РАНГОВАЯ СТАТИСТИКА ВСТРЕЧАЕМОСТИ СЛОВ В БОЛЬШОЙ ТЕКСТОВОЙ КОЛЛЕКЦИИ ♣

© В.А.Капустин А.А.Ямсен

Санкт-Петербургский государственный университет, Филологический факультет
vak@icape.nw.ru anna_zheleznova@mail.ru

Аннотация

Исследована ранговая статистика встречаемости словоформ на монотематической коллекции русскоязычных текстов объёмом около $6*10^7$ словоупотреблений. Показано, что распределение частоты встречаемости словоформ как функция ранга словоформы отличается как от классического закона Ципфа, так и от распределений, полученных другими авторами на русскоязычных коллекциях меньшего объёма. Предложена двухстепенная аппроксимация исследованного распределения. Количественные параметры этой аппроксимации (показатель степени в области высоких частот, равный 0.8 – 0.9, показатель степени в области низких частот, равный 1.8 – 1.9, положение области перехода от одного показателя к другому в диапазоне рангов от 4000 до 9000) близки к аналогичным результатам, полученным на Британском национальном корпусе современного английского языка. Полученные результаты соответствуют двухлексиконной модели языка Дороговцева-Мендеса [17].

1 Введение

Интерес к ранговым статистикам различных явлений существует давно [1]. Не избежали этого интереса и большие коллекции текстов [2–5], в том числе и электронные [6–12]. Основным предметом исследования в области ранговых статистик больших текстовых коллекция является частота встречаемости словоформ или лемм [2, 4 12]; немногие работы [12, 14] посвящены так называемым коллокациям — парам совместно встречающихся словоформ или лемм.

Знание ранговых статистик — частотных распределений словоформ — позволяет оптимизировать структуры инвертированных файлов, обеспечивающих поиск в текстовых коллекциях [11,15,16], однако за исключением [7,11] статистические исследования русскоязычных коллекций сосредоточены вокруг лингвистических проблем, — таких, как, например, выявление конкретных наиболее употребительных слов в языке в целом. В то же время анализ ранговых статистик позволяет сделать некоторые выводы о структуре отношений между словоформами в коллекции безотносительно к языку и к конкретному словарному составу того или иного интервала рангов [17,18].

В настоящей работе изложены результаты исследования распределения частот встречаемости словоформ на коллекции русскоязычных текстов объёмом около $6*10^7$ словоупотреблений. Полученные результаты существенно дополняют результаты аналогичных исследований небольших текстовых коллекций [7] и аналогичны полученным в [12] на Британском национальном корпусе английского языка.

Работа построена следующим образом:

- В разделе **Закон Ципфа** приведён краткий обзор состояния исследований в области ранговых статистик встречаемости слов.
- В разделе **Экспериментальные результаты** мы приводим полученные нами данные; для сравнения здесь же приведены данные из [6,7].
- В разделе **Обсуждение** проводится анализ полученных результатов.
- В **Заключении** сформулирован ряд проблем, нуждающихся в дальнейшем исследовании.

Поскольку мы не исследовали статистику распределения лемм, то в тексте работы мы будем пользоваться терминами «словоформа» и «слово» как синонимами.

2 Закон Ципфа

Г. Ципф в своей пионерской работе [2] установил, что частота F встречаемости слова в большой коллекции текстов примерно обратно пропорциональна его рангу r в частотном распределении («рангу слова»):

$$F = cr^{-1} \qquad (1)$$

где c — некоторая константа. В многочисленных исследованиях [7,13,19] установлено, что практически для всех коллекций имеет место не строгая обратная пропорциональность частоты и ранга, а

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006.

