АНАЛИЗ ДАННЫХ И СИГНАЛОВ
(Вычислительные системы)
1998 год Выпуск 163

УДК 519.769

СОЗДАНИЕ И ИССЛЕДОВАНИЕ КОМПЬЮТЕРНОГО СЛОВАРЯ ПАРОНИМОВ

Н.В. Саломатина

Введение

Наметившаяся в последние годы тенденция к формализации и автоматизации гуманитарных (в том числе языковых) исследований предполагает не только механический перенос в память компьютера накопленных данных, знаний, но и создание с помощью компьютера новых ресурсов, получение которых традиционным для этой области ручным способом затруднительно. Примером такого рода ресурса является описанный в данной работе компьютерный (или электронный) словарь паронимов, формирование которого потребовало формализации понятия "паронимы" и разработки эффективного алгоритма их поиска.

Одним из стимулов для создания электронного словаря паронимов послужило исследование подборки текстовых ошибок, не выявляемых существующими автоматическими орфографическими и синтаксическими корректорами. Количественный анализ ее показал, что наиболее часто встречающимися ошибками подобного рода являются паронимические. В связи с этим возникает вопрос о степени распространенности паронимов в языке, об их специфических особенностях, которые могут быть использованы для обнаружения паронимических ошибок, о наиболее "ошибкоопасных" словах.

Имеющиеся словаря паронимов русского языка (Вишняковой О.В. [1], Бельчикова Ю.А. и Панюшевой М.С. [2], Колесникова Н.П. [3]) составлены вручную, что не гарантирует полноты подборок. 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 Колесникова Н.П. 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 русским языком, а поэтому часть из них является паронимами, а часть — нет (например, "казаться" и "касаться"). Последние в словарь не включаются. Анализ нашей подборки ошибок обнаруживает, тем не менее, множество случаев, когда ошибка имела место, но соответствующие пары слов не фигурировали в словаре Колесникова Н.П. в качестве паронимов.

Целью работы является описание процедуры формирования электронного словаря паронимов русского языка и получение количественных характеристик вариативности как языка в целом, так и отдельных слов. Под вариативностью будем понимать способность одних слов переходить в другие из данного словаря в результате незначительного изменения их буквенного состава. Базой для создания электронного словаря паронимов послужил достаточно представительный словарь русского языка объемом свыше 100 тыс. слов [4].

1. Формализация понятия "паронимы"

Единого определения паронимов не существует. Мы будем придерживаться максимально широкой трактовки этого понятия, представленной в [5]: "...слова близкие друг другу по звучанию, частичное совпадение внешней формы которых является случайным, т. е. не обусловлено ни семантикой, ни словообразователь-

ными процессами". Для создания электронного словаря паронимов нам потребуется формально определить их, в связи с чем необходимо ввести подходящую меру близости между словами. Таковой, нам представляется, может служить редакционное расстояние, понимаемое как минимальное число допустимых операций, переводящих одно слово в другое [6]. В качестве допустимых (редакционных) операций могут быть использованы такие как замена, вставка, перестановка двух символов и т.п. Пусть S — исходный словарь канонических форм. Пару слов a и b из S будем считать паронимами, если величина редакционного расстояния между ними, отнесенная к длине минимального из них, не превышает фиксированного порога q. Величина q обычно не превосходит $1/3$, для длинных слов она существенно меньше. Приведенное определение, возможно, носит излишне общий характер, но конструктивно и в большинстве случаев согласуется с набранной нами значительной по объему подборкой паронимических текстовых ошибок, не выявляемых существующими автоматическими корректорами. В соответствии с приведенным определением каждое слово может находиться в отношении паронимии с несколькими словами. Поэтому удобно ввести еще одно определение. D–окрестностью слова a из словаря S назовем совокупность всех слов из S, удаленных от a (в смысле редакционного расстояния) не более чем на D. Тогда задача построения все более расширяющихся вариантов словаря паронимов сводится к вычислению D–окрестностей каждого слова, где D — регулируемый нарастающий параметр ($D = 1, 2, \ldots$ в предположении, что веса всех редакционных операций одинаковы и равны 1). На практике в большинстве случаев целесообразно будет ограничиваться значениями $D = 1, 2$. Очевидно, что при $D = 1$ допустимо лишь одно искажение на слово (паронимы будут отличаться друг от друга только одной заменой или вставкой–устранением символа). При $D = 2$ слова, образующие паронимическую пару, могут отличаться двумя заменами, или одной заменой и одной вставкой-устранением символа, или двумя вставками–устранениями. Перестановка может трактоваться как специфический вариант замен в двух соседних (а иногда разнесенных) позициях.

