Derivational meaning in language resources

Lukáš Kyjánek

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

UFAL PhD conference
May 3, 2021

Outline

- Derivation and derivational meaning
- Contemporary language resources for word-formation
- Labelling derivational meanings
  - Supervised machine-learning experiment
  - Unsupervised machine-learning experiment
- Empirical study based on the labelled data
- Challenges in the labelling

Derivational meaning

- *odesílat* $\xrightarrow[\text{agent}]{}$ *odesíla-tel* *(to send > sender)*
  - *odesílat* = activity
  - *odesílatel* = someone who does the activity

- One affix can convey many meanings
  - *úředník* $\xrightarrow[\text{female}]{}$ *úředn-ice* *(officer > female officer)*
  - *věznit* $\xrightarrow[\text{location}]{}$ *vězn-ice* *(to imprison > jail)*
  - *kytka* $\xrightarrow[\text{augmentative}]{}$ *kyt-ice* *(flower > bouquet)*

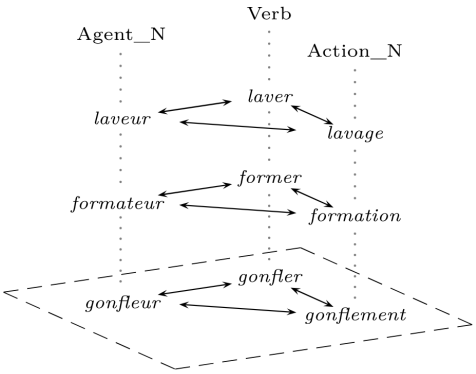
- One meaning can be conveyed by many affixes
  - *úředník* $\xrightarrow[\text{female}]{}$ *úředn-ice* *(officer > female officer)*
  - *šéf* $\xrightarrow[\text{female}]{}$ *šéf-ová* *(boss > female boss)*
  - *učitel* $\xrightarrow[\text{female}]{}$ *učitel-ka* *(teacher > female teacher)*
  - *ministr* $\xrightarrow[\text{female}]{}$ *ministr-yně* *(minister > female minister)*

Approaches to derivational morphology

Körtvélyessy et al. (2020:10-11)

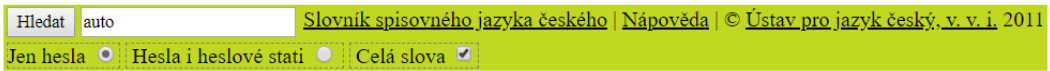
1. Direct derivatives (paradigm)
   dom → *dom-ov*
   → *dom-ček*
   → *dom-ík*
   → *dom-isko*
2. Subsequent derivatives (series)
   dom → *dom-ov* → *dom-ov-ina* → *dom-ov-in-ový*
   dom → *dom-ček* → *dom-ček-ový*
   dom → *dom-ík* → *dom-ík-ový*
   dom → *dom-isko* → *dom-isk-ový*
3. Semantic categories of each derivational step
   agent, female, location, quality, agmentative, etc.
4. Derivational network
   = derivatives derived from a simple underived word
   (combination of (1) and (2) and (3))

Bonami and Strnadová (2019:172)



Related work

- The universal set of labels utilisable across languages is not defined;
  - but there are some proposals of comparative concepts, e.g., Bagasheva (2017).
- Specialised resources for word-formation usually lack explicit labels.
  - Derivancze for Czech — 17 labels; (Pala and 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 and Namer 2014)
- Some pieces of information on derivational meaning occur in resources that cover primarily other phenomena, or in explanatory dictionaries;
  - so the information can be extracted and exploited.



auto, -a s. *automobil:* osobní, nákladní, sanitní a.; vojenské, pancéřové a.; rozhlasové, televizní a.; přijet autem; sednout do auta; vystoupit z auta; → **zdrob. autíčko**, -a s. (6. mn. -ách): dětské, šlapací a.; — autový přid.: a-á doprava *automobilová*

Meaning-Text Theory

- Paradigmatic lexical functions; as mathematical function $f(x) = y$
- e.g., function S_0 for substantivisation $S_0(\textit{analyzovat.VERB}) = \textit{analýza.NOUN}$
- Melčuk (1981), Wanner (1966), Apresjan (2011)

