sciendo
DOI 10.2478/jazcas-2019-0073

PROCESSING OF DERIVATIONAL FEATURES FOR (SEMI)AUTOMATIC CREATION OF DICTIONARY DEFINITIONS IN THE USER INTERFACE (CZEDD) FOR LEARNING CZECH AS A SECOND LANGUAGE: SUFFIX *-tel* AND *-ista*

ERIK CITTERBERG – ADRIANA VÁLKOVÁ
Faculty of Arts, Masaryk University, Brno, Czech Republic

CITTERBERG, Erik – VÁLKOVÁ, Adriana: Processing of derivational features for (semi)automatic creation of dictionary definitions in the user interface (CZEDD) for learning Czech as a second language: suffix *-tel* and *-ista*. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 444 – 455.

Abstract: This work-in-progress paper presents the tool CZEDD which enables the user to learn how to predict the meaning of words. The CZEDD consists of (semi) automatic definitions for derived words because a lot of these words have predictable lexical meaning. The tool will be intended for foreigners who learn the Czech language and it could be useful as a dictionary and/or translator in which the definitions based on the word's structure are stored. Two detailed case examples (the suffix *-tel*, and the suffix *-ista*) illustrate the approach.

Keywords: derivational morphology, Czech for foreigners, suffixes, lexical meaning, structural meaning, dictionary

1 INTRODUCTION

The Czech language is a Slavic language with richly developed morphology. Foreigners who learn the Czech language are confronted with it from the beginning. Next to inflectional morphology which studies how the forms of lexemes are created by morphemes (e.g. from the noun *pes*[1] 'a dog' forms *psovi* 'to a dog', *psi* 'dogs' and e.g. from the verb *mít* 'to have' forms *měli jsme* 'we had', *máš* 'you have' (form in singular)) it is necessary for complete knowledge of Czech to know how some of these morphemes build other different lexemes, not just their forms (e.g. noun *knihovna* 'a library' derived from noun *kniha* 'a book'). This is a part of the derivational morphology and the word-formation in general. However, there are many studies, from codification grammars to online tools, which handle inflectional morphology for

[1] In this text, we translated to English just those derived words which we have found in English dictionary Glosbe (see [6]), we did not try to create the neologism. When we don't find the shape in the dictionary, we write the verb which is semantically related to the agent name (e.g. *vychovatel* which means 'one who nurtures especially children' so we write it like this: *vychovatel* ← *vychovat* 'to nurture'.

Czech language and which show how to work with it in the teaching and/or learning of Czech language for foreigners, the textbooks included. The information of meaning, so-called structural meaning, of these word-formation morphemes are mostly found only in the specialized books about the word-formation (see e.g. [4], [5], [12]), in newer Czech grammars (see e.g. [2], [16]) and in online dictionaries (see e.g. [7], [14]). In this field of study are tools which show the derivational relations, for example, the DeriNet (see [13], [18]), Deriv or Derivancze (see [11]) and Morfio (see [3]) which show more of formal relations than semantic relations.

At this moment, there is no study which focuses on the predictability of lexical meaning of Czech derived words, especially for how much concrete suffixes are predictable or not, and its use for teaching.

In the following sections, we present the tool CZEDD (Czech electronic derivational dictionary) and processing of suffixes which are used in this application. CZEDD provides an option of working with word-formation as a part of grammar which may play a key role in learning (acquisition) the Czech language by the clearly determined meaning of affixes.

2 MOTIVATION

Native speakers can predict the meaning of words they have never heard before or they can subconsciously create a "new" word for the specific context using word-formation morphemes. This is because they know the meaning of a suffix analogically based on already known words, e.g. *publikovatel* 'a person who publishes something' ← *publikovat* 'to publish' with the analogy to words ending with *-tel*: *učitel* 'a teacher', *cestovatel* 'a traveller', etc.). The lexical meaning is a complex of the historical, social and other influences, and for its complete understanding the structural meaning is insufficient. On the other hand, a rough estimate of unknown word meaning could prove to be of value for the fluency of communication. We think the foreigners who will periodically use this app might become more aware of the structure of words. Moreover, the morphemes with word-formation function carry specific semantic information (e.g. *knih-ovna* 'a library'): *-ovna* is a name for a place) and more specific grammatical information of a part of speech and its properties (e.g. *knihovna*: noun, feminine, noun paradigm *žena* 'a woman'). However, foreigners find it difficult to recognize the paradigm of nouns.

