sciendo
DOI 10.2478/jazcas-2021-0057

BUILDING CZECH TEXTBOOK CORPORA (UcebKo) FOR WORD-FORMATION RESEARCH OF CZECH AS A SECOND LANGUAGE

ADRIANA VÁLKOVÁ
Masaryk University, Brno, Czech Republic

VÁLKOVÁ, Adriana: Building Czech textbook corpora (UcebKo) for word-formation research of Czech as a second language. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 631 – 640.

Abstract: This work-in-progress paper presents a specialized language corpus UcebKo built from textbooks of Czech for foreigners. The corpus integrates three subcorpora (UcebKo-A2, UcebKo-B1, and UcebKo-B2) which allow research of Czech as a second/foreign language at chosen language levels (A2, B1, and B2). In this case, the research is focused on word-formation, where the first results, i.e., mapping of derived words denoting persons, illustrate the approach and methodology used.

Keywords: word-formation, derivational morphology, textbook corpus, Czech as a second language, names of persons

1 INTRODUCTION

Most language corpora can be understood as extensive databases of parts of texts, in which it is possible to search and sort individual text units (sentences, word combinations, words, etc.) and observe them in their natural context ([1]). For most linguistic research, it is more appropriate to work with the so-called annotated corpus, where each corpus text unit (the so-called token) is provided with a lemma, the word form itself, and morphological information in the form of tags (i.e., information about the part of speech and its grammatical categories). In terms of general vs. specialized (specific) lexicon we differentiate between two types of corpora – general and specialized. General corpora are built for the sake of making generalizations (relating to morphology, lexicology, etc.) about the language. Specialized language corpora (in contrast to general corpora) always have a specifically defined purpose for which they are built – there are many types of specialized corpora (e.g. [2]).

In the case of specialized corpora made up of textbooks, the so-called textbook corpora (e.g., [3], [4]), the aim can be twofold – 1. to map the language (metalanguage) of textbooks, i.e., the linguistic and/or pedagogical research is aimed at all parts of the textbook or 2. to capture the vocabulary of the target group of textbooks. We have aimed at point two, i.e., to build a textbook corpus that would represent the Czech language of foreigners. A corpus like this can be understood as a simplified natural language (specifically the Czech language), where the degree of

simplification is determined by the language level (from A2 to B2). Corpus UcebKo could work as a basis of any research of Czech as a second language. In our case, the corpus works as a basis of word-formation research in which we want to map and then obtain (the most common) affixes for language levels A2, B1, and B2 and present them by using appropriate vocabulary.

2 MOTIVATION

Morphology plays a key role in the acquisition of Czech as a second language, as in other languages with a richly developed morphology (more than 75% of the Czech lexicon consist of derived words, see [5]). In addition to the inflectional morphology, which is focused on creating different forms of one word (e.g., from the word *otec* 'father' forms like *otcovi* 'to the father', *otcové* 'fathers') and which the student – a foreigner – encounters from the beginning, derivational morphology exists as a separate part of word-formation. Derivational morphology deals with the formation (or reproduction) of new words from already existing words (e.g. *mluvit* 'to speak' → *mluvčí* 'speaker'). As both morphologies are closely related – thanks to the suffix from which the word has been derived, it is possible 1. to identify the part of speech and 2. to classify the word within its paradigm (e.g., *cestovatel* 'traveller' is derived by the means of the suffix *-tel*, i.e., it is a noun which is inflected according to the *muž* 'man' paradigm). Some of the word-formation rules in Czech are mostly well acquired (e.g., adverbialization of adjectives: *krásný* 'beautiful' → *krásně* 'beautifully'), but most of them cause problems due to 1. the polyfunctionality of most suffixes (e.g., the suffix *-ka* with about 27 different meanings: a person (*manželka* 'wife'), an appliance (*sušička* 'a dryer'), a deminutive (*dcerka* 'little daughter') etc., 2. the irregular morphological alternations (e.g., *e/a*: *vejce* 'an egg' → *vaječný* 'made from eggs', [6]) and due to 3. many options of how to name the facts around (e.g., in Czech there are about 19 suffixes for naming a person according to the action which this person does).

