A Distributional Approach to Inflection vs. Derivation in Czech

Lukáš Kyjánek & Olivier Bonami

June 23–26, 2022
Delineating the border between inflection and derivation

• Changes of an affix affect grammatical or lexical meaning of a word, the former ones are treated as inflectional, while the latter ones as derivational categories
  - Affix -(e)s for 3rd singular person as inflection, e.g., teach → teaches
  - Affix -er for agent name as derivation, e.g., teach → teacher

• When delineating border between inflection–derivation, the available literature insists on
  - Either a categorical distinction and look for corresponding criteria (Anderson, 1982),
  - Or an elusiveness of the distinction, which is seen as gradient and/or multidimensional (Dressler 1989; Booij 1996; Haspelmath 1996; Corbett 2010; Spencer 2013; Štekauer 2015)

• Recent work has applied computational methods from distributional semantics (e.g. Boleda 2020) to the issue of the border between inflection and derivation (cf. Bonami and Paperno 2018, Rosa and Žabokrtský 2019), but consider smaller sets of morphological categories.
Three views of the inflection–derivation distinction

Monodimensional, categorical

Monodimensional, gradient

Multidimensional, categorical
Outline

Current State of Knowledge
  Distributional semantics
  Existing data resources for Czech

Data & Methods
  Semantic contrasts
  Prototypical sample

Results
  Global overview [monodimensional gradient]
  Specific features [multidimensional categorical]

Discussion

Conclusion

Appendix
Distributional semantics

• The distributional hypothesis (see Harris 1954, Firth 1957) from (Lenci, 2008, p. 3): "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear."

• Contemporary computational linguistics deduce semantic representations from large corpora to follow this idea.
Distributional semantics for morphology

- Proportional analogy, accessible through vector arithmetic (Mikolov et al., 2013), works to the extent that differences between pairs of words are similar.
- These **difference vectors** represent the shift in distribution from word to the next.
- Studying the similarity of these difference vectors, tells us about stability of contrasts.
Existing data resources for Czech

Distributional semantics

1. **Word2vec** (Mikolov et al., 2013)
2. **SYN v9 corpus** (Křen et al., 2021)
   - large representative corpus of Czech
   - 362M sentences; 4,719M tokens; 7.3M lemmas
   → We rely on the corpus pos-tag annotations as we train vectors for combinations of tokens and tags.

Morphological data

1. **MorfFlexCZ 2.0** (Hajič et al., 2020)
   - inflectional morphological lexicon
   - 125.3M lemma-tag-wordform triples
2. **DeriNet 2.1** (Vidra et al., 2021)
   - derivational morphological lexicon
   - 1M lemmas; 782,814 derivations
### Example from MorfFlexCZ: inflection of 'barber'

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Tag</th>
<th>Word form</th>
</tr>
</thead>
<tbody>
<tr>
<td>holič</td>
<td>NNMS1-----A----</td>
<td>holič</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS2-----A----</td>
<td>holiče</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS3-----A----</td>
<td>holiči</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS3-----A---1</td>
<td>holičovi</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS4-----A----</td>
<td>holiče</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS5-----A----</td>
<td>holiči</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS6-----A----</td>
<td>holiči</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS6-----A---1</td>
<td>holičovi</td>
</tr>
<tr>
<td>holič</td>
<td>NNMS7-----A----</td>
<td>holiče</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP1-----A----</td>
<td>holiči</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP2-----A----</td>
<td>holičů</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP3-----A----</td>
<td>holičům</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP4-----A----</td>
<td>holiče</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP5-----A----</td>
<td>holiči</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP6-----A----</td>
<td>holičích</td>
</tr>
<tr>
<td>holič</td>
<td>NNMP7-----A----</td>
<td>holiči</td>
</tr>
</tbody>
</table>
Example from DeriNet: derivation of 'barber'
Semantic contrasts available for Czech

We process 24 types of morphological contrasts (difference vectors)

