Theoretical stances shape empirical generalisations on inflection *vs.* derivation

Quantitative evidence from Czech

Lukáš Kyjánek & Olivier Bonami

Université Paris Cité
Centre National de la Recherche Scientifique
Laboratoire de Linguistique Formelle

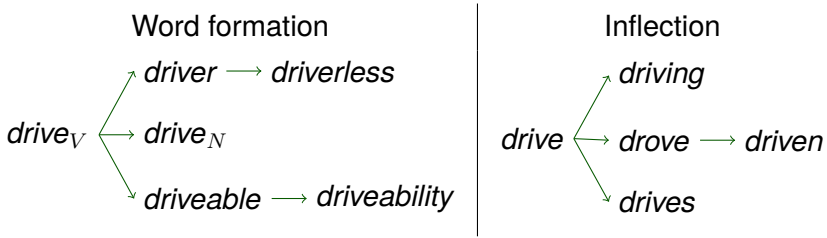
Outline

- Delineating the border between inflection and derivation
- Morphological representations used in the current research
- Main study: Comparison of the current morphological representations
  - Available data for Czech
  - Measurement of the influence on results
  - Differences and Discussion
- Case study: Word-formation meanings at the border
- Conclusions and Future work

- The distinction between inflection and derivation remains unresolved in morphology (Anderson 1982; Dressler 1989; Booij 1996; Haspelmath 1996; Corbett 2010; Spencer 2013; Štekauer 2015).
- Recent computational approaches have addressed this by contrasting the properties of sets of word pairs standing in the same morphological relation. However, different morphological representations are used across studies (Bonami and Paperno 2018; Rosa and Žabokrtský 2019; Copot et al. 2022; Haley et al. 2024).
- We exploit the issue of delineating the border between inflection and derivation to study the influence of different morphological representations on the results of the issue.

A basic theoretical divide

- Two general families of approaches within word-based morphology:
  1. **Rooted tree approaches**: every word is uniquely characterized by its relation to a unique designated ancestor.

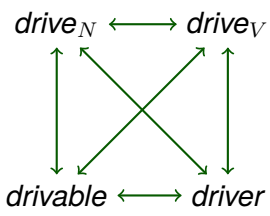


  - In word formation: Aronoff 1976 and many others.
  - In inflection: Boyé 2000, Albright 2003.

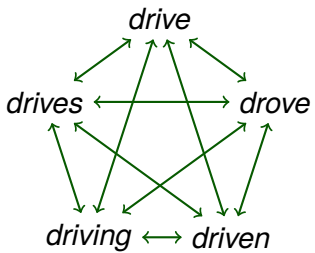
A basic theoretical divide

2. **Paradigmatic approaches**: every word is characterized by its place in a network of content-based relations.

Word formation



Inflection



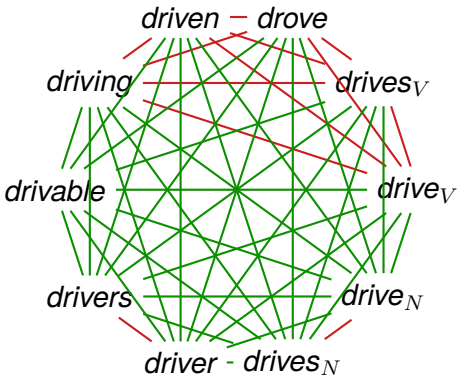
- In inflection: Matthews 1972 and many others
- In word formation: minority position since Robins 1959

A basic theoretical divide

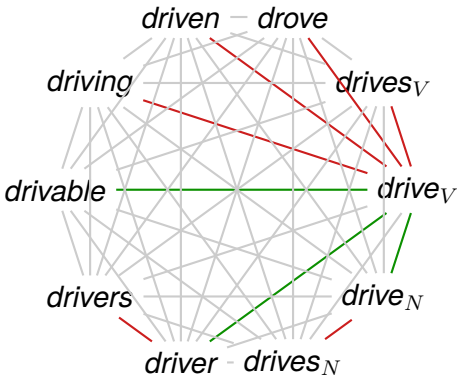
- Some mix and match the two approaches
  - E.g., tree-based word formation, paradigmatic inflection
- Others defend having a consistent approach to both
  - See, e.g., Bochner 1993, Bonami and Strnadová 2019 for a defense of a uniformly paradigmatic approach.

