ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1999 год Выпуск 166

УДК 519.769:801.54

КОЛИЧЕСТВЕННЫЕ ИССЛЕДОВАНИЯ МОРФЕМНОЙ СТРУКТУРЫ СЛОВ РУССКОГО ЯЗЫКА[1]

(на базе электронного словаря Д. Уорта)

Н.В. Саломатина

Введение

Известные способы представления в оперативной памяти словарей в различных системах, автоматически анализирующих тексты на русском языке, опираются не только на алгоритмические приемы(например,[1]), но и используют чисто языковые закономерности. Примерами подобного рода (использования закономерностей словоизменения в русском языке) могут служить работы [2, 3]. В памяти хранится лишь основа слова (слово без окончания) и информация о ее грамматическом классе. Номер класса, которому соответствует определенный набор окончаний, задает всю парадигму слова. Считается, что представление словарей в виде основ слов требует для хранения существенно меньше памяти (в несколько раз), чем пословное.

Следующим естественным шагом на пути привлечения регулярно воспроизводимых языковых единиц для компактного представления в оперативной памяти словарей является использование морфем. Речь идет, как и в [1–3], о словарях, статья (или запись) в которых сводится к цепочке символов алфавита

[1]Работа поддержана грантом РГНФ № 99–04–12026в

переменной, но ограниченной длины (несколько десятков символов).

Изучение морфемного строения слов языковедами нашло свое практическое выражение в создании различных словарей морфем. Например, словарь А.И. Кузнецовой и Т.Ф. Ефремовой [4] содержит не только инвентарь морфем, но и морфемные модели с указанием числа слов, которые им удовлетворяют (если число слов не меньше ста). Под морфемными моделями понимается представление слова в виде цепочки морфов, в которой корневой морф унифицирован, т.е. заменен на определенный символ. Использование словаря [4] для сжатия информации затруднено в силу того, что хотя членение на морфемы проводилось по формальному признаку (критерию структуры слова), а не семантическому, все же само членение осуществлялось вручную с учетом словообразовательного аспекта. Из-за этого возникает некоторая нерегулярность в членении. Например, в словах "озеленительный" и "обременительный" выделяются, соответственно, суффиксальные цепочки "-ен-и-тел-ь-н-" и "-ен-и-тельн-". Для автоматического анализа предпочтительнее было бы иметь единообразное представление слов на морфемном уровне и сведения о морфемных моделях, полученных на словаре с автоматическим членением слов на морфы.

Материалом, подходящим для получения необходимой информации, может служить словообразовательный словарь Д. Уорта [5]. Кроме того что слова в нем разбиты на морфемы автоматически, его объем существенно превосходит[4]: по числу слов в два раза, по числу морфем более, чем в два. Только корневых морфов в нем 13 тыс., тогда как [4] содержит всего порядка 5 тыс. морфем (корневых и аффиксальных). Некоторые существенные недостатки, за которые словарь [5] подвергался справедливой критике, были устранены при вводе его в компьютер канд. филол. наук Юдиной Л.С. Внесенные исправления касались, в частности, перераспределения слов с омонимичными корнями по соответствующим их семантике гнездам. Кроме того, все слова были дополнены частеречными характеристиками. Собственно морфемные модели слов в словаре не даны, однако легко могут быть из него получены. В данной работе исследуются

количественные характеристики морфемных моделей слов, извлеченных из словаря Д. Уорта, включая многокорневые структуры, не рассматривавшиеся в [4].

Задача выявления морфемных моделей слов

Информативная часть морфемных моделей слов по сути представляет собой цепочки аффиксов ("...выделенных в словоформе морфем — префиксов, инфиксов, суффиксов, — видоизменяющих значение остальной части слова" [6]) с окончаниями (включая нулевое). По данным [1] насчитывается около сотни разных префиксов и более пятисот суффиксов, встречающихся в определенных сочетаниях друг с другом и с корневыми морфами. Все многообразие слов словаря легко задать, таким образом, множеством корней и всевозможных цепочек обрамляющих их аффиксов, установив соответствующие ссылки–связи между корнями и аффиксальными цепочками. Если заменить в слове корневой морф на любой символ, например "*R*", то полученное обобщенное представление слова и будет являться его морфемной моделью (согласно терминологии, принятой в [4]). В этом случае слова w_1 = "наброшенный", w_2 = "налаженный", w_3 = "нахмуренный" и т.п. будут иметь один и тот же вид: m ="на–*R*–е–нн–ый" с константной аффиксальной цепочкой (аранжировкой) и переменной *R*, принимающей значения "брош", "лаж", "хмур". В сложных словах дополнительные корни не замещаются символом "*R*", а выписываются полностью в скобках: "(сам)о–*R*–н–ый" (*R* = "ход", например).