Students who learn Czech as a second language speak at least one other language (their mother tongue). With the knowledge of suffixes similar to that of native speakers, the foreigners should be able to understand the approximate meaning of an internationalism which is adapted to the Czech language by suffixes. Moreover, for Slavic students with similar mother tongue to Czech, it is possible to expect a quick understanding of these adaptations and such students should be able to acquire the derivation rules intuitively.

3 PROCESSING OF AFFIXES FOR CREATING DEFINITIONS IN THE CZEDD

3.1 Processing affixes

We focus on the most frequent and productive suffixes used for deriving nouns. We have processed nouns derived by adding monofunctional suffixes *-tel* and *-ista*. We have found the possible meaning of nouns derived by these suffixes according to the information about them in the online dictionary and from the specialized books, mentioned in the Introduction. We have tested this meaning on data of written Czech corpora SYNv6 (SYNv7) (see [8]) which enables us to find and work with the most frequent of them. The queries are specified in Corpus Query Language (CQL). For the words from the corpus, we have compared their structural meaning and the meaning found in the online dictionaries in Lexiko (see [17]) and evaluated the correspondence between them in percent. For the words for which the lexical and structural meaning are in acceptable correspondence, we are trying to find the most general definition.

3.2 Evaluation data from corpus

For suffix *-tel* we have processed 1 129 lemmas, i.e. all lemmas for the corpora query [tag="N.*"& lemma=".**tel*"][2] and we have found out that for words with lower frequency the structural meaning corresponds to their lexical meaning. Therefore, for the next suffix, suffix *-ista* specified by query [tag="N.I.*"& lemma=".**ista*"], we have processed only the 200 most frequent word forms. The word-formation research in corpus is described in e.g. [9], [10].

Suffix *-tel*

Out of all found lemmas, those which are not derived (e.g. epitel 'an epithelium') have been manually removed and 1 129 lemmas have been chosen to be further processed. This number includes the unprefixed and prefixed forms (prefix *do-, na-, o-, ob-, od-, po-, pod-, pro-, pře-, před-, při-, roz-, s-, u-, v-, vy-, vz-, z-, za-*)[3].

General and simplified structural meaning found in books: suffix *-tel* means, in general, an agens of some action with semantic features [+Person], [+Masculine], [+Animate] [+Agens]. This action is represented by the verb from which the noun is derived.

56,07% of 1 129 lemmas were found in the dictionaries. The lexical meaning does not correspond to the structural meaning in 3,01%, for e.g. *nakladatel* 'a publisher', *věřitel* 'a creditor', *buditel* 'a revivalist' and we have found there is a group of the impersonal nouns:

[2] We use regular expressions occurring in the SYN corpus. ".*" is interpreted as any character repeated from zero to potential infinity.

[3] For this corpus research, we have worked with prefixed nouns derived from prefixed verbs, i.e. we did not process the nouns with prefix *nad-* (e.g. *nadučitel* (freely translated 'more than teacher') and prefix *pod-* in words like *podučitel* (freely translated 'less than teacher').

- [-Person], [+Masculine], [-Animate], [+Agens]: typically, mathematics and business names *jmenovatel* ‘a denominator’, *dělitel* ‘a divisor’, *čitatel* ‘a numerator’, *násobitel* ‘a multiplier’, *menšitel* ‘a subtrahend’, *úročitel* ‘an interest rate’, *odmocnitel* ‘a square root’, *umořovatel* ‘one payment in a series of installments’, *odúročitel* ‘a discount rate’, *součinitel* ‘a coefficient’
- 3 words which could be [+Person] [+Masculine] [+Animate], [+Agens] or [-Person], [+Masculine], [-Animate], [+Agens]: *činitel* ‘a factor/an agent’, *ukazatel* ‘a pointer’, *zaměstnavatel* ‘an employer’[4].

