ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

1999 год Выпуск 166

УДК 519.769 : 801.314.4

ОПРЕДЕЛЕНИЕ И АНАЛИЗ БЛИЖАЙШИХ ОКРЕСТНОСТЕЙ КОРНЕЙ СЛОВ РУССКОГО ЯЗЫКА[1]

В.Д. Гусев, Н.В. Саломатина

Введение

Во многих приложениях, связанных с автоматизацией обработки текстовой информации, возникает необходимость во введении формальных[2] мер близости между словами. Укажем, в частности, на задачи обнаружения орфографических и паронимических ошибок, компактного представления больших словарей в памяти компьютера, формирования трудных тестовых словарей для систем распознавания и синтеза речи и др. Достаточно естественной формальной мерой близости между словами может служить *редакционное расстояние* [1]. Его прототипом является *метрика Левенштейна* [2], предложенная в связи с рассмотрением ошибок синхронизации в системах связи. Применительно к генетическим текстам та же метрика фигурирует под названием *"эволюционного" расстояния* [3]. Использование этой меры в различных языковых системах свидетельствует об ее универсальности и позволяет проводить межъязыковые аналогии. В

[1]Работа выполнена в рамках проекта № 99-04-12026в, поддержанного грантом РГНФ.

[2]Термин "формальная мера близости" несет в нашем случае двойную семантическую нагрузку: с одной стороны, он подразумевает алгоритмизуемость процедуры вычисления расстояний; с другой стороны, подчеркивает, что речь не идет о близости значений (смыслов) сравниваемых слов.

частности, полезными для нас в идейном плане оказались работы по изучению точечных мутаций в генетических текстах. Одним из конкретных практических результатов этих исследований явилось формирование матрицы близостей для аминокислотного алфавита.

Введение формальной меры близости между словами позволяет определить для каждого слова его ближайшую окрестность. Она может содержать как разрешенные для данного языка слова, так и запрещенные (ошибочные). В дальнейшем нас будут интересовать лишь слова, допускаемые данным языком. С учетом этого соглашения ближайшей окрестностью слова α будем называть совокупность всех слов языка, минимально отличающихся от α в заданной метрике. Состав окрестностей для слов разной длины, характер отличий слов-соседей друг от друга и распределение отличий по длине слова представляют интерес как в теоретическом отношении (изучение механизмов словообразования), так и в практическом (выявление наиболее и наименее "ошибкоопасных" слов и позиций, осмысленных агрегирований элементов алфавита и т.п.).

В [4] на материале достаточно объемного (свыше 100 тыс. канонических форм) словаря русского языка [5] проведено количественное исследование ближайших окрестностей всех слов для случая, когда слова-соседи отличаются лишь заменой одного символа в произвольной позиции[3]. Анализ подмножеств слов фиксированной длины показал четкую позиционную привязку распределения определенных типов замен по длине слова. Эта привязка носит характер доминирования той или иной замены в конкретной позиции. Наиболее характерные доминирования приходятся на начальные и конечные позиции слов и несут существенную информацию об их морфемной структуре. В то же время средние (корневые) позиции слов не обнаруживают заметного доминирования той или иной замены над остальными. Этому может быть дано двоякое объяснение. С одной стороны, можно предполагать, что корень, как наиболее устойчивая структурная единица слова,

[3] Аналогичная работа с анализом вставок и выпадений символов подготовлена к печати и выставлена в сети Internet по адресу http://iitm8.math.nsc.ru/paronym

менее вариативен, чем слово в целом, т.е. в среднем характеризуется меньшим числом "допустимых" замен, вставок и т.п. С другой стороны, возможно, и имеет место доминирование некоторых замен в определенных позициях корней, но при анализе слов оно завуалировано тем, что даже в словах фиксированной длины корни могут иметь различную позиционную привязку.

Для выявления того, какой из указанных факторов имеет место, целесообразно провести отдельное исследование вариативности корней по той же схеме, которая применялась для слов [4]. Как и в [4] вариативность понимается в самом широком смысле, а именно, как возможность перехода одного корня в другой в результате незначительного искажения его буквенного состава.

Целью работы является выявление *ближайших окрестностей* всех[4] *корней* русского языка и получение *количественных характеристик вариативности* как отдельных корней, так и их подмножеств (например, всех корней фиксированной длины). Базой для проведения работы послужил электронный *словообразовательный* словарь русского языка объемом свыше 100 тыс. единиц [5], в котором проведено членение слов на морфемы. Тем самым процедура выделения корней из слов становится простой формальностью.

1. Обозначения и определения

Пусть u и v — произвольные цепочки символов, составленные из элементов алфавита Σ. Для обозначения длин цепочек и размера алфавита будем использовать одинаковую символику: $|u|$, $|v|$, $|\Sigma|$. В нашем случае Σ — это русский алфавит, а u и v — пары слов или корней русского языка. *Редакционным расстоянием* между цепочками u и v называется минимальное число допустимых операций, переводящих одну цепочку в другую [1]. В качестве допустимых (редакционных) операций обычно фигурируют "замена", "вставка", "устранение" символа и некоторые

[4]Термин "всех" применяется по отношению к исходному словарю русского языка, достаточно объемному, но все же ограниченному

другие. Операции могут иметь разные веса, тогда минимизируется не число операций, переводящих u в v, а суммарная их стоимость. Мы ограничимся для простоты случаем двух операций ("замена", "вставка") и будем полагать веса этих операций равными единице.

Пусть S — словарь исходных канонических форм, $K = K(S)$ — словарь корней, выделенных из S. Очевидно, что $|K| \ll |S|$, поскольку в K каждый корень фигурирует по разу, а в S он может быть представлен во многих канонических формах. Подмножества всех слов длины j из S (или корней длины j из K) будем обозначать S_j (соответственно K_j). Цепочки u и v будем считать близкими, если $d(u,v)/\min(|u|,|v|) \le q$, где q — фиксированный порог, который выбирается с учетом длин u и v (обычно q не превышает $1/3$). *D-окрестностью* слова α из S (или корня β из K) назовем совокупность всех слов из S (соответственно корней из K), удаленных от α (или β) не более чем на D в метрике редакционного расстояния. Окрестность, соответствующую минимальному отличному от нуля значению D (в нашем случае это $D = 1$), будем называть *ближайшей* и обозначать $O(\alpha)$ для слова α из S или $O(\beta)$ для корня β из K. Нетрудно видеть, что ближайшую окрестность корня β, например, составляют все корни из $K_{|\beta|}$, отличающиеся от β заменой символа в одной из позиций, а также те корни из $K_{|\beta|+1}$, которые отличаются от β вставкой символа в любой из позиций.