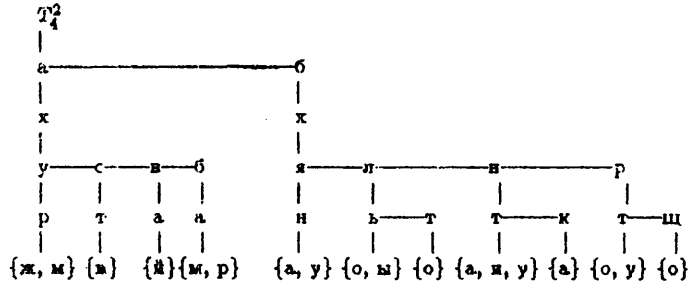
2. Построение словаря паронимов

Для создания словаря паронимов был использован электронный словарь русского языка объемом свыше 100 тыс. слов. Чтобы избежать сопоставления каждого слова с каждым (иначе вычислительные затраты будут пропорциональны квадрату числа слов в словаре) весь словарь S был разбит на подмножества слов одинаковой длины S_j, так что $S = \bigcup S_j$, где $j \geq 2$ — длины слов словаря. В этом случае достаточно ограничиться сопоставлением нескольких соседних (по длине представленных в них слов) подмножеств, гарантируя в то же время полноту поиска. В случае, когда в качестве редакционной операции используется только замена символа, поиск ведется внутри одного и того же подмножества. Именно этот вариант с $D = 1$ и был реализован при создании описываемой (первой) версии электронного словаря паронимов.

Для компактного представления подмножеств S_j использовалась традиционная структура данных — бинарное дерево [7]. Исследуемая на вариативность позиция считалась неопределенной, т.е. заменялась переменной x во всех словах из S_j.

Пусть s_i — i-е слово подсловаря S_j, $s_i = a_1 \dots a_k \dots a_j$, где k - позиция символа в слове ($k = 1, \dots, j$), а a_k — символ алфавита A. Нужно найти всех соседей s_i в окрестности $D = 1$, используя редакционную операцию замены в некоторой позиции k ($1 \leq k \leq j$), т.е. выявить все слова, содержащиеся в словаре и отличающиеся от данного лишь одной буквой в k-й позиции. При построении бинарного дерева T_j^k все слова из подмножества S_j словаря записываются в форме: $\hat{s}_i = a_1 \dots a_{k-1} x a_{k+1} \dots a_j$. Всякий раз при прохождении узла дерева, соответствующего позиции x, символ встречающийся в слове в этой позиции, запоминается в листе пройденной (или достроенной) ветви. Именно наборы символов, записанные в листьях дерева, задают возможные варианты паронимов для каждого s_i. Пример бинарного дерева, построенного по подмножеству слов S_4' = { "ажур", "аист", "айва", "амба", "амур", "арба", "банк", "бант", "баян", "бинт", "болт", "боль", "борт", "борщ", "бунт", "бурт", "буян", "быль "}, с x во второй позиции приведен ниже, на рисунке.

Так как соседи, образующиеся у слова в случае замены символа, имеют ту же длину, что и само слово, то полный обход построенного дерева позволяет выписать всех соседей, если они имеются, для каждого слова длины j в k–й позиции. Проведя таким образом исследование всех позиций, получим подсловарь паронимов, состоящий из списков слов, отличающихся заменой одного символа в произвольной позиции слова, для всех слов подсловаря S_j. Объединение всех подсловарей паронимов, полученных по всем S_j, представляет собой полный словарь паронимов для случая замены одного символа в слове.



