вполне правдоподобное объяснение, обозначив отношение между дескриптором и словом в этом случае как «свойство и эталонный носитель свойства». Естественно возникающие при этом вопросы типа «а почему не ландыш или лавровый лист?» показывают только, что интенсивность семантического взаимодействия в этой паре невысока, но в том, что она имеет место, никто не усомнился.

Итак, говоря о семантической связи, мы не имеем теперь права утверждать в каждом отдельном случае только, что связь есть или ее нет. Это слишком огрубленное, утрированное представление. Теоретически связь есть всегда, между любыми двумя лексическими единицами, но при этом она «есть более» либо «есть менее», в большей степени (непосредственная связь, или сильное семантическое взаимодействие) либо в меньшей (опосредованная связь, или слабое семантическое взаимо-

действие). Такое постепенное изменение ее интенсивности в общем поле семантического тяготения в лексике (словаре) говорит о допустимости ее измерения с использованием вероятностных принципов. Можно полагать, что существует некий тезаурусный порог силы семантической связи: при уменьшении ее ниже определенной величины словарная статья тезауруса распадается, и наоборот, по достижении ею некоторого рубежа лексические единицы, характеризующиеся данным или более высоким значением показателя силы семантической связи с соответствующим дескриптором, «притягиваются» к нему, начинают включаться в его статью (поле).

Значит, при построении тезауруса задача состоит в том, чтобы найти эту пороговую величину и научиться определять силу семантической связи между двумя единицами для соотнесения ее с пороговой.

Теперь можно возвратиться к осуждению результата первичного распределения слов по дескрипторам, полученного по принципу равносемности и охарактеризованного в начале настоящего параграфа. С учетом наших рассуждений о существе и типах семантических связей становится понятным, почему из наличия одинаковых семантических множителей (сем) еще не следует автоматически, что слова-носители этих сем «притягиваются» друг к другу с достаточной для фиксации их в тезаурусе силой.

Вероятный характер распределения семантических сил, которыми дескриптор удерживает вокруг себя семантическое поле, усиливается в нашем случае еще и из-за того, что есть известная приблизительность, статистически оправданная усредненность в приравнении, в конструктивной замене семы, семантического множителя — квазиосновой, образованной от полнозначного слова из дефиниции испытываемой на связность лексемы.

Таким образом, онтологическая вероятность того, достаточна или недостаточна сила семантической связи, складывается каждый раз с методически обусловленной вероятностью того, насколько справедливым является для данного конкретного случая усредненное представление семантического множителя. Следовательно, селективный критерий, который позволял бы из списков, полученных на основе недифференцированной равносемности формировать окончательную статью тезауруса, должен отражать вероятность достижения тезаурусного порога силой семантического взаимодействия и учитывать одновременно вероятностный характер совпадения идеальной семы с реальным множителем, представленным квазиосновой.

Поиск такого критерия — это поиск, естественно, случайный, использующий метод проб и ошибок. Было проведено большое число испытаний, осуществлена объемная работа с информантами и экспертами, опробовано множество разнообразных методик. Работа велась в трех направлениях, различающихся тем, что за основу брались разные характеристики семантических множителей. Во-первых, это количество совпадающих у слова и дескриптора множителей. Интуитивно кажется ясным, что чем больше общих элементарных множителей, тем сильнее семантическое притяжение в паре сопоставляемых слов. Вопрос можно поставить так: сколько совпадений необходимо и достаточно для достижения силой

семантической связи величины тезаурусного порога <sup>19</sup> — больше одного ( $>1$ ), или больше ли двух ( $>2$ ), или больше трех ( $>3$ ) и т. д.? Это направление поисков довольно быстро завело в тупик. Уже чисто формальный показатель работы этого критерия оказался неудовлетворительным: верхней границей числа одинаковых множителей может быть только «три», поскольку большее число совпадений дескриптор и слово обнаруживают лишь при их полной идентичности, т. е. в случае, когда слово само одновременно является и дескриптором, например *химия* и ХИМИЯ, *авиация* и АВИАЦИЯ и т. п. Однако совпадения в трех множителях оказываются не часты: на 200 первых по алфавиту слов их зафиксировано менее 80, а это означает, что больше половины единиц исходного словника не найдет при распределении своих дескрипторов, останется нераспределенным. Чтобы не приводить громоздких иллюстраций для подтверждения этого положения, отсылаем читателя к образцам первичного равносемного распределения, приведенным на с. 51—54, которые отражают принципиальную картину соотношения одного, двух и трех совпадений по множителям на всем словнике. При переходе от формальной оценки эффективности селективного критерия, основанного на трех совпадениях множителей сравниваемых слов, к его содержательной интерпретации мы также убеждаемся в его непригодности, поскольку далеко не во всех таких парах, по мнению экспертов, сила связи оказывается достаточной для фиксации ее в тезаурусе.

Если снизить минимально необходимое число общих семантических множителей и принять пороговую величину большей или равной двум ( $\geq 2$ ), то формальная сторона получаемого таким образом распределения не вызовет как будто недоумений: число слов, которые не найдут своего дескриптора, существенно уменьшится; количество дескрипторов, по которым будет распределяться каждое слово, колеблется при этом от 2 до 13, а в среднем окажется равным 6 или 7, что соответствует представлению об оптимальной структуре словарной дефиниции, необходимом и достаточном количестве ключевых понятий (концептов) в ней; и наоборот, объем тезаурусной статьи, т. е. число слов, относящихся к одному дескриптору, получится близким к стандартному и составит, при том небольшом исходном словнике, который мы имеем, 25—40 единиц. Но эта в целом удовлетворительная картина по внешним параметрам распределения теряет свою стройность, как только мы подвергнем результат содержательному анализу. По оценкам экспертов в парах, обладающих двумя одинаковыми множителями, семантическая связь достигает достаточной силы только в 50% проанализированных случаев. Иными словами, мы опять получаем значительный информационный шум на выходе. С другой стороны, определенная доля информации оказывается при таком критерии и потерянной, из-за того что из рассмотрения исключаются все слова с единичными совпадениями по множителям. Между тем, по мнению

<sup>19</sup> Определение этой величины мы можем проводить только качественно, пользуясь методом экспертных оценок. Когда все эксперты сходятся в том, что в данной паре, отношения между партнерами которой отвечают принятому критерию, семантическая связь является надежной и достаточной для включения слова в тезаурусную статью, испытываемое слово считается относящимся к данному дескриптору. При расхождении в оценках экспертов, несмотря на формальное соответствие некоторому критерию, слово в статью не включается.

экспертов и по интуитивным оценкам самого автора, семантические связи в парах, например, *анатомия* и *печень*, *анатомия* и *железа* сильнее, чем в паре *анатомия* и *ботаника*, хотя первые характеризуются одним общим множителем, а вторая — двумя. Аналогичны отношения в парах *авторитет* и *власть*, *авторитет* и *превосходство* vs *авторитет* и *дискредитация*; *агент* и *разведка* vs *агент* и *занятие*, *агент* и *служба* и т. п., где первыми названы пары с одним совпадающим множителем, но более интенсивным семантическим взаимодействием, а вторыми — пары с двумя множителями и более слабым взаимным притяжением. Таким образом, рассмотрение возможностей поисков критерия в первом направлении этим исчерпывается, и главным итогом предпринятых попыток было то, что произошло разрушение некоторой исходной «очевидности», того подожжения, которое вначале представлялось интуитивно бесспорным и справедливым, а именно, о наличии прямо пропорциональной зависимости между числом совпадающих множителей и силой семантической связи между словами — их носителями. Оказалось, что исходный тезис «чем больше общих элементарных множителей, тем сильнее семантическое взаимодействие в паре сопоставляемых слов» обладает лишь относительной истинностью: он может иногда оказаться верным, а иногда — нет. Это заставило нас обратиться к поискам селективного критерия в другом направлении.

Второй путь эксплуатирует статистическую характеристику семантических множителей, т. е. основывается на предположении, что частотность совпадающего множителя может служить показателем его значимости, его «веса» при тезаурусном распределении, позволяя дифференцировать и ранжировать все множители в зависимости от того, на сильное или слабое семантическое взаимодействие указывает каждый из них. Диапазон варьирования частот для обнаружения тезаурусного порога релевантности множителя был уже гораздо шире, чем в случае варьирования только числа множителей: от 2—3 появлений множителя на массиве дескрипторов до частот 12—15. Оптимум, т. е. наибольшая вероятность того, что множитель с данной частотой указывает на сильное семантическое взаимодействие, был нащупан в районе частотности, равной 5—7. Причем оказалось, что если принять в качестве пороговой частоты множителя равной 7, то «шум», или число неточных распределений, становится уже довольно значительным. Если же остановиться на 5, то мы терьем заметное количество слов — правильных, точных распределений. Поэтому и выбран был в качестве критерия показатель частоты множителя на массиве дескрипторов  $\leq 6$  «меньше или равен шести» <sup>20</sup>

Обращение к статистическим характеристикам семантических множителей позволило достичь большего приближения к идеально ожидаемой структуре тезаурусной статьи, чем это удавалось сделать на основе числа совпадений, и тем более с помощью одного лишь принципа равносемности. Однако нельзя сказать, что принятие уточненного критерия, когда необходимым и достаточным показателем сильного взаимодействия является

<sup>20</sup> Подробно методика поисков пороговой частоты и иллюстрация результатов распределения при различных величинах частоты совпадающего семантического множителя здесь не приводятся, так как они изложены в другой работе: Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка, с. 259—270.



совпадение хотя бы одного множителя с частотой  $\leq 6$ , дает вполне удовлетворительный результат. Два основных недостатка получаемого распределения — информационный шум, с одной стороны, и информационные потери, с другой, хотя и в значительно сокращенных масштабах, но продолжают оставаться и при использовании последнего критерия. В связи с этим было опробовано третье направление поисков и уточнения селективного критерия, составившее третий этап приближения модели распределения к некоему идеальному представлению.

Цель третьего этапа, или третьего направления поисков, — дальнейшее совершенствование селективного критерия. Из двух зол — информационного шума и потерь — было выбрано меньшее: представлялось целесообразным примириться с получающимся на втором этапе уровнем шума в расчете на то, что его можно будет устранить при постмашинном редактировании статей словаря. И тогда содержанием поисковой работы стало комбинирование уже рассмотренных критериев, направленное теперь только на снижение информационных потерь при распределении слов по дескрипторам. Ясно, что этого можно было достичь, расширяя, так сказать, кумулятивные возможности уже отработанного частотного критерия.

Основной неиспользованный резерв на этом пути — число совпадающих множителей, который не оправдал себя как абсолютный показатель сильного семантического взаимодействия, но, обладая относительной истинностью, мог помочь в решении нашей задачи.

Было намечено несколько комбинаций, учитывающих зависимость силы связи одновременно и от числа совпадений, и от частот совпадающих множителей.

Общий принцип при этом был таков: поскольку релевантность множителя, т. е. его способность быть показателем сильного взаимодействия, определяется пороговой частотой 6 на массиве дескрипторов, а при увеличении числа совпадений вероятность усиления семантической связи растет, то при числе одинаковых множителей  $> 2$  допустимый порог частоты может быть выше 6. Что касается его точной количественной оценки, то она устанавливалась, исходя из частоты, кратной 6. Наиболее сложный из проверяемых на этом этапе критериев отбора формулировался так: для фиксации в паре слово—дескриптор сильного семантического взаимодействия необходимо и достаточно выполнение одного из следующих условий:

- а) при совпадении их кодовых семантических записей по одному множителю его частота должна быть  $\leq 6$ ;
- б) при совпадении двух множителей максимально допустимая частотность 18 ( $\leq 18$ );
- в) при совпадении по трем множителям максимальная частотность может быть равна 36 ( $\leq 36$ ).

Проверка по этому критерию 100 слов, извлеченных по алфавиту из нашего словаря, дала следующие общие результаты: информационные потери уменьшились, но шум увеличился. В ряде случаев общее улучшение, увеличение точности и полноты было особенно существенным; в других — применение изощренного трехуровневого критерия не показало никаких изменений в распределении по сравнению с простым критерием, согласно которому частота совпадающего множителя не должна превы-

шать 6 (ср. пункт «а» усложненного критерия); наконец, наблюдалось определенное число таких изменений, которые ухудшали картину по сравнению с простым критерием, доводя уровень шума до слишком большой величины. Приведем некоторые иллюстрации каждого из этих случаев.

