Word-Formation Theories VI & Typology and Universals in Word-Formation V

A Distributional Approach to Inflection *vs.* Derivation in Czech

Lukáš Kyjánek & Olivier Bonami

June 23–26, 2022

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

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

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

+ derivation
− inflection

− derivation
+ inflection

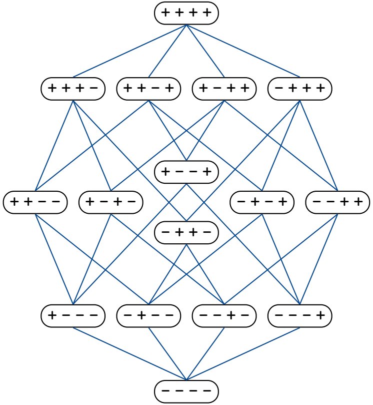
Monodimensional, categorical

canonical derivation

canonical inflection

Monodimensional, gradient

canonical derivation



canonical inflection

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]

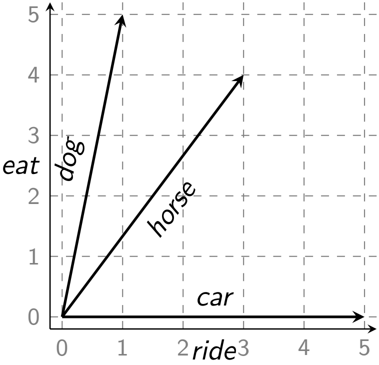
Discussion

Conclusion

Appendix

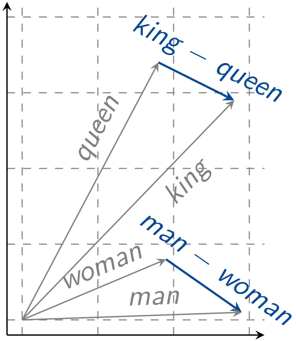
Distributional semantics

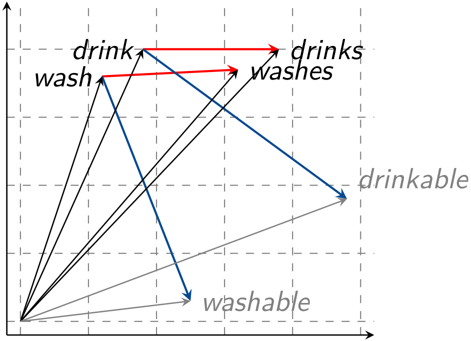
- The distributional hypothesis (see Harris 1954, Firth 1957) from (Lenci, 2008, p. 3): *"The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear."*
- Contemporary computational linguistics deduce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

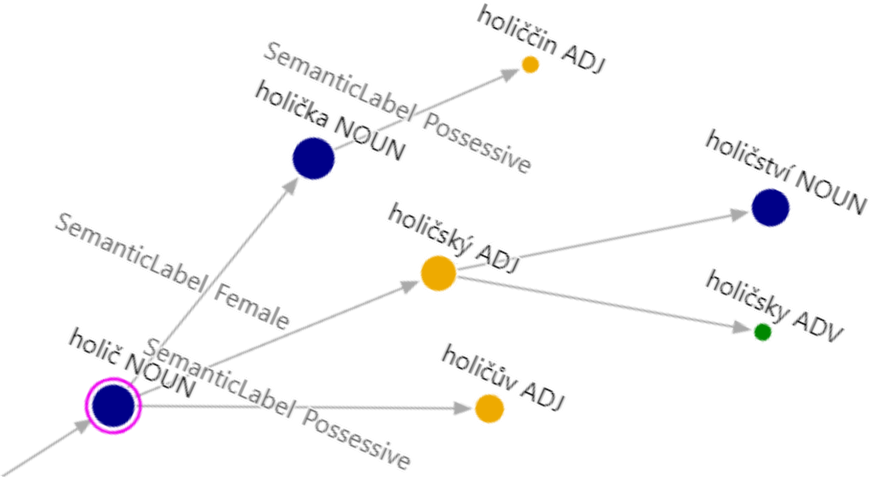
Morphological data

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'*

Lemma	Tag	Word form
holič	NNMS1-----A----	holič
holič	NNMS2-----A----	holiče
holič	NNMS3-----A----	holiči
holič	NNMS3-----A---1	holičovi
holič	NNMS4-----A----	holiče
holič	NNMS5-----A----	holiči
holič	NNMS6-----A----	holiči
holič	NNMS6-----A---1	holičovi
holič	NNMS7-----A----	holiče
holič	NNMP1-----A----	holiči
holič	NNMP2-----A----	holičů
holič	NNMP3-----A----	holičům
holič	NNMP4-----A----	holiče
holič	NNMP5-----A----	holiči
holič	NNMP6-----A----	holičích
holič	NNMP7-----A----	holiči

