Frequency of Low-Frequency Words in Text Corpora

Pavel Rychlý

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pary@fi.muni.cz

Abstract. Low-frequency words, esp. words occurring only once in a text corpus, are very popular in text analysis. Also many lexicographers draw attention to such words. This paper lists a detailed statistical analysis of low-frequency words. The results provides important information for many practical applications, including lexicography and language modeling.

1 Introduction

Text Corpora play a crucial role in current linguistics. They provide empirical data for studies on how a language is used. Almost any analysis of a text corpus uses some form of statistics and raw frequency of words is the most common.

Correct handling of the most frequent words is very important in many applications. Other applications ignore most frequent words, usually listed in a stop-list, and pay more attention to mid-frequent or low-frequent words.

There are linguistics studies where the key role of a text analysis have words occurring only once in the whole text (corpus). Such words are called *hapax legomena*. The related terms *dis legomenon*, *tris legomenon*, and *tetrakis legomenon* refer respectively to double, triple, or quadruple occurrences, but are far less commonly used.

This paper quantifies how important very low frequent words could be. It also discusses frequency distribution of zero-frequency words.

2 Corpora

To test frequency distributions of different phenomena we have build a set of corpora, each corpus with different size. Corpora was created by repeated random selection of a text unit from a master Corpus. The Manatee system [1] was used for the random selection of corpus parts and creation of subcorpora. The results presented in this paper use the British National Corpus (BNC) [2] as the master corpus.

We have tested the following units, each corresponds to a XML tag in the BNC source data: sentences (`<s>`), paragraphs (`<p>`), texts (`<text>` – this one covers only written part of the corpus, one document could be divided into several texts), and documents (`<bncdoc>`).

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 53–57, 2010.

Choosing smaller units for sampling results in more random corpus. On the other hand, bigger units creates corpora containing more natural texts. The selection of unit size make a big difference in frequency distributions of low-frequency words. Figures 1 and 2 shows percentages of hapax legomena in

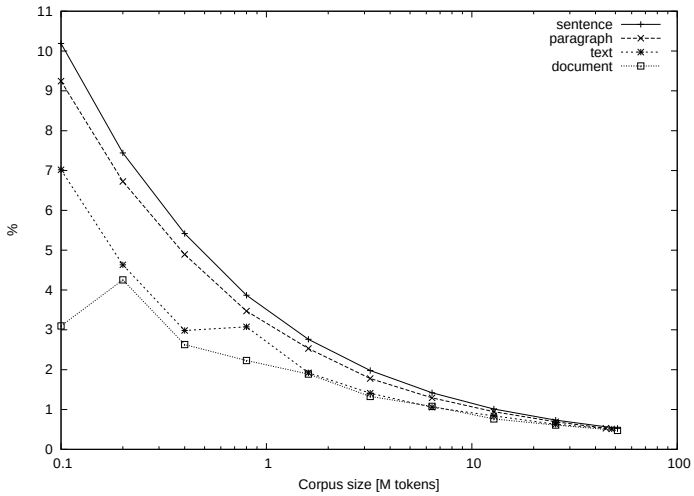

Fig. 1. Percentage of text covered by hapax legomena. Comparison of four sampling units

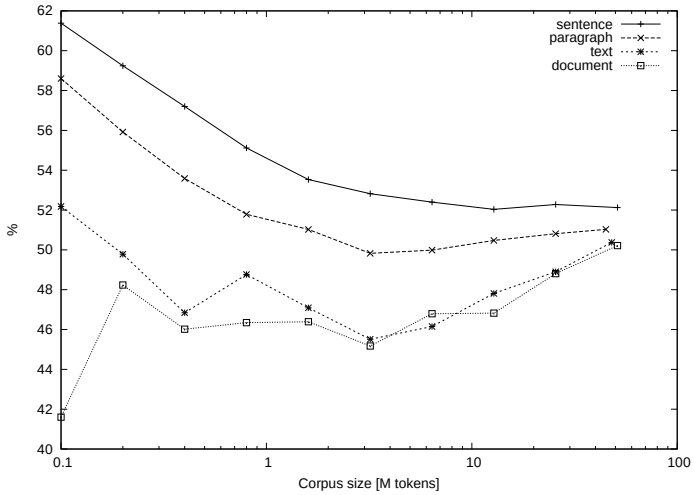

Fig. 2. Percentage of word types (size of the corpus lexicon) of hapax legomena. Comparison of four sampling units

corpora of different size and using different sampling units. First is based on token numbers, second on word types.

The size of a sampling unit is not important for high-frequency words, Figure 3 displays percentage of text which is covered by words occurring exactly 10 times in the corpus. We can see that even small corpora (200,000 tokens) have no difference in coverage for different sampling units.

3 Low-Frequency Words in Corpora

The important question about low-frequency words is how frequent are hapax legomena from a corpus measured in the whole language. We will simulate "the whole language" by much bigger corpus. For this purpose we have chosen sequence of 3 samples of the BNC, starting with 1 million tokens and following with 10 million and 100 million tokens.

