Word-formation structure of Czech words: a data-based research

Czech Science Foundation Grant Nr. GA19-14534S
2019–2021

Magda Ševčíková (PI), Jarmila Panevová, Zdeněk Žabokrtský,
Jonáš Vidra, Lukáš Kyjánek, Ruda Rosa,
Šárka Dohnalová, Adéla Kalužová

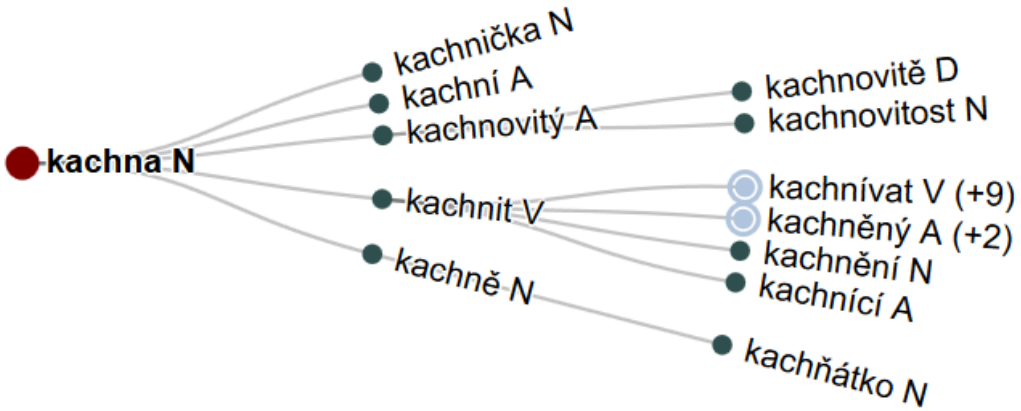
ÚFAL Seminar, Pec pod Sněžkou
September 16, 2019

Current tasks

- modelling derivations of Czech
  - DeriNet 2.0 data format
  - morphemic segmentation of Czech
  - semantic labelling of derivational relations
- derivation across languages
  - Universal Derivations
  - supervised morphological segmentation, improving the language model of fastText using morphological information
  - unsupervised lemmatization
- linguistic research
  - loan words in the word-formation system of Czech
  - direction of derivation in suffixless nouns
  - grammatical aspect in verbs with native and loan bases

DeriNet 2.0 format (Jonáš Vidra et al.)

- DeriNet database
  - derived lexemes linked with their base lexemes
    - 1M lexemes linked by 800k relations
  - lexemes with the same root morpheme organized into rooted trees
    - the simplest (unmotivated) lexeme as the root of the tree
    - the most complex ones in the leaves

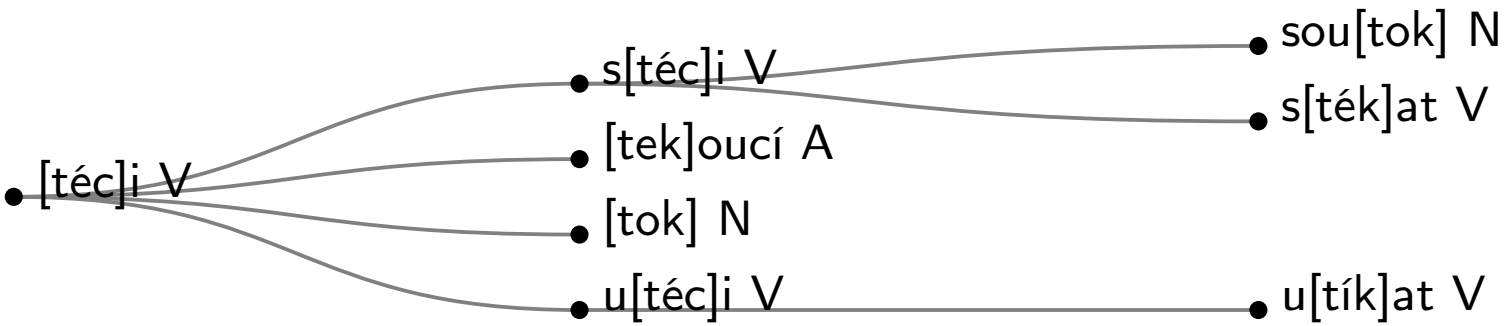


- new features in DeriNet 2.0
  - + morphological categories
  - + compounding
  - + morphemic segmentation
  - + semantic labelling

Morphemic segmentation

(Jonáš Vidra, Zdeněk Žabokrtský, Šárka Dohnalová, MŠ)

