К компьютерному моделированию варьирования русского корня

А.А. Кретов email: kretov@rgph.vsu.ru, О.Г. Артемова,
А.С. Шудрикова

Воронежский государственный университет

Аннотация. *В работе рассматриваются различные аспекты проблемы компьютерного анализа и синтеза вариативности русского корня.*

Ключевые слова: *Обработка текстов на естественном языке, русские корни, чередования фонов, полифоны, претезы, посттезы, фузиаты.*

Введение

Одной из существенных составляющих искусственного интеллекта (ИИ) является ОТЕЯз – Обработка Текстов на Естественном Языке (Natural Language Processing, NLP) — совместное направление компьютерных наук и математической лингвистики. Применительно к ИИ *анализ* текстов на естественном языке означает их понимание, т.е. переход от формы к содержанию, а *синтез* текстов на естественном языке — порождение их, т.е. переход от содержания к форме. Решение этих проблем необходимо для создания оптимальной формы взаимодействия компьютера и человека.

Одна из главных проблем ИИ и компьютерной лингвистики в целом связана с тем, что компьютер работает только с формой, а человека интересует, главным образом, содержание текстов. Поскольку естественному языку свойственна формально-содержательная асимметричность, проблема ОТЕЯза, а тем самым – и ИИ состоит в её преодолении.

Существуют разные подходы к формализации языка. Едины они лишь в одном: язык состоит из словаря и грамматики. Различие в подходах заключается в предпочтении, отдаваемому *словарю* как декларативной части алгоритма или *грамматике* как его операционной части.

Предпочтение *декларативной* части можно назвать «американским» или «голливудским» подходом, основывающемся на принципе «сила есть – ума не надо», предпочтение *операционной* части

можно назвать «русским» или «суворовским» подходом, исповедующим принцип «воюй не числом, а уменьем».

Оба подхода особенно ярко проявляются при решении вопроса о моделировании словаря. При *голливудском* подходе словарь задается списком, при *суворовском* списком задаются только корни и прочие значимые части слов (морфемы). Кстати говоря, впервые такое решение было применено в IV в. до н.э. индийским лингвистом Па́нини в его книге «Дхату-патха» (санскр. □□□□□□□□ dhātupāṭ ha – образовано из двух слов *дхату*: □□□□ «элемент, корень слова» и □□□ *патха*: □□□ означает «чтение» или «урок»). При этом оказывается, что более 300 000 слов санскрита, содержащиеся в самом полном санскритско-немецком словаре, образованы от 3 689 корней, выделенных методами Панини и обобщенных в 580 корней европейскими учеными [1]. По своему богатству русский язык не уступает санскриту, поэтому описывать лексику русского языка целесообразнее в виде исчисления.

Но и корни, сочетаясь с окончаниями, суффиксами и приставками, не остаются неизменными. И описывать варьирование их формы также целесообразно не в виде списка, а в виде исчисления.

Для этого необходимо иметь набор исходных элементов (фонем) и правил их преобразования в наблюдаемый в речи морф в зависимости от позиции, в которой находится корень (как морфема) с составляющими его форму фонемами.

Таким образом, нам необходимо установить множество фонем, из которых складываются корни в русском языке, и множество правил преобразования фонем в реальные звуки речи.

В готовом виде такая информация нигде не представлена, поэтому наша задача состоит в том, чтобы извлечь её из имеющихся описаний корней русского языка.

1. Методология исследования

Наиболее подходящим источником такой информации является «Словарь морфем русского языка» А.И. Кузнецовой и Т.Ф. Ефремовой. (М.: Русский язык, 1986 далее [СМоРЯ-1986]), в котором в явном виде даны корневые морфы, сведенные в морфемы [2]. И хотя в этом словаре отсутствует информация о фонемном составе корневых морфем (членение слов осуществляется в орфографической записи), имеющейся в нем информации достаточно для того, чтобы эти фонемы выявить. При установлении фонемного состава корневых морфем мы исходим из принципов, сформулированных в книге [3].