степенная зависимость с показателем, отличающимся от –1:

$$f = cr^{-k} \quad (2)$$

причём в большинстве случаев $k<1$. Все подобные распределения (частотность–ранг объекта) обычно объединяются под общим названием «закон Ципфа». Экспериментальные данные для малых ($r<10$) и больших рангов ($r>R/10$, R — наивысший для данной коллекции ранг — число различных слов в коллекции) заметно отличаются от соотношения (2). Это отличие было объяснено в [20] (см. также [19,21]) дискретным характером распределения и получило название поправок Мандельброта. Существуют многочисленные модели [18–21], приводящие к степенным распределениям, подобным (2), однако само это распределение при $k\leq 1$ имеет странное следствие: полное число словоупотреблений, несмотря на поправки Мандельброта, расходится с ростом числа возможных слов R. В [18] обсуждаются теоретические модели, которые могут приводить к такому поведению, а также принципиальный характер мезоскопичности коллекций.

В последнее время были предложены модели [17,18,12,9], в которых присутствует так называемое «обрезание» закона Ципфа для частотности словоупотреблений: начиная с некоторого ранга ($\sim 10^5$) показатель степени k должен стать >1, что обеспечивает сходимость полного числа словоупотреблений с ростом числа слов в коллекции. В [13,9] приведены экспериментальные данные по ранговым статистикам словоформ в больших коллекциях на английском (Британский национальный корпус, $4*10^8$ словоупотреблений) и баскском ($\sim 10^7$) языках, подтверждающие предложенные в [17,18,12,9] модели.

Исследования ранговых статистик словоупотреблений в коллекциях русских текстов немногочисленны [6–11] и используют относительно небольшие коллекции: до 10^5 словоупотреблений в [7], $1.6*10^7$ в [6] и $1.2*10^7$ в [11].

3 Экспериментальные результаты

Мы предприняли исследование распределения частотности употребления слов на монотематической коллекции русскоязычных текстов — коллекции федеральных законодательных актов Российской Федерации конца 90-х годов XX века. Эта коллекция состоит из 60 308 HTML-документов размером от 2 Кбайт до более чем 6 Мбайт, общий объём исходных документов около 1.5 Гбайтов.

Документы этой коллекции изобилуют буквенными обозначениями перечислений, инициалами, различного рода сокращениями. В настоящем исследовании мы воспользовались графематическим анализатором [23]. Этот инструмент довольно точно определяет использование кириллицы в перечислениях, а также выделяет последовательности слов, с высокой степенью вероятности являющиеся инициалами и фамилией. При расчёте частотности словоупотреблений мы игнорировали буквенные обозначения перечислений и инициалы.

Исследованная коллекция содержит также значительное количество сокращённых обозначений различных объектов, видов собственности, единиц измерений и пр., включающих косую черту: а/м, и/или, или/и, р/счет, а/о, кг/год, нГр/час, ИСО/МЭК и пр.. Все такие сокращённые обозначения трактовались следующим образом: если по одну из сторон косой черты находится одна буква, то сокращение рассматривалось как единая словоформа; если же все составные части сокращения состояли более чем из одной буквы, то каждая из них считалась отдельной словоформой.

Кроме сокращенных обозначений, документы коллекции содержат конкатенации слов через косую черту, появляющиеся, видимо, из-за особенностей технологии подготовки этих документов. В таблице 1 приведены некоторые примеры конкатенации слов через косую черту, полученные из документов коллекции.

Таблица 1. Примеры конкатенации через косую черту

буквы/цифры	брутто/нетто
текущего/бюджетного	расчетного/текущего
последнего/товар	оплачен/фрахт
машин/оборудования	ввоза/вывоза
уд/неуд	разрешения/лицензии
Соединенное королевство/Великобритания	
Дополнительная информация/представляемые документы	
Общая декларация/Предшествующий документ	

Такие конкатенации мы разбили на отдельные словоформы.

В документах коллекции часто встречается и другая форма конкатенации, обычно предназначенная для образования составных слов — объединение слов дефисом.

