Introducing Semantic Labels into the DeriNet Network

Magda Ševčíková and Lukáš Kyjánek

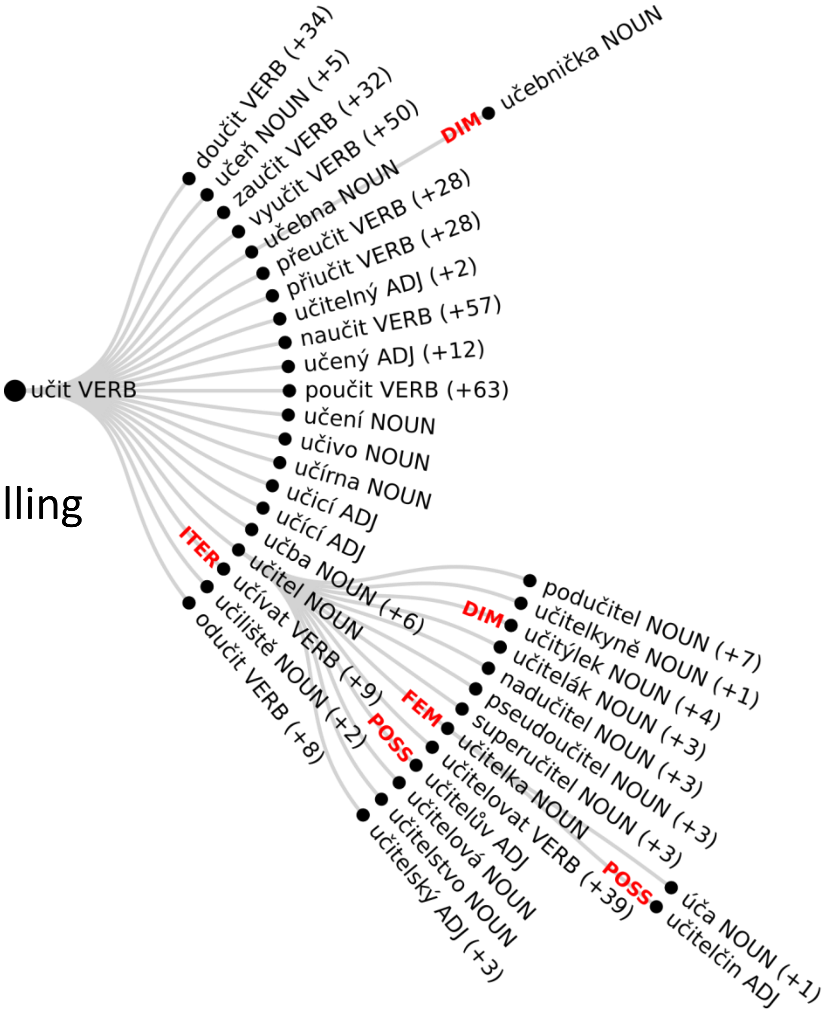
Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

