В.Л. ВВЕДЕНСКИЙ
Российский научный центр «Курчатовский институт», Москва
vvedensky@imp.kiae.ru

СЕТЬ КОРНЕЙ ГЛАГОЛОВ РУССКОГО ЯЗЫКА

Аннотация

Проведен анализ структуры множества глагол-образующих корней русского языка. Количество диалектно-устойчивых корней невелико и включает около 400 корней. Обычно корень имеет вид «согласный-гласный-согласный» и два крайних согласных являются хорошими систематизирующими факторами. Оказалось, что множество стабильных корней может быть сведено в плотно-упакованную двумерную систему так, что одинаковые согласные-фонемы образуют сеть примерно параллельных взаимно-пересекающихся линий. Такая организация множества корней наводит на мысль, что они хранятся в мозге в соответствии со структурой коры, образующей чередующиеся полоски, состоящие из колонок нейронов. Этот вид организации хранения лингвистической информации указывает, каким образом может быть построена искусственная нейронная сеть, позволяющая распознавать речь в условиях фонового шума в речевом интерфейсе.

ВВЕДЕНИЕ

Эта работа является продолжением исследований языка человека направленных на создание эффективного средства общения с компьютером через речевой интерфейс. В работе [1] было показано, что применение звуков речи (фонем) при образовании слов подчиняется строгим математическим закономерностям, общим для большой группы европейских языков, а, вероятно, и для всех человеческих языков. Это указывает на то, что в основе построения слов в разных языках лежат одни и те же причины, скорее всего вытекающие из устройства коры мозга человека. Закономерности, наблюдаемые при анализе структуры языка, можно использовать как инструмент для выяснения устройства механизма, осуществляющего функцию языка в мозге человека. Это, примерно, как по показаниям счетчиков, детектирующих рентгеновские лучи, рассеиваемые мозгом при сеансе томографии, можно восстановить картинку, изображающую этот самый мозг. Используя эту метафору, можно сказать, что при исследованиях важно найти правильное место, куда направлять лучи. Наш опыт показывает, что из сложного и многообразного явления, каким явля-

ется человеческий язык, на этом начальном этапе пристальное внимание следует уделить именно глаголам.

СТАБИЛЬНЫЕ КОРНИ ГЛАГОЛОВ

Оснований для приоритета рассмотрения глаголов несколько. Глаголы в большей степени, чем остальные части речи, образуют связанную внутри себя систему. При переводе на другой язык основные глаголы имеют большое число вариантов перевода, так что при обратном переводе, например на русский с английского языка, можно получить очень пространный список других глаголов [1]. Это относится практически ко всем базовым глаголам, которые смыкаются в почти непрерывную систему, покрывающую пространство понятий, описывающих набор действий, воспринимаемых человеком. Свойства этой системы доступны для количественного анализа.

Глаголы допускают построение иерархии, например, по количеству разнообразных переводов на другой язык (английский, например). В работе [1] приведено начало списка наиболее «продуктивных» глаголов:

Держать, Стоять, Вертеть, Плести, Выразить, Изрекать, Дышать, Лежать, Иметь, Бить, Тереть, Вязать, Дать, Поместить, Вить, Крыть, Делить.

Конечно, имеются и существительные с этими корнями:

Задержка, Стойка, Поворот, Плетка, Образ, Речь, Дыхание, Ложе, Имение, Битва, Терка, Повязка, Дача, Место, Виток, Крыша, Доля,

но это те же самые корни, а связную иерархию корней по глаголам установить проще.