Таблица 2. Примеры составных и «квазисложных» слов

какой-либо
яслей-садов
водителей-военнослужащих
саней-волокуш
фарой-прожектором
вет-блю
дом-интернат
СОВАМ-ТЕЛЕПОРТ
организаций-экспортеров
организацией-плательщиком
Смоленская-товарная
предприятия-производства
заявлений-справок
третий-получателю
извещений-расчетов

Кроме обычных составных слов (примеры которых приведёны в первой части таблицы 2), в документах коллекции встречаются и искусственные образования (названные нами «квазисложными» словами, см. вторую часть таблицы 2). Различение «настоящих» и «квазисложных» составных слов — трудная задача, но её строгое решение, по-видимому, не слишком сильно повлияет на распределение частотности словоформ. Поэтому все слова, содержащие дефис, были разбиты нами на отдельные составляющие.

Документы коллекции содержат и другие дефекты текста, затрудняющие их корректный статистический анализ. Наиболее заметны дефекты, связанные с представлением таблиц. Таблицы в документах коллекции зачастую отформатированы как обычный текст с использованием моноширинного шрифта, а выравнивание колонок достигается вставкой необходимого числа неразрывных пробелов ( ). Такой подход потребовал принудительных жёстких переносов длинных слов, не помещающихся в колонку — в результате программа словоделения выделяет заметное количество фрагментов слов, образовавшихся в результате таких переносов.

Таблица 3. Примеры форматирования фрагментов таблиц

В ячейке HTML-таблицы:

```
реги-<BR>ональ-<BR>ного
```

В обычном HTML-абзаце (элемент <p>; документ 500008.htm):

```
|Основные ме-
|Исполнитель,
|Срок    
|Источники 
|Финансовые  Ожидаемые 
|<BR>роприятия 
|соисполни-  
|исполне-
|финансиро-
|затраты на   
|конечные  
|<BR>
|Программы и 
|тели   
|ния     
|вания    
|реализацию   
|результаты|
```

В приведённом в таблице 3 примере вертикальная черта — знак вертикальной черты, применённый для вертикального графления таблицы; количество неразрывных пробелов уменьшено по сравнению с реальным документом.

Выявить фрагменты слов, заканчивающиеся дефисом, не представляет труда, и такие фрагменты нами были отброшены. Но выявление фрагментов слов, перенесённых на другую строку (выделены в табл. 3 жирным шрифтом), невозможно без сопоставления со словарём. Сопоставление со словарём нами не выполнялось, и такие фрагменты учтены как полноценные словоформы. Как правило, частотность таких фрагментов невелика (не более 20 на всю коллекцию), и учёт таких фрагментов «поднимает хвост» распределения.

Оформление таблиц в документах исследованной коллекции приводит ещё к одному артефакту: конкатенации слов и знаков подчёркивания:

```
оборота_____________________
```

Это явление обнаружено, и слова отделены от знаков подчёркивания.

Выявленные словоформы были приведены к нижнему регистру. В результате обнаруженное количество словоупотреблений составило 56 856 781 — это наибольшая коллекция, доступная нам для исследования в настоящее время. Заметим, что без «очистки» словоформ (считая словоформой произвольную последовательность кириллических букв) количество обнаруженных в той же коллекции словоупотреблений составило около 68 млн. — на 20% больше.

Для каждой словоформы было определено количество её появлений в коллекции — её частотность. Результат исследования в двойном логарифмическом масштабе представлен на рисунке 1.

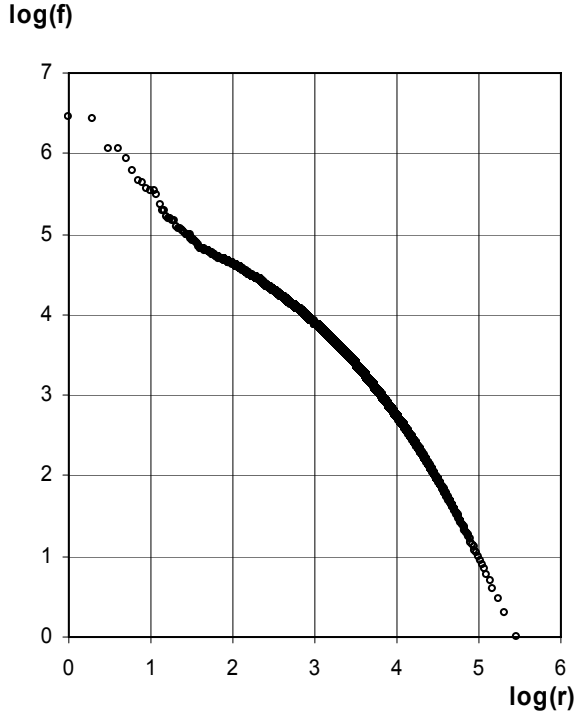

Рис. 1. Распределение частотности употребления слов

В исследованной коллекции нами обнаружено 351457 словоформ. Некоторые из них (с частотнос-