SLOVKO 2019, 25th October

CHARLES UNIVERSITY

Outline

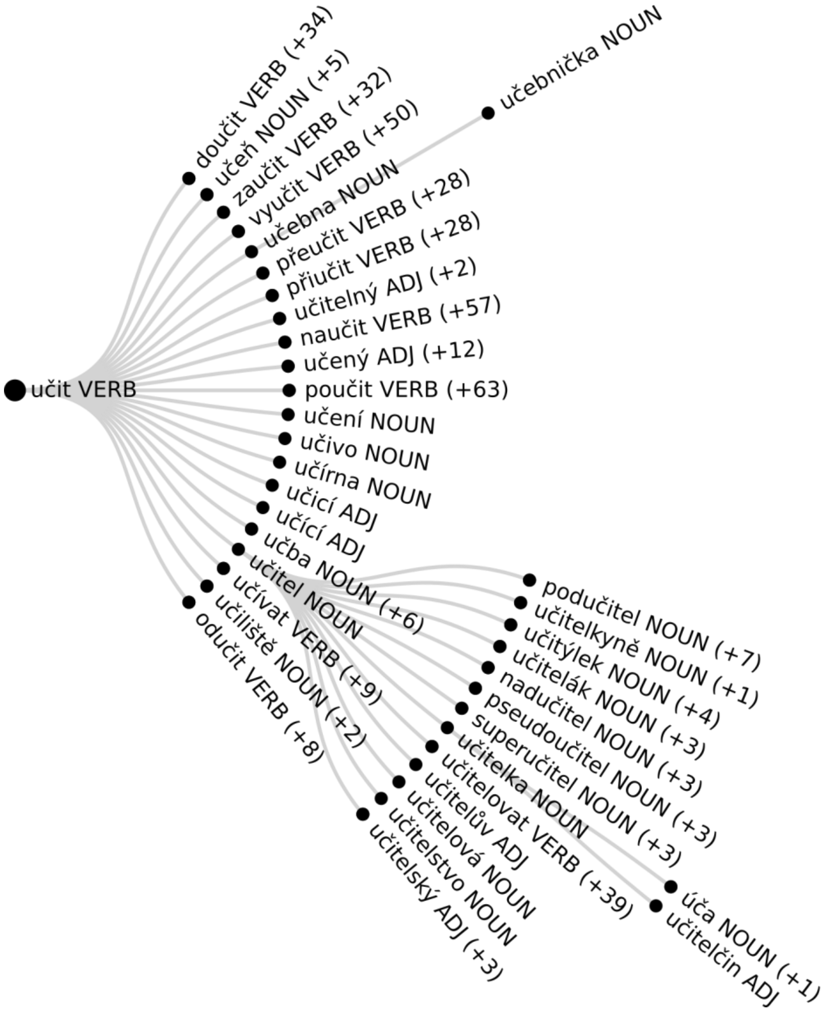
- DeriNet Network
- Derivation as a change of meaning
- The goal of the pilot experiment
- Machine Learning approach to semantic labelling
- Evaluation and conclusion



DeriNet Network

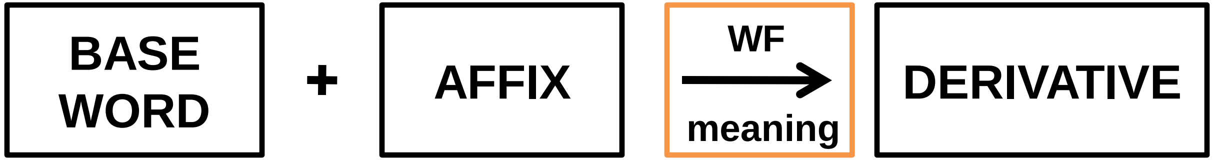
- A database of Czech lexemes (1M) connected by links (809k) corresponding to derivational relations (rooted tree)
- Developed since 2013; ufal.cz/derinet
- Organized according to morphemic and semantic complexity from the simplest to the most complex lexemes

- Semantic labeling experiments carried out on DeriNet 1.7 (Vidra et al. 2018)
- Resulting labels included into DeriNet 2.0 (Vidra et al. 2019) in LINDAT/CLARIAH CZ repository under the CC BY-NC-SA 3.0 http://hdl.handle.net/11234/1-2995



Derivation as a change of meaning

- Word-Formation meaning
  - The change in meaning that happens when attaching an affix to a base word
  - *pomoc-i* > *pomoc-ník* = 'a person who helps'
  - *poruč-i-t* > *poruč-ník* = 'a person who commands'
- Lexical meaning
  - The meaning of the word in the current use (listed in the dictionaries); may or may not be shifted
  - *pomocník* = 'helper'
  - *poručník* = 'legal guardian'



pomoc-i	*ník*	*pomocník*
poruč-i-t	*ník*	*poručník*

Homonymy and synonymy of affixes in Czech

- **Homonymy**: Formally identical affixes convey more than one meaning.

