A language resource specialized in Czech word-formation: Recent achievements in developing the DeriNet database

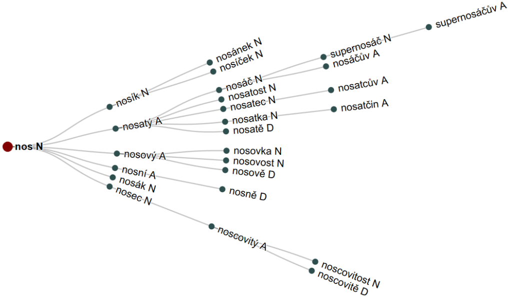
Magda Ševčíková, Adéla Kalužová, and Zdeněk Žabokrtský

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

SlaviCorp 2018
September 26, 2018

DeriNet database

- DeriNet database http://ufal.mff.cuni.cz/derinet
  - developed since 2013, current version 1.6, version 2.0 by the end of 2018
  - 1M+ lexemes extracted from the *MorfFlex CZ* dictionary
  - connected with 800k+ links representing derivational relations



Outline

1. Introduction
   - Derivational resources for Czech
   - Derivational resources for other languages
2. DeriNet database
   - Design decisions
   - Connecting the lexemes
   - Data format
   - Current version
   - Search tools
3. Case studies
   - Aspectual chains
   - Loan words
   - Derivational networks for Spanish and Polish
4. Conclusions

Derivational morphology in NLP of Czech

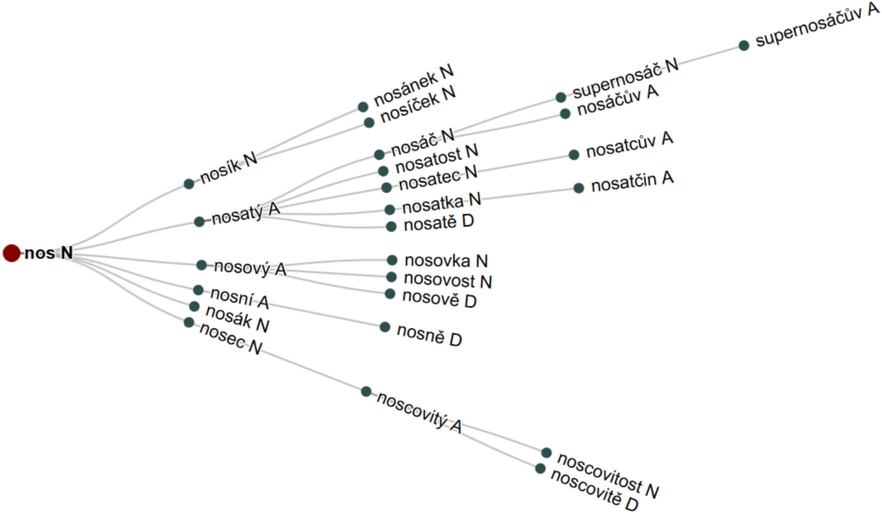
- derivational morphology underresourced in Czech
- specialized resources and tools
  - *Deriv* (Osolsobě et al. 2009)
  - *Morfio* (Cvrček & Vondřička 2013)
  - *Derivancze* (Pala & Šmerk 2015)
- basic derivational info included in resources of other types
  - *Ajka/Majka* analyser (Sedláček & Smrž 2001, Hlaváčková et al. 2009)
  - *Czech WordNet* (Pala & Hlaváčková 2007)
  - *MorfFlex CZ* dictionary (Hajič & Hlaváčová 2013)
  - deep-syntactic annotation of *PDT 2.0* (Hajič et al. 2006, Razímová & Žabokrtský 2006)

Derivational resources for other (Slavic+) languages

- attention to derivations in other languages rather recent, cf.
    - *CELEX* (en, de, nl; Baayen et al. 1995)
    - *DerivBase* (de; Zeller et al. 2013)
    - *CroDeriV* (Šojat et al. 2014)
    - *DerivBase.Hr* (Šnajder et al. 2014)
    - language-independent approach (Baranes & Sagot 2014)
    - *Démonette* (fr; Hathout & Namer 2014)
    - *Word Formation Latin* (Litta et al. 2016)
    - networks for Polish and Spanish (applying the DeriNet approach; Lango et al. 2018)