Совокупность цепочек (морфемных моделей) $M = \{m_i\}$, полученных при записи всех слов словаря русского языка (достаточно большого объема) в указанном виде, будем считать допустимой. Исследование M (множества допустимых цепочек или моделей) позволит продемонстрировать, каким образом осуществляется аффиксальная аранжировка корней в русском языке, и оценить возможность компактизации словаря за счет представления слов в оперативной памяти в виде морфемных структур.

Целью данной работы является описание количественных характеристик построенного множества допустимых цепочек m_i из M и выявленных закономерностей, которые могут быть полезны для компактного представления словарей.

Способ представления допустимых цепочек

Основные трудности при решении поставленной задачи связаны лишь с тем, что объем хранимых в оперативной памяти данных значителен. Для получения достаточно полного множества допустимых цепочек необходим довольно большой словарь. В нашем случае он состоял из более чем 100 тыс. слов. Каждому слову словаря сопоставляется его обобщенное представление в виде цепочки m_i. Цепочки, в свою очередь, упаковываются в бинарное дерево T [7]. В узлах дерева размещаются морфемы, составляющие цепочку m_i, и символ, замещающий корень. В листьях хранится информация о числе слов, описываемых m_i (разнокоренным словам может соответствовать одна и та же аффиксальная последовательность). Число слов, имеющих один и тот же обобщенный вид m_i, будем называть частотой встречаемости m_i и обозначать $f(m_i)$.

Изначально дерево пусто, т.е. имеет один выделенный узел T, называемый корнем дерева. Слово, представленное в виде модели m_i, путем прямого обхода дерева T (от корня T к листьям сначала через левое поддерево, затем через правое) поморфемно сравнивается с содержимым узлов дерева. В случае несовпадения морфа из m_i и узлового морфа дерева осуществляется переход на правый узел T. В случае отсутствия соответствующего

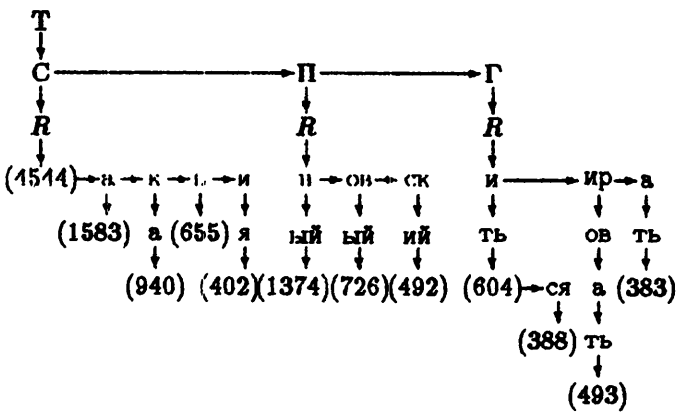

Рис. 1.

морфу из m_i узла в T производится его достройка. Таким образом, в результате рассмотрения всех слов словаря в виде m_i и поморфемного сравнения элементов m_i с содержимым узлов T, строится дерево, включающее все допустимые морфемные модели слов, а полный обход T позволяет их выписать.

Использованная традиционная структура представления данных в виде дерева позволяет компактно хранить в памяти все цепочки и достаточно быстро определять, существует ли цепочка, соответствующая анализируемому слову, в уже имеющемся списке и легко дополнять этот список в случае отсутствия в нем такой цепочки. Фрагмент дерева T, содержащий самые часто встречающиеся m_i, представлен выше на рис. 1. Стрелки, ориентированные сверху вниз, соответствуют левой ветви дерева, а слева направо — правой ветви. Список цепочек m_i, на которых построен фрагмент T, приведен далее, в табл. 3 следующего раздела. Цепочки начинаются с идентификатора части речи (заглавные буквы). Номенклатура идентификаторов представлена в подразделе 2 следующего раздела.