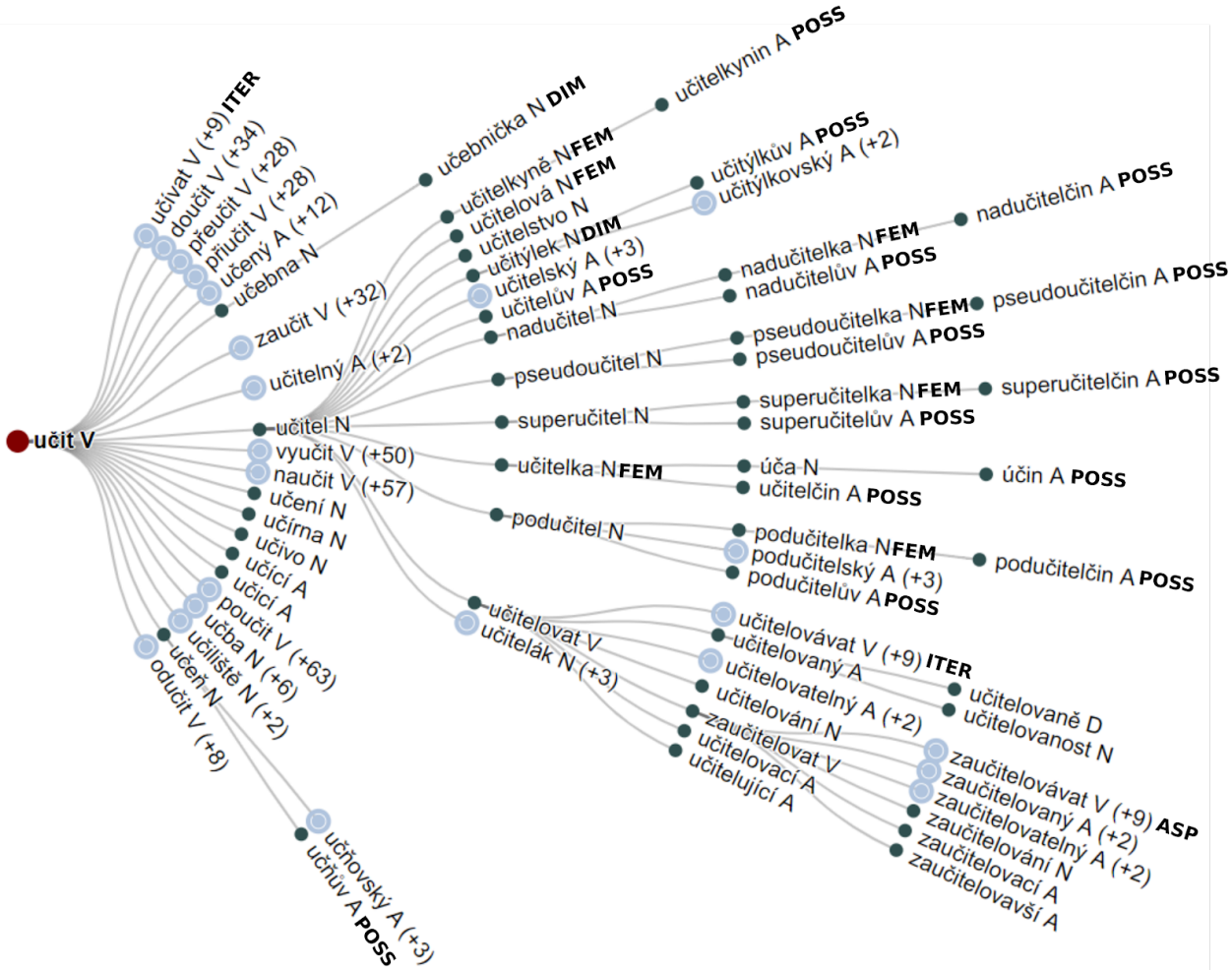
- dividing a word into a sequence of segments corresponding to morphemes
- complicated by allomorphy
  - cf. 8 root allomorphs in lexemes related to the verb *jíst*
    - *jís, jíd, jed, níd, nís, něd, jez, něz*
- delimitation of roots
  - semi-automatic identification of roots in 760 biggest trees
  - roots annotated in 250k lexemes in DeriNet 2.0



Semantic labelling of derivational relations (Lukáš Kyjánek, MŠ)

