sciendo
DOI 10.2478/jazcas-2019-0070

INTRODUCING SEMANTIC LABELS INTO THE DeriNet NETWORK

MAGDA ŠEVČÍKOVÁ – LUKÁŠ KYJÁNEK
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic

ŠEVČÍKOVÁ, Magda – KYJÁNEK, Lukáš: Introducing semantic labels into the DeriNet network. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 412 – 423.

Abstract: The paper describes a semi-automatic procedure introducing semantic labels into the DeriNet network, which is a large, freely available resource modeling derivational relations in the lexicon of Czech. The data were assigned labels corresponding to five semantic categories (diminutives, possessives, female nouns, iteratives, and aspectual meanings) by a machine learning model, which achieved excellent results in terms of both precision and recall.

Keywords: derivation, semantic category, comparative semantic concepts, suffix, machine learning

1 INTRODUCTION

Although word-formation in general and derivation in particular is defined as a process affecting both form and meaning of words, most language resources that focus on derivation lack explicit semantic information. The present paper describes a recent account of introducing semantic labels into the DeriNet network, which is a large, freely available resource modeling Czech derivation [25].

After a brief overview of how meaning is approached in selected theoretical treatments of derivation and how it is captured in existing language resources (Section 2), basic facts on the DeriNet network are summarized in Section 3. The pilot experiment on semantic labeling of derivational relations in DeriNet is described in Section 4. Adhering to basic principles of a cross-linguistic proposal of comparative semantic concepts in affixation [2], we have chosen five semantic categories to be assigned by a semi-automatic procedure using machine learning techniques. Results of the experiment and future steps are discussed in Section 5.

2 SEMANTICS OF DERIVATIONAL RELATIONS IN EXISTING DESCRIPTIONS AND LANGUAGE RESOURCES

2.1 Theoretical accounts of meaning in derivation

An elaborate description of word-formation in Czech was proposed by Dokulil [3] and, since then, broadly accepted and applied in all reference grammars of Czech,

incl. the representative volume by Dokulil et al. [4] and the latest reference grammars, e.g. [28]. As Dokulil's account proceeds primarily in the meaning-to-form direction, a sort of semantic classification is, in fact, an inherent part of descriptions of word-formation in Czech grammars. A closer look reveals, however, that the descriptions are organized, first, according to the part-of-speech category of the derivatives and, second, according to the part-of-speech category of the base words, and only then the meaning of affixes is taken into account. Since semantics is used as a third-level criterion, derivatives with the same derivational meaning are split into several subgroups if belonging to different part-of-speech categories. An even more detailed, though even more fragmented description of meaning, is provided by the recent dictionary of affixes used in Czech [26].

A theoretical approach to derivational meanings that we would like to adhere to in semantic labeling of the DeriNet network is anchored in linguistic discussion on comparative semantic concepts, which are argued to be more adequate for cross-linguistic studies than established grammatical categories rooted in descriptions of particular languages [6]. Applying this discussion to derivation, Bagasheva [2] proposes a set of 51 comparative semantic concepts (Table 1). The concepts, designed as language-independent, are not limited to a particular type of affixation (prefixation, infixation etc.) and are applied across part-of-speech categories.

ABILITY	*DESIDERATIVE*	*INCEPTIVE*	*PRIVATIVE*	*SIMILATIVE*
ABSTRACTION	*DIMINUTIVE / ATTENUATIVE*	*INSTRUMENT*	*PROCESS*	*SINGULATIVE*
ACTION	*DIRECTIONAL*	*ITERATIVE*	*PURPOSIVE*	*STATE*
AGENT	*DISTRIBUTIVE*	*LOCATION*	*QUALITY*	*SUBITIVE*
ANTICAUSATIVE	*DURATIVE*	*MANNER / VIEWPOINT*	*RECIPROCAL*	*TERMINATIVE*
AUGMENTATIVE / AMELIORATIVE / INTENSIVE	*DWELLER*	*ORNATIVE*	*REFLEXIVE*	*TEMPORAL*
CAUSATIVE	*ENTITY*	*PATIENT*	*RELATIONAL*	*UNDERGOER*
COLLECTIVITY	*EXPERIENCER*	*PEJORATIVE*	*RESULTATIVE*	
COMITATIVE	*FEMALE*	*PERCEPTIVE*	*REVERSATIVE*	
COMPOSITION	*HYPERONYMY*	*PLURIACTIONALITY*	*SATURATIVE / TOTAL*	
CUMULATIVE	*HYPONYMY*	*POSSESSIVE*	*SEMELFACTIVE*	