В русском языке (и в ряде других, например, немецком) корни могут быть охарактеризованы такой величиной, как «сродство» к приставкам. По этому качеству также можно ввести иерархию корней русских слов. Максимально это сродство у глаголов. На рис. 1 показана зависимость числа допустимых приставок от порядкового номера корня в иерархии. При построении этого графика рассматривались диалектно-стабильные корни языка, то есть такие, которые сохраняются даже в других родственных языках. Такой подход вытекает из задачи выделения стабильного ядра языка, которое можно было бы аккуратно анализировать с применением математических средств. Словарь морфем русского языка [2] содержит тысячи корней, многие из которых свойственны лишь русскому языку, и если выделить из них те, которые сохраняются и в польском, и в болгарском языках, число резко сократится. Какие-то механизмы удерживают

именно эти корни в языках и можно полагать, что это отражает способ их хранения в мозге. Выбор болгарского и польского языков в качестве «реперных» базируется на результатах сравнения европейских языков, проведенного в работе [3], говорящих о том, что они равно удалены от русского, но в разных направлениях. Различие направлений проявляется в том, что заметная часть корней у русского языка, общая с болгарским языком, но не с польским, и наоборот. Помимо глагол-образующих корней, сродство которых к приставкам показано на рис. 1, удается найти еще примерно 250 корней, таких как в словах *босой, сивый, боб*, которые не принимают приставок и не образуют глаголов. Есть основания полагать, что эти корни хранятся в коре в другом месте, не имеющем прямых связей с областями, генерирующими приставки. Главный результат проведенного отбора, необходимый для дальнейшего анализа, состоит в том, что можно выделить подмножество устойчивых корней глаголов в количестве 400-500 и упорядочить их по относительной важности.

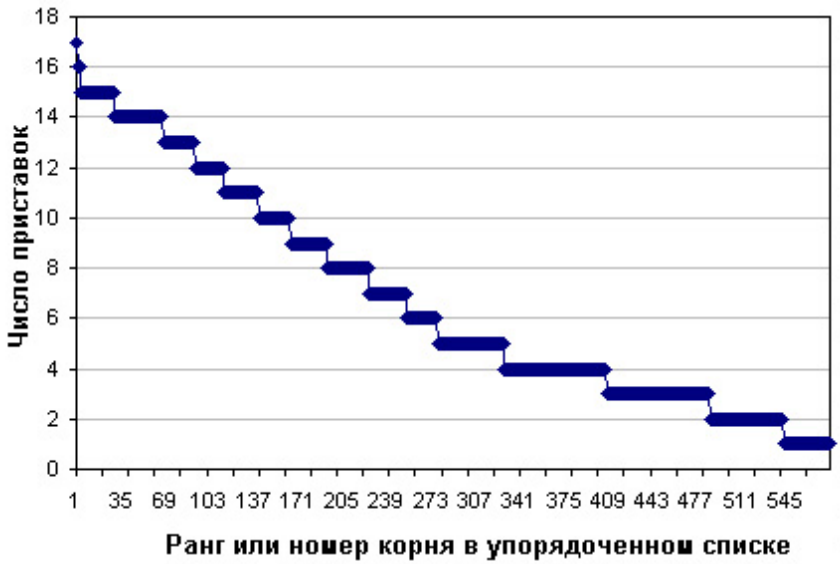

Рис. 1. Количество приставок, совместимых с различными корнями русского языка. Приставки приведены в порядке употребительности: по, за, о, у, про, на, раз, в, пере, с, от, при, до, под, об, из, не, без

ТРИ КОМПОНЕНТА КОРНЕЙ СЛОВ

Подавляющее число корней русского языка (и других европейских языков) имеет однотипную трехкомпонентную структуру. Корень начинается комплексом согласных (или одной), в середине находится комплекс, включающий Р, Л или В в комбинации с гласными, и заканчивается опять комплексом согласных. В ряде славянских языков ценральный комплекс может обойтись и без гласных, как чешские *VRH, TRH,* *VLK* – верх, рынок (торг), волк. Центральный комплекс обладает повышенной лабильностью, легко изменяясь как внутри языка, так и при переходе к родственному языку – *М(оЛо)К{о}* – *М(Ле)Ч{ный}* – *М(Le)K{o} (польский)* – *M(iL)K (англий ский).* Замечено, что при чтении текстов центральная часть слова (корня) аккуратно не воспринимается и даже перестановка букв в ней не сильно влияет на восприятие смысла написанного [4, обратите внимание на название статьи]. С другой стороны, «обрамляющие» согласные в первой и третьей позиции достаточно строго определяют звучание корня, так что гласная часто «напрашивается сама». Иногда она просто выпадает, как в словах *П-Нуть* или *Г-Нать*. Некоторые позиции в корне могут оставаться пустыми, как в словах *ПИ-ть* или *ЛИ-ть*. Этот пропуск следует учесть как особый знак (-) при анализе.