Example from DeriNet: derivation of *'barber'*



Semantic contrasts available for Czech

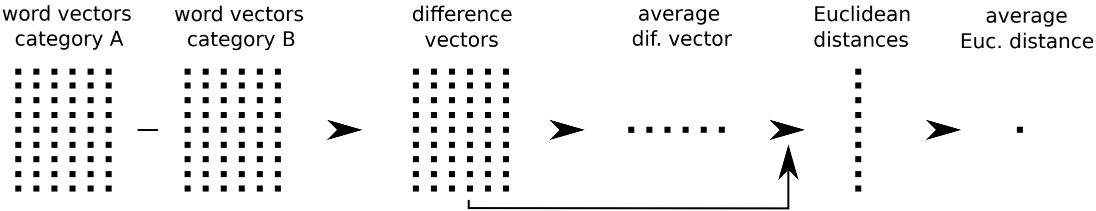
We process 24 types of morphological contrasts (difference vectors)

Word category A	**Word category B**	**Type of contrast**
Noun.**NOM**.FEM.SG	Noun.**GEN**.FEM.SG	core cases (N∼N)
Noun.**NOM**.FEM.PL	Noun.**GEN**.FEM.PL	core cases (N∼N)
...	...	...
Noun.**DAT**.FEM.SG	Noun.**LOC**.FEM.SG	non-core cases (N∼N)
Noun.**DAT**.FEM.PL	Noun.**LOC**.FEM.PL	non-core cases (N∼N)
...	...	...
Noun.**NOM**.FEM.SG	Noun.**DAT**.FEM.SG	mixed cases (N∼N)
Noun.**NOM**.FEM.PL	Noun.**DAT**.FEM.PL	mixed cases (N∼N)
...	...	...
Noun.NOM.FEM.SG	Noun.**DIM**.NOM.FEM.SG	diminutive (N∼N)
Noun.GEN.FEM.PL	Noun.**DIM**.GEN.FEM.PL	diminutive (N∼N)
...	...	...
Verb.inf	Noun.**AGENT**.NOM.MASC.SG	agent (V∼N)
Verb.inf	Noun.**AGENT**.NOM.MASC.PL	agent (V∼N)

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

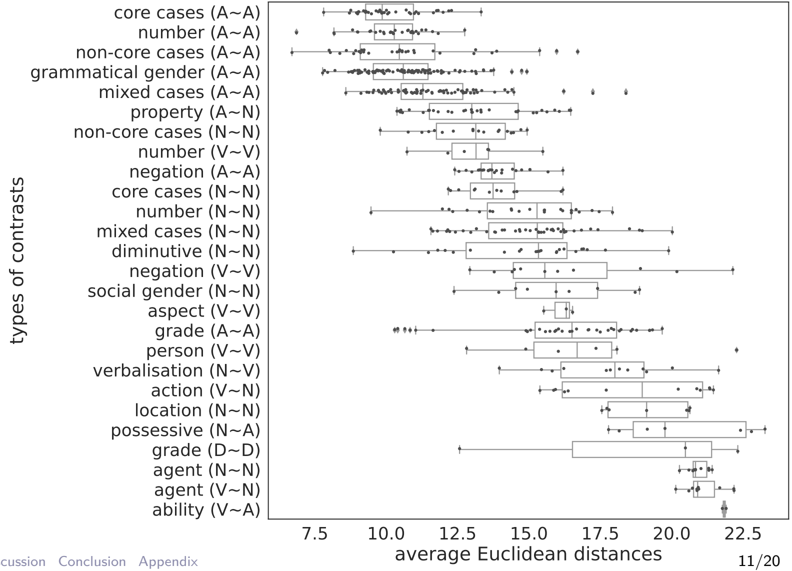
Method

- 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.