Расположение материала в примерах повторяет аналитические таблицы, которые строились при проверке эффективности разных критериев для 100 испытуемых слов. Они состоят из двух колонок (I и II), и левая содержит перечень дескрипторов, взаимодействие которых с заглавным словом интуитивно оценивалось как сильное. Интуитивное отнесение дескрипторов к заглавному слову было двух рангов — экспериментальное, отражающее результаты ассоциативного эксперимента, и экспертное, т. е. включающее только те дескрипторы из отмеченных в ассоциативном эксперименте (последних, естественно, было больше), связь которых с заглавным словом была подтверждена экспертами. Дескрипторы, отобранные по оценкам экспертов, помещались в верхней (A) половине колонки I, в нижней же ее половине размещались остальные дескрипторы с интуитивно установленной в эксперименте связью. Правая колонка II

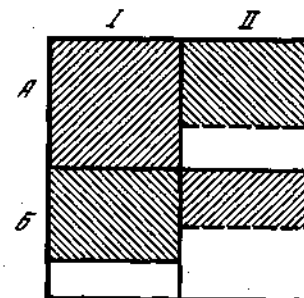


Рис. 1

включает дескрипторы, отвечающие формальным критериям связи с заглавным словом при автоматическом распределении. В ее верхней (AII) половине расположены единицы, отбор которых получил одобрение экспертов, в нижней — дескрипторы, которые по экспертным оценкам находятся в зоне слабого взаимодействия с заглавным словом и представляют собой информационный шум машинного распределения. В числе дескрипторов второй колонки звездочкой помечены те, которые добавлены при применении трехуровневого критерия, т. е. при использовании пунктов «б» и «в» (см. выше, с. 62). Таким образом, для каждого слова строится своеобразный квадрат (рис. 1), разделенный на четыре части, с асимметричным заполнением каждой из них. Левая часть (I) квадрата относится к ассоциативному эксперименту, правая (II) — к машинному. Верхняя половина квадрата (A) содержит позитивный результат экспериментов — как ассоциативного, так и машинного, — получивший подтверждение в экспертных оценках, нижняя половина (B) отражает уровень информационного шума в том и другом эксперименте. Эталонем сравнения является зона AI, и задача заключается в том, чтобы путем варьирования формального критерия отбора добиться максимального увеличения площади заштрихованной части в зоне AII и сокращения белого поля — потерь, при сохранении в разумных пределах (т. е. не выходящей за

рамки квадрата, а значит, не превышающей размеров полезной информации) заштрихованной части в зоне БII, которая отражает уровень шума.

1-я группа случаев: все формальные критерии дали один и тот же результат, который хорошо согласуется с интуитивными представлениями.

алкоголь		алфавит	
спирт	спирт	буква	буква
пиво	пиво	словарь	словарь
вино	вино	читать	читать
пьянство	пьянство	писать	писать
		перечень	перечень
		календарь	календарь
		порядок	—
		правило	—
			описание
ангина			
воспаление	воспаление		
нагноение	нагноение		
язва	язва		
рот	рот		
			губы
			слюна

2-я группа случаев: усложненный трехуровневый критерий (см. пункты «а», «б», «в» на с. 62) дал улучшение результата — либо по сравнению с интуитивными оценками, либо по сравнению с другими формальными критериями (добавления помечены звездочкой).

аккуратный		бактерия	
чистота	чистота	брожение	брожение
точность	точность	плесень	плесень
исполнительность	исполнительность	дезинфекция	дезинфекция
обязательный	* обязательный	заражение	заражение
—	* верность	* животное	* животное
—	* нравственность	дыхание	* дыхание
—	ювелирный	физиология	* физиология
—	исследователь	паразиты	* паразиты
—	бдительность	воздух	—
порядок	набожность	гигиена	—
	соревнование	питаться	—
воля		—	* живое
аппетит			
вкусный	вкусный	рыба	* рост
жевать	жевать	человек	* старость
глотать	глотать	произрастать	* лимфа
питаться	питаться	смерть	
жажда	* жажда	здоровье	
желание	* желание	опухоль	
		плодиться	

3-я группа примеров: изощренный критерий не дает заметного улучшения в результатах распределения или даже ухудшает их.

активный	азот	
движение	* движение	воздух
участвовать	* участвовать	газ
живое	—	взрыв
процесс	—	химия
	сила	смесь
	* интрига	—
	* действие	
	* подъем	ветер
		вода
	соревнование	белый
	* работать	лимфа
	* ум	невыразительный
	* общественный	слеза
	* результат	* звук
	* экономика	* насос
	* остановка	* трубопровод
аптека		* геология
фармакология	фармакология	* медь
больница	—	* атомный
—	медицина	* железо
—	вспрыскивание	* цинк
		* платина
	* ювелирный	* сера
	* работать	* олово
	* фотография	* никель
	малина	* алюминий
		* очищать
		* составлять
		* сложность
		* часть
		* крашение

Справедливости ради следует заметить, что среди асимметричных квадратов-слов преобладают такие, у которых гипертрофирована зона БI, а не БII, как в случае азота, т. е. наибольший информационный шум порождается в ассоциативных, а не в машинных экспериментах.

После проверки этого критерия были рассмотрены его варианты: один из них, при сохранении тех же самых трех условий, расширял зону частотности и исходил из того, что 18 и 36 являются не максимальными величинами частоты, соответственно при двух и при трех совпадающих множителях, а минимальными. Эффект от применения критерия отбора в такой формулировке был ничтожным, поскольку он практически не снижал потери, но заметно увеличивал уровень шума. Другой вариант того же критерия предполагал уменьшение степени кратности: допустимые максимальные частоты были установлены 12 и 18 соответственно. Получаемый при этом результат почти не отличался от распределения по принципу одного совпадающего множителя при частоте  $\leq 6$ . Наконец, на том же материале было проведено испытание самого широкого по своим кумулятивным

возможностям и самого простого по своему программному воплощению критерия, который формулировался следующим образом: для фиксации в паре слово—деSCRIPTOR сильной семантической связи необходимо и достаточно выполнение одного из двух условий:

- а) при совпадении одного множителя его частота не должна превышать величины 6 ( $\leq 6$ );
- б) при совпадении более чем одного множителя, т. е. двух и больше, частота не учитывается.

Использование этого критерия привело к максимальному сокращению потерь при известном возрастании, естественно, информационного шума. Сравнительная таблица эффективности разных критериев показала, что если, в надежде на постредктирование, не обращать внимания на увеличение шума в отдельных случаях, то наиболее приемлемый результат мы получаем при использовании последнего селективного критерия. Он и был положен в основу окончательного распределения материала по всему словарю, и примеры для входов  $K \rightarrow 3$  и  $3 \rightarrow K$ , приводившиеся в начале главы, взяты из окончательной редакции тезауруса.

Таким образом, после испытания и оценки феноменологической модели распределения слов по дескрипторам, которая строилась на принципе равносемности («если есть общий множитель, то есть семантическая связь»), были осуществлены поиски уточненного селективного критерия по трем направлениям. Они составили в итоге три последовательных этапа доводки модели распределения, ее приближения к идеально представляемому результату: этап учета числа совпадающих множителей, этап учета частоты встречаемости множителя на массиве дескрипторов и этап соединения числа совпадений с показателем частотности совпадающих множителей. Эти шаги последовательных приближений и дали основания для того, чтобы назвать полученную модель асимптотической.

Дальнейшее совершенствование модели и улучшение итогового распределения может идти по линии количественного изменения силы семантической связи между дескриптором и составляющими его статью словами и коррекции полученного ранее результата, исходя из определенной величины коэффициента силы связи. Способы подсчета таких коэффициентов многообразны, но все они так или иначе учитывают соотношение суммарного числа семантических множителей у пары слов с количеством одинаковых, совпадающих у них множителей. Для нескольких семантических полей в нашем тезаурусе эти коэффициенты подсчитаны по формуле:

$$P_{x_i x_j} = \frac{K}{M_i + M_j - K},$$

где  $x_i, x_j$  — два слова или слово и дескриптор;  $M_i, M_j$  — число семантических множителей в дефинициях слов  $x_i, x_j$ ;  $K$  — число совпадений по множителям.

Анализ распределения коэффициентов силы связи в двух дескрипторных статьях АБСОЛЮТНЫЙ и АВИАЦИЯ (их состав и значения коэффициентов для каждого слова приводятся в главе III) выявил неожиданную картину. Оказалось, что между определенным по указанной формуле коэффициентом силы связи и интуитивной оценкой качества семантического взаимодействия слова с дескриптором нет прямо пропор-

циональной зависимости. Иными словами, мы не можем сказать, что обязательно чем выше коэффициент силы связи в данной паре слов, тем увереннее констатируют эксперты в этой паре сильное семантическое взаимодействие. Подобное утверждение остается справедливым лишь до некоторого значения коэффициента, ниже которого связь может стать уже достаточно слабой, но может сохраниться и на высоком уровне, допускающем, по мнению экспертов, ее включение в тезаурус. Таким образом, вновь повторяется ситуация, отмеченная нами выше при обсуждении результатов ассоциативного эксперимента, который также показал отсутствие прямой зависимости между ассоциативной повторяемостью и числом совпадающих в паре слов семантических множителей. Если результаты распределения коэффициентов силы связи в двух названных полях представить в виде графика, где по оси абсцисс отложено убывание значения коэффициента в пределах от 1 до 0, а по оси ординат условная градация оценки экспертами отнесенности слова к дескриптору, то мы получим следующую картину (рис. 2).

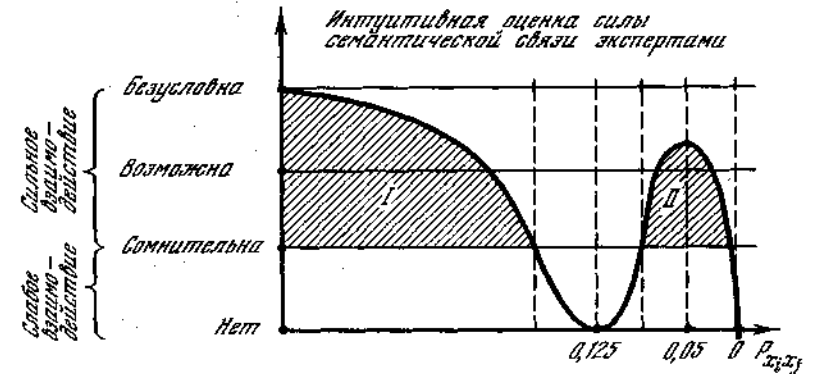


Рис. 2

На графике видно, что существует некоторая зона I значений коэффициента, которая хорошо согласуется с интуитивными оценками (на рисунке она заштрихована), и мы можем провести условный порог этих значений в районе 0,125 (этот порог будет, очевидно, варьироваться для разных полей). По достижении коэффициентом этого порога слова уже не оцениваются экспертами как сильно связанные со своим дескриптором. Однако при дальнейшем понижении значений коэффициента наступает второй всплеск кривой (заштрихованная зона II): эксперты снова констатируют сильную семантическую связь, пик которой приходится примерно на значение коэффициента 0,05.

Обобщать это наблюдение, видеть в нем постоянную закономерность и распространять на все статьи тезауруса, конечно, рано. Тем более рано, не накопив достаточного количества материала, пытаться дать этому явлению содержательную интерпретацию. Внешняя же его аналогия с ситуацией, возникшей при обсуждении результатов ассоциативного эксперимента, легко объяснима: интуитивная оценка — информантов ли, экспертов ли — и в том и в другом случае соотносилась с количественными показателями, опирающимися только на число множителей. Этот показатель может быть проще (ср. обыкновенный подсчет одинаковых мно-

жителей) или изощренное (ср. формулу для коэффициента силы связи), но он все равно использует только один параметр в сравниваемых дефинициях. Отсюда и принципиальное сходство результатов. Вероятно, в формулу коэффициента надо ввести и статистические характеристики семантических множителей, что позволит более адекватно представить картину семантических взаимодействий в словарной статье тезауруса.

### Глава III

## МЕТОДЫ АНАЛИЗА МЕТАЯЗЫКА СЛОВАРЯ

### § 1. СИСТЕМНЫЙ ПОДХОД К АНАЛИЗУ МЕТАЯЗЫКА СЛОВАРЯ

Интенсивное развитие системной проблематики в последние годы постепенно приводит к важным изменениям дисциплинарной структуры современной науки и техники. Многие факты свидетельствуют о том, что исследования по системному подходу, общей теории систем и системному анализу идут в направлении формирования особой научной дисциплины. Одновременно с этим все шире становится проникновение системных идей в специальные области науки и техники.