Набор стабильных глагол-образующих корней может быть разложен на такие элементы и можно определить их количество. Результат показан на рис. 2. Из вида графика напрашивается схема образования корней показанная на вставке. Действительно, для наиболее часто встречающегося в первой позиции звука *П* имеются комбинации с наиболее часто применяемыми в конце корня *д, н, (-), т*: *ПАДать, ПИНать, ПИть, ПЫТать*. В общем, комбинирование согласных в корнях следует этой «треугольной» идеализированной схеме, однако есть важное расхождение. Общее число стабильных корней около 400, а эта схема дает лишь около 200. Величины можно уравнять введением дополнительных «столбцов и строк», что даст большее число пересечений. Действительно, в русском языке есть такие слова. как *КОПать, КАПать, КУПить и КИПеть*, которые могут располагаться на пересечении двух столбцов *К* и двух строк *П*. И после этой процедуры остаются еще ряд комбинаций отсутствующих и лишних внутри треугольника, а также имеются комбинации из «редких» согласных, располагающиеся вне треугольника. Эти «дефекты» можно устранить или упорядочить, перемещая их на границы треугольника путем перестановок строк и столбцов. При этом треуголь-

ник деформируется, а столбцы и строки могут искривляться, сохраняя свою непрерывность.

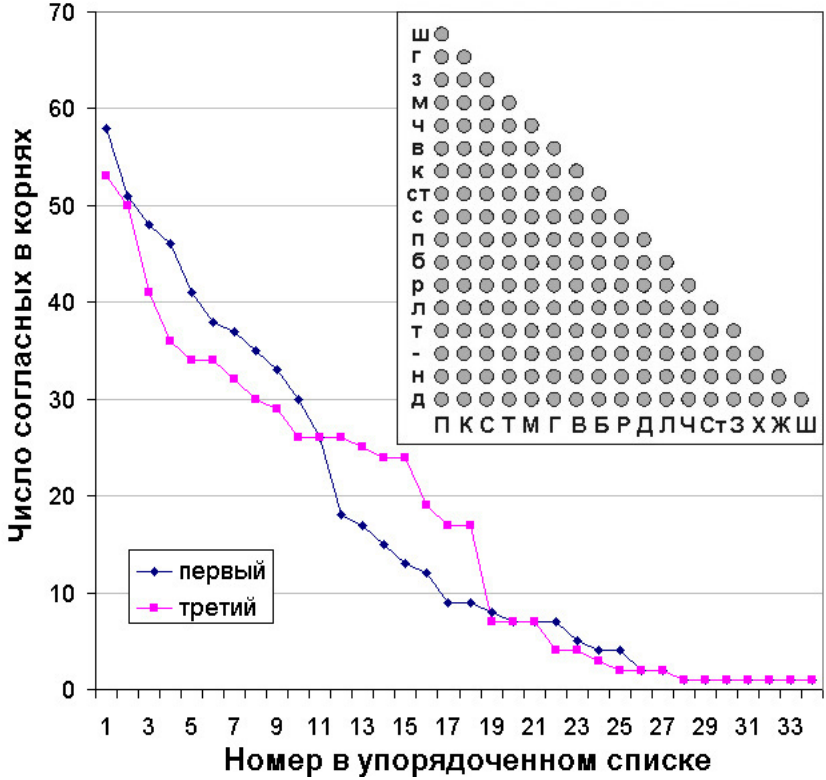

Рис. 2. Общее число согласных в стабильных глагол-образующих корнях русского языка в первой и третьей позиции. Порядок первых 17 согласных показан на вставке. Начальные согласные - по горизонтали, согласные в третьей позиции – по вертикали. Серые кружки изображают набор возможных комбинаций, дающих распределение, соответствующее графику. Это идеализированная ситуация, о реальных комбинациях – в тексте

