***A. Ch. Piperski***

**Simplicity Beats Sophistication:
An Evaluation of Adjusted Frequency Measures**

**Abstract.** Adjusted frequency measures are required to generalize frequency counts obtained from a corpus to the whole population it represents. However, no systematic evaluation of such measures has ever been made. In this paper, I describe a way of testing whether an adjusted frequency measure is good at predicting the frequency ranking of words in unseen data. 11 adjusted frequency measures are compared using different-sized corpora of eight languages. The results show that Range, one of the simplest adjusted frequency measures, combined with Plain Frequency at the second level of sorting, provides the best ranking. Average Reduced Frequency (*ARF*) outperforms all other measures, except for Range.

**Keywords:** adjusted frequencies, range, Average Reduced Frequency, evaluation, cross-validation.

**1. Introduction**

Compiling a frequency list does not seem to be a complicated task: take a tokenized corpus, calculate how many times each word type occurs, and sort the list in descending order. This approach is unproblematic if the corpus one uses is also one’s object of interest, i.e. if the corpus is equal to the population. However, in most cases we are interested not in the corpus itself, but rather in the population it represents. This means that a frequency list compiled from a corpus is mostly not a frequency list of the population, but a frequency list of the sample that approximates the frequency distribution in the population more or less successfully. One might say that building a frequency list of a language is absolutely impossible and that we can only construct a frequency list for a certain corpus, but this is not what public expects from lexicographers. For instance, Routledge has been publishing frequency dictionaries since 2008, starting with Portuguese [Davies & Preto-Bay 2008]; the series includes frequency dictionaries of 14 languages (American English, Arabic, Czech, Dutch, French, German, Japanese, Korean, Mandarin Chinese, Persian, Portuguese, Russian, Spanish, Turkish). These dictionaries would be of little interest for language learners if they were called “A Frequency Dictionary of the Turkish National Corpus” or “A Frequency Dictionary of Russian Internet Corpus” rather than “A Frequency Dictionary of Turkish” [Aksan et al. 2017], “A Frequency Dictionary of Russian” [Sharoff et al. 2013], etc. Thus, even though it is impossible to compile a frequency dictionary of a language, this is an aim that lexicographers and corpus linguists have always been trying to approach.

It is never an easy question whether a corpus is truly representative of the language as a whole or not, cf. the seminal paper by [Biber 1993]. Even if we assume it to be representative, there is another unavoidable problem: each text in the corpus has a topic, and the words relating to these topics occur in the corpus more often that one would expect in the language in general. Since [Kilgarriff 1997], this issue is known as the *whelk* problem: if a corpus contains a text on whelks, this word is going to be much more frequent in this corpus than in the language. For lexicographic purposes, e.g., in order to compile a frequency dictionary, it is important to know the “true” ranking of the words rather than a corpus-specific ranking, which means that we need to find a way to construct this “true” ranking based on a corpus we have at hand.

For some applications, it might be necessary to find out not only the ranking, but also the “true” frequencies of the words in a language; a similar task has been often addressed in Natural Language Processing where it is important to estimate frequencies of higher-order ngrams even though many of them remain unseen [Jurafsky & Martin 2009]. However, it is outside the scope of this paper to discuss how one can arrive at “true” frequencies, since humans using frequency dictionaries are only interested in rankings rather than exact counts; I am going to focus on adjusted frequency measures that can be used to obtain a robust ranking of words in a frequency list.

**2. Adjusted frequency measures**

The most comprehensive survey of adjusted frequency measures was compiled by [Gries 2008]. This paper was primarily focused on measures of dispersion, especially the new measure DP introduced by the author, but it contains two sections (2.2 and 2.3) on adjusted frequencies.

Adjusted frequency measures are classified into two groups. The first group includes measures that are based on dividing a corpus into equally-sized parts. The simplest adjusted frequency measure, which is also a measure of dispersion, is called Range (*R*). If we divide a corpus into *n* equally-sized parts, *R* is the number of parts that contain at least one instance of the word whose frequency is being computed. Other measures take into account the frequencies of a word in different parts of the corpus (*v*1, *v*2, … , *vn*; their mean is denoted as and their sum is *f*, the total frequency of the word) and mostly rely upon some measure of dispersion, e.g., standard deviation. Here is the list of these measures based on [Gries 2008], but also on the primary sources where some clarification was required [Carroll 1970; Kromer 2003]:

Juilland’s *U*: , where

Rosengren’s *AF*:

Carrol’s *Um*: , where (if *vi*­ = 0, assume )

Engvall’s measure:

Kromer’s *UR*: , where *ψ* is the digamma function, and C ≈ 0.58 is the Euler–Mascheroni constant.