Количественные характеристики морфемных моделей

Как уже было сказано выше, исследования проводились на материале словообразовательного словаря канонических форм Д. Уорта объемом более 100 тыс. слов. Однокоренные слова в словаре объединены в гнезда. Всего в нем более 13 тыс. гнезд, если считать , что алломорфы ("блеск", "блист", "блещ") и омонимичные корни ("круп 1": "крупа", "круп 2": "крупозный", "круп 3": "круп", "круп 4": "крупье") стоят во главе отдельных гнезд. Слова разбиты на морфемы автоматически, членение слова часто оказывается более дробным, чем в других словарях, однако и более единообразным. Например, последовательность суффиксов "-ничать" представлена цепочкой: "–н–ич–а–ть" (возможны варианты: "–н–ича–ть", "–нич–а–ть", "–нича–ть"), а "–изаторский" — цепочкой "–из–ат–ор–ск–ий" и т.п. Для построения компьютерных моделей M регулярность в поморфемном разбиении необходима, т.к. благодаря ей закладывается возможность автоматического членения анализируемого в тексте слова на морфемы.

1. В словаре обнаружено несколько слов, вовсе не имеющих корня, например, "зампред", "неуд". Число допустимых цепочек в гнездах (K) определяет свойство продуктивности (способности образовывать новые слова[6]) корня и меняется от 1 до 532. Гнезда, содержащие по одному слову, составляют 37,5% всех гнезд. Малопродуктивные корни, как правило, извлечены из иноязычных слов ("абцуг", "визави", "оксюморон" и т.п.), либо являются алломорфами ("камеш", "камуш" — варианты основного корня "камен", в гнезде которого содержится 21 цепочка). Двадцать наиболее продуктивных корней помещены в табл.1 с указанием числа допустимых цепочек K в соответствующих им гнездах. Корни "вод", "граф", "да", "дел", "мер", "кат" имеют омонимичные варианты, поэтому рядом с ними помещены слова, дающие представление о семантике корневого морфа. Примерно 3% гнезд (370) самых продуктивных корней (с $K \geq 50$) покрывают треть объема словаря.

Таблица 1

Список самых продуктивных корней русского языка

№ п/п	Корень	K	Пример слова	№ п/п	Корень	K	Пример слова
1	ВОД*	532	"водить"	11	ЖИ*	251	
2	ГРАФ	421	"графа"	12	ДЕЛ	245	"делать"
3	ДА*	355	"дать"	13	ВАЛ*	242	
4	МЕТР	355		14	ВАР	239	
5	НОС	301		15	БИ*	236	
6	ХОД*	294		16	ЗНА*	235	
7	РЕЗ*	281		17	МЕР	233	"мера"
8	ПРАВ	274		18	КАТ*	232	"катать"
9	ВЕД	266		19	РОД	231	
10	ВИД	255		20	ПИС*	226	

Символом "*" помечены корни, встретившиеся среди первых двадцати по продуктивности в [4]. Пересечение составляет 50%, причем самый продуктивный корень ВОД в [4] расположен на

восьмом месте. Возможно, существенная разница в результатах объясняется тем, что в допустимых цепочках в [4] различаются ударные и безударные морфемы. В этом случае указание ударного гласного, который может стоять в разных морфах m_i, порождает из одной цепочки несколько в зависимости от того, сколько вариантов ударения возможны в ней. Соответственно, и число цепочек в гнезде изменится, и как следствие — продуктивность корня. Учет ударности/безударности морфов весьма важен для анализа речи. При исследовании письменных текстов наличие ударения только усложняет анализ. Еще одним источником расхождения результатов исследований являются различия в членении слов на морфемы. Более дробное и унифицированное в словаре Уорта уменьшает разнообразие морфемных моделей и, следовательно, повышает их частоту встречаемости.