Зафиксируем конкретный корень $\beta = a_1 a_2 \dots a_{k-1} a_k a_{k+1} \dots a_j$, где j — длина β, k —номер позиции. Рассмотрим все корни из $O(\beta)$ вида $\beta' = a_1 a_2 \dots a_{k-1} a'_k a_{k+1} \dots a_j$, с $a'_k \ne a_k$, $a'_k \in \Sigma$. Совокупность $\{ a'_k \}$, упорядоченную определенным образом (например, по алфавиту), естественно назвать вектором замен в корне β по k й позиции. Однако при таком определении векторы для β и всех близких ему корней вида β' окажутся разными, что затруднит их анализ. Этого не произойдет, если включить в совокупность $\{ a'_k \}$ сам заменяемый символ a_k. С учетом этого соглашения *вектором замен* z_k будем называть упорядоченную по алфавиту совокупность элементов алфавита $a_k \bigcup \{a'_k\}$. Тогда, если в k-й позиции корня β допустимы замены, то длина вектора замен $d_k = |z_k| \ge 2$ и имеет место $z_k(\beta) = z_k(\beta')$.

Подобных коллизий, обусловленных бинарным характером операции "замена", не возникает при рассмотрении вставок. Пусть, как и раньше, $\beta = a_1a_2\dots a_{k-1}a_ka_{k+1}\dots a_j$, — произвольный корень из K_j. Выделим из $O(\beta)$ все корни вида $\beta'' = a_1a_2\dots a_{k-1}a''_ka_ka_{k+1}\dots a_j$, где $\beta'' \in K_{j+1}$, $a''_k \in \Sigma$. Будем говорить, что корни вида β'' получаются из β вставкой одного (каждый раз разного) элемента алфавита в k-ю позицию, $k = 1, 2, \dots, j, j+1$. Здесь $k = 1$ соответствует удлинению β на 1 символ слева, а $k = j+1$ — на один символ справа. Совокупность всех допустимых элементов $\{a''_k\}$ назовем *вектором вставок* корня β по k-й позиции и будем обозначать b_k. В отличие от векторов замен векторы вставок могут содержать по одному элементу.

ПРИМЕР 1. Корень "враж" допускает замены по всем позициям. Длины векторов замен для позиций $k = 1 \div 4$ составляют соответственно 5, 2, 2, 5. Векторы замен: z_1 = (Б, В, Д, К, П), z_2 = (Л, Р), z_3 = (А, Е), z_4 = (Г, Ж, Т, Ч, Щ). В ближайшую окрестность корня "враж" входят корни: "браж", "драж", "краж", "праж" (замена по первой позиции); "влаж" (по второй); "вреж" (по третьей); "враг", "врат", "врач", "вращ" (по четвертой).

Длины векторов вставок для этого же корня составляют 2, 1, 0, 0, 1 ($k = 1 \div 5$). Векторы вставок: b_1 = (О, У), b_2 = (О), b_5 = (Д); в позициях $K = 3, 4$ допустимых (анализируемых словарем) вставок не существует. Таким образом, с учетом вставок ближайшая окрестность корня "враж" пополняется еще четырьмя корнями: "овраж", "увраж" (богато иллюстрированное художественное издание большого формата) — вставки в позиции 1; "вораж" — поз. 2; "вражд" — поз. 5.

ПРИМЕР 2. Всего 14 корней длины 3 *не допускают замен* ни в одной из позиций: "вне" (внешний), "нэп" (аббревиатура), "жди" (иждивение), "дзе" (дзекание), "ибо" (предлог), "имя" (имя), "ишь" (частица), "фру" (фру), "цве" (выцветший), "этн" (этнический), "юкк" (юкка — вечнозеленое растение), "язв" (язва), "ямб" (ямб), "айв" (айва).

Корни "швед", "шхун", "пятн", "вдов" и пр. (длина $j = 4$), а также "гнезд", "сахар", "снабж", "дежур" и другие ($j = 5$)

не допускают однократных замен и вставок ни по каким позициям.

2. Построение ближайших окрестностей корней

На первом этапе из исходного словаря S, содержащего свыше 100 тыс. канонических форм [5], формируется словарь корней $K(S)$. Эта процедура носит преимущественно формальный характер, поскольку в S корневые морфемы уже выделены. Значительных усилий потребовала коррекция довольно многочисленных ошибок членения, выполненная канд. филол. наук Л.С. Юдиной. В словаре $K(S)$ все корни (в том числе омонимичные) представлены по разу. Из многокорневых словоформ в $K(S)$ включался основной корень.

Для построения ближайших окрестностей корней использовался тот же подход, что и для канонических форм. Он коротко описан в [4], а более детально — в электронной публикации, выставленной в сети Internet (http://i1m8.math.nsc.ru/paronym), а также в статье, принятой к печати журналом НТИ (серия 2) в 2000 г. Здесь изложим лишь идею подхода.

Словарь предварительно разбивается на подмножества корней одинаковой длины j, так что $K = \bigcup K_j$, $1 \le j \le 15$. В случае, когда в качестве редакционной операции используется "замена" символа, поиск ближайших соседей ведется для каждого K_j независимо. Если используется "вставка", ближайшие соседи выявляются путем сопоставления элементов двух соседних подмножеств — K_j и K_{j+1} ($j = 1, 2, \ldots$). Заметим, что выделяемые при малых j пары (группы) корней лишь формально можно называть соседями, поскольку они не удовлетворяют сформулированному выше критерию близости.