Так, например, один из корней в [СМоРЯ-1986] имеет 8 морфов: *лаг* (*возЛАГать*), *лег* (*ЛЕГла*), *лёг, лог* (*заЛОГ, наЛОГ, подЛОГ*), *леж*

(*ЛЕЖит*), *лёж* (*поЛЁЖивал*), *ложь* (*поЛОЖит*) и *леч* (*ЛЕЧь*). При этом сам корень представляют только первые 7 морфов. Из них три последних отличаются только звуком /ж/, который в позиции перед гласными переднего ряда (/и/, /е/) представляет фонему {Г}. Иными словами, звук /ж/ – это фонема {Г} перед {И-Е} на суффиксальном стыке, т.е. /ж/ в функции {Г}, сокращённо – /ж/^{г}. Правило этой трансформации корня можно записать следующим образом: если {Г} на суффиксальном стыке находится перед {И/Е}, то {Г} => /ж/.

Правило преобразования /лег/ > /лёг/ можно записать следующим образом: если {Е} находится под ударением перед твёрдым согласным, то {Е} => /’ó/ (т.е. реализуется в виде /о/-смягчающего).

Правило преобразования /лог/ > /лаг/, описывается так: перед ударным суффиксом =á, гласный корня «удлиняется» в /а/: если {О} перед =á, /о/ => /а/.

Что касается чередования *лег* ~ *лог*, то оно восходит к праиндоевропейской древности и описывается правилом: если слово с корнем *л(е/о)г-* находится в предложении позиции сказуемого (и, соответственно, выполняет его функцию), то корень реализуется морфом *лег-*, если нет, то – морфом *лог-*. Поскольку и в форме *подлёг*, и в форме *подлог* меняется не значение корня, а его грамматическое значение: глагол / имя, то чередование /е~о/ является не чередованием фонем, которые различают морфемы, а чередованием фонов, представляющих одну **гиперфонему**: {œ}.

Что касается сегмента *леч-*, то он представляет как минимум две материально выражаемые морфемы: корень *лег-* и суффикс инфинитива *-ть* (ср. *бы=ть, да=ть*): в соответствии с законом восходящей звучности сочетание /гт’/ в слоге **леГ=Ть* (4522) реализуется в виде *леч-* (462). При этом звук /ч/ представляет сразу две фонемы: {Г} корня и {Т} суффикса инфинитива; сокращенно: /ч/^{гт}. Такие бифункциональные звуки мы называем *бифонами* (т.е. бифункциональными фонами) и рассматриваем отдельно. Как целое, (например, *стол*) нельзя сопоставлять с его частями (*ножкой* или *столешницей),* так и *бифон* – целостное представление двух фонем нельзя сопоставлять с *фонами* – отдельными и независимыми представлениями его частей.

Процесс слияния фонов, представляющих фонемы соседних морфем, в один звук называют *фузией* (**ФУЗИЯ** (лат. *fusio* «слияние») «Тесное морфологическое соединение изменяемого корня с аффиксами, приводящее к стиранию границ между морфемами» [Ахманова 1966:515]), а результат *фузии* – неразрывно сплавившиеся морфы соседних морфем можно было бы назвать *фузиатом* – по аналогии с

предикация – предикат, фабрикация – фабрикат, концентрация – концентрат, конденсация – конденсат, дистилляция – дистиллят.

Таким образом, все семь морфов корня: *лег-, леж-, лёг-, лёж-, лог-, ложˉ лаг-*, и **фузиат** *лечˉ* закономерным и строго формализуемым образом выводятся из единого глубинного, т.е. фонемного представления {ЛŒГ}.

Строго формализуемый порядок варьирования корня позволил также обнаружить ещё не описанное в русистике явление. Если наряду с понятием (конкретный) *корень* ввести понятие *Корень* (вообще) – как системное обобщение формальных свойств всех русских корней, то обнаруживается, что ***Корень* в русском слове имеет две факультативные позиции: ничего не меняющий звук перед ним** (например, *ОСЬМушка ~ **в**ОСЬМушка; КОРа ~ **с**КОРняк*) **и ничего не меняющий звук после него**: например, *страд*()*-ать, стра(д>**ø**)-**х*** и *стра(д>**с**)-**т**ь*. По аналогии с терминами *протеза-эпентеза*, эти факультативные звуки, не нарушающие тождества и не меняющие семантики корня, можно назвать **претезой** и **посттезой**. По-видимому, это свойство русского корня имеет архаичный характер: так, например, русскому ***с**трах* в лужицком соответствует *trach*, а русскому ***с**традать* – луж. *tradać)*. Эти факты позволяют нам (по крайней мере, для праславянского языка) считать /с/ этого корня *претезой*. Под определение *претезы* подходит и S-mobile праиндоевропейского языка.

