МЕТОДЫ ОБНАРУЖЕНИЯ
ЭМПЕРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
2001 год Выпуск 167

УДК 519.769

КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ВАРИАТИВНОСТИ МОРФЕМНЫХ МОДЕЛЕЙ (на материале словаря канонических форм русского языка) [1]

Н.В. Саломатина

Введение

Настоящая работа является продолжением исследований вариативности структурных единиц естественного языка, в частности, слов. Под вариативностью понимается способность одних слов переходить в другие при незначительном искажении их буквенного состава. Искажения могут быть описаны с помощью редакционных операций, например, таких как вставка, замена и устранение элемента, составляющего слово, — символа, фонемы, морфа. Варьирование канонических форм и корней на символьном уровне (замена, вставка, устранение одной, двух или трех букв, а также применение к слову комбинаций указанных операций) рассмотрено в [1, 2].

Объект изучения данной работы — морфемные структуры канонических форм слов русского языка. Под морфемной структурой (моделью) понимается представление слова в виде цепочки морфов, в которой корневой морф унифицирован, т.е. заменен

[1] Работа выполнена в рамках проекта №00-06-80420, поддержанного грантом РФФИ.

на определенный символ. Например, слова "под-бор-к-а", "под-вод-к-а", "под-зем-к-а", "под-нож-к-а" имеют одну и ту же морфемную модель: "под-R-к-а", где R принимает одно из значений: "бор", "вод", "зем", "нож". По сути, представление словаря в виде морфемных моделей уже демонстрирует вариативность слов на морфемном уровне, а именно: способность слов переходить друг в друга при замене корневого морфа. Количественные характеристики морфемных моделей, полученные по словарю канонических форм Д. Уорта, опубликованы в [3].

Изучаемые в данной работе единицы варьирования — аффиксальные морфы. Они включают в себя множество префиксов (Pr), суффиксов (Sf), окончаний (F) и возвратных частиц (C). Количественные характеристики аффиксальных морфов, полученные на более ограниченном чем в данной работе словарном материале без учета позиции относительно корня и связности аффиксов в цепочке, опубликованы в [4]. Позиционные частоты встречаемости аффиксов в суффиксальных и префиксальных цепочках канонических форм приведены в работах [5, 6]. Количественное исследование вариативности морфемных моделей, насколько известно автору, не проводилось. Цель работы — количественно описать способность морфемных моделей переходить друг в друга при замене, вставке/удалении одного аффиксального морфа в модели.

Возможности практического применения исследований вариативности различных языковых единиц, в том числе морфов, намечены в [7]. В частности, морфемные модели могут быть использованы для сжатия словарей, текстов. По данным [3] представление слов в словаре в виде морфемных структур сокращает число записей более чем в два раза, а также уменьшает длину записи слова на длину корня. Исследование вариативности морфемных моделей представляет интерес в плане дальнейшего сокращения числа записей в словаре путем хранения всех близких вариантов модели в виде одной записи с заданным множеством элементов варьирования.

1. Определение близости морфемных моделей

Пусть m — морфемная модель из множества морфемных моделей M, т.е. цепочка морфов, в которой корневой

морф унифицирован. Морфемную модель m можно записать в обобщенном виде как последовательность аффиксов, разделенных символом корня (R): $m = p_{-l}p_{1-l}...p_{-1}Rs_1...s_k...s_{n-1}s_nfc$, где $p_{-l}, p_{1-l}, ..., p_{-1} \in Pr$, $s_1, ..., s_n \in Sf$, l, n — начальная и конечная позиция аффикса в слове относительно корневого морфа, $f \in F$, $c \in C$. Здесь нами использованы разные обозначения для суффиксов, чтобы различать словообразовательные (s_k) и словоизменительные (f и c) суффиксы, и тем самым специфицировать морфемные модели.

Любой из аффиксальных элементов может иметь "нулевое" значение, т.е. морфемные модели не всегда содержат префиксальную или суффиксальную часть. Например, p_1R: под – R ("под-бор"), Rs_1: R – ист ("арт-ист"), Rf: R – а ("плат-а").

Количественную оценку близости моделей m_i и m_j из M можно получить с помощью вычисления редакционного расстояния между ними [8]. В данном случае редакционное расстояние $d(m_i, m_j)$ равно минимальному числу редакционных операций, переводящих m_i в m_j. В качестве редакционных рассматриваются операции вставки (I), замены (S), удаления (D) составляющего элемента модели -- морфа. Указанные редакционные операции применяются только к аффиксальным (p_k, s_k, f, c) морфам. Символ "R", заменяющий корень, выступает в качестве разделителя в аффиксальной цепочке и не затрагивается редакционными операциями. Набор значений, которые может принимать "R" для разных морфемных моделей, вообще говоря, различен, но при вычислении редакционного расстояния различия в значении "R" не принимаются во внимание. Например, морфемные модели слов "за-брос-к-а" (m_1 = за – R – к – а), "вы-сыл-к-а" (m_2 = вы – R – к – а) отличаются одним префиксом и $d(m_1, m_2) = 1$.

