New Approach to Frequency Dictionaries — Czech Example

Jaroslava Hlaváčová*

*Institute of Formal and Applied Linguistics
Charles University
Malostranské nám. 25, 118 00 Prague, Czech Republic
hlava@ufal.mff.cuni.cz

Abstract

On the example of the recent edition of the Frequency Dictionary of Czech we describe and explain some new general principles that should be followed for getting better results for practical uses of frequency dictionaries. It is mainly adopting average reduced frequency instead of absolute frequency for ordering items. The formula for calculation of the average reduced frequency is presented in the contribution together with a brief explanation, including examples clarifying the difference between the measures. Then, the Frequency Dictionary of Czech and its parts are described.

1. Introduction

Frequency dictionaries are very popular for two main reasons:

- theoretical: they bring interesting insight into the vocabulary of the language — the core from one side (the most frequent words) and the periphery from the other one;
- practical: they can be directly used in practice, especially for selecting entries into new monolingual or bilingual dictionaries of various natures.

Frequency is one of the most popular characteristics of words. It is often the main criterion for lexicographers, who are deciding whether include a word into a dictionary, or not. The frequency of a word is easy to calculate — it is number of its occurrences in a text. If we compare frequencies of all words from a given text, we immediately see, which words are common. However, that result concerns only the text, we have used for our calculations, not the language as a whole. The frequency as a number of word occurrences depends on the text. Not only on its length, but on its subject, its author(s), style and other properties.
We can take a text collection containing various styles, authors and genres. If we could make the collection out of all existing texts, written as well as spoken, then we could calculate the real frequency of all words in the language. Such a task is of course impossible, we have to manage with a sample of texts — a language corpus.

2. The Corpus and its Treatment

The bigger the corpus, the more reliable facts about the language we can infer from it. However, the corpus size is not the only characteristics affecting the results. It depends also on the composition of the corpus, on proportions of its individual constituents. If we, for instance, included only fiction into the corpus, we would probably not get special terms, not even the most common ones. On the other hand, including only technical reports or newspapers, the number (frequency) of more common words would become askew. In other words, we need a representative corpus formed by a great variety of texts in order to cover the major part of language phenomena — for our purpose especially lexical ones.
The basis of our dictionary was the Czech National Corpus – its version SYN2000 [1]. Not only it is quite large (100 million word forms), but it includes the wide spectrum of different texts. They can be grouped into three main categories [2]:

1. fiction – 15%;
2. expert texts – 25%;
3. newspaper articles – 60%

The corpus SYN2000 is automatically lemmatized and morphologically tagged. Every word form from the corpus was assigned with a unique basic form — lemma, and an appropriate morphological tag (Hajič, 2004; Hajič and Hladká, 1998). For the frequencies, we worked with lemmas, not with word forms. It would be interesting to calculate frequency dictionary of word forms, too, but the frequency dictionary of lemmas will certainly have more practical applications.

2.1. Error Handling

The original corpus SYN2000 contains errors of several types, especially:

1. misspellings and typos;
2. morphological and disambiguation errors.

One possibility was to ignore them all and calculate the frequencies only automatically. This would be very straightforward and would not demand any human intervention but some numbers would be inaccurate. That's why we decided to make some "manual" corrections, according to the type of the error.

[1] http://ucnk.ff.cuni.cz/

[2] All the categories are further divided into more subtle subcategories, but they were not taken into consideration for the frequency dictionary, because their number — several tens — is not appropriate for our purpose.

2.1.1. Misspellings and Typos

This type of errors is very hard to discover. We would need to use a spellchecker, but it could not be done automatically, because there are quite a lot of words not included in any spellchecker dictionary, and still correct. A human would have to supervise the spellchecker, but there are approximately 2.5% or unrecognized words in the corpus, which is too many for any manual work. However, unrecognized words are mainly foreign names; number of spelling errors is not very high and does not affect the frequency results seriously, so we let them untouched.