Functional Generative Description & Prague Dependency Treebank

- Tectogrammatical layer, t-lemmas; nodes t-lemmas substitute morphological lemmas from the surface level
- e.g., *matčin* $\xleftarrow[\textit{possessive}]{}$ *matka*
- e.g., *pěkně* $\xleftarrow[\textit{deadj. adv.}]{}$ *pěkný*
- Sgall (1964), Mikulová et al. (2006), Hajič et al. (2020)

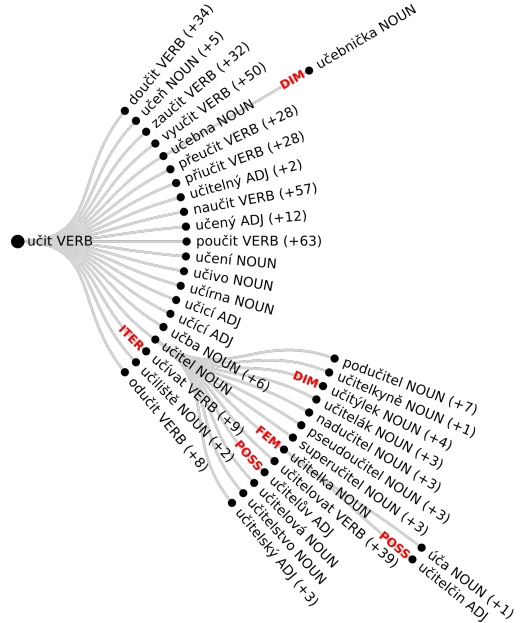
Supervised labelling of derivational meanings (Ševčíková and Kyjánek 2019)

- Pilot experiment: to add 5 labels limited to suffixation into DeriNet for Czech
  - *pes* $\xrightarrow[\text{diminutive}]{}$ *psík* *(dog > small dog)*
  - *učitel* $\xrightarrow[\text{female}]{}$ *učitelka* *(teacher > female teacher)*
  - *učitel* $\xrightarrow[\text{possessive}]{}$ *učitelův* *(teacher > teacher's)*
  - *chodit* $\xrightarrow[\text{iterative}]{}$ *chodívat* *(to walk (IPFV) > to walk repeatedly (IPFV))*
  - *obalit* $\xrightarrow[\text{aspect}]{}$ *obalovat* *(to wrap (PFV) > to wrap (IPFV))*
- Input data: 14,752 semantically labelled base-derivative pairs from SSJČ (Havránek 1960-1971), MorfFlexCZ (Hajič and Hlaváčová 2013), VALLEX 3.0 (Lopatková et al. 2016), and PMČ (Nekula et al. 2012); each label around 2.5 thousand pairs
- Features: part-of-speech categories, genders, aspects, possessivity tags, final character n-grams (2-6)

- Task: to classify the most probable semantic label
- Method: Multinomial Logistic Regression with newton-cg solver
- F1-score = 98.4%

Label	Derivations
Diminutive	5,383
Female	28,623
Possessive	87,087
Iterative	11,778
Aspect	15,186

- Already available since DeriNet 2.0



Unsupervised labelling of derivational meanings

- Plan: to label base lexemes from which female nouns are derived
  - *pekařka.NOUN* ←female— *pekař.NOUN* ←agent— *péci.VERB*
  - *vesničanka.NOUN* ←female— *vesničan.NOUN* ←dweller— *vesnice.NOUN*
  - *obžalovaná.NOUN* ←female— *obžalovaný.NOUN* ←experiencer— *obžalovaný.ADJ*
  - *adresátka.NOUN* ←female— *adresát.NOUN* ←patient— *adresovat.VERB*

- Two unsupervised approaches:
  1. Hierarchical clustering based on a given set of features
  2. Word embeddings combined with clustering based on distances of base-derivative pairs in the vector space

Competition in derivational morphology (Ševčíková, Kyjánek and Hladká 2021)

Agent noun formation (suffix rivalry)

