УДК 004.656

В. Д. Гусев, Н. В. Саломатина

Электронный словарь паронимов: версия 1*

На достаточно объемном словаре русского языка (свыше 100 тыс. канонических форм) исследуется способность одних слов переходить в другие в результате незначительного искажения их буквенного состава. Построен словарь ближайших соседей каждого слова (словарь "паронимов" в широком смысле). Проанализирована зависимость числа соседей от длины слова, характера искажения (замена, вставка, устранение символа) и его локализации. Выявлены характерные варианты замен и вставок в разных позициях. Рассмотрены возможности практического применения полученного словаря.

ВВЕДЕНИЕ

Традиционным продуктом лингвистических исследований являются различного рода словари (толковые, двуязычные, синонимов, антонимов и т. п.), создаваемые, как правило, вручную. В связи с широким распространением персональных компьютеров и созданием больших текстовых баз данных появилась возможность автоматизации подобного рода деятельности. При этом речь идет не только об ускорении создания соответствующего лингвистического продукта, но и о качественно новом уровне обеспечения его полноты, корректности и доступности. Примером такого рода разработки является описываемый в данной статье электронный (или компьютерный) словарь паронимов русского языка, формирование которого потребовало формализации понятия "паронимы" и реализация эффективного алгоритма их поиска.

Стимулом к созданию электронного словаря паронимов послужило исследование накопленной нами в течение ряда лет подборки текстовых ошибок, **не выявляемых** существующими **автоматическими** корректорами типа ОРФО 3.5, ОРФО 4.0, входящими в состав русской версии текстового процессора Microsoft Word 6.0 для Windows. Выяснилось, что часто встречающимися ошибками такого рода являются паронимические, стоящие на втором месте после ошибок согласования и управления. В связи с этим возникает вопрос о степени распространенности паронимов в языке и их специфических особенностях, о сравнении помехоустойчивости естественного языка и формальных кодовых систем, о возможности практического использования явления паронимии и ряд других.

Известные нам словари паронимов русского языка (О. В. Вишняковой [1], Ю. А. Бельчикова и М. С. Панюшевой [2], Н. П. Колесникова [3]) составлены вручную и весьма ограничены по объему. Первый словарь содержит порядка 1000 пар однокоренных паронимов, второй — свыше 200 гнезд (однокоренных), третий — порядка 1400 гнезд (по большей части двухсловных, однокоренных и разнокоренных). Ограниченность объемов двух первых словарей объясняется довольно узким толкованием термина "паронимы" (однокоренные слова, принадлежащие к одной части речи и в большинстве случаев семантически соотнесенные друг с другом). Словарь Н. П. Колесникова основан на более широкой трактовке термина *паронимы* и, как следствие, больше по объему, однако содержит спорное и не поддающееся формализации ограничение, сводящееся к тому, что не все сходные в звуковом отношении слова смешиваются лицами, владеющими языком. Поэтому некоторые из них являются паронимами, а другие — нет (например, *казаться* и *касаться*). Последние в словарь не включаются. Анализ нашей подборки паронимических ошибок обнаруживает, тем не менее, множество случаев, когда ошибка имела место, но соответствующая пара слов не фигурировала в словаре Н. П. Колесникова в качестве паронимов.

Целью работы является описание процедуры формирования электронного словаря паронимов русского языка и исследование его свойств. Результаты анализа представлены в виде количественных характеристик вариативности как языка в целом, так и фиксированных его подмножеств, включая отдельные слова. Под вариативностью мы понимаем способность одних слов переходить в другие из данного словаря в результате незначительного изменения ("искажения") их буквенного состава. Базой для создания электронного словаря паронимов послужил достаточно представительный словарь русского языка объемом свыше 100 тыс. слов [4].

Данная работа, являющаяся непосредственным продолжением и развитием [5], завершает этап формирования первой версии электронного словаря паронимов. Если в [5] рассматривались пары (или группы) слов, различающихся заменой лишь одного символа, то здесь исследуются слова, отличающиеся вставкой (устранением) одного символа. Проводится сравнительный анализ обоих случаев. Последующие версии словаря будут допускать совместное использование операций замены и вставки, что приведет к размыванию степени близости элементов, составляющих паронимическую связку, и, соответственно, к увеличению словаря паронимов.

В идейном плане данная работа перекликается с [6]. Там исследовалось явление омоформии (совпадение несловарных форм слов или словарной формы одного слова с несловарной другого, например, *дам* от *дать* и *дама*). Построенный на основе словаря А. А. Зализняка электронный словарь омоформ можно рассматривать как предельный случай словаря паронимов, когда учитывается словоизменительная парадигма, а расстояние между словами,

* Работа выполнена в рамках проекта № 99-04-12026в, поддержанного грантом РГНФ

образующими паронимическую связку, равно нулю. С алгоритмической точки зрения различие в способах построения обоих словарей сводится к тому, что в [6] используется техника отыскания точных (или "совершенных") повторов, тогда как в настоящей работе — более сложная техника обнаружения несовершенных повторов.

1. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ

Единого определения паронимов, как уже демонстрировалось выше, не существует. Мы будем придерживаться максимально широкой трактовки, содержащейся в [7]: "... слова близкие друг другу по звучанию, частичное совпадение внешней формы которых является случайным, т. е. не обусловлено ни семантикой, ни словообразовательными процессами". Чтобы использовать это определение в качестве рабочего инструмента, требуется ввести подходящую меру близости между словами. Таковой на начальном этапе может служить **редакционное расстояние**, понимаемое как минимальное число допустимых операций, переводящих одно слово в другое [8]. В качестве допустимых (редакционных) операций фигурируют "замена", "вставка", "перестановка" двух символов и т. п.