Морфемные модели m_i и m_j близки в смысле редакционного расстояния, если $d(m_i, m_j)/min(|m_i|, |m_j|) \leq q$, где q — фиксированный порог, существенно меньший 1, а $|m_i|, |m_j|$ — число морфов в моделях m_i, m_j. Мы будем использовать значения $q \leq 1/3$. Морфемные модели, удаленные от m не более чем на d в метрике редакционного расстояния, образуют ее d-окрестность, обозначаемую $V_d(m)$. Если $d = 1$, полную 1-окрестность модели m

составляют три подокрестности:

$$V_1(m) = (V^S(m)) \bigcup (V^I(m)) \bigcup (V^D(m)),$$

где $V^S(m)$, $V^I(m)$ и $V^D(m)$ включают в себя модели, отличающиеся от m соответственно заменой, вставкой и удалением одного морфа. В свою очередь, множество $V^S(m)$ состоит из совокупности морфемных моделей, отличающихся заменой морфа в определенной k-ой позиции модели: $V^S(m) = \bigcup_k V_k^S(m)$, аналогично: $V^I(m) = \bigcup_k V_k^I(m)$ и $V^D(m) = \bigcup_k V_k^D(m)$.

Список морфов в k-й позиции моделей из $V_k^S(m)$, включающий также k-й морф из m, представляет собой вектор замен (или подстановок) в этой позиции и обозначается $sub_k(m)$. Список морфов в k-й позиции моделей из $V_k^I(m)$ ($V_k^D(m)$) определяет вектор вставок (удалений) морфа в указанной позиции и обозначается $ins_k(m)$ ($del_k(m)$).

Ниже приведены примеры возможных векторов вставок и замен для морфемной модели m = под – R – к – а (см.следующую страницу).

Частичную окрестность $V_1(m)$ в приведенном примере составляют следующие модели (R — образцы допустимых корневых морфов):

под-R-к-а (R = бел), за-R-к-а (R = брос), пере-R-к-а (R = вал), вы-R-к-а (R = воз), с-R-к-а (R = бор), при-R-к-а (R = вяз), от-R-к-а (R = верт), про-R-к-а (R = дел), на-R-к-а (R = сад), по-R-к-а (R = зем), рас-R-к-а (R = кач), о-R-к-а (R = чист), пере-под-R-к-а (R = готов), под-за-R-к-а (R = прав), под-на-R-к-а (R = лад), под-R-в-к-а (R = ши), под-R-ен-к-а (R = дуб), под-R-ов-к-а (R = страх), под-R-в-а (R = ли), под-R-иц-а (R = ызб), под-R-щиц-а (R = бор), под-R-к-и (R = мыш), R-к-а (R = ков), под-R-а (R = ков);

$sub_{-1}(m)$ = (вы, за, на, о, от, пере, по, под, при, про, рас, с); $sub_1(m)$ = (в, иц, щиц) — замены в суффиксах; $sub_2(m)$ = = (а, и) — замены в окончаниях; $ins_{-2}(m)$ = (пере); $ins_{-1}(m)$ = = (на, за); $ins_1(m)$ = (в, ен, ов); $del_{-1}(m)$ = (под); $del_1(m)$ = (к).

В С Т А В К И								
-2		-1		1				№позиций для вставляемых морфов
				в				
		на		ен				
пере		за		ов				
⇓		⇓		⇓				
	под	-	R	-	к	-	а	
	вы				в		и	З
	за				иц			
	на				щиц			А
	о							
	от							М
	пере							
	по							Е
	при							
	про							Н
	рас							
	с							Ы
	-1				1		2	№позиций для заменяемых морфов

2. Предобработка множества морфемных моделей

Чтобы оптимизировать формирование окрестностей $V_k^S(m)$, $V_k^I(m)$, $V_k^D(m)$, необходимо предварительно провести разбиение множества всех моделей M на подмножества одинаковой длины с фиксированным для каждого подмножества положением корня. Для этого приведем каждую структуру m к обобщенному виду, в котором буквенный состав морфов не конкретизирован. Обозначим все префиксы символом p, а суффиксы в соответствии с их спецификой — s, f и c. Тогда m преобразуется к виду: $m' = pp...pRs...sfc$, в котором сохраняется лишь число префиксов и суффиксов исходной модели. Представление m в виде m' будем называть типовой моделью для m. Например, m_1 = не-по-R-а (непоседа) и m_2 = не-до-R-ть (недодать) имеют одинаковую типовую модель: $m' = ppRf$. Аналогично, m_1 = ис-по-R-ов-а-ть

(исповедовать), m_2 = не-до-R-ев-а-ть (недоумевать), m_3 = не-у-R-ва-ющ-ий (неунывающий) имеют $m' = ppRssf$ в качестве типовой модели. Следует отметить, что при одинаковой длине m'_1 и m'_2 и разном положении R будем иметь разные типовые модели: $ppRsf$, $pRssf$, $pRsss$ и т.д.

В соответствии с данным определением множество M может быть разбито на подмножества, содержащие типовые модели одинаковой длины и одинаковым положением R в модели: $M = \bigcup_{l=1}^{L} M_l$, L — число типовых моделей. Морфемные структуры с одинаковой типовой моделью и составляют каждое подмножество M_l ($|M_l|$ — число структур в M_l). Для построения $V_k^S(m)$ достаточно провести сравнение моделей, принадлежащих одному и тому же подмножеству M_l. Для выявления окрестностей $V_k^I(m)$, $V_k^D(m)$ (вставка/удаление аффикса) нужно сравнить каждое подмножество моделей с другим, в котором модели имеют на один аффикс больше (случай вставки) или меньше (в случае удаления).