There is no publication or textbook that systematically works with word-formation, or the existing textbooks do not provide a complete view of the word-formation system of Czech although their vocabulary could be many times more extensive if the students-foreigners acquired the word-formation principles in Czech.

We assume the results of this corpus research could be useful as a basis for any work with word-formation or for any word-formation project intended for students-foreigners.

3 TEXTBOOK CORPUS UcebKo

3.1 Corpus characteristics and composition

The UcebKo corpus is a specialized language corpus created from nine textbooks of Czech for foreigners including the keys thereto (that is, where the key

was available). This type of corpus represents the vocabulary that should be acquired by students-foreigners (contrary to the learner corpus which represents vocabulary that has been already acquired by students-foreigners, including the acquired mistakes). In general, UcebKo represents a sample of natural language (Czech) which is simplified based on the certain language level (A2–B2). It is possible to assume that textbooks capture rather the core of the lexicon than its periphery (which must be taken into account for linguistic research).

UcebKo integrates three subcorpora:

1. **UcebKo-A2** created from Czech textbooks for foreigners for level A2,
2. **UcebKo-B1** created from Czech textbooks for foreigners for level B1,
3. **UcebKo-B2** created from Czech textbooks for foreigners for level B2.

The designations according to the CEFR (see [7]) were used in all three corpora mentioned, in accordance with the textbooks they have been derived from. Level A1 was intentionally omitted, because according to a search of the textbooks of Czech for foreigners, they do not deal with word-forming phenomena at such a low language level (an exception is just a single textbook).

Every subcorpus always consists of three textbooks (see Table 1). In general, those authors who have written a textbook for more than one language level were preferred.

	UcebKo-A2	**UcebKo-B1**	**UcebKo-B2**
1	*Česky krok za krokem 1* 'Czech Step by Step 1' (from 13th chapter)	*Česky krok za krokem 2* 'Czech Step by Step 2'	*Čeština pro azylanty a cizince (B2)* 'Czech for asylum seekers and foreigners (B2)'
	L. Holá (2016) Praha, Akropolis	L. Holá – Bořilová, P. (2014) Praha, Akropolis	A. Adamovičová et al. (2006) Brno, SOZE
2	*Česky, prosím II.* 'Czech, please II.'	*Česky, prosím III.* 'Czech, please III.'	CZech it UP! B2
	J. Cvejnová (2012) Praha, Karolinum	J. Cvejnová (2016) Praha, Karolinum	D. Hradilová (2020) Olomouc, UPOL
3	*Čeština pro cizince A1 a A2* 'Czech for foreigners A1 and A2' (from 5th chapter)	*Čeština pro cizince B1* 'Czech for foreigners B1'	*Čeština pro cizince B2* 'Czech for foreigners B2'
	M. B. Kestřánková et al. (2017) Brno, Edika	M. B. Kestřánková et al. (2016) Brno, Edika	M. B. Kestřánková et al. (2020) Brno, Edika

Tab. 1. Composition of corpus

3.2 Corpus size

The UcebKo corpus spans 303,862 words and the size of each subcorpus is different (see Table 2). The smallest is the UcebKo-B2 subcorpus and the largest is the UcebKo-B1 subcorpus.

	UčebKo-A2	UčebKo-B1	UčebKo-B2
number of words	91,561	122,604	89,697
number of sentences	15,099	16,993	9,739
average sentence	6 words / sentence	7 words / sentence	9 words / sentence

Tab. 2. Size of the UcebKo corpus

The size of an average sentence (in terms of the number of words the sentence consists of) differs for each level (see 3rd line in Table 2) – data have been found thanks to statistical data of the corpus interface. An average sentence at the B2 level is 9 words, which is quite a long sentence and, therefore, it can be assumed that complex sentences will predominate.