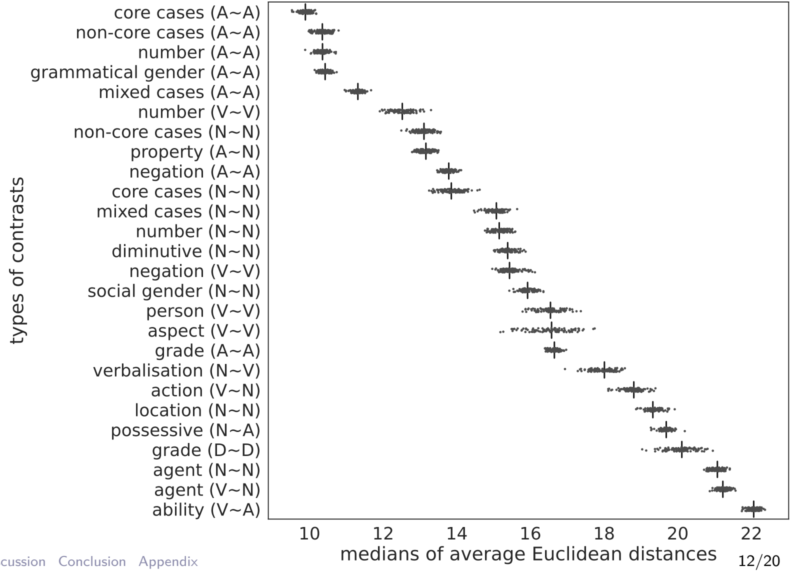
Results in a single sample

- 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.



Global overview of the inflection–derivation scale (monodimensional gradient)

- 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):

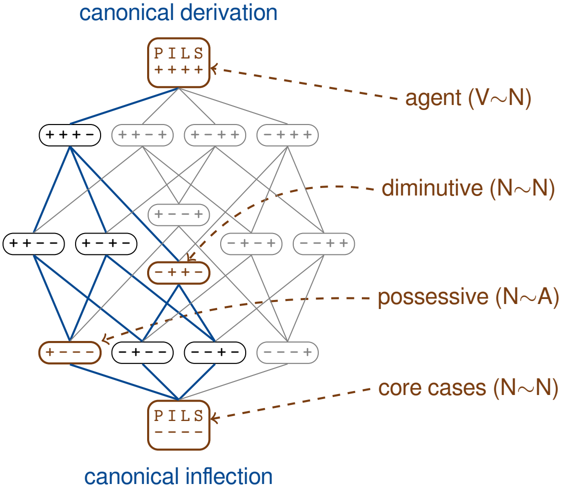
P+ different part of speech
I+ inherent (*vs.* contextual)
L+ different lexeme
S+ different semantic type (individual *vs.* eventuality *vs.* property)

Type of contrast	P	I	L	S
core cases (N∼N)	−	−	−	−
grammatical gender (A∼A)	−	−	−	−
…		…		
possessive (N∼A)	+	+	−	−
…		…		
diminutive (N∼N)	−	+	+	−
social gender (N∼N)	−	+	+	−
…		…		
action (V∼N)	+	+	+	−
property (A∼N)	+	+	+	−
…		…		
ability (V∼A)	+	+	+	+
agent (V∼N)	+	+	+	+

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%).



canonical derivation

P I L S
+ + + +

agent (V∼N)

+ + + −
+ + − +
+ − + +
− + + +

diminutive (N∼N)

+ − − +

+ + − −
+ − + −
− + + −
− + − +
− − + +

possessive (N∼A)

+ − − −
− + − −
− − + −
− − − +

core cases (N∼N)

