At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project)

Pavla Kochová, Zdeňka Opavská, Martina Holcová Habrová
Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i.
kochova@ujc.cas.cz, opavska@ujc.cas.cz, holcova@ujc.cas.cz

Abstract

The aim of the article is to present the new lexicographic project that is being implemented at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i. Since 2012, its Department of Contemporary Lexicology and Lexicography has worked on the creation of a new medium-sized monolingual dictionary of Czech with the working title *Akademický slovník současné češtiny* (The Academic Dictionary of Contemporary Czech). With its size and method of treatment, the dictionary ranks among academic dictionaries, i.e. dictionaries with an elaborated, standardised and structured explanation of the meaning of lexical units, with an adequately rich exemplification documenting the typical use of lexical units, with a sufficiently elaborated description of the basic semantic relations, mainly synonymy and antonymy, with the appropriate description of the grammatical properties of lexical units and with usage labels (a description of the stylistic, temporal, spatial, frequency and pragmatic markedness) of lexical units.

Keywords: lexicography; monolingual dictionary; macrostructure of a dictionary; microstructure of an entry

1 Introduction

Akademický slovník současné češtiny (hereinafter as the *ASSČ*) builds on the tradition of the general monolingual dictionaries of Czech that emerged at the Institute for the Czech Language in the 20th century.[1] More than forty years have passed since the publication of a dictionary of a larger size, i.e. the

1 This tradition has developed from the largest *Příruční slovník jazyka českého* (Reference Dictionary of the Czech Language; *PSJČ*, 1935–1957), through the medium-sized *Slovník spisovného jazyka českého* (Dictionary of the Standard Czech Language; *SSJČ*, 1960–1971) to the one-volume *Slovník spisovné češtiny pro školu a veřejnost* (Dictionary of Standard Czech for Schools and the Public; *SSČ*, 1st edition in 1978; 2nd, revised edition in 1994; 3rd, revised edition in 2003). The *PSJČ* is a scientific descriptive dictionary of a large size (ca 250,000 entries); it describes Czech vocabulary since 1880; it does not use run-on entries; examples are provided by quotations. The *SSJČ* is a medium-sized dictionary (192,908 entries); it captures the literary lexical standard of the time, but the range of the word list exceeds it (by including obsolete, infrequent words etc.); it describes contemporary Czech vocabulary (approximately from the 1930s, selectively from 1880); the SSJČ uses run-on entries; exemplification is mainly based on the minimal typical contexts. The *SSČ* is a smaller-size dictionary (2nd ed.: 45,366 entries) focusing on the widest range of users; it describes the central vocabulary of contemporary Czech (mainly from 1945) with an overlap to variously marked words; it has a normative character; exemplification is very limited and is based on the minimal typical contexts. On the history and characteristics of the modern Czech lexicography, see mainly the detailed study by Hladká (2007).

Slovník spisovného jazyka českého (since the publication of the first volume it has even been fifty years), which is very long considering vocabulary dynamics, linguistic methodology,[2] in terms of the platform for the creation of a dictionary as well as the medium for its publication.

2 The Basic Characteristics of the ASSČ

The *ASSČ* is a *medium-sized* dictionary,[3] with the expected number of 120–150 thousand lexical units. Its *aim* is to capture widespread contemporary Czech vocabulary used in public official and semi-official communication as well as in everyday (i.e. non-public, unofficial) communication. A natural part of the lexis described are terminological expressions, but not highly specialised terms. To a limited extent, the dictionary presents units utilised in professional and interest-group communication, namely if their use has been extended beyond their professional, interest milieu. Dialectal expressions have been included if they are common in a wider area and are used especially in oral communication or in literature. The expected *user* of the dictionary is a secondary-school educated native speaker; nevertheless, also those interested in Czech as a foreign language are marginally taken into account (since Czech is a language of a small nation, specialised monolingual dictionaries of a larger size for learners are not created). The dictionary being prepared will be continuously *published* on the Internet (on the website of the Institute of the Czech Language). After the work on the dictionary has been completed, it will be possible to publish the work as a whole in a book form.