Процесс поиска соседей среди элементов выделенного подмножества K_j (или пары подмножеств K_j и K_{j+1}) ведется итеративно по k, где k — номер позиции, в которой допускается замена (или вставка) символа. На k-й итерации произвольный корень $\beta = a_1 a_2 \ldots a_{k-1} a_k a_{k+1} \ldots a_j$ из K_j преобразуется к виду:

$$\begin{aligned} &\beta' = a_1 a_2 \ldots a_{k-1} x a_{k+1} \ldots a_j \text{ (в случае замен)}, \\ &\beta'' = a_1 a_2 \ldots a_{k-1} x a_k a_{k+1} \ldots a_j \text{ (в случае вставок)}, \end{aligned} \qquad (*)$$

где $x \notin \Sigma$. Преобразование к виду β' делает неразличимыми ("склеивает") корни длины j, отличающиеся только по k й позиции. Преобразование к виду β'' удлиняет корни из K_j на 1 символ, а после приведения корней из K_{j+1} к виду β' неразличимыми оказываются пары корней из K_j и K_{j+1}, отличающиеся лишь вставкой в k–й позиции. Тем самым задача выявления ближайших соседей сводится к отысканию точных повторов. В случае замен совпавшие "слова" отыскиваются во множестве $K'_j(k)$, составленном из элементов множества K_j, представленных в форме β'. В случае вставок совпавшие "слова" отыскиваются среди элсмснтов множества $K''_j(k) \bigcup K'_{j+1}(k)$, где $K''_j(k)$ содержит элементы из K_j, представленные в форме β''.

Сама процедура отыскания точных повторов может быть реализована разными способами: лексикографической сортировкой, хешированием и т.п. Нами использован способ упаковки "слов" из $K'_j(k)$ (в первом случае) и $K''_j(k) \bigcup K'_{j+1}(k)$ (во втором) в виде дерева, где каждому слову вида β'(или β'') соответствует путь от корня к листьям, при этом одинаковые префиксные части разных слов склеены, т.е. им соответствует общий путь. С целью экономии памяти полученное n–арное дерево ($n = \mid \Sigma \mid$), затем преобразуется в бинарное. Заметим, что описанная схема наилучшим образом подходит для отыскания именно ближайших соседей, расстояние между которыми равно 1.

3. Сравнительный анализ словарей S и $K(S)$ (канонических форм и корней)

Словарь Уорта и др. [5] содержит 100960 канонических форм. Из них выделено 13230 корней, включая омонимичные. В словаре $K(S)$ омонимичные корни представлены в одном экземпляре. Поэтому различных корней в $K(S)$ всего 11691. Длины корней меняются в диапазоне от $j = 1$ (19 корней) до 15 (2 корня), при этом $arg\max_j \mid K_j \mid = 5$, $\max_j \mid K_j \mid = \mid K_5 \mid = 2837$, т.е. наибольшим разнообразием характеризуются корни длины 5, составляющие почти четверть всех корней. Полностью зависимости числа корней $\mid K_j \mid$ и канонических форм $\mid S_j \mid$ от длины j приведены в табл. 1 (спектр величин $\mid S_j \mid$ обрезан на значении $j = 15$). Некоторые канонические формы из S совпадают со своими корнями.

Поэтому они представлены как в S, так и в $K(S)$. В табл. 1 указано число таких совпадений для разных j в абсолютном выражении (т.е. $|S_j \bigcap K_j|$), а также в процентах: $|S_j \bigcap K_j| / |K_j|$.

Нетрудно видеть, что при малых j, $j = 2 \div 4$, $|K_j| > |S_j|$ причем доля корней, образующих самостоятельные слова (т.е. совпадающих с канонической формой), невелика. Эти факторы, как увидим далее, оказываются существенными при объяснении различий в длинах векторов замен и вставок элементов из S и $K(S)$ (речь идет о рекордных показателях). В целом доля корней из $K(S)$, образующих самостоятельные слова, составляет 38,7%.

Таблица 1

Количественные характеристики словарей S и $K(S)$

j	$\lvert S_j \rvert$	$\lvert K_j \rvert$	$\lvert S_j \bigcap K_j \rvert$	$\lvert S_j \bigcap K_j \rvert / \lvert K_j \rvert$
1	—	19	—	—
2	31	252	31	12,3%
3	435	1322	360	27,2%
4	1593	2451	765	31,2%
5	3678	2837	1020	35,9%
6	5030	2179	972	44,6%
7	9193	1284	638	49,7%
8	12605	799	442	55,3%
9	13790	365	206	56,4%
10	13750	108	56	51,8%
11	11732	44	22	50%
12	9227	16	10	62,6%
13	6532	8	2	25%
14	4578	5	2	40%
15	2965	2	0	0
$\sum$	96139	11691	4526	38,7%

Поскольку в дальнейшем речь пойдет о статистике допустимых замен и вставок в разных позициях корней, важно знать, коррелирована ли эта статистика с частотой употребления элементов алфавита в корнях (т.е. в множестве $K(S)$). Информация

Таблица 2

Частоты встречаемости элементов алфавита в S и $K(S)$
(α — элемент алфавита, $F(\alpha)$ — его частота, $r(\alpha)$ — его ранг)

№	α	$F(\alpha)$ в K	$F(\alpha)$ в S	$r_S(\alpha)$	Δ
1	А	5922	85130	2	-1
2	Р	5520	59270	7	-5
3	Е	4362	68021	6	-3
4	О	4057	87412	1	3
5	Т	4017	75306	3	2
6	Л	3778	37195	12	-6
7	Н	3778	69394	5	2
8	И	3759	71501	4	4
9	К	3458	38299	11	-2
10	С	2907	55380	8	2
11	М	2197	19649	19	-8
12	П	2080	33076	13	-1
13	У	2015	20960	17	-4
14	Д	1851	21770	16	-2
15	Б	1581	14820	21	-6
16	Г	1513	12368	23	-7
17	В	1397	43613	9	8
18	Ф	866	4644	28	-10
19	З	864	18191	20	-1
20	Ш	858	7391	24	-4
21	Ч	743	14189	22	-1
22	Ц	629	6758	25	-3
23	Ж	618	6722	26	-3
24	Х	589	5745	27	-3
25	Ь,Ъ	491	39451	10	15
26	Я	399	19920	18	8
27	Ю	383	2235	30	-3
28	Ы	337	30016	15	13
29	Э	328	1540	31	-2
30	Й	249	30310	14	16
31	Щ	139	4110	29	2

о частоте встречаемости элементов алфавита в S и $K(S)$ приведена в табл. 2. Элементы алфавита упорядочены по убыванию частоты их встречаемости в $K(S)$, т.е. ранг r_K элемента $\alpha \in \Sigma$ определяется его порядковым номером в этом упорядочении. Ранг r_S того же элемента в упорядочении, полученном на S, указывается в отдельном столбце. Представляет также интерес разность рангов одних и тех же элементов в обоих упорядочениях: $\Delta = r_K(\alpha) - r_S(\alpha)$ (см. последний столбец).