Как было выяснено в [3], множество M морфемных моделей, покрывающих словарь в 100 тыс. слов Д. Уорта, содержит около 30 тыс. структур. Количество подмножеств с одинаковой структурой m' для каждого подмножества) уже не столь велико — 119. Около половины подмножеств маломощные, каждое из них содержит менее 10 морфемных моделей. Максимальное число моделей в подмножестве достигает почти 2,5 тыс. Список типовых моделей m', к которым сводятся более 50 % всех морфемных моделей словаря, представлен в табл. 1.

Типовые модели, выписанные в третьем столбце таблице, упорядочены согласно указанному в предыдущем столбце параметру $|M_l|$, т.е. числу морфемных моделей m в M_l с одной и той же типовой структурой m'. Так, типовая модель $m' = pRssf$ является самым часто употребляемым образцом для построения морфемных моделей, например, таких как: m_1 = вы-R-е-нн-ый, m_2 = у-R-е-ни-е, m_3 = за-R-ов-а-ть. В четвертом столбце таблицы помещены данные о числе n слов словаря с типовой структурой m'. Около 70 % слов словаря имеют именно те типовые модели, которые перечислены в табл. 1. В последнем столбце выписан ранг r типовой модели при упорядочении структур по

n. Следует отметить, что числа l и r совпадают лишь в четвертой строке таблицы, т.е. единственная типовая модель *Rssf* занимает четвертую позицию при упорядочении типовых моделей и по числу слов и по числу морфемных моделей в m'.

Таблица 1

Интегральные характеристики типовых морфемных структур

Номер п/п (l)	$\|M_l\|$	Типовая модель m'	n	r
1	2471	*pRssf*	8947	3
2	1680	*pRsf*	14105	1
3	1365	*pRsssf*	2155	11
4	1058	*Rssf*	7181	4
5	955	*Rsssf*	3404	8
6	896	*ppRsf*	1380	22
7	790	*ppRssf*	827	15
8	693	*pRssssf*	1010	19
9	522	*pRf*	1392	14
10	461	*pRs*	1447	13
11	450	*Rsssssf*	1878	12
12	444	*Rsf*	9255	2
13	433	*Rss*	2257	10
14	411	*pRssfc*	987	21
15	399	*pRsfc*	6132	5
16	316	*ppRsfc*	420	26
17	312	*pRss*	1031	18
18	309	*ppRsssf*	328	33
19	272	*Rs*	4427	7
20	255	*pRssssssf*	227	32
21	245	*Rsss*	1342	16
22	197	*Rssssssf*	414	27

В остальных случаях l и r могут существенно отличаться: диапазон $l - r$ лежит в пределах от 12 (*Rs*) до -16 (*ppRsf*). Если определение "продуктивный" в отношении модели понимать как

"...являющийся образцом словообразования" [9], то по разнице рангов видно, что продуктивность типовой модели как образца для построения морфемных моделей и слов словаря различна.

Самые частые типовые модели, помещенные в таблице, имеют не слишком сложную морфемную структуру, если считать, что сложность модели коррелирует с числом аффиксальных элементов в ней. Сложные типовые модели содержат до четырех префиксов и до семи суффиксов (без учета окончания и возвратной частицы). Причем максимальное число суффиксов в модели — семь — реализуется как в бесприставочном варианте, так и с одним, двумя, тремя префиксами. Например, пере – о – с – *R* – е – тел – ь – ств – ов – а – ни – е, *R* = вид, не – *R* – й – ств – и – тел – ь – н – ост – ь, *R* = де, *R* – он – ал – из – ир – ов – а – нн – ый, *R* = раци. Средняя длина слова (с учетом корня) равна примерно четырем морфам (модели — пяти). Число разных корней $|R|$ равно 13231, суффиксов $|Sf|$ — 466, префиксов $|Pr|$ — 73, окончаний $|F|$ — 17.

3. Способ выявления вариантов морфемных моделей

Для формирования 1–окрестностей морфемных моделей удобен способ представления данных в виде бинарных деревьев [10]. Для поиска моделей, отличающихся от заданной заменой одного морфа, рассмотрим множество M_l, которому она принадлежит. Всех представителей m_i $(1 \le i \le |M_l|)$ множества M_l характеризует, как уже было сказано, одинаковый по численности состав суффиксальных и префиксальных морфов. Заменив морф в исследуемой на наличие соседей позиции k буквой x, представим модели из M_l в виде:

$m_i = p_{-j} \ldots p_{-1-k} x p_{1-k} \ldots p_{-1} R s_1 \ldots s_n f c,$

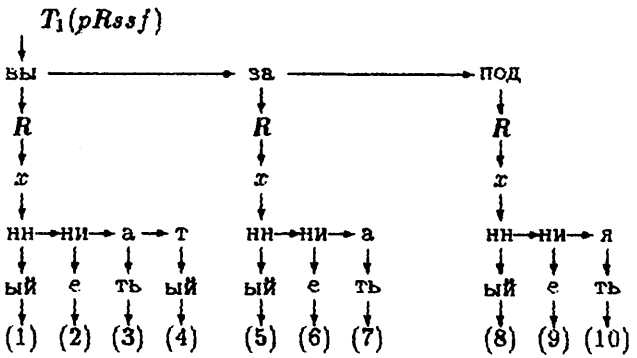
$x = p_{-k}, (k \le 4)$, если замена проведена в префиксальной части модели. Аналогично поступаем и в случае замен в суффиксальной части модели, но учитываем специфику заменяемого морфа (s, f или c):

