УДК 81’32

НИЗКОЧАСТОТНЫЕ СЛОВА В РУССКОМ ЯЗЫКЕ И ПОДХОДЫ К МОДЕЛИРОВАНИЮ ОБЩЕЯЗЫКОВОЙ ЧАСТОТНОСТИ [1]

Ольга Владимировна Блинова
к. филол. н., доцент кафедры общего языкознания
Санкт-Петербургский государственный университет
199034, Санкт-Петербург, Университетская набережная 7/9. o.blinova@spbu.ru

Целью настоящей статьи является выработка методики формирования списков низкочастотных слов путем сравнения данных, предоставляемых русскими корпусами. В статье сравниваются частотные списки лемм, представленные в «Новом частотном словаре русской лексики» на базе НКРЯ и полученные на базе веб-корпуса ruTenTen11. Перед сравнением исходные списки были преобразованы; в результате преобразования получены списки общей длиной, соответственно, 51 681 слово и 457 935 слов. Сопоставлены списки слов, имеющих относительные частоты <5 ipm по данным хотя бы одного из корпусов. В качестве нижнего порога отсечения вынужденно выбрано значение абсолютной частоты, равное 37. Посчитаны значения мер «coverage» (охват) и «enrichment» (обогащение). Выяснилось, что мера «coverage», отражающая долю перекрытия между списками лемм, принимает значение в 9,4%.

Ключевые слова: русский язык; языковые корпусы; веб-корпусы; частотный список лемм; общеязыковая частотность; низкочастотные слова; лексическая сложность.

1. Постановка задачи

Информация о частотности слов применяется в самых разнообразных исследованиях (см., например, [Gries, Divjak 2012]). В настоящей статье рассматривается проблема определения так называемой общеязыковой частотности слова, точнее, более узкая проблема выделения слов с низкой общеязыковой частотностью (такие слова будем называть также редкими).

Термин «общеязыковая частотность» (general-language frequency) подразумевает, что встречаемость слова устанавливается для языка вообще. Со всей ясностью понимая, что корпус – это не язык, к эмпирическому решению проблемы установления общеязыковой частотности приходится подходить с корпусными данными в руках.

Такие данные предоставляют в том числе большие русские корпусы, то есть частотные списки на их основе. Кроме того, важнейшим источником информации об общеязыковой частотности слов является «Новый частотный словарь русской лексики» (НЧСРЛ), основанный на жанрово-сбалансированном подкорпусе Национального корпуса русского языка объемом в 92 млн. графических слов [Ляшевская, Шаров 2009].

Информация об общеязыковой частотности необходима при оценке лексической сложности текстов корпуса русских локальных документов CorRIDA в рамках реализации проекта по изучению объективной и перцептивной сложности русских официальных документов (см., например, [Белов и др. 2018]). Так, лексическая сложность может быть определена через установление в тексте доли редких (resp., низкочастотных) слов [Ляшевская 2017]. Корпус CorRIDA (1,5 млн. слов) собран в 2018 г. и включает тексты, взятые из Интернета. Словник НЧСРЛ не покрывает всего лексического разнообразия, наблюдаемого в текстах корпуса CorRIDA, кроме того, у нас нет данных об общеязыковой частотности для вхождений, наблюдаемых внутри CorRIDA с высокой периодичностью (среди таких вхождений, например, аббревиатуры *ГБУЗ* ‘государственное бюджетное учреждение здравоохранения’, *ОМС* ‘обязательное медицинское страхование’, *ЦРБ* ‘центральная районная больница’, *РБ* ‘районная больница’, *ДМС* ‘дополнительное медицинское страхование’, *ЛПУ* ‘лечебно-профилактическое учреждение’, *СНИЛС* ‘страховой номер индивидуального лицевого счета’ и др.). Таким образом, даже если взять НЧСРЛ в качестве источника информации об общеязыковой частотности и пользоваться традиционным «порогом отсечения» низкочастотных слов (тогда слова с относительной частотой ниже 5 ipm

будут считаться «низкочастотными»), охарактеризовать тексты корпуса с указанных позиций не получится.