В этих условиях неизмеримо возрастает роль философского и методологического анализа системных исследований. В широком комплексе методологических проблем, связанном с развитием современных системных исследований, имеется ряд ключевых вопросов, к обсуждению которых исследователи — философы и специалисты конкретных наук — возвращаются вновь и вновь. Это вопросы о взаимоотношении диалектики и методов исследования систем и структур, о статусе и способах построения общей теории систем, о природе системного подхода и формах реализации его принципов и т. п. Необходимость постоянного возвращения к теоретическому осмыслению этих вопросов вызвано прежде всего тем, что по мере прогресса специальных системных исследований открываются их новые аспекты и стороны.

Сказанное особенно справедливо для периода современной научно-технической революции, когда в орбиту системных исследований интенсивно вовлекаются все новые и новые области науки, техники и практической деятельности.

В результате первых, более или менее систематических попыток анализа предпосылок и основных проблем системных исследований, проведенных на основе диалектико-материалистических философских принципов, в конце 60-х годов было сформулировано понимание системного подхода как методологической концепции, как одного из трех крупных методологических направлений, связанных с изучением системных объектов (наряду со структурно-функциональным анализом и структурализмом), как определенного научно-исследовательского подхода, имеющего междисциплинарный характер и направленного на разработку новой общественной стратегии исследования целостных объектов<sup>1</sup>.

<sup>1</sup> См., напр.: Блауберг И. В., Садовский В. Н., Юдин Э. Г. Системный подход: предпосылки, проблемы, трудности. М., 1969; Они же. Системный подход в современной науке. — В кн.: Проблемы методологии системного исследования. М., 1970.

В названных работах при выделении основных направлений развития современного системного подхода в качестве одного из таких направлений называется философская проблематика системного подхода, определение его познавательных возможностей, формирование общих (мировоззренческих) принципов системного анализа. В этом отношении, следовательно, системный подход не отличается от кибернетического, семиотического и других подходов, имеющих междисциплинарный характер.

Такое понимание системного подхода опиралось на убеждение в том, что современное методологическое знание не представляет собой некое однородное образование. Бурному развитию науки, усложнению ее структуры, существенному увеличению в ней роли теоретического, абстрактного рассуждения, широкой математизации и формализации современной науки и т. п. сопутствовал интенсивный процесс дифференциации методологического знания, разработки наряду с общеподлинными методологическими принципами конкретно-научных методологических понятий и концепций. Эта новая ситуация в методологии науки была подвергнута анализу, в результате чего было выделено четыре основных уровня современного методологического знания<sup>2</sup>.

1. Уровень философской методологии — анализ общих принципов познания и категориального строя науки в целом. Эта сфера методологии представляет собой раздел философского знания и разрабатывается специальными методами.

2. Уровень общенаучных методологических принципов и форм исследования, куда включаются как содержательные общенаучные концепции (методологические принципы кибернетики и т. д.), так и формальные методологические теории (логика науки, разрабатываемая на основе применения аппарата математической логики, и т. д.). Поскольку общенаучные методологические концепции не претендуют на решение мировоззренческих общеподлинных задач, их разработка осуществляется в сфере нефилософского знания, а именно — в рамках современной логики и методологии науки.

3. Уровень конкретно-научной методологии, на котором анализируются методы, принципы и процедуры исследования, применяемые в специальных научных дисциплинах.

4. Уровень методики и техники исследования — описание способов получения релевантной информации, условий проведения экспериментов, учета погрешностей, методов обработки экспериментальных данных и т. д.

Это разграничение уровней современного методологического знания позволило уточнить ранее сформулированное понимание статуса системного подхода и его отношение к философской методологии — диалектике. Такое уточнение было проведено в начале 70-х годов. При этом принципиальное положение о нефилософском характере системного подхода осталось без изменений: системный подход не может быть отнесен как таковой к уровню философской методологии, сам по себе он не связан непосредственно ни с разработкой мировоззренческой проблематики, ни с выполнением функции философской критики форм и принципов научного познания<sup>3</sup>. Системные исследования представляют собой одну из быстро разви-

<sup>2</sup> Блауберг И. В., Юдин Э. Г. Становление и сущность системного подхода. М., 1973; Лекторский В. А., Швырев В. С. Методологический анализ науки (типы и уровни). — В кн.: Философия. Методология. Наука. М., 1972.

<sup>3</sup> Блауберг И. В., Юдин Э. Г. Указ. соч.

вающихся сфер современного конкретного научного познания, их теоретическое описание в рамках системного подхода или общей теории систем также ограничено проблемами специально-методологического порядка и ни в коей мере не может претендовать на противопоставление и тем более на замену философской методологии<sup>4</sup>.

В рамках такого общего понимания статуса системного подхода идея различных уровней методологии дала возможность охарактеризовать системный подход не только негативно (как нефилософскую методологию), но и позитивно (как общенаучное междисциплинарное методологическое знание). В соответствии с этим системный подход оказывается особым предметом исследования на втором из указанных уровней методологии, т. е. на уровне общенаучных методологических принципов и форм знания. Принципы системного подхода, таким образом, имеют большую общность, чем методологические утверждения, формулируемые и применяемые в конкретных областях научного и технического знания, но они при этом не выходят за рамки специально-научного знания и тем самым не претендуют на философскую общность.

Разграничение уровней методологии позволяет не только выделить различные типы методологического анализа, но и установить взаимосвязь между ними. В частности, оно предполагает, что философская методология — диалектика имеет основополагающее значение для любых форм методологического знания. Важным условием такого разграничения является также тезис о преимущественном влиянии методологии более общего уровня на уровне методологии, имеющие дело с менее обобщенными методологическими утверждениями. Так, например, обобщенные принципы, и в частности системный подход, оказывают несомненное воздействие на формирование методологического знания на уровне конкретно-научной методологии и на уровне методики и техники исследования. Учет этих связей между уровнями методологии дает возможность построить более детальную картину взаимоотношения диалектики и системного подхода<sup>5</sup>.

Диалектика является основой системно-структурной методологии, ее философским базисом; в свою очередь развитие системно-структурной методологии способствует обогащению и конкретизации методологического потенциала диалектики в тех ее разделах, которые связаны с философско-методологическими характеристиками сложноорганизованных объектов действительности — систем и структур<sup>6</sup>.

Такое понимание взаимоотношения диалектики и системного подхода базируется, в частности, на том, что философские идеи системности, целостности, структурности, универсальности и многообразия форм связи и т. д. органически присущи диалектическому методу и пронизывают все его важнейшие понятия и принципы. Эти идеи лежат и в основе метода восхождения от абстрактного к конкретному, и принципа единства логического и исторического, и в диалектическом требовании при исследовании объекта выявлять координацию и субординацию его связей, определять специфические связи анализируемого объекта, и в принципе

единства генетического и структурного анализа, и т. д. Эти же идеи, взятые с точки зрения их конкретного научного содержания, разрабатываются в рамках системного подхода и общей теории систем. На этой основе был сформулирован общий вывод: «как системный подход не вправе претендовать на решение философских проблем метода научного познания, так и диалектика не стремится подменить собой конкретно-научную проблематику системного подхода и общей теории систем<sup>7</sup>».

Изложенное понимание статуса системного подхода и его взаимоотношение с диалектикой, как нам представляется, достаточно полно учитывает и принципиальное для марксистской философии утверждение о примате философской методологии над другими формами методологического знания, и современный уровень теоретического осознания сущности и специфики методов подхода.

Предлагаемая нами точка зрения по системному подходу к анализу метаязыка словаря состоит в том, что надо изучать форму целостных совокупностей как составленных из не очень четко определенных элементов, а свойства последних не столько определяют эту форму, сколько определяются ею.

Членение системы, т. е. представление ее в виде множества подсистем, определяется не произволом наблюдателя, в внутренних свойствах системы. Удобные для описания системы членения — это те, которые отражают сущность системы.

Часто имеются такие совокупности, где один и тот же объект повторяется несколько раз. Например, в тексте много раз могут повторяться одни и те же слова, в нашем случае — семантические множители в определении дескрипторов и слов. Семантические множители могут быть упорядочены по убыванию их числа по информационным массивам дескрипторов и слов. Зависимость численности, соответствующей данному элементу, от его порядкового номера (ранга) при расположении элементов по убыванию этой численности называется ранговым распределением. Оказывается, что ранговое распределение имеет обычно весьма устойчивую форму<sup>8</sup>.

При этом существенным оказывается тот факт, что ранговое распределение фиксируется при членении естественной системы, а не путем анализа больших конгломератов однородных систем. Целостность здесь важнее объема выборки. Это явно противоречит гипотезе о статистической природе ранговых распределений и позволяет рассматривать существование ранговых распределений как проявление целостности системы<sup>9</sup>.

Каталог дескрипторов и словник, обладающие указанными ранее свойствами (семантическая непрерывность и т. д.), представляют собой самостоятельные замкнутые системы. Сравнивая ранговые распределения семантических множителей по информационным массивам дескрипторов и слов видим, что объем выборки не играет существенной роли. Так, например, семантический множитель «зем», ранг которого равен 1, имеет

<sup>4</sup> Садовский В. Н. Основания общей теории систем. Логико-методологический анализ. М., 1974.

<sup>5</sup> Садовский В. Н. Указ. соч.

<sup>6</sup> Блауберг И. В., Юдин Э. Г. Указ. соч.

<sup>7</sup> Садовский В. Н. Указ. соч., с. 48.

<sup>8</sup> Аранов М. В., Ефимова Е. Н., Шрейдер Ю. А. О смысле ранговых распределений. — НТИ. Сер. 2, 1977, № 11.

<sup>9</sup> Аранов М. В., Шрейдер Ю. А. Классификация и ранговые распределения. — НТИ. Сер. 2, 1977, № 11.

частоты по массиву дескрипторов максимальную — 69, по массиву слов — 102. Множитель «дви», ранг которого равен 8, по массиву дескрипторов имеет частоту 52, а по массиву слов максимальную — 183. С увеличением порядкового номера множителя ранг его в массиве дескрипторов и слов совпадает.

Это можно интерпретировать как высокую степень связи между семантическими множителями дескрипторов и слов.

Таким образом, мы считаем, что ранговые распределения можно рассматривать не только как распределение случайных величин, привычных для классической статистики, но и в совершенно иной понятийной «парадигме». Выполнение на данном объекте (тексте) рангового распределения — это, с нашей точки зрения, признак «правильности» (хорошей организации) данного текста, взятого как единое целое. При таком подходе, например, свойство законченного языкового текста быть «целостным» рассматривается в том же ряду (парадигме), что и его свойство быть построенным по законам грамматики данного языка. Мета-язык словаря, представленный в правой части толкового словаря, в известном смысле отражает семантическое ядро русского языка. Можно предположить, что статистические структуры отдельных текстов русского естественного языка могут быть согласованными друг с другом, а именно: слово, имеющее «высокий статус» в одном тексте, с большой вероятностью сохраняет этот «статус» и в другом тексте.

Статистический анализ правой части толкового словаря дал возможность получить ранговые распределения семантических множителей. В настоящее время понятие рангового распределения в информатике стало вполне привычным.

Идея распределения информационных потоков по закону Ципфа-Мандельброта принята общественным мнением и является теоретической основой изучения этих потоков<sup>10</sup>. Этот же тип распределения имеется и в лингвистике.

## § 2. СТАТИСТИЧЕСКИЙ АНАЛИЗ

Дальнейшее развитие и повышение эффективности научно-исследовательских работ во всех областях обществоведения связано с переработкой непрерывно увеличивающихся объемов социально-экономической, исторической и иной информации. Развитие современной научно-технической революции вызвало в жизни настолько большой поток научной и производственной информации, что традиционных методов ее накопления, систематизации и переработки становится явно недостаточно. Специалист физически не в состоянии прочесть или хотя бы бегло просмотреть все, что издается даже по его узкой специальности. Эффективное решение проблемы переработки больших массивов информации возможно при создании информационно-переводящих кибернетических устройств с передачей ряда трудоемких и рутинных операций по сбору, первичной обработке и хранению информации электронно-вычислительными машинами.

Привлечение ЭВМ к массовой переработке информации меняет социальный статус языка, расширяет его коммуникативно-общественные функции. Продолжая оставаться важнейшим орудием общения между людьми, естественный язык становится той основой, на которой организуется диалог между человеком и ЭВМ. Автоматизация процессов обработки текстовой информации представляет собой, по существу, моделирование понимания текстов и сопоставления их содержания, т. е. некоторую упрощающую имитацию человеческого мышления. Для ее внедрения необходимы создание специального информационного языка и разработка алгоритмов перевода содержания документов на этот язык. Это обуславливает необходимость дальнейшего всемерного расширения применения вычислительной техники для существенного улучшения информационной работы в обществоведении.