Наибольший интерес в табл. 2. представляют элементы с аномально высокими (по модулю) разностями рангов в обоих упорядочениях. Большие положительные значения Δ соответствуют элементам алфавита "Й", "(Ь, Ъ)", "Ы", "Я", "В". Это означает, что данные элементы доминируют в аффиксах, окружающих корневую морфему, т.е. часто входят в состав префиксов, суффиксов, окончаний. Большие (по модулю) отрицательные значения Δ характеризуют буквы, чаще встречающиеся в корнях, чем в аффиксах (М, Ф, Г, Б, Л, Р). В целом по отрицательным значениям Δ различия между словарями не столь ярки, как по положительным.

4. Количественные характеристики вариативности корней

Как уже отмечалось выше, ближайшая окрестность каждого корня (если она непустая) состоит из элементов двух типов: а) корней, отличающихся от исходного заменой символа в одной из позиций; б) корней, отличающихся вставкой символа в одну из позиций. Случай а) характеризуется векторами замен по разным позициям, случай б) — векторами вставок. Разные корни из K_j могут характеризоваться одинаковыми допустимыми векторами замен (или вставок) в определенных позициях. Статистики этих векторов для разных подмножеств и являются предметом нашего рассмотрения.

4.1. *Зависимость соседей от длины корня.* Здесь и далее будем использовать следующие обозначения: j — длина корня, k — номер позиции в корне ($1 \le k \le j$ — для замен, $1 \le k \le j+1$ — для вставок), K_j — множество корней длины j в словаре; $| K_j |$ — их число; K_j^1 (соответственно K_j^2) — подмножество

корней из K_j с непустой 1–окрестностью по заменам (вставками), $O'_j(\beta)$ (соответственно $O''_j(\beta)$) множество ближайших соседей корня длины j, отличающихся от него одной заменой (вставкой), $Rec'_j = \max_\beta \mid O'(\beta) \mid$ (соответственно $Rec''_j = \max_\beta \mid O''(\beta) \mid$) — рекордные значения числа соседей по заменам (вставкам) для корней длины j.

В табл.3 приведены данные о числе корней длины j с непустой 1–окрестностью: а) по заменам; б) по вставкам; в) по заменам и/или вставкам. Формально таблица начинается со значения $j = 2$, хотя говорить о соседях корней столь малой длины можно лишь условно, и заканчивается значением $j = 11$, поскольку корней большей длины очень мало (см. табл. 1) и они практически не допускают варьирования.

Отметим наиболее существенные закономерности, наблюдаемые в табл.3.

1. Примерно 61% всех корней допускают хотя бы в одной из позиций замену символа, переводящую данный корень в другой из анализируемого словаря. Для канонических форм аналогичный показатель был гораздо ниже (примерно 35%). Это частично объясняется тем, что искаженный корень в общем случае требует другого аффиксального окружения, чем исходный, поэтому "допустимая" замена в корне может оказаться "недопустимой" для содержащей его словоформы.

Корни из K_j^1 с непустой 1–окрестностью по заменам можно классифицировать по числу соседей: примерно 26% корней имеют одного соседа, 14% — двух, 9% — трех и т.д. Аналогично, 48,5% корней из K_j^2 с непустой 1-окрестностью по вставкам имеют одного соседа, 18,8% — двух, 8,8% — трех и т.д.

2. Вариативность корней по заменам (61%) значительно выше, чем по вставкам (23%). Подмножества K_j^1 и K_j^2 существенно пересекаются, поэтому объединение их (т.е. возможность использования как замен, так и вставок при формировании ближайшей окрестности) не слишком увеличивает долю корней с непустой 1–окрестностью (с 61% до 63%). Если к рассматриваемым двум операциям добавить еще одну — устранение символа, доля кор–

Таблица 3

Число корней с непустой ближайшей окрестностью

j	$\lvert K_j \rvert$	Замены			Вставки			Замены+вставки		
		$\lvert K_j^1 \rvert$	$\frac{\lvert K_j^1 \rvert}{\lvert K_j \rvert}$	Rec_j'	$\lvert K_j^2 \rvert$	$\frac{\lvert K_j^2 \rvert}{\lvert K_j \rvert}$	Rec_j''	$\lvert K_j^1 \cup K_j^2 \rvert$	$\frac{\lvert K_j^1 \cup K_j^2 \rvert}{\lvert K_j \rvert}$	Rec_j
2	252	252	100%	28	243	96%	30	252	100%	50
3	1322	1308	99%	40	991	75%	20	1312	99,2%	53
4	2451	2280	93%	25	964	39%	12	2309	94%	28
5	2837	2047	72%	17	369	13%	4	2104	74%	19
6	2179	878	40%	6	91	4%	3	931	43%	7
7	1284	274	21%	4	17	1,3%	2	288	22,4%	4
8	799	88	11%	2	7	0,8%	2	94	11,7%	2
9	365	29	7,9%	2	1	0,3%	1	30	8,2%	2
10	108	6	5,5%	1	0	0	0	6	5,5%	1
11	44	2	4,5%	1	0	0	0	6	4,5%	1
$\sum$	11660	7162	61%		2702	23%		7349	63%	

ней с непустой 1-окрестностью возрастет до 68%. Аналогичный показатель для канонических форм равен примерно 43%.

3. С увеличением j доля корней с непустой ближайшей окрестностью монотонно снижается, т.е. устойчивость корней (равно как и словоформ) к рассматриваемым типам искажений повышается. С увеличением длин корней уменьшается в среднем и число элементов, формирующих ближайшую окрестность. То же самое справедливо по отношению к рекордным показателям.

