В. Д. Соловьев
V. D. Solovyev

ЧАСТОТНО-ОСНОВАННЫЙ ПОДХОД К ЯЗЫКОВОЙ ДИНАМИКЕ[1]

FREQUENCY-BASED APPROACH FOR LANGUAGE DYNAMICS

Аннотация. В статье дается обзор проведенных в КФУ в 2011–2012 гг. исследований динамики частотности языковых единиц на материале корпуса Google Books Ngram. Корпус содержит тексты на 9 языках с 1550 по 2009 г. общим объемом более 500 млрд. слов. Исследования велись в следующих основных направлениях, представленных в данной работе: уточнение законов Ципфа и Хипса, оценка скорости изменения лексического состава языков, динамика средней длины слов.

Abstract. The paper provides an overview of research carried out in Kazan University in 2011–2012 years on the dynamics of frequency of language units on the material of the Google Books Ngram corpus. The corpus includes texts on 9 languages from 1550 to 2009 year, totaling more than 500 billion words. Research was held in the following main directions, represented in the present work: refinement of the Zipf and Heaps laws, evaluation of rate of change of the languages lexical structure, dynamics of the average length of words.

Законы Ципфа и Хипса: уточнение и объяснение

Закон Ципфа, хотя и известен уже давно и много исследовался, до сих пор не имеет полного объяснения. Использование сверхбольшого корпуса текстов Google Books Ngram (**http://books.google.com/ngrams/**) позволило получить уточнение этого закона. Анализ распределения частот употребления слов в европейских языках показал наличие двух

[1] При поддержке РФФИ, грант № 12-06-00404-а

степенных участков. Для наиболее распространённых слов показатель степени 1 (как и в стандартной формулировке закона), для редких слов колеблется в интервале 1.5 – 2.4 для различных языков.

Для объяснения степенного порядка асимптотики предложены статистические модели порождения текста с независимыми буквами, а также обобщенные марковские модели. В рамках марковской модели удалось строго установить, при каком виде матрицы переходных вероятностей можно ожидать степенную асимптотику, а в каких случаях будут другие вероятностные законы (например, субэкспоненциальный). Получена явная формула для показателя степени в законе Ципфа.

Закон Хипса связывает размер коллекции (общее число вхождений слов) со словарем коллекции (общее число различных слов в коллекции) и выражается степенной функцией. Впервые проведена проверка применимости закона Хипса к сверхбольшому корпусу текстов. На рис. 1. приведена зависимость числа разных слов в корпусе от его размера и ее аппроксимации на различных участках. Оказалось, что показатель степени в этих аппроксимациях близок к среднему значению, установленному Хипсом – 0,5, но все же несколько отличается от него. Для очень больших корпусов он принимает значение, меньшее 0,5, для малых – большее 0,5.

Установлено, что наилучшее соответствие закону Хипса наблюдается для английского языка. Показано, что быстрый рост лексикона в соответствии с законом Хипса в значительной степени обеспечивается очень редкими, уникальными словами, в том числе названиями и собственными именами. Установлено, что показатель степени в законе Хипса, во-первых, имеет для всех проанализированных языков тренд в сторону уменьшения с течением времени; во-вторых, показатель испытывает значительные колебания с периодом 80–100 лет для английского языка, 75–100 для немецкого, 60–70 для французского и 50–70 для русского языка.

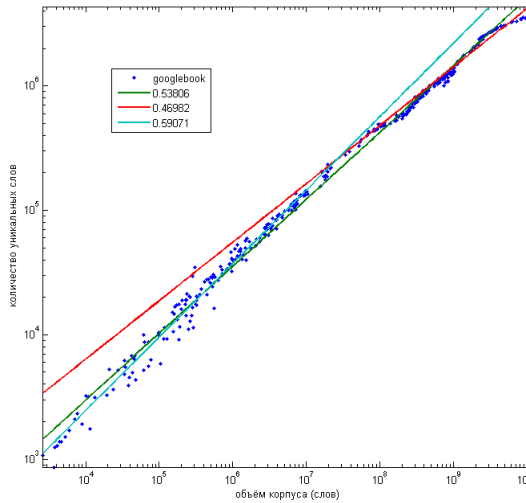


Рис. 1. Аппроксимации числа различных слов в текстах разными прямыми в зависимости от размера корпуса

Эволюция лексикона языков

Использование сверхбольшого корпуса позволило поставить вопрос о скорости эволюции всей лексики языка в целом. Ранее все исследования проводились на небольшом числе наиболее стабильных слов. Так гипотеза Сводеша относится только к ядру лексики (100 или 200 лексем).