Tab. 1. Comparative semantic concepts proposed by Bagasheva [2]

Another inspiring approach to derivational meanings, though not applied in our work, was elaborated in the Meaning-Text Theory. Derivational relations between words are captured by a subset of Lexical Functions, which are defined as mathematical functions whose arguments and values are lexical units. Lexical Functions were applied in the Explanatory Combinatorial Dictionary ([15], [16]).

2.2 Meaning in language resources focusing on derivational morphology

Even if several resources focusing on derivational morphology have been made available for selected European languages in the last decade (see [13] for a detailed overview), semantic issues are, to the best of our knowledge, addressed in more or less explicit way only in some French resources (cf. Morphonette and Démonette; [7], [8]) and in the Czech resource Derivancze [20].

Derivancze is a tool that searches dictionary data for derivations. For an input word, this tools provides its base word and a word or words immediately derived from it, if available in the data. Each of 255 thousand derivational relations contained in this resource was assigned a semantic label. The set of a total of 17 semantic labels was extracted from existing resources, esp. from the morphological analyzer [18] and Czech WordNet [19].

Labels in Derivancze differ from our approach described below in that some of them are more fine-grained (e.g. female surnames are labeled differently from female counterparts of common nouns in Derivancze), they seem to be limited to a particular part of speech (e.g. the diminutive label is attested with nouns only) and, moreover, they cannot be used as a feature for searching the data.

Pieces of information that relate to semantics of Czech derivations can be found also in resources which do not have word-formation or derivation as their primary focus; cf. morphological analyzers and corresponding dictionaries and tools ([5], [22], [23]) and general and specialized lexicographic resources ([9], [14]). Several of them were used in compilation of the training and test data sets for our labeling experiment (cf. Section 4.2).

3 DERINET

The DeriNet network has been developed since 2013 as a database of Czech words connected with links corresponding to derivational relations [25]. In DeriNet, 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 them, pointing from the base word to the derived one. Each derivative has at most one base word. Thus, a primary (unmotivated) word is the root of the tree and all its derivatives are organized according to their morphemic and semantic complexity from the simplest to the most complex ones; see the tree structure in Fig. 3.

Lexemes in DeriNet were extracted from the MorfFlex CZ dictionary [5], which covers a major part of the lexicon of contemporary Czech including proper names, archaic words, low-frequency words and regular, automatically generated coinages without respect to whether they are attested in a corpus. Derivational relations between lexemes were created semi-automatically under manual control, preferring high precision to recall.

The current version, DeriNet 2.0 [30], contains more than 1 million lexemes connected with more than 809 thousand derivational relations. The DeriNet data are available for download (Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, CC BY-NC-SA 3.0), and can be searched online by the DeriSearch tool.[1]

4 A SEMI-AUTOMATIC APPROACH TO ASSIGNING WORD-FORMATION RELATIONS WITH SEMANTIC LABELS

4.1 Linguistic decisions on the design of the experiment

The task of introducing semantic labels into DeriNet was a challenge mainly due to the size of the data, which does not allow for a large-coverage manual annotation, and second, due to that the resource is still under construction (edges are either added, or deleted in course of revisions). The task was thus designed as a semi-automatic procedure and limited to only five semantic categories in this pilot phase, namely to:

- *DIMINUTIVE*: in line with the proposal of comparative semantic concepts (and unlike the corresponding semantic label in Derivancze), we assume this category to be expressed by words belonging to different part-of-speech categories in Czech by using suffixes (cf. examples for this and the other labels in Table 2),
- *FEMALE*: this category subsumes female counterparts of both masculine animate common nouns and proper nouns in Czech, both derived by suffixation,
- *POSSESSIVE*: in our experiment this semantic category is limited to denominal derivation of possessives (with the affixes *-ův* and *-in*) that relate to an individual, and is thus narrower than the respective comparative semantic concept (since not including possessives related to a group of individuals or to a species like *pes* 'dog' > *psí* as in *psí srst* 'dog hair'),
- *ITERATIVE*: following up the long-lasting linguistic debate on this category (reviewed by Ševčíková and Panevová [24] with respect to the DeriNet data), this category is assumed to be limited to imperfective verbs derived from imperfectives in Czech by different suffixes,
- *ASPECT*: this label, as the only one in our experiment, has no counterpart in the repertoire of comparative semantic concepts; it relates to a previous decision made in the course of the build-up of DeriNet to include pure aspectual counterparts into the data since the category of aspect is, unlike other inflectional categories of verbs, conveyed by derivational morphemes [24]; this semantic label is applied to suffixation of verbs from verbs when changing aspect.

In accordance with the focus of DeriNet, the semantic labels are meant to reflect the structural, word-formation meaning while lexical shifts are not taken into

[1] http://ufal.mff.cuni.cz/derinet/search

consideration [27]. The aim of the labeling experiment was to apply the labels to the entire DeriNet data.

example	label to assign
pes 'dog' > *psík* 'small dog'	*DIMINUTIVE*
žlutý 'yellow' > *žluťoučký* 'yellowish'	*DIMINUTIVE*
málo 'little' > *maličko* 'very little'	*DIMINUTIVE*
spát 'to sleep' > *spinkat* 'to sleep' (baby talk)	*DIMINUTIVE*
učitel 'teacher' > *učiteka* 'female teacher'	*FEMALE*
Jaroslav (male first name) > *Jaroslava* (female first name)	*FEMALE*
Novák (male surname) > *Nováková* (female surname)	*FEMALE*
učitel 'teacher' > *učitelův* 'teacher's'	*POSSESSIVE*
učitelka 'female teacher' > *učitelčin* 'female teacher's'	*POSSESSIVE*
chodit 'to walk.IPFV' > *chodívat* 'to walk.IPFV repeatedly'	*ITERATIVE*
kupovat 'to buy.IPFV' > *kupovávat* 'to buy.IPFV repeatedly'	*ITERATIVE*
chytit 'to catch.PFV' > *chytat* 'to catch.IPFV'	*ASPECT*
štěkat 'to bark.IPFV' > *štěknout* 'to give a bark.PFV'	*ASPECT*

Tab. 2. Examples of semantic categories

4.2 Compilation of the training and test data sets

For the machine learning experiment, training and test data sets were prepared in four subsequent steps. First, relevant base-derivative pairs were extracted from existing language resources, namely from the monolingual dictionary of Czech [9] (examples of all five categories; see Fig. 1), from MorfFlex CZ [5] (instances of diminutives, possessives and female names; Fig. 2), and from the VALLEX dictionary [14] (examples to be assigned the *ITERATIVE* and *ASPECT* labels). Only those pairs were included that are attested in DeriNet and a derivational link is established in the data.

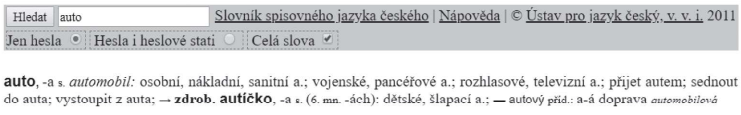

Fig. 1. Entry of the noun auto 'car' in the monolingual dictionary [9] (`https://ssjc.ujc.cas.cz/`), containing the diminutive autíčko 'small car' (incl. the semantic category)

word	morphological lemma	comment
astrofyzička	astrofyzička^(^FM*3k)	remove 3 letters, add 'k', and get: astrofyzik
bioložka	bioložka^(^FM*3g)	remove 3 letters, add 'g', and get: biolog
cizinka	cizinka^(^FM*2ec)	remove 2 letters, add 'ec', and get: cizinec

Fig. 2. Candidates of female nouns extracted from MorfFlex CZ [5]. Semantic category (^FM for female noun) and base words are encoded in the morphological lemma (see the "comment").