тью менее 20 и рангами $>10^5$) явно не принадлежат русскому языку, и их появление может быть объяснено особенностями формирования коллекции.

На следующем рисунке (рис. 2) представлены для сравнения результаты Гельбуха и Сидорова [7] (текст №13, в [7] не приведены числовые данные, поэтому значения для этой коллекции сняты с графика и являются приближёнными) и Шарова [6] (в работе [6] опубликованы значения частот, приведённые к 1 млн. словоупотреблений; в связи с нелинейной зависимостью распределения от объема коллекции мы не стали приводить эти значения к фактическому объёму коллекции, исследованному в [6]).

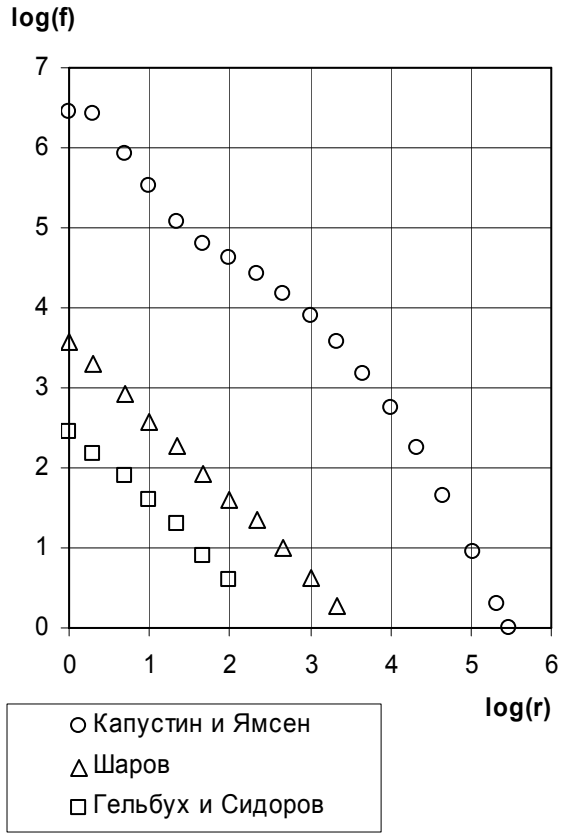

Рис. 2. Сравнительные результаты трёх исследований

Результаты всех трёх исследований неплохо соотносятся. Проведённый нами анализ показал, что на исследованной коллекции, так и на коллекции [7], показатель степени в законе Хипса [22] близок к 1 (в [6] анализ роста числа уникальных словоформ с ростом коллекции не проводился).

Подгонка показателя степени закона Ципфа (2) даёт следующие результаты:

Таблица 4. Одностепенное приближение

Исследование	**k**
Гельбух и Сидоров	0.93
Шаров	0.99
Капустин и Ямсен	1.12

Для коллекции Гельбуха и Сидорова в [7] приведено значение показателя k, равное 0.90. Расхождение может объясняться как тем, что мы использовали приближённые значения частот, так и тем, что для подгонки k мы учитывали меньше точек, чем было учтено в [7]. Для коллекции Шарова мы также учли только представленные на рис. 2 точки. Детальный анализ численных данных [6] показывает, что для рангов, б*о*льших представленных на рис. 2, наблюдается «загиб распределения вниз»: все полученные в [6] значения частот в этой области рангов лежат ниже аппроксимирующей прямой с показателем степени, равным 0.99, а использование этих точек для подгонки показателя степени (2) приводит к значению >1. К сожалению, в [6] приведены значения частот только до ранга 69307 (более одного появления слова на миллион словоупотреблений, что соответствует более чем 16 появлений слова в коллекции Шарова), что отвечает очень небольшой области рис. 2, лежащей правее и ниже последней приведённой на этом рисунке точки данных Шарова.