Между тем, не вполне ясно, какие слова в принципе можно считать редкими. Так, например, в работе М.В. Хохловой «Large Corpora and Frequency Nouns» [Khokhlova 2016] указано, что, во-первых, «выбор низкочастотных существительных довольно сложен», во-вторых, показатели частотности в НЧСРЛ и в больших корпусах (Russian Web Corpus, ruTenTen, ruTenTen Sample) для низкочастотных слов сильно различаются (и различия между указанными корпусами и НЧСРЛ для низкочастотных слов более очевидно, чем для высокочастотных).

В настоящей статье сравниваются данные о частотности слов, взятые из двух источников: НЧСРЛ и корпуса ruTenTen11 (15 млрд слов). Представляется, что на этом этапе можно ограничиться использованием лишь одного из крупных русских веб-корпусов, так как веб-корпусы Russian Web Corpus, ruTenTen и ruTenTen Sample при сравнении между собой показали достаточно высокую согласованность в том, что касается показателей частотности, см. [там же].

Целью статьи является выработка методики формирования списков низкочастотных слов путем сравнения данных, предоставляемых русскими корпусами.

2. Некоторые подходы к проблеме выделения низкочастотных слов

Распределение частот слов (а также последовательностей слов, например, биграмм, триграмм и т. д.) в корпусах и коллекциях текстов обычно формулируется как некая универсалия типа «несколько очень высокочастотных слов и длинные “хвосты” очень редких слов» (см., например, [Baroni 2008: 812]). Закономерность распределения частот описывает закон Ципфа, устанавливающий обратно пропорциональную зависимость между частотой слова и его рангом (порядковым номером в ранжированном списке, где все слова корпуса/коллекции расположены по убыванию частоты) [Piantadosi 2014]. Закон предсказывает быстрое снижение частоты, это снижение становится медленнее с ростом ранга, при этом наблюдаются очень длинные «хвосты» слов с одинаковыми низкими частотами [Baroni 2008: 813]. Одним из практических последствий наблюдаемого распределения частот является так называемая «проблема разреженности данных» (data sparsity problem): независимо от того, насколько велик корпус, большинство слов будет иметь низкую частотность и даже большие корпусы не смогут осуществить выборку всего словаря языка [там же].

На практике составители частотных словарей пользуются пороговыми значениями, ограничивающими список слов по частоте употребления [Ляшевская 2016: 236]. Например, для 100-миллионных корпусов принято ограничивать частотные списки словами с относительной частотой около 5 ipm [там же]. Частотный порог в НЧСРЛ для алфавитного списка составляет 0,4 ipm.

В разнообразных работах, привлекающих сведения о частотности лексических единиц, для выделения высокочастотных/частотных, среднечастотных и низкочастотных слов используются различные пороговые значения. Это значения, характеризующие позиции элементов в ранжированном частотном списке (т. е. ранги), значения относительных, логарифмированных (или даже просто абсолютных) частот, значения, учитывающие распределение слов по сегментам или документам корпуса (т. е. дисперсию) [Ахутина 2014: 71–73]. Скажем, в [Láznička, Janda 2019] низкочастотными считаются слова с ipm < 7; в [Zhao, Jurafsky 2009] в качестве низкочастотных рассматриваются слова из нижнего квартиля ранжированного частотного списка. В работе [Bell et al. 2009] частотный диапазон разделен на три части: высокочастотные слова, среднечастотные слова и низкочастотные слова. Первый порог деления выбран в точке 5-го процентиля списка полнозначных слов, а второй порог деления – в точке 95-го процентиля списка полнозначных слов. В значительном количестве работ пороговые значения (как и вообще основания для выбора слов из частотных списков как «высокочастотных» или «низкочастотных») не эксплицируются.

3. Сбор и представление данных

В настоящей статье сравниваются алфавитный список лемм, данный в csv-версии НЧСРЛ, и данные корпуса ruTenTen11 из семьи TenTen с платформы Sketch Engine [Kilgariff et al. 2014]. В используемой версии Sketch Engine есть возможность выгружать списки слов не длиннее 1000 строк. Поэтому для получения наиболее полного частотного списка была выбрана стратегия, согласно которой из ruTenTen11 выгружались частотные списки лемм, начинающихся на возможные двухбуквенные сочетания (последовательно: *аб, ав, аг* и т. д.). Список двухбуквенных сочетаний, возможных в начале лемм, получен на основе 52-тысячного списка лемм НЧСРЛ. Для однобуквенных лемм был выполнен отдельный поиск.