$m_i = p_{-j} p_{1-j} \ldots p_{-1} R s_1 \ldots s_{k-1} x s_{k+1} \ldots s_n f c,$

$m_i = p_{-j} p_{1-j} \ldots p_{-1} R s_1 \ldots s_n x c.$

$m_i = p_{-j} p_{1-j} \ldots p_{-1} R s_1 \ldots s_n f x.$

Множество записанных таким образом моделей организуется в виде бинарного дерева, в узлах которого помещаются морфы, а также x, замещающий конкретные морфы. В листьях дерева собираются те морфы, которые были замены в модели на x. Полный обход дерева позволяет построить окрестности $V_k^S(m_i)$. Таким образом, для поиска всех соседей m_i по k-й позиции достаточно провести сравнение моделей внутри одного подмножества M_l. Пример фрагмента дерева приведен на рисунке. Объединив всех соседей m_i по всем позициям, получим полную 1–окрестность модели по заменам. Перебор по всем l позволяет получить полное представление о вариативности моделей. Вариативность моделей, допускаемая заменой в них корня, исследовалась, как уже говорилось, в [3].



В приведенный на рисунке фрагмент V_1^S дерева $T_1(pRssf)$ включены тридцать моделей, в том числе, например: вы – R – е–нн–ый, вы–R–а–нн–ый, вы–R–ива–ни–е, вы–R–а–ни–е, за–R–е–нн–ый, за–R–а–нн–ый, под–R–е–нн–ый, под–R–ыва–ни–е, под–R–л–я–ть. Информация, собранная в листьях и пронумерованная цифрами в круглых скобках, включает кроме самих аффиксов еще и частеречное значение слов, в моделях которых они встретились, а также продуктивность самой модели, т.е. число корней, подставляемых в модель:

(1): (е/прч/132), (а/прч/66), (я/прч/8);

(2): (ива/с/102), (ыва/с/46), (а/с/39);

(3): (ов/г/7), (ев/г/3);
(4): (ну/п/14), (о/прч/8), (у/прч/1);
(5): (е/прч/84), (а/прч/68), (ё/прч/54);
(6): (ива/с/76), (а/с/45), (е/с/40), (ыва/с/37);
(7): (ов/г/58), (ев/г/15), (к/г/16);
(8): (е/прч/50), (ё/прч/33), (а/прч/30);
(9): (ива/с/12), (а/с/7), (е/с/6), (ыва/с/5);
(10): (л/г/9), (н/г/2).

Здесь использованы следующие обозначения: "с" — существительное; "п" — прилагательное; "г" — глагол; "прч" — причастие.

В случае вставки/удаления морфа в модели сравнение проводится между двумя M_l, отличающимися по длине на один префиксальный или один суффиксальный морф. Все $m_i \in M_l$ записываются в виде:

$m_i = p_{-j}...p_{-1-k}xp_{-k}...p_{-1}Rs_1...s_nfc,$
$p_{-k}, x \in Pr,\ s_k \in Sf,\ f \in F\ ,\ c \in C;$

аналогично,

$m_i = p_{-j}p_{1-j}...p_{-1}Rs_1...s_{k-1}xs_k...s_nfc,\ x \in Sf;$

$m_i = p_{-j}p_{1-j}...p_{-1}Rs_1...s_nxfc,\ x \in F$

и т.п., если исследуется возможность вставки морфа в k-ю позицию модели. Сравнение проводится с тем M_l, в котором на один морф больше из того же множества (Pr, Sf, F, C), что и вставляемый x. В остальном процесс формирования V_k^l подобен случаю V_k^S.

4. Количественные характеристики вариативности морфемных моделей

4.1. *Особенности (свойства) 1-окрестностей.* Вариативность моделей исследовалась на подмножестве M, содержащем самые часто встречающиеся (не менее 10 раз) в словаре Д. Уорта [11] морфемные структуры, типовые модели большей части которых представлены в табл. 1. Они составляют почти 60 %

всех выявленных в словаре морфемных моделей. Подмножество слов, покрываемых исследуемыми моделями, включает более 80 % словарного состава. В рассмотрение не вошли сложные (многокорневые) слова, фразеологизмы, сложные предлоги, содержащие более одного слова, и т.п.

Понятно, что количественно преобразование морфемных моделей с помощью операции "вставка" можно описать, используя операцию "удаление", и наоборот. Множество удаляемых и вставляемых морфов при этом будет совпадать. Поэтому количественные характеристики будут рассматриваться для той или другой операции в зависимости от того, что проще получить в вычислительном плане. Пусть, например, m_1 = вы − R − от − а − ть, m_2 = вы − R − ов − а − ть, m_3 = вы − R − ев − а − ть. При удалении суффиксов в ближайшей к корню позиции все три модели перейдут в одну — $\bar{m}$ = вы − R − а − ть. Число компонентов в векторе вставок V^I = (от, ов, ев) для $\bar{m}$ равно числу моделей, из которых удалены морфы.