- **Bonami and Paperno 2018**: fully paradigmatic approach, i.e. they compare sets of pairs of words such that:
  - The two words are in any two cells in the paradigm of the same lexeme.
  - The two words are each in one cell in the paradigm of two different lexemes related by derivation.
- **Haley et al. 2024**: tree-based approach, i.e. they compare sets of pairs of words such that:
  - One is the citation form for a lexeme, the other one is another form of the same lexeme.
  - Both are citation forms of two lexemes related by derivation.

Paradigmatic approach



Tree-based approach



Another type of difference

Derivation

Bonami and Paperno 2018

physics ∼ *physic**ian***
phonetics ∼ *phonetic**ian***
biology ∼ *biolog**ist***
morphology ∼ *morpholog**ist***
geometry ∼ *geomet**er***
philosophy ∼ *philosoph**er***
. . . ∼ . . .

Content-based

Haley et al. 2024

physic ∼ *physic**ian***
phonetics ∼ *phonetic**ian***
. . . ∼ . . .

biology ∼ *biolog**ist***
morphology ∼ *morpholog**ist***
. . . ∼ . . .

geometry ∼ *geomet**er***
philosophy ∼ *philosoph**er***
. . . ∼ . . .

Form-based

Inflection

Both

wash ∼ *washed*
miss ∼ *missed*
sing ∼ *sang*
ring ∼ *rang*
leave ∼ *left*
keep ∼ *kept*
. . . ∼ . . .

Content-based

MAIN STUDY:

Comparison of morphological representations

Data available for Czech

- **Word embeddings** (Kyjánek and Bonami 2022) based on
  - Word2vec (Mikolov et al. 2013)
  - **SYN v9 corpus** (Křen et al. 2021)
    - 362M sentences, 4,719M tokens; 7.3M lemmas
- **MorfFlexCZ 2.0** (Hajič et al. 2020)
  - inflectional morphological lexicon
  - 125.3M lemma-tag-wordform triples
- **DeriNet 2.1** (Vidra et al. 2021)
  - derivational morphological lexicon
  - 1M lemmas; 782,814 derivations

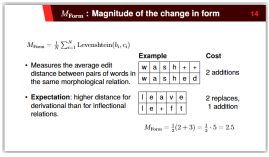
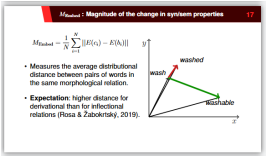
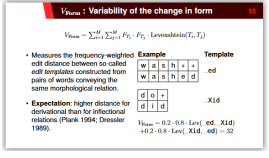
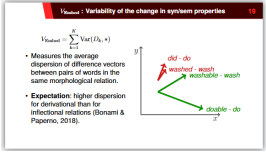
Data prepared for our study

- We analyze morphological categories, exemplifying canonical inflection, derivation, and intermediate cases. We analyse the following types of contrasts:

Cases (1-7)$_{N,A}$	Negation$_{A,V}$ (*ne-*)	Location$_N$ (*-írna, -ště, -ebna, -[á|a]rna, -ovna, -elna, -isko*)
Tense (Pa, Pr, Ft)$_V$	Possessivity$_A$ (*-ův, -in*)	Masculine-Feminine$_N$ (*-ová, -ka, -yně, -ice, -ovna, -ezna*)
Number (Pl, Sg)$_{N,A,V}$	Action$_N$ (*-ní, -ace*)	Diminutive$_N$ (*-ka, -ička, -áček, -(n)ek, -ík, -enka, -ečka*)
Gender (M, I, F, N)$_A$	Verbalisation$_V$ (*-ovat*)	Agent$_N$ (*-ář, -ák, -(n)ík, -tel, -ař, -ač, -ce, -ič, -eč, -ec*)
Grade (1, 2, 3)$_{A,D}$	Ability$_A$ (*-telný*)	
Person (1, 2, 3)$_V$	State$_N$ (*-ost*)	+/- derivation for the purpose of this research
Aspect (P, I)$_V$		+/- inflection for the purpose of this research

- For each contrast (e.g., case: $\mathrm{nominative} \rightarrow \mathrm{instrumental}$), we randomly sample 100 pairs of words (freq > 50).