4 Обсуждение

Наши данные по частотности словоупотреблений имеют заметное отличие от степенного закона (2): относительное отклонение в логарифмическом масштабе составляет в среднем около 10%, заметно превышая 50% для точки с наивысшим учтённым рангом (292 277). Поэтому по аналогии с [9,12] мы предположили, что начало распределения (для рангов, меньших некоторого ранга R) имеет показатель, близкий к традиционному ($k_1 = 0.8 - 0.9$), а хвост (для рангов, больших R) — существенно б*о*льший показатель k_2:

$$F(r) = \begin{cases} c_1 r^{-k_1} & r \le R \\ c_2 r^{-k_2} & r > R \end{cases} \qquad (3)$$

Пять параметров формулы (3) связаны соотношением

$$c_2 = c_1 + (k_2 - k_1) * \log(R) \qquad (4)$$

поэтому для проверки гипотезы (3) была выполнена четырёхпараметрическая $\{c_1, k_1, k_2, R\}$ подгонка двух степенных зависимостей, сопрягающихся в точке R.

Измеренные значения (как ранги, так и частоты) изменяются в области подгонки на 6 порядков. Это делает более или менее бессмысленным минимизацию суммы квадратов отклонений теоретической кривой (3) от измеренных точек, поскольку для больших рангов (и, соответственно, малых частот) сами частоты (и, соответственно, отклонения теоретической кривой от измеренных точек) на 6 порядков отличаются от значений (и отклонений) в области малых рангов (высоких частот). Мы решили минимизировать сумму квадратов отклонений логарифмов частот в зависимости от логарифмов рангов — по аналогии с подходом, использованным Гельбухом и Сидоровым в [7].

4.1 Подгонка двухстепенной зависимости по всем измеренным точкам

При избранном подходе возникает другая проблема — измеренные точки расположены неравномерно, их плотность по оси log(r) растёт пропорционально логарифму ранга. Если не компенсировать это неравномерное распределение точек, то значения высоких рангов (малых частот) будут доминировать в минимизируемой целевой функции. Фактически мы минимизировали следующее выражение:

$$\sum_{r\le R}\left(\varphi(r)-C_1-k_1 r\right)^2 10^{-r} + \sum_{r>R}\left(\varphi(r)-C_2-k_2 r\right)^2 10^{-r} \qquad (5)$$

Здесь

$C_i = \log(c_i)$ $(i=1,2)$

$\varphi(r) = \log(f(r))$

$f(r)$ – измеренное значение частоты словоформы с рангом r

Основание логарифма равно 10.

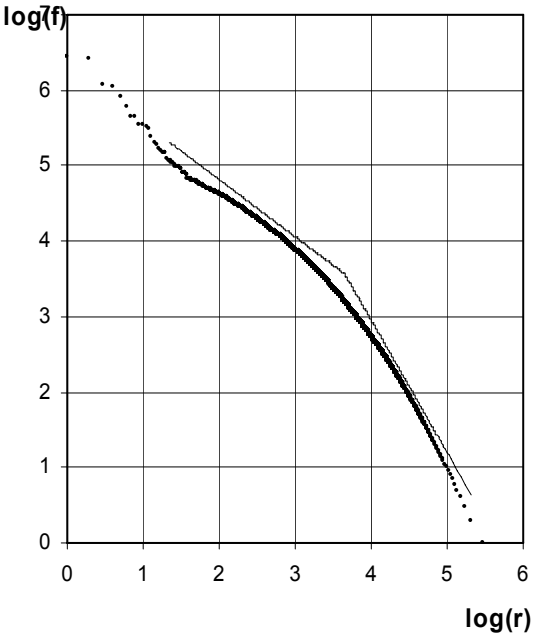

Рис. 3. Два степенных закона — подгонка по всем измеренным точкам (за исключением флуктуаций Мандельброта)

Результат подгонки представлен на рис. 3. (области флуктуаций Мандельброта исключены из расчёта; ломаная линия, полученная в результате подгонки, приподнята на 0.2 над экспериментальной кривой).