2.1.2. Morphological and Disambiguation Errors

There are some words in Czech with two or more possible spellings. Typical doublets are *citron* / *citrón*, *komunizmus* / *komunismus*. Moreover, we find in the corpus even incorrect spellings (*dýchat* / *dejchat* – English *to breathe*) and still want to recognize them under the same lemma as the correct one(s).

It was necessary to go through all these possibilities manually and embrace them under the same lemma. In fact, as a side effect, this brought some hints for improving basic morphological dictionary of Czech.

Disambiguation errors were more serious. Czech has a lot of homonymous word forms that needed to be disambiguated. The disambiguation was made statistically, which naturally was not errorless. That's why we decided to check all the homonymous forms. We got them from the morphological dictionary. It was not possible to check and correct them manually, because some homonymous forms are very frequent; for instance the word form *bez* can be either preposition (*without*) or nominative / accusative of noun (a bush - *black elder*) and their common frequency in the corpus is 85,541. We checked manually only random sample of 200 occurrences of every homonymous form and counted the ratio of the possibilities. If the ratio was less than 5%, the smaller alternative was not taken into account and all the occurences of the word form were assigned to the more frequent lemma. If the ratio was higher, we added its numeric value as a note to the dictionary entry to warn the user, that the calculated frequencies could be affected by the homonymy. We know that the results still are not correct, but it is probably better than without the manual changes.

We will not go into more details about this subject, because it is not entirely language independent. Only the languages with similar degree of homonymy could take any advantage from it. The detailed description of all the manual processing is in (Čermák and Křen, 2005).

The final corrected version of the corpus SYN2000 became available for all users, so that they could use both the Dictionary and the Corpus as compatible data for their own research.

3. About Frequencies

Having the corpus, we can easily count frequencies of all its words. If it is representative and big enough, we can trust the results more but it will never overcome unevenness of word distribution. There are always texts with unusual accumulation of a special word. Then, that word gets much higher frequency in the corpus, than would correspond with its frequency in the language. There are always texts with unusual concentration of a special word (a hero of a novel, a newly discovered species in an article of a popular journal, a name of an unknown village where something important took place, ...). Lexicographers wanting to select entries into their (never unlimited) dictionaries know the problem very well. Especially towards the lower frequencies, the order has to be manually corrected. It always happens that some special words have in the corpus higher frequency than is their frequency in the language, and the lexicographers have to count upon their individual language experience and intuition (, which, moreover, is never the same for more persons). In fact, they would need commonness of words rather than their frequency.

It was the reason, why we used for our dictionary not the (absolute) frequency as the primary criterion, but the **average reduced frequency** (Savický and Hlaváčová, 2002). It points out those words that occur in (a few) clusters in the corpus. The average reduced frequency (ARF) of such words is much smaller than the ARF of words with the same frequency but with even distribution in the corpus. In this way we can better approach the concept of word commonness.

3.1. Average Reduced Frequency

We will present here the principles of ARF only briefly. The detailed description of its derivation can be found in (Savický and Hlaváčová, 2002).

The corpus consists of so called positions. Every position is occupied by one and only one word. We number the corpus positions with numbers 1 to N. Thus, N is the lenght of the whole corpus.

Let us have a word with frequency f in our corpus. We will split the whole corpus into f segments of the same lenght N/f (for simplicity we can at the beginning suppose that N is divisible by f). If our word was spread evenly in the corpus, every segment would contain one and only one occurence of the word. Usually, the situation is different; some segments contain more than one occurence, others contain none. The number of segments occupied by at least one occurence, will be called **reduced frequency**.

The reduced frequency has one bad property. Its value for a word occuring in a small cluster is either 1 or 2, depending on the position of the cluster within the corpus. If the whole of the cluster is situated inside a segment, the reduced frequency would be one, if the border between segments falls in the middle of the cluster, the reduced frequency of the word is 2. To avoid this imperfection and make the measure more objective, we calculate **average reduced frequency**, as the arithmetic mean over all possible beginnings of the first segment. For this purpuse we imagine the corpus not as a line segment, but as a circle. After the last corpus position, the first one comes. Then, we can move the beginning of the first segment along the whole circle and count reduced frequencies for every its position.