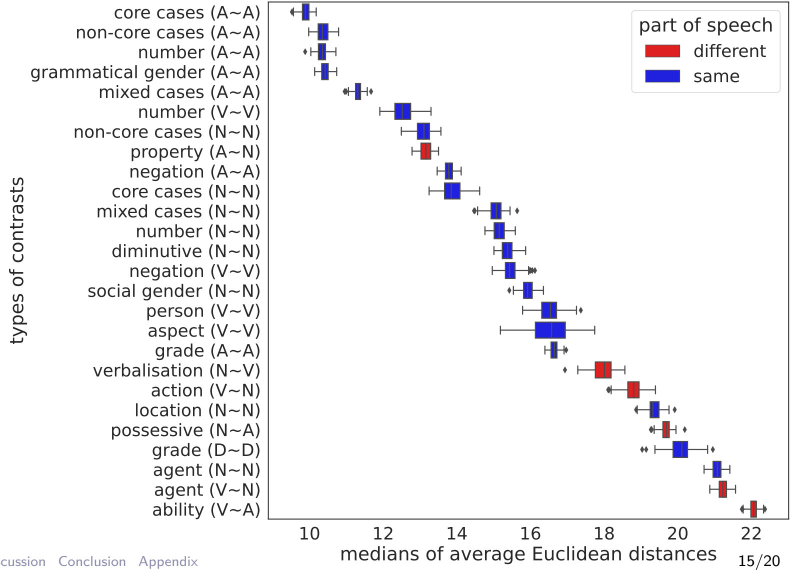
P I L S
− − − −

canonical inflection

Feature P : part of speech

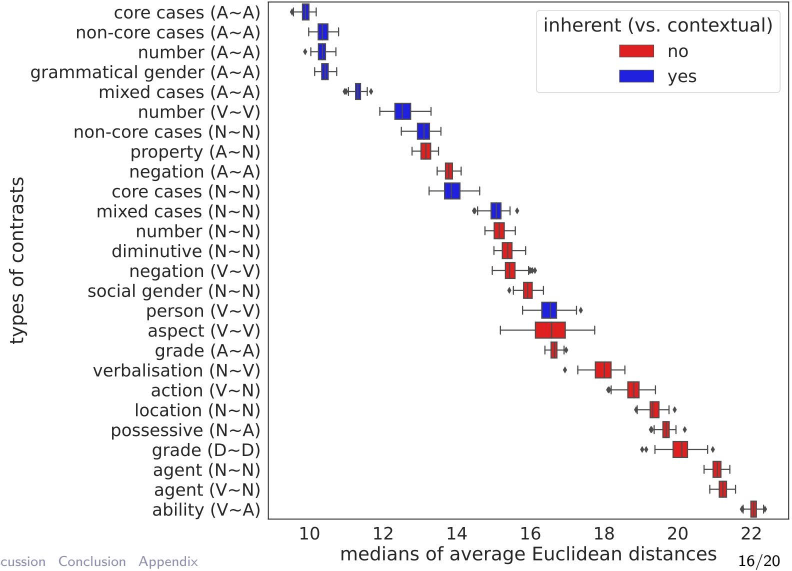
- Most of the pos-changing contrasts have higher distances, except for **property (A∼N)**.

- The opposite exceptions:
  - location (N∼N)
  - grade (D∼D)
  - agent (N∼N)



Feature I : inherent (*vs.* contextual)

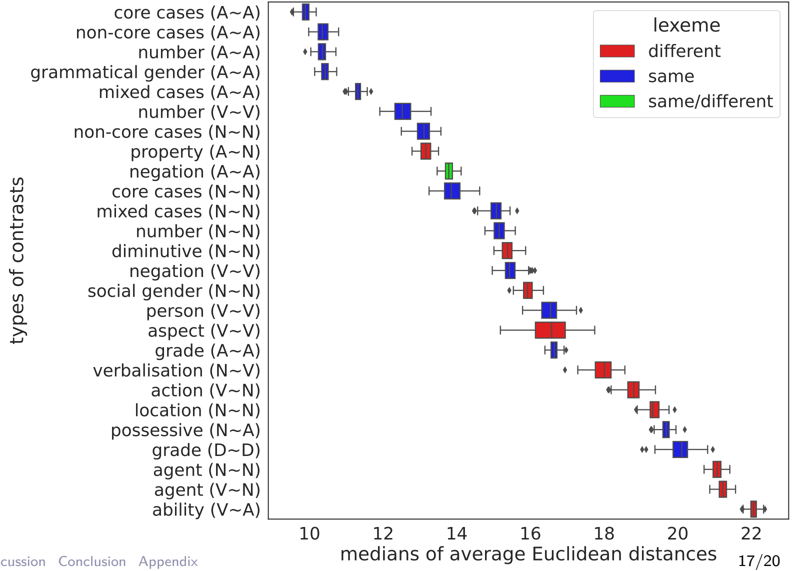
- Most of the canonical inflectional contrasts have lower distances, except for **person (V∼V)**.
- The opposite exceptions:
  - property (A∼N)
  - negation (A∼A)



