A data-based approach to competition in word-formation: diminutives and gender marking across seven languages

Morphological meeting at Laboratoire de linguistique formelle
Paris 2022

Lukáš Kyjánek
KYJANEK@UFAL.MFF.CUNI.CZ
https://lukyjanek.github.io/

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

This work was supported by the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935).
START Grant

• 03/2021 – 03/2022

• Morphological research into competition in Germanic, Romance, and Slavic langs.
  • G: Dutch English German | R: French Spanish | S: Czech Russian

• Mgr. Magda Ševčíková, PhD. (mentor)
• Mgr. Lukáš Kyjánek (PI) : semantics in derivational morphology, language resources
• Mgr. Jan Bodnár : morphological segmentation
• Mgr. Emil Svoboda : compounding
• Mgr. Jonáš Vidra : linguistic transfer methods, language resources
Outline

1. Basic notions

2. Data Resources
   • DeriNet
   • Universal Derivations
   • DeriNet.RU
   • Universal Segmentations

3. Methodology
   • Searching for spelling variants (in Czech)
   • Labelling derivational meanings (in Czech)
   • Analysing agent noun formation (in Czech)
   • Transfering word-formation networks (from Czech)

4. Ongoing work
   • Analysis of gender marking formation (in Czech)
   • Comparison of diminutiveness and gender marking across languages
Basic notions
Approaches to derivational morphology

Kőrtvélyessy et al. (2020:10-11)

1. Direct derivatives (paradigm)
   \[\text{dom} \rightarrow \text{dom-ov}\]
   \[\rightarrow \text{dom-ček}\]
   \[\rightarrow \text{dom-ík}\]
   \[\rightarrow \text{dom-isko}\]

2. Subsequent derivatives (series)
   \[\text{dom} \rightarrow \text{dom-ov} \rightarrow \text{dom-ov-ina} \rightarrow \text{dom-ov-in-ový}\]
   \[\text{dom} \rightarrow \text{dom-ček} \rightarrow \text{dom-ček-ový}\]
   \[\text{dom} \rightarrow \text{dom-ík} \rightarrow \text{dom-ík-ový}\]
   \[\text{dom} \rightarrow \text{dom-isko} \rightarrow \text{dom-isk-ový}\]

3. Semantic categories of each derivational step
   agent, female, location, quality, agumentative, etc.

4. Derivational network
   = derivatives derived from a simple underived word
   (combination of (1) and (2) and (3))
Derivational meaning

- **odesílat** → **odesíla-tel** (to send > sender)
  - **odesílat** = activity
  - **odesíla-tel** = someone who does the activity

- One affix can convey many meanings
  - **úředník** → **úředn-ice** (officer > female officer)
  - **věznit** → **vězn-ice** (to imprison > jail)
  - **kytka** → **kyt-ice** (flower > bouquet)

- One meaning can be conveyed by many affixes
  - **úředník** → **úředn-ice** (officer > female officer)
  - **šéf** → **šéf-ová** (boss > female boss)
  - **učitel** → **učitel-ka** (teacher > female teacher)
  - **ministr** → **ministr-yně** (minister > female minister)
Data Resources
DeriNet

• Lexical network which models word-formation relations in the lexicon of Czech

• Over 1 milion lexemes
  • 782 thousand derivational relations
  • 50 thousand links for orthographic variants
  • 1,952 links for compounding
  • 144 relations of conversion
  • 295 relations of univerbisation

• http://www.ufal.cz/derinet
Universal Derivations

- Collection of harmonized lexical networks capturing word-formation, especially derivation, in a cross-linguistically consistent annotation scheme for many languages (UDer 1.1 contains 31 harmonized resources covering 21 languages)


- Harmonisation process:
  - Assembling the existing resources
  - Scoring derivational relations
  - Finding maximum spanning tree
DeriSearch v2:  http://www.ufal.cz/derisearch
DeriNet.RU

- Lexical network which models word-formation relations in the lexicon of Russian

- Over 337 thousand lexemes connected by more than 164 thousand derivational relations into 172 thousand derivational families

- Created on the basis of:
  - Grammar-based model of derivational rules from Russian grammar books, e.g.,
    - Rule 343 (noun + and > noun)
    - анархия [anarchy] > анархист [anarchist]
  - Harmonisation procedure (improved)
Universal Segmentations

• Collection of lexical resources capturing morphological segmentations harmonised into a cross-linguistically consistent annotation scheme for many languages (17 harmonized resources providing 48 data sets covering 37 languages)