Учет *претез* и *посттез* при изучении варьировании корня важен для того, чтобы не приписывать им статуса фонов – представителей фонем в речи. Те и другие являются несистемными спорадическими явлениями, как правило, древнего происхождения, и должны задаваться списком.

2. Исследование

Для формального определения состава и звукового варьирования русских корневых морфем, представленных в СМоРЯ-1986, проведено их позиционное упорядочивание, позволившее осуществить автоматизированное выявление чередований, составляющих фактологическую базу исследования количества и качества фонем в современном русском языке. Пример позиционного распределения фонов в корневых морфах представлен в Таблице 1.

Таблица 1

№	Корень в словаре	морфы	П1	П2	П3	П4	П5
1	ко́пот	ко́пот	к	ó	п	оъ	т
2	(ка́пч,	кАпøч	к	áо	п	øъ	чтй
3	копт,	копøт	к	о	п	øъ	т
4	копч)	копøч	к	о	п	øъ	чтй

При определении позиций фонов в морфах соблюдается принцип максимального совпадения фонов. Во всех четырех морфах наблюдается совпадение в позициях 1 и 2. В строках 1 и 3 фоны совпадают и в позиции 5. ***Ч*** в П5 (строки 2, 4) является **бифоном** $Ч^{mй}$, а сегменты *капч-, копч-* – **фузиатами.** В позициях П2 и П4 наблюдаем ряды чередований *ó:á:о,* свидетельствующий о фонеме {О}, и *о:ø,* свидетельствующий о фонеме {Ъ} (О-беглом). Разнодлинные корни указывают на наличие в корне чередования ненулевых фонов с нулем звука. Автоматизация анализа чередований в корне обеспечивается превращением разнодлинных корней в равнодлинные. С этой целью применяется ручная вставка символа *ø,* соответствующего нулевому альтернанту.

Выявление чередований осуществляется с помощью таблиц MS Excel и их встроенных функций. Алгоритм анализа базы данных СМоМЯ-86 включает:

– разбиение корней, упорядоченных по числу алломорфов в морфеме, на соответствующие множества: от 2 до 8;

– анализ на отдельном листе таблиц MS Excel каждого множества отдельно с соблюдением последовательности анализа от меньшего множества к большему множеству (от двухморфных корней к восьмиморфным корням);

– для каждого морфа создаются столбцы, количество которых соответствует максимальному количеству букв. Столбцы для каждого морфа обозначаются своим цветом. В каждой ячейке каждого столбца соответствующего морфа размещается одна буква. Для заполнения ячеек используется функция ПСТР;

– после заполнения всех ячеек с помощью функции ЕСЛИ создаются 8 столбцов сравнений, обозначаемых с1, с2, с3, … с8. При совпадении значений в соответствующих позициях функция ЕСЛИ имеет значение 0, при несовпадении позиций – 1. Для определения суммы значений столбцов в каждой строке создается отдельный столбец СумС;

– после сортировки данных в порядке возрастания по столбцу СумС и в порядке убывания по столбцам с1-с8, в столбцах с1-с8 убираются все 0;

– осуществляется отдельный анализ каждой позиции по следующему алгоритму (пример дан для 4-х морфных корней):

С1 – сравнение позиции Х морфемы 1 с позицией Х морфемы 2;

С2 – сравнение позиции Х морфемы 1 с позицией Х морфемы 3;

С3 – сравнение позиции Х морфемы 1 с позицией Х морфемы 4;

С4 – сравнение позиции Х морфемы 2 с позицией Х морфемы 3;

С5 – сравнение позиции Х морфемы 2 с позицией Х морфемы 4;

С6 – сравнение позиции Х морфемы 3 с позицией Х морфемы 4;

– обеспечить алфавитную последовательность фонов в чередовании.