Представление S_4' в виде бинарного дерева.

Чтобы получить список соседей, образующихся в результате однократных вставок символа в k–й позиции слова s_i из S_j, следует, записав s_i в виде: $\hat{s}_i = a_1 \dots a_{k-1} x a_k a_{k+1} \dots a_j$, осуществить его поиск по T_{j+1}^k. Осуществив эту процедуру для всех $1 \le k \le j+1$ (вставка в позиции $j+1$ соответствует записи x в конце слова длины j) и $j \ge 2$, получим списки слов–паронимов, отличающихся вставкой одного символа.

Для получения списков паронимов, отличающихся выпадением одного символа, требуется осуществить поиск в T_j^k каждого слова из S_{j-1}, представленного в форме: $\hat{s}_i = a_1 \dots a_{k-1} x a_k a_{k+1} \dots a_{j-1}$.

Объединение результатов по всем позициям k при фиксированном j и всем S_j позволяет составить списки паронимов, отличающихся выпадением одного символа.

Для случаев $D > 1$ может быть использована техника поиска по групповому частично специфицированному запросу [8].

3. Количественные характеристики словаря паронимов

1. **Число соседей.** Исследуемый электронный словарь паронимов составлен на базе словообразовательного словаря Д. Уорта, содержащего 100 960 канонических форм, с применением редакционной операции замены символа для случая $D = 1$. Предварительно словарь Д. Уорта был разбит на 32 подмножества в соответствии с длинами составляющих его слов, которые изменяются от 2 до 34, исключая слова длины 32. Слова такой длины в словаре не встретились. Самые короткие (всего 31 слово в словаре) — существительные, частицы: "ад", "ус", "щи", "бы", "ли"; самые длинные — сложные прилагательные: "территориально-производственный", "частнопредпринимательский". Во втором столбце табл. 1 указан объем полученных в результате разбиения словаря подмножеств в зависимости от длин содержащихся в них слов. Количество слов в подмножестве дано в % от общего числа слов в словаре. Ни у одного из слов длиной от 22 до 34 символов нет соседей. Это, как правило, слова сложной морфемной структуры с двумя или несколькими корнями, часто пишущиеся через дефис. В сумме они составляют 0,3 % от всех слов словаря. Поэтому данные о них в таблицу не помещены.

Поскольку основой словаря паронимов, в котором каждое слово сопровождается списком всех его соседей, имеющих ту же длину и отличающихся одной буквой, послужил словарь канонических форм, то словоизменительная парадигма слова в списках соседей отсутствует (считается, что элементы ее паронимами не являются).

В процессе построения словаря паронимов было выяснено, что 35 % слов словаря Д. Уорта имеют соседей. Каждое из них допускает хотя бы в одной из позиций определенную подстановку (замену), переводящую это слово в осмысленное слово из того же словаря (в данном случае из того же подмножества). Очевидна зависимость наличия соседей у слова от его длины (см. третий столбец табл. 1, где число слов, имеющих соседей, указано в % от общего числа слов данной длины). Почти все короткие слова имеют соседей. Максимальное суммарное (по всем позициям) число соседей — двадцать — имеет слово "бок": в первой позиции — "док", "сок", "кок", "нок", "рок", "ток", "фок", "шок", во

второй — "бак", "бєк", "бук", "бык", в третьей — "бой", "боа", "боб", "бон", "бор", "бот", "бош", "бог".