3 The Dictionary Development Method: Selected Aspects

The essential *material basis* is the synchronic corpus of written texts SYN of the Institute of the Czech National Corpus of a size of 2.2 milliard words. Other material resources are the electronic archives of the company Newton Media, a. s. (the archives of both nationwide and regional printed periodicals and transcripts of current affairs television and radio programmes), the internet and the databases of the Institute of the Czech Language.[4]

2 In connection with lexicography and lexicology, it is mainly the creation and development of computational and corpus linguistics and corpus lexicography (language corpora, excerpt databases, electronic archives, special software tools, eg. for an analysis of collocations the Word Sketch Engine – Kilgarriff et al. 2004). Cf. Čermák, Blatná 1995; Čermák 2010.

3 It should be emphasised that the dictionary being developed is not a lexical database. The relation between a lexical database and a monolingual dictionary is understood in accordance with Hanks (2010: 581): "A lexical database is a fundamental background resource for use in the creation of many important linguistic artefacts – dictionaries, course books, computer programs for natural language processing among them. A great monolingual dictionary has a different function: it brings together speakers of a language, it has a socially integrative function, making explicit the basis of words and meanings and usage, which all uses of the language rely on."

4 An excerpt database of neologisms (focused on new lexical phenomena), a database of specialised vocabulary, the Pralex – preparatory lexical database and the Modern Czech lexical archives created in 1911–1991.

4 The Macrostructure of the Dictionary

The word list of the *ASSČ* is built using a different lexicographic technique than before.[5] It draws on a set of three balanced corpora, SYN 2000, SYN 2005 and SYN 2010. The entries are selected from an automatically generated word list mainly based on the frequency criterion and the criterion of the commonness of their usage (i.e. only widespread lexical units are included; specialised terms, professional and slang expressions etc. are included only selectively). On the other hand, the word list has been expanded on the basis of word-formation relations (members of word-formation groups) and on the basis of co-hyponymic and other relations (members of lexical-semantic classes have been added).

Unlike in earlier dictionaries (*SSJČ*, *SSČ*), *derivatives* (relational adjectives, adverbs, names of properties), which used to be added to the lemmas as run-on entries, are listed as separate entries now. The new method (including the explanation of the meaning and exemplification) makes it possible to give an adequate lexicographic description, which however requires a detailed, often demanding analysis, cf. the explanation of the meaning in the entry for the relational adjectives (*badatel* n. "researcher" → *badatelský* adj. "vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství" =pertaining to researchers, researching • consisting of researchers • intended for researchers, for researching), see figure 1.

badatelský přid.
vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství; syn. vědecký: *intenzivní badatelská práce; badatelské projekty; moderní badatelské přístupy; badatelské zaujetí; badatelský tým; nastupující badatelská generace; poskytovat badatelské a knihovnické služby; špičkové badatelské pracoviště*
□ *badatelský list* tiskopis sloužící k evidenci údajů o badateli a jeho výpůjčkách z knihovny, archivu ap.

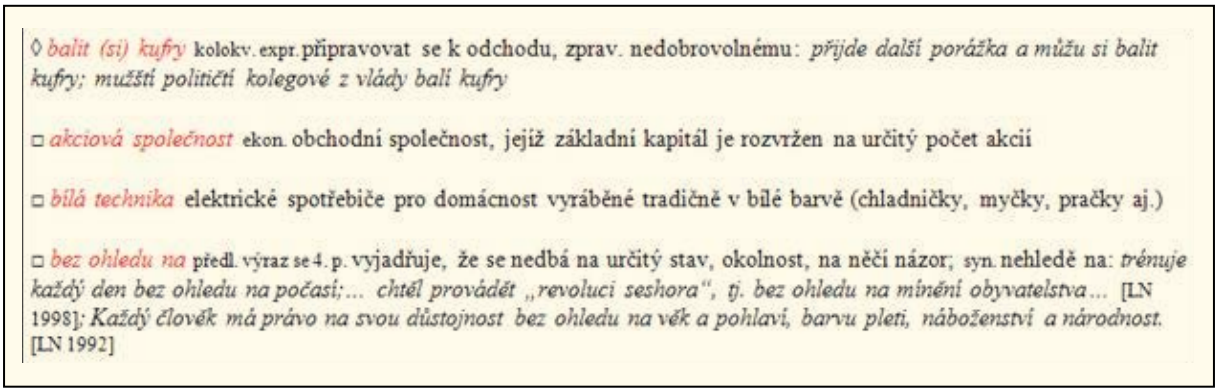
Figure 1: Entry *badatelský*.