В Таблице 2 представлены выявленные в русских корнях чередования гласных (V), согласных(С) и сонантов (S).

Таблица 2

Черед	Раз	Тип	Черед	Раз	Тип	Черед	Раз	Тип
а:о	136	v	а:о:ы	4	v	е:ё:ь:ø	2	v
к:ч	129	c	е:и:о:ь	4	v	е:ё:о	2	v
е:ё	129	v	е:ё:ю	4	v	е:и:ø	2	v
г:ж	93	c	и:о:ь	4	v	ё:ø	2	v
о:ø	78	v	и:ø	4	v	о:е	2	v
е:ø	50	v	у:ø	4	v	б:с	1	c
и:ы	20	v	у:ю	4	v	к:т	1	c
д:ø	19	c	с:т:ø	3	c	к:ц:ч:ш	1	c
в:ø	15	s	в:ы	3	s	к:ч:щ	1	c
ц:ч	12	c	м:ø	3	s	к:ш	1	c
т:ø	11	c	ё:ь	3	v	с:т:ч	1	c
е:ё:ø	10	v	б:ø	2	c	т:ч:ø	1	c
е:о	9	v	г:ø	2	c	р:ø	1	s
е:ь	9	v	г:ж:ø	2	c	а:е:ё:и:о:ø	1	v
о:ы:ø	9	v	г:к	2	c	а:е:ё:и:о	1	v
к:ø	8	c	д:с:ø	2	c	а:е:ё:и:я	1	v
и:о:ø	7	v	д:т:ø	2	c	а:е:и:о:ø	1	v

н:ø	6	s	з:ø	2	c	а:е:о:ø	1	v
а:о:ø	6	v	к:т:ц:ч	2	c	а:е	1	v
е:и:о:ø	6	v	к:ц	2	c	е:ё:и:ø	1	v
к:ч:ø	5	c	к:ц:ч	2	c	е:ё:и:о	1	v
а:е:о	5	v	п:ø	2	c	е:ё:о:ø	1	v
е:о:ø	5	v	с:ø	2	c	е:и:о	1	v
о:ы	5	v	с:т:ч:ø	2	c	е:ы:ø	1	v
г:ж:з:ø	4	c	а:е:и:ø	2	v	е:и	1	v
г:ж:ч	4	c	а:и:о:ø	2	v	и:й:ы	1	v
д:с	4	c	а:и:о:ь	2	v	и:й	1	v
к:т:ø	4	c	а:и:ø	2	v	о:у:ы	1	v
с:т	4	c	а:и:о	2	v	о:у	1	v
а:е:ё:о	4	v	е:ё:и:о:ø	2	v	у:ы	1	v

Анализ данных из Таблицы 2 свидетельствует о том, что в русских корнях преобладает чередование гласных. Таких чередований выявлено 558. Почти вдвое меньше случаев чередования у согласных фонем – 333. Чередование сонантов отмечено в 28-ми случаях. Из 90 представленных комбинаций фонов наиболее частотными являются следующие: *а:о* (134), *к:ч* и *е:ё* (129). Замыкают пятерку самых частотных чередований *г:ж* (93), *о:ø* (78) и *е:ø* (50). Высокочастотными также являются чередования *и:ы* (20), *д:ø* (19), *в:ø* (15), *ц:ч* (12), *т:ø* (11), *е:ё:ø* (10), *к:ø* (8), *и:о:ø* (7). Большинство представленных в Таблице 2 чередований – 25 встретилось по одному разу. Среди них преобладают чередования гласных фонов, составляющие 17 комбинаций. Это – *а:е:ё:и:о:ø, а:е:ё:и:о, а:е:ё:и:я, а:е:и:о:ø, а:е:о:ø, а:е, е:ё:и:ø, е:ё:и:о, е:ё:о:ø, е:и:о, е:ы:ø, е:и, и:й:ы. и:й, о:у:ы, о:у, у:ы*. Один раз встречается чередование сонанта с нулем звука – *р:ø*. Остальные 7 сочетаний представляют чередование согласных фонов: *б:с, к:т, к:ц:ч:ш, к:ч:щ, к:ш, с:т:ч, т:ч:ø*. Следующие 24 чередования встречаются по два раза. Среди них преобладают чередования согласных, включая 5 чередований согласного с нулем звука – 13: *б:ø, г:ø, г:ж:ø, г:к, д:с:ø, д:т:ø, з:ø, к:т:ц:ч, к:ц, к:ц:ч, п:ø, с:ø, с:т:ч:ø*. Еще 11 чередований представляют чередования гласных, включая одно чередование гласной с нулем звука: *а:е:и:ø, а:и:о:ø, а:и:о:ь, а:и:ø, а:и:о, е:ё:и:о:ø, е:ё:ь:ø, е:ё:о, е:и:ø, ё:ø, о:е*. Следующую по численности группу составляют чередования, встретившиеся 4 раза.