Таблица 1

Интегральные характеристики слов словаря Д. Уорта с ненулевым числом соседей

Число символов в слове	Всего слов в словаре (в %)	Всего слов с числом соседей > 0 (в %)	max соседей у слова
2	<0.1	93	6
3	0.4	95	20
4	1.6	85	19
5	3.6	68	16
6	6.0	53	17
7	9.1	48	12
8	12.4	46	15
9	13.7	39	12
10	13.6	36	12
11	11.6	30	9
12	9.3	23	11
13	6.5	16	7
14	4.5	10	5
15	3.0	5	4
16	1.8	3	2
17	1.2	1	1
18	0.7	2	1
19	0.4	1	1
20	0.2	–	–
21	0.1	1	1

Короткие слова, не имеющие соседей, — это, прежде всего, иноязычные ("атом", "лье", "фру"), старорусские ("се", "выя", "блуд"), а также и просто редко употребляемые ("алый", "внук", "вдох"). Однако в этот список попали и довольно часто встречающиеся в текстах слова, например, "дитя", "дно", "зло", "опыт".

Их можно трактовать как наиболее устойчивые к ошибкам типа "замена" слова. Частота встречаемости определялась по частотному словарю русского языка (под редакцией Л. Н. Засориной).

По суммарному числу соседей во всех позициях слова в словаре паронимов распределяются следующим образом: примерно 50 % слов имеют по одному соседу, 22 % слов имеют по два соседа, 11 % слов — по 3 соседа, 6% слов — по 4 соседа, 4% слов — по 5 соседей, более семи соседей — у менее, чем 1 % слов.

Если рассмотреть, как меняется число соседей у слова при замене символа в определенной позиции, то можно отметить следующие закономерности: в коротких словах ($l = 3, 4$; l — длина слова) наибольшее число соседей у слова встречается при замене символа в начальной и конечной позиции. Это, как правило, замены начальных и конечных символов в корневых морфах. В более длинных словах ($l = 5, \ldots, 8$) число соседей у слова уменьшается при замене символа в позициях близких к конечной, т. к. в этих словах есть стандартные суффиксы и окончания канонических форм.

У слова "заживать", например, соседи в первой позиции — "важивать", "хаживать", "саживать", "наживать", в третьей — "забивать", "завивать", "закивать", "заливать", "запивать", "зашивать", в четвертой — "зажевать", в пятой — "зажигать", "зажимать", "зажинать", в шестой — "заживить", во второй и в последних двух — соседей нет. Начала же слов часто совпадают с корневым морфом, и число соседей при замене символов в начальной позиции у слова по-прежнему велико. При образовании соседей довольно распространенным случаем является омонимия на морфемном уровне — совпадение при замене одной буквы структур одного слова "префикс + корень" и "корень + суффикс" с корневым морфом другого слова.

Интересно отметить, что по последовательности чисел, соответствующих максимальному количеству соседей в определенных позициях слов фиксированной длины, можно определить, какого звукового качества (гласный или согласный) символы преимущественно встречаются в этой позиции. Число гласных почти в два раза меньше числа согласных и количество соседей, образующихся при замене гласных, существенно меньше, чем число

соседей, образующихся у слова при замене согласных, в случае когда "гласный" заменяется на "гласный" или "согласный" на "согласный". Возможные случаи замены "гласный" на "согласный" не нарушают эту закономерность. Например, у 8-буквенных слов максимальное число соседей при замене символа в позициях с первой по восьмую таково: 5, 4, 6, 4, 6, 3, 1, 1; соответственно, в словах длины 8 с ненулевым числом соседей в четных позициях преимущественно встречаются гласные, в нечетных — согласные.

Слова с $l > 8$ символов часто начинаются с приставки, поэтому число соседей при замене символа в слове в начальных позициях невелико. Для 12-буквенных слов, например, максимальное число соседей при замене символа в позициях с первой по двенадцатую таково: 3, 2, 5, 2, 5, 3, 6, 2, 2, 1, 1, 1. Чем больше число соседей у слова, тем меньше таких слов в словаре вне зависимости от того, в какой позиции произведена замена символа.

2. Характерные замены символов.