Сформированный таким способом частотный список лемм на базе корпуса ruTenTen11 состоит из 457 935 строк. В нём содержатся только сведения об абсолютной частоте, поэтому относительные частоты леммам нужно было присваивать самостоятельно (что и было сделано из расчета, что объем корпуса составляет 14 553 856 113 слов). Данных, которые могли бы служить основанием для суждений о распределении лемм по документам или сегментам корпуса, в моем распоряжении нет. В списке присутствуют ошибки лемматизации, лексические дублеты, например *абажур* (абс. частота 21 618) и *абажурах* (абс. частота 299), неорфографичные письменные репрезентации слов, например *абажаю* (абс. частота 487).

Перед сравнением частотные списки из НЧСРЛ и из ruTenTen11 следовало подвергнуть предобработке, в частности, заменить все «ё» на «е», так как в НЧСРЛ «ё» не используется, и суммировать частоты омографов, так как в списке ruTenTen11 леммы типа *але* (междометие) и *але* (частица), *бездомный* (существительное) и *бездомный* (прилагательное), *будто* (союз) и *будто* (частица) и т. д. не разведены.

Затем в каждом из списков каждому вхождению был присвоен ранг. Ранг присваивался по принципу, реализованному в НЧСРЛ (т. е. у каждого следующего слова в ранжированном частотном списке ранг больше на единицу, а леммы с одинаковым количеством вхождений имеют разные ранги) [Ляшевская 2016: 229].

Кроме того, были посчитаны средние значения ipm (imp mean) для всех лемм, содержащихся в списке, а сам список был отсортирован по убыванию imp mean. В результате для объединенного частотного списка лемм НЧСРЛ и ruTenTen11 был сформирован традиционный «rank/frequency profile» (см. Таблицу 1).

Таблица 1

Фрагмент объединенного частотного списка, отн. частота

Lemma	ipm RNC	ipm ruTenTen11	ipm mean	rank RNC	rank ruTenTen11
и	35801,800	34622,753	35212,276	1	1
в	31374,200	34440,504	32907,352	2	2
на	15870,800	16575,012	16222,906	5	3
не	18028,000	13288,503	15658,251	3	4
что	16098,300	9658,545	12878,422	4	7
с	11316,000	11219,013	11267,507	9	5
быть	12160,700	10148,599	11154,650	7	6
он	11791,100	7827,361	9809,230	8	8
я	12684,400	5881,327	9282,863	6	11
а	8226,600	5940,911	7083,756	10	10
по	5790,800	7567,131	6678,966	13	9
как	6626,700	5686,558	6156,629	11	14
это	6174,500	5812,341	5993,420	12	13
этот	5414,000	5203,444	5308,722	15	17
к	5389,000	4973,073	5181,037	16	19
они	4850,000	5388,613	5119,306	19	16
но	5382,900	4109,520	4746,210	17	21
который	4209,000	5142,417	4675,709	23	18
для	3229,300	5816,463	4522,882	32	12
из	4314,100	4365,481	4339,790	21	20

4. Анализ данных

Сравнение данных, представленных в «rank / frequency profile» показало, что сводный список содержит в общей сложности 41 252 леммы с ненулевыми значениями относительных частот в каждом из сравниваемых списков. При этом, как показывает Таблица 2, где леммы ранжированы по ipm mean, леммы с пороговым ipm = 0,4 в НЧСРЛ демонстрируют в ruTenTen11 разброс значений относительных частот в диапазоне от 33,508 ipm (слово *отел*) и 10,025 ipm (слово *утилита*) до 0,0003 ipm (слово *членораздельно*).