2. Полученный список допустимых m_i составил не многим более 29 тыс. цепочек (29131). Если в цепочке m_i сохранять информацию о частеречной принадлежности слова, то m_i примет вид $m_i^* =$ "ЧРЗ m_i", где ЧРЗ — идентификатор части речи, принимающий значения: С — существительное, П — прилагательное, Г — глагол, Н — наречие, ПРЧ — причастие, Ф — фразеологизм, СР - - степени сравнения, ЧС — числительное, Д — деепричастие, М — междометие, ПР — предлог, СЗ — союз, Ч — частица. Например, из представления $m_i^* =$ $=$ "П на–*R*–е–нн–ый" следует, что аффиксальная цепочка получена из слова, относящегося к прилагательным. В случае указания ЧРЗ в m_i число допустимых цепочек возрастет до 29926. Это означает, что формообразующие свойства аффиксальных цепочек лишь у трех процентов допустимых цепочек (в среднем по всем ЧРЗ) не позволяют установить однозначно их принадлежность к определенной части речи. В качестве иллюстрации можно привести такие цепочки:

а) $m_i =$ "*R*-а", $f(m_i) = 1598$, из них:
$m_i^* =$ "С *R*–а" имеют $f(m_i^*) = 1583$,
$m_i^* =$ "Н *R*–а" — $f(m_i^*) = 7$,
$m_i^* =$ "М *R*–а" — $f(m_i^*) = 3$,
$m_i^* =$ "Д *R*–а" — $f(m_i^*) = 3$,
$m_i^* =$ "ЧС *R*–а" — $f(m_i^*) = 2$;

б) m_i = "R-ь", $f(m_i) = 693$, из них:
m_i^* = "С R-ь" имеют $f(m_i^*) = 655$,
m_i^* = "Г R-ь" — $f(m_i^*) = 10$,
m_i^* = "Н R-а" — $f(m_i^*) = 10$,
m_i^* = "ЧС R-ь" — $f(m_i^*) = 6$,
m_i^* = "М R-ь" — $f(m_i^*) = 5$,
m_i^* = "ПР R-а" — $f(m_i^*) = 3$,
m_i^* = "СЗ R-ь" —$f(m_i^*) = 3$,
m_i^* = "Ч R-ь" — $f(m_i^*) = 1$.

Самыми сложными случаями в различении частеречного значения являются цепочки, соответствующие причастиям и прилагательным. Например,

m_i = "R ир ов а нн ый", $f(m_i) = 332$, из них:
m_i^* = "ПРЧ R-ир-ов-а-нн-ый" имеют $f(m_i^*) = 203$,
m_i^* = "П R-ир-ов-а-нн-ый" — $f(m_i^*) = 129$.

3. Большая часть допустимых цепочек m_i соответствует лишь одному слову в словаре Д. Уорта, т.е. $f(m_i) = 1$. Лишь 24 % цепочек имеют $f(m_i) > 1$. Цепочкой, представляющей самое большое количество разнокоренных слов, является "вырожденная" — m_1 = "R". Ее частота встречаемости $f(m_1) = 4677$ (по частям речи: 4544 — С, 58 — М, 26 — Н, 17 — Ч, 12 — ПР, 6 — СЗ, 6 — П, 5 — МС (местоимение), 2 — ЧС, 1 — СП (субстантивированное прилагательное)). Здесь частеречная характеристика определяется с не меньшей чем 3 % точностью по принадлежности к m_1. Более подробно зависимость доли m_i в словаре, т.е. $m_i/|M|$ в %, от $f(m_i)$ можно проследить по табл. 2.

Таблица 2

Доля допустимых цепочек в словаре Д.Уорта с заданной частотой встречаемости

$f(m_i)$							
> 1	$= 2$	$= 3$	$= 4$	$= 5$	$= 6$	$= 7$	$= 8$
24%	10%	4%	2%	1,3%	0,8%	0,7%	0,5%
$f(m_i)$							
$= 9$	$= 10$	> 10	> 20	> 30	> 40	> 50	> 100
0,4%	0,4%	3,8%	2%	1,6%	1,2%	0,9%	0,4%

Частотная характеристика $f(m_i)$ отражает число возможных ссылок от разных корней на цепочку m_i. Оно меняется, следовательно, от 1 до $\max\limits_{m_i} f(m_i)$. Примеры двадцати самых частых структур, покрывающих 56 % слов словаря, приведены в табл. 3.