Было установлено (см. табл. 2), что 32 % моделей имеют соседей в случае замены одного морфа, и из 39 % моделей можно, удалив морф, получить существующую в словаре модель меньшей длины, т.е. в отличие от канонических форм стали преобладать вставки/делеции. В приводимой ниже таблице использованы следующие обозначения: $N_{sub} = max_{k,i}|sub_k(m_i)|$ — реализуемое в одной из позиций $1 \leq k \leq l$ максимальное число соседей, которое имеет модель, если произвести в этой позиции замену морфа; $N_{del} = max_k(|del_k(m_i)|)$ — максимальное суммарное число моделей, для которых удаление морфа в k-ой позиции не приводит к запрещенной морфемной структуре, $1 \leq i \leq |M|$, $|M|$ — число морфемных моделей; P_{sub} — доля моделей, имеющих соседей при использовании операции "замена" (суммарно по всем позициям); P_{del} — аналогичная характеристика для операции "удаление". Числа N_{sub} и N_{del} соответствуют количеству соседей в случае, если аффиксы рассматриваются без учета частеречного значения моделей, из которых они удалены или в которых они заменены.

Таблица 2

Характеристики вариативности морфемных структур

Номер п/п(l)	Типовая модель	Замены		Удаления	
		P_{sub}	N_{sub}	P_{del}	N_{del}
1	pRssf	41%	45	27%	45
2	pRsf	30%	56	15%	56
3	pRsssf	44%	38	48%	38
4	Rssf	34%	100	27%	100
5	Rsssf	34%	43	44%	43
6	ppRsf	43%	25	68%	25
7	ppRssf	43%	20	84%	20
8	pRsssssf	39%	32	44%	32
9	pRf	17%	44	18%	44
10	pRs	21%	26	24%	26
⋮	⋮	⋮	⋮	⋮	⋮

С учетом частеречных значений N_{sub} и N_{del} увеличатся в 1,3 — 1,5 раза, что можно продемонстрировать на примере модели m = по − R − н − о, описывающей наречие "подекадно" и предлог "подобно". Если учесть частеречное значение m, то в векторе замен, содержащем предлог "по", будут представлены отдельными компонентами наречия (н/) и предлоги (пр/) — $sub_{-1}(m)$ = (н/по, пр/по).

Из табл. 2 видно, что количество моделей, имеющих соседей, находится в определенной зависимости от числа аффиксальных морфов. Например, если рассмотреть подмножество моделей с одним префиксом, то в случае использования редакционной операции "замена" при увеличении цепочки суффиксов от нуля до трех доля моделей, имеющих соседей, растет от 17 % до 44 %. При дальнейшем удлинении суффиксальной цепочки доля моделей с непустой окрестностью снижается до 39 %. Ту же тенденцию можно наблюдать и на менее частотных типовых моделях (цифра — доля моделей с соседями):

Rsf —	22	Rssf —	34	Rss —	21
pRsf —	30	Rsssf —	34	Rsss —	33
ppRsf —	43	Rssssf —	29	Rssss —	30
pppRsf —	37	Rsssssf —	24	Rsssss —	30

Максимумы числа соседей N_{sub} и N_{del} совпадают для первых десяти моделей. Однако, в позициях, где максимум не достигается, довольно часто $|sub_k(m)|$ и N_{del} различаются. Совпадение N_{sub} и N_{del} является проявлением преимущественно аффиксального способа словообразования — для самых частых моделей морфы в позиции с максимальным варьированием могут быть как заменены, так и удалены (для них в словаре имеются подходящие модели с меньшим числом морфов). Для иллюстрации рассмотрим модель с типовой структурой *pRssf* и x в первой позиции (эта модель имеет максимум соседей в данной позиции — 81 (с учетом частеречного значения модели)):

$$m = x - R - \text{е} - \text{нн} - \text{ый} \left\{ \begin{array}{ll} 132 & \text{прч/вы} \\ 87 & \text{прч/за} \\ 79 & \text{прч/на} \\ 78 & \text{прч/пере} \\ 59 & \text{прч/про} \\ 53 & \text{п/по} \\ 50 & \text{прч/под} \\ 49 & \text{прч/по} \\ 48 & \text{прч/от} \\ 43 & \text{п/с} \\ \ldots & \ldots \end{array} \right.$$

В примере приведены десять самых продуктивных моделей-соседей, цифра перед вектором замен указывает, какое количество разных корней может быть подставлено в модель. Среди моделей с типовой структурой *Rssf* присутствует $R - \text{е} - \text{нн} - \text{ый}$, а это означает, что из всех моделей $x - R - \text{е} - \text{нн} - \text{ый}$ (x включает 81 разную приставку с учетом частеречного зачения) может быть удален 81 префикс.

Следует отметить, что количественно варьирование слов и морфемных моделей заметно отличается для $d = 1$, особенно в

случае применения операции "удаление". Осмысленные замены символов в словах возможны в 35 % случаев, морфов в моделях — в 32 %, т.е. возможности посимвольного варьирования слов и поморфемного — моделей количественно оцениваются примерно одинаково. Осмысленные удаления символов можно реализовать для 16 % всех слов, а удаления морфов — для 39 % моделей. Более высокая устойчивость моделей к устранению морфа, чем слов к устранению символа объясняется, как уже упоминалось, тем, что именно вставка/удаление аффиксального морфа, а не символа, лежит в основе словообразования в русском языке.

4.2. *Позиционная вариативность морфемных моделей.* Как уже отмечалось, в данной работе исследуются морфемные модели канонических форм со стандартными окончаниями. Поэтому менее всех морфов вариативны окончания. У многих моделей варианты у словоизменительных аффиксов возникают лишь в случае, когда совпадают окончания у моделей с разными частеречными значениями. Однако, встречаются и случаи несовпадения окончаний, например, в модели "не–за–R–и–тел–ь–н–x" x = ый (у прилагательных) может заменяться на x = о (у наречий). Вставка окончания может как поменять частеречное значение модели, так и оставить его без изменения: у – R – ист — модель существительных, а у – R – ист – x при x = ый определяет прилагательное, тогда как x = а — существительное.