Конечно, существуют классические проверенные техники сопоставления корпусов, в частности, через сравнение наблюдаемых в разных корпусах частот слов [Kilgarriff 2001; Kilgarriff,

Rose 1998], однако цель настоящей статьи не заключается в сравнении корпусов per se. Поэтому было бы правильно выбрать методику, в рамках которой сравниваются не абсолютные, относительные частоты или ранги конкретных лемм в частотных списках для корпусов попарно, но оценивается общий состав частотных списков и различия между ними. Такую методику предложили М. Барони и его соавторы в [Baroni et al. 2009] (см. также [Lindemanna, San Vicente 2015]).

Методика подразумевает сравнение корпусов с применением двух мер. Первая называется «coverage» (охват, покрытие), вторая называется «enrichment» (обогащение). Меры призваны показать, в какой степени список лемм в одном корпусе «покрывается» списком лемм в другом корпусе. Точнее, согласно М. Барони, , «охват корпуса X относительно корпуса Y (охват (Y/X)) представляет собой долю types <долю уникальных токенов – *О.Б.*>, которые находятся выше порога Синклера (Sinclair cutoff) как в X, так и в Y, по отношению к общему числу types, находящихся выше порога Синклера в X»; «обогащение корпуса X относительно корпуса Y (обогащение (Y/X)) измеряет долю слов, которые находятся выше порога Синклера в корпусе Y, но ниже этого порога в корпусе X, по отношению к общему числу types, находящихся ниже порога в корпусе X» [Baroni et al. 2009: 11].

Таблица 2

Фрагмент объединенного частотного списка (20 лемм с низкими значениями ipm mean)

Lemma	ipm RNC	ipm ruTenTen11	ipm mean
этнонациональный	0,4	0,009	0,205
птицеперерабатывающий	0,4	0,009	0,204
рыбодобывающий	0,4	0,009	0,204
тыняновский	0,4	0,009	0,204
зэковский	0,4	0,007	0,204
лживо	0,4	0,007	0,204
угреться	0,4	0,007	0,204
офонареть	0,4	0,007	0,204
едреный	0,4	0,007	0,203
вгиковский	0,4	0,007	0,203
жэковский	0,4	0,005	0,203
вдовствующий	0,4	0,005	0,203
знакомо	0,4	0,004	0,202
рюшечка	0,4	0,004	0,202
цыпленок-бройлер	0,4	0,003	0,202
гэкачепист	0,4	0,003	0,201
эфросовский	0,4	0,002	0,201
вмуровать	0,4	0,002	0,201
нюни	0,4	0,001	0,200
членораздельно	0,4	0,000	0,200

Показатель «coverage» рассчитывается по формуле

(1) $coverage\ (C1/C2) = (N1 \cap N2)\ /N1,$

где $C1$, $C2$ – корпусы, $N1$ – количество лемм с абсолютной частотой ≥ 20 в корпусе 1, $N2$ – количество лемм с абсолютной частотой ≥ 20 в корпусе 2.

Показатель «enrichment» рассчитывается по формуле

(2) $enrichment\ (C1/C2) = N2/S1,$

где $N2$ – количество лемм с абсолютной частотой ≥ 20 в корпусе 2, $S1$ – количество лемм с абсолютной частотой < 20 в корпусе 1.

В списке НЧСРЛ нет лемм с частотой < 20, так как порогом отсечения тут является относительная частота 0,4 ipm (примерно соответствующая для корпуса объемом 92 млн графических слов абсолютной частоте, равной 37). Соответственно, сравнение с использованием порога Синклера в 20 вхождений работать не будет.

Для того чтобы сравнение стало возможным, решено взять из общего списка ruTenTen только список лемм, чья абсолютная частота ≥ 37 (примерно такая абсолютная частота наблюдается у слов с относительной частотой 0,4 ipm в НЧСРЛ). Кроме того, в рамках настоящей статьи решено сравнивать частотные списки лемм, имеющих относительную частоту < 5 ipm хотя бы в одном из корпусов, так как именно такие слова можно предварительно считать низкочастотными.

Таким образом, была выбрана следующая стратегия формирования частотных списков, подлежащих сравнению: на первом этапе были выбраны все леммы с относительной частотой < 5 ipm (точнее, были выбраны все леммы, имеющие ipm < 5 хотя бы в одном из сравниваемых списков), затем из списка ruTenTen11 были исключены все леммы с абсолютной частотой $f < 37$.