<table>
<thead>
<tr>
<th>Word category A</th>
<th>Word category B</th>
<th>Type of contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun. NOM. FEM. SG</td>
<td>Noun. GEN. FEM. SG</td>
<td>core cases (N∼N)</td>
</tr>
<tr>
<td>Noun. NOM. FEM. PL</td>
<td>Noun. GEN. FEM. PL</td>
<td>core cases (N∼N)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Noun. DAT. FEM. SG</td>
<td>Noun. LOC. FEM. SG</td>
<td>non-core cases (N∼N)</td>
</tr>
<tr>
<td>Noun. DAT. FEM. PL</td>
<td>Noun. LOC. FEM. PL</td>
<td>non-core cases (N∼N)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Noun. NOM. FEM. SG</td>
<td>Noun. DAT. FEM. SG</td>
<td>mixed cases (N∼N)</td>
</tr>
<tr>
<td>Noun. NOM. FEM. PL</td>
<td>Noun. DAT. FEM. PL</td>
<td>mixed cases (N∼N)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Noun. NOM. FEM. SG</td>
<td>Noun. DIM. NOM. FEM. SG</td>
<td>diminutive (N∼N)</td>
</tr>
<tr>
<td>Noun. GEN. FEM. PL</td>
<td>Noun. DIM. GEN. FEM. PL</td>
<td>diminutive (N∼N)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Verb.inf</td>
<td>Noun. AGENT. NOM. MASC. SG</td>
<td>agent (V∼N)</td>
</tr>
<tr>
<td>Verb.inf</td>
<td>Noun. AGENT. NOM. MASC. PL</td>
<td>agent (V∼N)</td>
</tr>
</tbody>
</table>

**core cases:** nom, gen, acc; **non-core cases:** dat, voc, loc, ins; **mixed cases:** contrasts between core and non-core cases
Method

• We sampled 200 word pairs (freq>50) for each contrast and calculated average difference vectors on the basis of individual difference vectors. Then we measured Euclidean distances between the individual difference vectors and the average difference vector. The individual distances were averaged to obtain one average Euclidean distance for the analysed contrast.
Results in a single sample

- Points in the graph show dispersion of the average Euclidean distances for individual contrasts.
- Boxes in the graph aggregate the contrasts into individual types of contrasts.
- The results (sorted by medians) correspond well to the linguistic tradition, but variances of the types of contrasts are high in a single sample.
• Making a bootstrap of medians of individual types of contrasts (100 iter.) shows more stable results with lower variances.

• Inherent inflection and category changing vs. denotation changing derivation stand at the extremes.

• Intermediate situations (e.g., diminution, social gender) stand between the extremes but are hardly comparable.
Assigning properties of the contrasts (multidimensional categorical)

We assign 4 properties to the contrasts (inspired by Bauer 2004 and Spencer 2013):

- **P+** different part of speech
- **I+** inherent (vs. contextual)
- **L+** different lexeme
- **S+** different semantic type (individual vs. eventuality vs. property)

<table>
<thead>
<tr>
<th>Type of contrast</th>
<th>P</th>
<th>I</th>
<th>L</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>core cases (N∼N)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>grammatical gender (A∼A)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>possessive (N∼A)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diminutive (N∼N)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>social gender (N∼N)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>action (V∼N)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>property (A∼N)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ability (V∼A)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>agent (V∼N)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Predictions of partial order

- We expect partial order in the feature lattice to predict differences in vector dispersion, e.g.
  - agents should be more dispersed than diminutives
  - diminutives should be more dispersed than core case contrasts
  - no prediction for possessive adjectives vs. diminutives, as these are not ordered.

- Most of the predicted multidimensional categorical comparisons follow the expected partial order made by the assigned properties (82%).
Most of the pos-changing contrasts have higher distances, except for property (A~N).