Длина векторов замен. Чтобы понять, в результате замены каких символов друг на друга одно слово преобразуется в другое слово словаря, рассмотрим все возможные наборы букв, составленные при исследовании таких преобразований. Пусть a_k^i — k-й символ слова s_i. Совокупность символов $z_k = a_k^i \bigcup \{\bigcup_{m=1}^{n} a_k^m\}$ (a_k^m — символы в k-й позиции каждого из n слов-соседей, образующихся у s_i при замене символа в этой позиции) назовем вектором замен в k-й позиции. Например, у слова "корка" в первой позиции составлен вектор замен, имеющий пять элементов: $z_1 = \{к, г, н, п, ш\}$, во второй — три: $z_2 = \{о, у, и\}$, в третьей и четвертой — восемь и четыре, соответственно: $z_3 = \{р, в, й, л, н, п, ш, ч\}$ и $z_4 = \{к, д, м, ч\}$. В результате подстановки в указанном слове в определенной позиции любой буквы из соответствующего вектора замен получается осмысленное слово. В частности, при подстановке букв в четвертой позиции в слове "корка" из вектора замен {к, д, м, ч} получаем три соседа: "корда", "корма" и "корча".

Обозначим длину z_k — число элементов в векторе замен — через $d(z_k)$. Тогда $\sum_k (d(z_k) - 1)$ равна суммарному числу соседей

у слова по всем позициям. В полученном нами словаре паронимов $d(z_k)$ изменяется от двух до одиннадцати.

Проведенный анализ векторов замен показал, что все z_k, рассмотренные в сумме по всем словам словаря и позициям в слове, с $d(z_k) > 6$ — встречаются не более одного раза. Для z_k с частотой встречаемости больше единицы

$$\max_{k,l} d(z_k) = \begin{cases} l, & \text{если } l = 3, \ldots, 6, \\ 6, & \text{если } l = 7, \ldots, 12, \\ 3, & \text{если } l = 13, 14, \\ 2, & \text{если } l \geq 15. \end{cases}$$

Количество различных z_k существенно уменьшается с ростом $d(z_k)$, тогда как число возможных z_k растет степенным образом. Максимальное число разных векторов замен (суммарно по всем позициям), равное 278, наблюдается у слов с $l = 6$ и реализуется на z_k с $d(z_k) = 2$ (примерно 30 % от числа возможных).

Буквенный состав векторов замен. Интересно рассмотреть элементный состав векторов замен с точки зрения наличия в нем гласных (Г) и согласных (С) букв. Всего в словаре (в сумме по всем позициям слов всех длин) 48 разных СГ-типов векторов замен, которые отличаются длиной — от двух до одиннадцати элементов — и СГ-составом. Мягкий знак рассматривается как специфический символ в СГ записи. Например, слова "коря" и "корь" имеют вектор замен в СГ-форме в конечной позиции — {Г, Ь}. Будем считать замены однородными, если вектор замен состоит только из гласных или только из согласных. Абсолютная частота встречаемости F (в сумме по всем позициям в слове и словам всех длин) нескольких самых часто встречающихся типов векторов замен, представленных в СГ-виде, в зависимости от длины вектора замен (от $d(z_k) = 2$ до $d(z_k) = 5$) дана в табл. 2.

Из табл.2 видно, что при коротком векторе замен $d(z_k) = 2$ наиболее частыми являются однородные замены, когда гласный заменяется на гласный, согласный на согласный. В сумме они составляют свыше 90 %. Данный факт может служить мотивацией использования явления паронимии для компактного представления словарей, поскольку при однородном векторе замен