Зафиксируем в качестве редакционной операцию "замена" морфа. Будем считать, что число моделей, имеющих соседей в исследуемой позиции, характеризует ее позиционную вариативность. Если рассмотреть варьируемость суффиксальной части цепочки морфов, то для моделей почти всех типов справедлива следующая закономерность: доля моделей с непустой окрестностью убывает при удалении от корня. В префиксальной части, наоборот, доля моделей с соседями растет при удалении от корня (самая вариативная позиция — начало слова). Таким образом, модели, имеющие соседей, чаще отличаются заменой префикса, стоящего в начале цепочки морфов, или суффикса, ближайшего к корню.

Количественно вариативность моделей по позициям можно проследить по табл.3, где кроме доли моделей с непустой окрестностью в указанной позиции второй строкой даны сведения о максимальном числе соседей в позиции.

В целом, по совокупности моделей почти в каждой из них вариативность ближайших к корню префиксальных позиций ниже, чем суффиксальных. При рассмотрении более удаленных от корня позиций более вариативными становятся префиксальные.

Позиционная вариативность моделей при удалении из них морфа, в основном, подчиняется тем же закономерностям, что и в случае замены. Как уже говорилось, такой результат неслучаен в силу того, что удаляемые морфы — суть компоненты вектора вставок, а вектор вставок с числом компонентов больше единицы является одновременно вектором замен в морфемной модели, содержащей на один морф больше, чем та, в которую может быть произведена вставка.

4.3. *Морфный состав и частота встречаемости векторов замен и вставок.* Компоненты векторов замен и вставок отражают особенности вариативности моделей, имеющих соседей. Так как $|del_k| = 1$, то при рассмотрении векторов больший интерес представляет ins_k. По операции"замена" с учетом всех морфемных моделей из выбранного подмножества по совокупности позиций всего получено 3127 разных векторов замен, из них 534 встретились более 1 раза. Для операции"вставка" — всего 2516 векторов, 1015 из них встретились более 1 раза. В табл. 4 приведены самые частые векторы замен и вставок с примерами типовых моделей, в которых они чаще всего встречаются (в первом столбце в скобках указано, сколько раз вектор встретился в модели). Символами n_S и n_I обозначены частоты встречаемости векторов замен и вставок суммарно по совокупности моделей и по всем позициям в модели.

Самые частые векторы замен (кроме (о, у)) состоят из суффиксальных морфов, тогда как векторы вставок (имеющие значительно большую частоту встречаемости) — как из префиксальных, так и из суффиксальных морфов. Самые частые префиксальные векторы замен — (о, у), (без, не), (по, с), (по, про) встретились 21, 17, 16 и 15 раз, соответственно. Необходимо от-

Таблица 3

Вариативность типовых моделей по позициям

Номер п/п	Тип модели	Заменяемый морф							Примеры
		P	P	S	S	S	S	P	
1	pRssf		11%	18%	17%			10%	*x* – *R* – е – вш – ий, *x* = (по, на, про, о, вы)
			81	30	6			4	
2	pRsf		7%	15%				15%	за – *R* – *x* – а, *x* = (к, щиц, ин, чиц)
			68	17				9	
3	pRsssf		13%	18%	12%	9%		7%	про – *R* – л – *x* – ни – е, *x* = (е, ива)
			55	15	5	3		3	
4	Rssf			11%	18%			13%	*R* – ь – н – *x*, *x* = (ый, я, ой, о, ий)
				106	12			6	
5	Rsssf			15%	16%	8%		3%	*R* – *x* – ов – а – ть, *x* = (ир, ств, из, к...)
				43	10	5		3	
6	ppRsf	20%	17%	21%				3%	*x* – до – *R* – н – ый, *x* = (не, на, без, пред, по)
		17	25	7				3	
7	ppRssf	19%	15%	16%	7%			5%	*o* – *x* – *R* – л – ива – ть, *x* = (бес, без, за)
		22	20	5	3			2	
8	pRsssf		13%	17%	7%	7%	4%	5%	не – *R* – и – *x* – ь – н – ый, *x* = (тел, ал, л)
			52	9	4	4	4	4	
9	pRf		6%					13%	за – *R* – *x*, *x* = (а, ти, ь, о, ий...)
			73					28	
10	pRs		6%	14%					с – *R* – *x*, *x* = (щик, ок, чик...)
			20	21					