В полученном сводном списке оказалось 307 218 лемм, однако он содержит заметное количество ошибок лемматизации и другого шума. Размах абсолютных частот (т. е. разность между максимальным и минимальным значениями), наблюдаемая для ruTenTen11, составляет 4 417 216, что, конечно, весьма много. В Таблице 3 даны 10 позиций с максимальными значениями абсолютных частот в ruTenTen11; эта таблица, как представляется, хорошо показывает различия между сравниваемыми частотными списками.

Таблица 3

Фрагмент объединенного частотного списка ($f \geq 37$)

Lemma	Freq RNC	Freq ruTenTen11	ipm RNC	ipm ruTenTen11
деньга	101,2	4417253	1,1	303,511
м	358,8	3570631	3,9	245,339
см		3276740		225,146
две		3264920		224,334
тыс		3105772		213,399
руб		3030732		208,243
втора	92,0	2942075	1,0	202,151
ст		2530321		173,859
др		2097536		144,122
скачивать		1862475		127,971

Понятно, что некоторые позиции соотнесены скорее случайно. Так, трудно предположить, что именно репрезентации леммы *втора* (второй голос в музыкальной партии, вторая партия, ipm в НЧСРЛ = 1) 2 942 075 раз встретились в составе веб-корпуса. Кроме того, морфоанализаторы присваивают словоформам глаголов разные леммы (ср. *скачать – скачивать*) [Ляшевская 2016: 228], что служит еще одной причиной существенных расхождений между сравниваемыми частотными списками. Наконец, леммы с относительной частотой 300, 200 и др. ipm в ruTenTen11 никак нельзя считать «низкочастотными».

При всем этом некоторые выводы с применением мер «coverage» и «enrichment» получить можно. Так, если учитывать, что $N1$ (количество лемм с $f \geq 37$) в списке на базе ruTenTen11 равно 307 218 и при этом совпадений в списках $N1 \cap N2$ – 28 759, мы получим «coverage» = 0,094. Имея в виду, что $S1$ (количество слов с $f < 37$) в ruTenTen11 = 137915, мы получим «enrichment» = 0,21.

5. Заключение

Итак, в настоящей статье сравнивались частотные списки лемм, представленные, во-первых, в НЧСРЛ, созданном на базе выборки из НКРЯ (ruscorpora.ru), во-вторых, в списке на базе веб-корпуса ruTenTen11 (sketchengine.eu). В том числе были сопоставлены списки слов, имеющих относительные частоты < 5 ipm по данным хотя бы одного из корпусов.

Для меры «coverage» (показывающей, в какой степени список лемм в одном корпусе «покрывается» списком лемм в другом корпусе) получено значение 0,094, для меры «enrichment» (измеряющей долю слов, которые находятся выше порога в 37 вхождений в одном корпусе, но ниже этого порога в другом корпусе) получено значение 0,21. Это значит, что доля перекрытия между списками составляет лишь 9,4% (т. е. список на базе НЧСРЛ имеет низкий охват списка на базе

ruTenTen для слов с абсолютной частотой ≥ 37). Что касается интерпретации результатов значения меры «enrichment», то получается, что список на базе НЧСРЛ предоставляет довольно мало информации о низкочастотных леммах относительно списка ruTenTen11.

Таким образом, становится понятным, что именно следует изменить в методике сравнения данных корпусов для получения списков низкочастотных слов. Во-первых, пороговые значения следует вводить для сводного списка лемм, ранжированного по значениям «ipm mean». Во-вторых, необходимо привлечь к сравнению частотный список лемм, полученный на материале относительно жанрово-сбалансированного корпуса. Такой русский корпус существует – это «Тайга» [Shavrina, Shapovalova 2017]. В-третьих, из списков, полученных на материале веб-корпусов, стоит удалить позиции, содержащие ошибки лемматизации и другой шум. В-четвертых, списки лемм, относящихся к словоформам глаголов, следует рассматривать отдельно. Именно такую методику и предполагается применять в дальнейшем.

Примечания

[1] Исследование проведено при поддержке гранта РНФ, проект № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

Список литературы

Ахутина Т.В. Нейролингвистический анализ лексики, семантики и прагматики. М.: Языки славянской культуры, 2014. 424 с.