В настоящее время возрастает интерес к частотным словарям, эффективность использования которых для решения разных прикладных и исследовательских задач, как и сам круг этих задач, постоянно возрастают<sup>11</sup>. Статистическая лингвистика, формулирующаяся на базе частотных словарей, занимается лингвистической и статистической интерпретацией распределения слов в текстах и на основании этого дает заключение о статистических закономерностях языка.

Статистика выступает и как самостоятельная наука, поскольку изучает встречающийся в действительности объект — массовое явление, и в то же время как метод в изучении конкретного массового явления, независимо от сферы, к которой оно принадлежит. В лингвистике статистика проявляет себя в статусе метода, поскольку в результате изучения явлений текста, речи она устанавливает общие закономерности.

Значительное расширение статистической базы лингвистики возможно только за счет автоматизации обработки текстов. Это, естественно, обуславливает необходимость более активного использования в лингвистических исследованиях вычислительной техники. Такое использование дает возможность значительно повышать качество научно-исследовательских и экспериментальных работ и сокращать сроки их выполнения. Если учесть, что активное применение вычислительной техники в лингвистике началось сравнительно недавно, то следует отметить определенные успехи на этом пути. Теперь уже не нужно доказывать лингвистам желательность, а подчас и необходимость применения в исследованиях ЭВМ — практика показывает его эффективность и целесообразность. Спектр проблем, изучаемых лингвистами за последние годы, существенно увеличился, соответственно расширились и зоны применения ЭВМ и математических методов для расширения этих проблем. К разбивавшейся подробно автоматической лексикографии следует добавить автоматическую обработку текстов, реферирование, аннотирование, машинный перевод и др. Все это потребовало модификации традиционных процедур и разработки ряда новых, более адекватных природе изучаемых объектов приемов обработки текстовой информации.

<sup>10</sup> Мандельброт Б. Теория информации и психоллингвистическая теория частот слов. — В кн.: Математические методы в социальных науках. М., 1973.

<sup>11</sup> Денисов П. Н., Морковкин В. В., Сафьян Ю. А. Комплексный частотный словарь русской научной и технической лексики. М., 1978; Головин Б. Н. Язык и статистика. М., 1971; Частотный словарь русского языка. М., 1977; Частотный словарь индексирования. Под общ. ред. Л. В. Сахарного. Пермь, 1974.



В настоящей работе приведен один из вариантов получения частотного словаря семантических компонентов с помощью ЭВМ.

В данном случае ЭВМ применяется для автоматической обработки текстовой информации определенных дескрипторов и слов в словарях, формализации описания языковых структур, моделирования процессов анализа текста, накопления статистики семантических компонентов. Это исследование получит практическое приложение при разработке тезауруса русского литературного языка. Приемы выделения семантических компонентов, данные об их частотности можно также использовать при разработке ИПС и банков данных по общественным наукам, в частности банка социологических данных.

Статистический анализ текстов и применение ЭВМ в этом анализе позволяет выявить, например:

- 1) наиболее частотные грамматические морфемы;
- 2) наиболее частотные словоформы языка;
- 3) наиболее употребительные в исследуемых текстах слова и группы слов;
- 4) все возможные в изучаемом языке контексты, в которых встречаются заданное слово или группа слов;
- 5) наиболее употребительные классы слов данного языка;
- 6) статистику употреблений типов предложений на уровне классов слов.

По статистическим характеристикам словаря семантических компонентов можно определить ту часть общего толкового словаря, в которую входят компоненты с высокой частотой появления — либо в массиве дескрипторов (с семантическими компонентами), либо в массиве слов (с семантическими компонентами) (см. предыдущую главу).

Статистическая структура словаря представляет собой таблицу распределения частот (см. табл. 1). Первые две графы дают полную статистическую информацию о связи «ранга» (порядкового номера группы компонентов с данной частотой) и общей частоты. В третьей графе таблицы даются сведения о количестве компонентов с данной частотой. Так, по одному разу зарегистрированы частоты от 69 до 52 (компоненты с порядковыми номерами от 1 до 8). В графе «накопленная абсолютная частота» содержатся сведения о сумме частот группы компонентов, в которую входят все компоненты от самого частотного (первого «по рангу») до данного компонента. Так, для группы компонентов с порядковым номером 8 эта сумма равна 483. По этой величине можно определить, какую долю в определениях дескрипторов составляет данная группа компонентов. Накопленная относительная частота (см. последнюю графу таблицы) служит для определения покрытия дефиниций группой компонентов с данными частотами. Так, для компонента с порядковым номером 8 эта величина равна 0,0258. Это значит, что 8 самых частотных компонентов покрывают 2,58% дефиниций. Для группы компонентов с порядковым номером 15 накопленная относительная частота равна 0,0504. В эту группу входят 18 компонентов, что мы узнаем из графы «накопленное число компонентов». Они покрывают 5,04% дефиниций дескрипторов.

Для отбора группы частотных компонентов используются данные графы «число компонентов с данной частотой». Выбирается группа компонентов заданного количества и далее определяется покрытие дефиниций

дескрипторов этой группой. Так, для 8 самых частотных компонентов накопленное относительное число компонентов равно 0,0014. Это означает, что они составляют 0,14% всего словаря. 15 частотных компонентов составляют 0,3% словаря. Аналогично мы находим интересные нас величины и по табл. 2 распределения частот по массиву слов с семантическими компонентами. Вероятно, эти данные более адекватным образом характеризуют структуру словаря.

Частотный словарь семантических множителей содержит 5562 различных компонента, образующих 18 649 словоупотреблений в дефинициях дескрипторов и 7870 компонентов, образующих 47 778 словоупотреблений в дефинициях слов. Отсюда можно получить коэффициент лексического разнообразия словаря. Это отношение числа разных компонентов ( $V$ ) к числу всех словоупотреблений ( $N$ ) для массива дескрипторов с семантическими компонентами:

$$C = \frac{V}{N} = \frac{5562}{18649} = 0,2982.$$

Для массива слов с семантическими компонентами:

$$C = \frac{7870}{47778} = 0,1647.$$

Для сравнения приведем данные разных частотных словарей по этому коэффициенту<sup>12</sup>.

Частотный словарь языка А. С. Пушкина:

$$C = \frac{21197}{544777} = 0,039.$$

Частотный словарь русского языка, составленный Э. А. Штейнфельд:

$$C = \frac{24224}{400000} = 0,0605.$$

Частотный словарь русского языка, составленный Л. Н. Засориной:

$$C = \frac{39268}{1056382} = 0,0371.$$

Из сравниваемых данных видно, что величина  $C$  зависит от размера выборки и от размера полученного списка разных единиц.

Величина  $C$  изменяется от 1 (в том случае, когда  $N = V$ , т. е. в определениях дескрипторов и слов все компоненты разные, что может произойти при минимальном  $N$ ) и стремится к бесконечно малой при неограниченных объемах выборки, т. е. при неограниченном увеличении количества дескрипторов или слов. Это объясняется тем, что наиболее интенсивно словарь семантических компонентов растет вначале, а увеличение объема дескрипторов и слов, соответственно увеличение объема семантических компонентов в определениях этих дескрипторов и слов, приводит к большому повторению встречавшихся ранее компонентов и соответственно к замедлению роста словаря семантических компонентов.

Таблицы распределения семантических компонентов и их частот дают интересные сведения о качественном расслоении семантических множителей.

<sup>12</sup> См.: Частотный словарь русского языка. М., 1977.

Таблица 1

Распределение частот по массиву дескрипторов с семантическими множителями

Ранг	Абсолютная частота	Число компонентов с данной частотой	Накопленная абсолютная частота	Накопленное число компонентов		Относительная частота	Накопленная относительная частота
				абсолютное	относительное		
1	2	3	4	5	6	7	8
1	69	1	69	1	0,0001797	0,0036999	0,0036999
2	67	1	136	2	0,0003595	0,0035926	0,0072926
3	66	1	202	3	0,0005393	0,0035390	0,0108316
4	60	1	262	4	0,0007191	0,0032173	0,0140490
5	59	1	321	5	0,0008989	0,0031637	0,0172127
6	57	1	378	6	0,0010787	0,0030564	0,0202691
7	53	1	431	7	0,0012585	0,0028419	0,0231111
8	52	1	483	8	0,0014383	0,0027883	0,0258995
9	51	2	585	10	0,0017979	0,0027347	0,0313689
10	47	3	726	13	0,0023372	0,0025202	0,0385297
11	46	1	772	14	0,0025170	0,0024666	0,0413963
12	44	1	816	15	0,0026968	0,0023593	0,0437556
13	43	1	859	16	0,0028766	0,0023057	0,0460614
14	41	1	900	17	0,0030564	0,0021985	0,0482599
15	40	1	940	18	0,0032362	0,0021448	0,0504048
16	38	4	1092	22	0,0039554	0,0020376	0,0585554
17	37	1	1139	23	0,0041352	0,0019840	0,0605399
18	36	3	1237	26	0,0046745	0,0019303	0,0663306
19	35	1	1272	27	0,0048543	0,0018767	0,0682074
20	33	4	1404	31	0,0055735	0,0017695	0,0752855
21	32	1	1430	32	0,0057533	0,0017159	0,0770014
22	31	4	1560	36	0,0064724	0,0016622	0,0836505
23	30	1	1590	37	0,0066522	0,0016086	0,0852592
24	29	4	1706	41	0,0073714	0,0015550	0,0914794
25	28	3	1790	44	0,0079108	0,0015014	0,0955836
26	27	5	1925	49	0,0086097	0,0014477	0,1032226
27	26	4	2029	53	0,0095289	0,0013941	0,1087993
28	25	6	2154	58	0,0104279	0,0013405	0,1155021
29	24	8	2346	66	0,0118662	0,0012869	0,1257976
30	23	9	2553	75	0,0134843	0,0012333	0,1366974
31	22	10	2773	85	0,0152822	0,0011796	0,1486942
32	21	10	2983	95	0,0170801	0,0011260	0,1595549
33	20	8	3143	103	0,0185185	0,0010724	0,1685345
34	19	8	3295	111	0,0199568	0,0010188	0,1766850
35	18	16	3583	127	0,0226335	0,0009651	0,1921282
36	17	13	3804	140	0,0251708	0,0009115	0,2035787
37	16	24	4188	164	0,0294857	0,0008579	0,2245696
38	15	20	4488	184	0,0330816	0,0008043	0,2406563
39	14	28	4880	212	0,0381157	0,0007507	0,2516762
40	13	40	5400	254	0,0453074	0,0006970	0,2895597
41	12	31	5772	283	0,0508809	0,0006434	0,3095072
42	11	51	6333	334	0,0600503	0,0005898	0,3395892
43	10	65	6983	399	0,0717367	0,0005362	0,3774436
44	9	69	7604	468	0,0841423	0,0004825	0,4077430
45	8	105	8444	573	0,1030204	0,0004289	0,4527856
46	7	118	9270	691	0,1242358	0,0003753	0,4976775
47	6	143	10128	834	0,1495460	0,0003217	0,5436854
48	5	204	11148	1038	0,1866235	0,0002681	0,5977800
49	4	309	12384	1347	0,2421790	0,0002144	0,6640570
50	3	524	13956	1871	0,3363897	0,0001608	0,7483511
51	2	1002	15960	2873	0,5165408	0,0001072	0,8556099
52	1	2689	18649	5562	1,0000000	0,0000536	1,0000000