Method – four measures by Haley et al. 2024

	Form	Syntactic/semantic prop.
Magnitude of the change	M_{Form} 	M_{Embed} 
Variability of the change	V_{Form} 	V_{Embed} 

M_{Form} : Magnitude of the change in form

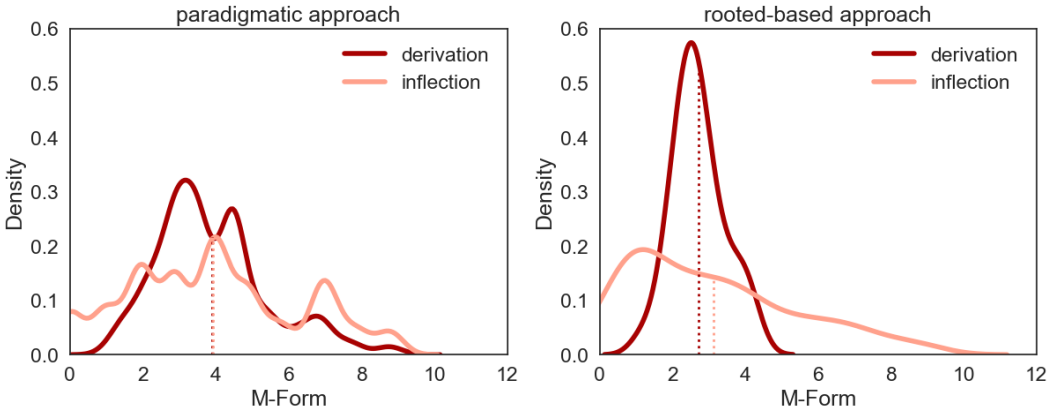
$$M_{\text{Form}} = \frac{1}{N}\sum_{i=1}^{N} \text{Levenshtein}(b_i, c_i)$$

- Measures the average edit distance between pairs of words in the same morphological relation.
- **Expectation**: higher distance for derivational than for inflectional relations.

Example						Cost
w	a	s	h	+	+	2 additions
w	a	s	h	e	d	
l	e	a	v	e		2 replaces, 1 addition
l	e	+	f	t		

$$M_{\text{Form}} = \frac{1}{2}(2+3) = \frac{1}{2}\cdot 5 = 2.5$$

Results: M_{Form}



Distributions differ, but overall, neither approach documents a difference in central tendency between inflection and derivation.

V_{Form} : Variability of the change in form

$$V_{\text{Form}} = \sum_{i=1}^{M} \sum_{j=1}^{M} F_{T_i} \cdot F_{T_j} \cdot \text{Levenshtein}(T_i, T_j)$$

- Measures the frequency-weighted edit distance between so-called *edit templates* constructed from pairs of words conveying the same morphological relation.
- **Expectation**: higher distance for derivational than for inflectional relations (Dressler 1989; Plank 1994).