Only some lexical types are treated as *run-on entries*. These include words derived by adding feminine suffixes (*herečka* ← *herec* "actress ← actor"), diminutives (*pejsek* ← *pes* "doggie ← dog") and frequentative verbs (*balívat* ← *balit* "to pack"), where the semantic structure of the derivative does not differ from the lemma. Cf. the lemma *bouček* "small beech" treated as a run-on entry in the entry buk "beech" (see figure 5).

In the ASSČ, greater autonomy has been given also to *multi-word lexical units*. The treatment distinguishes between: phraseological units (*balit si kufry* "to pack one's bags") and non-phraseological units (terminological – *akciová společnost* "joint-stock company"; non-terminological – *bílá technika* "white goods"; multi-word grammatical expressions – *bez ohledu na* "regardless of" preposition etc.). In the dictionary, these are listed in an one-word lemma entry, but it is taken into account that they are

5 The word list of the PSJČ relied on a comprehensive and sophisticated excerption of 5 million excerpts. The word list of the SSJČ built on the word list of the preceding dictionary, i.e. PSJČ, and its own excerption. Similarly, the latest of these modern dictionaries, the one-volume SSČ, proceeded from the word list of the SSJČ and its own excerption.

independent formal-semantic lexical units; therefore, the meaning explanation and exemplification are provided for a large part of them (always for phrasemes; for non-phraseological units, the explanation is given where the meaning is not compositional). The independence of multi-word lexical units is indicated also by the method of their presentation in the entry (a highlighted multi-word lemma, labelling with special symbols), see figure 2 and figure 3.


◊ *balit (si) kufry* kolokv. expr. připravovat se k odchodu, zprav. nedobrovolnému: *přijde další porážka a můžu si balit kufry; mužští političtí kolegové z vlády balí kufry*

□ *akciová společnost* ekon. obchodní společnost, jejíž základní kapitál je rozvržen na určitý počet akcií

□ *bílá technika* elektrické spotřebiče pro domácnost vyráběné tradičně v bílé barvě (chladničky, myčky, pračky aj.)

□ *bez ohledu na* předl. výraz se 4. p. vyjadřuje, že se nedbá na určitý stav, okolnost, na něčí názor; syn. nehledě na: *trénuje každý den bez ohledu na počasí;... chtěl provádět „revoluci seshora", tj. bez ohledu na mínění obyvatelstva...* [LN 1998]; *Každý člověk má právo na svou důstojnost bez ohledu na věk a pohlaví, barvu pleti, náboženství a národnost.* [LN 1992]

Figure 2: Multi-word lexical units balit (si) kufry, akciová společnost, bílá technika and bez ohledu na.

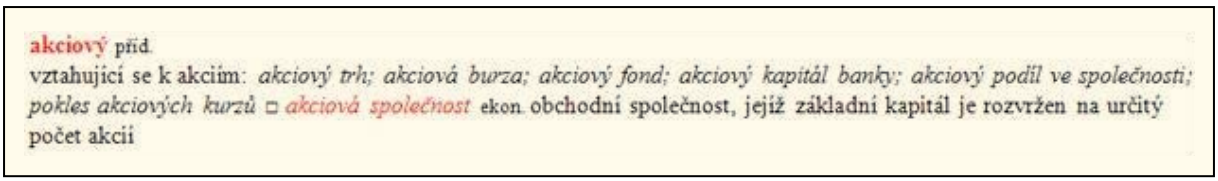

akciový přid.
vztahující se k akciím: *akciový trh; akciová burza; akciový fond; akciový kapitál banky; akciový podíl ve společnosti; pokles akciových kurzů* □ *akciová společnost* ekon. obchodní společnost, jejíž základní kapitál je rozvržen na určitý počet akcií

Figure 3: Multi-word lexical unit *akciová společnost* listed in the one-word entry *akciový*.