We have found that the lexical meaning is more specific in 4,34% (e.g. *spisovatel* ‘a writer’) and concurrently we have not found the nouns for which the lexical meaning is more general than their structural meaning.

The nouns derived by the suffix *-tel* from perfect verbs (e.g. *vydražitel* ‘an auctioneer’ ← *vydražit* ‘to auction off’) expresses the action that has been done or will be done. The nouns derived from imperfect verbs (*vyšetřovatel* ‘an investigator’ ← *vyšetřovat* ‘to investigate’) means the action is in progress.

We have found that 74,3% nouns are derived from imperfect verbs and 25,7% nouns from 1 129 lemmas are derived from perfect verbs. But in 5,31% we have found

a. the nouns derived from perfect verbs behave like nouns derived from imperfect verbs:

- names for professions (e.g. *vychovatel* ← *vychovat* ‘to nurture’, *zastupitel* ‘a representative’ ← *zastoupit* ‘to deputize’
- name for a person for whom this action is typical of (*zastavitel* ← *zastavit* ‘to pawn’, *chovatel* ‘an animal keeper’ ← *chovat* ‘to keep’) but not typical as a job

and b. nouns derived from imperfect verbs, but behave like the nouns derived from perfect verbs:

- e.g. *pachatel* ‘an offender’ derived from imperfect verb *páchat* ‘to offend’ but with the definition for the perfect verb *pachatel = ten, kdo spáchal* ‘one who committed a crime’, *zakladatel* ‘a founder’ is *ten, kdo něco založil* ‘one who founded an organization’, *zastupitel* ‘a representative’ is *ten, kdo zastupuje* ‘one who represents’.

Most of the nouns are derived from verbs of III–V[5] verbal classes, though we have found two exceptions: *přistižitel* ‘one who caught an offender in an act’ derived

[4] The target of the next study will be finding the context of words which can be animate and inanimate as well and we want to find which is the predominant interpretation. We could not use the tags of the SYN corpus tagset to recognize it, because of inaccuracies of the disambiguation.

[5] According to traditional division of Czech verbs based on their forms in present tense into five verbal classes.

from verb *přistihnout* ‘to catch’ belonging to the II verbal class; and noun *přemožitel* ‘one who overcame someone or something’ derived from verb *přemoct* ‘to overcome’ belonging to the I verbal class.

Suffix *-ista*

General (and simplified) structural meaning: name for a person with semantic features [+Person], [+Masculine] and [+Animate].

We have found that in the group of the 200 most frequent lemmas, it is not possible to predict the lexical meaning in 50,5% of the lemmas, while the structural meaning can be applied in 49,5% of the lemmas:

- 27,5% nouns derived from nouns ending with *-ismus* (e.g. *fašista* ‘a fascist’ ← *fašismus* ‘a fascism’)
- 9% names for instrumental players (e.g. *kytarista* ‘a guitarist’ ← *kytara* ‘a guitar’)
- 8,5% names for sports players (e.g. *fotbalista* ‘a footballer’ ← *fotbal* ‘football’)
- 4,5% for nouns derived from words *-istika* (e.g. *cyklista* ‘a cyclist’ ← *cyklistika* ‘cycling’)

3.3 Generation of definitions

We have created definitions for nouns derived by adding the suffix *-tel* depending on verbal aspect. We can distinguish between the perfect and imperfect verbs thanks to the DeriNet. Definitions have been created by specifying the endings and according to the existing or not existing prefix for both suffixes, i.e. *-tel* and *-ista*.

First step – verbal aspect recognition

At first, we focused on verbal form without the prefix[6] in two previous steps

- e.g. *zpracovatel* ‘a processor’ ← *zpracovat* ‘to process’ ← *pracovat* ‘to work’
  definition *ten, kdo zpracoval nebo zpracuje* ‘one who processed or processes’

Also, based on the existence of imperfective verbs in the scope of two previous derivational steps, we have found the nouns derived from secondary imperfective forms:

- e.g. *dotazovatel* ← *dotazovat* ← *dotázat* ← *tázat* ‘to ask’
  definition *ten, kdo dotazuje* ‘one who asks’.