Feature `L` : lexeme

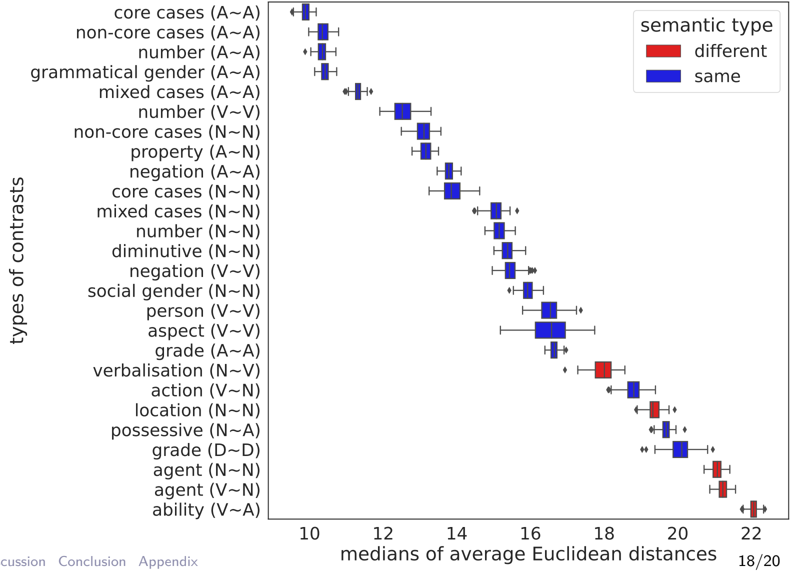
- Most of the contrasts represented by different lexemes have higher distances, except for **property (A∼N)**, **diminutive (N∼N)**, and **social gender (N∼N)**.

- The opposite exceptions:
  - person (V∼V)
  - grade (A∼A)
  - possessive (N∼A)
  - grade (D∼D)



Feature S : semantic type

- 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.



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.

core cases (A∼A)
non-core cases (A∼A)
number (A∼A)
grammatical gender (A∼A)
mixed cases (A∼A)
number (V∼V)
non-core cases (N∼N)
→ property (A∼N)
negation (A∼A)
core cases (N∼N)
mixed cases (N∼N)
number (N∼N)
diminutive (N∼N)
negation (V∼V)
social gender (N∼N)
person (V∼V)
aspect (V∼V)
grade (A∼A)
verbalisation (N∼V)
action (V∼N)
location (N∼N)
possessive (N∼A)
grade (D∼D)
agent (N∼N)
agent (V∼N)
ability (V∼A)

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.

core cases (A∼A)
non-core cases (A∼A)
number (A∼A)
grammatical gender (A∼A)
mixed cases (A∼A)
number (V∼V)
non-core cases (N∼N)
property (A∼N)
negation (A∼A)
core cases (N∼N)
mixed cases (N∼N)
number (N∼N)
diminutive (N∼N)
negation (V∼V)
social gender (N∼N)
→ person (V∼V)
aspect (V∼V)
grade (A∼A)
verbalisation (N∼V)
action (V∼N)
location (N∼N)
possessive (N∼A)
grade (D∼D)
agent (N∼N)
agent (V∼N)
ability (V∼A)

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.
- 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).

- core cases (A∼A)
- non-core cases (A∼A)
- → number (A∼A)
- grammatical gender (A∼A)
- mixed cases (A∼A)
- → number (V∼V)
- non-core cases (N∼N)
- property (A∼N)
- → negation (A∼A)
- core cases (N∼N)
- mixed cases (N∼N)
- → number (N∼N)
- diminutive (N∼N)
- → negation (V∼V)
- social gender (N∼N)
- person (V∼V)
- aspect (V∼V)
- → grade (A∼A)
- verbalisation (N∼V)
- action (V∼N)
- location (N∼N)
- possessive (N∼A)
- → grade (D∼D)
- agent (N∼N)
- agent (V∼N)
- ability (V∼A)

Conclusion

- 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.

References I

Stephen R. Anderson. Where's morphology? *Linguistic Inquiry*, 13:571–612, 1982.