5 The Microstructure of an Entry

An entry in the *ASSČ* consists of the following parts: the lemma (including variant forms), information on homonyms, pronunciation, the etymology of the lexical unit, grammatical information (word class, morphology, valency), the usage label, the explanation of the meaning (including synonyms and antonyms), exemplification, notes (e.g. encyclopaedic information, further etymological information)[6] and cross-reference to (semantically, grammatically) related entries.

In the *ASSČ*, *grammatical information* (see figure 4) is treated more comprehensively than in previous monolingual dictionaries.[7] The morphological data in the ASSČ entries include mainly doublet forms and forms where the users may hesitate. The information on valency is systematically given for verbs

6 On the usage of notes, see e.g. the Oxford Dictionary of English (Soanes, Stevenson 2005), in Czech lexicography the neological dictionaries (Martincová et al. 1998, 2004).

7 In some respects, this transcends the genre of a general monolingual dictionary; on the other hand, it accommodates the users, who expect this type of information in a dictionary.

(both right and left valency), selectively also for nouns and adjectives. The valency information is semantically specified, if necessary, in the explanation of the meaning, or in the examples.

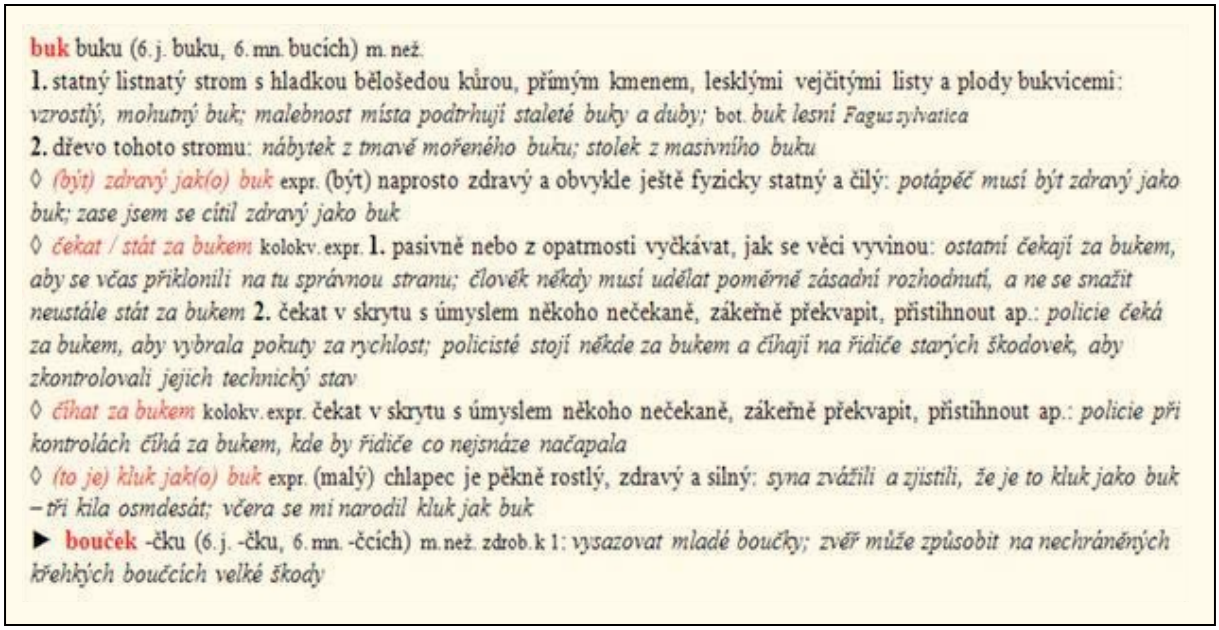
bafat (3. j. bafá, bafe, rozk. (ne)bafej!, čin. bafal, podst. jm. bafání) ned. expr.
4. (kdo || ~) (zprav. o psu nebo jiné psovité šelmě) vydávat jednotlivě vyrážené zvuky baf, haf; syn. štěkat: *pes bafal jako divý; Malý pokojový psík ňafá podstatně vyšším hlasem, než jakým bafá mohutná doga.* [Týdeník Rozhlas 2010]

Figure 4: Sense 4 of the lemma *bafat*.