3.3 Criteria for textbook selection

The textbooks from which the corpora are made up have been chosen according to the six criteria that were set with regard to the purpose for which the corpora have been built:

1. **criterion determines the target group of textbook users – adults** (because the results of corpus research will be used in projects which are primarily intended for adult students-foreigners),
2. **criterion is the type of textbook should not be purely grammatical but should be more conversational** (because we want to capture as much of the natural context of the derived words as possible, not the grammatical rules, etc.),
3. **criterion is the recency of the textbook and it was decided not to incorporate a textbook older than 15 years** (because we want to work with the most recent Czech lexicon),
4. **criterion is the number of textbooks – three textbooks for each subcorpus**. This criterion was chosen due to impossibility to obtain more than three textbooks for the B2 language level. Distinction between B1 and B2 allows word-formation research according to language levels. Moreover, we presume the most frequent words (the core of the lexicon) do not change with a larger corpus size (cf. same as in general Czech corpora),
5. **criterion is that each subcorpus consisted of textbooks from different authors** (because we assume, a corpus built from textbooks of more authors is more objective than a corpus built from textbooks of only one author),
6. **criterion is that the textbooks were actually used in practice**, i.e., in teaching, etc.

3.4 Access to corpus

The corpus is accessible at the website of Sketch Engine (https://ske.fi.muni.cz/) only for verified users who have submitted a statement in which they undertake to use the corpus just for research purposes and also to publish corpus parts with only complete citations.

4 BUILDING OF THE UcebKo CORPUS

4.1 The corpus-building process

The textbook corpus UcebKo was created in SketchEngine ([8]), a corpus interface, thanks to which an already annotated corpus was created (Sect. 1). All Czech corpora created in SketchEngine have been annotated by the morphological analyzer Majka ([9]) and Desamb ([10]). The process of corpus building can be described as the result of five steps:

- **Obtaining the textbooks**.
- **Textbook scanning**.
- **OCR / Copy text**: The textbooks were converted from their scanned form to plain text using a program with the OCR function.
- **Cleaning text**: In this step, it was necessary to 1. check the obtained text against the original text from the textbooks and 2. set the criteria for what to keep in the text and what to remove (more info in 4.2 Cleaning text).
- **Creating corpus**: The text document was uploaded to the corpus interface which created the corpus automatically. This process takes between several seconds and a maximum of several minutes, so this is why it is the easiest step in the whole process of building a corpus.

4.2 Cleaning the text

First, it was necessary to check the obtained text against the original text in the textbook, because different errors may have occurred during the scanning and/or OCR phase. There were errors especially like bad text recognition due to a graphically processed background on which the text was written or incorrectly recognized diacritics over some words etc. After checking the texts, criteria were set by which it was determined exactly what would be kept in the text and what would be removed from the text. It was necessary to clarify the aim for which the corpus was being built in order to determine the criteria for cleaning the text, i.e., to create a database of texts which would represent simplified Czech corresponding to a certain language level. Therefore, only whole sentences were kept in the text and the word combinations or free-standing words were removed. Next, the language of mediation, inscribed pronounciation, and grammar explanation was removed because it does not represent natural language (natural context of words) and finally, the examples of poems were removed because they do not usually reflect current lexicon.

5 WORD-FORMATION RESEARCH IN UcebKo

5.1 Aim of word-formation research

The aim of word-formation research is 1. to map the quantity of the words derived from suffixes that denote the persons in the vocabulary of foreigners and 2. to capture the suffixes (the so-called word-forming types) from which these words are derived at

language levels A2, B1, and B2. It can be assumed that productive word-formation types will occur across all language levels. However, the aim is to find out which word-formation types are involved. The resulting data will provide a view on the Czech lexicon from the word-formation perspective presented by concrete numbers.

5.2 Methodology of word-formation research

The described research could be understood as a process that consists of three steps: 1. obtaining suffixes from books focused on Czech word-formation, 2. searching for words derived from these suffixes in the corpus, 3. final (quantitative) analysis of the found data.