Focus on derivation

- derivation predominates over compounding in Czech
  - based on Dokulil's (1962) approach to derivation (Štekauer 1998)
- lexemes extracted from the *MorfFlex CZ* dictionary
  - limited to nouns (N), adjectives (A), verbs (V), and adverbs (D)
  - represented as nodes
- a derivational relation between two lexemes represented as an edge connecting two nodes
  - one base lexeme for each derivative
- derivationally related words form a tree structure
  - an unmotivated lexeme is the root of the tree
  - increasing morphemic and semantic complexity of the derivatives


nos N
nosík N
nosánek N
nosíček N
nosatý A
nosáč N
supernosáč N
supernosáčův A
nosáčův A
nosatost N
nosatec N
nosatcův A
nosatka N
nosatčin A
nosatě D
nosový A
nosovka N
nosovost N
nosově D
nosní A
nosně D
nosák N
nosec N
noscovitý A
noscovitost N
noscovitě D

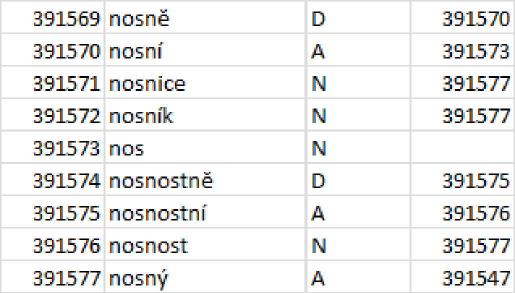
Connecting lexemes with derivational links

1. semi-automatic procedure searching base-derivative pairs
   - using suffix-substitution rules, e.g.
     Adj-ý>N-ost: *závislý*$_A$ ‘dependent’ → *závislost*$_N$ ‘dependency’
     V->N-el: *učit*$_V$ ‘to teach’ → *učitel*$_N$ ‘teacher’
   - suffix-substitution rules extracted from the data or compiled manually
2. extraction of derivational information from existing resources
   - *MorfFlex CZ*
   - *Vallex* valency lexicon (Lopatková et al. 2018)
   - www.wiktionary.org
   - monolingual dictionaries (*Slovník spisovného jazyka českého*)
3. Machine Learning methods
   - applied to partially annotated data

»»» all base-derivative pairs confirmed manually

Data format

- .tsv format
  - tab separated values
  - for each lexeme:
    - unique ID
    - lemma
    - POS
    - ID of the base word



391569	nosně	D	391570
391570	nosní	A	391573
391571	nosnice	N	391577
391572	nosník	N	391577
391573	nos	N	
391574	nosnostně	D	391575
391575	nosnostní	A	391576
391576	nosnost	N	391577
391577	nosný	A	391547

- data published in the Lindat/Clarin repository
  - `http://hdl.handle.net/11234/1-2873`
  - Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA)

Current version: DeriNet 1.6

lexemes	**1,027,832**	**incl. 33,236 compounds**
N	452,374	incl. 14,924 compounds (NC)
A	357,444	incl. 17,265 compounds (AC)
D	162,019	incl. 353 compounds (DC)
V	55,995	incl. 694 compounds (VC)

derivational links	**803,404**							
	V2A	208,053	V2N	61,791	A2A	987	D2A	56
	A2N	164,028	V2V	44,960	A2V	604	D2N	8
	A2D	159,568	N2N	43,796	D2D	95	V2D	7
	N2A	117,401	N2V	1,984	N2D	61	D2V	5

trees	**224,428**
roots	224,428 – 33,236 out of them are compounds – 127,062 out of them capitalized – some of them unmotivated words

Search tools

- DeriSearch

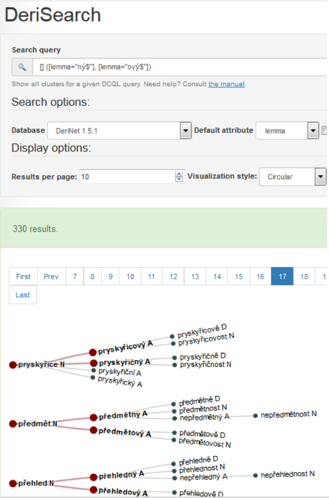
  `http://ufal.mff.cuni.cz/derinet/search`

  - by Jonáš Vidra
  - features of the nodes, tree structure
  - cf. `[] ([lemma="ný$"], [lemma="ový$"])`
  - another three visualization modes (Vidra & Žabokrtský 2017)
  - usable for other resources

- DeriNet Viewer

  `http://ufal.mff.cuni.cz/derinet/viewer`

  - by Milan Straka
  - grouping trees according to their shape, depth etc.