Таблица 2

Распределение частот по массиву слов с семантическими множителями

Ранг	Абсолютная частота	Число компонентов с данной частотой	Накопленная абсолютная частота	Накопленное число компонентов		Относительная частота	Накопленная относительная частота
				абсолютное	относительное		
1	2	3	4	5	6	7	8
1	183	1	183	1	0,0001270	0,0038302	0,0038302
2	181	1	364	2	0,0002541	0,0037883	0,0076185
3	180	2	724	4	0,0005082	0,0037674	0,0151534
4	161	1	885	5	0,0006353	0,0033697	0,0185231
5	160	1	1045	6	0,0007623	0,0033488	0,0216719
6	137	1	1182	7	0,0008894	0,0028674	0,0247394
7	136	1	1318	8	0,0010165	0,0028464	0,0275859
8	129	1	1447	9	0,0011435	0,0026999	0,0302859
9	112	1	1559	10	0,0012706	0,0023441	0,0326300
10	110	2	1779	12	0,0015247	0,0023023	0,0372347
11	107	2	1993	14	0,0017789	0,0022395	0,0417377
12	103	1	2096	15	0,0019059	0,0021558	0,0436695
13	102	1	2198	16	0,0020330	0,0021348	0,0460044
14	101	1	2299	17	0,0021601	0,0021138	0,0481183
15	99	2	2497	19	0,0024142	0,0020720	0,0522625
16	97	2	2691	21	0,0026683	0,0020302	0,0563229
17	96	1	2787	22	0,0027954	0,0020092	0,0583322
18	95	1	2882	23	0,0029224	0,0019883	0,0603205
19	91	2	3064	25	0,0031766	0,0019046	0,0641299
20	90	1	3154	26	0,0033036	0,0018837	0,0660136
21	89	2	3332	28	0,0035578	0,0018627	0,0697392
22	87	1	3419	29	0,0036848	0,0018209	0,0715601
23	85	1	3504	30	0,0038119	0,0017790	0,0733391
24	83	1	3578	31	0,0039390	0,0017372	0,0750763
25	81	4	3911	35	0,0044472	0,0016953	0,0816577
26	80	2	4071	37	0,0047013	0,0016744	0,0852065
27	76	1	4147	38	0,0048284	0,0015905	0,0867972
28	75	3	4372	41	0,0052096	0,0015697	0,0915065
29	74	1	4446	42	0,0053367	0,0015488	0,0930553
30	73	2	4592	44	0,0055908	0,0015278	0,0961111
31	71	2	4734	46	0,0058449	0,0014860	0,0990832
32	70	2	4874	48	0,0060991	0,0014651	0,1020134
33	69	4	5150	52	0,0066073	0,0014441	0,1077901
34	68	1	5218	53	0,0067344	0,0014232	0,1093134
35	67	3	5419	56	0,0071156	0,0014023	0,1134204
36	66	3	5617	59	0,0074968	0,0013813	0,1175645
37	65	3	5747	61	0,0077509	0,0013604	0,1202854
38	64	3	5939	64	0,0081321	0,0013396	0,1243040
39	63	2	6065	66	0,0083862	0,0013185	0,1265412
40	62	4	6313	70	0,0088945	0,0012975	0,1321319
41	61	3	6496	73	0,0092757	0,0012767	0,1355621
42	59	4	6732	77	0,0097839	0,0012348	0,1405016
43	58	2	6848	79	0,0100381	0,0012139	0,1433295
44	57	4	7076	83	0,0105463	0,0011930	0,1481016
45	56	1	7132	84	0,0106734	0,0011720	0,1492737
46	55	3	7297	87	0,0110546	0,0011511	0,1527271
47	54	2	7405	89	0,0113087	0,0011302	0,1545876
48	53	2	7511	91	0,0115628	0,0011092	0,1572062
49	52	6	7823	97	0,0123252	0,0010883	0,1637364
50	51	5	8078	102	0,0125606	0,0010674	0,1690736
51	60	8	8478	110	0,0139771	0,0010465	0,1774456
52	49	3	8625	113	0,0143583	0,0010255	0,1805224

Таблица 2 (окончание)

1	2	3	4	5	6	7	8
53	48	4	8817	117	0,0148665	0,0010046	0,1845410
54	47	10	9287	127	0,0161372	0,0009837	0,1943781
55	46	9	9701	136	0,0172808	0,0009627	0,2030432
56	45	13	10286	149	0,0185326	0,0009418	0,2152873
57	44	9	10682	158	0,0200762	0,0009209	0,2235757
58	43	9	11069	167	0,0212198	0,0008999	0,2316756
59	42	13	11615	180	0,0226776	0,0008790	0,2431035
60	42	6	11861	186	0,0236340	0,0008581	0,2482523
61	40	9	12221	195	0,0247776	0,0008372	0,2557871
62	39	9	12572	204	0,0255212	0,0008162	0,2631336
63	38	7	12838	211	0,0266166	0,0007953	0,2687010
64	37	10	13208	221	0,0280813	0,0007744	0,2764452
65	36	10	13568	231	0,0293519	0,0007534	0,2835800
66	35	10	13918	241	0,0306226	0,0007325	0,2913056
67	34	8	14190	249	0,0316391	0,0007116	0,2965986
68	33	10	14520	259	0,0325097	0,0006906	0,3035055
69	32	14	14968	279	0,0346886	0,0006697	0,3132822
70	31	17	15495	290	0,0366487	0,0006488	0,3243124
71	30	27	16305	317	0,0402795	0,0006279	0,3412658
72	29	19	16856	336	0,0426937	0,0006069	0,3527983
73	28	11	17164	347	0,0440914	0,0005860	0,3592448
74	27	21	17731	368	0,0467598	0,0005651	0,3711122
75	26	22	18303	390	0,0495552	0,0005441	0,3830842
76	25	19	18778	409	0,0515695	0,0005232	0,3930260
77	24	29	19474	438	0,0556543	0,0005023	0,4075934
78	23	20	19934	458	0,0581966	0,0004813	0,4282724
79	22	24	20462	482	0,0612452	0,0004604	0,4282724
80	21	22	20924	504	0,0640406	0,0004395	0,4375421
81	20	35	21624	539	0,0684879	0,0004186	0,4525932
82	19	39	22365	578	0,0734434	0,0003976	0,4681024
83	18	45	23175	629	0,0791613	0,0003767	0,4850558
84	17	44	23923	667	0,0847522	0,0003558	0,5007116
85	16	47	24675	714	0,0907242	0,0003348	0,5164510
86	15	68	25695	782	0,0993646	0,0003139	0,5377998
87	14	71	26689	853	0,1083862	0,0002930	0,5586043
88	13	83	27768	936	0,1185326	0,0002720	0,5811879
89	12	68	28584	1004	0,1275730	0,0002511	0,5982669
90	11	90	29574	1094	0,1380088	0,0002302	0,6185878
91	10	113	30704	1207	0,1533672	0,0002092	0,6426388
92	9	153	32081	1360	0,1726081	0,0001883	0,6714596
93	8	159	33353	1519	0,1930114	0,0001674	0,6980827
94	7	210	34823	1729	0,2196950	0,0001465	0,7286500
95	6	239	36257	1968	0,2500635	0,0001255	0,7586639
96	5	355	38032	2323	0,2951715	0,0001046	0,7960149
97	4	466	39896	2789	0,3543837	0,0000837	0,8350286
98	3	732	42092	3521	0,4473951	0,0000627	0,8805912
99	2	2337	44766	4858	0,6172808	0,0000418	0,9365584
100	1	3012	47778	7870	1,0000000	0,0000209	1,0000000

Из табл. 3 видно, что среди 834 самых частых компонентов дескрипторы получаются из 439 компонентов (частоты от 6 до 69), что составляет 29,22% всех дескрипторов. А среди 2689 компонентов с частотой 1 дескрипторы получаются из 440 компонентов, что составляет 29,25% всех дескрипторов. Как видно, эти проценты почти равны, но самые частые компоненты, из которых получаются дескрипторы, составляют 52,63% от всех компо-

Таблица 3

Распределение семантических компонентов по их частотам по массиву документов

Частота	Количество компонентов, из которых получаются дескрипторы	Количество компонентов с данной частотой	Количество компонентов, из которых получаются дескрипторы, в % к компонентам с данной частотой	Количество компонентов, из которых получаются дескрипторы, в % ко всем дескрипторам
1	440	2689	16,34	29,25
2	264	1002	26,34	17,57
3	161	524	30,72	10,71
4	117	309	37,86	7,76
5	81	204	39,70	5,35
6—69	439	834	52,63	29,22

нентов с данными частотами, а компоненты с частотой 1, из которых получаются дескрипторы, составляют только 16,36% от всех компонентов с этой частотой.

Отсюда можно сделать два вывода. Во-первых, семантические множители, полученные из дескрипторов, имеют более высокую частоту. Для методических и других прикладных задач можно использовать данный вывод, отбирая по частотному словарю семантических множителей списки наиболее частотных компонентов. Во-вторых, одинаковая распределенность компонентов в начале и в конце их частотного списка показывает известную независимость смыслового критерия, по которому проходил отбор дескрипторов, от статистических закономерностей.

Статистический анализ словаря позволяет также сделать вывод как о степени полноты семантического ядра русского языка, выбранного для анализа, так и о возможности предсказания новых компонентов, входящих в это ядро. Семантические компоненты с нулевой частотой могут быть использованы при уточнении и изменении состава дескрипторов и слов.

Семантические компоненты служат основой дефиниционной связи слов с дескрипторами. Наличие сведений о частоте компонентов позволяет применять математические методы при разработке критерия семантической близости слов с дескрипторами. Методика оценки силы семантической связи между дескрипторами и словами путем определения количественных характеристик — по имеющейся статистической информации — закодированных определений дескриптора и слова открывает определенные перспективы для объективного исследования текста мета-языка словаря.

Большое значение для устранения отдельных недостатков специальных тезаурусов и дальнейшего совершенствования структуры и эффективности их применения имеет работа по машинному построению и анализу русского тезауруса. При подготовке социологического тезауруса можно будет использовать методику, разработанную для тезауруса естественного языка. Решение этой задачи будет означать серьезный прогресс этой важной отрасли общественных наук и даст возможность расширить автоматизацию социологических исследований.

Возьмем несколько примеров установления семантической отнесенно-

сти слов к дескрипторам из списка предварительного распределения дескрипторов по словам, т. е. обратного распределения.

Так, слова *абсолютный* попало в поле дескриптора АБСОЛЮТНЫЙ по девяти семантическим компонентам (абсол, безотно, безусл, внесрав, неотн, несрав, полн, соверше), в поле СЛАВА — по одному (безусл), в ОБЯЗАТЕЛЬНЫЙ — по одному (безусл), ПРОКЛЯТЬЕ — по двум (безусл, полн), ОСОБЕННОСТЬ — по одному (независ), СУДЬБА — по одному (независ), НЕЗАВИСИМЫЙ — по одному (независ), НАПОЛНЯТЬ — по одному (полн), СОВЕРШЕНСТВО — по двум (полн, соверш), ЗАВЕРШЕННОСТЬ — по двум (полн, соверше), ЗАВЕРШИТЬ — по двум (полн, соверш), РОСТ — по одному (соверше). Интуитивная отнесенность этого слова к дескрипторам при более чем одном совпадающем семантическом компоненте — полная для дескрипторов АБСОЛЮТНЫЙ, ЗАВЕРШЕННОСТЬ, СОВЕРШЕННОСТЬ, тогда как для дескриптора ПРОКЛЯТЬЕ такой отнесенности нет. При совпадении только по одному семантическому компоненту имеется частичная отнесенность этого слова к дескриптору ОБЯЗАТЕЛЬНЫЙ.

Частотный анализ позволяет определить вес семантического компонента по информационному массиву дескрипторов. Рассмотрим условия связи слова с дескриптором с учетом не только количества совпадающих семантических компонентов, но и их частоты.

1. Совпадение только одного семантического компонента у дескриптора и слова дает примерно 95% отнесенных к дескрипторам слов.

2. При совпадении одного семантического компонента с частотой меньше пяти независимо от того, есть ли другие совпадения и сколь высокочастотны они: например, слово *абсолютный* попадает в дескриптор ПРОКЛЯТИЕ; слово *авиация* не попадает в дескрипторы САМОЛЕТ и КРЫЛО, но попадает в дескрипторы ПРОБЛЕМА, БУХГАЛТЕРИЯ; слово *автомат* не попадает в дескриптор ОГНЕСТРЕЛЬНОЕ. Слова *автор* и *авторучка* не попадают ни в один дескриптор; *агрессия* не попадает в ЗАВОЕВАНИЕ, НАПАДЕНИЕ, ПРИНУЖДЕНИЕ и ВМЕШАТЕЛЬСТВО.

3. При совпадении одного семантического компонента с частотой меньше пяти, но при условии еще хотя бы одного или больше совпадений со сколь угодно большой частотой: слово *авторитет*, например, попадает в дескриптор ДИСКРЕДИТАЦИЯ.

4. Два общих семантических компонента и больше с частотой меньше пяти: слово *агроном* не попадает в дескрипторы КОЛХОЗ и ЗЕМЛЕДЕЛИЕ, но попадает в СЕЛЬСКОХОЗЯЙСТВЕННЫЙ.