Таблица 3

Наиболее частые морфемные модели слов

№ п/п	m_i^*	$f(m_i^*)$	Примеры корней, реализуемых в структурах
1	"С *R*"	4544	берег, ключ, пласт
2	"С *R*–а"	1583	дорог, езд, забот
3	"П *R*–н–ый"	1374	закон, нрав, плав
4	"С *R*–к–а"	940	лун, книж, лод
5	"П *R*–ов–ый"	726	бриллиант, крем, план
6	"С *R*–ь"	655	блаж, ден, суш
7	"Г *R* и –ть"	604	буд, винт, говор
8	"Г *R*–ир–ов–а–ть"–	493	баланс, дресс, ремонт
9	"П *R*–ск–ий"–	492	ветеран, гаван, свет
10	"С *R* и –я"	402	гарант, кулинар, станц
11	"Г *R*–и–ть–ся"	388	гнезд, мир, плат
12	"Г *R*–а–ть"	383	дел, кат, леж
13	"С *R*–н–ик"	370	втор, лапот, сапож
14	"С *R* –ок"–	319	береж, мед, струч
15	"Г *R* ир ов а ть –ся"–	308	виз, план, тренер
16	"С *R*–изм"–	291	капитал, морф, сад
17	"С *R*–ист"–	290	лог, танк, юмор
18	"С *R* –ец"–	278	гор, крест, секрет
19	"С *R*–ик"	274	гвозд, куст, ствол
20	"Г за *R* и –ть"	273	вар, готов, ключ

Знаком "–" помечены структуры, не встречающиеся в [4]. Самые частые структуры, естественно, принадлежат трем основным частям речи и соответствуют однокоренным словам, в основном, бесприставочным. Приставки в часто встречающихся допустимых цепочках присущи, в основном, глаголам. В первый десяток по частоте встречаемости входят такие префиксы: "за",

"по", "на", "про", "пере", "вы", "от", "с", "у", "при". В первой сотне упорядоченного по убыванию частоты списка допустимых структур кроме глаголов встречается всего одно причастие с приставкой. Суффиксальные цепочки в часто встречающихся структурах не слишком длинные — в среднем в них содержится три четыре элемента. Таким образом, самыми частыми оказываются короткие допустимые цепочки с простейшей структурой.

4. У сложных слов, имеющих вторые корни, $f(m_i^*) < 34$. Например, у m_i^* = "П (одн)о-R-н-ый" (R = "акт", "борт", "струн") — $f(m_i^*) = 33$; у m_i^* = "П (тр)ёх-R-н-ый" (R = "лист", "мест", "гран") — $f(m_i^*) = 31$; у m_i^* = "С (пол)у-R" (R = "бак", "бог", "свет") — $f(m_i^*) = 31$; у m_i^* = "С (авт)о-R" (R = "граф", "мат", "стоп") — $f(m_i^*) = 28$; у m_i^* = "П (электр)о-R-н-ый" (R = "бур", "воз", "лит") — $f(m_i^*) = 22$.

Всего в словаре встретилось около 16% слов, содержащих более одного корня, включая собственно сложные слова, сложные предлоги, фразеологизмы. Если унифицировать все корни в слове, т.е. все, что заключено в скобках обзначить R_1, R_2 (основному корню по-прежнему соответствует символ R без индекса) и т.д., например, m_i^* = "П (электр)о-R-н-ый" привести к виду m_i^{**} = "П R_1-о-R-н-ый", то число структур, покрывающих сложные слова, существенно уменьшится (с 12783 до 4984). Самые частые многокорневые морфемные модели содержат два корня. В табл.4 помещены примеры структур с $f(m_i^{**}) > 100$. При сравнении перечисленных в ней самых частых многокорневых моделей со списком, приведенным в табл.3, обращает на себя внимание то, что в большинстве случаев часто встречающиеся однокорневые модели преобразуются в самые частые двухкорневые присоединением дополнительного корня (через соединительные гласные "о", "е", или, реже, без них) без изменения аффиксального оформления в морфемной модели. Например, таким образом получена одна из частых структур m_1^{**} = = "П R_1-о-R-н-ый" ($f(m_1^{**}) = 674$) из m_1^* = "П R-н-ый" ($f(m_1^*) =$ = 1374). Многокорневые структуры с аффиксальными аранжировками, не встречающимися в списке двадцати самых частых однокорневых структур (табл. 3), в своей основе, тем не менее, также содержат однокорневые с высокой частотой встречаемости.