- 8 top-frequent suffixes forming agent nouns (SYN2015); manually created data
- Data set divided into training, evaluation, and hold-out subsets
- Settings of hyper-parameters of Logistic regression were obtained from the first experiment on dataset containig all features
- Other experiments used 5 different subsets of features, but the same settings

target_noun	***viník***	**target_noun_suffix**	***-ník|-ík***
base_number_syllables	*1*	paradigm_type	*NNA-V-*
base_number_prefixes	*0*	freq_target_noun	*1188*
base_shared_theme	*x*	freq_parent_noun	*6758*
base_ending	*n*	freq_parent_adj	*2274*
base_ending_cvs	*consonant*	freq_parent_oth	–
base_ending_vertical	*nasal*	freq_parent_v1	*689*
base_ending_horizontal	*alveolar*	freq_parent_v2	–
parent_noun	*vina*	freq_slots	*VxAN*
parent_adj	*vinný*	v1_theme	*i*
parent_oth	–	v1_aspect	*imp*
parent_v1	*vinit*	v1_conjug	*4*
parent_v2	–	v2_theme	–
inanim_noun	*no*	v2_aspect	–
v1_suf_asp_counterpart	*no*	v2_conjug	–

Table: Absolute numbers of individual agent suffixes in our data set.

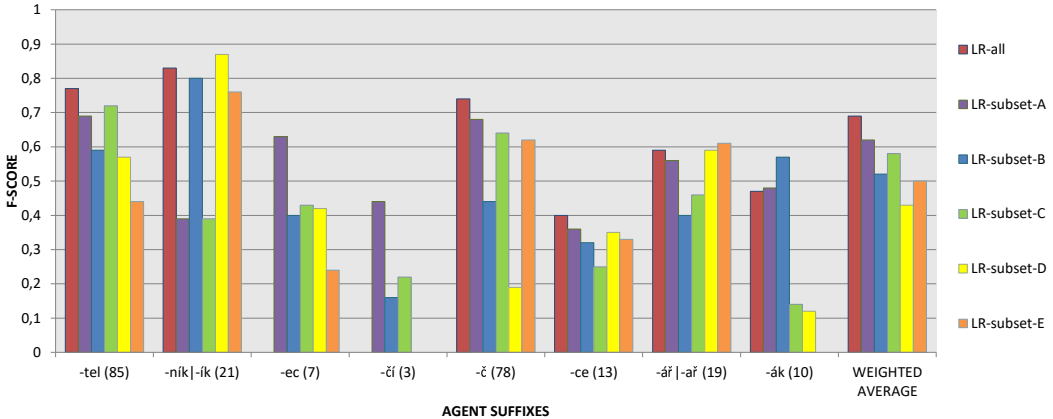
Suffix	*-tel*	*-č*	*-ník|-ík*	*-ář|-ař*	*-ce*	*-ák*	*-ec*	*-čí*	TOTAL
Count	426	388	106	96	66	50	32	14	**1,178**

Subsets

- Subset A: formal characteristics
- Subset B: phonological characteristics
- Subset C: morphological characteristics
- Subset D: morphological family characteristics
- Subset E: quantitative characteristics

Examples of results

- There must be more relevant features not included
- The combination of features from different linguistic areas is necessary to model competition
- Results of *-ář|-ař* and *-ce* seems relatively balanced: instances are likely complex regarding competition



Challenges

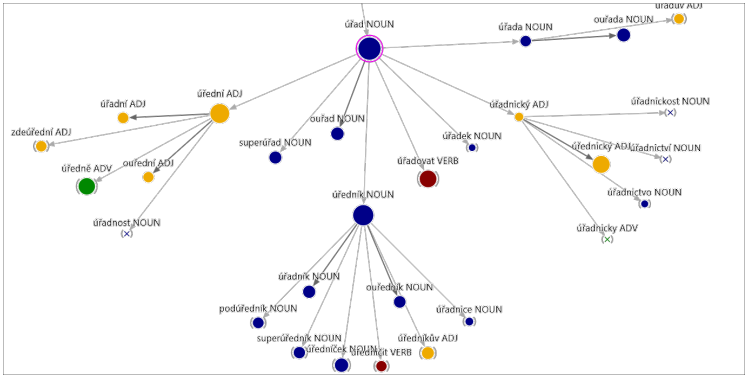
- Lemmatisation: inflection *vs.* derivation
  - Causes systematic differences in sets of lemmas across languages
  - e.g., *lexical negation* in Czech (inflectional) and in Russian (derivational)
  - Complicates analyses, especially cross-linguistic ones
- Spelling variants
  - Causes an increase in the number of forms (lemmas)
  - e.g., *radioizotop vs. radioisotop* (*-s-/-z-*)
  - Complicates both labelling and the subsequent analyses

Spelling variants in Czech (available as *DeriNet DEVEL* in DeriSearch v2)