5. Один семантический компонент с частотой меньше шести, независимо от того, есть ли другие совпадения: например, слово *административный* не попадает ни в один дескриптор.

6. Один семантический компонент с частотой меньше семи, независимо от других совпадений: слово *алюминий* попадает в дескрипторы МОЛОТОК, КУЗНИЦА и МЕДЬ. Слово *академия* при всех рассмотренных условиях связи не попадает ни в один дескриптор. Оно распределяется по дескрипторам только при частоте совпадающих компонентов больше десяти.

Для устранения этих противоречий при установлении семантической отнесенности слов к дескрипторам можно опираться на «доминирующий

компонент» в дефиниции слова, т. е. надо учитывать ранг семантического компонента внутри определения данного слова. Если отказаться от понятия определяющего, или доминирующего, компонента и рассматривать все семантические компоненты, из которых складывается определение данного слова, как равноправные, то возникает трудность при распределении слова по дескрипторам.

Это было видно при рассмотрении условий связи слов с дескрипторами. Так, например, в слове *авиация* большой вес имеют компоненты: авиа, шипа, возд, лет, на основании которых оно относится к дескрипторам АВИАЦИЯ, КРЫЛО, САМОЛЕТ, ЛЕТАТЬ, а компоненты прак, теор, по которым *авиация* попадает в ПРОБЛЕМА и БУХГАЛТЕРИЯ, оказываются второстепенными. На основании этого можно дать оценку числу и составу семантических компонентов у дескрипторов и слов, которая и свою очередь характеризует их дефиниции. Таким образом, ранжирование семантических компонентов внутри каждой дефиниции есть предпосылка более точного распределения слов по дескрипторам. Это условие, а также количество совпадающих компонентов у слова и дескриптора, как и их частоты, являются в конечном счете основой системной организации словаря.

Второй решаемой в нашем исследовании задачей с широким применением статистических параметров текста является определение коэффициентов подобия между двумя дефинициями. Основная гипотеза, которая лежит в основе этой задачи, состоит в том, что если в двух дефинициях используются одни и те же семантические компоненты, то эти дефиниции принадлежат к одной области понятий; или отражают один когнитивный феномен объективной реальности. Можно выделить две функции, которые выполняет эта гипотеза в процессе построения тезауруса русского языка.

Во-первых, выделяется гносеологическая функция, состоявшая в том, что, руководствуясь этой гипотезой, по формальным статистическим критериям выделяется класс дефиниций, принадлежавший одной смысловой области.

Во-вторых, выделяется рефлексивная функция, нацеленная на поиск таких дефиниций, в которых явно выражено некорректное использование семантических компонентов.

При реализации этих функций на этапе анализа конкретного эмпирического материала мы учитывали как семантическую многозначность понятий, так и тот факт, что «информация, содержащаяся в понятиях более высоких уровней рефлексии, не исчерпывается информацией, содержащейся в тех понятиях, через которые даются определения и пояснения, и накопленная таким образом новая информация и делает возможным уточнить через «выводимые» понятия сами первоисходные «выводящие» понятия»<sup>12</sup>.

Здесь уместно отметить разницу между определением понятия конкретного явления в конкретной области знания в процессе конструирования теории и определением понятия в процессе создания общерусского словаря. В процессе конструирования теории определяемое понятие прежде всего ориентировано на более глубокий смысл передачи содержания того или иного аспекта исследуемого объекта. В процессе же включения

<sup>12</sup> См.: Частотный словарь русского языка. М., 1977.

<sup>13</sup> Мельников Г. П. Системология и языковые аспекты кибернетики. М., 1979, с. 34.

понятий в общерусский словарь основная задача состоит в том, чтобы это понятие вписалось в языковую систему. Именно на реализацию этого аспекта, на эту специфику ориентирована методика машинного анализа социологического словаря.

Связь между двумя понятиями устанавливается с помощью коэффициентов подобия. Коэффициенты образуют класс искусственных показателей связи, используемых для определения близости между двумя текстами (предложениями). Учитывая разнообразие в структуре и содержании элементов данного текста, правомерно предположить, что будет определенное разнообразие коэффициентов, отражающих связь между двумя текстами.

Один из первых таких коэффициентов был введен Т. В. Трофимовой и А. Н. Попескул<sup>14</sup>. Этот коэффициент имеет следующий вид:

$$S_{x_i x_j} = \frac{K}{M},$$

где  $K$  — количество совпадений по значению значимых слов в предложениях  $x_i$  и  $x_j$ ;  $M$  — минимальное значение количества значимых слов в предложениях  $x_i$  и  $x_j$ .

Значение  $S_{x_i x_j}$  может меняться от 0 до 1. При этом значение 1 коэффициент будет иметь в том случае, когда определение, содержащее меньшее число значимых слов, будет покрываться определением, содержащим большее число значимых слов. Т. е. если  $A$  — множество значимых слов первого понятия, а  $B$  — множество значимых слов второго понятия и при этом  $A \supset B$ , то, следовательно,  $A \supset B$ , но  $B \not\subset A$ .

Таким образом, это подобие не тождественно при коэффициенте, равном единице. Тем не менее этот коэффициент позволяет выделить класс таких множеств понятий  $B$ , который полностью включается в класс понятий  $A$ . Этот коэффициент при машинной реализации дает значительный объем информации для определения принадлежности, например, слов к дескрипторам.

Второй коэффициент, введенный нами, также меняется от 0 до 1. Его отличие от коэффициента  $S_{x_i x_j}$  состоит в том, что при значении коэффициента  $R_{x_i x_j}$  равном единице, два понятия тождественны по значимым словам.

$$R_{x_i x_j} = \frac{K}{M_i + M_j - K},$$

где  $K$  — количество совпадений по значению значимых слов в предложениях  $x_i$  и  $x_j$ ;  $M_i$ ,  $M_j$  — количество значимых слов в предложениях  $x_i$  и  $x_j$ .

Как легко можно заметить, знаменатель в нашем коэффициенте содержит меру разнообразия слов. Введенный коэффициент показал высокую степень смысловой близости понятий, попадающих в один класс по формально-лексическим признакам. В то же время встречается значительная группа дефиниций, которые попадают формально в данный класс, но по смыслу к нему не принадлежат. Как правило, в данном

случае легко обнаруживаются ошибки в дефиниции, среди которых наиболее часто встречаются неправильное словоупотребление, суживание понятий, в понятии выпячена одна, пусть даже характерная, черта и т. д. Уже при решении нашей задачи с помощью ЭВМ было просмотрено и осуществлено несколько сот тысяч попарных сравнений дефиниций, что позволило обнаружить несколько сот ошибок в определениях, уточнить состав дескрипторного словаря и каталога ключевых слов.

Рассмотрим распределения слов тезауруса русского языка для двух дескрипторов АБСОЛЮТНЫЙ и АВИАЦИЯ. Эти распределения получены на основании следующего критерия отнесенности слов и дескрипторов. Слово и дескриптор считаются семантически связанными, если у них имеются два или больше совпадающих семантических множителя, или если имеется один совпадающий множитель, но частота его меньше или равна шести. В приведенном ниже примере слова у дескрипторов АБСОЛЮТНЫЙ и АВИАЦИЯ расположены с учетом коэффициента  $P$ . Как видно из полученных данных, предложенный коэффициент отражает, на наш взгляд, смысловую близость дескриптора и слова. С увеличением  $P$  увеличивается и их смысловая близость. Коэффициент  $P$  принимает значение, равное 1, тогда, когда у дескриптора и слова имеется одинаковое количество семантических множителей и все они совпадают. Так, например, дескриптор АБСОЛЮТНЫЙ имеет в своем определении девять множителей: абсол, безуслов, независ, внесрав, соверше, полн, безотно, неотн, несрав. Слово *абсолютный* в своем определении имеет те же девять компонентов. В то же время коэффициент  $P$  для дескриптора *авиация* и слова *авиация* равен 0,7. Это происходит вследствие того, что в определение дескриптора АВИАЦИЯ входят семь семантических множителей: возд, средс, передви, флот, лет, аппа, авиа, а у слова *авиация* девять множителей: авиа, возд, средс, передви, флот, теор, прак, лет, аппа.

Следовательно, количественный и качественный состав дефиниций слов влияет на положение их в дескрипторной статье. Проведенный анализ этого состава позволил уточнить определения слов. Из приведенных примеров, а также из остальных распределений видно, что можно выбрать такое пороговое значение коэффициента  $P$  для данного поля, при меньших значениях которого можно считать, что нет смысловой близости между дескриптором и словом.

#### Дескриптор АБСОЛЮТНЫЙ

Слова	$P$	Слова	$P$
<i>абсолютный</i>	1,0000000	<i>совсем</i>	0,0833333
<i>вполне</i>	0,2000000	<i>независимость</i>	0,0765230
<i>совершенно</i>	0,1818181	<i>посторонний</i>	0,0765230
<i>безусловно</i>	0,1666666	<i>ровно</i>	0,0765230
<i>совершенство</i>	0,1333333	<i>кругом</i>	0,0666666
<i>точный</i>	0,1176470	<i>обязанность</i>	0,0666666
<i>круглый</i>	0,1111111	<i>прямой</i>	0,0625000
<i>зрелый</i>	0,1000000	<i>обязательный</i>	0,0586235
<i>особый</i>	0,0909090	<i>определенный</i>	0,0586235
<i>положительный</i>	0,0865565	<i>физкультура</i>	0,0586235
<i>независимый</i>	0,0833333	<i>точно</i>	0,0555555

<sup>14</sup> Попескул А. Н., Трофимова Т. В. Распльвчатые множества и распознавание смысла текста. — В сб.: Автоматическая переработка текста методами прикладной лингвистики. Кишинев, 1977, с. 5.

оплодотворение	0,0526315	развитие	0,0416666
обязательство	0,0526315	рост	0,0416666
ровный	0,0476190	развиться	0,0370370
самостоятельный	0,0476190	расти	0,0344827
материализм	0,0454545	строгий	0,0344827
материя	0,0434782	чистый	0,0294117
объект	0,0434782	развить	0,0270270
судьба	0,0434782		

#### Дескриптор АВИАЦИЯ

Слова	P	Слова	P
авиация	0,7777777	офицер	0,0833333
летать	0,3750000	подводник	0,0833333
планер	0,3333333	хвост	0,0833333
самолет	0,3000000	лейтенант	0,0765230
летчик	0,2222222	магнитофон	0,0765230
пилот	0,2222222	патефон	0,0765230
полет	0,2222222	снаряд	0,0765230
сообщение	0,2000000	камера	0,0740740
воздушный	0,1818181	оторваться	0,0740740
планировать	0,1818181	звонить	0,0714285
лететь	0,1764705	передатчик	0,0714285
приземлиться	0,1666666	дорожка	0,0625000
вертолет	0,1536461	крыло	0,0625000
корабль	0,1536461	капитан	0,0625000
везти	0,1426571	механизм	0,0625000
ездить	0,1426571	печатать	0,0625000
ехать	0,1426571	прибор	0,0625000
потолок	0,1250000	флот	0,0586255
фотоаппарат	0,1250000	автомат	0,0555555
фотографировать	0,1111111	телефон	0,0555555
посадка	0,1052631	записать	0,0526315
транспорт	0,1052631	магазин	0,0526315
поход	0,1000000	приемник	0,0526315
ракета	0,0952380	пушка	0,0526315
дорога	0,0909090	трубка	0,0476190
пылесос	0,0909090	спутник	0,0454545
матрос	0,0833333	аппарат	0,0434782
моряк	0,0833333	касса	0,0357142

## Глава IV

### МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ЭКСПЕРИМЕНТА

#### § 1. КОНЦЕПЦИЯ МАТЕМАТИЧЕСКОЙ И ИНФОРМАЦИОННОЙ БАЗЫ ЭКСПЕРИМЕНТА

Разработка машинной технологии и математического обеспечения проводилась на основе и в рамках концепции автоматизированной системы анализа информации, разрабатываемой в ИСИ АН СССР с 1971 г.<sup>1</sup> Под автоматизированной системой анализа информации (АСАИ) понимается интегрированный комплекс программ, обеспечивающих преобразование информации любым набором имеющихся программ в технологически допустимой последовательности. Структурно разрабатываемая система состоит из восьми подсистем.

Рассмотрим структуру АСАИ под углом зрения решения нашей задачи — построения «Русского семантического словаря».