4. Абсолютный рекорд по числу допустимых вставок достигается для $j = 2$, а по числу замен — для $j = 3$. Так, одним из двух рекордистов по числу допустимых замен является корень β = "мал". Длины векторов замен для β: $| z_1 | = 17$, $| z_2 | = 7$, $| z_3 | = 19$, т.е. замены допустимы в каждой из трех позиций, максимальное число замен (19-1) приходится на третью позицию, чуть меньше (17-1) — на первую, т.е. на позиции, содержащие согласные звуки. Число элементов в ближайшей окрестности $| O^1(\beta) | = (17-1)+(7-1)+(19-1) = 40$, если будут произведены замены. Для иллюстрации приведем вектор замен по 3-й позиции: $z_3(\beta)$ = (В, Г, Д, Ж, З, Й, К, Л, М, Н, Р, С, Т, Ф, Х, Ц, Ч, Ш, Щ). Вектор замен однородный, т.е. состоит из одних согласных. Заметим для сравнения, что рекордистом по числу соседей среди канонических форм является слово α = "бок" ($| z_1 | = 9$, $| z_2 | = 5$, $| z_3 | = 9$, $O^1(\alpha) = 8 + 4 + 8 = 20$; z_1 = (Б, Д, К, Н, Р, С, Т, Ф, Ш), z_2 = (А, Е, О, У, Ы), z_3 = (А, Б, Г, Й, К, Н, Р, Т, Ш)). Абсолютными рекордистами по суммарному числу соседей с однократными заменами и вставками являются среди корней длины 2 — "ар" (50 соседей: 20 по заменам, 30 по вставкам), длины 3 — "мар" (53 соседа: 40 по заменам, 13 по вставкам), длины 4 — "карт" (28 соседей: 23 по заменам, 5 по вставкам), длины 5 — "колон" (19 соседей: 16 по заменам, 3 по вставкам) и т.д.

Более высокая в среднем, а также по рекордным показателям, насыщенность ближайших окрестностей у корней по сравнению с каноническими формами наблюдается для значений $j = 2 \div 4$, при $j = 5$ показатели примерно одинаковые, а при длинах $j \geq 6$ доминируют уже канонические формы. В значительной мере

это объясняется комбинаторными соображениями. Из табл. 1 видно, что при $j \leq 4$ $|K_j| > |S_j|$, т.е. поиск ближайших соседей у корней ведется среди большего числа претендентов, чем у канонических форм. При $j = 5$ наступает перелом, а при $j \geq 6$ уже $|S_j| \gg |K_j|$. К тому же слова с длиной 6 и более уже имеют достаточно богатое аффиксальное окружение, в котором чаще всего происходят замены.

Аналогичные соображения в сочетании с общим (для корней и канонических форм) принципом уменьшения числа соседей при увеличении длины структурной единицы применимы и для объяснения существенного различия доли корней и канонических форм с непустой 1–окрестностью (63% против 38%). Из той же табл. 1 видно, что наибольшим многообразием отличаются у корней подмножества K_4 и K_5, а у канонических форм — S_8, S_9 и S_{10}. Но при $j = 4$ и 5 эффект уменьшения числа соседей с увеличением j еще не проявлен столь сильно, как при $j = 8, 9, 10$. Таким образом, основная доля корней находится в диапазоне длин, благоприятных для существования соседей, а основная доля канонических форм — в диапазоне длин, характеризующихся довольно незначительным числом соседей. Этим и объясняются количественные различия в вариативности корней и канонических форм.

4.2. *Распределение замен по позициям корня.* В данном разделе исследуется распределение векторов замен z_k в позициях $k = 1 \div j$ для корней из K_j $(j \geq 3)$. Будем использовать следующие обозначения $n = |z_k|$ — длина вектора замен; $M_n(z_k)$ — число разных z_k длины $n (n \geq 2)$, реализуемых в k-й позиции корней из K_j; $F_n(z_k)$ — полное число (с учетом повторений) векторов длины n в k-й позиции $(F_n(z_k) \geq M_n(z_k))$; $z^*_{k,n}$ — наиболее часто встречающийся вектор замен длины n в k-й позиции, $f(z^*_{k,n})$ — его частота. Будем говорить, что вектор $z^*_{k,n}$ доминирует в позиции k, если его частота существенно превышает частоту следующего в порядке убывания вектора замен (будем обозначать его $z^{**}_{k,n}$), т.е. имеет место $f(z^*_{k,n}) \gg f(z^{**}_{k,n})$.

Проиллюстрируем введенные понятия на примере корней длины 3 из подмножества K_3:

$k=1$:	$M_2(z_1)=32;$	$F_2(z_1)=44;$	$z^*_{1,2}$ = (О, Я);	$f(z^*_{1,2})=3;$
	$M_3(z_1)=21;$	$F_3(z_1)=21;$	$\longrightarrow f(z^*_{1,3})=f(z^{**}_{1,3})=\ldots=1;$	
	$M_4(z_1)=16;$	$F_4(z_1)=17;$	$z^*_{1,4}$ = (В, Л, М, Р);	$f(z^*_{1,4})=2;$
	⋮	⋮	⋮	⋮
	$M_{18}(z_1)=1;$	$F_{18}(z_1)=1;$	$z^*_{1,18}$ = (Б,В,Г, Д,Ж,З,К,Л, М, Н, П, С, Т, Ф, Х, Ц, Ч, Ш).	
$k=2$:	$M_2(z_2)=36;$	$F_2(z_2)=104;$	$z^*_{2,2}$ = (А, И);	$f(z^*_{2,2})=11;$
			$z^{**}_{2,2}$ = (А, О);	$f(z^{**}_{2,2})=8;$
	$M_3(z_2)=32;$	$F_3(z_2)=75;$	$z^*_{2,3}$ = (А, О, У);	$f(z^*_{2,3})=12;$
			$z^{**}_{2,3}$ = (А, И, У);	$f(z^{**}_{2,3})=8;$
	$M_4(z_2)=24;$	$F_4(z_2)=61;$	$z^*_{2,4}$ = (А, И, О, У);	$f(z^*_{2,4})=12;$
			$z^{**}_{2,4}$ = (А, Е, И, О);	$f(z^{**}_{2,4})=12;$
	$M_5(z_2)=18;$	$F_5(z_2)=44;$	$z^*_{2,5}$ = (А,Е,И, О,У);	$f(z^*_{2,5}))=17;$
			$z^{**}_{2,5}$ = (А,Е,И, О,Ы);	$f(z^{**}_{2,5})=4;$
	⋮	⋮	⋮	⋮
	$M_9(z_2)=3;$	$F_9(z_2)=3;$	$z^*_{2,9}$ = (А,Е,И, О,У,Ь,Ю,Я);	$f(z^*_{2,9})=1;$
$k=3$:	$M_2(z_3)=33;$	$F_2(z_3)=43;$	$z^*_{3,2}$ = (Л, Р);	$f(z^*_{3,2})=4;$
			$z^{**}_{3,2}$ = (Н, Р);	$f(z^{**}_{3,2})=3;$
	$M_3(z_3)=20;$	$F_3(z_3)=21;$	$z^*_{3,3}$ = (А, О, Ы);	$f(z^*_{3,3})=2;$
	⋮	⋮	⋮	⋮
	$M_{19}(z_3)=1;$	$F_{19}(z_3)=1;$	$z^*_{3,19}$ = (В,Г,Д, Ж,З,Й,К,Л, М,Н,Р,С,Т,Ф, Х,Ц,Ч,Ш,Щ);	$f(z^*_{3,19})=1.$