Ограничимся в первом приближении лишь каноническими словоформами, **исключив** тем самым **словоизменительную парадигму**, приводящую к появлению большого количества формально близких, но не интересующих нас в рамках данного исследования групп слов. Пусть S — словарь исходных канонических форм, $d(a,b)$ — редакционное расстояние между словами a и b, $|u|$ — длина произвольного слова u. Пару слов a и b из S будем считать паронимами, если

$$d(a,b)/\min(|a|,|b|) \le q, \qquad (1)$$

где q — фиксированный порог, обычно не превышающий 1/3. Приведенное определение конструктивно и в большинстве случаев согласуется с набранной нами значительной по объему подборкой паронимических текстовых ошибок.

В соответствии с (1) каждое слово может находиться в отношении паронимии с несколькими словами. D-окрестностью слова a из S назовем совокупность всех слов из S, удаленных от a (в смысле редакционного расстояния) не более чем на D. В этих терминах задача построения все более расширяющихся вариантов словаря паронимов сводится к вычислению D-окрестностей каждого слова, где D — монотонно нарастающий параметр ($D = 1,2,\ldots$ в предположении, что веса всех редакционных операций одинаковы и равны 1).

На практике в большинстве случаев целесообразно ограничиться двумя редакционными операциями ("замена" и "вставка", или "устранение", символа) и значениями $D = 1,2$. Очевидно, что при D=1 допустимо лишь одно искажение на слово. Иными словами, паронимы будут отличаться друг от друга только одной заменой (этот случай рассмотрен в [5]) или вставкой-устранением символа (этот случай рассматривается в данной работе и сопоставляется с [5]).

При D=2 слова, образующие паронимическую связку, могут различаться двумя заменами, или одной заменой и вставкой символа, или двумя вставками. Перестановка может трактоваться как специфический вариант замен в двух соседних (а иногда разнесенных) позициях.

2. ПОСТРОЕНИЕ СЛОВАРЯ ПАРОНИМОВ

Поскольку исходный словарь русского языка достаточно объемен (свыше 100 тыс. слов), желательно избежать сопоставления каждого слова с каждым, иначе вычислительные затраты будут пропорциональны квадрату числа слов. Поэтому словарь S разбивается на подмножества слов одинаковой длины j, так что $S = \cup S_j$ ($j \ge 2$). Не трудно видеть, что если в качестве редакционной операции используется замена символа, сопоставлению подлежат только слова внутри каждого из подмножеств S_j, а если вставки и D=1 — слова каждой пары соседних подмножеств S_j и S_{j+1} ($j = 2,3,\ldots$).

Процесс поиска соседей для элементов выделенного подмножества S_j строится итеративно по k, где k — номер позиции, в которой допускается замена (или вставка) символа. В случае замен имеем $1 \le k \le j$; для вставок $1 \le k \le j+1$. Значение $k = 1$ соответствует удлинению слова из S_j на один символ слева, $k = j+1$ — удлинению на один символ справа. На k-й итерации произвольное слово $a = a_1 a_2 \ldots a_k \ldots a_j$ из S_j преобразуется к виду:

$$a' = a_1 a_2 \ldots a_{k-1} x a_{k+1} \ldots a_j \text{ (в случае замен)}, \qquad (2)$$

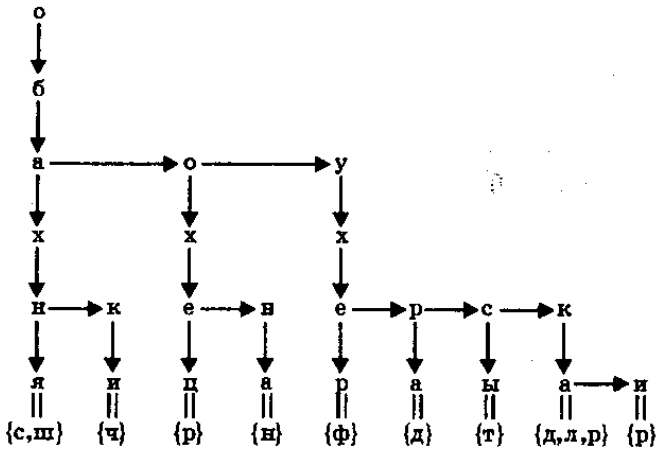
$$a'' = a_1 a_2 \ldots a_{k-1} x a_k a_{k+1} \ldots a_j \text{ (в случае вставок)},$$

где x — символ, не принадлежащий исходному алфавиту. Преобразование (2) делает неразличимыми ("склеивает") слова длины j, различающиеся только по k-й позиции. В случае вставок это преобразование делает неразличимыми пары слов из S_j и S_{j+1}, различающиеся вставкой в k-й позиции. Тем самым задача поиска несовершенных повторов, каковыми являются паронимы, сводится к существенно более простой задаче отыскания точных повторов. На k-й итерации в случае замен мы будем отыскивать совпавшие слова во множестве $S'_j(k)$, составленном из элементов множества S_j, представленных в форме a'. В случае вставок совпавшие слова отыскиваются среди элементов множества $S''_j(k) \cup S'_{j+1}(k)$, где $S''_j(k)$ содержит элементы из S_j, представленные в форме a''.