1. Операционная подсистема выполняет следующие функции: ввод и контроль лингвистической информации (словарей, дескрипторов, каталогов слов и т. д.); обеспечение непрерывной обработки информации в соответствии с разработанной логико-информационной схемой решения задач по анализу словарной информации; обновление, корректировка и получение необходимых сведений о состоянии банка лингвистической информации и математического обеспечения.

2. Подсистема стандартизации<sup>2</sup> выполняет следующие основные функции: приведение исходных данных к единому универсальному виду с единым универсальным описанием их структуры.

Основными средствами, с помощью которых реализуются эти функции, служат программа редактирования, перекодирования, сортировки, отбор информации по заданному критерию и т. д. Так, например, согласно методическим разработкам, проведенным в рамках концепции АСАИ<sup>3</sup>, введенная в ЭВМ информация была на естественном языке. Работа с информацией такого уровня формализации не позволила эффективно использовать ЭВМ, усложняла процесс создания математического обеспечения, повышала затраты машинного времени. Исходя из этого, была разработана система кодирования словников дескриптора и слов каталогов дескрипторов, слов и семантических множителей, которая позволила значительно упростить технологию следующих преобразований. Кроме того, широко использовались процедуры корректировки каталогов и словников при устранении различного рода ошибок, исключения неполноценных и включения новых элементов. Для удобства последующего анализа был разработан стандартный документ.

<sup>1</sup> См. более подробно: Молчанов В. И. Применение ЭВМ в социологическом исследовании. — Социологические исследования, 1974, № 1; Он же. Социальная информация и управление предприятием. М., 1977; Молчанов В. И., Афанасьев В. А. Социальная статистика и пути автоматизации статистического анализа. — В кн.: Методологические и методические проблемы применения статистики в социологии. М., 1978.

<sup>2</sup> Более подробно см.: Афанасьев В. А., Величко А. Н., Молчанов В. И. Подсистема стандартизации информации. — В кн.: Анализ социологической информации с применением ЭВМ. М., 1976.

<sup>3</sup> Молчанов В. И. Система кодирования социологической информации. — В кн.: Анализ социологической информации с применением ЭВМ, ч. 1. М., 1973.



Под документом понимается как дефиниция дескрипторов и слов, так и дескриптор и слово соответствующих каталогов. Стандартный документ — это форма представления данных с нормативной структурой, которая позволяет осуществлять анализ статистическими и логическими методами, абстрагируясь на этом этапе от конкретного содержания.

В подсистеме предусмотрена возможность изменения содержания стандартного документа, а также формирование рабочих массивов для решения конкретных задач. Изменения в содержании стандартного документа могут относиться как ко всему документу, так и к отдельным его признакам (например, составу семантических множителей). Могут быть изменения в самих множителях (например, частота множителя в словаре дескрипторов).

При построении рабочих массивов документы могут отбираться по всем структурным элементам. Так, например, при работе с дефинициями дескрипторов отбор осуществлялся по самому дескриптору, по семантическим множителям, входящим в дескриптор, по частоте семантических множителей, входящих в данный дескриптор, и т. д.

3. Подсистема статистического анализа. В рамках этой подсистемы разработан ряд программ, позволяющих получить статистическую оценку некоторых параметров словарной информации. Так, например, были получены частотные распределения семантических множителей по словарю дескрипторов и словнику. Эти частоты, как уже было отмечено, использовались при определении критерия информативности семантических множителей при установлении взаимосвязи между дескрипторами и словами. Для определения тесноты взаимосвязи между словом и дескриптором использовался предложенный авторами коэффициент подобия.

4. Подсистема контент-анализа. В рамках этой подсистемы осуществляется разработка ряда логико-смысловых методов анализа текстовой информации. Одним из наиболее распространенных методов анализа текстовой информации, к которой относятся и исходные дефиниции дескрипторов и слов, является метод контент-анализа.

Суть метода контент-анализа состоит в следующем. Выделяется единица анализа — сегмент текстового материала, подлежащий изучению. Строится словарь категорий, в рамках которых материал формализуется. Категории операционализируются. Кроме единицы анализа выделяются единицы счета — операциональные признаки категорий словаря. В рамках одной единицы анализа категория подсчитывается не более одного раза. После того как текст разбит на единицы анализа и последние категоризованы, подсчитываются частоты вхождения в текст категорий. Как правило, берутся относительные частоты. Кроме того, производится индексирование текста, основанное на совместных вхождениях категорий. Таким образом, основная функция подсистемы контент-анализа состоит в приведении информации, содержащейся в неформализованном виде, к виду, удобному для последующего анализа.

Авторами разрабатывается системно-целевой метод. Суть этого метода состоит в том, что при кодировании информации с помощью семантических множителей кодируется не только конкретное содержание дефиниций, но ряд системных свойств данного понятия, которое включено в систему понятий как отдельной области человеческого знания. так и

языка в целом. Одним из важных средств вычленения системных параметров является научная картина мира, лексикографическая параметризация языка, логико-математический анализ лингвистических структур.

В соответствии с этими принципами технология обработки словаря на ЭВМ включает следующие этапы:

1. Перфорация данных на перфокартах.
2. Формирование каталогов в памяти ЭВМ, включающее операции ввода каталогов в машину, распечатку, контроль и корректирование.
3. Формирование информационных массивов дескрипторов и слов, включающее операции ввода в машину, перекодирование, выявление ошибок, исправление ошибок и распечатку массивов.
4. Получение частот семантических множителей по информационным массивам.
5. Формирование словарной статьи частотного словаря и ее печать.
6. Формирование дескрипторной статьи в цифрах. Декодирование шифров дескрипторов, слов и семантических множителей. Придание статье заданной формы и вывод на печать.

Рассмотрим вопросы, связанные с подготовкой машинных носителей информации. Необходимо было выполнить большой объем перфорационных работ по следующим массивам: каталог семантических множителей, каталог дескрипторов, перечень дескрипторов с их определениями (семантическими множителями) и перечень слов с их определениями. Как известно, высокая точность подготовки данных на машинных носителях требует повышенных затрат труда. Затраты начинают быстро расти после превышения 95% точности. Этап подготовки машинных носителей является только одним из этапов, на котором возможно нарушение точности формирования информационного массива. Учитывая все сказанное выше, мы решили снизить трудоемкость формирования массивов на магнитной ленте путем использования ЭВМ для контроля правильности перфорации исходных данных и корректировки выявленных ошибок непосредственно на магнитной ленте по мере их выявления. При этом, конечно, часть ошибок оставалась, но они, как и ошибки кодирования, были убраны на последующих этапах применения коэффициентов и критериев. Большую роль в механизации контроля подготовленных данных сыграло использование каталогов.

Каталоги содержат расположенное в заданной последовательности описание определенных элементов информационного массива. Например, каталог слов содержит все слова из массива словника в виде буквенного состава и цифрового шифра слова. При решении наших задач эффективнее оперировать в ЭВМ словами (дескрипторами, семантическими множителями), представленными не в буквенном виде, а в виде упрощенных цифровых шифров. Каталоги составлены для следующих элементов информационных массивов: дескрипторов, слов и семантических множителей. Элементы в каталоге упорядочены в зависимости от решаемой задачи или по возрастанию шифров, или по алфавиту, или иным способом.

Формирование каталогов в ЭВМ проводится специальной машинной программой. Каталог, подготовленный на перфокартах, вводится в машину и подвергается в ней всестороннему контролю на наличие в тексте элементов только букв, на возрастание шифров элементов и нахождение шифров в заданном числовом интервале. Текст каталога выводится на

печать и досконально сверяется с исходным текстом. Обнаруженные ошибки исправляются. Большое последующее значение каталогов оправдывает столь тщательную их подготовку.

Каталоги для последующей работы формируются на магнитной ленте в виде массивов документов — например, массив документов каталога дескрипторов, или проще — каталог дескрипторов. Одним документом каталога описывается один элемент структуры информационного массива. Он указывает: шифр элемента, буквенный состав элемента, номер области, к которой он относится, его частоту на массиве словника и частоту на массиве дескрипторов. Форма документов всех каталогов типовая. Разработанные программы позволяют корректировать каталоги, т. е. дополнять или убирать документы из каталога, выводить каталоги на печать, переформировывать каталоги, т. е. по-иному упорядочивать элементы внутри каталога, например, по частоте на массиве дескрипторов.

Следующим этапом является формирование информационных массивов слов и дескрипторов с относящимися к ним семантическими множителями в памяти машины. Ввиду большого объема данных работа разбита на несколько частей и выполняется несколькими программами. Вначале данные с перфокарт вводим в машину без изменений и записываем на магнитную ленту. Структура информации на перфокартах такова: шифр слова (или дескриптора), номер области, буквенные составы семантических множителей, относящихся к данному слову (дескриптору). Семантические множители отделены один от другого пробелами. В конце группы ставится разделитель. На магнитной ленте каждая группа оформляется в виде одного документа. В программе ввода предусмотрен контроль: на наличие шифра дескриптора или слова по соответствующему каталогу, на наличие не букв в текстах семантических множителей. Обнаруженные ошибки выводятся на печать, а ошибочные компоненты отбрасываются. После анализа ошибок специальная программа позволяет внести в массив необходимые исправления. Сформированные таким образом информационные массивы становятся базовыми для всей последующей обработки. Следующая программа проводит замену текстов семантических множителей на шифры по каталогу семантических множителей. Эта процедура позволяет выявить в массиве наличие множителей, отсутствующих в каталоге, а также множители, в текстах которых при перфорации данных были допущены ошибки. Все это позволяет провести дальнейшую корректировку массива, а для выявления ошибок использовать машину.

## § 2. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

### Программа ВВКАТ

(Формирование каталогов)

Предназначена для ввода каталогов, подготовленных на перфокартах, и формирования их на магнитной ленте в стандартной форме. Форма документов каталогов дескрипторов и обрубков на перфокартах представлена в отношении форм документов. Каталог слов формируется из информационного массива «слова—семантические компоненты».

Процедура предусматривает следующие виды контроля:

1. Повторение шифра.
2. Возрастание шифров.
3. Наличие не букв в тексте элемента (слова, семантические компоненты, дескрипторы).

Проводится подсчет частот букв в текстах элементов. Тексты семантических компонентов, имеющие более 10 букв, урезаются до 10 букв.

### Работа с программой на ЭВМ

1. Установить МЛ (магнитную ленту) для записи каталога.
2. Подготовить перфокарту АБИМЯКТ

А — вид каталога

- = 1 каталог семантических компонентов,
- = 2 каталог слов,
- = 3 каталог дескрипторов.

Б — признак печати

- = 0 нет печати каталога,
- = 1 выполняется печать сформулированного документа.

ИМЯКТ — имя массива каталога при записи на МЛ.

3. При обнаружении ошибок происходит останов работы машины и печать вида ошибки на пишущей машине (ПМ) и АЦПУ. После выяснения ошибки работу можно продолжать.

В конце массива перфокарт каталога необходимо положить перфокарту, содержащую девятки в колонках с 1 по 20, что будет признаком конца массива перфокарт.

### Программа ВВИНФ

(Ввод информационных массивов)

Предназначена для ввода информационных массивов (дескрипторы с семантическими компонентами или слова с семантическими компонентами), подготовленных на перфокартах, и записи их на МЛ. (Форма документа, при которой семантические компоненты — в виде текстов). Тексты семантических компонентов на перфокарте имеют произвольное число символов. В результате работы программы в документе на МЛ тексты семантических компонентов имеют постоянную длину, а именно 10 символов (отсутствующие символы заполнены пробелами).

Процедура предусматривает следующие виды контроля:

1. Наличие шифра дескриптора или слова в соответствующем каталоге.
  2. Совпадение шифра дескриптора (слова) на всех перфокартах одного документа.
  3. Наличие не буквы в текстах семантических компонентов.
- При работе программы проводится подсчет частот обрубков в документах.

### Работа с программой на ЭВМ

1. Установить МЛ 1 с текстом каталога, соответствующего информационному массиву (каталог дескрипторов или слов).
2. Установить МЛ 2 для записи информационного массива.
3. Подготовить перфокарту АБИМЯКТИМЯЗП

А — вид контроля при вводе  
= 0 без контроля возрастания в массиве шифров дескрипторов (слов,  
= 1 с контролем возрастания.

Б — признак печати  
= 0 без печати документов,  
= 1 печать документа, записываемого на МЛ.  
ИМЯТ — имя считываемого каталога (дескрипторов или слов),  
ИМЯЗП — имя информационного массива, записываемого на МЛ.

4. При обнаружении ошибок происходит останов работы машины и печать вида ошибки на ПМ и АЦПУ. Документ, содержащий ошибку, исключается из обработки. После выяснения ошибки работу можно продолжить.