DeriSearch

Search query

[] ([lemma="ný$"], [lemma="ový$"])

Show all clusters for a given DCQL query. Need help? Consult the manual.

Search options:

Database: DeriNet 1.5.1 | Default attribute: lemma

Display options:

Results per page: 10 | Visualization style: Circular

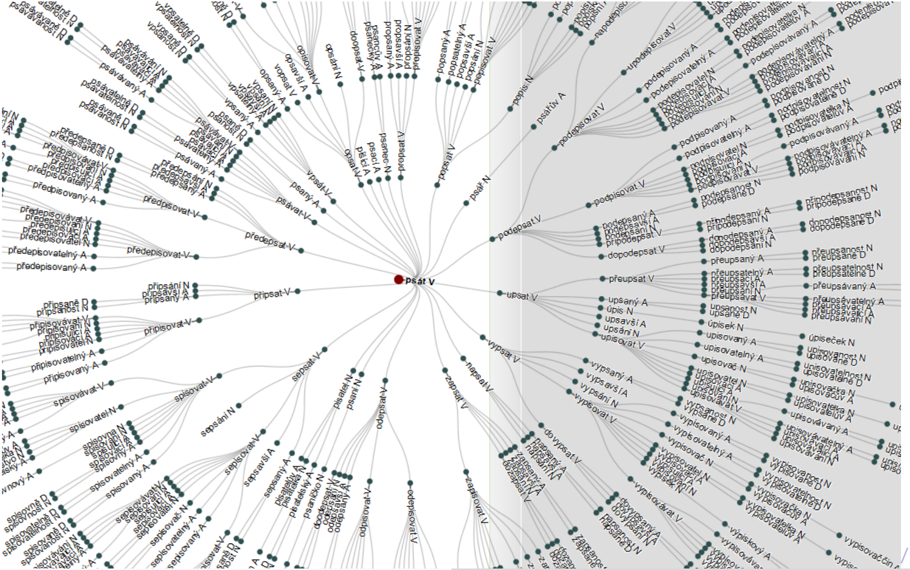
330 results.

Case studies

- linguistic research
  - aspectual chains
    - Ševčíková & Panevová 2018
  - derivational behavior of loan words in Czech
    - Ševčíková 2017
- Natural Language Processing
  - semi-automatic creation of derivational networks
    - Lango et al. 2018

Aspectual chains

- derivation of verbs in Czech
  - verbs mostly derived from verbs
  - prefixation predominates over suffixation
    - up to 18 prefixes attested with a verbal stem
  - form large derivational families
  - derivationally related verbs differ in meaning and/or in aspect
- 55k+ verbs in DeriNet organized according to a simple set of criteria (Žabokrtský et al. 2017)
- for the sake of the analysis, a subset of the DeriNet data compiled that contained only verbs attested in the SYNv6 corpus (Křen et al. 2017)


psát V
napsat V
vypsat V
upsat V
přeupsat V
podepsat V
podepisovat V
psaň N
psaťův A
popis N
popsat V
přepsat V
opsat V
vpsat V
psaný A
psávat V
předepsat V
předepisovat V
předepisovaný A
předepisovatelný A
připsat V
připisovat V
připisovaný A
připisovatelný A
spisovat V
spisovatel N
sepsat V
sepsání N
sepisovat V
sepsavší A
sepsaný A
pisatel N
psaní N
odepsat V
odpisovat V
odepisovat V
zapsat V
zapisovat V
dovypsat V
vypsaný A
vypsavší A
vypsání N
vypisovat V
vypisovaný A
vypisovatelný A
vypisovač N
upsaný A
úpis N
upsavší A
upsání N
upisovat V
upisovaný A
upisovatelný A
upisovač N
úpisek N
úpiseček N
přeupsaný A
přeupsatelný A
přeupsávaný A
dopodepsat V
podepsaný A
přípodepsat V
podepisovaný A
podepisovatelný A
podpisovat V
podpisovaný A
podpisovatelný A
upodepisovat V
napodepisovat V
propsat V
psací A
písař N
psanec N
přepisovat V