Median (*M*) is another simple measure that can be mentioned here. If the median of the values *v*1, *v*2, … , *vn* is multiplied by *n*, this results in an estimate of the frequency of the word in question.

Among these seven measures, Range and Median have a drawback with respect to ranking. The application of these measures results in numerous ties, since *R* has only *n* possible values, and *M* will very likely be the same for many low-frequency items. If one intends to use these measures for ranking, one can introduce a second level of sorting. Namely, for words with the same *R* or *M*, Plain Frequency *f* serves as a tie breaker. Technically, this can be implemented by adding *f* / *l* to *R* and to *M* for each word, where *l* is the length of the corpus. For instance, if a word with *f* = 650 occurs in 90 out of 100 parts in a corpus that contains 1,000,000 words, its *R* is assumed to be 90.00065 rather than 90. Such a word would be sorted above a word with *f* = 500 that also occurs in 90 parts of the corpus, because its *R* would be 90.0005.

Another group of measures is based on analyzing the distances between the occurrences of a word *w* in a corpus of size *l*. These distances are denoted as *d1*, *d2*, … , *dn*, where *di* is the interval between the (*i* − 1)-th and the *i*-th occurrence of *w* and *d*1 is the distance between the last and the first occurrence under the assumption that the corpus is periodically repeated. This group of measures was introduced by [Savický & Hlaváčová 2002] and includes the following three measures:

Average Reduced Frequency (*ARF*):

Average Waiting Time Frequency (*fAWT*):

Average Logarithmic Distance Frequency (*fALD*):

In spite of the fact that there are so many adjusted frequency measures, a rigorous comparative evaluation of these measures has never been conducted, the most notable exception being a paper by Gries [Gries 2010] that studies correlations between these measures. [Savický & Hlaváčová 2002] compare the stability of their three measures across different corpora and come to a conclusion that *fAWT* is the least stable, whereas the stability of *ARF* and *fALD* depends on how variable the frequency of a word is (*fALD* should be used for words with substantial variation).

**3. Evaluating adjusted frequency measures**

To make an evaluation of frequency measures, we need to find out how well a frequency dictionary compiled from a corpus represents the population from which this corpus was taken. Obviously, we have no access to the population as a whole, but we can check how well the ranking we obtained using some frequency measure on a training set from a corpus describes an unseen sample from the same corpus.

For the experiment, eight corpora of eight languages were taken. They are listed in Table 1.

*Table 1.*Corpora used for the experiment

|  |  |  |
| --- | --- | --- |
| Language | Corpus name | Tokens |
| Arabic | NYUAD Arabic UD | 841,460 |
| Catalan | AnCora  | 533,150 |
| Czech | Prague Dependency Treebank 3.0 | 1,509,236 |
| English | Brown | 1,161,192 |
| German | Hamburg Dependency Treebank | 3,055,010 |
| Icelandic | Icelandic Parsed Historical Corpus | 1,016,527 |
| Russian | SynTagRus | 1,107,741 |
| Spanish | AnCora | 551,456 |

Each corpus was split into five parts in order to perform 5-fold cross-validation. Four parts were used as training set, and the remaining part served as test set. For the training set, 46 frequency lists were compiled using different frequency measures:

1) **Plain frequency**; **Distance-Based measures:** 2) Average Reduced Frequency (*ARF*); 3) Average Waiting Time Frequency (*fAWT*); 4) Average Logarithmic Distance Frequency (*fALD*); **Part-based measures:** 5) Juilland’s *U*; 6) Rosengren’s *AF*; 7) Carroll’s *Um*; 8) Engvall’s measure; 9) Kromer’s *UR*; 10) Median frequency; 11) Range.

The ranking in the resulting frequency lists was compared to the plain frequency list of the test set. To perform the comparison, we take those types that were attested in the training set at least five times, and calculate Spearman’s rank-order correlation coefficient ρ for the 46 frequency lists obtained from the training set against the plain frequency list of the test set.