Second step – creating the definition

Suffix *-tel*

Figure 1 shows the steps used for generating definitions. There are a few exceptions: *chovatel* ‘an animal keeper’, *klovatel* ← *klovat* ‘to peck’, *snovatel* ← *snovat* ‘to weave’, *plovatel* ← *plovat* ‘to float’, *kovatel* ← *kovat* ‘to smith’, which are individually specified.

[6] Prefix *ne-* is not computed, because it does not change the verbal aspect.

[^ch]ovatel	prefix	NO	„ten, kdo .*uje“		
		YES	the string contains a verb .*it or .*nout or [aá]t?	YES	„ten, kdo .*uje“
				NO	„ten, kdo .*oval or .*uje“

[^o][iíyýaá]vatel	„ten, kdo .*[íýá]vá“

[^o]ěvatel	„ten, kdo .*ívá“

-itel	prefix	NO	the string contains a verb .*u.it?	„ten, kdo .*u.í“
			the string contains a verb .*[^u].[eěi]t?	„ten, kdo .*í“
			the string contains a verb .*u.ovat a .*ou.it?	„ten, kdo .*ou.il nebo .*ou.í“
		YES	the string contains a verb .*ou.it?	„ten, kdo .*ou.il nebo .*ou.í“
			the string contains a verb .*ovat & not existing the verbal form .*it?	„ten, kdo .*oval“
			the string contains a verb .*[eě]t?	„ten, kdo .*[eě]l nebo .*í“
			the string contains a verb .*[^ou].*it?	„ten, kdo .*il nebo .*í“

-[^z^b]atel	prefix	NO	„ten, kdo .*á“
		YES	„ten, kdo .*al nebo .*á“

-zatel	prefix	NO	„ten, kdo .*že“
		YES	„ten, kdo .*zal nebo .*že“

-batel	prefix	NO	„ten, kdo .*bá (.*be)“
		YES	„ten, kdo .*bal nebo .*bá (.*be)“

-p[ií]satel	prefix	NO	„ten, kdo píše“

Fig.1. Rules for definition generation for nouns *-tel*

Suffix *-ista*:

We have created a definition for nouns derived from nouns ending with *-ismus* (e.g. *komunista* 'a communist' ← *komunismus* 'communism' and for nouns derived from nouns ending with *-istika* (e.g. *cyklista* 'a cyclist' ← *cyklistika* 'cycling'). We have also created a definition for foreign adapted words with meaning "name for sports players", but they are derived from the base word, not as in two previous groups. Definitions are created according to base word endings (e. g. *hokejista* 'a hockey player' ← *hokej* 'hockey' and *fotbalista* 'a footballer' ← *fotbal* 'football'):

- noun derived from nouns *-ismus*: definition *stoupenec* [.**ismu*] ('a follower of [.*]')
- noun derived from nouns *-istika*: definition *ten, kdo se zabývá* [.**istikou*] ('one who is an enthusiast of [.*]')
- adapted words of foreign origin ending with *-ej*: *hráč* [.**eje*] and *-al*: *hráč* [.**alu*] ('a player of [.*]')

We have found the suitable correspondence in meaning and word structure for words with meaning "instrumental players" but we have not found the way how to write a rule which will apply to most of these words.

4 TECHNICAL REALIZATION

4.1 Technical realization

The CZEDD could be considered both a conventional dictionary with automatically generated definitions of words, organized according to predefined typology of derivation, and a user interface built for interaction with the DeriNet with extended functionality of implementing Majka [15] and Ajka [1] morphological analyzers.

As a wrapper for all technologies, the Flask Python framework has been used. The simplest form of the CZEDD is pregenerated database with derivational and morphological information for chosen words contained within the Derinet network.

To interact with the database a web application has been created to connect the CZEDD database, the Derinet and Ajka. This web application serves as a user-friendly interface for searching within the database.

The user can enter a word or a text as an input which is then checked against the database and the DeriNet. Additional information is then provided from Ajka.