Относительные отклонения для всех точек не превышают 5%. Полученные значения степенных показателей составляют 0.76 для рангов, не превышающих 4490, и 1.76 для б*о*льших рангов. Наличие «квазисловоформ» несколько уменьшает значение показателя k_2.

4.2 Подгонка двухстепенной зависимости по точкам, равномерно распределённым в логарифмической шкале рангов

Вместо взвешивания отклонений можно выбросить часть измеренных точек с тем, чтобы оставшиеся были распределены равномерно по оси абсцисс [9]. Результат такой подгонки изображён на рис. 4.

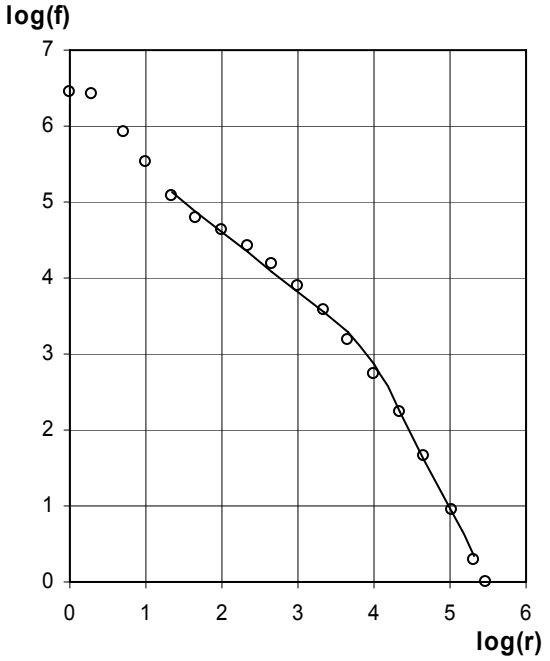

Рис. 4. Два степенных закона — подгонка по равномерно распределённым точкам

И опять относительные отклонения для всех точек не превышают 5%. Полученные значения степенных показателей составляют 0.78 для рангов, не превышающих 6068, и 1.85 для б*о*льших рангов.

4.3 Сравнение двухстепенных приближений

Полученные результаты поразительным образом достаточно хорошо согласуются с результатами [9], полученными на Британском национальном корпусе английского языка (табл. 5 и 6)

Таблица 5. Двухстепенные приближения (без учёта областей флуктуаций Мандельброта)

Исследование	k_1	k_2	R
Капустин и Ямсен – полная подгонка	0.76	1.76	4490
Капустин и Ямсен – подгонка по равномерно распределённым точкам	0.78	1.85	6068
Cancho, Solé [9]	1.06	1.97	$10^3<R<10^4$

Таблица 6. Двухстепенные приближения (с учётом областей флуктуаций Мандельброта)

Исследование	k_1	k_2	R
Капустин и Ямсен – полная подгонка	0.89	1.90	8760
Капустин и Ямсен – подгонка по равномерно распределённым точкам	0,89	1,90	8246

Мы не приводим графики моделей распределения, учитывающих при подгонке области флуктуаций Мандельброта.

К отклонению от закона Ципфа (2), описываемому формулой (3), могут приводить различные причины, не последней из которых может являться монотематичность исследованной коллекции. Однако сходство полученных нами и в [9] параметров двухстепенного закона (3) говорит, скорее, в пользу общих механизмов [17], лежащих в основе наблюдаемого в обоих случаях отклонения от (2): в лексиконе, образующем тексты коллекций, есть небольшое ядро, состоящее из ~10^4 словоформ, и периферия, образуемая остальными словоформами. Способы употребления слов ядра и периферии существенно различаются (слова ядра чаще употребляются совместно друг с другом; слова периферии употребляются почти исключительно совместно со словами ядра [17]), что и приводит к изменению показателя степени распределения частот словоупотреблений на границе ядра лексикона коллекции. При этом конкретный состав лексикона, видимо, не имеет значения — слова «организации» и «деятельности» для юридических текстов также являются ядерными, — как и предлоги и союзы.