Сама процедура отыскания точных повторов сводится к компактному представлению множеств $S'_j(k)$ (в первом случае) и $S''_j(k) \cup S'_j(k+1)$ (во втором) в виде дерева, где каждому слову вида a' (или a'') соответствует свой путь от корня к листьям. Одинаковые префиксные части разных слов склеены, т. е. им соответствует общий путь. Конкретное слово a' в общем случае является результатом склеивания d разных слов из S_j, различающихся лишь по k-й позиции. Соответственно, в листе, которым заканчивается это слово, размещается список из d различных букв ($d \ge 1$), замещенных символом x. Если $d \ge 2$, этот список будем называть **вектором замен** для k-й позиции слова a' и обозначать z_k.

Заметим, что при отыскании вставок сначала строится дерево для слов из $S''_j(k)$ (при этом листовые списки пусты), а затем к нему добавляются лишь те слова из $S'_{j+1}(k)$, которые заканчиваются в одном из уже имеющихся листьев, поскольку только они могут образовывать паронимы. Остальные слова из $S'_{j+1}(k)$ в дерево не включаются с целью экономии памяти. Именно на этапе добавления к начальному дереву слов из $S'_{j+1}(k)$ в листьях формируются списки букв, замещенных символом x в k-й позиции. Список, соответствующий цепочке a'', будем называть **вектором вставок** слова $a_1a_2\ldots a_j$ по k-й позиции и обозначать b_k. В отличие от векторов замен векторы вставок могут содержать по одному элементу.

Процесс формирования деревьев на k-й итерации сводится к однократному просмотру всех элементов множеств $S'_j(k)$ (в случае замен) и $S''_j(k) \cup S'_{j+1}(k)$ (в случае вставок). Объединение результатов всех j итераций дает полный список паронимов для подмножества S_j исходного словаря. Строящиеся деревья в принципе могут содержать до n ветвлений в отдельных узлах, где n — размер исходного алфавита. Поскольку истинное число ветвлений заранее не известно, n-арное дерево преобразуется в бинарное с целью экономии памяти. Пример такого бинарного дерева для случая отыскания вставок приведен на рисунке. Параметр $j = 4$, $k = 3$, в качестве S_4 и S_5 взяты реальные подмножества этих множеств: $\tilde{S}_4 = \{$*баня, баки, боец, бона, буер, бура, бусы, бука, буки*$\}$ и $\tilde{S}_5 = \{$*басня, башня, бачки, борец, бонна, будра* (практически вышедшее из употребления), *буфер, бутсы, будка, булка, бурка* (в смысле "одежда"), *бурки* (в смысле "обувь")$\}$.



Дерево для обнаружения слов из $\tilde{S}_4$ и $\tilde{S}_5$, отличающихся вставкой в третьей позиции

Здесь вершины одного уровня, связанные линиями, соответствуют разветвлениям в исходном дереве. Так вершины *а, о, у* второго уровня связаны каждая с вершиной *б* в исходном дереве. Поэтому, например, путь от корня дерева к листу $\{$*д, л, р*$\}$ соответствует цепочке "бу x ка". Подстановка вместо x элементов вектора вставок $b_3 = \{$*д, л, р*$\}$ выявляет пары слов $\{$*бука-будка, бука-булка, бука-бурка*$\}$, отличающихся вставкой одного символа.

При поиске паронимов в более широкой окрестности ($D > 1$) описанная выше схема также применима, но трудоемкость существенно возрастает. В этом случае может быть использована техника поиска по групповому частично специфицированному запросу [9].

3. КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ СЛОВАРЯ ПАРОНИМОВ

Как уже говорилось выше, при $D = 1$ словарь паронимов распадается на две части в соответствии с тем, какая из операций — замена или вставка — использовалась при определении ближайшего соседа. Мы рассмотрим подсловарь со вставками и проведем сравнительный анализ с подсловарем, основанным на заменах.

3.1. Зависимость числа паронимов от длины слова

Используемый нами в качестве исходного словарь Д. Уорта содержит 100 960 канонических форм. В соответствии с длинами фигурирующих в нем слов он разбивается на подмножества S_j ($2 \le j \le 34$ за исключением $j = 32$ — слов такой длины в словаре не встретилось). Самые короткие слова (их всего 31) — существительные и частицы: *ад, ус, щи, бы, ли* и т. п.; самые длинные — сложные прилагательные типа "территориально-производственный" и т. п.

В табл. 1 приведены данные о числе слов M_j с непустой D-окрестностью ($D = 1$) для каждого из подмножеств S_j. Выделены три случая: а) D-окрестность формируется с использованием операции замены; б) с использованием операции вставки; в) с использованием обеих операций. Иначе говоря, в случае а) учитываются слова из S_j, которые имеют хотя бы одного соседа, отличающегося от данного слова заменой символа в любой позиции; в случае б) слово должно иметь соседа (одного или более), отличающегося вставкой, а в случае в) допускаются соседи любого типа. Случай в) соответствует объединению пересекающихся (в общем случае) подмножеств, выделяемых по схемам а) и б).

Отметим следующие закономерности, наблюдаемые в табл. 1.

1. **Степень проявления паронимии в языке весьма высока.** Примерно 35% всех канонических форм допускают хотя бы в одной из позиций подстановку (замену) символа, переводящую данное слово в другое слово того же словаря. Аналогичный показатель для вставок существенно ниже — примерно 11%. Подмножества слов, находящихся в отношении паронимии по заменам и вставкам, существенно пересекаются (особенно при малых j). Поэтому объединение их не слишком увеличивает долю слов, допускающих переход в другое (осмысленное) слово в результате однократного применения любой из рассматриваемых операций. Она составляет 38,2%.