Таблица 4

Морфемные модели сложных слов

m_i^{**}	$f(m_i^{**})$	R_1	R
"С R_1-о-R"	1278	авт, электр	кар, кран
"П R_1-о-R-н-ый"	674	мног, одн	мест, знач
"С R_1-о-R-и-я"	426	ге, стере	метр, граф
"П R_1-о-R-ич-ес-к-ий"	317	хрон, сейсм	метр, лог
"С R_1-о-R-к-а"	202	вод, мал	мер
"С R_1-о-R-а"	168	зо, авт	баз
"П R_1-о-R-ый"	159	бел, черн	бров, грив
"П R_1-R-ич-ес-к-ий"	155	мета,радио	физ
"П R_1-R-н-ый"	128	поли, экзо	гам
"С R_1-о-R-и-е"	125	слаб, мал	ум, душ
"С R_1-о-R-е-ни-е"	119	кров, вод	теч
"П R_1-е-R-н-ый"	105	корн, луч	образ

Все они, кроме одной, входят в первую сотню упорядоченного по частоте встречаемости списка однокорневых морфемных моделей.

Слова, имеющие более трех корней, покрываются 586 морфемными структурами. Максимальное число корней в слове — четыре. Всего в словаре встретилось 17 структур с четырьмя корнями, каждая описывает одно слово, т.е. является уникальной. В качестве примеров четырехкорневых структур можно привести следующие:

1) m_1^{**} = "С R_1-о-R_2-о-R_3-о-R-к-и", где R_1 = "авт", R_2 = "мот", R_4 = "вел", R = "гон";

2) m_2^{**} = "П R_1-о-R_2-ист-о-R_3-о-R-н-ый", где R_1 = "золот", R_2 = "хлор", R_3 = "вод", R = "род";

3) m_3^{**} = "С $R_1 - R_2 - R_3 - R$", где R_1 = "радио", R_2 = "фото", R_3 = "теле", R = "граф".

5. Если говорить о количестве возможных аранжировок ($|M|$) корней, включая случаи как однокорневых, так и многокорневых морфемных моделей, то у разных частей речи оно далеко неодинаково. В табл. 5 представлены сведения о числе допустимых морфемных цепочек по всевозможным частям речи для $N > 100$ (N — число слов с заданным частеречным значением).

Таблица 5

Зависимость числа слов и цепочек m_i от их частеречного значения (ЧРЗ)

№ п/п	ЧРЗ	N	$\|M\|$	k
1	С	43729	13806	3,17
2	Г	27428	4441	6,18
3	П	22974	8674	2,65
4	ПРЧ	5010	1073	4,67
5	Н	1792	1040	1,72
6	Ф	246	238	1,03
7	СР	145	40	3,62
8	М	130	53	2,45
9	Д	117	75	1,56
10	Ч	102	72	1,42
11	ЧС	101	67	1,51
Итого:		101774	29579	3,4

В табл.5 даны также коэффициенты k ($k = N/|M|$), показывающие во сколько раз число морфемных моделей меньше числа слов у различных частей речи. Частеречные значения помещены в порядке убывания числа слов, к ним относящихся. Просмотрев четвертый столбец, можно заметить, что разнообразие структур у глаголов и причастий меньше, чем у других частей речи. Среди трех основных частей речи лишь у прилагательных $k < 3$, а в среднем по всем частям речи (последняя строка таблицы) больше трех.

В словаре Д.Уорта слова, относящиеся к основным частям речи (существительные, прилагательные, глаголы), составляют 91 % от всего объема словаря. То есть, самые многочисленные группы слов, объединенные по частеречному значению, обладают и меньшим разнообразием в обобщенном представлении m_i^*. Если не включать в рассмотрение встретившиеся по разу морфемные модели слов, то k резко возрастет: у существительных — до 10, у глаголов — до 16, у прилагательных — до 8,5.

