sciendo
DOI 10.2478/jazcas-2025-0027

DO FREQUENCY TYPES MATTER IN LEXICOGRAPHY?

MAREK BLAHUŠ[1] – VOJTĚCH KOVÁŘ[2] – FRANTIŠEK KOVAŘÍK[3]
[1]Lexical Computing CZ, s.r.o., Brno, Czech Republic
(ORCID: 0009-0009-4096-4158)
[2]Lexical Computing CZ, s.r.o., Brno, Czech Republic & Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0005-0307-9046)
[3]Lexical Computing CZ, s.r.o., Brno, Czech Republic & Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0009-0003-8002-8360)

BLAHUŠ, Marek – KOVÁŘ, Vojtěch – KOVAŘÍK, František: Do Frequency Types Matter in Lexicography? Journal of Linguistics, 2025, Vol. 76, No 1, pp. 303 – 311.

Abstract: Word frequency in a corpus can be calculated in several different ways. Amongst the most common frequency types are the absolute frequency, the document frequency, ALDF and ARF. This paper focuses on comparing these four types in terms of "word correctness." For determining whether a word is correct or not, we use the data gathered for the Czech lexicon used for the recent Czech Dictionary Express project. In this project, each of the top 100,000 most frequent headwords was reviewed by several Czech native speakers, who decided whether the word should be accepted or rejected or has some minor issues. The quality of the "word correctness" is further discussed in the paper.

Keywords: corpus annotation, semi-automatic dictionary drafting, Dictionary Express, word frequency, frequency type, absolute frequency, document frequency, ALDF, ARF, Czech

1 INTRODUCTION

Word frequency is a number heavily used in corpus linguistics for statistics. It represents the word count across the corpus. The frequency of a word, a lemma or a token illustrates its distribution, determines the score of a collocation, and constitutes the base for frequency wordlists.

Frequency wordlists are lists of words (lemmas, tokens, etc.) sorted from the most frequent ones down to the least frequently used words (typically with one occurrence).

There are different strategies for counting a word's frequency. This paper revolves around four of the most typically used word frequency types, and examines how differences in word frequency can correlate with the occurrences of typos, non-words, words of a different language than the corpus, non-standard words and incorrectly lemmatized and/or POS-tagged words, as well as the rest – the "correct" words, in the Czech Web (csTenTen12+17+19) corpus (Suchomel 2018). For distinguishing whether a word is "correct" or faces some issues, we use annotation data gathered manually from Czech native speakers in the Czech Dictionary Express project.

Chapter 2 briefly introduces the principle of Dictionary Express projects, the manual annotation of Czech headwords, and the criteria of "correctness" of headwords. Chapter 3 the purpose of frequency types, their differences and their usage. Chapter 4 presents the correlation statistics between higher or lower frequency of each type and the "correctness" rate of headwords of these frequencies.

2 HEADWORD ANNOTATION

2.1 Dictionary Express

Dictionary Express (DE) is a series of dictionary making projects, which focus on rapid semi-automatic dictionary drafting methods (Kovařík et al. 2024). Each DE project concentrates on a different language and divides the dictionary making process into simple tasks such as building the vocabulary, selecting proper word forms for every headword, word sense disambiguation etc. As opposed to the "traditional dictionaries", created one entry at a time, the DE dictionaries are done in stages matching the tasks: the first stage includes going through the whole set of headwords and creating a proper vocabulary; the next stage includes going through the whole vocabulary and choosing the proper forms; etc.

Each stage is prepared automatically, using data from large language corpora (with tens of billions of tokens), preferably lemmatized and POS-tagged. The data is then manually annotated by a team of native speakers without academic education in linguistics, called the *annotators*.