- How is the meaning of the base lexeme modified by attaching an affix?
- complicated by homonymy (polyfunctionality) and synonymy of affixes
  - *vlnka, hráčka, mluvka, skládka*
  - *učitelka, kolegyně, lékárnice, švagrová*
- 5 semantic labels to assign (Bagasheva 2017)
  - DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE, ASPECT
- labels assigned with 150k relations across the DeriNet 2.0 data by a Machine Learning experiment

Semantic labelling

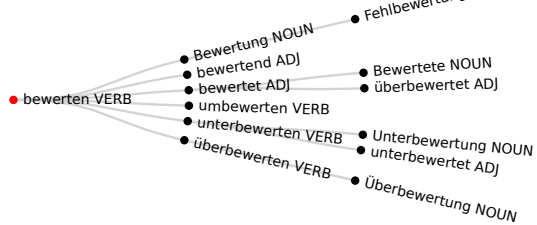


Universal Derivations (UDer) (Lukáš Kyjánek et al.)

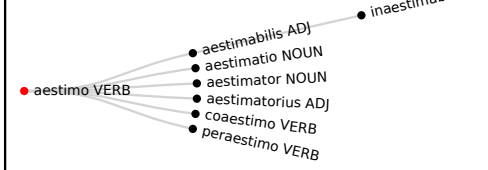
- a collection of harmonized derivational resources
  - admittedly imitative title
- target format: DeriNet 2.0 format
- release in Lindat repository
- currently 11 resources for 11 different languages
  - Démonette 1.2 (French)
  - DeriNet 2.0 (Czech)
  - DeriNet.ES (Spanish)
  - DeriNet.FA (Persian)
  - DErivBase 2.0 (German)
  - English WordNet 3.0 (English)
  - EstWordNet 2.1 (Estonian)
  - FinnWordNet 2.0 (Finnish)
  - Nomlex-PT 2017 (Portuguese)
  - Polish WFN (Polish)
  - Word Formation Latin (Latin)