Laurie Bauer. The function of word-formation and the inflection-derivation distinction. In *Words in their Places. A Festschrift for J. Lachlan Mackenzie*, pages 283–292. Vrije Universiteit, Amsterdam, 2004.

Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234, 2020. doi: 10.1146/annurev-linguistics-011619-030303. URL https://doi.org/10.1146/annurev-linguistics-011619-030303.

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

Geert Booij. Inherent versus contextual inflection and the split morphology hypothesis. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1995*, pages 1–16. Springer Netherlands, Dordrecht, 1996. 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.

Greville G. Corbett. Canonical derivational morphology. *Word Structure*, 3(2):141–155, 2010. doi: 10.3366/word.2010.0002. URL https://doi.org/10.3366/word.2010.0002.

Wolfgang U. Dressler. Prototypical differences between inflection and derivation. *STUF - Language Typology and Universals*, 42(1):3–10, 1989. doi: doi:10.1515/stuf-1989-0102. URL https://doi.org/10.1515/stuf-1989-0102.

J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. MorfFlex CZ 2.0, 2020. URL http://hdl.handle.net/11234/1-3186. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL https://doi.org/10.1080/00437956.1954.11659520.

Martin Haspelmath. Word-class-changing inflection and morphological theory. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1995*, pages 43–66. Springer Netherlands, Dordrecht, 1996. 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.

References II

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. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Rudolf Rosa and Zdeněk Žabokrtský. Attempting to separate inflection and derivation using vector space representations. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 61–70, Prague, Czechia, September 2019. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. URL https://aclanthology.org/W19-8508.

Andrew Spencer. *Lexical Relatedness*. Oxford University Press, 2013. ISBN 9780199679928.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. DeriNet 2.1, 2021. URL http://hdl.handle.net/11234/1-3765. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pavol Štekauer. 14. the delimitation of derivation and inflection. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Volume 1 Word-Formation: An International Handbook of the Languages of Europe*, pages 218–235. De Gruyter Mouton, 2015. doi: doi:10.1515/9783110246254-016. URL https://doi.org/10.1515/9783110246254-016.

Appendix A: Euclidean distance *vs.* Cosine distance I

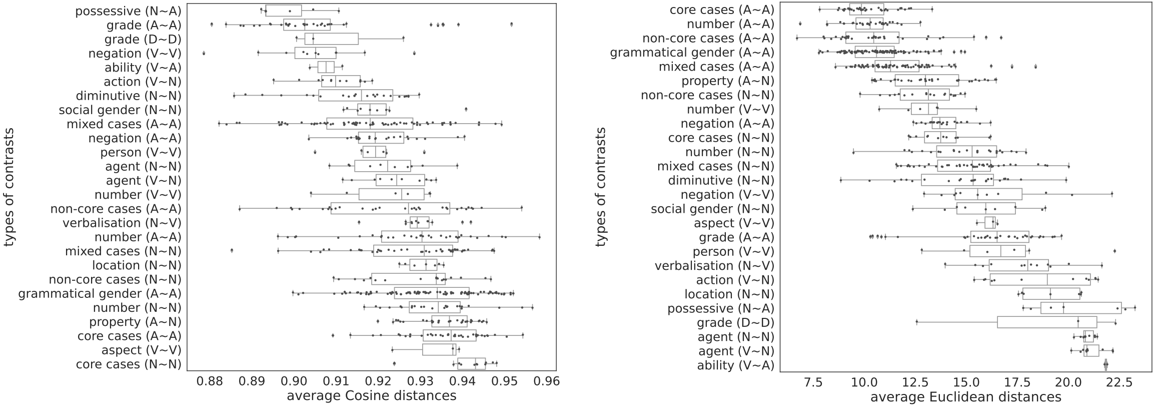


Figure: Prototypical sample before bootstrapping (from left: Cosine, Euclidean distances).