Продолжение таблицы 3

Номер п/п	Тип модели	Удаляемый морф							Примеры
		P	P	S	S	S	S	F	
1	pRssf		13%	20%	4%			0.2%	$x - R - $ е $-$ ни $-$ е, $x =$ (у, о, по, за...)
			81	30	4			1	
2	pRsf		11%	8%				2%	воз $- R - x -$ ть, $x =$ (а, и, е, я...)
			68	17				2	
3	pRsssf		15%	39%	4%	1%			ре $- R -$ ат $- x -$ н $-$ ый, $x =$ (ив, ор)
			55	15	2	1			
4	Rssf			10%	13%			4%	$R -$ а $-$ л $- x$, $x =$ (о, а, ь, ый)
				106	110			5	
5	Rsssf			31%	19%	6%		0.1%	$R -$ ат $- x -$ н $-$ ый, $x =$ (ив, ор, ич, ур...)
				43	10	3		1	
6	ppRsf	46%	42%	7%					$x -$ с $- R -$ и $-$ ть, $x =$ (по, пере, при...)
		17	25	7					
7	ppRssf	54%	47%	13%	1%				не $- x - R -$ а $-$ ни $-$ ый, $x =$ (до, о, об, из...)
		22	20	5	1				
8	pRsssssf		16%	35%	2%	0.7%			за $- R -$ к $- x -$ а $-$ ни $-$ ый, $x =$ (ов)
			52	9	1	1			
9	pRf		12%					14%	по $- R - x$, $x =$ (ть, а, ти, у...)
			73					28	
10	pRs		16%	16%					под $- R - x$, $x =$ (ок, чик, щик...)
			29	21					

метить, что число элементов, составляющих самые частые векторы, почти всегда минимально (для замен — 2, для вставок — 1). Самый частый вектор замен с числом компонентов существенно больше минимального, встретившийся у 11 моделей, имеет в своем составе 9 морфов: (а, ва, е, и, ива, ну, о, ыва, я). Он же является одним из самых частых многокомпонентных векторов вставок. Самый частый вектор вставок, имеющий два компонента в своем составе — (л, н), встретился 35 раз.

Таблица 4

Самые частые векторы замен и вставок

Примеры m'	Замены		Примеры m'	Вставки	
	вектор	n_S		вектор	n_I
pRssf(33)	(а, и)	205	ppRssf(119)	(не)	479
pRssf(19)	(а, ыва)	98	ppRsf(45)	(по)	212
pRssf(23)	(и, ива)	94	pRsssf(56)	(л)	190
ppRssf(20)	(а, е)	76	pRssf(28)	(н)	180
pRssf(13)	(о, ый)	59	ppRsf(36)	(с)	169
pRsf(26)	(ой, ый)	51	pRsssf(35)	(ов)	147
ppRsf(13)	(и, я)	47	Rsssf(21)	(ь)	139
pRssf(16)	(и, ива, я)	35	Rssf(21)	(а)	138
pRsssss(9)	(а, ва)	26	ppRssf(22)	(о)	124
pRsf(19)	(и, ий)	25	pRssf(22)	(в)	116
ppRsf(6)	(а, и, я)	25	ppRsf(18)	(на)	101
pRsssf(8)	(из, ир)	23	pRssf(22)	(со)	95
pRsf(5)	(н, ов)	22	pRssf(13)	(т)	91
ppRsssf(4)	(о, у)	21	Rsssf(29)	(к)	90
Rsssf(7)	(н, нн)	21	ppRssf(19)	(у)	86

Распределение векторов замен и вставок по моделям. Нельзя сказать, в целом, что состав векторов замен и вставок может быть определен типом модели. Ниже указано число типовых моделей для самых частых векторов замен и вставок. Видно, что оно довольно велико.

Замены:

(а,и)	(а,ыва)	(и,ива)	(а,е)	(о,ый)	(ой,ый)	(и,я)	(и,ива,я)
26	13	15	15	15	6	10	4
(а,ва)	(и,ий)	(а,и,я)	(из,ир)	(и,ов)	(о,у)	(н,нн)	
6	5	11	5	9	9	8	

Вставки:

(не)	(по)	(л)	(н)	(с)	(ов)	(ь)	(а)
22	29	24	24	29	19	23	23
(о)	(в)	(ва)	(со)	(т)	(к)	(у)	
29	23	19	16	23	14	18	

Количественно по-разному в разных типовых моделях представлены векторы замен и вставок, но это, в основном, зависит от мощности множеств M_l, т.е. от числа морфемных моделей в них. На этом фоне выделяются словоизменительные суффиксальные векторы замен (ой, ый), (и, ий), которые встречаются в ограниченном числе типовых моделей. В то же время они входят в состав компонентов более длинных векторов замен, например, (а, и, ий, ой, ый) и т.п. Аналогично ведут себя и словообразовательные суффиксальные векторы замен, например, (и, ива, я). Вектор (из, ир), требующий определенного суффиксального обрамления, встречается только в моделях с длинной цепочкой суффиксов.

Тяготение некоторых векторов замен и вставок к моделям определенного типа сильнее выражено у префиксальных морфов. Примеры векторов замен и вставок:
вектор замен (без, не) представлен в 17 моделях, соответствующих 6-ти типовым моделям (тогда как (по, с) — в 16 моделях, типовых — 11);
вектор вставок (не) встречается в 479 моделях(типовых — 22), а вектор (по) — в 212 моделях (типовых — 29), (с) — в 169 моделях (типовых — 29).

Распределение векторов по позициям. В префиксальных частях моделей векторам замен свойственны большое число компонентов и не слишком большая частота встречаемости. Самые частые модели, имеющие префиксальную часть, содержат от одного до трех префиксов. Если в модели больше одного префикса, то при всем многообразии векторов замен заметна зависимость их состава от рассматриваемой позиции в модели по отношению к корню и началу слова. Например, *sub* = (без, не), встре-

чаясь в многопрефиксных цепочках, тяготеет к началу слова (в скобках указана частота встречаемости вектора): $p(7)pRssf$, $p(3)pRsf$, $p(2)pRsssf$, $p(2)pRssssf$, $p(2)pRsssssf$, а sub = (о, у) может встретиться в любой позиции префиксальной цепочки — $p(2)p(2)Rsssf$, $p(2)p(2)Rssfc$, $pp(1)Rssf$, $p(1)pRsf$, $ppp(1)Rssf$.