Second, diminutive and possessive suffixes that were not covered by the exploited resources were searched for in reference books (esp. [11], [17], [26]) and used to identify relevant derivatives in the DeriNet data. After a manual annotation, these instances were added to those extracted from the dictionaries.

These instances, which positively substantiated the relations under consideration, were complemented by negative examples (to be assigned none of the five semantic categories) in the third step. The negative examples were extracted from DeriNet under manual control and assigned a sixth label (*none*). In this way, a data set consisting of a total of 14,752 both positive and negative instances was compiled, see Table 3.

label	*DIMINUTIVE*	*FEMALE*	*POSSESSIVE*	*ITERATIVE*	*ASPECT*	*none*
count	3,303	1,449	2,252	1,555	3,719	2,474

Tab. 3. Portions of examples for each semantic label and the none label within the data set of a total of 14,752 instance

In the fourth step, the data set consisting of positive and negative examples was assigned features in Table 4. All features were binarized according to the labeled data, which increased dimensionality, mainly because of the n-gram features (encoded as one-hot).

The data were then divided into three data sets (with no overlaps):

- 80% of the data were used as a training data set for training the machine learning model,
- 10% of the data served as a development test data set to find adequate probability thresholds for each semantic label,
- 10% of the data were used as an evaluation test data set for evaluation of the model.

4.3 Development of the machine learning model

Starting with a preliminary set of supervised machine learning experiments using the Python 3 scikit-learn module [21], Multinomial Logistic Regression (MLR) has been chosen as the most promising method for the semantic labeling task, showing better results than Decision Tree and Naive Bayes methods.

feature	values	comment
part-of-speech category of the derivative	N (noun), A (adjective), V (verb), D (adverb)	- source: DeriNet 1.7 [29]
part-of-speech category of the base word	N (noun), A (adjective), V (verb), D (adverb)	- source: DeriNet 1.7
gender of the derivative	M (masculine animate), I (masculine inanimate), F (feminine), N (neuter)	- assigned with nouns - source: MorfFlex CZ [5]
gender of the base word	M (masculine animate), I (masculine inanimate), F (feminine), N (neuter)	- assigned with nouns - source: MorfFlex CZ
aspect of the derivative	PFV (perfective), IPFV (impfective), B (biaspectual)	- assigned with verbs - source: MorfFlex CZ, SYN2015 corpus [12], VALLEX [14]
aspect of the base word	PFV (perfective), IPFV (impfective), B (biaspectual)	- assigned with verbs - source: MorfFlex CZ, SYN2015 corpus, VALLEX
possessivity tag	1 (with possessives), 0 (others)	- assigned with the derivative - source: MorfFlex CZ
final n-grams of the derivative	final bi-, tri-, tetra-, penta-, hexagrams	
final n-grams of the base word	final bi-, tri-, tetra-, pentagrams	
semantic label of the derivational relation	*DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE, ASPECT, none*	- source: monolingual dict. [9], MorfFlex CZ, VALLEX, manual annotation

Tab. 4. Features to assign with the training and test data

MLR is a generalization of the logistic regression method to multiple target tasks. For all given features of each class, MLR estimates adequate regression parameters. As an output, the MLR method returns probability values based on logistic sigmoid function for each target class ([1], [10]).

To set the MLR model for prediction of semantic labels, *newton-cg* solver was used, which is predefined in scikit-learn. The number of iterations was increased up to one thousand to converge. The goal of the MLR model to be trained was to classify examples according to the most probable semantic label, taking into account the highest possible precision.

Based on the performance of the method on the development test data set, the following thresholds were determined for individual semantic labels in order to further increase the precision: 0.75 for *DIMINUTIVE*, 0.4 for *FEMALE*, 0.4 for *POSSESSIVE*, 0.5 for *ITERATIVE*, and 0.4 for *ASPECT*. If the probability of the most probable semantic label predicted by the model was below the particular threshold, the semantic label was not accepted (changed to *none*). The results of MLR on the training and evaluation test data sets are reported in Table 5.

	accuracy	precision	recall	f1-score
training data set	0.992	0.991	0.992	0.991
evaluation test data	0.986	0.984	0.984	0.984
sample of predicted data	0.971	0.962	0.963	0.962