The average reduced frequency is calculated according to the following formula:

$$ARF = \frac{1}{v}\sum_{i=1}^{f}\min\{d_i, v\}$$

where $v = N/f$ and d_i designate the distance between two following occurences of the word in the corpus. Particularly, if n_1,n_2,...,n_f are numbers of positions, where the word occurs, then $d_i = n_i - n_{i-1}$ for every $i = 2, \ldots, f$ and $d_1 = n_1 + (N - n_f)$, which is the distance between the last and the first occurence of the word in the cyclic order of the corpus described above.

3.1.1. Properties of the ARF

Though there is word "frequency" in the name of the measure, average reduced frequency can have (and usually has) non–integer value. ARF has value from the interval $< 1, f >$.

Only the words with absolute frequency 1 have the lowest possible $ARF = 1$.

Only the words with entirely evenly distribution within the corpus can reach the highest value of ARF, namely the value of the absolute frequency f. However, only words with absolute frequency 1 reach in reality that value. There was no word with higher frequency in the Czech National Corpus that was distributed entirely evenly in the whole of the corpus.

A word occuring only in one small cluster has its ARF slightly higher than 1. It depends more on the length of the cluster than on its absolute frequency, how great the difference betweent the both frequencies will be. A word occuring in two small clusters has its ARF slightly higher than 2, and similarly for the following small integers. The more evenly distributed word, the less difference between the absolute and the average reduced frequencies.

In (Savický and Hlaváčová, 2002) there are presented three different measures overcoming the drawbacks of the absolute frequency. Besides ARF, there are: AWT — average waiting time:

$$AWT = \frac{1}{2}\left(1 + \frac{1}{N}\sum_{i=1}^{f} d_i^2\right)$$

and ALD — average logarithmic distance:

$$ALD = \frac{1}{N}\sum_{i=1}^{f} d_i \log_{10} d_i.$$

The ARF was chosen for two reasons:

1. this measure became part of the corpus manager Bonito [3] that is mainly exploited by users of CNC;
2. it is the most straightforward from the three measures.

Let us show the difference between the absolute and average reduced frequencies on examples from the dictionary. We will have a look at two words with the same frequency — 223. The first word is *molekulový* (in English *molecular*), the second one *nahromadit* (in English *to accumulate*).

[3] http://nlp.fi.muni.cz/projects/bonito/

The figures 1 and 2 show their distribution within the corpus. The horizontal axis represents the whole of the corpus; its beginning (the first word of the corpus) lays at the leftmost point. The shades of grey distinguish the three basic genres – fiction, expert texts and newspapers. You can see from the pictures that the corpus was designed so that the genres were kept together.

Number of word occurrences are registered on the vertical axis. They do not have the same scale on the both pictures — it is always calculated according to the data read from the corpus. [4]

You can see that the first word (*molecular*) occurs mainly in that part of the corpus, where the expert texts are gathered, while the second one (*to accumulate*) is distributed much more evenly within the whole corpus. If we took into account only the frequencies, we would not be able to see the difference between their commonness in the language.

At the pictures, we can see the ARF for the both examples and compare their values.

4. Description of the Frequency Dictionary of Czech

Having explained the general principles, let us have a look at the dictionary itself (Čermák et al., 2004).

It has two versions — electronic one on a CD, and paper one in a book.

4.1. The Book

The book consists of 5 lists:

1. Frequency Dictionary of Common Words (alphabetically ordered) – 50,000 items
2. Frequency Dictionary of Common Words (ordered according to absolute frequency) – 20,000 items
3. Frequency Dictionary of Common Words (ordered according to average reduced frequency) – 20,000 items
4. Frequency Dictionary of Proper Names (ordered according to average reduced frequency) – 2,000 items
5. Frequency Dictionary of Abbreviations (ordered according to average reduced frequency) – 1,000 items

and 3 appendices:

1. Frequency List of Delimiters
2. Frequency List of Graphemes
3. Lexical Cover of Texts

[4] The pictures were taken from the corpus manager Bonito that was developed by Pavel Rychlý from the Masaryk University in Brno, Czech republic.