Example **Template**

w	a	s	h	+	+
w	a	s	h	e	d

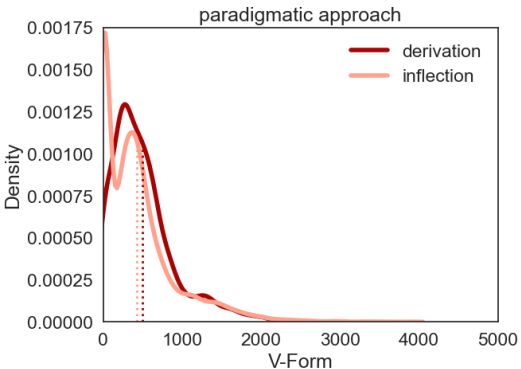
`_ed`

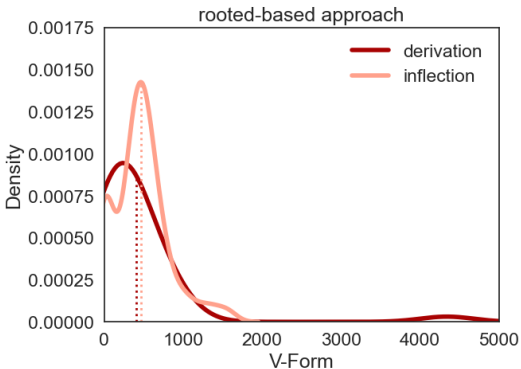
d	o	+
d	i	d

`_Xid`

$$V_{\text{Form}} = 0.2 \cdot 0.8 \cdot \text{Lev}(\texttt{\_ed}, \texttt{\_Xid}) + 0.2 \cdot 0.8 \cdot \text{Lev}(\texttt{\_Xid}, \texttt{\_ed}) = 32$$

Results: V_{Form}

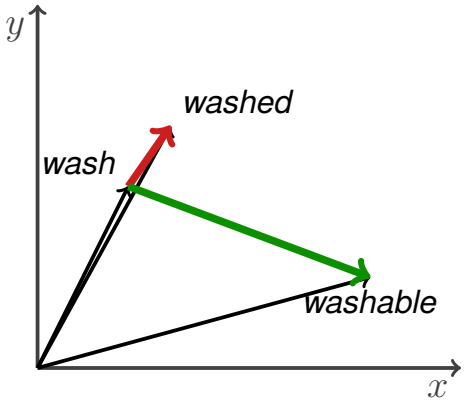




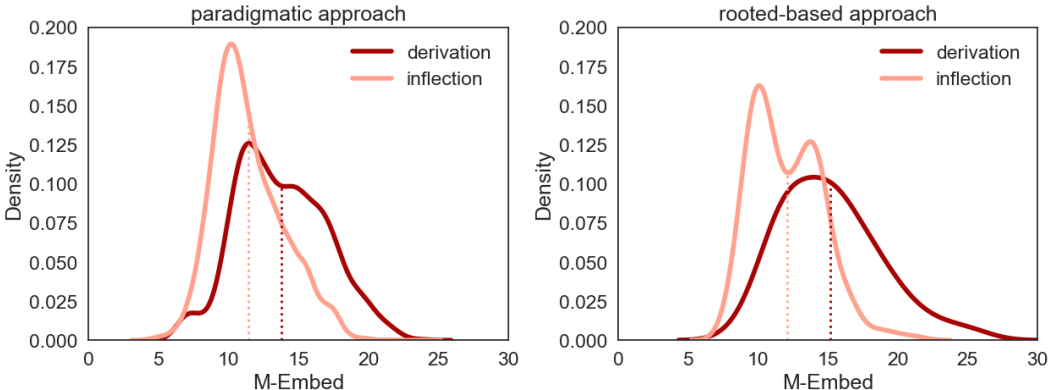
The two approaches lead to very similar results, finding no noticeable difference between inflection and derivation.

$$M_{\text{Embed}} = \frac{1}{N}\sum_{i=1}^{N} ||E(c_i) - E(b_i)||$$

- Measures the average distributional distance between pairs of words in the same morphological relation.
- **Expectation**: higher distance for derivational than for inflectional relations (Dressler 1989; Rosa and Žabokrtský 2019).