Aspectual chains: four most frequent patterns

1. simplex imperfective – prefixed perfective
   - *psát* ‘to write.impf’> *napsat* ‘to write.pf’
   - *pršet* ‘to rain.impf > *napršet* ‘to rain (down).pf’
     pršet ‘to rain’ > *zapršet* ‘to rain (a little).pf’
2. simplex impf – prefixed pf – secondary impf
   - *psát* ‘to write.impf’ > *odepsat* ‘to write back.pf’ > *odepisovat* ‘to write back.impf’
3. prefixed pf – secondary impf
   - *odeslat* ‘to send off.pf’ > *odesílat* ‘to send off.impf’
4. (a) simplex impf – suffixed pf – prefixed pf
   - *štěkat* ‘to bark.impf’ > *štěknout* ‘to bark.pf’ > *vyštěknout* ‘to snap.pf’

   (b) simplex impf – prefixed pf – pf with two prefixes
   - *čistit* ‘to clean.pf’ > *vyčistit* ‘to clean.pf’ > *dovyčistit* ‘to clean.pf’

Loan words

- internationalisms (Jiráček 1984) are members of larger or smaller derivational families in West-Slavic languages (Waszakowa 2003)
- a case study on nouns in *-ismus* in Czech (Ševčíková 2017)
  - nouns in *-ismus* share their root with a different number of derivatives formed by different suffixes

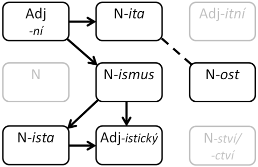
šamanismus – šaman – šamanista – šamanistický
darwinismus – Darwin – darwinista – darwinistický
rusismus – rusista – rusistický – rusistika
kanibalismus – kanibal – kanibalský
alkoholismus – alkohol – alkoholik – alkoholický
fotbalismus – fotbal – fotbalista – fotbalistický

Loan words: corpus data analysis

- all nouns *-ismus* from the SYN2015 corpus
  - reduction from 1,219 to 749 types due to orthographic variability
  - selected formations that share the root with the *-ismus* nouns extracted from the corpus
- analysing the size and inner structure of the derivational families
- there are correlations between how a particular derivational family looks like and what meaning the involved derivatives have
- word-formation meaning of the suffix described by patterns

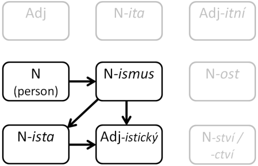
Loan words: word-formation patterns

naivismus, objektivismus



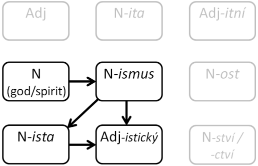
Pattern 1: “approach / movement”

darwinismus, marxismus



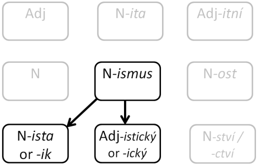
Pattern 3: “approach by someone”

šamanismus, višnuismus



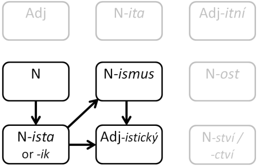
Pattern 4: “belief in someone”

astigmatismus, autismus



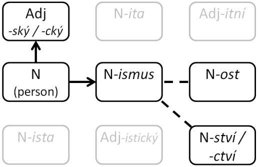
Pattern 5: “condition”

alkoholismus, kariérismus



Pattern 6: “inclination”

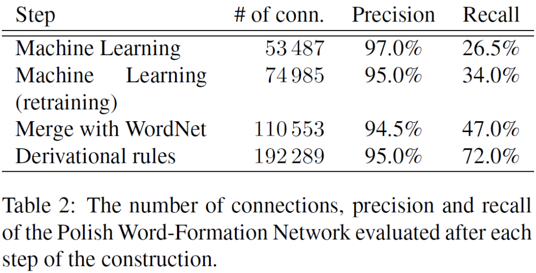
barbarismus, kanibalismus



Pattern 7: “being someone”

Derivational networks for Spanish and Polish (i/ii)

- Lango et al 2018:
  - semi-automatic construction methods, applicable to underresourced languages
  - manual annotation of a small training set, Machine Learning techniques, suffix-substitution rules, Polish WordNet

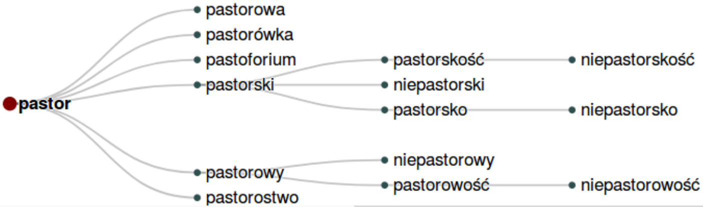


Step	# of conn.	Precision	Recall
Machine Learning	53 487	97.0%	26.5%
Machine Learning (retraining)	74 985	95.0%	34.0%
Merge with WordNet	110 553	94.5%	47.0%
Derivational rules	192 289	95.0%	72.0%

Table 2: The number of connections, precision and recall of the Polish Word-Formation Network evaluated after each step of the construction.