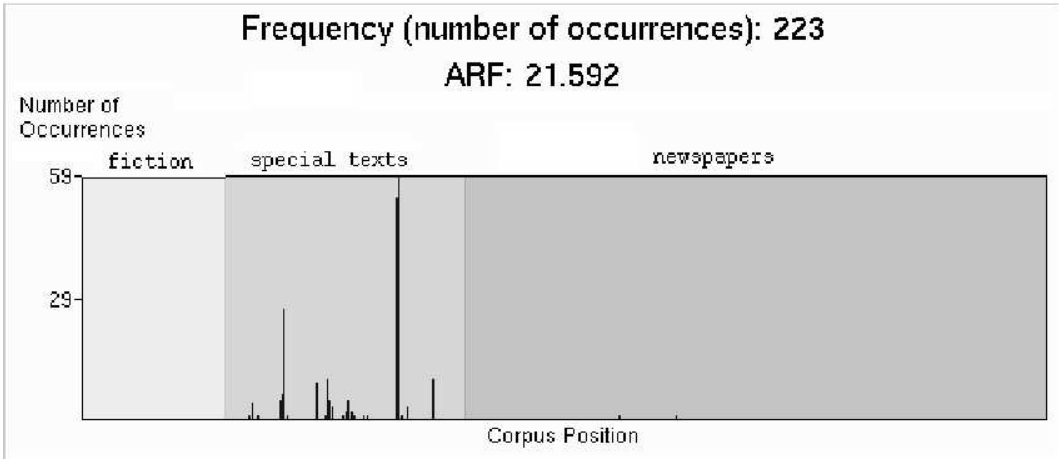

Figure 1: Distribution of all the occurrences of the word *molekulový* in the corpus SYN2000.

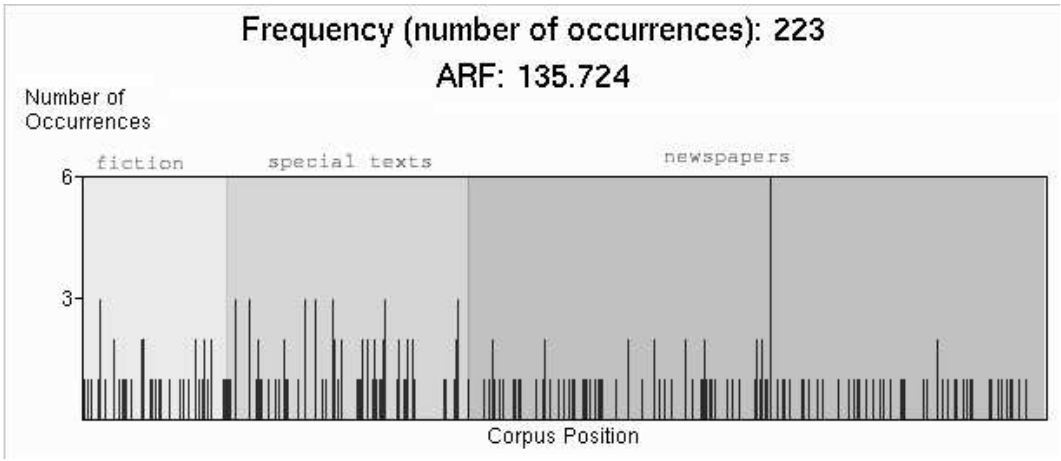

Figure 2: Distribution of all the occurrences of the word *nahromadit* in the corpus SYN2000.