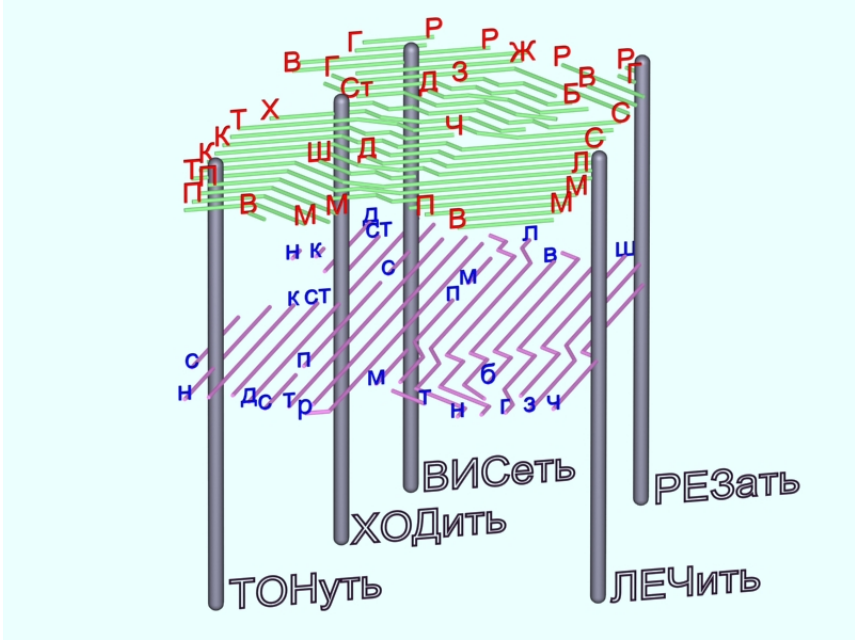

Рис. 3. Колончатая система стабильных глагол-образующих корней русского языка. Колонки плотно заполняют пространство ограниченное системой линий, показаны лишь некоторые из слов

СЕТЬ КОРНЕЙ

Результат упорядочения системы стабильных корней русского языка показан на рис. 3 и 4. Вид и размер полученной сети (измеренный числом полос) наводит на мысль о связи множества корней с характерными структурами, обнаруженными в коре мозга при изучении зрительной системы [5]. В простейшем случае это чередующиеся полоски доминирования правого и левого глаза. Характерный шаг такой полосатой структуры 0,4 мм. Полный размер отдельной «полосатой области» 10-15 мм, т.е. число полосок примерно тоже, что и для множества стабильных корней. Аналогичную, но более сложную структуру образуют полоски доминирования ориентации предъявляемого стимула (например, 30, 60 и 90 градусов

от вертикали) – они накладываются друг на друга. Это как раз тот случай, что мы имеем для корней глаголов. Основой для такого «чередующегося» представительства в коре является характер прорастания аксонов от клеток-детекторов ориентации в высшие отделы коры, где они порождают упорядоченные узоры. Это как папиллярные узоры на коже или полосатость шкуры тигра или зебры. В случае человеческой речи детекторы, по-видимому, настроены на отдельные фонемы, они то попеременно представлены в высших отделах коры, как показывают рис. 3 и 4.