- Identified 50,581 relations of spelling variants in DeriNet
- Extracted examples from SSJČ, MorfFlexCZ, VALLEX
- *n*-sets of spelling variants were found and their representative forms were selected using regular expressions and manual annotations

Examples:

- ***ú**řad*, ***ou**řad*
- *předhřát*, *pře**de**hřát*
- *ohra**ž**ování*, *ohra**z**ování*
- *jakkoli**v***, *jakkoli*
- *doping**g**ový*, *dopin**k**ový*
- *bý**k***, *be**j**k*
- *Tcha**iv**an*, *Tcha**jw**an*
- *odbydl**et***, *odbydl**it***
- ***troj**nožka*, ***tří**nožka*
- ***ž**ebřina*, ***ř**ebřina*
- *berl**a***, *berl**e***
- (?) *bezkolej**ný***, *bezkolej**ový***
- (?) *bezhlesý*, *bezhles**n**ý*
- (?) *bleďou**č**ký*, *bleďou**n**ký*
- (?) *bočn**ý***, *boční*
- (?) *drobí**n**ek*, *drobí**t**ek*
- (?) *rozechvě**n**ý*, *rozechvě**l**ý*



Main goals

- Formalising derivational meanings
  - To find any other data resources capturing derivational meanings
  - To analyse sets of derivational meanings labelled in the existing resources
  - To compare granularity of labels/meanings given by linguists and ML methods
- Labelling of derivational meanings in language resources
  - To label derivational meaning using supervised methods
  - To investigate possibilities of unsupervised methods in the labelling
  - To uncover challenges related to the labelling of derivational meanings
- Studying derivational meanings and their competition across languages
  - To measure influence of features for modelling derivational meanings
  - To observe competition of affixes within the same derivational meaning

References I

- Apresjan, Ju. D. 2011. K novoj versii teorii leksičeskich funkcij (LF). In: *Meždunarodnaja konferencija, posvjaščennaja 50-letiju Peterburskoj tipologičeskoj školy*, 21–26.
- Bonami, O., Strnadová, J. 2019. Paradigm Structure and Predictability in Derivational Morphology. *Morphology*, 29, 167-197. Springer. ISSN: 1871-5656.
- Fellbaum, Ch., Osherson, A., Clark, P. E. 2007. Putting semantics into WordNet's" morphosemantic" links. In: *Language and Technology Conference*, 350–358. Springer.
- Filko, M., Šojat, K., Štefanec V. 2019. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistics*, 115, 83-104. ISSN: 0032-6585.
- Hajič, J. et al. 2020. Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0). Data/Software, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Hajič, J., Hlaváčová, J. 2013. MorfFlex CZ. Data/Software, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Hathout, N., Namer, F. 2014. Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology*, 11, 125-162.
- Havránek, B. (ed.). 1960–1971. Slovník spisovného jazyka českého. Praha, Academia.
- Körtvélyessy, L., Bagasheva, A., Štekauer, P. 2020. *Derivational Networks Across Languages.* De Gruyter Mouton. ISBN: 9783110686494.
- Lopatková M. et al. 2016. VALLEx 3.0. Data/Software, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

References II

- Meľčuk, I. A. 1981. Meaning-Text Models: A Recent Trend in Soviet Linguistics. *Annual Review of Anthropology*, 10, 27–62.
- Mikulová, M. et al. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. URL: `https://ufal.mff.cuni.cz/pcedt2.0/publications/t-man-en.pdf`
- Nekula, M. et al. 2012. Příruční mluvnice češtiny. 2nd edition. Praha, NLN.
- Pala, K., Šmerk, P. 2015. Derivancze—Derivational Analyzer of Czech. In: *International Conference on Text, Speech, and Dialogue*, 515-523. Springer.
- Sgall, P. 1964. Zur Frage der Ebenen in Sprachsystem. TLP 1, 95–106.
- Ševčíková, M., Kyjánek, L. 2019. Introducing Semantic Labels into the DeriNet Network. *Journal of Linguistics*. Bratislava: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, 412-423. ISSN: 0021-5597.
- Ševčíková, M., Kyjánek, L., Hladká, B. 2021 in press. Agent noun formation in Czech: An empirical study on suffix rivalry. In: Conference Paradigmo.
- Vidra et. al. 2019. DeriNet. 2.0. Data/Software, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Wanner, L. (ed.). 1996. *Lexical Functions in Lexicography and Natural Language Processing*.