DeriNet

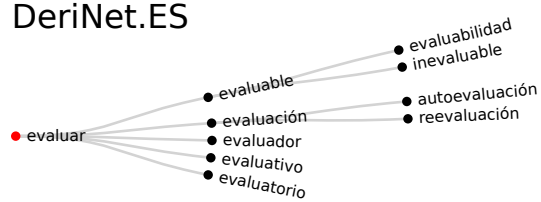
DErivBase



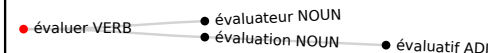
Word Formation Latin



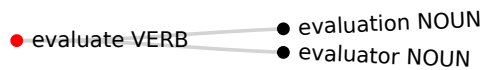
DeriNet.ES



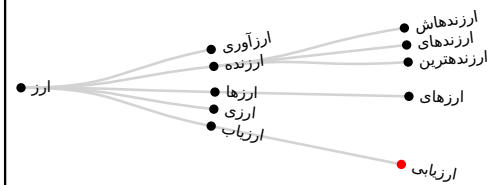
Démonette



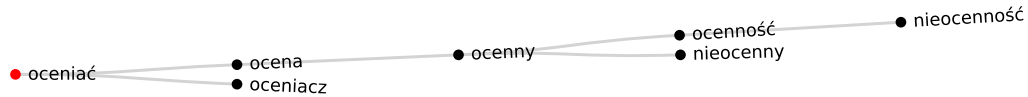
English WordNet



DeriNet.FA



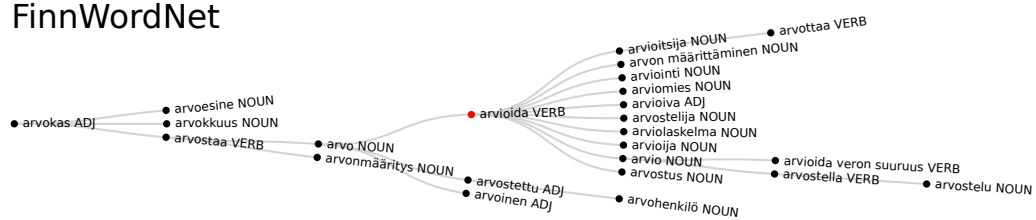
The Polish Word-Formation Network



EstWordNet

NomLex-PT

FinnWordNet



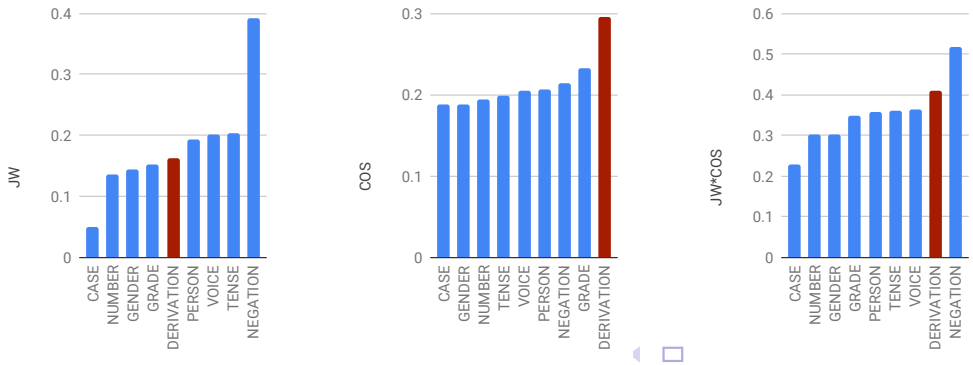
Supervised morphological segmentation, fastText improvement (Ebrahim Ansari et al.)

- manually creating a rich morphological segmentation lexicon for Persian
  - implementation and evaluation of various algorithms for supervised morphological segmentation on Persian, Czech and Finnish
  - morphological tree construction using segmented lexicon (supervised and unsupervised segmentation)

- fastText improvement using morphological information
  - extracting highest significant segments using fastText
  - considering those strong segments as unigrams in fastText algorithm
  - designing a bootstrapping strategy to find and insert those atomic subwords gradually

Unsupervised lemmatization (Ruda Rosa, Zdeněk Žabokrtský)

- original motivation
  - For studying derivation across languages, we need their lemma sets...
  - ... but what we see in the corpora are inflected forms.
- Can we separate inflection from derivation without annotated data?
- intuition: inflected forms are in some sense closer to each other
- experiment: distance of word forms approximated as a combination of
  - Jaro-Winkler edit distance of word forms
  - cosine distance of word embeddings

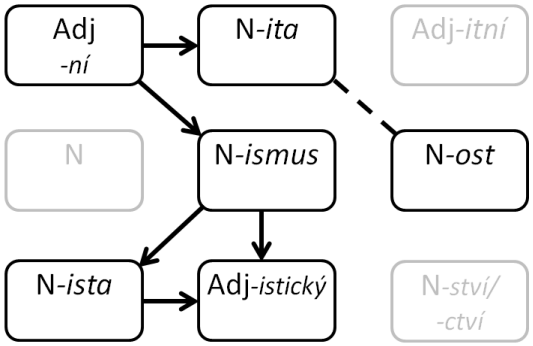


Paradigmatic approach to word-formation

- text corpora
  - words in their natural context as produced by language users
  - token-type distinction
  - provide insight into syntagmatic relations in language
- lexical resources
  - word lists are not produced by language users
  - types in mutual relations
  - provide insight into **paradigmatic** relations in language

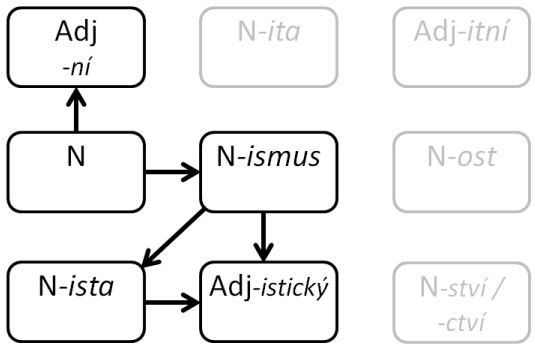
Loan words in the word-formation of Czech

ex. *naivismus, objektivismus*



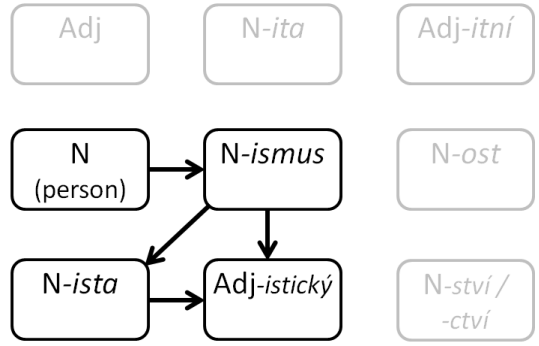
Pattern 1: “approach / movement”