For example, if we consider Czech words *být*, *v*, *a*, *se*, *na*, *ten*, *že*, *z*, *s*, *který* and use the first four parts of the Czech corpus as training set and the last one as test set, we get the following rankings on our training data:

*ARF*: *být*, *v*, *a*, *se*, *na*, *ten*, *z*, *s*, *který*, *že*;

*fAWT*: *být*, *v*, *a*, *se*, *na*, *s*, *který*, *z*, *ten*, *že*.

The same types in the test set are ordered as follows:

Test: *být*, *v*, *a*, *se*, *na*, *že*, *ten*, *z*, *s*, *který*.

In this case, ρ(*ARF*, Test) = 0.88 and ρ(*fAWT*, Test) = 0.77, which shows that *ARF* is better at predicting the actual ranking of these words in the unseen data that *fAWT*. The actual calculation in this case was based on 15,124 lemmas that occurred in the training set at least five times (out of 51,703 lemma types in total), and *ARF* also outperformed *fAWT*, the two measures scoring 0.745 and 0.736 respectively.

Each of the seven part-based measures was computed using 5, 10, 20, 50, 100, and 200 equally-sized parts. It is worth establishing the best number of parts for each measure and then proceed with comparing 11 rather than 46 measures. For each measure, we find with what number of parts this measure performs better than with any other number of parts (i.e., gets a higher ρ in 21 or more cases out of 8 × 5 = 40). Table 2 illustrates this for Juilland’s *U*.

*Table 2.*Pairwise comparison of Juilland’s *U*
with different number of parts.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of parts | 5 | 10 | 20 | 50 | 100 | 200 |
| 5 |  | 5 | 6 | 10 | 20 | 22 |
| 10 | 35 |  | 15 | 19 | 23 | 26 |
| 20 | 34 | 25 |  | 28 | 30 | 32 |
| 50 | 30 | 21 | 12 |  | 35 | 37 |
| 100 | 20 | 17 | 10 | 5 |  | 35 |
| 200 | 18 | 14 | 8 | 3 | 5 |  |

For instance, this table shows that Juilland’s *U* with 10 parts outperforms Juilland’s *U* with 100 parts 23 times out of 40, but it outperforms Juilland’s *U* with 20 parts only 15 times out of 40. The table makes clear that the best number of parts for Juilland’s *U* is 20, because all five numbers in the corresponding row are greater than 20.

The optimal numbers of parts for part-based measures are as follows:

Juilland’s U 20 Engvall 50 Range 100

Rosengren’s AF 50 Kromer’s UR 50

Carroll’s Um 50 Median 5

Further comparison between the 11 measures follows the same procedure. The results are shown in Table 4:

*Table 4.* Pairwise comparison of 11 frequency measures.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rank | Measure | Range-100 | *ARF* | Kromer’s *UR*-50 | Rosengren’s *AF*-50 | Juilland’s *U*-20 | Carroll’s *Um*-50 | *fALD* | Engvall-50 | *fAWT* | Median-5 | Plain |
| 1 | Range-100 | 0 | 23 | 26 | 29 | 29 | 29 | 34 | 29 | 37 | 38 | 40 |
| 2 | *ARF* | 17 | 0 | 25 | 27 | 30 | 29 | 31 | 27 | 38 | 40 | 40 |
| 3 | Kromer’s *UR*-50 | 14 | 15 | 0 | 34 | 34 | 35 | 25 | 37 | 36 | 37 | 40 |
| 4 | Rosengren’s *AF*-50 | 11 | 13 | 6 | 0 | 24 | 29 | 22 | 36 | 35 | 37 | 40 |
| 5 | Juilland’s *U*-20 | 11 | 10 | 6 | 16 | 0 | 21 | 23 | 22 | 36 | 37 | 40 |
| 6 | Carroll’s *Um*-50 | 11 | 11 | 5 | 11 | 19 | 0 | 21 | 22 | 34 | 37 | 40 |
| 7–8 | *fALD* | 6 | 9 | 15 | 18 | 17 | 19 | 0 | 20 | 38 | 33 | 38 |
| 7–8 | Engvall-50 | 11 | 13 | 3 | 4 | 18 | 18 | 20 | 0 | 34 | 37 | 40 |
| 9–10 | *fAWT* | 3 | 2 | 4 | 5 | 4 | 6 | 2 | 6 | 0 | 20 | 33 |
| 9–10 | Median-5 | 2 | 0 | 3 | 3 | 3 | 3 | 7 | 3 | 20 | 0 | 38 |
| 11 | Plain | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 7 | 2 | 0 |