Tab. 5. Evaluation of the trained MLR model on the training data, evaluation test data, and manually annotated random sample of 2,000 relations from predicted data

The MLR model was applied to the previous version of the DeriNet data (DeriNet 1.7; [29]), which were previously assigned the same features as the training and test data (except for the semantic label feature; see Table 4). The MLR model assigned one of the five semantic labels to 150,521 derivational relations in total. The *POSSESSIVE* label was the most frequent one (predicted with 88,620 derivational relations), followed by the *FEMALE* label (28,510 rel.), *ASPECT* (15,459 rel.), *ITERATIVE* (11,890 rel.), and *DIMINUTIVE* (6,042 rel.).

The precision and recall of the labeling procedure were evaluated on a randomly selected sample of 2,000 relations assigned either one of the five semantic categories or the *none* label; see Table 5 for evaluation of the sample as a whole and Table 6 for details on individual labels.

gold / predicted	*DIMINUTIVE*	*FEMALE*	*POSSESSIVE*	*ITERATIVE*	*ASPECT*	*none*
DIMINUTIVE	62	0	0	0	0	4
FEMALE	1	296	0	0	0	3
POSSESSIVE	0	0	905	0	0	1
ITERATIVE	0	0	0	135	4	0
ASPECT	0	0	0	3	170	1
none	1	39	1	0	0	374
precision	0.969	0.982	0.999	0.985	0.987	0.948
recall	0.983	0.941	0.999	0.988	0.987	0.976

Tab. 6. Confusion matrix based on manual annotation of a random sample of 2,000 relations and precision and recall calculated for individual labels in the sample

Semantic labels, as assigned in the machine learning experiment, are part of the current version of the DeriNet network (DeriNet 2.0, [30]). Semantic labels can be used for searching the data by the DeriSearch tool. A sample tree containing semantic labels is displayed in Fig. 3.

5 DISCUSSION AND FUTURE WORK

The word-formation system of Czech is characterized by homonymy of affixes,[2] on the one hand, and synonymy of affixes, on the other. Many affixes convey more

[2] The term "homonymy" [17] or "polyfunctionality" [26] is preferred to "polysemy" in recent accounts.

than one meaning, cf. *-ka* is used to express the diminutive meaning in *skříňka* 'small cupboard' while in *učitelka* 'female teacher' it falls under the *FEMALE* category, but it occurs also in instrument nouns (*žehlička* 'iron') and other formations. From the opposite perspective, a particular meaning is usually expressed by several, formally different affixes, cf. the suffixes *-ka, -yně, -ice, -ová* in the *FEMALE* category.