The starting point of word-formation research was to obtain suffixes denoting a person. The list of suffixes was obtained by searching dictionaries or grammar specializing in Czech word-formation ([11] and [12]). Words containing the suffixes were searched for in the corpus by two queries – first, by a query containing morphological tags and then a query without morphological tags, where the string of characters was searched for at the end (e.g., *-tel*). The second query was performed as a check that takes into account the expected error rate of automatic natural language processing. The error rate found is mainly based on the basic properties of natural language, which is 1. the linguistic homonymy (e.g., the word *mluvčí* 'speaker' was not found by tags specifying the noun), and the fact that 2. the word form is missing in the dictionary of morphological analysis (see [13], [14]) (e.g., the word *antitalent* 'dullard' was not found by tags specifying the masculine animate noun).

After this corpus searching, all masculine animate nouns were searched and analyzed by their endings. In this way, the suffixes *-ál* (e.g. *profesionál* 'professional'), *-át* (e.g., *adresát* 'addressee'), *-eň* (e.g., *vězeň* 'prisoner') were found. The found suffixes were subsequently searched for in the largest dictionary of affixes for Czech (see [15]), in which they were found.

5.3 Derived words denoting persons

A total of 13,486-word forms (1,564-lemmas) were found by using the morphological tags for the masculine animate nouns. However, this number includes 1. a group of words which were incorrectly assigned as animate masculine (e.g., *knedlík* 'dumpling') and 2. words that do not denote persons (e.g., *pták* 'bird') which are necessary to remove for analysis. Also, proper individual names (first names, e.g., *Adam* and surnames, e.g., *Novotný*) have been removed, because they do not, in contrary to the rest of the words, denote persons according to certain circumstances or characters. Thanks to manual analysis it was found:

- 643-lemmas (8,350-word forms) were found as words denotating persons,
- out of 643-lemmas, 78.5% (505-lemmas) of lemmas were derived from other word(s),
- out of 505-lemmas, 91.9% (464-lemmas) of lemmas were derived from suffixes.

It was necessary to separate the derived words from loanwords adapted into Czech by suffixes. These loanwords (containing suffixes) constitute 14.8% (95-lemmas) of a total of 643-lemmas. These words are also mentioned here because they could be easily acquired by foreigners with a knowledge of word-formation principles (suffixes).

The number of words denoting persons and the words derived from suffixes differ in the individual subcorpora of UcebKo:

- **UcebKo-A2**: contains 263-lemmas denoting persons, of which 78.5% are derived,
- **UcebKo-B1**: contains 401-lemmas denoting persons, of which 77.1% are derived,
- **UcebKo-B2**: contains 409-lemmas denoting persons, of which 80.1% are derived.

5.4 Identified suffixes with the meaning denoting a person

A total of 28 suffixes with the meaning denoting a person have been found in the UcebKo corpus (see Table 3). Suffixes are sorted by their relative frequencies ([16]) because the work with absolute frequencies is not possible due to the different size of each subcorpus. Most of these suffixes are represented in all three textbook subcorpora, i.e., UcebKo-A2, UcebKo-B1 and UcebKo-B2 (see numbers 1–18 in Table 3). However, the number of suffixes found in the subcorpora is different:

- in A2 21 suffixes were found,
- in B1 20 suffixes were found,
- and in B2 25 suffixes were found.

Some suffixes (marked by *) have been found with a word-formation function (e.g., *student* ‘student’ ← *studovat* ‘to study’) and/or with a lexical function (e.g., *pacient* ‘ patient’ > lat. ‘patiēns‘), it depends on the concrete words. Suffixes with lexical function are part of loanwords and in this sense, the suffixes work as formal instruments of adaptation into Czech. Loanwords are not (naturally) included in the word-formation analysis presented in Table 3.

		UcebKo	**UcebKo-A2**	**UcebKo-B1**	**UcebKo-B2**
	suffix	number of lemmas with this suffix (and their relative frequency)			
1	*-tel*	**68 (7251.2)**	14 (2337.2)	22 (2316.4)	32 (2597.6)
2	*-ik/-ník*	**117 (4516.3)**	28 (1157.6)	40 (1541.5)	49 (1817.2)
3	*-ec (-ovec, -inec)*	**79 (4436)**	18 (1376.1)	31 (1566.0)	30 (1493.9)
4	*-č*	**27 (3847.8)**	7 (1190.4)	11 (1386.5)	9 (1270.9)
5	*-ař*	**29 (2305.8)**	7 (709.9)	14 (1150.0)	8 (445.9)
6	*-ent**	**14 (2297.9)**	3 (819.1)	6 (709.6)	5 (769.2)