Приведенные данные можно дополнить следующими интегральными показателями:

$$M(z_1)=\sum_n M_n(z_1)=171; \quad F(z_1)=\sum_n F_n(z_1)=184;$$
$$M(z_2)=\sum_n M_n(z_2)=132; \quad F(z_2)=\sum_n F_n(z_2)=324;$$
$$M(z_3)=\sum_n M_n(z_3)=170; \quad F(z_3)=\sum_n F_n(z_3)=182;$$

Из последних данных следует, что максимальное число допустимых замен в корнях длины 3 приходится на вторую позицию

($F(z_2) = 324$). Далее, векторы замен в 1–й и в 3–й позиции состоят преимущественно из согласных, а во 2–й — из гласных. Это говорит о том, что большая часть корней длины 3 имеет структуру СГС (С — согласный, Г — гласный). Очень показательны рекордные по длине векторы замен $z^*_{1,18}$, $z^*_{2,6}$ и $z^*_{3,19}$. Фактически они задают разбиение алфавита на согласные ($z^*_{1,18} \bigcup z^*_{3,19}$) и гласные ($z^*_{2,6}$), что несет в себе определенный дешифровочный потенциал. И, наконец, единственный случай доминирования имеет место для вектора $z^*_{2,5}$ = (А, Е, И, О, У): $f(z^*_{2,5}) = 17 \gg f(z^{**}_{2,5}) = 4$. Это достаточно необычный случай доминирования, поскольку с увеличением длины n вектора замен $z^*_{2,n}$ его частота не понижалась, а повышалась. Связки корней, допускающих подстановки такого типа, могут быть представлены в виде "б$z^*_{2,5}$с", "ж$z^*_{2,5}$р", "с$z^*_{2,5}$м" и т.п.

Проведя анализ рассмотренных выше показателей для подмножеств корней K_j со значениями $j > 3$, можно суммировать полученные результаты следующим образом.

1. Для корней с длиной 4 зависимость числа допустимых замен от номера позиции имеет характер выпуклой кривой, как и в случае $j = 3$, т.е. большая часть замен приходится на внутренние ($k = 2$ и 3) позиции корня. При $j = 5$ кривая имеет колебательный ("переходный") характер. Для корней с длинами $j \geq 6$ зависимость числа допустимых замен от номера позиции приобретает характер вогнутой кривой, т.е. наибольшую изменчивость демонстрируют крайние ($k = 1$ и $k = j$) позиции корня, причем с увеличением j доминирующей становится последняя позиция. В этом отношении корни радикально отличаются от канонических форм: у последних с увеличением их длины доминирующей по числу замен становилась первая позиция ("осадить — усадить — всадить — сосадить" и т.п.).

2. Векторы замен при разных j и k — в подавляющем большинстве однородные, т.е. состоят из одних гласных, либо согласных. Неоднородность иногда наблюдается у коротких векторов замен. В первой и последней позициях корня векторы замен чаще всего имеют тип "С", для внутренних позиций характерно чередование типов. Так, в корнях длины 5 доминирующие типы векторов замен распределены по позициям $k = 1 \div 5$ следующим

образом: С-Г-С-Г-С. Поскольку $\max_j |K_j| = 5$, это говорит о том, что в русском языке много корней с симметричной и ритмичной СГ-структурой.

3. Подтвердилось предположение о том, что в корнях, как и в канонических формах, наблюдается привязка определенных векторов замен к определенным позициям. Она проявляется в виде доминирования частот соответствующих векторов. При анализе канонических форм фиксированной длины факт доминирования некоторых векторов в корневых позициях был завуалирован различиями в позиционной привязке корней внутри слова.

В табл. 4 приведены наиболее характерные примеры доминирования отдельных векторов замен z_k^* в различных позициях корней. Для контраста указаны ближайшие к ним по частоте векторы z_k^{**}. Интересно отметить, что в первых позициях корней разной длины не наблюдается явного доминирования каких-либо векторов замен, тогда как в последних, как правило, наблюдается: для значений $j \leq 6$ доминирует вектор (К, Ч), а при $j \geq 7$ — (Т, Ц). Векторы замен типа (Г, Г) фигурируют в качестве доминирующих лишь во внутренних позициях корней. Лидерство здесь принадлежит вектору (А, О), хотя, как видно из табл. 2, наиболее частыми гласными в корнях являются "А" и "Е".

Доминирование некоторых векторов замен в определенных позициях объясняется в значительной мере *эффектом чередования* гласных или согласных в корнях. В первую очередь это касается позиции $k = j$ с векторами замен z_j = (К, Ч) в относительно коротких корнях (вле(z_j), дья(z_j), азбу(z_j), рыно(z_j), башма(z_j), копья(z_j)) и z_j = (Т, Ц) — в длинных (продук(z_j), инспек(z_j), абстрак(z_j), интервен(z_j), ...). Иногда чередованием объясняется и доминирование более длинных ($n > 2$) векторов замен, например, z_8 = (К, Ц, Ч): бурла(z_8), криги(z_8) и т.п.