Если вместо вставки использовать операцию устранения символа, то результаты окажутся несимметричными. Около 16% слов словаря (вместо 11% при вставках) допускают переход в другое слово того же словаря в результате выпадения символа. Отсутствие симметрии объясняется тем, что из двух (или более) разных слов может получиться одно и то же слово путем устранения одного символа. Объединение подмножеств слов, находящихся в отношении паронимии по заменам, вставкам и устранениям символа, покрывает (с учетом пересекаемости этих подмножеств) около 43% объема исходного словаря.

Таблица 1

Доля слов с непустой 1-окрестностью в подсловарях S_j

j	$\mid S_j \mid$	$\frac{\mid S_j \mid}{N}$	Замены		Вставки		Зам.+вст.	
			M_j	$\frac{M_j}{\mid S_j \mid}$	M_j	$\frac{M_j}{\mid S_j \mid}$	M_j	$\frac{M_j}{\mid S_j \mid}$
2	31	0,0..	29	93	23	74	31	100,0
3	435	0,4	413	95	296	68	429	98,6
4	1593	1,6	1364	85	670	42	1414	88,8
5	3678	3,6	2517	68	887	24	2645	71,9
6	6030	6,0	3182	53	1259	21	3464	57,4
7	9193	9,1	4421	48	1643	18	4833	52,6
8	12605	12,4	5850	46	1861	15	6348	50,4
9	13790	13,7	5405	39	1580	11	5902	42,8
10	13750	13,6	5035	37	1315	10	5445	39,6
11	11732	11,6	3536	30	594	5	3775	32,2
12	9227	9,1	2150	23	413	4	2356	25,5
13	6532	6,5	1034	16	188	3	1137	17,4
14	4578	4,5	462	10	81	2	529	11,6
15	2965	2,9	157	5	31	1	184	6,2
16	1827	1,8	55	3	16	0,9	71	3,9
17	1200	1,2	14	1,2	9	0,7	23	1,9
18	737	0,7	16	2,2	4	0,5	20	2,7
19	426	0,4	6	1,4	2	0,5	8	1,9
20	229	0,2	—	—	—	—	—	—
21	158	0,1	2	1,3	1	0,6	3	1,9
Итого:			35648	35,3	10872	10,8	38617	38,2

* j — длина слова, $\mid S_j \mid$ — число слов в S_j, N — объем исходного словаря, дроби даются в %.

2. **Способность к образованию паронимов зависит от длины слова.** Почти все короткие слова имеют соседей (см. проценты, приведенные в последнем столбце для $j = 2, 3, 4$). С увеличением длины слова эта способность монотонно падает. Слова, длина которых превышает 21 символ, уже не имеют соседей. Это, как правило, слова со сложной морфемной структурой, имеющие два и более корней, часто пишущиеся через дефис. Доля их не превышает 0,3% от объема словаря.

3. **Число соседей у слова также зависит от его длины.** Объективным показателем может служить среднее число соседей у слов фиксированной длины. Усреднение проводится лишь по тем словам из S_j, которые имеют соседей. Соответствующий показатель почти монотонно убывает с увеличением j. Для случая вставок, к примеру, имеем следующий ряд значений:

Длина слова	2	3	4	5	6	7	8	9	10
Среднее число соседей на слово	4,2	2,4	1,86	1,55	1,60	1,53	1,47	1,38	1,30

Если отказаться от дифференциации по длинам, число соседей можно охарактеризовать и другим способом. Так, в варианте со вставками 70% всех фиксируемых слов имеют одного соседа, 19% — двух, 6%— трех, 3% — четырех, 1% — пять соседей и свыше пяти соседей имеет 1% слов. В варианте с заменами примерно 50% слов имеют по одному соседу, 22% — по два, 11% — по 3,6% — по 4,4% — по 5 соседей и т. д.

Если рассматривать не средние значения, а отдельные слова-рекордисты, то здесь для $j = 2 \div 5$ рекорд (максимальное число соседей) сохраняется примерно на одном уровне, а далее начинает медленно падать с увеличением длины слова.

Для варианта со вставками это выглядит следующим образом:

Длина слова	2	3	4	5	6	7	8	9	10	11	12
Максимум соседей	10	11	11	11	9	9	8	8	6	5	3

В варианте с заменами рекордные значения существенно выше и убывают с увеличением длины слова не столь монотонно:

Длина слова	2	3	4	5	6	7	8	9	10	11	12	13
Максимум соседей	6	20	19	16	17	12	15	12	12	9	11	7

Приведем примеры слов-рекордистов разной длины.

Замены: $j = 3 \rightarrow$ *бок*. Размерности векторов замен по трем позициям: $\mid z_1 \mid = 9$, $\mid z_2 \mid = 5$, $\mid z_3 \mid = 9$. Состав векторов замен: $z_1 = \{$*б, д, с, к, н, р, т, ф, ш,*$\}$ $z_2 = \{$*о, а, е, у, ы,*$\}$ $z_3 = \{$*к, й, а, б, г, н, р, т, ш.*$\}$ Число соседей (на уровне замен) у слова *бок* равно двадцати: по первой позиции — *док, сок, кок, нок, рок, ток, фок, шок;* по второй — *бак, бек, бук, бык;* по третьей — *бой, боа, боб, бон, бор, бот, бош, бог.*

$j = 6 \rightarrow$ *полить*. Аналогично предыдущему имеем: $z_1 = \{$*п, м, с, х, з, д,*$\}$ $z_2 = \{$*о, а, и, я, ы,*$\}$ $z_3 = \{$*л, в, б, ж, п, ч, ш,*$\}$ $z_4 = \{$*и, о, с,*$\}$ в остальных позициях замены отсутствуют. Число соседей — 17: по первой позиции — *молить, солить, холить* и т. д.; по второй — *палить, пилить, пялить* и т. д. Заметим, что все векторы замен (за исключением z_4) — однородные, т. е. состоят либо из одних согласных, либо из гласных. Согласную *с* в z_4, нарушающую однородность, в эволюционном плане можно считать "исчезающей" (*полсть* — устаревшая форма от *полость*).