2.2 Annotation

In the first stage, the annotators go through a list of headwords (i.e. pairs of lemma and part of speech), which are automatically lemmatized and POS-tagged by specialized tools. The annotators assign each headword one of these possible "flags":

- *don't know the word* if they do not understand the word;
- *not Czech* if they know of the word but the word isn't part of the Czech language (based on their native speakers' intuition);
- *non-standard* if the word is not part of the standard Czech language (we take Czech *spisovný jazyk* as the standard, although again the annotators' language intuition is determinant);
- *wrong lemma* if the lemma is incorrect (including words with incorrect lemmatization, words in their non-lemma form and words with typos);
- *wrong POS* if the POS is incorrect;
- *ok* if the lemma and the POS are correct;
- *name* if the lemma and the POS are correct and the word is a proper name.

The annotators don't see the context of the words and are not allowed to look up the word in any other dictionary or on the internet.

This way, each headword has got at least two flags from two different annotators.

2.3 Revision

The headwords that were annotated with a variety of flags (i.e. with an insufficient inter-annotator agreement) and the ones whose majority flag was *non-standard*, *wrong lemma* and *wrong POS* had to be revised.

A group of experienced annotators (called "inspectors") went through each of these headwords, and according to the flags previously assigned to them and their corpus context, they decided whether the word is correct or incorrect or should be revised to another lemma or POS.

2.4 "Correctness" criteria

The wordlist for Czech Dictionary Express was created using document frequency. It contained the 100,000 most frequent headwords of the Czech Web corpus. After the revision, each of the headwords was either considered correct (marked *ok* or *name*) or incorrect (marked *don't know the word* or *not Czech* or revised to a correct headword).

3 FREQUENCY TYPES

Word frequency can be counted in a number of ways. This paper examines four of the most commonly used frequency types: absolute frequency, document frequency, ALDF and ARF.

Absolute frequency is the number of occurrences a word has in a corpus (Sketch Engine 2024). For smaller corpora with a specific topic, this can be an effective and simple way to count the words and analyze the vocabulary statistically. Absolute frequency, however, can be easily manipulated if a single word is used a lot of times in a single document or in a narrow area of texts. In other words, it ignores the word burstiness.

Word burstiness is the quality of the distribution of a word, i.e. whether it is used only in a closed area (it "bursts" somewhere) or whether it is spread throughout the corpus (or the language) (Rychlý 2011). Some words can be used many times in only a few documents. Absolute frequency of these words is high, but their distribution over the whole corpus or language use is narrow.

For taking word burstiness into consideration, the lexicographer can use other frequency types, such as document frequency, ALDF and ARF.

Document frequency is the number of documents a word occurs in at least once. This makes the widely distributed words more frequent than the ones that are only used in a few documents.

ALDF, or average logarithm distance frequency, reflects the average distance between the occurrences of the word. For two words with the same absolute frequency, ALDF is lower for the word only used in a small number of texts or text areas (Sketch Engine 2022). ARF, or average reduced frequency, though counted in a different manner, serves a similar goal.

Choosing a proper frequency type that does or does not take word burstiness into account can make a big difference when examining a small area of words or differences between particular words or their usage. But what about bigger tasks, such as choosing words for a mono-lingual dictionary? The next chapter discusses the role of the frequency types in building a dictionary lexicon.

4 FREQUENCY WORDLIST DIFFERENCES

4.1 Relation between word frequency and its "correctness"

As suggested in chapter 3, we consider a word "correct" if most of the annotators agreed it is a standard part of the language or if an inspector revised it to be correct after seeing its previous annotations and its context. We mark the "correctness" with quotations, since this is not a measure of whether a word should or shouldn't be considered a stable and directive part of the language system, but only a consideration based upon the intuition of several native speakers.

The 100,000 most frequent headwords according to the document frequency have been differentiated this way. In Fig. 1, we see how the percentage of "correct" words is related to higher frequency. (For easier calculation, the frequency wordlist of 100,000 headwords has been divided into "percentiles" of 1,000 words. The numbers on the X axis represent these groups. To get the document frequency rank of a headword in a particular area, multiply the number by 1,000.) On the left are the headwords at the top of the frequency wordlist, on the right the words with the frequency rank up to 100,000.