The Figures 7, 6 and 8 show frequency distribution of words with selected fix frequency (hapax or tris legomena) in ten times bigger corpus. In all the graphs x-axis lists all different frequencies in which any selected word occurs in the bigger corpus. The y-axis then measures how many of selected words have such feature. For example, a point at $[x = 10, y = 15]$ means, there are 15 different words which occurs exactly 10 times in the bigger corpus. Words are selected from the smaller corpus and occurs there exactly ones or three times (depending on the graph).

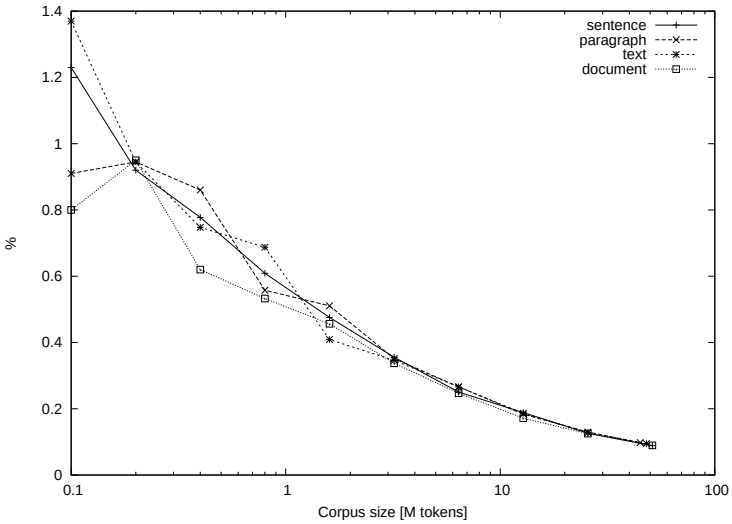

Fig. 3. Percentage of text covered by words occurring exactly 10 times in the corpus. Comparison of four sampling units

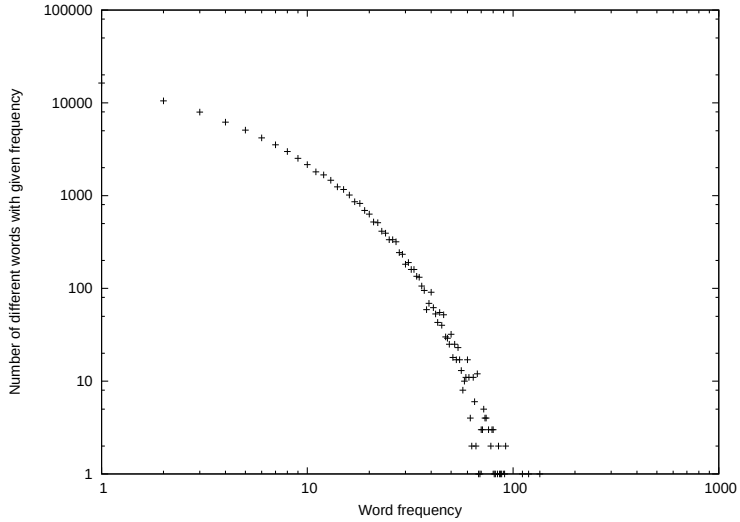

Fig. 4. Frequency of hapax legomena form 10M corpus in 100M corpus. Sentence as an sampling unit

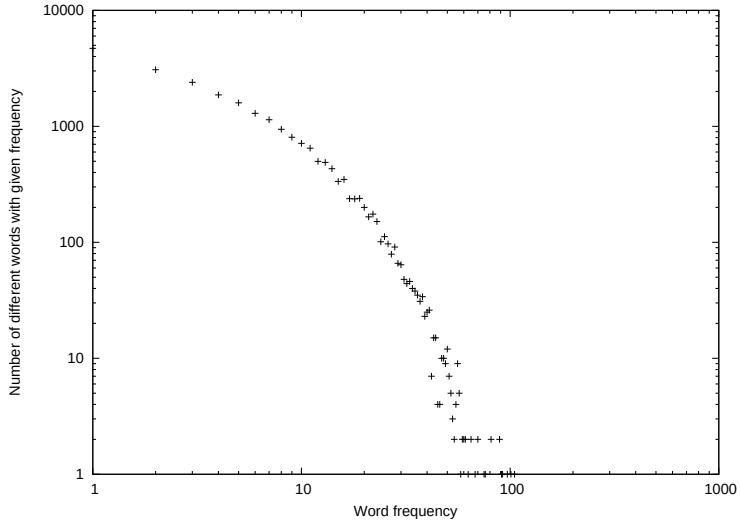

Fig. 5. Frequency of hapax legomena form 1M corpus in 10M corpus. Sentence as an sampling unit