		UcebKo	UcebKo-A2	UcebKo-B1	UcebKo-B2
	suffix	number of lemmas with this suffix (and their relative frequency)			
7	*-ista*	**49 (1936.3)**	15 (600.6)	20 (856.4)	14 (479.3)
8	*-ce*	**44 (1439.3)**	10 (382.2)	18 (522.0)	16 (535.1)
9	*-ář*	**56 (1234.9)**	10 (207.5)	20 (481.2)	26 (546.2)
10	*-ák*	**33 (1010.6)**	6 (207.5)	15 (424.1)	12 (379.0)
11	*-an*	**35 (794.7)**	9 (207.5)	14 (252.8)	12 (334.4)
12	*-ik**	**27 (702.7)**	8 (196.5)	11 (261.0)	8 (245.2)
13	*-ér**	**13 (616.3)**	3 (283.9)	5 (187.5)	5 (144.9)
14	*-or (-tor, -átor)**	**20 (330.2)**	4 (54.6)	9 (228.3)	7 (312.1)
15	*-ant*	**17 (288.9)**	5 (65.5)	5 (89.7)	7 (133.7)
16	*-a**	**10 (250.7)**	2 (43.6)	5 (73.4)	3 (133.7)
17	*-íř**	**3 (219.2)**	1 (54.6)	1 (97.8)	1 (66.8)
18	*-ina (-otina)*	**4 (96.9)**	1 (10.9)	1 (57.0)	2 (29.0)
19	*-áč*		-	-	3 (200.6)
20	*-eň*		-	-	1 (89.1)
21	*-át**		-	2 (16.3)	1 (44.5)
22	*-ál*		-	1 (8.1)	1 (22.2)
23	*-oun*		-	-	1 (22.2)
24	*-ita**		1 (21.8)	-	-
25	*-án**		-	-	1 (11.1)
26	*-och*		-	-	1 (11.1)
27	*-ka*		1 (10.9)	-	-
28	*-l*		1 (10.9)	-	-

Tab. 3. Suffixes for derivation of words denoting a person found in UcebKo

As it is possible to see (Table 3), the words denoting persons are most often derived from the suffixes *-tel* (e.g., *učitel* 'teacher'), *-ík/-ník* (e.g., *mladík* 'a young man', *zákazník* 'customer'), *-ec* (e.g., *cizinec* 'foreigner') and *-č* (e.g., *rodič* 'parent') at language levels A2–B2, and moreover, the suffix *-tel* was found to be the most frequent. It must be said that these suffixes, which occur across all subcorpora, play a key role in Czech language acquisition because students-foreigners are confronted with them almost all the time when learning Czech (from A2 to B2).

6 CONCLUSION AND FUTURE WORK

The submitted paper has presented a process of building a textbook corpus called UcebKo, which has been built with the purpose of having language material for word-formation research in Czech as a second language. The corpus UcebKo integrates three subcorpora (UcebKo-A2, UcebKo-B1, and UcebKo-B2) which

allow conducting a separate research of the Czech vocabulary of foreigners at language levels A2, B1, and B2. The described research was focused on the mapping of suffixes from which words denoting names for persons are derived.

The process of building the corpus was described as a five-step process: 1. to obtain textbooks, 2. to scan them, 3. to do an OCR to get the text alone, 4. to clean the text and 5. to create the corpus in the corpus interface of SketchEngine. The cleaning of the text was found to be the most arduous phase of the corpus building. In this phase, it was necessary to decide what to remove (structures without sentence form, the mediation language, the written pronunciation, the examples of poems, and grammar explanations) and what to keep (only the sentence structures that are not grammatical in nature only).