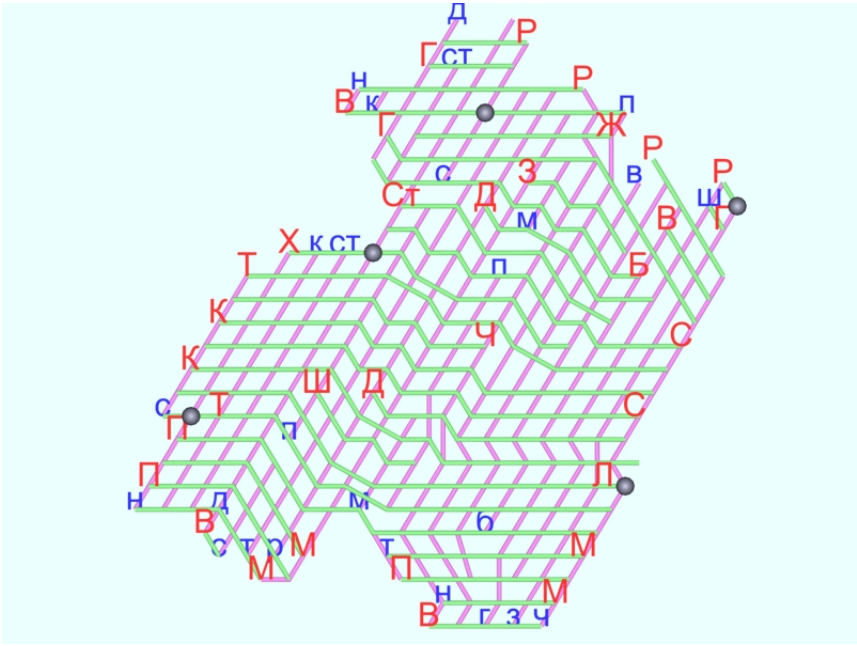

Рис. 4. Вид сверху на систему, показанную на рис. 3. Линии, идущие примерно справа-налево, соответствуют согласным из начала корня слова (заглавные буквы), а линии, идущие примерно сверху-вниз, – соответствуют согласным в конце корня (строчные буквы). Видно, что они образуют плотную сеть, в узлах которой расположены колонки, соответствующие стабильным корням слов русского языка

Мы полагаем, что колонка нейронов в коре является элементом хранения отдельного корня, и поскольку произноситься или восприниматься в связной речи может лишь один глагол, лишь одна колонка из этой области может быть активирована единовременно. Современная функциональная магнитная томография позволяет визуализировать строение таких областей коры, например, полос доминирования глаз в зрительной коре [6], что позволяет надеяться на прогресс в понимании процессов, лежащих в основе языка человека. С другой стороны, обнаруженные закономерности представления языка в коре позволяют произвести правильный выбор искусственных нейронных сетей, предназначенных для распознавания речи в компьютерных интерфейсах. Пока еще можно говорить, что имевшиеся до сих пор подходы к проблеме распознавания речи не дали эффективных результатов, вероятно, по причине неверного согласования с особенностями организации живой человеческой речи.

Список литературы

1. В.Л. Введенский, Математические закономерности словообразования в европейских языках // Нейроинформатика 2005, VII Всероссийская Научно-Техническая Конференция, Сборник научных трудов, 2005, часть 2, с. 263-270.

2. А.И. Кузнецова, Т.Ф. Ефремова, Словарь морфем русского языка. Русский язык, Москва, 1986.

3. V.L. Vvedensky, Proximity space of the European Languages. Text Processing and Cognitive Technologies, v.11, 2005, p.376-378.

4. J. Grainger, C. Whitney, Does the huamn mnid raed wrods as a wlohe? Trends in Cognitive Sciences 2004, v.8, p.58-59.

5. D.H. Hubel, Eye, Brain, and Vision. W.H. Freeman & Company, 1995.

6. R.S. Menon, S.-G. Kim, Spatial and temporal limits in cognitive neuroimaging with fMRI, Trends in Cognitive Sciences, 1999, v.3, p.207-216.