Вставки: $j = 4 \rightarrow$ *есть*. Размерности векторов вставок по позициям $k = 1 \div (j + 1)$ равны соответственно $\{11, 0, 0, 0, 0\}$, т. е. вставки возможны лишь в первой позиции (удлинение слова слева). Вектор вставок $b_1 = \{$*н, в, с, у, д, т, ж, л, м, ш, ч.*$\}$ Число соседей равно 11: *несть, весть, сесть, уесть, десть* (вышедшая из употребления единица счета писчей бумаги, равная 24 листам), *тесть, жесть, лесть, месть, шесть, честь.*

$j = 5 \rightarrow$ *поить*. Размерности векторов вставок — $\{3, 1, 7, 0, 0, 0\}$. Векторы вставок по позициям: $b_1 = \{$*о, с, у,*$\}$, $b_2 = \{$*л,*$\}$, $b_3 = \{$*б, в, л, ж, п, ш, ч,*$\}$, по остальным позициям вставки невозможны. Число соседей — 11: по первой позиции — *опоить, споить, упоить*; по второй — *плоить* (устаревшее — делать на чем-либо ряды параллельных волнообразных складок); по третьей — *побить, повить, полить, пожить, попить, почить, пошить.*

Данный пример в совокупности с последним примером по заменам (см. *полить*) иллюстрирует природу взаимосвязи между вставками и заменами. Нетрудно видеть, что вектор замен z_3 для третьей позиции слова *полить* совпадает с вектором вставок b_3 в третьей позиции слова *поить*. Такого рода совпадения имеют место, когда длина вектора вставок в какой-либо позиции больше или равна двум. В этом случае более длинные слова

в парах, различающихся вставкой, образуют пары (или группы) слов, связанных заменой (из сходства *поить — полить, поить — побить* и т. п. следует, что *полить — побить* также являются соседями, но уже в метрике замен). Сформулированная закономерность в существенной мере объясняет факт незначительного увеличения объема словаря паронимов при добавлении операции вставки (см. последний столбец табл. 1).

Отмеченная связь между вставками и заменами ослабевает при больших длинах слов, когда размерности векторов вставок не превышают единицу. Из той же табл. 1 видно, что при $j \geq 16$ вставки и замены действуют независимо и их доли (в %) суммируются.

3.2. Зависимость числа соседей от номера позиции

Речь идет о длинах векторов вставок и замен для разных позиций. В целом закономерности для обоих типов операций похожи. Проиллюстрируем их на примере вставок.

Количественной характеристикой зависимости числа соседей от номера позиции k может служить $|\bar{b}_k|$ — среднее число соседей по k-й позиции для слов фиксированной длины. Усреднение проводится по всем словам из S_j, допускающим вставку хоть в одной позиции. В табл. 2 приведены значения $|\bar{b}_k|$ для слов длиной от двух до 13 символов.

Не трудно видеть, что при малых длинах слов ($j = 2 \div 4$) максимальное число вставок допускается в начальной ($k = 1$) и конечной ($k = j + 1$) позициях, т. е. легче удлинить слово слева или справа, чем сделать вставку в середине. Примером такого рода удлинений слева может служить цепочка: *ад, лад, клад, оклад, доклад* (вопрос о максимально длинных цепочках остается открытым).

При средних длинах слов ($j = 5 \div 9$), по-прежнему, превалирует первая позиция, а вклад конечной резко уменьшается. Зависимость $|\bar{b}_k|$ от k становится монотонно убывающей. При больших длинах ($j \geq 10$) повышается роль второй позиции (из-за подстановок типа *по-про, в-вы*, например *подержаться — продержаться, вкатываться — выкатываться* и т. п.). Вставки в последних позициях уже практически запрещены (исключения составляют уменьшительные формы *(программа — программка)* и варианты типа *агротехник — агротехника)*.

Монотонность убывания $|\bar{b}_k|$ с ростом k нередко нарушается в словах большой длины (см., например, значение $|\bar{b}_5|$ для $j = 13$). Подобные эффекты объясняются спецификой морфемной структуры длинных слов. Так, существенный вклад в аномалию $|\bar{b}_5|$ при $j = 13$ вносят слова с длинной приставкой *пере: переучиваться — перешучиваться* и др. Важно отметить, что здесь вставка указывает на **межморфемную границу**. В словах большой длины, где вставки носят, как правило, единичный характер, этот эффект проявляет себя наиболее ярко *(франко-русский — франко-прусский, малосмысленный — малоосмысленный, двустворчатый — двухстворчатый* и т. п.).

4. БУКВЕННЫЙ СОСТАВ, ЧАСТОТА ВСТРЕЧАЕМОСТИ И ПОЗИЦИОННАЯ ПРИВЯЗКА ВЕКТОРОВ ВСТАВОК

В данном разделе обсуждается вопрос о том, какие векторы вставок встречаются часто, какие редко, существует ли привязка определенных векторов к конкретным позициям и чем это обусловлено. В основном рассматриваются одноэлементные векторы вставок, поскольку 70% слов, допускающих вставки, имеют всего по одному соседу. Более того, наиболее частые векторы вставок размерности 2 и выше формируются обычно из наиболее частых одноэлементных вставок, характерных для данной позиции.