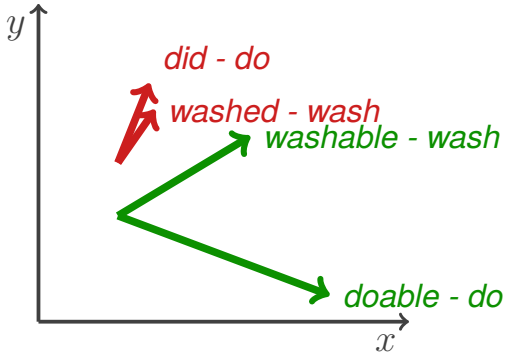
Results: M_{Embed}



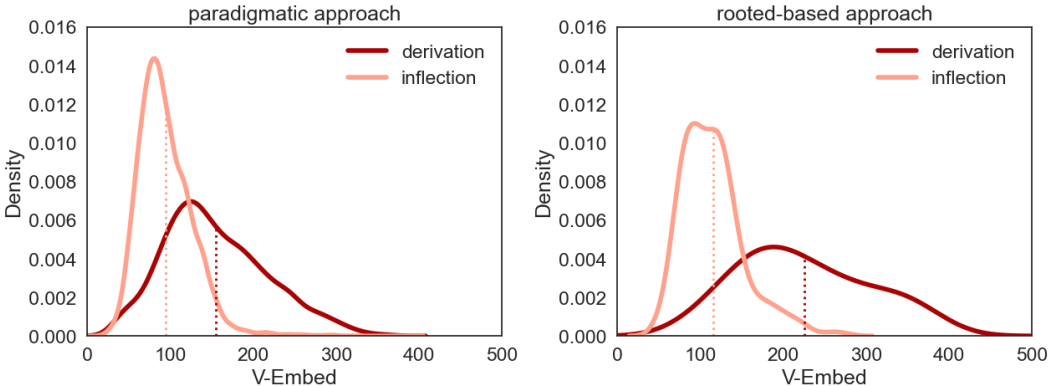
Both approaches find higher distances for derivation than inflection, with a sharper effect in the tree-based approach.

$$V_{\text{Embed}} = \sum_{k=1}^{K} \text{Var}(D_k, *)$$

- Measures the average dispersion of difference vectors between pairs of words in the same morphological relation.
- **Expectation**: higher dispersion for derivational than for inflectional relations (Bonami and Paperno 2018).



Results: V_{Embed}



Both approaches find higher dispersion for derivation than inflection, with a sharper effect in the tree-based approach.

CASE STUDY:

Diminutive formation

vs.

Gender paired personal nouns

- *Tree-based representation* seems to be relatively coarse compared to the *paradigmatic representation*. If an affix is polyfunctional, many word-formation meanings can be encoded in the same category; e.g.,

  učitel (*teacher*) $\rightarrow$ *učitel-ka* (*female teacher*) – masculine-feminine pair

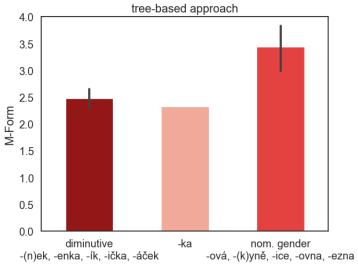
  skříň (*cupboard*) $\rightarrow$ *skříň-ka* (*small cupbard*) – diminutive

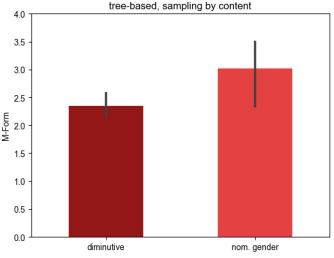
  both belonging to the tree-based category *N-N* with the affix *-ka*

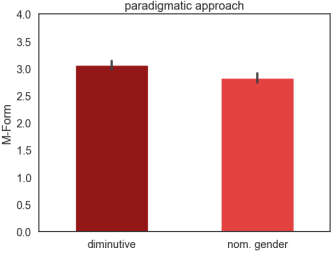
  What effect do both representations have on the results?

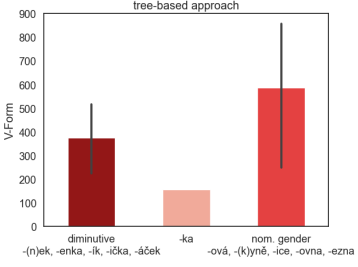
- We verified that and also added a *tree-based sampling by form* representation that distinguishes different *-ka*.