When giving the *lexical meaning* of the entries, the *ASSČ* proceeds from the basic concept of determining the species classification – genus proximum – and differential semantic elements – differentia specifica (bearing in mind that besides notional elements also pragmatic elements need to be described). A part of the lexical meaning, however, are also those semantic elements that cannot be considered as necessary distinctive features but which mirror the complex of information on the denoted extra-linguistic reality that the language users have on the level of common knowledge. To a certain extent, the explanation of the meaning may hence contain "encyclopaedic" data (especially those that are objectively reflected in the word-formation structure of a word, in set similes and other phrasemes and in semantically derived meanings, on the basis of a metaphor).[8] In order to eliminate circularity, the explanation by means of synonyms is limited to a minimum, only to some slang, expressive or dialectal words.

The *exemplification part* of an entry includes both typical examples illustrating typical usage and extended examples that show less common, unusual and sometimes even authorial use of the word (mainly in the case of less frequent words and those belonging to peripheral areas of vocabulary). In addition, the examples are to illustrate grammatical information (especially on valency) and demonstrate (semantic) collocability. The exemplification may further contain those connotations which are not included in the explanation of the meaning but which the user (proto)typically connects with the unit concerned.

8 Dolník (2012: 45); cf. Buzássyová, Jarošová (2006: 27–28).



buk buku (6. j. buku, 6. mn. bucích) m. než.
1. statný listnatý strom s hladkou bělošedou kůrou, přímým kmenem, lesklými vejčitými listy a plody bukvicemi: *vzrostlý, mohutný buk; malebnost místa podtrhují staleté buky a duby;* bot. *buk lesní Fagus sylvatica*
2. dřevo tohoto stromu: *nábytek z tmavě mořeného buku; stolek z masivního buku*
◊ *(být) zdravý jak(o) buk* expr. (být) naprosto zdravý a obvykle ještě fyzicky statný a čilý: *potápěč musí být zdravý jako buk; zase jsem se cítil zdravý jako buk*
◊ *čekat / stát za bukem* kolokv. expr. **1.** pasivně nebo z opatrnosti vyčkávat, jak se věci vyvinou: *ostatní čekají za bukem, aby se včas přiklonili na tu správnou stranu; člověk někdy musí udělat poměrně zásadní rozhodnutí, a ne se snažit neustále stát za bukem* **2.** čekat v skrytu s úmyslem někoho nečekaně, zákeřně překvapit, přistihnout ap.: *policie čeká za bukem, aby vybrala pokuty za rychlost; policisté stojí někde za bukem a číhají na řidiče starých škodovek, aby zkontrolovali jejich technický stav*
◊ *číhat za bukem* kolokv. expr. čekat v skrytu s úmyslem někoho nečekaně, zákeřně překvapit, přistihnout ap.: *policie při kontrolách číhá za bukem, kde by řidiče co nejsnáze načapala*
◊ *(to je) kluk jak(o) buk* expr. (malý) chlapec je pěkně rostlý, zdravý a silný: *syna zvážili a zjistili, že je to kluk jako buk – tři kila osmdesát; včera se mi narodil kluk jak buk*
► **bouček** -čku (6. j. -čku, 6. mn. -čcích) m. než. zdrob. k 1: *vysazovat mladé boučky; zvěř může způsobit na nechráněných křehkých boučcích velké škody*

Figure 5: Entry *buk* "beech".

6 Software

Unlike earlier monolingual dictionaries created in the Institute of the Czech Language, the *ASSČ* has been compiled since the very beginning by means of specialised lexicographic software for dictionary creation (DWS). After various possibilities were considered, it was decided that new, special software needs to be developed as the creation of the dictionary has its significant specifics and the programme must be flexible. The software development has received grant support from the Ministry of Culture of the CR within the National and Cultural Identity (NAKI) applied research and development programme. (For more details on the software, see Barbierik et al. 2013; Barbierik et al. 2014.)