Таблица 2

СГ-состав и частота встречаемости
наиболее высокочастотных векторов замен

$d(z_k)$	СГ-состав	F	Примеры векторов замен
2	СС	13670	зн, бт, дп, лр, св
	ГГ	6064	аи, ои, яи, ае, еи
	ГС	1444	ус, ос, ыз, ив, ыс
	ГЬ	156	аь, еь, оь, яь, уь
	СЬ	50	вь, мь, сь, ть, чь
3	ССС	2053	спд, мпк, впч, жвп, нтп
	ГГГ	460	аиу, аое, аеу, еия, еио
	ГСС	312	усв, осв, усш, осш, укп
	ГГС	132	оус, оув, ояк
	ССЬ	17	мнь, всь
4	ССЗС	570	зптл, висд, зптч, бвлш, дрзш
	ГГЗС	109	оувс
	ГГГГ	42	аеоу, аиоу, еоуы, аиоы, аеиу
	ГСЗС	36	увгс, обвд
	ГГГС	6	аияк, аиот
5	ССЗСС	211	бвжлш, вджпш, жлпст, бвплш, бдмпт
	ГСЗСС	13	обвдж
	ГГГГГ	7	аеиоу
	ГГГСС	3	всоуя
	ГГСЗС	3	аезмя

слова-паронимы, как правило, имеют одинаковые грамматические характеристики, которые не понадобится дублировать.

В векторах замен с $d(z_k) \geq 4$ неоднородные замены встречаются реже, чем однородные типа "согласный" на "согласный", но чаще, чем однородные типа "гласный" на "гласный". Самый длинный и единственный вектор замен $z_k = \{$у, н, в, с, д, т, ж, л, м, ш, ч$\}$ с $d(z_k) = 11$ соответствует максимуму соседей в первой позиции слова "уесть". В третьем столбце табл.2 помещены примеры в буквенной форме самых частых векторов замен (с упорядочением по частоте встречаемости, которая всегда больше единицы). Легко заметить, что часто встречающиеся более длинные векторы замен, состоящие, например, из гласных, содержат в своем составе более короткие: $\{a, u\} \subset \{a, u, y\} \subset \{a, u, o, y\} \subset \{a, e, u, o, y\}$. Запятые, разделяющие элементы векторов замен, и обрамляющие фигурные скобки в табл. 2 (и далее — в табл. 3)

для экономии места опущены. Аналогичная тенденция вложенности при удлинении вектора замен проявляется, хоть и не столь последовательно, и в том случае, когда векторы состоят только из согласных $\{д, н\} \subset \{д, н, с\} \subset \{в, д, н, с\}$, или имеют смешаный состав: $\{у, с\} \subset \{у, с, в\} \subset \{у, с, в, г\}$. Собственная частота встречаемости букв, из которых состоят высокочастотные векторы замен, не всегда самая высокая, особенно в векторах замен, содержащих только согласные.

Если рассмотреть векторы замен по подмножествам слов одинаковой длины, то самое большое разнообразие векторов замен, в том числе и в СГ-форме, а также максимум их длины (что уже отмечалось при рассмотрении всех соседей) приходится на короткие слова. С ростом длины слова количество разных векторов замен уменьшается, равно как и максимум их длин. Неизменно самыми частыми являются векторы $z = \{СС\}$ у слов всех длин, кроме самых коротких ($l = 2, 3$) и слов с $l = 17$, где самые частые замены — гласный на гласный. Максимальное число разных векторов замен у слова обнаружено в словах с $l = 8$ в позиции три. Самые частые векторы {СС} и {ГГ} также наблюдаются у 8-буквенных слов в позициях один и четыре, соответственно.

В табл. 3 помещены самые часто встречающиеся векторы замен в определенных позициях слов с их абсолютными частотами. По горизонтали указаны номера позиций в слове, а по вертикали — длины слов. Незаполненные клеточки в табл. 3, соответствующие элементному представлению векторов замен с записью "ед." в строке частот в некоторых позициях, означают, что все встретившиеся в этой позиции слова векторы замен единичные. Прочерки соответствуют отсутствию соседей в данной позиции слова. Слова длины два имеют только единичные векторы замен.

Состав векторов замен в первую очередь определяется морфемной структурой слова, а не дифференциальными фонетическими признаками входящих в них букв. Векторы замен, позиционно относящиеся к аффиксам, отражают зависимую от словообразования паронимию, тогда как векторы замен в корневых морфах — в основном, процесс образования семантически не связанных слов-соседей.