We can see that the more frequent a word is, the more likely it is going to be considered "correct". This relation is very linear, at least for the 100,000 most frequent headwords.

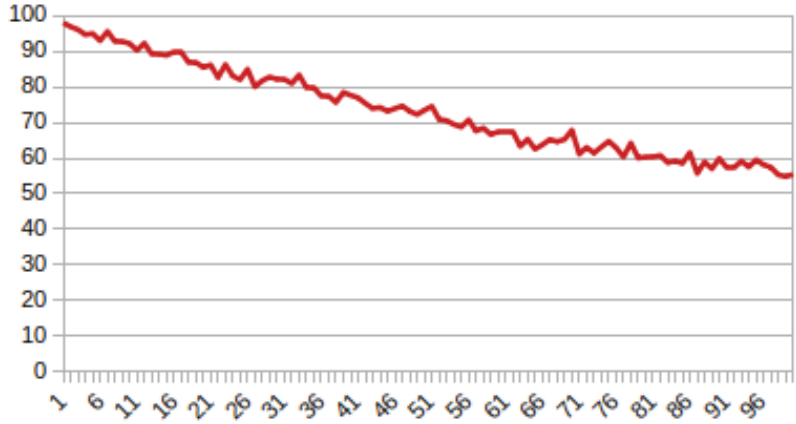

Fig. 1. Relation between the rank in document frequency divided by 1,000 (X axis) and "correctness" percentage (Y axis)

Fig. 2 shows a similar graph, but wordlists of all four frequency types are present now, represented by a separate color. The lines copy a very similar trajectory, except for the right ends of the wordlists. The data of the wordlists other than that of the document frequency are getting more scarce, because only the words from the 100,000 document frequency wordlist have been used, so some of the words from the ends of other wordlists are missing (as explained further, see Tab. 1), and thus more noise can be expected.

This means for the 100,000 most frequent headwords, there aren't many differences between the frequency types considering the "correctness" of the headwords.

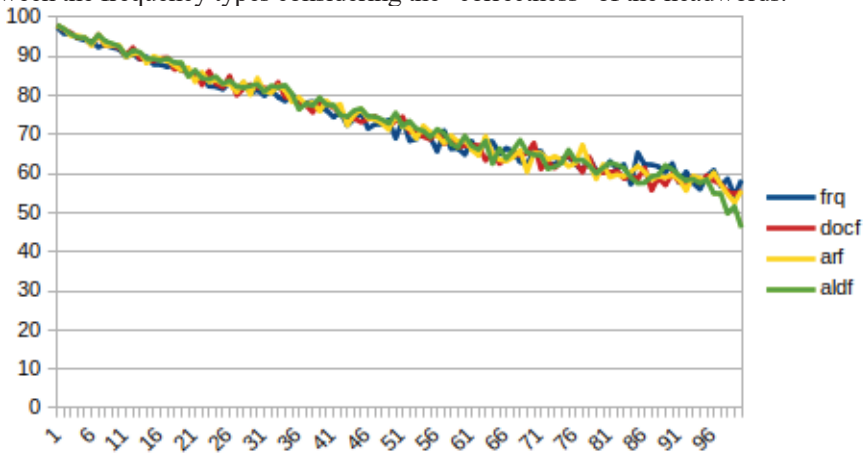

Fig. 2. Relation between the rank in the wordlist of frequency of a given type divided by 1,000 (X axis) and "correctness" percentage (Y axis)

The lexicon of the Czech DE project is based on the document frequency wordlist. Tab. 1 presents the word differences between the 100,000 document frequency wordlist and the wordlists of the other frequency types. Each number represents the number of words that are in the document frequency wordlist and are missing from the wordlist of a particular frequency, and vice versa.

As we can see in Tab. 1, the ALDF and ARF frequency wordlists are more similar to the document frequency wordlists than the one of absolute frequency. This should come as no surprise since both ALDF and ARF as well as the document frequency reflect not only the word count of a headword, but also its burstiness.