Derivational networks for Spanish and Polish (ii/ii)

- Spanish Word-Formation Network 0.5
    - 160k lexemes with 18k+ links
- Polish Word-Formation Network 0.5
    - 260k+ lexemes with 190k+ links
- available under the CC-BY-ND license at `http://ufal.mff.cuni.cz/derinet`



Conclusions, next steps

- DeriNet 1.6
  - 1M+ Czech lexemes connected with 800k+ derivational links
  - compounds identified but not connected with bases
  - usable in both linguistic research and NLP tasks
- DeriNet 1.6 –> DeriNet 2.0
  - increase the number of derivational links
  - substantial changes in the data structure
    - representation of compounds
    - links to more motivating lexemes
    - semantic labelling of derivational links
- derivational data for other languages

`http://ufal.mff.cuni.cz/derinet`

References

- Baayen, R. H. et al. (1995): *The CELEX lexical database* (release 2). Data/software. Philadelphia, PA: LDC.
- Baranes, M. & Sagot, B.: A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of LREC 2014*, pp. 2793–2799.
- Cvrček, V. & Vondřička. P. (2013): Nástroj pro slovotvornou analýzu jazykového korpusu. In: *Gramatika a korpus 2012*. Hradec Králové.
- Křen, M. et al. (2015): *SYN2015*. Dostupný z WWW: `http://www.korpus.cz`
- Dokulil, M. (1962): *Tvoření slov v češtině 1: Teorie odvozování slov*. Praha.
- Hajič, J. & Hlaváčová, J. (2013): *MorfFlex CZ*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9
- Hajič, J. et al. (2006): *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.
- Hathout, N. & Namer, F.: Démonette, a French Derivational Morpho-Semantic network. *LiLT* 2014:11, 125–168.
- Hlaváčková, D. et al. (2009): Relations between Formal and Derivational Morphology in Czech. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Czech in Formal Grammar*. München, pp. 79–87.
- Lango, M. et al. (2018): Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). In *Proceedings of LREC 2018*.
- Litta, E. et al. (2016): Formatio formosa est. Building a Word Formation Based Lexicon for Latin. In *CLiC-it 2016*, pp. 185–189.
- Lopatková, M. a kol. (2017): *Valenční slovník českých sloves Vallex*. Praha.
- Osolsobě, K. et al. (2009): Exploring Derivational Relations in Czech with the Deriv Tool. In *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Bratislava, pp. 152–161.
- Pala, K. & Hlaváčková, D. (2007): Derivational Relations in Czech WordNet. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*. Prague, pp. 75–81.
- Pala, K. & Šmerk, P. (2015): Derivancze – Derivational Analyzer of Czech. In: *TSD 2015*, pp. 515–523.

- Razímová, M. & Žabokrtský, Z. (2006): Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In *Proceedings of the LREC 2006 Workshop on Annotation Science*, pp. 12–19.
- Sedláček, R. & Smrž, P. (2001): A New Czech Morphological Analyzer ajka. In: *TSD 2001*, pp. 100–107.
- Ševčíková, M. (2017): The suffixes -ismus and -ita in nouns in Czech. *Societas Linguistica Europaea conference*
- Ševčíková, M. & Panevová, J. (2018): Derivation of Czech verbs and the category of aspect. *Linguistica Copernicana*
- Šnajder, J. et al.: DerivBase.Hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of LREC 2014*, pp. 3371–3377.
- Štekauer, P. (1998): *An Onomasiological Theory of English Word-formation*. John Benjamins.
- Šojat, K. et al.: CroDeriV: a New Resource for Processing Croatian Morphology. In *Proceedings of LREC 2014*, pp. 3366–3370.
- Vidra, J. & Žabokrtský, Z. (2017): Online Software Components for Accessing Derivational Networks. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*. Milan, pp. 129–139.
- Zeller, B. et al.: DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL 2013*, pp. 1201–1211.

The presented work was supported by the Czech Science Foundation grant GA16-18177S.