The suffixes from which the words denoting persons are derived at language levels A2, B1, and B2 have been found and presented in the form of lists. Eighteen suffixes have been found at all researched language levels: *-a, -ák, -an, -ant, -ař, -ář, -ce, -č, -čí, -ec, -ent, -ér, -ián, -ik, -ík, -ina, -íř, -ista, -or (-tor/-átor), -tel*. Moreover, it was found that most often persons are named by the suffixes *-tel, -ík/-ník, -ec*, and *-č*. The data could be compared with data from general corpora in the future (but it will be undertaken as individual research due to the large polyfunctionality of the suffixes).

The next research will be focused on describing the meanings of the found derivatives by their semantic features. The result will be presented in lists intended for language levels A2, B1, and B2. Derived words processed in this way could be useful in the field of didactics of Czech as a foreign language – for lecturers of Czech (for creating word-formation exercises) and for students-foreigners (such as knowledge of how to name a person in concrete circumstances).

ACKNOWLEDGEMENTS

This work was supported by grant No. TL003000293 (Word-formation Analysis Software Tool for Teaching Czech for Foreigners) through the Technology Agency of the Czech Republic.

References

[1] Oliva, K., and Doležalová, D. (2004). O korpusu jako o zdroji jazykových dat. In Korpus jako zdroj dat o češtině. Brno, Masarykova univerzita, pages 7–10.

[2] Cvrček, V. (2021). Struktura Českého národního korpusu. In Wiki Český národní korpus. Accessible at: https://wiki.korpus.cz/doku.php/cnk:struktura.

[3] Vališová, P. (2013). Učebnicový korpus a jeho využití pro výuku češtiny jako cizího jazyka. In J. Klímová, Gramatika a korpus 2012: 4. mezinárodní konference. Hradec Králové. Accessible at: http://utkl.ff.cuni.cz/~rosen/public/GC2012/Konferencni_prispevky/ValisovaPavlina.pdf.

[4] Meunier, F., and Gouverneur, C. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material.

[5] Dokulil, M. (1962). Tvoření slov v češtině. 1, Teorie odvozování slov. Praha, Nakladatelství Československé akademie věd.
[6] Ševčíková, M. (2018). Modelling Morphographemic Alternations in Derivation of Czech. The Prague Bulletin of Mathematical Linguistics, 110, pages 7–42. Accessible at: https://ufal.mff.cuni.cz/pbml/110/art-sevcikova.pdf.
[7] Ivanová, J. (2002). Společný evropský referenční rámec pro jazyky: jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme. Olomouc, Univerzita Palackého v Olomouci.
[8] Kilgarriff, A., Rychlý, P., Jakubíček, M., Rundell, M. et al.: Sketch Engine [Computer Software and Informatiom Resource]. Accessible at: http://www.sketchengine.co.uk.
[9] Jakubíček, M., Kovář V., and Šmerk, P. (2011): Czech Morphological Tagset Revisited. In A. Horák, P. Rychlý (eds.), Proceedings of Recent Advances in Slavonic Natural Languages Processing. Brno: Tribun EU, 2011, pages 29–42, 14 p. ISBN 978-80-263-0077-9.
[10] Šmerk, P. (2008): K morfologické desambiguaci češtiny. Accessible at: https://is.muni.cz/auth/th/wteg5/teze.pdf. Advanced Master's thesis. Masaryk University, Faculty of Informatics.
[11] Štícha, F. et al. (2013). Velká akademická gramatika spisovné češtiny. Praha, Academia.
[12] Karlík, P., Nekula, M., and Pleskalová, J. (2016). Nový encyklopedický slovník češtiny. Praha, Nakladatelství Lidové noviny. Accessible at: https://www.czechency.org/slovnik/.
[13] Osolsobě, K. (1996). Algoritmický popis české morfologie a strojový slovník češtiny. Brno, Masarykova univerzita. Disertační práce.
[14] Brno Morphological Analyzer Ajka. Accessible at: https://nlp.fi.muni.cz/projekty/wwwajka/.
[15] Šimandl, J. (2016). Slovník sufixů užívaných v češtině. Praha, Univerzita Karlova, Karolinum. Accessible at: http://www.slovnikafixu.cz.
[16] Kováříková, D. (2021). Frekvence. In Wiki Český národní korpus. Accessible at: https://wiki.korpus.cz/doku.php/pojmy:frekvence.