Белов С.А. и др. Корпус русских локальных документов и актов CorRIDA: цели формирования, состав, структура / С.А. Белов, О.В. Блинова, В.Б. Гулида, В.И. Зубов, Е.Ю. Ларионова, П.С. Толстикова // Компьютерная лингвистика и вычислительные онтологии. 2018. Вып. 2. С. 112–120.

Ляшевская О.А. Корпусные инструменты в грамматических исследованиях русского языка. М.: Издательский Дом ЯСК: Рукописные памятники Древней Руси, 2016. 520 с.

Ляшевская О.Н. К определению сложности русских текстов // XVII Апрельская международная научная конференция по проблемам развития экономики и общества: в 4 кн. М: Издательский дом НИУ ВШЭ, 2017. Кн. 4. С. 408–418.

Ляшевская О.Н., Шаров С.А. Новый частотный словарь русской лексики, CSV-версия словаря. 2009. [Электронный ресурс]. URL: http://dict.ruslang.ru/freq.php (дата обращения: 12.10.2019).

Baroni M. Distributions in text // Lüdeling A., Kytö M. (eds.). Corpus Linguistics: An International Handbook. Berlin; New York: Walter de Gruyter, 2008. Vol. 2. P. 803–822.

Baroni M. et al. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Webcrawled Corpora / M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta // Language Resources and Evaluation. 2009. Vol. 43(3). P. 209–226.

Gries S.T., Divjak D (eds.). Frequency Effects in Language Learning and Processing. Berlin: De Gruyter Mouton, 2012. Vol. 1. 248 p.

Khokhlova M.V. Large Corpora and Frequency Nouns // Компьютерная лингвистика и интеллектуальные технологии. 2016. Вып. 15(22). С. 237–250.

Kilgarriff A. Comparing Corpora // International Journal of Corpus Linguistics. 2001. N 6(1). P. 97–133.

Kilgarriff A., Rose T. Measures for Corpus Similarity and Homogeneity // Proceedings of the Third Conference on Empirical Methods for Natural Language Processing. Granada, 1998. P. 46–52.

Kilgarriff A. et al. The Sketch Engine: Ten Years On / A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel // Lexicography. 2014. Vol 1, iss. 1. P. 7–36.

Lázníčka M., Janda V. Grammatical Profiling of Czech Nouns: What do Cases Tell Us about Nouns' Meanings // The 15th International Cognitive Linguistics Conference (ICLC-15): Online Book of Abstracts. [Электронный ресурс]. URL: https://iclc 2019.site /wpcontent/uploads/abstracts/corpus/ICLC15_paper_688.pdf (дата обращения: 17.11.2019).

Lindemanna D., San Vicente I. Building Corpus-Based Frequency Lemma Lists // Procedia – Social and Behavioral Sciences. 2015. Vol. 198. P. 266–277.

Piantadosi S.T. Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions // Psychonomic Bulletin & Review. 2014. Vol. 21. P. 1112–1130.

Shavrina T., Shapovalova O. To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser // Proceedings of the International Conference «Corpus Linguistics – 2017». St. Petersburg, 2017. P. 78–84.

RUSSIAN LOW-FREQUENCY WORDS AND APPROACHES TO MODELING GENERAL LANGUAGE FREQUENCY

Olga V. Blinova
Assistant Professor, General Linguistics Department
Saint Petersburg State University

The paper is aimed at developing a methodology for forming lists of low-frequency words by comparing the data from Russian corpora. The paper compares lemma frequency lists from the "Frequency dictionary of modern Russian language" and the frequency data from the ruTenTen11 corpus. Before the comparison, the original lists were converted; as a result, lists with the total length of 51 681 words and 457 935 words respectively were formed. Lists of lemmas with the relative frequencies of <5 ipm according to at least one of the corpora are compared. For the lower cut-off threshold, the absolute frequency value of 37 was selected. The values of the measures "coverage" and "enrichment" are counted. The results show that the "coverage", which reflects the overlap between the lists of lemmas, amounts to only 9,4%.

Keywords: Russian language; linguistic corpora; web corpora; lemma frequency lists; general language frequency; low-frequency words; lexical complexity.