4.4 Определеине точности

Проведённые нами измерения и расчёты (впрочем, как и аналогичные измерения и расчёты других авторов [2,4–7,9,11]) не позволяют надёжно определить точность полученных результатов. Для определения точности необходимо выполнить измерения и подгонку формулы (3) на ансамбле коллекций достаточно большого объёма. При этом возникнет две новые проблемы: обеспечение однородности ансамбля и трудоёмкость вычислений (выполненные нами вычисления потребовали нескольких часов расчётов).

5 Заключение

Нами обнаружено существенное отличие ранговой статистики встречаемости словоформ в большой текстовой коллекции от известного закона Ципфа. Наблюдаемое отличие хорошо моделируется композицией (3) двух степенных зависимостей с различными показателями степени для малых и больших рангов. Переход от одной зависимости к другой лежит в области рангов 10^3 – 10^4, что может быть интерпретировано моделью, предложенной в [17], как наличие ядерного множества словоформ, структура употребления которых существенно отличается от структуры употребления большинства слов из лексикона коллекции.

Для более надёжного анализа гипотезы о существовании ядерного множества словоформ необходимо исследовать структуру совместного использования словоформ — коллокации. Для Британского национального корпуса исследование парных коллокаций было выполнено в [12], а в [17] была представлена модель, объясняющая как двухстепенное поведение частот словоупотреблений, так и зависимость частоты парных коллокаций от их рангов. Из этой модели следует, что слова достаточно большой коллекции образуют так называемый «тесный мир», в котором должно наблюдаться значительное число тройных коллокаций. Это предсказание, однако, пока не проверено экспериментально. Исследование совместной встречаемости слов может привести к новым способам навигации по коллекции — по словам с высокой степенью связи.

Нами выполнено предварительное исследование парных коллокаций на той же коллекции. Картина распределения коллокаций напоминает аналогичную картину, полученную в [12]. Детальный анализ коллокаций будет представлен нами в отдельной работе.

Теоретический интерес также представляет изучение частотных распределений встречаемости лемм и коллокаций лемм. Использование морфологического анализатора также может улучшить структуру «хвоста» распределения за счёт удаления квазислов.

Конкретное значение показателей степеней в (3), в особенности показатель степени убывания частоты словоупотреблений на больших рангах, может существенно повлиять, например, на процедуры сжатия пост-листов в инвертированных файлах, обеспечивающих поиск в коллекциях большого объёма.

Мы планируем продолжить исследование ранговых статистик различных распределений на больших текстовых коллекциях, в том числе и с применением компьютерных кластеров и, возможно, Grid.

Литература

[1] Pareto V. Cours d'économie politique professé à l'université de Lausanne, vol. 1, Lusanne: F. Rouge, 1896. — p. 341

[2] Zipf G. K. Human Behaviour and the Principle of Least-Effort. — Cambridge MA: Addison-Wesley, 1949