Скорость изменения лексического состава оценивалась для английского, русский, немецкий, французский и испанский языков, начиная с 1800 года, так как более ранних данных мало и они статистически недостоверны. Изменение лексического состава языка оценивалось через расстояния между векторами частот всех слов языка в разные моменты времени. Использовалась метрика Кульбака-Лейблера изменений частотных распределений встречаемости слов за 10-летние

периоды времени. Нормированная скорость изменения лексического состава V_{norm} рассчитывается по формуле:

$$V_{norm}(t) = \frac{1}{T} \frac{D(\vec{p}(t), \vec{p}(t+T))}{H(t)}$$

где T – интервал времени (10 лет), D(p(t), p(t+T)) – значение метрики Кульбака-Лейблера для распределений p частот слов в годы t и t+T, H(t) – энтропия частотного распределения.

В качестве примера на рис. 2 приведена скорость изменения лексического состава английского языка.

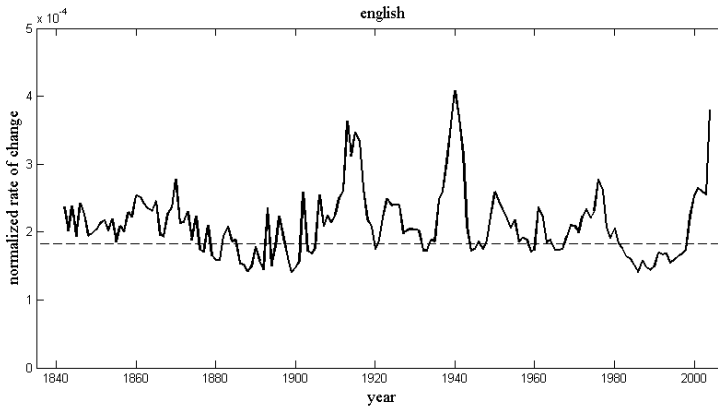


Рис. 2. Динамика скорости изменения лексики английского языка

На графике чётко выделяются всплески, соответствующие двум мировым войнам, а также значительные изменения лексики в последние 10–15 лет. Также можно видеть некоторое понижение скорости изменения частотного состава в годы «викторианской эпохи» (1860–1900). Однако за исключением указанных значительных пиков, скорость изменяется в относительно малых пределах (интерквартильный размах составляет 6.1e-5 при медианном значении 2.15e-4, то есть типичные вариации данного параметра лежат в пределах 13–14%).

Аналогичная картина наблюдается и для других языков, хотя прямые сопоставления осложняются различиями, связанными с морфологическим строем языков. Тем не менее, можно выделить некоторые закономерности, так британский английский демонстрирует больший разброс скорости изменения лексики, причем большие, чем в американском английском всплески приходятся на годы мировых войн, что видимо, отражает большую степень вовлеченности в них Великобритании.

По нашей методике подсчетов за сто лет в американском языке меняется 2,66% слов, в британском – 4,11%, в русском – 5,74%, в немецком – 5,42%, во французском – 3,43%, в испанском – 2,97%. Таким образом, как и для списков Сводеша, вся лексика меняется с относительно постоянной скоростью.

Следует отметить, что скорость и динамика эволюции лексики, применимо к нашему исследованию, зависит также от количества исследуемых лексических единиц. Чем меньше выборка анализируемой лексики, тем медленнее она эволюционирует, а изменения носят более скачкообразный характер.

С помощью предложенной методологии можно изучать процессы расхождения диалектов. Нами оценивалось расстояние между частотными распределениями слов для двух вариантов английского языка. На рис. 3 пунктирной линией представлены сглаженные данные.