The database was generated from the DeriNet with series of filters, mainly in the form of regular expressions, which enabled us to find words for which we have sufficient rules to create a suitable definition as well as provide additional information about them. Results of this process was a list of words divided into types, which were further analyzed with rules defined for specific derivational types and then saved into a table in database which can then be queried by users via the CZEDD web application.

Lemmatization and morphological analysis has been done with Majka and for missing words, a function that was checking Ajka api was implemented.

In case the user searches for an unknown word which is not included neither in our CZEDD database, nor in the DeriNet network, it is saved for future review if labeled as one of processed derivational types.

The concept of processing each word was based on creating class objects in Python for words.

As a base, class 'Word' has been created. In the first step based on word endings, an internal pseudo-derivational type has been determined. Then the word has been tagged by Majka and in case it was not found within its data an http request to Ajka has been submitted to retrieve a morphological tagging from there. Both morphological analyzers use tag format developed at Masaryk University. Tag enabled us to determine several attributes: lemma, gender, number, paradigm, part of speech.

The DeriNet was then queried for base word and retrieval of derivational branch up to second verb. This also enabled us to check if a prefix could be identified. English translation of each word within the derivational branch has been extracted by sending an http request to Glosbe API [6].

In case a word has been identified to be within our derivational typology, a class based on Word class has been created. This class was named TypeWord. Additional attributes include definitions in Czech and English languages created by applying rule based substitutions based on their prefix and pseudo-derivational type attribute.

For text input containing multiple words a Text class has been created which is an object of Word and TypeWord objects with additional dictionary attributes for storing original unprocessed words with their lemma as a value.

All this information has been provided for all words within the DeriNet network and the CZEDD database has been then generated. The web application serves as a user interface for this database as well as a searching tool for words within the DeriNet network.

4.2 DeriNet – Derivational network

The DeriNet is a lexical network, which comprises core word-formation relations. The network is currently limited to derivational relations. The network has been extracted from an existing corpus of contemporary Czech and semi-automatically generated using existing data resources (corpora and lexical resources).

Generated candidate pairs of a derived word and its base word were checked manually before creating an edge in the network, unless they came from a highly reliable resource.

The relations between derived words and their base words are modeled as an oriented graph. Nodes of the graph correspond to lexemes. Edges represent derivational steps between lexemes. The orientation of edges reflects the word-

formative process: the edge points from a base lexeme to a derived lexeme. Each lexeme can have at most one base lexeme.

The DeriNet is publicly available on the Internet at `http://ufal.mff.cuni.cz/derinet`. It can be used under the terms of the Creative CommonsAttribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA). The data is available in a simple line-oriented format as well as in a self-documenting XML-based format.

4.3 Majka – Morphological tool

For lemmatization and morphological analysis in the CZEDD web application and in the CZEDD database generation two morphological analyzers were used. These tools contain different data. When possible, Majka was used due to much higher speed and Ajka was used for words not contained within Majka.

It has several functionalities including lemmatization, morphological tagging as well as generating word forms for given lemma based on given tag.

4.4 Glosbe – Online dictionary

There are many commercial solutions available for bilingual Czech to English dictionary. However, in case of open source solutions for simple Czech to English dictionary, the only on-going project that we know of is the Glosbe online dictionary.

Glosbe is a simple multilingual online dictionary that runs as a community project managed by a small development team based in Poland. Its purpose is to create an extensive polylingual general purpose dictionary. Among other things, it contains examples of usage for words taken from several sources as well as general definitions for certain suffixes.

Data included in the Glosbe dictionary are under various licenses: CC-BY-SA, FDL and custom license. Data source is always indicated next to data if it is needed due to the license.

5 CZEDD – CZECH ELECTRONIC DERIVATIONAL DICTIONARY

5.1 What is the CZEDD?

The CZEDD is a user interface which enables to understand the principle of semantic and formal connections between Czech derived words. There are two basic functions – 1. insert word and 2. insert text. The CZEDD works like a special bilingual dictionary with definitions based on word structure (see the CZEDD as a dictionary). The function "**Insert text**" provides the processing of text in which derived words are colour marked and for the colour marked words the same process is applied as in the first function.

