Michal Škrabal/Michaela Lišková/Martin Šemelík

ON DEFINING VOCABULARY IN A MONOLINGUAL ONLINE DICTIONARY

Some remarks from the lexicographical practice on the *Academic Dictionary of Contemporary Czech*

Keywords Defining vocabulary; definition; metalanguage; hapax legomenon; Academic Dictionary of Contemporary Czech

1. Theoretical background

Our study focuses on defining vocabulary in the context of monolingual lexicography with the *Academic Dictionary of Contemporary Czech* (ADCC) as the main subject of interest. It is anchored in corpus linguistic research based on the Czech National Corpus (Charles University) and lexicographical practice at the Czech Language Institute (Czech Academy of Sciences). We aim to demonstrate how even a simple statistic can improve dictionary definitions from the user perspective and to offer some recommendations for the authors' future work.

The ADCC is an alphabetical, monolingual general-purpose dictionary. In every monolingual general-purpose dictionary, the meaning description of lexical units plays a crucial role. Consequently, "a systematically selected range of words to be used for describing the content of a larger number of words" (Svensén 2009, p. 246), the defining vocabulary, poses a relevant research topic, especially in connection with the user aspect.

Although there is not a strictly predefined metalanguage for the meaning description for the ADCC, it can be said, however, that (a) the defining vocabulary consists of lexemes, which are included in the ADCC main register and that (b) within certain lexical-semantic classes, standardised basic "pillar" words for the expression of the *genus proximum* are determined (cf. Kochová/Opavská 2016, p. 88). Lexicographers understand and comply with the basic rule that metalanguage should not be too complicated. Apart from distinctive meaning features, some other elements are considered a part of lexical meaning in the ADCC. Among those, we find non-distinctive, facultative features that reflect a complex of knowledge that language users have at the level of common knowledge about denominated non-linguistic facts.

On the contrary, in the Anglo-Saxon tradition, the defining vocabulary is a frequent feature of many dictionaries, especially those that can be termed as learners' dictionaries (Kamiński 2021 for English; Töpel 2021 for German). These have defined their defining vocabularies "to ensure that the definitions are clear and easy to understand and that words used in explanations are easier than the words being defined" (LDoCe p. B17; cf. Xu 2012). Within the Czech lexicography, we have noted only one exception so far (significantly under an English influence): Sinclair's et al. (1998) *English-Czech Explanatory Dictionary*, which is a bilingualised dictionary based on the original work *Collins COBUILD Student's Dictionary* (1990), provided with a COBUILD Word List that consists of words which appear in mean-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ing descriptions at least ten times. In this list, there are 1860 lemmas, respectively 2591 words. (Sinclair et al. 1998, pp. IX, 1162).

Within our experiment, we compare entries with four initial letters, i.e. A–Č, in the ADCC. After publication, A-entries were criticised for being too encyclopaedic, but this could also be due to the prevalence of words of foreign origin, often terms. Conversely, Č-entries are mostly of domestic origin. Besides, the ADCC's conception has changed in the meanwhile. These changes included, among others, the following:

a) On the basis of stricter inclusion rules, the terms are given less prominence as compared to previous practice.

b) With regard to the user aspect, we avoid cognitively overloaded definitions. Definitions undergo a gradual process of "de-encyclopedisation".

2. Quantitative analysis of the ADCC's metalanguage

We used the following procedure for our analysis:

1) In the dictionary editorial system, we exported the definitions of all currently published entries from the Definitions field (omitting synonyms, for which there is another column).

We lemmatised the individual text files and performed simple frequency statistics, which resulted in Table 1.

Initial letter	Tokens	Types	Type-token ratio	H-point	Hapaxes	Hapax-token ratio	Number of entries
A	39,822	7,202	0.181	63	3,490	0.088	2,897
B	56,832	9,546	0.168	72	4,458	0.078	3,806
C	17,702	4,434	0.25	40	2,342	0.132	1,308
Č	20,004	4,304	0.215	46	2,156	0.108	1,236
A–Č	134,360	16,063	0.12	122	6,967	0.052	9,247

Table 1: Frequency statistics of the ADCC's metalanguage (entries A–Č)

Reviewers' comments and instructions given in the reviews should be taken into account.

2) We compared individual groups with a focus on prominent content words that form the pillar of every definition) and, at the same time, on *hapax legomena* that are most prone to be eliminated from ADCC's defining vocabulary.