-ka:	*skříň > skříň-ka*	*cupboard > small cupboard*	*diminutive*
	učitel > učitel-ka	*male teacher > female teacher*	*female noun*
	obal-i-t > obál-ka	*to wrap > envelope*	*instrument noun*

- **Synonymy**: A particular meaning is expressed by several, formally different affixes.

female noun:	*hráč > hráč-ka*	*male player > female player*
	ministr > ministr-yně	*male minister > female minister*
	šéf > šéf-ová	*male boss > female boss*

Semantically labeled derivational resources

- Addressed in more or less explicit way in:
  - Derivancze for Czech (17 labels; Pala & Šmerk 2015)
  - CroDeriV for Croatian (14 labels; Filko et al. 2019)
  - Database from English WordNet (14 labels; Fellbaum et al. 2007)
  - Démonette for French (4 labels; Hathout & Namer 2014)

The goal

- Long-term goal:

To add explicit semantic labels
(based on semantic comparative concepts by Bagasheva, 2017)
into DeriNet

- Challenges:
  - 1M+ lexemes in DeriNet, derivational links still added/deleted
  - Homonymy and synonymy of affixes

- Pilot experiment:

A semi-automatic Machine Learning procedure limited to
five semantic categories when conveyed by suffixation

Selected semantic labels

• ***DIMINUTIVE***	*pes > psík*	*dog > small dog*
	žlutý > žluťoučký	*yellow > yellowish*
• ***FEMALE***	*učitel > učitelka*	*teacher > female teacher*
	Jaroslav > Jaroslava	*(male first name) > (fem. first name)*
	Novák > Nováková	*(male surname) > (fem. surname)*
• ***POSSESSIVE***	*učitel > učitelův*	*teacher > teacher’s*
	učitelka > učitelčin	*female teacher > female teacher’s*
• ***ITERATIVE***	*chodit > chodívat*	*to walk (IPFV) > to walk repeatedly (IPFV)*
	kupovat > kupovávat	*to buy (IPFV) > to buy repeatedly (IPFV)*
• ***ASPECT***	*obalit > obalovat*	*to wrap (PFV) > to wrap (IPFV)*
	štěkat > štěknout	*to bark (IPFV) > to bark (PFV)*

Machine Learning procedure

DATA → FEATURES → ML MODEL

- 14,752 base-derivative pairs extracted from DeriNet 1.7
- negative examples extracted also from DeriNet 1.7
  - their amount determined empirically

Machine Learning procedure

DATA → FEATURES → ML MODEL

- **Semantic label**
  - For: base-derivative pair
  - From:
    Slovník spisovného jazyka českého (Havránek 1960-1971)
    Morphological dict. MorfFlex CZ (Hajič & Hlaváčová 2013)
    The valency lexicon VALLEX 3.0 (Lopatková et al. 2016)
    Příruční mluvnice češtiny (Nekula et al. 2012)

pes.N > psík.N	***DIMINUTIVE***
učitel.N > učitelka.N	***FEMALE***
učitel.N > učitelův.A	***POSSESSIVE***
chodit.V > chodívat.V	***ITERATIVE***
obalit.V > obalovat.V	***ASPECT***

Machine Learning procedure

DATA → FEATURES → ML MODEL

- **Part-of-speech category**
  - For: the base word and derivative
  - From: DeriNet

- **Gender**
  - For: the base word and derivative (nouns only)
  - From: MorfFlex CZ

*pes.**N.m_anim*** > *psík.**N.m_anim***

*učitel.**N.m_anim*** > *učitelka.**N.fem***

*učitel.**N.m_anim*** > *učitelův.**A***

*chodit.**V*** > *chodívat.**V***

*obalit.**V*** > *obalovat.**V***

Machine Learning procedure

DATA → FEATURES → ML MODEL

- **Aspect**
  - For: the base word and derivative (verbs only)
  - From: MorfFlex CZ, SYN2015, VALLEX