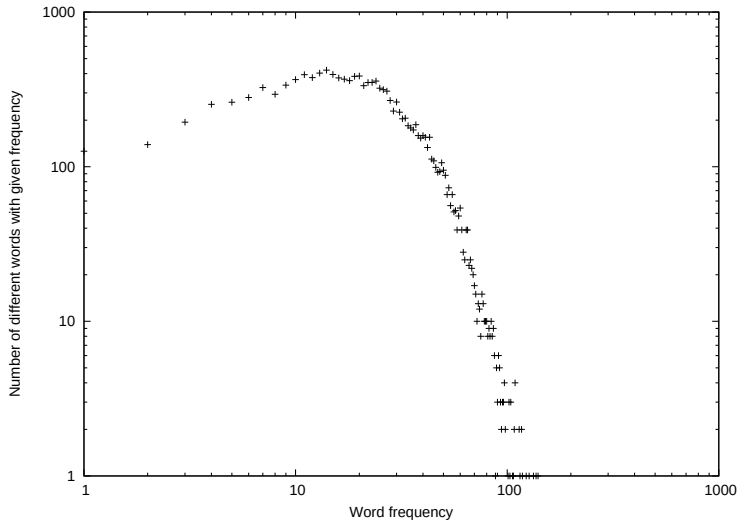

Fig. 6. Frequency of tris legomena form 10M corpus in 100M corpus. Sentence as an sampling unit

4 Conclusion

Even though Low-frequency words form a big part of a corpus lexicon (hapax legomena about 50%), they are not significant in the text. Especially big corpora provide enough data for all relevant words.

Sampling unit have big impact on the frequency distribution of low frequency words. Selection of only sentences or paragraphs instead of whole texts or documents can destroy frequency distribution of words, esp. low-frequency ones.

Acknowledgements

This work has been partly supported by the Ministry of Education, Youth and Sports of Czech Republic under the project LC536 and within the National Research Programme II project 2C06009, and by the Czech Grant Agency under the projects P401/10/0792 and 407/07/0679.

References

1. Rychlý, P.: Manatee/Bonito–A Modular Corpus Manager. RASLAN 2007 Recent Advances in Slavonic Natural Language Processing (2007).
2. Aston, G., Burnard, L.: The BNC handbook: exploring the British National Corpus with SARA. Edinburgh Univ Pr (1998).

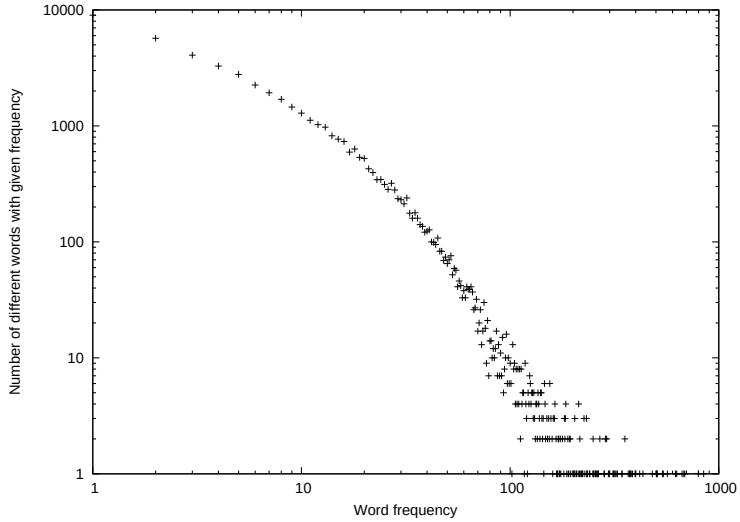

Fig. 7. Frequency of tris legomena form 10M corpus in 100M corpus. Document as an sampling unit

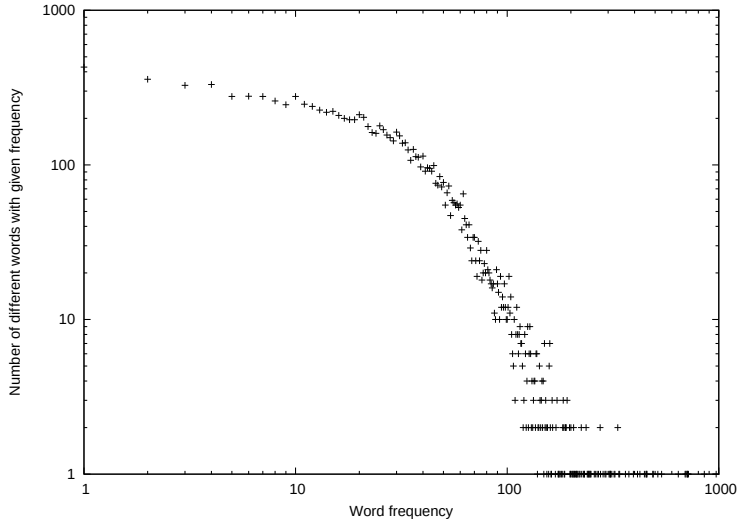

Fig. 8. Frequency of hapax legomena form 10M corpus in 100M corpus. Document as an sampling unit