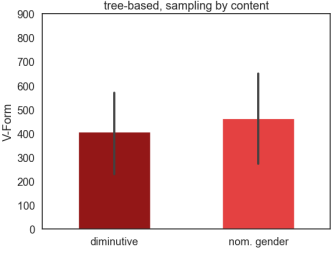
Results: M_{Form} & V_{Form}

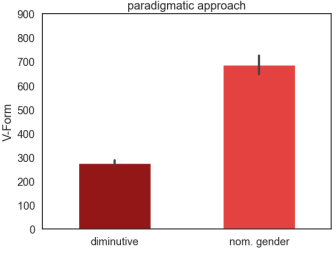




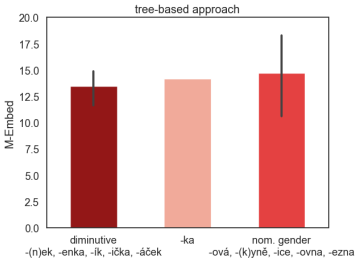


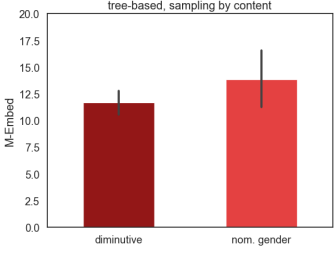


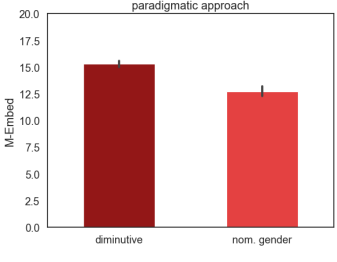


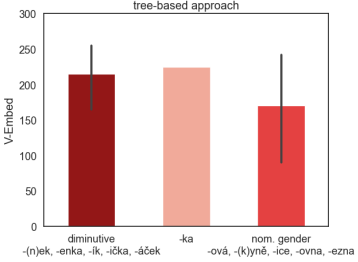


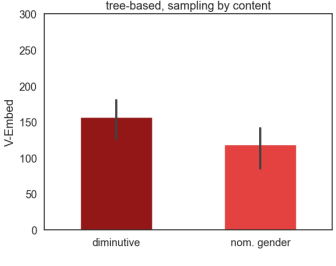
Results: M_{Embed} & V_{Embed}

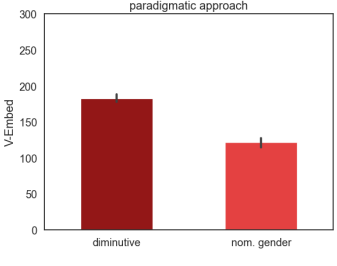












Discussion, Conclusions & Future work

Discussion, Conclusions and Future work

- The main part of the study shows:
  - There is no huge discrepancy between the approaches and measures when modelling inflection *vs.* derivation.
  - The tree-based approach seems to overestimate differences between inflection and derivation.
- The case study brings more interpretable data and shows the diversity.
  - Lumping of forms can be misleading (cf. middle graph).
  - Sometimes, the approaches disagree; sometimes, they do not…
- We showed that theoretical stances of how morphological data is organised shape empirical generalisations. However, more research needs to be done, not across languages, but into the details of this area.

Albright, Adam (2003). “A quantitative study of Spanish paradigm gaps”. In: *WCCFL 22 Proceedings*.

Anderson, Stephen R. (1982). “Where’s morphology?” In: *Linguistic Inquiry* 13, pp. 571–612.

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. The MIT Press. ISBN: 0-262-51017-0.

Bochner, Harry (1993). *Simplicity in Generative Morphology*. Berlin: Mouton de Gruyter.

Bonami, Olivier and Denis Paperno (2018). “Inflection vs. derivation in a distributional vector space”. In: *Lingue e Linguaggio* 17, pp. 173–195. URL: *https://halshs.archives-ouvertes.fr/halshs-01957367*.