The CZEDD can be used in the teaching of Czech word-formation, as an e-learning tool, especially for more advanced students. A different way how to use the CZEDD is for translations from Czech to English (especially for beginners) and as a translator from Czech to Czech (for advanced students).

5.2 CZEDD as a translator

CZEDD can be used as a translator, especially for the newly created words, neologisms or the words created just for the concrete situation (context). Most translators cannot work with these types of words. CZEDD provides bilingual translator interface: from Czech to English.

5.3 CZEDD as a dictionary

In CZEDD you can find a grammatical information about searched word (see Figure 2). This dictionary is available through the "**Insert word**" function.

1. definition
2. part of speech
3. noun gender
4. noun paradigm
5. base word
6. derivation process

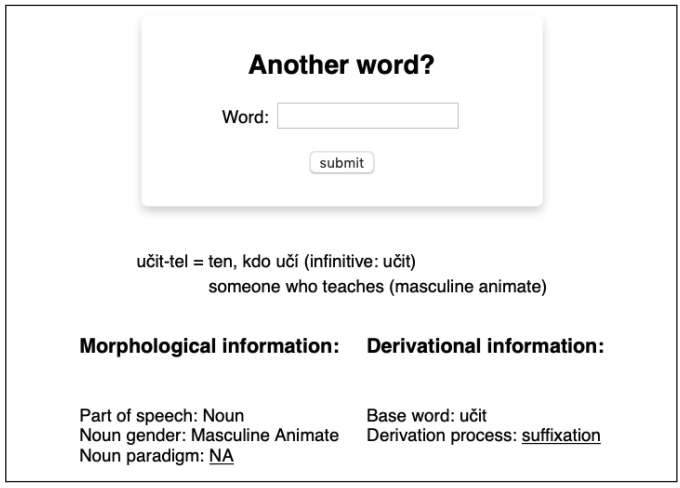

Fig. 2. Processed word

6 DISCUSSION AND FUTURE WORK

We have tried to create definitions for derived words from their structural meaning. We have processed 1 129 nouns derived by suffix *-tel* and the first 200 most frequent nouns derived by suffix *-ista*. General definition was created for most nouns derived by *-tel* according to the verbal aspect: *ten, kdo* [*dělá*] *nebo* [*udělal*/*udělá*] ('one who [does] or [has done/will do]'). However, it was necessary

to separate two groups of nouns derived by *-ista* depending on their base word: a. *-ismus* (*komunista* 'a communis' ← *komunismus* 'communism'): *stoupenec* [.**ismu*] ('a follower of ...'), b. *-istika* (*cyklista* 'a cyclist' ← *cyklistika* 'cycling'): *ten, kdo se zabývá* [.**istikou*] ('one who is an enthusiast of...'). We have also processed nouns with meaning "sports players" according to the endings of their respective base (non-derived) words: *-al*, *-ej* (*fotbalista* 'a footballer' ← *fotbal* 'football'; *hokejista* 'a hockey player ← *hokej* 'hockey' with definition *hráč* [.**alu*/.**eje*] 'a player of...').

The lexical meaning was not in correspondence with the structural meaning for nouns derived by the suffix *-tel,* which amounts to 3,01%. This percentage contains 10 nouns which are not primarily animate: *jmenovatel* 'a denominator', *dělitel* 'a divisor', *čitatel* 'a numerator', *násobitel* 'a multiplier', *menšitel* 'a subtrahend', *úročitel* 'an interest rate', *odmocnitel* 'a square root', *umořovatel* 'one payment in a series of installments', *odúročitel* 'a discount rate', *součinitel* 'a coefficient'; and three nouns which could mean a person, or an inanimate object: *činitel* 'a factor/an agent', *ukazatel* 'a pointer', *zaměstnavatel* 'an employer'.

As expected, most of nouns (74,3%) are derived from imperfect verbs and only 25,7% nouns are derived from perfect verbs. Nouns derived from perfect verbs have an identical meaning as the nouns derived from imperfect verbs in 5,31%.