7 Conclusion

In the creation of the *ASSČ*, we are seeking new paths for the resolution of the issues that lexicographers have always faced as well as those of the modern present. Although the preparation of a good monolingual dictionary is a Sisyphean task – "the pursuit of perfection in lexicography is doomed to constant failure" –,[9] it must be attempted. "A dictionary of the national language is one of the basic needs of an educated man." (J. Jungmann, the preface to *Slovník česko-německý* (A Czech-German Dictionary)).

9 We borrowed the metaphor at the end from Hanks (2005: 254).

8 References

Barbierik, K. et al. (2013). A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In K. Gajdošová, A. Žáková (eds.), *Natural Language Processing, Corpus Linguistics, E-learning, Proceedings of the conference Slovko 2013, Bratislava, 13–15 November 2013.* Lüdenscheid: RAM-Verlag 2013, pp. 9–26.

Barbierik, K. et al. (2014). Simple and Effective User Interface of Dictionary Writing System. In *Euralex 2014 Proceedings, Bolzano 15–19 July 2014*.

Buzássyová, K., Jarošová, A. (eds.) (2006). *Slovník súčasného slovenského jazyka A–G.* (First edition.) Bratislava: Veda.

Czech National Corpus – SYN2000. Institute of the Czech National Corpus, Prague 2000. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN2005. Institute of the Czech National Corpus, Prague 2005. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN2010. Institute of the Czech National Corpus, Prague 2010. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN. Institute of the Czech National Corpus, Prague. Accessed at: http://www.korpus.cz [07/04/2014].

Čermák, F. (2010). Notes on Compiling a Corpus-Based Dictionary. In *Lexikos* 20 (*AFRILEX-reeks/series* 20:2010), pp. 559–579.

Čermák, F., Blatná, R. (eds.) (1995). *Manuál lexikografie.* Jinočany: H & H.

Dolník, J. (2012). Lexikálna pragmatika. In K. Buzássyová, B. Chocholová & N. Janočková (eds.), *Slovo v slovníku. Aspekty lexikálnej sémantiky – gramatika – štylistika (pragmatika). Na počesť Alexandry Jarošovej.* Bratislava: Veda, pp. 41–49.

Hanks, P. (2005). Johnson and Modern Lexicography. In *International Journal of Lexicography*, 18(2), pp. 243–266.

Hanks, P. (2010). Compiling a Monolingual Dictionary for Native Speakers. In *Lexikos* 20 (*AFRILEX-reeks/series* 20:2010), pp. 580–598.

Hladká, Z. (2007). Lexikografie. In J. Pleskalová, M. Krčmová et al. (eds.), *Kapitoly z dějin české jazykovědné bohemistiky.* Prague: Academia, pp. 164–198.

Jungmann, J. (1835 (1834) – 1839). *Slovník česko-německý.* (5 vol.) Prague: Knížecí arcibiskupská knihtiskárna.

Kilgarriff, A. et al. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6–10 2004.* Lorient: Université de Bretagne-Sud, pp. 105–116.

Martincová, O. et al. 1998. *Nová slova v češtině. Slovník neologizmů 1.* Prague: Academia.

Martincová, O. et al. 2004. *Nová slova v češtině. Slovník neologizmů 2.* Prague: Academia.

Příruční slovník jazyka českého 1935–1957. Prague: Státní pedagogické nakladatelství / SPN.

Slovník spisovné češtiny pro školu a veřejnost (1978). (Second, revised edition 1994; third, revised edition 2003.) Prague: Academia.

Slovník spisovného jazyka českého (1960–1971). (First edition.) Prague: Nakladatelství ČSAV.

Soanes, C., Stevenson, A. (eds.) (2005). *Oxford Dictionary of English* (Second, revised edition.) Oxford: Oxford University Press.

Acknowledgements

The article has been written within the grant project of the National and Cultural Identity (NAKI) applied research and development programme A New Path to a Modern Monolingual Dictionary of Contemporary Czech (DF13P01OVV011).