- **Possessivity tag**
  - For: the derivative (adjectives only)
  - From: MorfFlex CZ

pes.N > psík.N

učitel.N > učitelka.N

*učitel.N > učitelův.A.**pos_true***

*chodit.V.**ipfv** > chodívat.V.**ipfv***

*obalit.V.**pfv** > obalovat.V.**pfv***

Machine Learning procedure

DATA → FEATURES → ML MODEL

- **Final n-grams**
  - For: the base word and derivative
  - Bi-, tri-, tetra-, penta-, hexa-grams

učitel > učitelův

*-el, **-tel**, -itel, -čitel, -učitel*

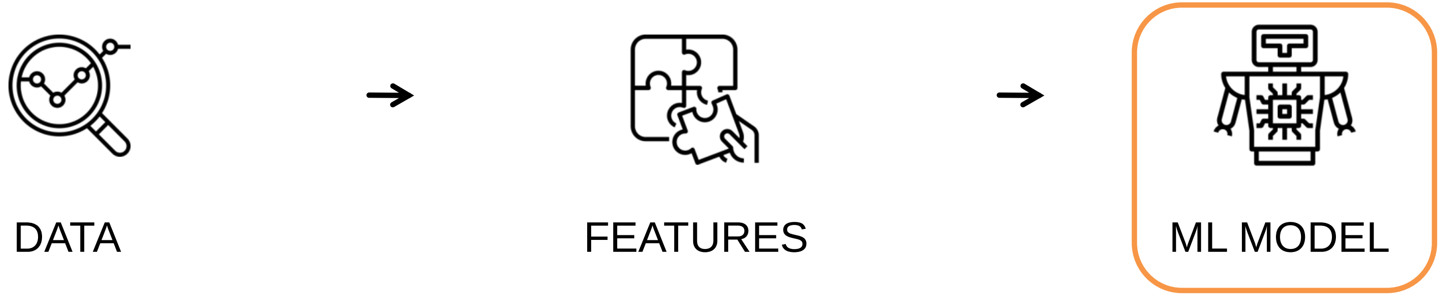
***-ův**, -lův, -elův, -telův, -itelův*

obalit > obalovat

***-it**, -lit, -alit, -balit, -obalit*

*-at, -vat, **-ovat**, -lovat, -alovat*

Machine Learning procedure



DATA → FEATURES → ML MODEL

- Classifying the most probable semantic label (highest possible precision)
- Multinomial Logistic Regression (MLR) with newton-cg solver

- Model trained on *training data set (80 %)*
- Probability thresholds adjusted according to *development data set (10 %)*
- Tested on *testing data set (10 %)*