ex. *kapitalismus, extrémismus*



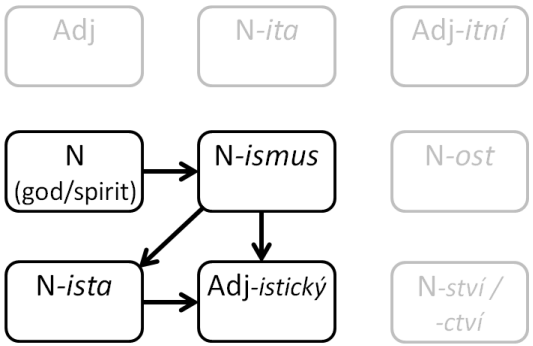
Pattern 2: “approach / movement”

ex. *darwinismus, marxismus*



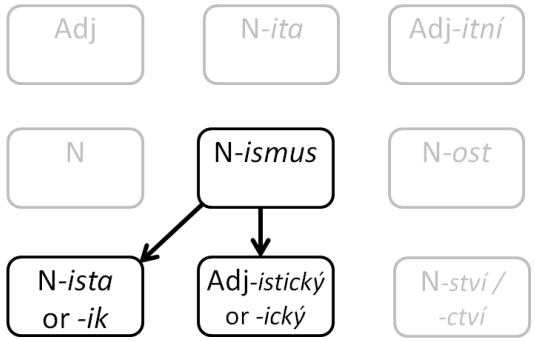
Pattern 3: “approach by someone”

ex. *satanismus, šamanismus*



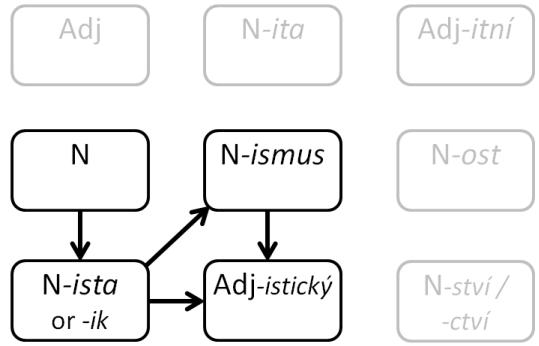
Pattern 4: “belief in someone”

ex. *astigmatismus, autismus*



Pattern 5: “condition”

ex. *alkoholismus, kariérismus*



Pattern 6: “inclination”

Direction of derivation in suffixless nouns

- case study on the direction of derivation between Czech suffixless nouns and corresponding verbs
- two patterns revealed:
  - aspectual pairs formed by **suffixation** correspond to nouns with **verbal roots**

    - *skok* 'jump' <– *skákat* : *skočit* 'to jump.IPFV/PFV'
    - *pád* 'fall' <– *padat* : *padnout* 'to fall.IPFV/PFV'
    - *pocit* 'feeling' <– *pocítit* : *pociťovat* 'to feel.IPFV/PFV'

    = **deverbal (verb-to-noun) derivation**

  - verbs whose aspectual counterparts are derived by **prefixation** correspond to nouns with **nominal roots**

    - *sůl* 'salt' –> *solit* : *osolit* 'to add salt.IPFV/PFV'
    - *sníh* 'snow' –> *sněžit* : *nasněžit* 'to snow.IPFV/PFV'
    - *kámen* 'stone' –> *kamenovat* : *ukamenovat* 'to stone.IPFV/PFV'

    = **denominal (noun-to-verb) derivation**

Grammatical aspect in loan verbs

- the suffixation vs. prefixation strategy of forming aspectual counterparts
  - rooted in the diachrony
  - present in the synchronic word-formation system, observable in loan verbs
    - *pád* – *padat* : *padnout*, *klik* – *klikat* : *kliknout*
    - *kámen* – *kamenovat* : *ukamenovat*, *Google* – *googlovat* : *vygooglovat*
- a case study on loan verbs in Czech documented a prevalence of the prefixation strategy
  - loan verbs in Czech resemble native denominal verbs, i.e. loan roots seem to be interpreted as nominal ones in Czech
  - it correlates with Moravcsik's (1975, 1978) typological hypothesis (Wichmann & Wohlgemuth 2008; Haspelmath 2008)
    - verbs are not borrowed as verbs between languages but are accepted rather as nouns
    - they are subsequently turned into verbs in the recipient language

DeriMo 2019 workshop

- **Second International Workshop on Resources and Tools for Derivational Morphology**
  - `https://ufal.mff.cuni.cz/derimo2019/`
  - September 19–20
  - ÚFAL, room S1

  - following up on the first workshop
    organized by Marco Passarotti & Eleonora Litta
    in Milano, October 2017