В суффиксальных позициях также наблюдается зависимость состава векторов замен от позиции (самые вариативные — первая после корня, либо первая перед окончанием). Однако эта зависимость проявляется более слабо, чем в префиксальных позициях. Например, вектор замен sub = (а, ыва) свойственен для конечной позиции в суффиксальной цепочке: $pRss(19)f$, $pRss(19)fc$, $pRss(4)s(14)f$, $pRsss(13)fc$, $ppRss(4)f$, $ppRss(3)fc$; тогда как вектор sub = (а, е) встречается во всех позициях суффиксальной цепочки: $pRs(8)s(3)f$, $Rs(5)ss(6)f$, $pRs(3)s(5)sf$, $Rs(3)s(1)f(4)$, $pRs(1)f(5)$, $pRs(1)ss(1)s(1)f$.

Векторы вставок ведут себя примерно так же, как и векторы замен. Еще одна общая тенденция — чем больше число компонентов в векторе замен или вставок, тем более обусловлена его позиционная встречаемость. Среди коротких векторов вставок имеются как характерные для определенных позиций моделей, так и встречающиеся в любом месте префиксальной или суффиксальной цепочки. Например, ins = (не) свойственен для начала слова: $p(115)p(4)Rssf$, $p(76)p(3)Rsf$, $p(64)pRsssf$, $p(29)ppRssf$, $p(22)p(7)Rf$, $p(22)pRsssssf$, в отличие от него ins = (по) более равномерно распределен по префиксальной цепочке — $p(10)p(35)Rssf$, $p(19)p(14)Rsf$, $p(12)p(13)Rsfc$, $p(5)p(14)Rsssf$, $p(3)p(4)Rsssssf$, $p(3)pp(2)Rsf$.

Заключение

На материале словаря канонических форм (более 100 тыс. слов) проведено исследование вариативности их морфемных моделей, т.е. вариативности слов на следующем после фонемного уровне значимых единиц языка. Установлено, что около трети моделей имеют соседей в случае замены и более трети — в случае удаления (вставки) одного морфа. Полученные закономерности варьирования очевидным образом обусловлены способом словообразования в данном языке. Прослеживается зависимость числа моделей с непустой 1-окрестностью от типа модели, а именно

от числа морфов в структуре и от того, в какой позиции цепочки морфов производится замена или удаление. Самыми вариативными позициями (в смысле наличия соседей у модели) являются в префиксальных цепочках — начальная, в суффиксальных — первая после корня и (в меньшей степени) последняя перед окончанием.

Для моделей, имеющих соседей, выявлены самые частые векторы замен и вставок. Исследована зависимость состава векторов замен и вставок от типа модели и позиции в цепочке морфов. Самые частые векторы замен встречаются в цепочках суффиксов, вставок — в цепочках префиксов. Явной зависимости состава векторов замен от типа модели не установлено. Однако, наблюдается позиционная предопределенность некоторых векторов замен и вставок.

Полученные закономерности позволяют количественно охарактеризовать процессы словообразования не только внутри одного словообразовательного гнезда, но и в совокупности по всем корневым гнездам словаря. Представляет интерес и возможность практического использования полученных результатов для многоступенчатого сжатия словарей, когда словарь представляется совокупностью списков корней и морфемных моделей, имеющих ссылки на доопределяющие их до словоформ корни. Те модели (по нашим данным — более трети), что имеют варианты, образованные заменой и/или вставкой/удалением морфа, достаточно представлять одной основной записью с набором векторов замен и вставок.

Литература

1. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Определение и анализ ближайших окрестностей корней слов русского языка // Обнаружение эмпирических закономерностей. — Новосибирск, 1999. — Вып. 166: Вычислительные системы — С. 80–103.

2. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Электронный словарь паронимов: версия 1 // НТИ, сер.2. Информационные процессы и системы. — 2000. — 6. — С. 34–41.

3. САЛОМАТИНА Н.В. Количественные исследования морфемной структуры слов русского языка (на базе электронного

словаря Д. Уорта). //Обнаружение эмпирических закономерностей. — — Новосибирск, 1999. — Вып. 166: Вычислительные системы — С. 104–118.

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

5. САЛОМАТИНА Н.В., ЮДИНА Л.С. О некоторых статистических характеристиках префиксов // Анализ текстов и сигналов. — Новосибирск, 1987. — Вып. 123: Вычислительные системы — С. 84–100.

6. САЛОМАТИНА Н.В., ЮДИНА Л.С. Фонетическая организация морфем (на статистическом материале суффиксов) // Тез. докл. 15-го Всесоюз. семинара (АРСО-15). — Таллин, 1989. С. 297-298.

7. ГУСЕВ В.Д., САЛОМАТИНА Н.В. Количественные исследования вариативности языковых единиц // Труды международной научно-практической конференции KDS-2001. Том 1. — Санкт-Петербург, 2001. — С. 186–193.

8. WAGNER R.A., FISHER M.J. The string — to — string correction problem // J. ASM. — 1974. — Vol. 21, 1. — P. 168–173.

9. АХМАНОВА О.С. Словарь лингвистических терминов. — М.: Сов. энциклопедия, 1969. — 606 с.

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

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

Поступила в редакцию
7 октября 2001 года