The opposite exceptions:
- location (N~N)
- grade (D~D)
- agent (N~N)
Feature I: inherent (vs. contextual)

- Most of the canonical inflectional contrasts have lower distances, except for person (V~V).
- The opposite exceptions:
  - property (A~N)
  - negation (A~A)
Most of the contrasts represented by different lexemes have higher distances, except for **property (A~N)**, **diminutive (N~N)**, and **social gender (N~N)**.

The opposite exceptions:
- **person (V~V)**
- **grade (A~A)**
- **possessive (N~A)**
- **grade (D~D)**
• Most of the contrasts denoting different semantic type have higher distances, except for action (V~N), possessive (N~A), and grade (D~D).

• There are no opposite exceptions.
• Property ($A \sim N$) type of contrast is modelled like inflection in distributional semantics (have lower distance) but we would expect it will behave more like action ($V \sim N$) type of contrast. The two ends up on different parts of the scale.
Discussion

• Property (A∼N) type of contrast is modelled like inflection in distributional semantics (have lower distance) but we would expect it will behave more like action (V∼N) type of contrast. The two ends up on different parts of the scale.

• Person (V∼V) contrast has surprisingly high distance, indicating derivational behaviour; it may be caused by the complicated resolution of person in past participles.
• Property (A∼N) type of contrast is modelled like inflection in distributional semantics (have lower distance) but we would expect it will behave more like action (V∼N) type of contrast. The two ends up on different parts of the scale.

• Person (V∼V) contrast has surprisingly high distance, indicating derivational behaviour; it may be caused by the complicated resolution of person in past participles.

• There are differences across part of speech for the same type of contrast
  - negation (A∼A) vs. (V∼V), and
  - number (A∼A) vs. (V∼V) vs. (N∼N), and
  - grade (A∼A) vs. (D∼D).
• We exploited models of distributional semantics to approach the issue of inflection–derivation distinction on a larger set of semantic contrasts in Czech.

• The results clearly show the inflection-derivation divide as gradient and/or multidimensional.
  - Inherent inflection and category changing, denotation changing derivation stand at the opposite extremes with a few exceptions.
  - Intermediate situations and the properties of the same category across parts of speech (e.g., number on nouns or adjectives, negation on verbs and adjectives) stand between the two extremes.

• This is an instance of convergence of computational modelling and linguistics, which leads us to new theoretical questions.
Acknowledgement

Thank you for your attention.

http://ufal.cz/node/2248

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). It was using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project.


Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. SYN v9: large corpus of written czech, 2021. URL http://hdl.handle.net/11234/1-4635. LINDA T/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.


Appendix A: Euclidean distance vs. Cosine distance

Figure: Prototypical sample before bootstrapping (from left: Cosine, Euclidean distances).
Appendix A: Euclidean distance vs. Cosine distance II

Figure: Bootstrapping (from left: Cosine, Euclidean distances).
Figure: Bootstrapping, feature POS (from left: Cosine, Euclidean distances).
Figure: Bootstrapping, feature INHERENT (from left: Cosine, Euclidean distances).
Appendix A: Euclidean distance vs. Cosine distance

Figure: Bootstrapping, feature LEXEME (from left: Cosine, Euclidean distances).
Appendix A: Euclidean distance vs. Cosine distance VI

Figure: Bootstrapping, feature SEMANTIC TYPE (from left: Cosine, Euclidean distances).
Appendix B: Comparison of the features

<table>
<thead>
<tr>
<th>Feature expectation</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine correct</td>
<td>98</td>
</tr>
<tr>
<td>Cosine incorrect</td>
<td>135</td>
</tr>
<tr>
<td>Euclidean correct</td>
<td>190</td>
</tr>
<tr>
<td>Euclidean incorrect</td>
<td>43</td>
</tr>
<tr>
<td>Same correct results</td>
<td>63</td>
</tr>
<tr>
<td>Same incorrect results</td>
<td>8</td>
</tr>
<tr>
<td>Different results</td>
<td>162</td>
</tr>
</tbody>
</table>

Cosine distance does not model the multidimensional distinction properly, while Euclidean distance does so.