4.1.1. Alphabetic List

The most important and the largest is the first part that includes 50,000 most common Czech words. We will call it in the rest of this contribution the ***central*** part of the dictionary. It does not contain proper names — they were gathered in the special list (number 4 in the above list). This basic list is alphabetically ordered. The words were selected from the corpus according to their average reduced frequency (ARF) — it means that the dictionary contains 50,000 most common words from the corpus SYN2000, (according to ARF). We are convinced that this characteristics is much more appropriate for the words than the (absolute) frequency.

However, for various comparisons, we have put the number expressing the absolute frequency (FRQ) into the list, too.

In fact, neither the frequency, nor the ARF are important. What is important, are the ranks according to the both measures; they are, of course, incorporated into the list as well.

The last thing that was calculated for every entry, is its frequency in the three main text categories listed in the description of the corpus SYN2000. As their proportions in the corpus are not the same, the frequencies were normalized. The normalized frequency expresses the frequency that the word would have, if all the genres were represented equally in the corpus (it means if 1/3 of the corpus was represented by fiction, 1/3 by expert texts and 1/3 by newspapers). For better comparison among different words, the normalized frequencies were converted into ratios (in %).

Let us have a look at the two entries from our previous examples. Though their pure frequencies (FRQ) are the same (223), their reduced frequencies (ARF) differ, and accordingly differ their respective ranks (see table 1). The numbers in the last 3 columns of the table 1 mean, that 99% of all the occurencies of the word *molecular* were found in the section of expert texts, 1% occured in newspapers and 0% in the fiction. The second example, *to accumulate* demonstrates more even representation in the three genres. It corresponds with the figures presented above.

It has the following implication. If somebody wanted to create a small pocket bilingual dictionary of say 20,000 entries, he would probably include the word *to accumulate*, but not the word *molecular*. It can be directly seen from the rank of ARF.

word	English translation	rankARF	ARF	rankFRQ	FRQ	fiction (%)	expert (%)	newspapers (%)
molekulový	molecular	37 502	22	18 959	223	0	99	1
nahromadit	to accumulate	14 970	136	18 915	223	34	43	23

Table 1: Example of two words from the CNC with the same frequency 223.

4.1.2. Frequency Ordering

The next two lists (2 and 3) are ordered according to respective frequencies — FRQ and ARF. For better orientation, the individual items contain not only the respective frequency but also the both frequency ranks. The detailed information about the representation in the three genres is not included, it has to be found in the central part.

The third list — that one ordered according to ARF — is a subset of the central list. The same statement cannot be said about the second list, ordered according to the absolute frequency (FRQ). There are 15 words with the frequency $FRQ < 20,000$, but with $ARF > 50,000$. It is the reason, why those 15 words appear in the smaller second list, but do not appear in the central part of the dictionary. All of them are special terms from various fields (physics, computer science, biology, electrical engineering, ...). They are mainly foreign words, some of them almost do not need translation into English (e.g. *suprematismus*, *repertorium*, *rezistor*, *heparin*). We must admit that these words really do not belong to the 50,000 most common Czech words.

4.1.3. Proper Names and Abbreviations

The frequency dictionaries of proper names and abbreviations are both ordered according to the average reduced frequency, and for every item they contain the same information as the central part, including representation in the three genres. The ARF was especially important in the case of proper names, because in fiction, there is often a hero having a huge absolute frequency in one novel but occurring nowhere else. Despite of the high frequency of such words, they do not belong to the most common Czech proper names.

The list of the most common proper names is not classified into any categories. There are names of persons, towns, countries, companies and other, gathered in one list.

Famous politicians and sportsmen entered the list just because of the great amount of newspapers in the corpus. In fact, this part of the dictionary is mainly a testimony of the corpus' time of origin. The same can be stated about the dictionary of abbreviations. It was the main reason why we did not include them into the central list, as was the traditional praxis for the most of older frequency dictionaries (see for instance the old Czech frequency dictionary (Jelínek et al., 1961)).