Methodology
Searching for spelling variants

Examples:
- úřad, ouřad
- předhřát, předehřát
- ohražování, ohrazování
- jakkoliv, jakkoli
- dopingový, dopinkový
- býk, bejč
- Tchałvan, Tchajwan
- odyditi, odydit
- trojnožka, triñožka
- žebřina, řebřina
- berla, berle
- (?) bezkolejný, bezkolejový
- (?) bezhlesy, bezhlesný
- (?) bledoučí, bledounký
- (?) bočný, boční
- (?) drobnínek, drobítek
- (?) rozechvěný, rozechvělý
Labelling derivational meanings

- Pilot experiment: to add 5 labels limited to suffixation into DeriNet for Czech
  - *pes* → *psík* (dog > small dog)
  - *učitel* → *učitelka* (teacher > female teacher)
  - *učitel* → *učitelův* (teacher > teacher’s)
  - *chodit* → *chodívat* (to walk (IPFV) > to walk repeatedly (IPFV))
  - *obalit* → *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%

<table>
<thead>
<tr>
<th>Label</th>
<th>Derivations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diminutive</td>
<td>5,383</td>
</tr>
<tr>
<td>Female</td>
<td>28,623</td>
</tr>
<tr>
<td>Possessive</td>
<td>87,087</td>
</tr>
<tr>
<td>Iterative</td>
<td>11,778</td>
</tr>
<tr>
<td>Aspect</td>
<td>15,186</td>
</tr>
</tbody>
</table>

- Already available since DeriNet 2.0
Analysing agent noun formation

- 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 containing all features
- Other experiments used 5 different subsets of features, but the same settings

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

<table>
<thead>
<tr>
<th>Suffix</th>
<th>-tel</th>
<th>-č</th>
<th>-nik/-ík</th>
<th>-ář/-alř</th>
<th>-ce</th>
<th>-ák</th>
<th>-ec</th>
<th>-čí</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>426</td>
<td>388</td>
<td>106</td>
<td>96</td>
<td>66</td>
<td>50</td>
<td>32</td>
<td>14</td>
<td>1,178</td>
</tr>
</tbody>
</table>

Magda Ševčíková & Lukáš Kyjánek & Barbora Vidová Hladká 2021 (ParadigMo2)
http://ekladata.com/_IPboYlm_KrJSWKInfym8wEURc/ParadigMo-2-Booklet-of-abstracts.pdf#page=71
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
Transferring word-formation networks

- proof-of-concept method for creating word-formation networks by transferring information from another language
- creates a low-precision and moderate-recall network in a language, for which no manual annotations need to be available
Ongoing work
Analysis of gender marking formation

- What are the base lexemes of the female representatives, and what is their distribution in terms of morphosyntactic categories, word-formation properties, and frequencies over time?

  - **simplex:**
    - matka$_{N,fem}$ (mother) – otec$_{M,masc}$ (father)
    - vdovala$_{N,fem}$ (widow) > vdocic$_{M,masc}$ (widower)
  
  - **derivatives:**
    - uctelka$_{N,fem}$ (female teacher) < uctel$_{N,masc}$ (teacher)
    - krasa$_{N,fem}$ (beautiful woman) < krasa$_{N,fem}$ (beauty)
    - behna$_{N,fem}$ (floozy) < behat$_{V}$ (to run)
    - svetlovlaska$_{N,fem}$ (fair-haired woman) < svetlovasy$_{A}$ (fair-haired)
  
  - **conversion:**
    - pruvodci$_{N,fem}$ (conductress) <= pruvodci$_{N,masc}$ (conductor)
    - hajna$_{N,fem}$ (female ranger) <= hajny$_{M,masc}$ (ranger)

Initial Data
- extracted all words with feminine gender from representative corpus SYN2020 (female nouns have only feminine gender in Czech)

Manual Annotations
- proper/common nouns
- social gender
- semantic type

Feature equipment
- word-formation process
- base lexeme(s)
- masculine counterparts
- diminutive derivatives
- stems, …

Analyses
- How much influence the individual features have on the selection of the word-formation process and derivational affix when coining a word?
Comparison of diminutiveness and gender marking across languages

- to quantify which strategies are used across the 7 languages to convey diminutiveness and gender marking => we need the same data across languages

- data: starts with a derivatives labelled as Diminutive/Female from DeriNet (cs) and translating them into other languages
  - several techniques of machine translation: neural systems, bilingual dictionaries, custom dictionaries from parallel corpora, other resources

- analyses: ... soon 😊
¿Distributional semantics?
Thank you.
References