Абсолютных запретов на использование каких-либо букв русского алфавита в качестве допустимых вставок практически не существует. Общесловарная статистика вставок по **всем позициям** содержит все элементы алфавита (за исключением твердого знака), однако частоты использования их сильно различаются. Наиболее часто в качестве вставки фигурируют буквы *р* (1341 раз) , *с* (1067), *д* (934), *о* (899), *н* (681), *к* (642), *т* (603), *ы* (532), *а* (502); наиболее редко — *ю* (22), *ф* (24), *ц* (32), *э* (34), *щ* (47), *ж* (51). Эта статистика существенно коррелирует с общесловарной статистикой употребления символов (особенно по низкочастотным элементам). Однако имеются символы, ранги которых в обоих упорядочениях значительно различаются:

Символ α	*д*	*ы*	*у*	*р*	*с*	*ч*	*ь*	*й*	*е, ё*	*а*	*и*	*я*
$r_{слов}(\alpha) - r_{вст}(\alpha)$	13	7	7	6	6	6	-12	-11	-9	-7	-7	-6

Таблица 2

Среднее число вставок по k-й позиции в словах длины j ($1 \leq k \leq j + 1$).

k / j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	2,3	0,26	1,65											
3	0,92	0,35	0,35	0,78										
4	0,50	0,29	0,33	0,32	0,43									
5	0,52	0,33	0,19	0,20	0,10	0,21								
6	0,71	0,28	0,18	0,14	0,08	0,11	0,10							
7	0,51	0,33	0,20	0,16	0,12	0,10	0,06	0,05						
8	0,53	0,29	0,24	0,15	0,08	0,09	0,06	0,03	0,02					
9	0,44	0,38	0,19	0,12	0,09	0,05	0,04	0,04	0,01	0,01				
10	0,32	0,33	0,27	0,14	0,11	0,08	0,02	0,02	0,01	0,00	0,01			
11	0,27	0,34	0,21	0,10	0,12	0,09	0,06	0,02	0,01	0,01	0,00	0,01		
12	0,25	0,29	0,21	0,11	0,11	0,10	0,06	0,04	0,00	0,00	0,01	0,00	0,00	
13	0,41	0,21	0,08	0,06	0,16	0,08	0,11	0,02	0,02	0,00	0,00	0,03	0,00	0,01

Символы с положительной разностью рангов относительно чаще фигурируют в качестве вставок, чем это можно было ожидать, исходя из их встречаемости в словаре. Обычно для каждого из них существует предпочтительная позиция, в которой данный символ доминирует над остальными. За подобным доминированием всегда просматривается определенный механизм образования вставки. Так, подстановки *по — под*, *на — над* приводят к доминированию вставки *д* в третьей позиции *(посадить — подсадить, написать — надписать)* и т. п.

Символы со значительной отрицательной разностью рангов относительно редко фигурируют в качестве вставок. Например, такие символы как *ь* и *й* заметно проявляют себя только в позициях $k = j + 1$, где j — длина слова *(намолот — намолоть, больно — больной)*, поэтому их лишь с натяжкой можно считать вставками.

Наконец, символы с разностью рангов, равной или близкой к нулю, ведут себя в роли вставок нейтрально: они не проявляют тенденции к устойчивому доминированию в определенной позиции, но и не избегают каких-либо позиций. Таковы буквы *б, л, м, п* и другие (исключение составляет *н*, см. ниже).

Отмеченные выше случаи доминирования той или иной вставки в конкретной позиции обычно связаны с **морфемной структурой** слов (наличием приставок, суффиксов, окончаний). Как следствие, они проявляют себя при значительных длинах слов ($j \geq 6$). Приведем примеры наиболее характерных доминирований. Они приходятся на начальные ($k = 1,\ 2,\ 3$) и конечные ($k = j - 2,\ j - 1,\ j,\ j + 1$) позиции:

Номер позиции	Доминирующие вставки	Примеры слов-соседей
$k = 1$	*с, о, у, в* (лидеры меняются при разных j)	*бить — сбить — обить — убить — вбить*
$k = 2$	*р, ы* (подстановки *по — про*, *в — вы*)	*полить — пролить* *вход — выход*
$k = 3$	*д* (подстановки *по — под*, *на — над*)	*посадить — подсадить* *наземный — надземный*
⋮	⋮	⋮
$k = j - 2$	*н* (*н — нн*)	*строченый — строченный*
$k = j - 1$	*н, ч*	*вопросик — вопросник* *строка — строчка*
$k = j$	*к*	*трава — травка* *половина — половинка*
$k = j + 1$	*а, ь, й*	*ворон — ворона* *собрат — собрать* *прямо — прямой*

В указанных позициях характер доминирования достаточно устойчив (сохраняется в большом диапазоне длин слов) и ярок (наблюдается заметное различие между частотами доминирующих — взяты в скобки — и недоминирующих символов-вставок в фиксированной позиции). Приведем несколько примеров:

Длина слова	Номер позиции	Частоты вставок, ранжированные по убыванию
$j = 6$	$k = 1$	[*с* — 80, *у* — 49, *о* — 46, *в* — 34], *п* — 15, ...
	$k = 5$	[*н* — 24, *ч* — 13], *а* — 8, *с* — 7, ...
	$k = 6$	[*к* — 118], *н* — 7, *ц* — 5, ...
$j = 7$	$k = 5$	[*н* — 73], *р* — 19, *с* — 13, ...
	$k = 6$	[*н* — 89], *а* — 17, *и* — 9, *ч* — 8, ...
	$k = 7$	[*к* — 78], *н* — 5, *ц* — 3, ...
	$k = 8$	[*а* — 20, *ь* — 15, *й* — 14], *н* — 5, ...
$j = 9$	$k = 1$	[*с* — 90, *о* — 81], *п* — 61, *у* — 52, ...
	$k = 2$	[*р* — 157, *ы* — 127], *т* — 50, *о* — 24, ...
	$k = 3$	[*д* — 95], *р* — 18, *с* — 13, ...
	$k = 4$	*и* — 19, *с* — 18, *о* — 16, *ы* — 15, ...