4.1.4. Delimiters

Ten most frequent delimiters are presented in the same manner as the central list, ordered according to ARF. From the list, we can for instance infer that delimiters "." and "," are used in all types of texts, while "?" and "!" are much more typical for fiction, brackets for expert texts.

4.1.5. List of Graphemes

Only absolute frequencies were counted for the graphemes. It is interesting to compare the order of graphemes with the similar order calculated 20 years ago (Těšitelová, 1985) on the basis of much smaller corpus (540,000 words). The two orders are similar, but not the same.

4.1.6. Lexical Cover of Texts

This small table proves very famous fact that even small number of the most frequent words cover the majority of texts. Thus, for instance, the first 10,000 most frequent words covers more than 91% of the whole corpus SYN2000.

For the calculation of this table the absolute frequency was used.

4.2. The CD

The CD contains three lists:

1. Common Words – 50,000 items
2. Proper Names – 2,000 items
3. Abbreviations – 1,000 items

The content of each list is identical with its counterpart in the paper book. In addition, the CD is equipped with the browser *EFES* that enables users to handle the data more effectively than it is possible with its paper version. Of course it is possible to reorder the data according to any of the 8 items included in the lists — ARF, FRQ, rank ARF, rank FRQ, relative representation in the three main genres, and alphabetically. It is the reason, why the two smaller lists (list 2 and 3 in the listing of the previous section) are not presented separately on the CD data. However, there is one tiny difference — those 15 words mentioned earlier (with $ARF > 50,000$ and $FRQ < 20,000$) are not present on the CD, because they do not belong to the central list.

The main function of the browser is to enable searching the data. The simplest search is alphabetical. As opposed to the paper version, we can search not only the words from their beginning, but also according to an end substring, or even an inside substring. Thus, we can make for instance our own frequency dictionary of individual suffixes or roots. The only drawback is, that it cannot deal with regular expressions.

We can also search the lists according to other criteria and combine them into more complicated search conditions. It is possible to state the intervals for the individual numeric categories. We can for instance find all the items with $rankARF > 30,000$ and $rankFRQ < 10,000$. In such a way, we discover the words with very uneven distribution within the corpus. We can include into our queries also constraints on the genre representation.

The results can be stored at external media for future uses.

5. Conclusion

We have presented the big project of the Frequency Dictionary of Czech. We have showed its main features that make it different from other similar projects. The decisions are justified in the text of this contribution.

The uniqueness of the dictionary consists in using not absolute, but average reduced frequency for ordering words. It overcomes incidental unevenness of word distribution within the corpus that distorts the credibility of results. We are convinced that using other than absolute frequency makes the frequency dictionary more appropriate for the direct use for compiling other sorts of dictionaries or encyclopedias.

6. Acknoledgements

This work was supported by the grants 1ET101120503 and 1ET101120413 of the Grant Agency of the Academy of Sciences of the Czech Republic.

7. References

J. Hajič and B. Hladká. 1998. Tagging inflective languages. prediction of morphological categories for a rich, structured tagset. In *Proc. ACL-Coling'98*, pages 483–490. Longman, Montreal, Canada.

J. Hajič. 2004. *Disambiguation of Rich Inflection. (Computational Morphology of Czech)*. Praha, Karolinum. ISBN 80-246-0282-2.

J. Jelínek, J.V. Bečka, and M. Těšitelová. 1961. *Frekvence slov, slovních druhøu a tvarøu v českém jazyce*. Státní pedagogické nakladatelství, Praha.

P. Savický and J. Hlaváčová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics*, 9:215–231.

M. Těšitelová. 1985. *Kvantitativní charakteristiky současné češtiny*. Praha, Academia, Praha.

F. Čermák and M. Křen. 2005. New generation corpus-based frequency dictionaries: The case of czech. *International Journal of Corpus Linguistics*, 10:453–467.

F. Čermák et al. 2004. *Frekvenční slovník češtiny*. Praha, NLN. ISBN 80-7106-676-1.