Таблица 3

Распределение векторов замен по позициям в слове

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	пт 2	еу,оу 7	ксф 2											
4	со,св 5	оу 17	ао 11	аь 13										
5	нз 27	ае,ау 22	лт 9	ио 26	аи 9									
6	кп,сув 12	оу 29	ои 26	еи 22	ио 18	аь 12								
7	нз 83	бт 48	ои 20	ае,ео 16	аи 66	оы 20	иь 10							
8	су 43	бт 156	ои 59	ае 38	лр 26	аи 129	оы 26	йд 8						
9	су 82	бт 63	ио 19	лр 22	аи 50	дз 12	иа 95	нц 5	йк 4					
10	дп 117	бт 176	ио 88	уо 36	лр 25	аи 73	лр 10	ия 55	нц,оы 2	як 5				
11	дп 54	бт 51	ио 144	аи,вм 14	ао,аи 13	лр 19	ия 63	тв 4	аи 18	иь 3	ед.			
12	нз 135	бт 65	ио 56	аи 10	лр 14	ыи 11	лр 12	ия 41	нч 3	ия 19	ед.	ед.		
13	нз 45	бт 16	ио 51	вм,ск 7	аи 6	лр 10	иы 4	вч,гч 3	ия 3	ия 6	ед.	ед.	–	
14	нз 29	бт 9	ио 16	еу 2	лр 4	ау 3	кч,гч 3	иы 2	чш 2	ия 2	ед.	ия 3	ед.	–
15	др 8	бд 2	ио 8	ед.	рс 3	ао,рм 2	ау 3	ед.	–	–	ия 5	–	–	–
16	др,кп 2	ед.	лт 2	–	–	–	–	–	–	–	–	ия 3	–	–
17	рн 3	аи 2	–	–	–	–	–	–	–	–	–	–	–	–
18	др 4	–	–	–	–	–	–	–	–	–	–	–	–	–
21	ед.	–	–	–	–	–	–	–	–	–	–	–	–	–

В первой позиции слов с $l > 4$ самые частые векторы замен содержат, в основном, начальные буквы взаимозаменяемых приставок (или сами приставки) и частиц: "на", "за"; "с", "у" , "в"; "дс", "по"; "де", "ре", "не". Во второй позиции самые частые векторы замен тоже состоят большей частью из вторых символов приставок (для слов с $l > 6$): "от", "об". Аналогично в третьей позиции — "про", "при".

Исследуя изменения в составе векторов замен по позициям, легко проследить как меняется морфемная структура слова при росте его длины. В позициях, близких к конечной, векторы замен — части суффиксов (или сами суффиксы): "ок", "ик"; "и", "я" ("скупать", "скупить"), "ива", "ыва", "ен", "еч", "ин", "ищ"; а также части заменяющихся окончаний: "ой","ый".

С удлинением слова самые частые векторы замен, отражающие варьирование в суффиксах, встречаются в позициях с большим номером. Состав векторов замен в промежуточных позициях — между приставкой и суффиксом — отражает возможные замены в корневых морфах. В конечной позиции слов векторы замен могут возникать в результата омонимии основ как разных частей речи "бита" и "бить", так и одинаковых "коксохимия", "коксохимик" (иногда со сменой ударения в словоформе: "ломота" и "ломоть").

Очевидна привязка некоторых самых частых векторов замен, например, $\{б, т\}$ и $\{о, и\}$ к определенной позиции в слове (второй и третьей, соответственно). Еще одним таким характерным примером может служить вектор неоднородных замен $\{с, о, у\}$, встретившийся 83 раза:

	Длина слова	Част. встр.	Номер поз.	Част. встр.
	4	2	1	2
	5	2	1	2
	6	5	1	5
	7	19	1	19
СОУ	8	10	1	10
(83)	9	18	1	18
	10	15	1	15
	11	8	1	8
	12	3	1	3
	13	1	1	1

Четкая привязка встречаемости вектора замен только к первой позиции слова определяется тем, что слова-соседи ("сбить", "обить", "убить") имеют взаимозаменяемые приставки "с", "о", "у", участвующие в словообразовательном процессе. Однако есть векторы замен, не имеющие столь явной позиционной устойчивости. Иллюстрацией может служить вектор $z = \{к, м\}$, встретившийся 189 раз:

	Длина слова	Част. встреч. в слове	Частота встречаемости в позиции № 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7
	4	3	2	–	–	1	–	–	–
	5	22	7	–	3	3	6	–	–
	6	14	3	2	2	3	3	1	–
	7	15	4	2	3	1	1	1	3
	8	19	7	–	10	2	–	–	–
КМ	9	21	2	12	5	2	–	–	–
(189)	10	32	1	1	29	–	1	–	–
	11	22	2	6	6	7	–	1	–
	12	21	–	1	18	–	2	–	–
	13	10	2	1	1	6	–	–	–
	14	5	1	–	3	–	1	–	–
	15	1	–	–	–	1	–	–	–
	16	3	3	–	–	–	1	–	–
	$\sum$	189	34	25	80	29	15	3	3

Распределение частоты встречаемости вектора замен $\{к, м\}$ по позициям в данном примере отражает общую тенденцию изменения числа соседей у слов словаря по позициям.

Существуют векторы замен, встречающиеся во всех позициях слов (лишь у слов с $3 \leq l \leq 8$). Их число невелико (всего 40) и $d(z_k) = 2$, а состав зависит от l. Так, $z_k = \{с, т\}$ встречается в словах всех длин, кроме $l = 3, 5$. Самый частый $z_k = \{а, и\}$ — в словах с $l = 7$.

Заключение

Полученный словарь паронимов по сути является списком ближайших окрестностей слов русского языка. Он содержит ка-

нонические формы, которые при замене в них определенным образом одной буквы (из вектора замен) преобразуются в осмысленные слова, что часто способствует подмене одного слова другим в тексте. Число "ошибкоопасных" слов довольно велико — в словарь паронимов попала треть слов словаря русского языка, и это при условии, что использовалась лишь единственная редакционная операция (замена) и рассматривалась ближайшая окрестность ($D = 1$). По этому параметру наблюдается принципиальное различие с известными [1, 2, 3] словарями паронимов. Основная масса слов имеет от одного до трех соседей. Замены символов, в результате которых они образуются, — чаще однородные по качеству звучания (гласный или согласный). Хотя соседи возникают у слов в случае замены символов в морфах всех типов, но, в основном, превалируют в корнях (в начальных и конечных позициях) и аффиксах, списки которых были составлены.

Электронный словарь паронимов может быть использован:

а) для получения более реалистичных оценок вероятности обнаружения ошибки существующими автоматическими корректорами, не учитывающими возможность перехода (в результате замены, вставки и т. д.) осмысленного слова в осмысленное же;

б) для компактного представления традиционных электронных словарей (всех соседей слова можно хранить в виде векторов замен, избегая к тому же, в большинстве случаев, дублирования сопутствующей грамматической информации);

в) для автоматизации поиска рифм, заголовков, оборотов, построенных на игре "слов" и т. п.

Просматриваются реальные перспективы развития словаря в направлении *расширения* числа *редакционных операций* и *комбинированного* их использования, *учета словоформ* наряду с каноническими формами, использования *транскрибирования* для выявления *"речевых паронимов"*.

Автор выражает искреннюю благодарность Гусеву Владимиру Дмитриевичу за привлечение внимания к проблеме количественного исследования проявлений вариативности в языковых системах, а также за существенные замечания, сделанные в ходе подготовки статьи к печати.

Литература

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

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

5. БСЭ. Т. 19, — М.: Советская энциклопедия, 1975. — 647 с.

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

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

8. ГУСЕВ В.Д., НЕМЫТИКОВА Л.А. Алгоритмы поиска в текстовых базах данных по групповому частично специфицированному запросу // Искусственный интеллект и экспертные системы. — Новосибирск, 1996. — Вып. 157: Вычислительные системы. — С.12-39.

Поступила в редакцию
29 декабря 1998 года