	Words missing in Doc. F.
Abs. F.	4962
ARF	1722
ALDF	1927

Tab. 1. Differences in wordlists of document frequency and of other frequency types

Tab. 2 presents the percentage of "correct" headwords within the 10,000, 50,000, 80,000 and approximately 100,000 most frequent headwords based on absolute frequency, document frequency, ARF and ALDF. (Since only the 100,000 most frequent headwords based on document frequency have been annotated, the statistics of headwords from the ends of the 100,000 wordlists of absolute frequency, ALDF and ARF are missing. Only the 95,038 most frequent words from the absolute frequency wordlist, the 98,278 most frequent words from the ARF wordlist and the 98,073 most frequent words from the ALDF wordlist have been decided to be "correct" or "incorrect". The ends of these 100,000 wordlists are still waiting to be properly annotated and revised by the annotators.)

	10,000	50,000	80,000	cca 100,000
Abs. F.	94.08%	83.33%	76.78%	74.07%
Doc. F.	94.65%	83.95%	77.07%	73.28%
ARF	94.70%	84.11%	77.29%	73.82%
ALDF	94.85%	84.44%	77.62%	74.09%

Tab. 2. The percentage of "correct" headwords in different frequency wordlists

We do not see a big difference between absolute frequency and the other types, even though absolute frequency seemed to be different from the other types considering the words of its 100,000 frequency wordlist (Tab. 1).

From the 100,000 document frequency wordlist, 73,278 have been marked "correct" and 6,518 headwords have been added as the result of the correction of "incorrect" headwords in the revision phase. This means that based on the quality of the corpus, the word lemmatization, POS tagging and language factors, a dictionary of 80,000 "correct" headwords needs approximately a 100,000-word wordlist. Considering the curve of the frequency-"correctness" relation in Fig. 1 and Fig. 2 and its predictable continuation, a dictionary of 100,000 "correct" headwords could require some 150,000 words from the frequency wordlists. Although there are differences between the wordlists, as shown in Tab. 1, these do not exceed 5% of the wordlists.

There could be, however, a bigger difference in the less frequent headwords, i.e. the headwords after the rank 100,000 of the document frequency wordlist. This is to be examined in future research focusing on the headwords after the frequency rank 100,000 and whether these headwords show different frequency-"correctness" relations than the more frequent ones.

4.2 Wordlist difference examples

For each wordlist, the words can be separated into 5 categories based on our research:

- *present accepted* are words that are in the 100,000 wordlist of a frequency type and are considered "correct";

- *present rejected* are words that are in the 100,000 wordlist of a frequency type and are considered "incorrect";
- *missing accepted* are words that are not in the 100,000 wordlist of a frequency type and are considered "correct";
- *missing rejected* are words that are not in the 100,000 wordlist of a frequency type and are considered "incorrect";
- and *missing from document frequency* are words that are present in the 100,000 wordlist of a frequency type other that document frequency and are not in the 100,000 document frequency wordlist – these words have not been yet marked "correct" or "incorrect" since only the document frequency wordlist has been annotated and revised, and are subject to further research.

The main subject of quality comparison between the wordlists has become the *present accepted* category, since these are the words a lexicographer would prefer to have in the dictionary yet are not included in some of the wordlists. Most of these are words from the end of the 100,000 most frequently used headwords.

The absolute frequency wordlist contains more company names and web page URLs than the other types, e.g. *Vareni.cz (noun)*, *Echo24 (noun)*, *Skyscanner (noun)*, *Ulož.to (noun)* and *ČSDF.cz (noun)*, whereas it is lacking many less frequent words such as *vypoklonkovat (verb)* – "to bow sb. out", *libující (adjective)* – "relishing", *utuchat (verb)* – "to weaken (literary)", *třímající (adjective)* – "holding (literary)", or *polovičatě (adverb)* – "halfway, poorly". This should come as no surprise, since company names and URLs can be very frequent in a small number of texts (their frequency is high, yet their overall distribution is low) and the common Czech words the absolute frequency wordlist is lacking are distributed more evenly across the whole corpus, although their frequency isn't as high.