Последняя строка ($j = 9,\ k = 4$) приведена для контраста, в ней отсутствуют доминирующие элементы.

Проведенный количественный анализ доминирующих вставок в фиксированных позициях подтверждает тезис о том, что "осмысленное" варьирование слов тесно связано с их морфемной структурой и выявляет основные механизмы образования слов-соседей.

5. СГ-СОСТАВ ВЕКТОРОВ ВСТАВОК

Данный раздел в определенном смысле суммирует и дополняет результаты предыдущего, поскольку речь, по-прежнему, идет о составе векторов вставок, но уже в агрегированном алфавите {С — согласный, Г — гласный}. Отдельно рассматривается "мягкий знак", но доля его в общем балансе вставок незначительна. Именно в этом алфавите удобно анализировать векторы вставок из двух и более элементов, поскольку число разновидностей их велико. Вектор вставок с $|b_k| \geq 2$ будем называть однородным, если он состоит только из одних гласных или согласных, и неоднородным — в противном случае. Представляет интерес количественная оценка проявлений неоднородности, поскольку в этом случае слова-соседи часто различаются по своим грамматическим характеристикам.

Пусть, как и ранее, j — длина слова, k — номер позиции, в которой появляется вставка ($1 \leq k \leq j + 1$), b_k — вектор вставок в k-й позиции, составленный из элементов {С, Г}, $F(b_k)$ — частота его встречаемости для слов из S_j. Анализ векторов вставок выявляет следующие закономерности.

1. В каждом конкретном слове вставка между двумя гласными, как правило, имеет тип "С" (*поить — побить, повить, полить* и т. д.), а между согласными — тип "Г" (*ткать — такать, тикать, тыкать* и т. п.). На стыках между "СГ" и "ГС" возможны разные варианты (*как — каюк — каяк, бак — банк, барк, баск, пот — порт, пост, поэт*).

2. Для одноэлементных векторов вставок имеет место $F(\text{С}) > F(\Г)$ практически для всех j и k (кроме $k = j + 1$), т. е. вставки согласных преобладают над вставками гласных, часто значительно. Так, для $j = k = 6$ имеем $F(\text{С}) = 139$, тогда как $F(\Г) = 0$; для $j = 8$, $k = 3$ – $F(\text{С}) = 328$, $F(\Г) = 49$ и т. д. При $k \neq j + 1$ наблюдается лишь несколько отклонений от этого правила, в частности, для $j = 11$, $k = 8$: здесь $F(\text{С}) = 0$, а $F(\Г) = 10$ (*переизбрать – переизбирать* и т. п.).

В позиции $k = j + 1$ (расширение слов длины j вправо) чаще всего $F(\Г) > F(\text{С})$ ($j = 3, 5, 6, 8, 9$ и т. д.) из-за доминирования гласных в окончаниях (*механик – механика*, *забыть – забытье*). Заметную роль играет для этой позиции вставка мягкого знака, составляющего до 15–20% всех вставок при $j = 6, 7$.

3. При выполнении в целом соотношения $F(\text{С}) > F(\Г)$ (а часто и более сильного — $F(\text{С}) >> F(\Г)$) для одноэлементных векторов b_k ($k = 1 \div j$) баланс между согласными и гласными в разных позициях может варьировать весьма заметно. В частности, для слов с длиной $j \geq 6$ существует позиция k^*, обычно приходящаяся на середину слова, в которой отмечается локальный рост числа вставок гласных по сравнению с соседними позициями. Так, для $j = 7$ имеем:

$k = 3$:	$F(\text{С}) = 186$,	$F(\Г) = 19$;
$k = 4$:	$F(\text{С}) = 120$,	$F(\Г) = 85$;
$k = 5$:	$F(\text{С}) = 145$,	$F(\Г) = 14$,

т. е. k^*=4. Вставка гласной в этой позиции по большей части имеет место между двумя согласными (*назвать – называть*, *нервный – неровный* и т. д.).

4. При $|b_k| \geq 2$ проявления неоднородности наиболее заметны в позициях $k = 1, 2$. Доля неоднородных по СГ-составу векторов в этих позициях колеблется от 5 до 25% ($k = 1$) и от 3,5 до 6% ($k = 2$). В остальных позициях эффектом неоднородности можно пренебречь.