5. В конце массива перфокарт необходимо положить перфокарту, содержащую девятки в колонках с 1 по 20, что будет признаком конца массива перфокарт.

### Программа ПЕРЕК

(Перекодирование текстов семантических компонентов в шифры в информационных массивах)

Предназначена для замены текстов обрубков на шифры в документах, полученных процедурой ввода информационных массивов и записи новых документов на МЛ. Форма результирующего документа представлена в описании форм документов.

Процедура предусматривает следующие виды контроля:

1. Наличие семантических компонентов, отсутствующих в каталоге.
2. Совпадение шифров семантических компонентов в одном документе.

Ошибочные семантические компоненты исключаются из документа, о чем имеются сообщения на АЦПУ.

Работа с программой на ЭВМ

1. Установить МЛ 1 с текстом каталога семантических компонентов.
2. Установить МЛ 2 с текстом информационного массива.
3. Установить рабочую МЛ 3 для записи промежуточного массива.
4. Запись перекодированного массива можно проводить на МЛ 2, иначе установить МЛ 4 для записи.

5. Подготовить перфокарту ИМЯКТИМЯИНИМЯПРИМЯЗП

ИМЯКТ — имя каталога семантических компонентов,

ИМЯИН — имя информационного массива,

ИМЯПР — имя промежуточного массива,

ИМЯЗП — имя перекодированного массива.

6. Подготовить вторую перфокарту (во всех случаях одинаковую — ОТТТТ).

7. Последовательность указаний об установке магнитных лент:

- а) чтение каталога семантических компонентов,
- б) чтение информационного массива,
- в) запись промежуточного массива,
- г) чтение каталога семантических компонентов,
- д) чтение промежуточного массива,
- е) запись перекодированного массива.

### Программа ПЕЧКА

(Печать каталогов

и информационных массивов)

Предназначена для вывода на АЦПУ записанных на МЛ каталогов и информационных массивов. Используется также для печати частот, записанных в каталог. Процедура предусматривает контроль возрастания шифров в каталогах. При нарушении возрастания строка каталога отмечается звездочкой.

Работа с программой на ЭВМ

I. Режим печати каталогов.

1. Установить МЛ с текстом каталога,
2. Подготовить перфокарту АБИМЯКТ

А — вид каталога

- = 1 — каталог семантических компонентов,
- = 2 — каталог слов,
- = 3 — каталог дескрипторов.

Б — объем печатаемого каталога

= 0 печатать весь каталог

= 1 задаются границы печатаемого массива (между наименьшим и наибольшим шифрами)

Границы указываются на дополнительной перфокарте двумя пятизначными шифрами.

ИМЯКТ — имя массива печатаемого каталога.

II. Режим печати информационных массивов.

1. Установить МЛ с текстом информационного массива.
2. Установить МЛ с текстом каталога, соответствующего информационному массиву (каталог слов или дескрипторов).

3. Подготовить перфокарту АБИМЯИНИМЯКТ

А — вид печатаемого массива:

- = 4 — массив слова с семантическими компонентами,
- = 5 — массив дескриптора с семантическими компонентами,

Б — объем печатаемого массива:

- = 0 — печатать весь массив,
- = 1 — задаются границы печатаемого массива (по шифрам слов или семантических компонентов). Задание делается так же, как при печати каталогов.

ИМЯИН — имя печатаемого массива информации,

ИМЯКТ — имя массива каталога, соответствующего информационному массиву.

### Программа ФИКУС

Предназначена для подсчета частот семантических компонентов по информационным массивам дескрипторов и слов.

Работает в двух режимах:

- 1) Подсчет частот и вывод полученных частот на печать вместе с шифрами семантических компонентов каталога компонентов.

Программа ФИКУС разработана Н. И. Ростягаевой.

2) Подсчет частот и запись полученных частот в каталог семантических компонентов с записью каталога компонентов с частотами на магнитную ленту.

#### Работа с программой на ЭВМ

I. Подсчет частот и вывод частот на печать.

1) Установить МЛ с текстом каталога компонентов.

2) Установить МЛ с текстом информационного массива (дескрипторов или слов с компонентами).

3) Подготовить пакет перфокарт:

а) ИМЯКТИМЯИМ

ИМЯКТ — имя массива каталога компонентов,

ИМЯИН — имя информационного массива.

б) 1

II. Подсчет частот и запись каталога с частотами на МЛ.

1) Установить МЛ с текстом каталога компонентов.

2) Установить МЛ с текстом информационного массива.

3) Установить МЛ для записи сформированного массива.

4) Подготовить пакет перфокарт:

а) ИМЯКТИМЯИН

ИМЯКТ — имя каталога компонентов,

ИМЯИН — имя информационного массива (дескрипторов или слов с компонентами).

б) 2.

в) ИМЯКИ — имя массива для записи на МЛ (записывается каталог компонентов с занесенными в него частотами компонентов для дескрипторов или слов).

г) 1 если информационный массив — слова с компонентами,

2 если информационный массив — дескрипторы с компонентами.

**Примечание.** Для получения массива «Каталог компонентов с частотами компонентов для дескрипторов и слов» на первой перфокарте пакета следует указать имя каталога компонентов с записанными ранее частотами для одного из массивов.

#### Программа ОБМАС

Предназначена для объединения массивов информации, записанных по стандартной форме.

Исходные данные:

I. N массивов информации, которые надо объединить, записанные на МЛ.

#### Работа с программой на ЭВМ

1) Установить МЛ с массивами информации, которые надо объединить.

2) Установить МЛ для записи объединенного массива.

3) Подготовить пакет перфокарт:

а) НИМЯОМ

N — пятизначное значение количества объединяемых массивов,

ИМЯОМ — имя объединенного массива

б) Карта с именем 1-го объединяемого массива

в) Карта с именем 2-го объединяемого массива

г) Карта с именем N-го объединяемого массива

Все массивы имеют пятизначные обозначения.

#### Программа РОБУС

Предназначена для распределения семантических компонентов по убыванию их частот по массиву дескрипторов или слов. Во время работы программы компоненты, имеющие нулевую частоту, из распределения исключаются. Внутри каждой частотной группы сохраняется расположение компонентов по алфавиту.

#### Работа с программой на ЭВМ

1) Установить МЛ с текстом каталога компонентов с записанными в нем частотами.

2) Установить МЛ для записи формирующего массива.

3) Подготовить пакет перфокарт:

а) ИМЯКОИМЯКУ

ИМЯКО — имя массива каталога компонентов с частотами,

ИМЯКУ — имя массива каталога компонентов, распределенного по убыванию частот

б) NN

N — пятизначное значение макс частоты компонента у дескриптора или слова.

NI — 00001 для частот по массиву дескрипторов,

NI — 00003 для частот по массиву слов.

#### Программа РАСДО

Предназначена для получения массива семантических компонентов с относящимися к ним дескрипторами.

#### Работа с программой на ЭВМ

1) Установить МЛ с информационным массивом дескрипторов с относящимися к ним семантическими компонентами.

2) Установить Л с текстом каталога семантических компонентов.

3) Установить МЛ для записи формирующегося массива.

4) Подготовить пакет перфокарт: ИМЯИМИЯККИМЯФМ

ИМЯИМ — имя информационного массива,

ИМЯКК — имя каталога семантических компонентов,

ИМЯФМ — имя формирующегося массива.

#### Программа РАСДП

Предназначена для получения промежуточного массива слов с относящимися к ним дескрипторами.

#### Работа с программой на ЭВМ

1) Установить МЛ с массивом семантических компонентов с относящимися к ним семантическими компонентами.

2) Установить МЛ с информационным массивом слов с относящимися к ним семантическими компонентами.

3) Установить МЛ для записи формирующегося массива.

4) Подготовить пакет перфокарт: ИМЯМКИМЯИМИЯМФМ

ИМЯМК — имя массива семантических компонентов с относящимися к ним дескрипторами,

ИМЯИМ — имя информационного массива слов с относящимися к ним семантическими компонентами,

ИМЯФМ — имя формирующегося массива.

**Программа РАСДС**

Предназначена для распределения дескрипторов по словам с подсчетом количества совпадающих компонентов у дескриптора и слова, а также для записи полученного массива на МЛ.

**Работа с программой на ЭВМ**

1. Установить МЛ с промежуточным массивом слов с относящимися к ним дескрипторами.

2. Установить МЛ для записи формулирующего массива слов с относящимися к ним дескрипторами с учетом количества совпадающих компонентов у дескрипторов и слов и самих совпадающих компонентов.

3. Подготовить пакет перфокарт:

а) ИМЯПМИМЯФН

ИМЯПМ — имя промежуточного массива,

ИМЯФМ — имя сформированного массива

б) N1, N2

N1 — пятисимвольное значение шифра первого слова,

N2 — пятисимвольное значение шифра последнего слова.

**Программа СОРОВ**

Предназначена для сортировки семантических множителей по возрастанию шифров в группах дескрипторов-компонентов и словокомпонентов в информационных массивах.

**Программа РАССД**

Предназначена для распределения слов по дескрипторам. Исходной информацией для этой программы является массив информации, полученный после применения программы РАСДС.

**Программа ПОИСК**

Проводит, осуществляет, обеспечивает нахождение необходимого шифра в каталогах дескрипторов, слов и семантических множителей.

**Программа ПЕКНИ**

Предназначена для печати частотного словаря семантических множителей.

**Программа ДЕСШИ**

Предназначена для определения общих семантических множителей для дескриптора и слов к нему относящихся.

**Программа СОРДО**

Предназначена для сортировки документов информационного массива по шифру дескриптора (слова).

**Программа ПАРЧС**

Предназначена для формирования следующих параметров: вычисление числа слов в статье дескриптора и числа семантических множителей в дефиниции слова.

**Программа СОРДЧ**

Предназначена для сортировки семантических множителей в дефиниции дескриптора по возрастанию частот множителей в массиве дескрипторов.

**Программа ПЕЧСЛ**

Предназначена для печати текста словаря. Осуществляет декодирование шифров дескрипторов, слов и семантических множителей. Комплекует статью дескриптора в соответствии с заданной формой. Печатает текст словаря по этой форме.

## ОГЛАВЛЕНИЕ

Введение . . . . .	3
Глава I	
<b>Основные направления использования ЭВМ в лексикографии</b>	
§ 1. Лексикографическая параметризация языка как тенденция в современном словаростроении и как объективная предпосылка его автоматизации . . . . .	5
§ 2. Автоматическая компиляция параметров и создание банков лексикографических данных . . . . .	13
§ 3. Машинный перевод и аналитические словарные параметры . . . .	22
§ 4. Лексикографические аспекты искусственного интеллекта . . . .	31
Глава II	
<b>Автоматическое построение русского тезауруса как способ семантического анализа метаязыка толкового словаря</b>	
§ 1. Лингвистическая технология конструирования тезауруса и общая характеристика результата . . . . .	40
§ 2. Семантическая связь и содержательный смысл селективных критериев . . . . .	50
Глава III	
<b>Методы анализа метаязыка словаря</b>	
§ 1. Системный подход к анализу метаязыка словаря . . . . .	68
§ 2. Статистический анализ . . . . .	72
Глава IV	
<b>Математическое обеспечение эксперимента</b>	
§ 1. Концепция математической и информационной базы эксперимента . .	85
§ 2. Программное обеспечение . . . . .	88

## АНАЛИЗ МЕТАЯЗЫКА СЛОВАРЯ С ПОМОЩЬЮ ЭВМ

*Утверждено к печати  
Научным советом по лексикологии и лексикографии  
Институтом языкознания*

Редактор издательства *Н.Н. Барская*  
Художник *И.Е. Сайко*. Художественный редактор *Т.П. Поленова*  
Технический редактор *Н.М. Петракова*. Корректор *О.А. Разуменко*

ИБ № 25039

Подписано к печати 27.07.82. Формат 60 x 90 1/16. Бумага офсетная №2  
Гарнитура литературная (фотонабор). Набор изготовлен  
в типографии №2 издательства "Наука"  
Печать офсетная. Усл.печ.л. 6,0. Усл.кр.-отт. 6,3. Уч.изд.л. 7,5  
Тираж 5000 экз. Тип.зак. 1708. Цена 70 коп.

Издательство "Наука", 117864 ГСП-7, Москва В-485, Профсоюзная ул., д. 90  
Ордена Трудового Красного Знамени 1-я типография издательства "Наука"  
199034, Ленинград В-34, 9-я линия, 12