As mentioned in 3.1, the absolute frequency wordlist is more different than all the other wordlists, although the "correctness" of its words is similar. The words missing from the other wordlists that are present in the absolute frequency wordlists, however, are not of the same quality as vice versa. In a dictionary, it would be preferable to include the less frequent words which are missing from the absolute frequency wordlist over the company names and web page, i.e. proper names of various origin.

Comparing the ARF and ALDF wordlists, the ARF wordlist does seem to have more company names and web page URLs, include more proper names in general, and also include more words of a foreign origin, such as *crowdfunding (noun)*, *selfíčko (noun)* – "selfie" and *magenergie (noun)* – "mana (fantasy)", whereas the ALDF wordlist has more originally Czech words similar to the ones missing from absolute frequency, e.g. *skotačící (adjective)* – "frolicking", *setrvávající (adjective)* – "remaining (literary)", *usekat (verb)* – "to cut off" or *brždění (noun)* – "braking (e.g. with brakes)".

This leads to the conclusion that the ALDF could be the preferred frequency type for building a dictionary lexicon if choosing from absolute frequency, ALDF

and ARF. However, the research cannot be considered complete until the headwords from the end of ALDF, ARF and absolute frequency wordlist (the ones *missing from document frequency*) are annotated and marked "correct" or "incorrect". After this, conclusions can be made about the differences between all the wordlists, including document frequency, which has been used as the base for the DE lexicon.

5 CONCLUSION

We have examined four frequency wordlists containing the 100,000 most frequent headwords, calculated using absolute frequency, document frequency, ALDF and ARF. We have found some differences between the wordlists which could have a small impact on dictionary drafting and on building a dictionary lexicon.

The annotations, revisions and the quality of "correctness" were only gathered for the 100,000 most frequent headwords of the document frequency wordlist. A complete statistic of "correctness" in the 100,000 wordlists for all four types of frequencies should be a matter of subsequent research. Further research should be also made for the words after the 100,000 ranks and whether these words show different frequency-"correctness" relations than the more frequent words.

From examining the example differences between wordlists of different frequency types, it seems ALDF could be the preferred frequency type for building a Czech dictionary from a large web corpus. However, a vocabulary of good quality could also be achieved combining wordlists of all frequency types, and annotating words from the 100,000 wordlists of all frequency types. Considering the low rate of differences between the frequency wordlists of the 100,000 most frequent words, which do not exceed 5% of the wordlists, this would not make the dictionary making process noticeably more complex or time-consuming.

References

Kovařík, F., Kovář, V., and Blahuš, M. (2024). On Rapid Annotation of Czech Headwords: Analysing the First Tasks of Czech Dictionary Express. Online. In: Kristina Š. Despot – A. Ostroški Anić – I. Brač: Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress. Cavtat: Institut za hrvatski jezik, 2024, pp. 336–344. Accessible at: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf [cit. 27/03/2025].

Rychlý, P. (2011). Words' Burstiness in Language Models. In: A. Horák, P. Rychlý: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011. Brno: Tribun EU, 2011, pp. 131–137. Accessible at: https://nlp.fi.muni.cz/raslan/2011/paper17.pdf [cit. 27/03/2025].

Sketch Engine. ALDF – Average Logarithm Distance Frequency. Online. Sketch Engine. 2022, [28/02/2023]. Accessible at: https://www.sketchengine.eu/aldf-average-logarithmic-distance-frequency/ [cit. 27/03/2025].

Sketch Engine. Frequency. Online. Sketch Engine. 2024, [12/11/2024]. Accessible at: https://www.sketchengine.eu/glossary/frequency/ [cit. 27/03/2025].

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2018, pp. 111–123. Accessible at: https://www.sketchengine.eu/wp-content/uploads/cstenten17.pdf [cit. 27/03/2025].