In the future, we want to add online exercises which will enable students to strengthen their knowledge of Czech word-formation, especially the derivation. It will be adjusted to their concrete language level due to Common European Framework (CEFR). We plan to provide examples of use by adding the sentences from the corpus.

In the future, the easiest way of extending the scope of CZEDD would be naturally, via using already created scripts for regenerating more complete databases as its source materials, such as the DeriNet, keep growing.

With continuous work on defining more derivation types, we will be able to find new rules for generating not only more automatic definitions, but also using rule-based approach with cross verification with other resources for further extension of the DeriNet network.

Further didactic functionality is also possible as well as making the results and created materials more accessible with our own API. This approach will make possible both further extensions via third party applications as well as creating a more user-friendly iterations of CZEDD itself.

ACKNOWLEDGMENTS

This work was supported by the project of specific research Czech language in unity of synchrony and diachrony – 2019 (MUNI/A/1061/2018).

References

[1] Brno Morphological Analyzer Ajka. Accessible at: https://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm

[2] Čechová, M. (1996). Čeština – řeč a jazyk. Praha, ISV nakladatelství.

[3] Cvrček, V. and Vondřička, P. (2013). Morfio – aplikace pro analýzu slovotvorných vztahů. Praha, FF UK. Accessible at: http://morfio.korpus.cz.

[4] Daneš, F., Dokulil, M., and Kuchař, J. (eds.) (1967). Tvoření slov v češtině 2. Odvozování podstatných jmen. Praha, Academia.

[5] Dokulil, M. (1962). Tvoření slov v češtině. 1, Teorie odvozování slov. Praha, Nakladatelství Československé akademie věd.

[6] Glosbe – multilingual online dictionary. Accessible at: https://cs.glosbe.com/cs/en.

[7] Karlík, P., Nekula, M., and Pleskalová, J. (2016). Nový encyklopedický slovník češtiny. Praha, Nakladatelství Lidové noviny. Accessible at: https://www.czechency.org/slovnik/.

[8] Křen, M. et al. (2018). Korpus SYN, verze 7. Praha, Ústav českého národního korpusu FF UK. Accessible at: https://www.korpus.cz.

[9] Osolsobě, K. (2011). Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny. Brno, Masarykova univerzita.

[10] Osolsobě, K. (2011). Korpus jako zdroj dat pro studium slovotvorby. In Petkevič, V. – Rosen, A. (eds.), Korpusová lingvistika Praha 2011 – 3. Gramatika a značkování korpusů, pages 10–23, Praha.

[11] Pala, K., and Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In TSD 2015, pages 515–523. Accessible at: https://link.springer.com/chapter/10.1007/978-3-319-24033-6_58.

[12] Skoumalová, Z., Dokulil, M., and Panevová, J. (1997). Obsah – výraz – význam: Výbor z lingvistického díla Miloše Dokulila I. Praha, Univerzita Karlova, Filozofická fakulta.

[13] Ševčíková, M., and Žabokrtský Z. (2014). Word-Formation Network for Czech.(LREC). Accessible at: http://www.lrecconf.org/proceedings/lrec2014/pdf/501_Paper.pdf.

[14] Šimandl, J. (ed.) (2016). Slovník afixů užívaných v češtině. Praha, Karolinum. Accessible at: http://www.slovnikafixu.cz/.

[15] Šmerk, P. (2007). Fast Morphological Analysis of Czech. In Petr Sojka and Aleš. Proceedings of Third Workshop of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pages 13–16, Brno, Masaryk University.

[16] Štícha, F. et al. (2013). Velká akademická gramatika spisovné češtiny. Praha, Academia.

[17] Webové hnízdo o novodobé české slovní zásobě a výkladových slovnících LEXIKO. Accessible at: https://lexiko.ujc.cas.cz/heslare/.

[18] Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J., and Limburská, A. (2016). Merging Data Resources for Inflectional and Derivational Morphology in Czech (LREC). Accessible at: http://ufal.mff.cuni.cz/~straka/papers/2016lrec_derinet.pdf.