Bonami, Olivier and Jana Strnadová (2019). “Paradigm structure and predictability in derivational morphology”. In: *Morphology* 29.2.

Booij, Geert (1996). “Inherent versus contextual inflection and the split morphology hypothesis”. In: *Yearbook of Morphology 1995*. Ed. by Geert Booij and Jaap van Marle. Dordrecht: Springer Netherlands, pp. 1–16. ISBN: 978-94-017-3716-6. DOI: *10.1007/978-94-017-3716-6_1*. URL: *https://doi.org/10.1007/978-94-017-3716-6_1*.

Boyé, Gilles (Jan. 2000). “Problèmes de morpho-phonologie verbale en français, en espagnol et en italien”. Theses. Université Paris-Diderot - Paris VII. URL: *https://theses.hal.science/tel-00276756*.

Copot, Maria et al. (2022). “Idiosyncratic Frequency as a Measure of Derivation vs. Inflection”. In: *Journal of Language Modelling* 10.2.

Corbett, Greville G. (2010). “Canonical derivational morphology”. In: *Word Structure* 3.2, pp. 141–155. DOI: *10.3366/word.2010.0002*. eprint: *https://doi.org/10.3366/word.2010.0002*. URL: *https://doi.org/10.3366/word.2010.0002*.

Dressler, Wolfgang U. (1989). “Prototypical Differences between Inflection and Derivation”. In: *STUF - Language Typology and Universals* 42.1, pp. 3–10. DOI: *doi:10.1515/stuf-1989-0102*. URL: *https://doi.org/10.1515/stuf-1989-0102*.

Hajič, Jan et al. (2020). *MorfFlex CZ 2.0*. URL: *http://hdl.handle.net/11234/1-3186*.

Haley, Coleman et al. (Dec. 2024). “Corpus-based measures discriminate inflection and derivation cross-linguistically”. In: *Journal of Language Mod-*

elling 12.2, pp. 477–529. DOI: 10.15398/jlm.v12i2.351. URL: https://jlm.ipipan.waw.pl/index.php/JLM/article/view/351.

Haspelmath, Martin (1996). “Word-class-changing inflection and morphological theory”. In: *Yearbook of Morphology 1995*. Ed. by Geert Booij and Jaap van Marle. Dordrecht: Springer Netherlands, pp. 43–66. ISBN: 978-94-017-3716-6. DOI: 10.1007/978-94-017-3716-6_3. URL: https://doi.org/10.1007/978-94-017-3716-6_3.

Křen, Michal et al. (2021). *SYN v9*. URL: http://hdl.handle.net/11234/1-4635.

Kyjánek, Lukáš and Olivier Bonami (2022). *Package of word embeddings of Czech from a large corpus*. URL: http://hdl.handle.net/11234/1-4920.

Matthews, P. H. (1972). *Inflectional Morphology. A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge: Cambridge University Press.

Mikolov, Tomáš et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Plank, Frans (1994). *Inflection and derivation*.

Robins, R. H. (1959). “In defense of WP”. In: *Transactions of the Philological Society* 58, pp. 116–144.

Rosa, Rudolf and Zdeněk Žabokrtský (Sept. 2019). “Attempting to separate inflection and derivation using vector space representations”. In: *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Prague, Czechia: Charles University, Faculty of

Mathematics, Physics, Institute of Formal, and Applied Linguistics, pp. 61–70. URL: https://aclanthology.org/W19-8508.

Spencer, Andrew (2013). *Lexical Relatedness*. Oxford University Press, p. 448. ISBN: 9780199679928.

Štekauer, Pavol (2015). “14. The delimitation of derivation and inflection”. In: *Volume 1 Word-Formation: An International Handbook of the Languages of Europe*. Ed. by Peter O. Müller et al. De Gruyter Mouton, pp. 218–235. DOI: doi:10.1515/9783110246254-016. URL: https://doi.org/10.1515/9783110246254-016.

Vidra, Jonáš et al. (2021). *DeriNet 2.1*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathe-

matics and Physics, Charles University. URL: *http://hdl.handle.net/11234/1-3765*.