**4. Discussion**

The results presented in Table 4 are to some extent surprising. Many years of research into adjusted frequencies have given rise to many sophisticated adjusted frequency measures, but the best measure turns out to be a very simple one, namely Range with Plain Frequency used as tie breaker. The superiority of *ARF* to other distance-based measures was already hinted at by [Savický and Hlaváčová 2002] in the paper where these measures were introduced, and it is confirmed by our experiment. A high place occupied by Kromer’s *UR* is also worth noting. [Gries 2008] criticizes this measure and says that Kromer’s claim [Kromer 2003] to its psycholinguistic appropriateness is not corroborated by any evidence. However, the fact that this measure is so good at predicting unseen data speaks in its favor. The criticisms against Juilland’s *D* as a dispersion measure [Biber et al. 2016] are supported by the fact that Juilland’s *U*, which is based on *D*, does not turn out to be among the best adjusted frequency measures.

Obviously, Range has some drawbacks that prevent one from recommending it as an ultimate adjusted frequency measure. Our experiment shows that it is good for ranking purposes, but it is not the best measure if we need to estimate not only the ranking, but also frequencies, e.g. for keyword extraction. *ARF* is much better suited to this purpose, and it is probably not surprising that this measure is used in the Czech National Corpus as well as in SketchEngine rather than Range. The inferiority of other measures is not as dramatic as it may seem; in fact, even using Plain Frequency results in reasonably good frequency lists.

There are also some limitations of the experiment that I must address. First, it is not clear how the performance of adjusted frequency measures relates to corpus size; this issue requires further investigation. Second, the experiment does not tell anything about adjusted frequency measures in case where a training corpus is divided into different-sized categories. Third, the order of the texts in a corpus may be of critical importance for adjusted frequency measures, and this issue was not touched upon. However, these concerns do not undermine the validity of the experiment as a whole. One may conclude that Range combined with Plain Frequency as the next level of sorting provides the best ranking, which makes it questionable whether other part-based distance measures are actually needed for lexicographic purposes. As for distance-based measures, *ARF* fares almost as well as Range, and it is advisable to use this measure in cases where not only rankings, but also frequency estimates are required.

**References**

1. Aksan, Y., Aksan, M., Mersinli, Ü., & Demirhan, U. U. (2017). A frequency dictionary of Turkish: Core vocabulary for learners. Routledge, Taylor & Francis Group.

2. Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

3. Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland’s D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, *21*(4), 439–464. https://doi.org/10.1075/ijcl.21.4.01bib

4. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O’Reilly Media, Inc.

5. Carroll, J. B. (1970). An Alternative to Juilland’s Usage Coefficient for Lexical Frequencies. *ETS Research Bulletin Series*, *1970*(2), i–15. https://doi.org/10.1002/j.2333-8504.1970.tb00778.x

6. Davies, M., & Preto-Bay, A. M. R. (2008). A frequency dictionary of Portuguese: Core vocabulary for learners. Routledge.

7. Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

8. Gries, S. Th. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In S. Th. Gries, S. Wulff, & M. Davies (Eds.), Corpus-linguistic applications (pp. 197–212). Brill. https://doi.org/10.1163/9789042028012\_014

9. Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd ed). Pearson Prentice Hall.

10. Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*(2), 135–155. https://doi.org/10.1093/ijl/10.2.135

11. Kromer, V. (2003). A usage measure based on psychophysical relations. *Journal of Quantitative Linguistics*, *10*(2), 177–186. https://doi.org/10.1076/jqul.10.2.177.16718

12. Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

13. Sharoff, S., Umanskaya, E., & Wilson, J. (2013). A frequency dictionary of Russian: Core vocabulary for learners. Routledge.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Alexander Piperski**

Russian State University for the Humanities / HSE University (Russia).

***E-mail: apiperski@gmail.com***