Среди них также преобладают чередования гласных, включая два чередования с нулем звука – 8: а:е:ё:о, а:о:ы, е:и:о:ь, е:ё:ю, и:о:ь, и:ø, у:ø, у:ю. Чередование согласных представляют следующие 5 сочетаний: *г:ж:з:ø, г:ж:ч, д:с, к:т:ø, с:т*. Две равноразмерные группы из четырех различных чередований встречаются по пять (*к:ч:ø, а:е:о, е:о:ø, о:ы*) и по три (*с:т:ø, в:ы, м:ø, ё:ь*) раза. Еще две равноразмерные группы представляют чередования, встретившиеся по 9 (*е:о, е:ь, о:ы:ø*) и по 6 (*н:ø, а:о:ø, е:и:о:ø*) раз. Первую группу составляют чередования гласных, вторую – два чередования гласных и одно чередование сонанта с нулем звука.

Что касается количества альтернантов, то здесь, безусловно, лидируют двучленные чередования – 43 сочетания, из которых альтернанты гласных фонов представлены 20-ью различными комбинациями, альтернанты согласных фонов – 18-ью, альтернанты сонантов – 5-ью. Из них чередование с нулем звука представлено 9 раз у согласных фона, у гласных – 5 раз, у сонантов – 4 раза. Следующими по числу альтернантов являются трехчленные чередования – 28: чередования гласных – 17, чередования согласных – 11. 15 четырехчленных чередований включают 11 комбинаций гласных и 4 комбинации согласных фонов. 4 пятичленных чередования и одно шестичленное чередование отмечены исключительно у гласных.

Заключение

Формально варьирование корня может быть подразделено на количественное, качественное и качественно-количественное.

В функциональном аспекте алломорфы можно подразделить на агглютинативные и фузионные.

Состав агглютинативных алломорфов обеспечивают аллофоны, связь фузионных алломорфов с соседними алломорфами обеспечивают полифоны (бифоны, трифоны, квадрофоны).

Массовое обследование чередований в корне обеспечивает выделение корней, представленных качественными алломорфами.

Для корректного определения позиций фонем в корне необходимо соблюдать ряд требований: 1) оптимально совмещать морфы, представляющие один корень; 2) при определении состава фонем исключать из рассмотрения последовательности, содержащие полифоны; 3) устранять метатезы и сдвиги звуков в речевой цепи с восстановлением дометатезной позиции; 4) перед определением позиций в корне в позиции после букв, обозначающих гласные, и в абсолютном начале слова буквы Е, Ё, Ю, Я следует заменить на сочетание Й с Е, О, У или А, соответственно. Таким образом, моделирование фонемного состава морфем в слове обеспечивается

исключительно применением фонематической, т.е. этимологической, транскрипции, что в полной мере соответствует сути фузионных языков, одним из которых и является русский язык.

Список литературы

1. Гасунс М. Ю. Состав и строй древнеиндийских корней: история изучения. – Автореф. … канд. филол. наук. – М., 2014. – 22 с.
2. Кузнецова А.И., Ефремова Т.Ф. Словарь морфем русского языка: Ок. 52 000 слов. – М.: Рус. яз, 1986. – 1136 с.
3. Кретов А.А. Системная русская фонемология : монография / А. А. Кретов; Воронежский государственный университет. – Воронеж: Издательский дом ВГУ, 2020. – 198 с. – (Серия «Системная русская лингвистика»; Т. 1).