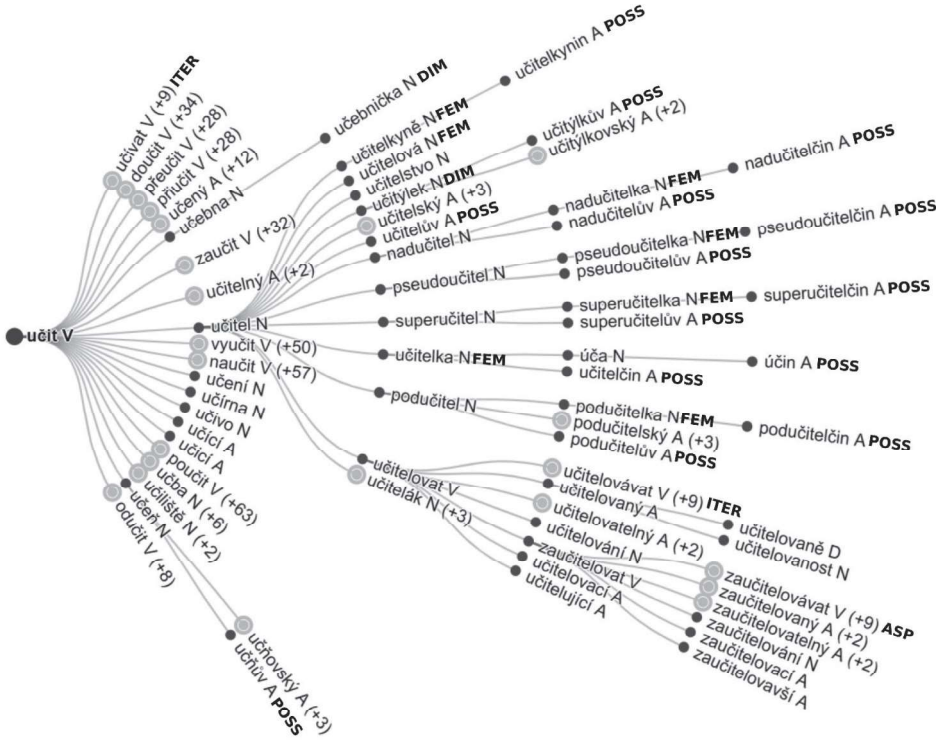

Fig. 3. The derivational tree with the root *učit* 'to teach' assigned with semantic labels predicted by the experiment. Semantic labels (in bold) are displayed with the derivative. Number of further derivatives, if hidden in the figure for clarity, is given in parentheses

Even though the presented labeling experiment was limited to a small number of semantic categories, its excellent results in terms of both precision and recall document, in our opinion, that the make-up of the DeriNet network (derivational families organized into rooted tree structures) and the features included in the machine learning experiment provided a sufficient basis for the resolution of homonymy in most cases.

A detailed analysis of incorrectly predicted labels draws attention to mostly peripheral or borrowed formations that are still formally and morphologically close to the correct representatives of the particular semantic categories but differ in meaning. See Table 7 for examples of relations that were incorrectly assigned the

FEMALE or the *DIMINUTIVE* label. Examples like *textař* 'lyricist' > *textařina* 'profession of a lyricist' point out the usefulness of animateness as a morphological feature of feminine nouns (not available in MorfFlex CZ) since only animate feminines are to be considered female counterparts of animate masculine nouns.

base word	derivative	incorrectly predicted label
ježek 'hedgehog'	*ježura* 'echidna'	*FEMALE*
fořt 'forest warden'	*fořtovna* 'forest warden's lodge'	*FEMALE*
profesor 'professor'	*profesura* 'professorship'	*FEMALE*
textař 'lyricist'	*textařina* 'profession of lyricist'	*FEMALE*
smrt 'death'	*smrtka* 'Death'	*DIMINUTIVE*
had 'snake'	*hadice* 'hose'	*DIMINUTIVE*

Tab. 7. Examples of pairs with incorrect labels

The labels *ASPECT* and *ITERATIVE* were not sufficient to cover a handful of examples in which a perfective verb is captured as a derivative of another perfective in DeriNet (e.g. *oloupat* 'to peel.PFV' > *oloupnout* 'to peel.PFV', *chytit* 'to catch.PFV' > *chytnout* 'to catch.PFV'). These relations correspond to the semelfactive semantic concept in Bagasheva's set; the respective label will be included in the next round of semantic labeling.

6 CONCLUSIONS

The semi-automatic procedure introducing semantic labels into the DeriNet network, which was described in the present paper, was carried out as a pilot experiment to verify its applicability to large, specifically organized data. The approach was limited to five semantic categories that are conveyed (mainly) by ambiguous suffixes and, with the exception of derivation of possessives, do not change the part-of-speech category of the base word. The fact that the assigned categories are rooted in the proposal of comparative semantic concepts might not be obvious in this pilot phase, as we chose basic categories that are involved not only in Bagasheva's proposal. However, the choice of a particular linguistic background is essential for perspectives of further development and usability of the data.

The labeling task started with extraction of relevant features from existing resources in order to compile high-quality training and test data sets with enough examples of each category in an efficient way. The machine learning model was designed with the aim to be replicable after any changes in the DeriNet data and to be extendable to other labels. More than 150 thousand semantic labels were predicted by the model, by achieving both an excellent precision and recall. Analysis of the data with both correctly and incorrectly predicted labels is expected to be relevant for our next steps as well as, importantly, for linguistic insights into derivations.

ACKNOWLEDGMENTS

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation and by the Student Faculty Grant UKMFF/160753/2018-2/SFG of the Faculty of Mathematics and Physics, Charles University. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project (LM2015071, LM2018101).