Эффект чередования вносит определенный вклад и в доминирование вектора замен z = (А, О) во внутренних позициях корней (б(z)лт, д(z)лб и т.п.), но этот вклад не всегда является решающим (к(z)рп, б(z)рщ, бар(z)н и др.). Доминирование вектора замен z = (Л, Р) в позиции 2 при $j = 4$ также, по видимому, не связано с эффектом чередования (у(z)ан, г(z)уб, к(z)ад, ...). Дан-

ный вектор доминирует и в корнях длины 5 в позиции 3, причем часто в окружении одинаковых гласных (го(z)од, ба(z)ак, се(z)ед и т.п.). Это усиливает характерную для корней длины 5 симметричную (в алфавите (С, Г)) структуру: СГСГС.

Таблица 4

Векторы замен z_k^*, доминирующие в позициях $k = 1 \div j$ корней длины j

j	k	z_k^*	$f(z_k^*)$	z_k^{**}	$f(z_k^{**})$	Примеры корней
3	2	(А, Е, И, О, У)	17	(А, Е, И, О, Ы)	4	тz_2^*ш
4	2	(Л, Р)	58	(А, О)	34	гz_2^*ад
	3	(А, О)	27	(А, И)	13	брz_3^*в
	4	(К, Ч)	18	(Л, Р)	10	влекz_4^*
5	2	(А, О)	34	(А, И)	18	пz_2^*пот
	3	(Л, Р)	29	(Л, Т)	8	палz_3^*ом
	4	(А, О)	45	(Е, О)	20	волнz_4^*к
	5	(К, Ч)	26	(Г, Ж)	20	папирz_5^*
6	5	(А, О)	13	(А, И)	8	станцz_5^*е
	6	(К, Ч)	33	(Т, Ц)	20	рюкзакz_6^*
		(К, Ц, Ч)	7	(К, Т, Ч)	1	критикz_6^*
7	7	(Т, Ц)	18	(Г, Ж)	8	инспекz_7^*
				(К, Ч)	8	
8	8	(Т, Ц)	11	(К, Ч)	3	композицz_8^*
9	9	(Т, Ц)	8	(К, Ч),(Г, Ж)	1	компетенцz_9^*

4. Если при каждом фиксированном j просуммировать векторы замен по разным позициям, а потом упорядочить их по убыванию частоты встречаемости, то новых лидеров, отличных от представленных в табл. 4, не появляется. Так при $j = 4$ лидируют векторы (Л, Р) и (А, О) (частоты встречаемости — 78 и 67, соответственно). При $j = 5$ они вновь впереди, но меняются местами ($f(A,O) = 87$, $f(Л,Р) = 53$). Лидерство этих двух векторов объясняется тем, что они доминируют (или близки к этому) сразу по нескольким позициям (см. $j = 4$ и 5 в табл. 4). При $j = 6$ лидером является вектор замен (К, Ч) с суммарной частотой 33. Но из табл.4 видно, что все эти замены реализовались в позиции $k = j = 6$, т.е. данный вектор замен четко идентифицирует

фицирует последнюю позицию корня. Аналогичным свойством обладает вектор замен (Т, Ц) для корней длины 7, 8 и 9.

5. Если просуммировать векторы замен по всем j и k, а потом упорядочить их по убыванию частоты встречаемости, картина будет такая:

$\mid z \mid$	z^*	$f(z^*)$	z^{**}	$f(z^{**})$	z^{***}	$f(z^{***})$
2	(А, О)	196	(Л, Р)	154	(К, Ч)	95
3	(А, Е, О)	24	(А, О, У)	22	(А, Е, И), (А, Е, У)	17
4	(А, И, О, У)	18	(А, Е, О, У)	16	(А, Е, И, О)	16
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	(А, Е, И, О, У, Ы, Ю)	4	(А, Е, И, О, У, Ы, Я)	3	(А, Е, И, О, У, Ы, Э)	2

При $|z| \geq 8$ все векторы имеют $f = 1$ и почти все тип "С".

Иными словами, вплоть до значения $j = 7$ лидирующие векторы замен имеют тип "Г", однако основное разнообразие обеспечивают векторы типа "С". Абсолютными лидерами являются векторы (А, О), (Л, Р) и (К, Ч) (см. первую строку).

4.3. *Распределение вставок по позициям корня.* В данном разделе используется та же система обозначений, что и в предыдущем, только вектор замен z_k, $k = 1 \div j$, $| z_k | \geq 2$, заменен вектором вставок b_k, $k = 1 \div j+1$, $| b_k | \geq 1$. Отметим, что в целом результаты по заменам и вставкам коррелированы вследствие общей закономерности характерной как для канонических форм, так и для корней. Применительно к корням ее можно сформулировать следующим образом. Если для какого-либо корня длины j имеет место $| b_k | \geq 2$, то образующиеся в результате вставки корни длины $j+1$ отличаются друг от друга заменой по k й позиции и вектор замен совпадает с вектором b_k. Поскольку свыше половины ($\sim 51,5\%$) корней, допускающих вставку, имеют по два и более соседей, условие $| b_k | \geq 2$ выполняется довольно часто.

Суммируем основные результаты по распределению допустимых вставок.

1. Наиболее длинные векторы вставок имеют корни длины 2. Рекордные значения по позициям: $| b_1 | = 18$ (тип С), $| b_2 | = 8$

(тип Г), $|\ b_3\ | = 19$ (тип С). При $j > 2$ наиболее длинные векторы вставок наблюдаются для $j + 1$-й позиции. С увеличением j рекордные показатели монотонно убывают: $j = 3 \longrightarrow |\ b_4\ | = 14$; $j = 4 \longrightarrow |\ b_5\ | = 12$; $j = 5 \longrightarrow |\ b_6\ | = 4$; $j = 6 \longrightarrow |\ b_7\ | = 2$.

2. Наибольшим суммарным (по всем позициям) числом вставок характеризуются корни длины 3. Корни с длиной $j \geq 9$ уже практически не допускают вставок. Наибольшее число вставок при разных j наблюдается в позициях $k = j$. При этом, начиная со значения $j = 4$, в качестве вставок в этой позиции превалируют гласные (Е, О и т.д.). Дело в том, что корни достаточно большой длины ($j \geq 4$), как правило, оканчиваются на согласный, поэтому именно такой тип вставок характерен для предпоследней позиции. Преобладание гласных (О, Е) во вставках, в основном, отражает наличие в словаре вариантов корней с беглыми гласными. Например, "башн" и "башен", "букв" и "буков", "министр" и "министер". Иногда вставка (О, Е) может менять семантику корня: "балт" и "балет", "борд" и "бород".