3) On the basis of qualitative analysis, as well as a comparison with the defining vocabularies from English dictionaries, we aim to make recommendations for authors' future work. As far as hapaxes are concerned, these can be as follows:

   (a) A word can be deleted with no substitute:

   borka I vrchní odumřelá *zkorkovatělá* vrstva kůry kmene dřevin ('the upper dead, *corky* layer of bark of a tree trunk')

   The word *zkorkovatělá* is not only rare and terminological but also redundant in the sense that even without this particular word the meaning delimitation is functional.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(b) A word may be substituted for another, more common word**:**

blaťácké zlato měkký sýr s pružnou konzistencí, *zlatooranžově* zbarveným povrchem a hořkomandlovou, nakyslou chutí (‘a soft cheese with an elastic consistency, a *golden-orange* surface and a bitter, sourish taste’)

The word *zlatooranžově* is rare (*zlatooranžov.** 88 hits (ipm 0.01) in SYN v10). Its more frequent synonym *žlutooranžově* (‘yellow-orange’, *žlutooranžov.** 1,374 hits (ipm 0.23) in SYN v10) is, as additional analyses have revealed, factually more appropriate.

(c) A definition needs to be re-formulated:

brukev 2. rostlina (odrůda brukve zelné) s listy na dlouhých řapících a *ztlustlým* stonkem, pěstovaná jako zelenina; syn. kedlubna 1 (‘a plant (a variety of kohlrabi) with leaves on long leafstalks and a *thickened* stem, grown as a vegetable; syn. turnip cabbage 1’)

The word *ztlustlý* is not common (*ztlustl.** 337 hits (ipm 0.06) in SYN v10). Additional evaluation of the definition resulted in the conclusion that the word *bulva/hlíza* (‘tuber’) should be included in the definition leading to post-editional measures taken in this respect.

3. Conclusion

Our study shows how a simple frequency statistic of defining vocabulary can improve lexicographic definitions and serve the user-friendliness of a dictionary. One can also ask more general questions, e. g.: Where do the words in the defining vocabulary actually come from? Do they necessarily overlap with the “core general vocabulary” (Brezina/Gablasová 2015) of a given language? In fact, such a vocabulary already exists for Czech (Čermák/Křen 2011), and besides its pedagogical use, its application to the field of lexicography is logically suggested. The result of this case study will be considered in the further modification of the ADCC’s conception.

References

ADCC (2017–2020): Akademický slovník současné češtiny. A–Č. Prague. http://slovnikcestiny.cz/uvod.php (last access: 24-03-2022).

Brezina, V./Gablasova, D. (2015): Is there a core general vocabulary? Introducing the New General Service List. Applied Linguistics 36 (1), pp. 1–22.

Collins COBUILD student’s dictionary (1990). London.

Čermák, F./Křen, M. (2011): Frekvenční slovník češtiny. Prague.

Kamiński, M. P. (2021): Defining with simple vocabulary in English dictionaries. Amsterdam.

Kochová, P./Opavská, Z. (2016): Kapitoly z koncepce Akademického slovníku současné češtiny. [released by Jazairiová, P./Opavská, Z.]. Prague.

LDoCE (1995): Longman dictionary of contemporary English. 3rd Edition. Munich.

Sinclair, J. et al. (1998): Anglicko-český výkladový slovník. Prague.

Svensén, B. (2009): A handbook of lexicography. The theory and practice of dictionary-making. Cambridge.

SYN v10 (2022): Křen, M. et al.: Korpus SYN, verze 10 z 22. 2. 2022. Prague. http://www.korpus.cz (last access: 24-03-2022).

Töpel, A. (2021): Der Definitionswortschatz im einsprachigen Lernerwörterbuch des Deutschen. Anspruch und Wirklichkeit. Tübingen.

Xu, H. (2012): A critique of the controlled defining vocabulary in Longman dictionary of contemporary English. In: Lexikos 22, pp. 367–381.

Contact information

Michal Škrabal
Institute of the Czech National Corpus, Charles University, Prague
michal.skrabal@ff.cuni.cz

Michaela Lišková
Czech Language Institute, Czech Academy of Sciences, Prague
liskova@ujc.cas.cz

Martin Šemelík
Czech Language Institute, Czech Academy of Sciences, Prague
semelik@ujc.cas.cz

Acknowledgements

The preparation of this article was financed within the statutory activity of the Czech Language Institute of the Czech Academy of Sciences (RVO No. 68378092) and the European Regional Development Fund-Project “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.