References

[1] Agresti, A. (2002). Categorical Data Analysis. 2nd edition. New York, John Wiley & Sons.
[2] Bagasheva, A. (2017). Comparative semantic concepts in affixation. In Competing Patterns in English Affixation, pages 33–65, Bern, Peter Lang.
[3] Dokulil, M. (1962). Tvoření slov v češtině: Teorie odvozování slov. Praha, ČSAV.
[4] Dokulil, M. et al. (1986). Mluvnice češtiny 1. Praha, Academia.
[5] Hajič, J., and Hlaváčová, J. (2013). MorfFlex CZ. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: http://hdl.handle.net/11858/00-097C-0000-0015-A780-9
[6] Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross-linguistic studies. Language, 86(3), pages 663–687.
[7] Hathout, N. (2010). Morphonette: a morphological network of French. CoRR, arXiv, abs/1005.3902.
[8] Hathout, N., and Namer, F. (2014). Démonette, a French derivational morpho-semantic network. Linguistic Issues in Language Technology, 11(5), pages 125–168.
[9] Havránek, B. (ed.; 1960–1971). Slovník spisovného jazyka českého. Praha, Academia.
[10] Hosmer, D. W., and Lemeshow, S. (2000). Applied Logistic Regression. 2nd edition. New York, John Wiley & Sons.
[11] Karlík, P. ed. (2016). Nový encyklopedický slovník češtiny. Praha, NLN.
[12] Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. Praha, ÚČNK FF UK. Accessible at: http://www.korpus.cz
[13] Kyjánek, L. (2018). Morphological Resources of Derivational Word-Formation Relations. Technical report no. 2018/TR-2018-61. Praha, ÚFAL MFF UK.
[14] Lopatková M. et al. (2016). VALLEX 3.0. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: http://hdl.handle.net/11234/1-2307
[15] Meľčuk, I. (2006). Explanatory Combinatorial Dictionary. In Open Problems in Linguistic and Lexicography, pages 225–355, Monza, Polimetrica.
[16] Mel'čuk, I., and Žolkovskij, A. K. (1984). Tolkovo-kombinatornyj slovar' russkogo jazyka. Vienna, Wiener Slawistische Almanach. Sonderband 14.
[17] Nekula, M. et al. (2012). Příruční mluvnice češtiny. 2nd edition. Praha, NLN.
[18] Osolsobě, K. et al. (2002). A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages. In Proceedings of LREC 2002, pages 1254–1259, Paris, ELRA.
[19] Pala, K., and Hlaváčková, D. (2007). Derivational Relations in Czech WordNet. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pages 75–81, Prague, ACL.

[20] Pala, K., and Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In International Conference on Text, Speech, and Dialogue, TSD 2015, pages 515–523, Berlin, Springer.
[21] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pages 2825–2830.
[22] Sedláček, R., and Smrž, P. (2001). A New Czech Morphological Analyser ajka. In International Conference on Text, Speech and Dialogue, TSD 2001, pages 100–107, Berlin, Springer.
[23] Straková et al. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In Proceedings of ACL 2014: System Demonstrations, pages 13–18.
[24] Ševčíková, M., and Panevová, J. (2018). Derivation of Czech verbs and the category of aspect. Linguistica Copernicana, 2018(15), pages 79–93.
[25] Ševčíková, M., and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In Proceedings of LREC 2014, pages 1087–1093, Paris, ELRA.
[26] Šimandl, J. ed. (2016). Slovník afixů užívaných v češtině. Praha, Karolinum.
[27] Štekauer, P. (2005). Meaning Predictability in Word Formation: Novel, context-free naming units. Amsterdam, John Benjamins.
[28] Štícha, F. et al. (2018). Velká akademická gramatika spisovné češtiny 1. Praha, Academia.
[29] Vidra, J. et al. (2018). DeriNet 1.7. Praha, ÚFAL MFF UK. Accessible at: `http://ufal.mff.cuni.cz/derinet`
[30] Vidra, J. et al. (2019). DeriNet 2.0. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: `http://hdl.handle.net/11234/1-2995`