Machine Learning procedure

DATA → FEATURES → ML MODEL

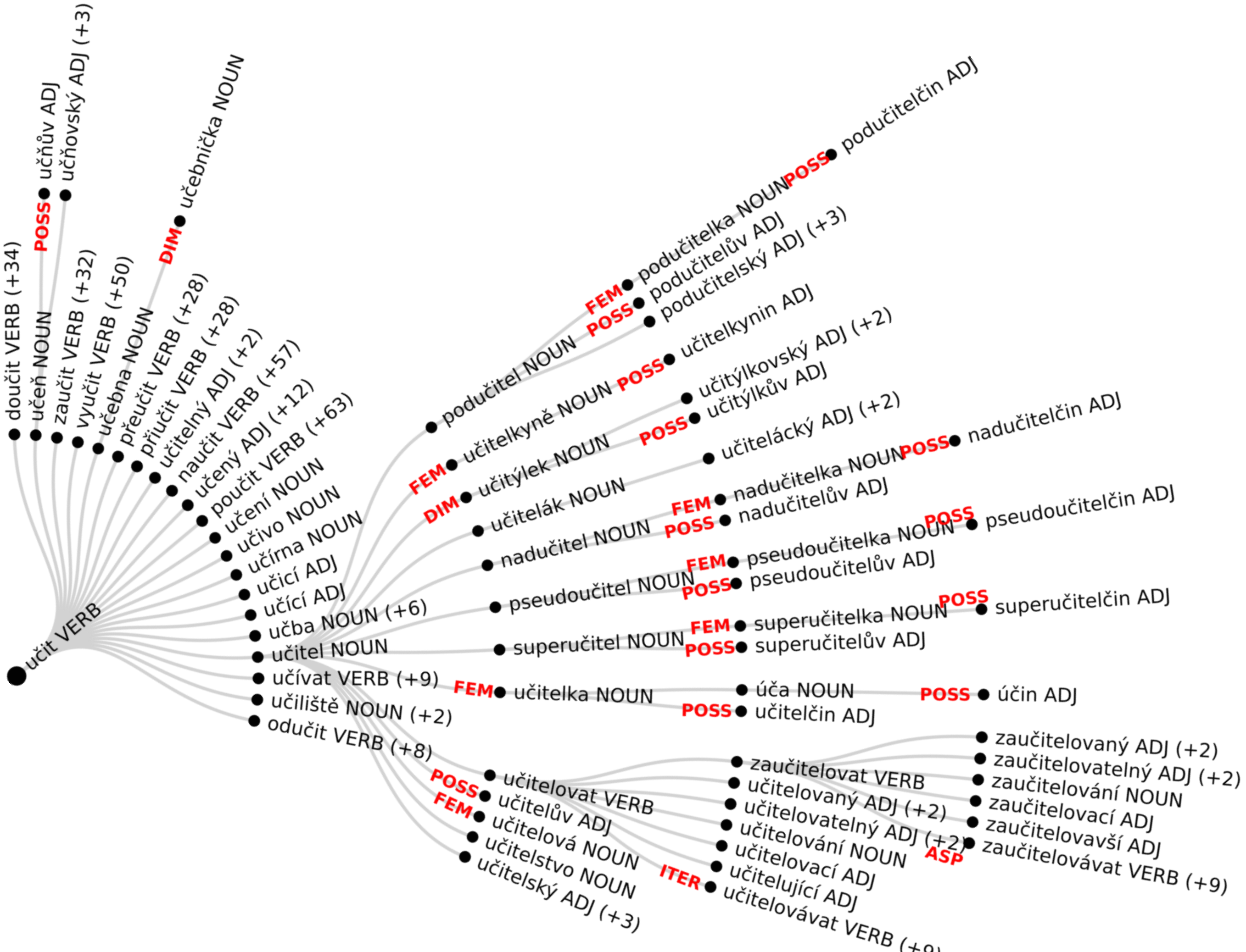
Evaluation on the *testing data set*

	Accuracy	Precision	Recall	F1-score
Baseline	0.827	0.813	0.827	0.792
MLR model	**0.986**	**0.984**	**0.984**	**0.984**

Applying the MLR model

- *Predicted data* = base-derivative relations from DeriNet 1.7
- 150,521 relations assigned one of the five semantic labels

DIMINUTIVE	***FEMALE***	***POSSESSIVE***	***ITERATIVE***	***ASPECT***
6,042	28,510	88,620	11,890	15,459


učit VERB
doučit VERB (+34)
učeň NOUN
POSS
učňův ADJ
učňovský ADJ (+3)
zaučit VERB (+32)
vyučit VERB (+50)
učebna NOUN
DIM
učebnička NOUN
přeučit VERB (+28)
přiučit VERB (+28)
učitelný ADJ (+2)
naučit VERB (+57)
učený ADJ (+12)
poučit VERB (+63)
učení NOUN
učivo NOUN
učírna NOUN
učicí ADJ
učící ADJ
učba NOUN (+6)
učitel NOUN
učívat VERB (+9)
učiliště NOUN (+2)
odučit VERB (+8)
podučitel NOUN
FEM
podučitelka NOUN
POSS
podučitelčin ADJ
POSS
podučitelův ADJ
podučitelský ADJ (+3)
FEM
učitelkyně NOUN
POSS
učitelkynin ADJ
DIM
učitýlek NOUN
učitýlkovský ADJ (+2)
POSS
učitýlkův ADJ
učitelák NOUN
učitelácký ADJ (+2)
naučitel NOUN
FEM
naučitelka NOUN
POSS
naučitelčin ADJ
POSS
naučitelův ADJ
pseudoučitel NOUN
FEM
pseudoučitelka NOUN
POSS
pseudoučitelčin ADJ
POSS
pseudoučitelův ADJ
superučitel NOUN
FEM
superučitelka NOUN
POSS
superučitelčin ADJ
POSS
superučitelův ADJ
FEM
učitelka NOUN
úča NOUN
POSS
účin ADJ
POSS
učitelčin ADJ
učitelovat VERB
zaučitelovat VERB
zaučitelovaný ADJ (+2)
zaučitelovatelný ADJ (+2)
zaučitelování NOUN
zaučitelovací ADJ
zaučitelovavší ADJ
ASP
zaučitelovávat VERB (+9)
učitelovaný ADJ (+2)
učitelovatelný ADJ (+2)
učitelování NOUN
učitelovací ADJ
učitelující ADJ
ITER
učitelovávat VERB (+9)
POSS
učitelův ADJ
FEM
učitelová NOUN
učitelstvo NOUN
učitelský ADJ (+3)