3. Имеет место доминирование определенных векторов вставок в некоторых позициях (см. табл. 5), но оно проявлено слабее, чем в случае замен.

Сопоставление таблиц 4 и 5 проясняет характер взаимосвязи между вставками и заменами. Так, при $j = 3$ во второй позиции доминирует вектор замен (А, Е, И, О, У) (см. табл. 4). Он же, но уже в качестве вектора вставок, доминирует во второй позиции корней длины 2, переводя корни типа СС в корни типа СГС (см. табл. 5). Аналогично, при $j = 3$ во второй позиции доминирует вектор вставок (Л, Р) (см. табл. 5). Он переводит корни длины 3 в корни длины 4, отличающиеся друг от друга лишь заменами "Л" на "Р" (или наоборот) во второй позиции. Из табл. 4 мы видим, что как раз этот тип замен доминирует при $j = 4$, $k = 2$.

В варианте с заменами не было обнаружено доминирования каких-либо векторов в первой позиции корней. В варианте со вставками такое доминирование имеет место: корни длины $j + 1$ часто образуются из корней длины j добавлением буквы "С" слева (см. $k = 1$ для $j = 3$ и 4 в табл. 5).

4. Суммирование и упорядочение по частоте всех векторов вставок по разным позициям при фиксированном j обнаруживает

Таблица 5

Примеры доминирования конкретных векторов вставок в различных позициях корней

j	k	b_k^*	$f(b_k^*)$	b_k^{**}	$f(b_k^{**})$	Примеры вставок
2	2	(А, Е, И, О, У)	5	(А, Е, И, О, Ы)	3	жр⟶жb_2^*р
3	1	(С)	52	(А)	37	вящ⟶b_2^*вящ
	2	(Р)	118	(Л)	47	топ⟶тb_2^*оп
		(Л, Р)	37	(Л, Т)	4	кич⟶кb_2^*ич
	3	(Р)	76	(Н)	24	кож⟶коb_2^*ж
4	1	(С)	41	(К)	23	наст⟶b_2^*наст
	2	(О)	59	(Е)	32	нрав⟶нb_2^*рав
	3	(Р)	28	(Л)	10	шест⟶шеb_2^*ст
	4	(Е)	84	(О)	46	весн⟶весb_2^*н
5	5	(Е)	25	(О)	13	глянц⟶глянb_2^*ц
	6	(Т)	12	(Н)	8	талан⟶таланb_2^*
6	6	(Е)	7	(О)	6	полотн⟶полотb_2^*н

два ярких (но не новых) доминирования. При $j = 3$ доминирующей является вставка "Р" (f("P") = 241), которая лидирует во второй, третьей и четвертой позициях, а также вектор (Л, Р) с частотой 43, характерный лишь для второй позиции. Иными словами, корни длины 4 часто образуются из корней длины 3, содержащих, как правило, одну гласную, добавлением (вставкой) буквы "Р" в любую позицию кроме первой.

При $j = 4$ доминируют вставки в виде гласных "Е" (f("E") = = 131) или "О" (f("O") = 131) в позициях 2 и 4. Они переводят четырехбуквенные корни с одним гласным (структуры типа СГСС, ССГС) в пятибуквенные со структурой СГСГС.

5. Общая статистика векторов вставок по разным j и k дает следующие результаты. Среди векторов длины 1 лидируют "Р" (f("P") = 377), "Е" (f("E") = 233), "С" (f("C") = 222), "О" (f("O") = 214); среди векторов длины 2 — (Л, Р) с частотой 49 и (А, О) с частотой 27. Вставки "Р", "Е", "Л" являются корнеспецифичными, они имеют значение $\Delta < 0$ (см. табл. 2). Вставки "С", "О", "А" характерны и для аффиксов, они имеют значение $\Delta > 0$.

З а к л ю ч е н и е

Важной в теоретическом и прикладном отношении межъязыковой проблемой является исследование вариативности структурных единиц языка на разных иерархических уровнях. В рамках этой общей проблемы анализируется способность корней слов русского языка переходить в другие корни (в пределах фиксированного словаря) путем незначительных искажений, носящих комбинаторный характер (замена, вставка, устранение символа и т.п.). Количественной мерой вариативности корня служит число его ближайших (в метрике редакционного расстояния) корней-соседей. Подобного рода постановки лежат в русле направления, получившего условное название "лингвистическая комбинаторика" [6].

На материале словаря Д.Уорта объемом свыше 100 тыс. слов показано, что вариативность корней существенно выше вариативности содержащих их канонических форм — единиц более высокого иерархического уровня. Порядка 63% корней характеризуются непустой ближайшей окрестностью, т.е. имеют хотя бы

одного соседа, отличающегося от исходного корня не более чем одной заменой или вставкой. Аналогичный показатель для канонических форм 38%. Рекордные показатели по числу ближайших соседей значительно выше у корней, чем у канонических форм. Различен и характер распределения допустимых искажений по позициям внутри структурной единицы. У корней наиболее вариабельной является последняя позиция, у канонических форм — первая.

Выделены доминирующие типы замен и вставок для различных позиций корня. Многие из них (но не все) объясняются эффектом чередования гласных (согласных) звуков в корнях. Лингвистам хорошо известны эти явления, но здесь они получили количественную оценку. Показана закономерная связь между заменами и вставками в определенных позициях.

Результаты работы представляют интерес для изучения взаимосвязи элементов различных иерархических уровней в языковых системах, выработки критериев автоматического выделения структурных единиц из слитных текстов, а также для исследования моделей корнеобразования.

Литература

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

2. Левенштейн В.И. Двоичные коды с использованием выпадений, вставок и замещений символов// ДАН СССР. - 1965. — Т.163, № 4. — С. 845 — 848.

3. Sellers P.H. On the theory and computation of evolutionary distances // SIAM J. Appl. Math. — 1974. — Vol.26, № 4.— P. 787 — 793.

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

5. Worth D., Kozak A., Jonson D. Russian Derivation Dictionary. — New-York, 1970. — 747 p.

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

Поступила в редакцию
28 марта 2000