Appendix A: Euclidean distance *vs.* Cosine distance II

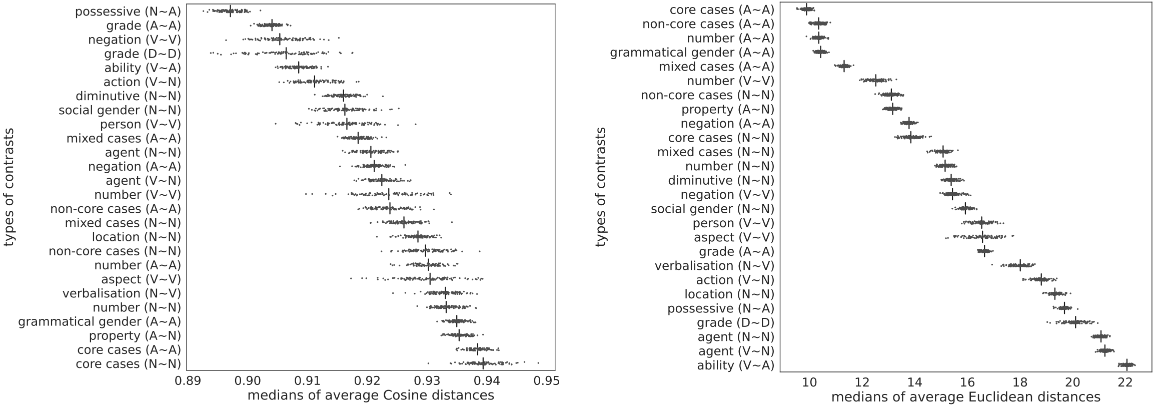


Figure: Bootstrapping (from left: Cosine, Euclidean distances).

Appendix A: Euclidean distance *vs.* Cosine distance III

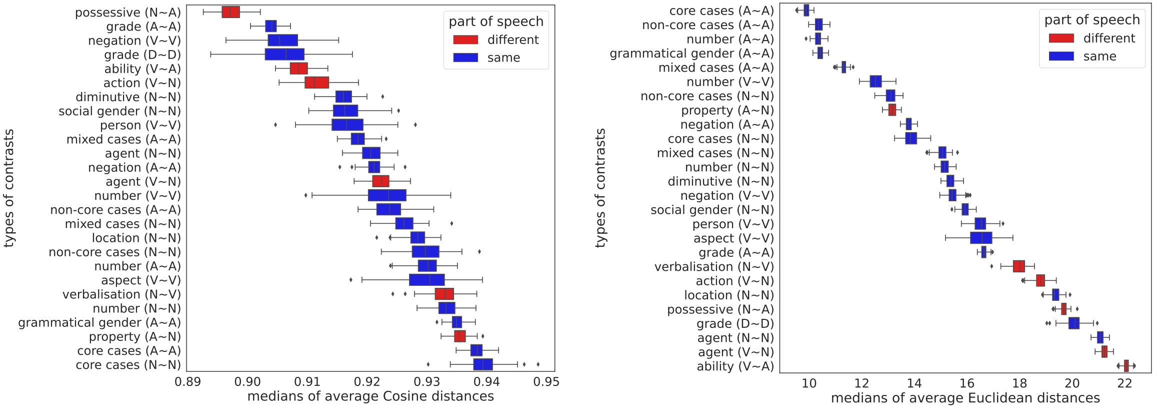


Figure: Bootstrapping, feature POS (from left: Cosine, Euclidean distances).

Appendix A: Euclidean distance *vs.* Cosine distance IV

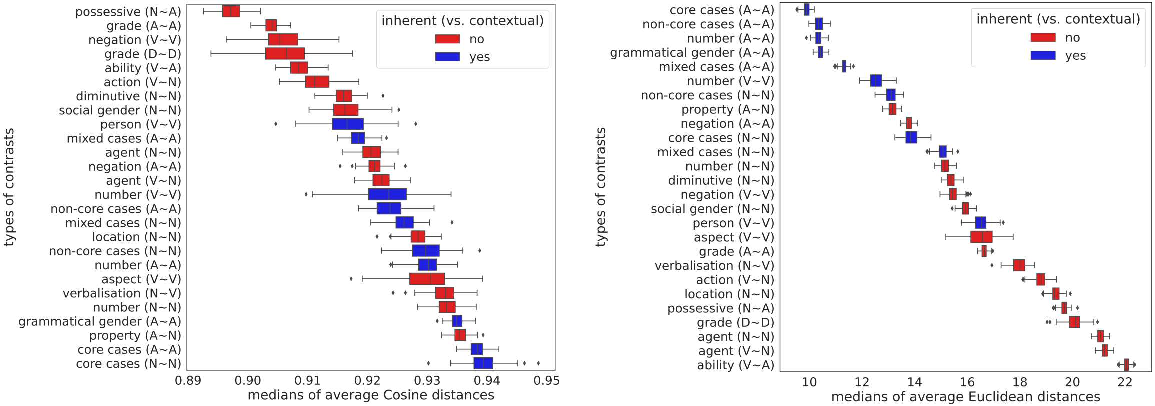


Figure: Bootstrapping, feature INHERENT (from left: Cosine, Euclidean distances).