[3] Le code Voynich / Ed. Pierre Barthélemy. — Jean-Claude Gawsewitch Editeur: Paris, 2005. — ISBN 2-35013-022-3. — См. также http://webtext.library.yale.edu/beinflat/pre1600.ms408.htm.
[4] Частотный словарь современного русского литературного языка // Штейнфельдт Э. А., сост. — М., 1973.
[5] Частотный словарь русского языка // Засорина Л. Н., ред. — М.: Русск. яз.,1977. — 936 с.
[6] Шаров С. А. Частотный словарь. — http://www.artint.ru/projects/frqlist.asp
[7] Гельбух, А. Ф., Сидоров Г. А. Коэффициенты законов Ципфа и Хипса для русского и английского языков // Научно-техническая информация, Сер. 2, Информационные процессы и системы. — 2001. — N 9. — С. 32-36. — ISSN 0548 0027
[8] Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language. // Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics (February 18–24, 2001, Mexico City) // Lecture Notes in Computer Science. — Springer-Verlag, 2001. — N 2004. — p. 332–335. — ISSN 0302-9743. — ISBN 3-540-41687-0
[9] Cancho R. F., Solé R. V. Two regimes in the frequency of words and the origin of complex lexicons // Journal of Quantitative Linguistics. — 2001. — N8. — p.165-17. — Santa Fe Institute Preprint SFI-00-12-068.
[10] Sharoff S. Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. — Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain. — http://gandalf.aksis.uib.no/lrec2002/
[11] Губин М. В. Изучение статистики встречаемости терминов и пар терминов в текстах для выбора методов сжатия инвертированного файла // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Четвёртой Всероссийской научной конференции RCDL'2002 (Дубна 15–17 октября 2002 г.). — Дубна: ОИЯИ, 2002. — Т.2. — с.26–38
[12] Cancho R. F., Solé R. V. The small world of human language // Proceedings of The Royal Society of London. Series B, Biological Sciences. — 2001. — N 268(1485). — p. 2261–2265
[13] Solé R. V., Corominas B., Valverde S. and Steels L. Language Networks: their structure, function and evolution. — Santa Fe Institute Preprint SFI-05-12-042
[14] Поиск биграмм. — http://www.aot.ru/demo/bigrams.html
[15] Кнут Д. Э. Искусство программирования. Т. 3. Сортировка и поиск. — М.: Вильямс, 2005. — ISBN: 5 8459 0082-4. — с. 469
[16] Trotman A. Compressing inverted files. Information Retrieval. — V. 6(1). — 2003. — p. 5–19.
[17] Dorogovtsev S. N., Mendes J. F. F. Language as an evolving word web // Proceedings of The Royal Society of London. Series B, Biological Sciences. — 2001. — N 268(1485). — p. 2603-2606
[18] Dorogovtsev S. N., Mendes J. F. F. Evolution of Networks: From Biological Nets to the Internet and WWW. — Oxford University Press, Oxford, ISBN: 0198515901 2003. —http://www.fyslab.hut.fi/~sdo/evolution_of_networks.pdf
[19] Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А. О смысле ранговых распределений. — http://www.kudrinbi.ru/public/442/index.htm
[20] Мандельброт Б. Фракталы, случай и финансы: Пер. с англ. — Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2003. — 255 с.
[21] Шлык В. А. Он оставил царапину на поверхности всего. К 80-летию Бенуа Мандельброта. — http://www.computer-science.ru/index.php?id=78
[22] Heaps H. S. Information retrieval: Computational and theoretical aspects. — 1978. — p. 206–208
[23] Сокирко А.В. Графематика. — http://www.aot.ru/docs/graphan.html

Rank Statistics of Word Occurrence in a Big Text Collection

Victor Kapustin, Anna Jamsen
Saint-Petersburg State University, Faculty of Philology

We study rank statistics of occurrence of word forms in a big collection of text documents in Russian. The collection size is 60 308 documents, ~$6*10^7$ word occurrences. The found distribution of word form frequency as a function of word form rank differs considerably both from the classic Zipf law and from the results of other researches done on Russian text collection of smaller sizes. We use two powers approximation to fit the experimental results. The powers that fit the data (0.8 – 0.9 in high frequency region, 1.8 – 1.9 in low frequency region with a crossover at the rank of 4000 — 9000) closely corresponds with a similar research done on the British National Corpus of contemporary English. The results obtained agree with Dorogovtsev-Mendes two-lexicons model of language [17].

♣ Авторы признательны Консорциуму «Кодекс» и, в особенности, М. В. Губину за предоставленную возможность исследования коллекции Федерального законодательства Российской Федерации.
Авторы благодарят участников Семинара по корпусной лингвистике ИЛИ РАН (рук. В. П. Захаров) и в особенности Г. Я. Мартыненко за полезное обсуждение.