Уникальные структуры, составляющие 76 % всех допустимых и описывающие лишь одно слово словаря, покрывают 22 % всех слов словаря. Уникальные модели имеют почти все фразеологизмы, наречия, деепричастия (в табл. 5 — части речи с $k < 2$). Среди основных частей речи — это сложные по морфемному составу слова (имеющие более одного корня или более одной приставки). Следующий пример наглядно демонстрирует, как префиксное дополнение самых частых m_i^* влияет на их частоту встречаемости $f(m_i^*)$:

Префиксная цепочка	$m_i^* =$ "С *R*"	"П *R*–ов–ый"	"Г *R*–и–ть"
—	4544	726	604
под	66	1	33
под–раз	2	—	1
по	65	5	254
по–за	1	—	1
при	64	3	—
при–с	1	—	3

Среди единичных m_i^*, не относящихся к сложным словам и словам с префиксной цепочкой, основную массу составляют структуры с уникальными суффиксальными последовательностями, часто содержащими редко встречающиеся суффиксы ("*R*–ос–" (тор–ос–и–ть), "*R*–он–" (астен–он–и–я)) в ближайших к корню позициях. При отсутствии уникальных суффиксов в цепочке ближайшую к корню позицию занимает, как правило, суффикс, встречающийся в структурах с невысокой частотой $f(m_i^*)$.

Полученные количественные характеристики не всегда удается сопоставить с представленными в словаре [4] из–за отсутствия в нем точных (а иногда и каких–либо) соответствующих количественных данных. Как уже указывалось, есть расхождения в перечне наиболее частых морфемных моделей и продуктивных корней, вызванные различиями в объеме и представлении исследуемого материала.

Заключение

На материале словообразовательного словаря русского языка объемом более 100 тыс. слов построено множество морфемных моделей русского языка. Проведенные исследования словообразовательных гнезд показали, что продуктивность корневых морфов далеко не одинакова — от одного слова до нескольких сотен. Примерно 60% корней порождают более одного слова. Однако, доля самых продуктивных корней с числом слов в гнезде не меньше 50 составляет всего 3 %.

С другой стороны, число всевозможных морфемных моделей, покрывающих словарь, довольно велико — более 29 тыс. Самая часто встречающаяся модель описывает слово, состоящее только из корня. Другие наиболее часто встречающиеся морфемные модели характеризуются не слишком большой длиной и простотой морфемной структуры.

Около 76 % (примерно 22 тыс.) аффиксальных цепочек являются уникальными, т.е. встречаются только в одном слове словаря. И, как правило, они описывают слова, не часто встречающиеся в текстах. Цепочки с большой частотой встречаемости (по словарю) соответствуют моделям как частых словоупотреблений в текстах, так и редких. Частоты встречаемости слов в текстах определялись по словарю Л.Н. Засориной. Число морфемных моделей в среднем по всем частям речи в 3,4 раза меньше, чем слов в словаре, что может быть использовано для дальнейшей (в развитие [1–3]) компактизации словарей. Среди основных частей речи выделяются глаголы, у которых число морфемных моделей в 6 раз меньше числа слов.

Кроме того, в самом представлении слов в виде морфемных моделей заключена дополнительная полезная для анализа текстов информация. С помощью перечня аффиксальных цепочек в анализируемых словах текста можно выделить их смыслоопределяющую часть – корень. А путем проверки ссылок–связей, установленных между морфемными структурами и корнями, легко установить, верно ли выделен корень и разрешить случаи омонимии, возникающие при вычленении аффиксальной цепочки в словоформе текста. Исследование семантики фразы на уровне

составляющих ее корней представляется более обозримым, так как известно, что число корней на порядок меньше числа слов.

Литература

1. БОЛЬШАКОВ И.А., ЕМЕЛИН Е.В. Алгоритм минимизации графового представления словарей // Изв. АН СССР. Техническая кибернетика.– 1987, № 4.– С. 3—13.

2. БЕЛОНОГОВ Г.Г., КУЗНЕЦОВ Б.А. Языковые средства автоматизированных информационных систем. – М.: Наука, 1983. 287 с.

3. АШМАНОВ И.С. Архитектура и технология промышленной реализации прикладных лингвистических систем: Автореф. дис. . . канд. технич. наук. – Переяславль–Залесский, 1995.

4. КУЗНЕЦОВА А.И., ЕФРЕМОВА Т.Ф. Словарь морфем русского языка. М.: Русский язык, 1986. 1133 с.

5. WORT D., KOZAK A., JONSON D. Russian Derivation Dictionary. — New–York, 1970. — 747 p.

6. АХМАНОВА О.С. Словарь лингвистических терминов. – М.: Сов. энциклопедия, 1969. – 606с .

7. КНУТ Д. Искусство программирования. Т.1. — М.: Мир, 1976. 735 с.

Поступила в редакцию
29 февраля 2000 года