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

  - 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

sevcikova@ufal.mff.cuni.cz

DeriNet database @ http://ufal.mff.cuni.cz/derinet
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 & Hlaváčková 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)
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
Connecting lexemes with derivational links

1. semi-automatic procedure searching base-derivative pairs
   - using suffix-substitution rules, e.g.
     Adj-ý > N-ost: závislý\textsubscript{A} ‘dependent’ → závislost\textsubscript{N} ‘dependency’
     V- > N-el: učit\textsubscript{V} ‘to teach’ → učitel\textsubscript{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

sevcikova@ufal.mff.cuni.cz

DeriNet database @ http://ufal.mff.cuni.cz/derinet
Data format

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

- 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

<table>
<thead>
<tr>
<th>lexemes</th>
<th>1,027,832 incl. 33,236 compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>452,374 incl. 14,924 compounds (NC)</td>
</tr>
<tr>
<td>A</td>
<td>357,444 incl. 17,265 compounds (AC)</td>
</tr>
<tr>
<td>D</td>
<td>162,019 incl. 353 compounds (DC)</td>
</tr>
<tr>
<td>V</td>
<td>55,995 incl. 694 compounds (VC)</td>
</tr>
</tbody>
</table>

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

| 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**
  - 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/derinetviewer
  - by Milan Straka
  - grouping trees according to their shape, depth etc.
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)
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’ > dvoýčistit ‘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

sevcikova@ufal.mff.cuni.cz

DeriNet database @ http://ufal.mff.cuni.cz/derinet
Loan words: word-formation patterns

**naivismus, objektivismus**

- **Adj** -ní
- **N**-ta
- **N**-ismus
- **Adj-isticky**
- **N**-sta

Pattern 1: “approach / movement”

**darwinismus, marxismus**

- **Adj**
- **N**-ta
- **Adj-istní**
- **N**-ismus
- **Adj-isticky**
- **N**-sta

Pattern 3: “approach by someone”

**šamanismus, višnuismus**

- **Adj**
- **N**-ta
- **Adj-istní**
- **N**-ismus
- **Adj-isticky**
- **N**-sta

Pattern 4: “belief in someone”

**astigmatismus, autismus**

- **Adj**
- **N**-ta
- **Adj-istní**
- **N**-ismus
- **Adj-isticky** or -ický
- **N**-sta or -ik

Pattern 5: “condition”

**alkoholismus, kariérismus**

- **Adj**
- **N**-ta
- **Adj-istní**
- **N**-ismus
- **Adj-isticky**
- **N**-sta or -ik

Pattern 6: “inclination”

**barbarismus, kanibalismus**

- **Adj** -ský / -cký
- **N**-ta
- **Adj-istní**
- **N**-ismus
- **Adj-isticky**
- **N**-sta or -ik

Pattern 7: “being someone”

---

sevcikova@ufal.mff.cuni.cz

DeriNet database @ http://ufal.mff.cuni.cz/derinet
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

<table>
<thead>
<tr>
<th>Step</th>
<th># of conn.</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning</td>
<td>53 487</td>
<td>97.0%</td>
<td>26.5%</td>
</tr>
<tr>
<td>Machine Learning (retraining)</td>
<td>74 985</td>
<td>95.0%</td>
<td>34.0%</td>
</tr>
<tr>
<td>Merge with WordNet</td>
<td>110 553</td>
<td>94.5%</td>
<td>47.0%</td>
</tr>
<tr>
<td>Derivational rules</td>
<td>192 289</td>
<td>95.0%</td>
<td>72.0%</td>
</tr>
</tbody>
</table>

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


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