В целом изменения расстояний носят достаточно регулярный характер. Вначале мы видим увеличение расстояния со временем, как отражение естественного расхождения диалектов. Затем, однако, в районе 1950–1955 гг. расхождение сменяется сближением, и, причём, достаточно быстрым, так как к 2000–му году мы практически возвращаемся к уровню различия 1840–го года. По-видимому, наблюдаемое сближение есть результат глобализации. Также установлено, что эволюция лексики в британском английском вначале опережает эволюцию в американском – до 1860–1880 гг. затем они меняются местами практически на целый век, и только в самые последние годы разрыв сокращается.

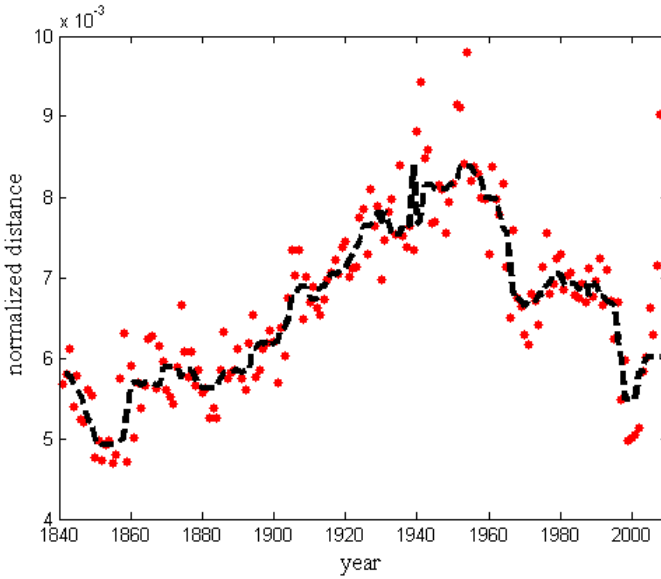
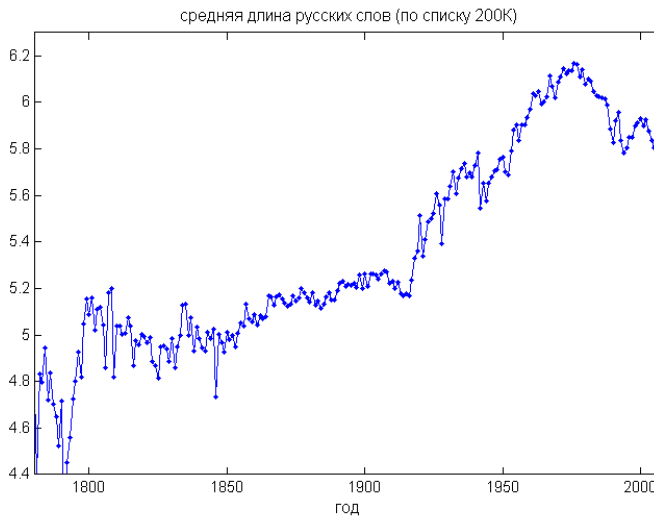


Рис. 3. Расстояния между частотными распределениями слов в британском и американском диалектах английского языка

Изменение средней длины слов

Проанализировано изменение средней длины слов в европейских языках в течение последних двух веков. Ранее динамика длины слов рассматривалась только на значительно больших временных отрезках и связывалась с изменением морфологического типа языка. Как видно на рис. 4, средняя длина слов медленно росла на протяжении 19-го века, начала быстро расти после революции, что продолжалось, примерно, до 1975 г., после чего она стала также быстро падать. В попытке объяснить эту динамику рассмотрены слова, оказывавшие наибольший вклад в увеличение и уменьшение средней длины слов.

Показано, что увеличение средней длины слов на протяжении большей части 20–го века связано с появлением и частым употреблением большого числа новых длинных слов: революция, советский и т.д. А уменьшение средней длины в



конце 20–го – начале 21–го века соотносится с резким падением частоты их употребления.

Удивительно, но аналогичным образом меняется и средняя длина слов в английском языке, только в нем влияние оказывают иные слова: development, education, information и т.д. Таким образом, на среднюю длину слов сильное влияние оказывают базовые концепции, определяющие судьбу общества.

Рис. 4. Динамика средней длины слов в русском языке.