Conclusion

- High precision and recall indicate that features selected for the ML approach were able to solve the homonymy/synonymy of affixes in most cases.

- Examples of pairs with <u>incorrect</u> labels:

ježek > ježura	*hedgehog > echidna*	***FEMALE***
profesor > profesura	*professor > professorship*	***FEMALE***
smrt > smrtka	*death > Death*	***DIMINUTIVE***

- Analysis of the data with predicted labels is expected to be relevant for our next steps as well as for linguistic insights into derivations.
- Semantic labels were included in the current version of DeriNet 2.0 and can be used for searching the data by the DeriSearch tool.

Semantic labels available in DeriNet 2.0

in LINDAT/CLARIAH CZ repository
under the **CC BY-NC-SA 3.0**

http://hdl.handle.net/11234/1-2995

or ufal.cz/derinet

Acknowledgement

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

- Agresti, A. (2002). *Categorical Data Analysis. 2nd edition. New York: John* Wiley & Sons.
- Bagasheva, A. (2017). Comparative semantic concepts in affixation. In *Competing Patterns in English Affixation, 33–65, Bern: Peter Lang.*
- Dokulil, M. (1962). *Tvoření slov v češtině: Teorie odvozování slov.* Prague: ČSAV.
- Dokulil, M. et al. (1986). *Mluvnice češtiny 1.* Prague: *Academia.*
- Hathout, N., Namer, F. (2014). Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology,* 11, 125–162.
- Fellbaum, Ch., Osherson, A., Clark, P. E. (2007). Putting Semantics into WordNet's "Morphosemantic" Links. In Language and Technology Conference. Springer, 350–358.
- Filko, M., Šojat, K., Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology, 71–80. Prague: Charles University.
- Hajič, J., Hlaváčová, J. (2013). *MorfFlex CZ. LINDAT/CLARIN digital library* at ÚFAL MFF UK, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.
- Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross-linguistic studies. *Language, 86(3), 663–687.*
- Havránek, B. (ed.; 1960–1971). *Slovník spisovného jazyka českého.* Prague: Academia.
- Karlík, P. et al. (ed.; 2017). *Nový encyklopedický slovník češtiny*. Prague: *NLN.*
- Křen, M. et al. (2015). *SYN2015: reprezentativní korpus psané češtiny.* Prague: ÚČNK FF UK, http://www.korpus.cz.
- Lopatková M. et al. (2016). *VALLEX 3.0. LINDAT/CLARIN digital library at* ÚFAL MFF UK, http://hdl.handle.net/11234/1-2307.
- Nekula, M. et al. (2012). *Příruční mluvnice češtiny. 2nd edition.* Prague: *NLN.*

- Pala, K., Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In International Conference on Text, Speech, and Dialogue, TSD 2015, 515–523, Berlin: Springer.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12, 2825–2830.*
- Straková et al. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of ACL 2014: Systém Demonstrations, 13–18.*
- Ševčíková, M., Žabokrtský, Z. (2014). Word-Formation Network for Czech. In *Proceedings of LREC 2014, 1087–1093, Paris: ELRA.*
- Šimandl, J. ed. (2016). *Slovník afixů užívaných v češtině.* Prague: *Karolinum.*
- Vidra, J. et al. (2018). *DeriNet 1.7.* Prague: *ÚFAL MFF UK,* http://ufal.mff.cuni.cz/derinet.
- Vidra, J. et al. (2019). *DeriNet 2.0. LINDAT/CLARIN digital library at ÚFAL* MFF UK, http://hdl.handle.net/11234/1-2995.