6. О ВОЗМОЖНОСТИ ПРАКТИЧЕСКОГО ИСПОЛЬЗОВАНИЯ ЭЛЕКТРОННОГО СЛОВАРЯ ПАРОНИМОВ*

6.1. Лингвистическая комбинаторика

Данный термин фигурирует в названии книги М. М. Маковского [10]. Актуальность комбинаторного подхода к изучению языковых единиц он мотивирует тем, что "именно комбинаторные преобразования являются неисчерпаемым источником возникновения в языке новых слов и значений". Созданный электронный словарь паронимов позволяет количественно оценить основные тенденции этого процесса. В сочетании же с другими словарями (толковыми, частотными и пр.) он обеспечивает возможность проведения более сложных исследований для ответа на вопросы типа: "каков тот предел, после достижения которого замена, вставка, перестановка, увеличение или уменьшение элементов комбинации (например, фонетических составляющих слова) ведут к изменению значения?" [10].

Следует отметить также, что электронный словарь паронимов может быть использован в учебно-методических целях для конструирования различного рода лингвистических задач и головоломок, требующих значительного перебора (одна из них, связанная с максимальными лево- и правосторонними расширениями слов, упомянута в разделе 3.2).

6.2. Обнаружение ошибок паронимического типа

Эти ошибки весьма распространены и не выявляются с помощью существующих автоматических корректоров. Вместе с тем, проводимые нами исследования помехоустойчивости лексического состава русского языка в сочетании с контекстным анализом подборки паронимических ошибок позволяют уже сейчас очертить множество наиболее "ошибкоопасных" (в интересующем нас смысле) слов и оценить возможности автоматического обнаружения "ложных паронимов" с помощью методов, не требующих полного грамматического разбора предложения. Более подробно этот вопрос рассмотрен в [11].

6.3. Сжатие словарей

В некоторых приложениях (например, таких как автоматическое обнаружение орфографических ошибок) требуется хранить в оперативной памяти компьютера достаточно объемные словари русского языка. При дефиците оперативной памяти актуальной становится проблема сжатия словарей. Для этих целей может быть использовано явление паронимии (всех соседей слова можно хранить в виде векторов замен и вставок, избегая к тому же в большинстве случаев дублирования сопутствующей грамматической информации). Более трудоемкой, однако, становится процедура поиска нужного слова, которая требует отдельного рассмотрения.

6.4. Формирование "трудных" тестовых словарей для систем распознавания и синтеза речи

Задача состоит в том, чтобы из словаря паронимов выбрать подсловарь, содержащий связки слов, мало различающихся по артикуляционно-акустическим характеристикам несовпадающих в них звуков. Для этого нами создана фонетическая версия электронного словаря паронимов на основе предварительно затранскрибированного (т. е. переведенного в звуковую запись) словаря русского языка. Объем этого словаря обеспечивает пользователю достаточную свободу выбора тестовых подсловарей в автоматическом режиме.

ЗАКЛЮЧЕНИЕ

Проблема вариативности структурных единиц является одной из центральных для эволюционирующих во времени языковых систем. Под вариативностью применительно к словам русского языка мы понимаем способность слова переходить в другое

* С интерактивным вариантом ресурса можно познакомиться в сети Internet по адресу: http://il.m8.math.nsc.ru/ paronym (в кодировке Win-1251)

слово из заданного словаря в результате незначительного его искажения с помощью операций типа "замена", "вставка", "устранение" символа и т. п. Пары слов, различающихся в указанном смысле, мы называем паронимами, имея в виду наиболее широкую трактовку этого термина.

Работа посвящена исследованию количественных характеристик вариативности как отдельных слов, так и языка в целом. Базой исследования послужил электронный словарь паронимов русского языка. Его создание потребовало формализации понятия "паронимы" и разработки эффективного алгоритма их поиска. Проведен сравнительный анализ подсловарей, построенных с использованием: а) только операции замены; б) операций вставки/устранения символа. Рассмотрены возможности практического использования электронного словаря паронимов.

СПИСОК ЛИТЕРАТУРЫ

1. Вишнякова О. В. Словарь паронимов русского языка.— М.: Рус. яз.. 1984 .— 348 с.

2. Бельчиков Ю. А., Панюшева М. С. Словарь паронимов современного русского языка.— М.: Рус. яз., 1994 .— 455 с.

3. Колесников Н. П. Словарь паронимов русского языка.— Тбилиси, 1971 .— 427 с.

4. Worth D., Kozak A., Jonson D. Russian Derivation Dictionary.— N.-Y., 1970 .— 747 p.

5. Саломатина Н. В. Создание и исследование компьютерного словаря паронимов // Анализ данных и сигналов. Вып. 163. Вычислительные системы.— Новосибирск, 1998 .— С. 97–112.

6. Сумбатян М. А., Хазагеров Г. Г. Типы русских омоформ и их автоматическое разведение // НТИ. Сер. 2 .— 1997 .— № 12 .— С. 35–37.

7. БСЭ. Т. 19 .— М.: Сов. энциклопедия, 1975 .— 647 с.

8. Wagner R. A., Fisher M. J. The string — to — string correction problem // J. ACM.— Jan. 1974 .— Vol. 21, № 1 .— P. 168–173.

9. Гусев В. Д., Немытикова Л. А. Алгоритмы поиска в текстовых базах данных по групповому частично специфицированному запросу // Искусственный интеллект и экспертные системы. Вып. 157. Вычислительные системы.— Новосибирск, 1996 .— С. 12–39.

10. Маковский М. М. Лингвистическая комбинаторика: опыт топологической стратификации языковых структур.— М.: Наука, 1988 .— 231 с.

11. Гусев В. Д., Саломатина Н. В. Анализ ошибок, не выявляемых автоматическими корректорами // Квантитативная лингвистика и семантика (КВАЛИСЕМ-99): Тез. докл. II Межвузовской конф., Новосибирск, 12–15 октября 1999 г.— С. 8–12.

Материал поступил в редакцию 29.12.99.