Appendix A: Euclidean distance *vs.* Cosine distance V

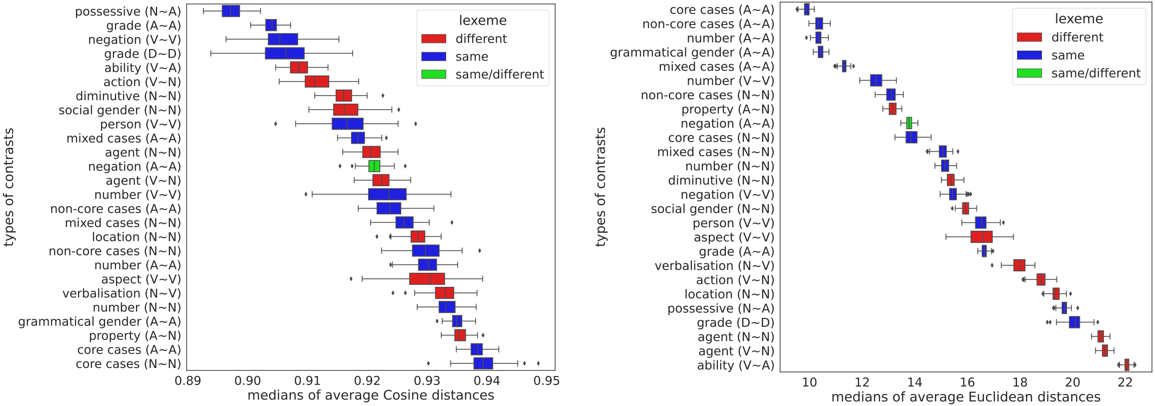


Figure: Bootstrapping, feature LEXEME (from left: Cosine, Euclidean distances).

Appendix A: Euclidean distance *vs.* Cosine distance VI

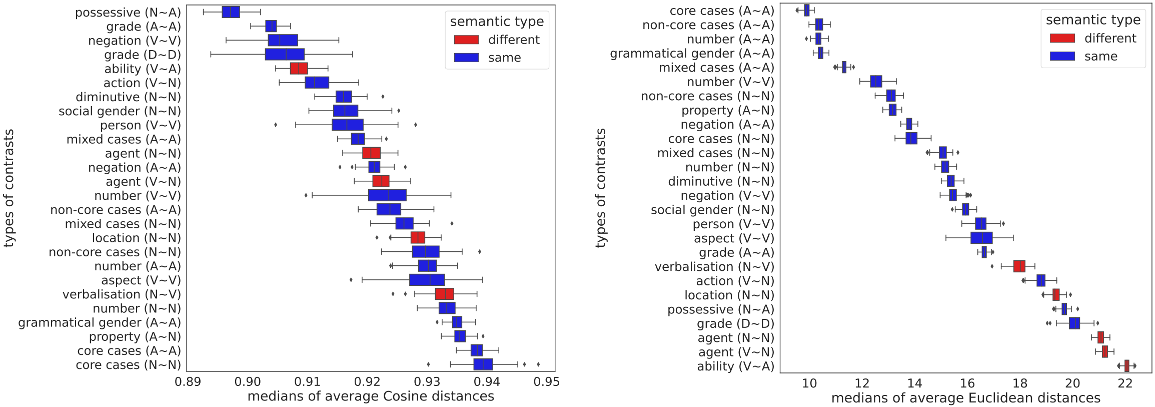


Figure: Bootstrapping, feature SEMANTIC TYPE (from left: Cosine, Euclidean distances).

Appendix B: Comparison of the features

Feature expectation	**Counts**	
Cosine correct	98	42%
Cosine incorrect	135	58%
Euclidean correct	**190**	**82%**
Euclidean incorrect	43	18%
Same correct results	63	27%
Same incorrect results	8	3%
Different results	162	70%

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