Morphosyntactic features in distributional space

Olivier Bonami & Marine Wauquier & Lukáš Kyjánek

July 8, 2022

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

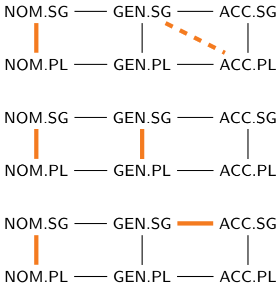
Université Paris Cité
Laboratoire de Linguistique Formelle
Centre National de la Recherche Scientifique

Featurally structured paradigms

- Many authors define inflectional paradigms in terms of their organization into orthogonal features, cf. Wunderlich and Fabri (1995, p. 266):

"A paradigm is an n-dimensional space whose dimensions are the attributes (or features) used for the classification of word forms. In order to be a dimension, an attribute must have at least two values. The cells of this space can be occupied by word forms of appropriate categories."

- Implicit assumptions:
  - Some pairs of forms in a paradigms are in direct pairwise contrast, while others are not.
  - Some contrasts within the paradigm are parallel in that they involve the same variation in the same feature(s).
  - Some contrasts within the paradigm are orthogonal in that they involve variation in different features.



Limitations of feature orthogonality I

- Evidently, some situations do not lead to a system of orthogonal features.
  - Neutralization: a dimension that disappears for some feature values.
    E.g. Russian verbs (and adjectives):

	SG	PL
MAS	igral	
FEM	igrala	igrali
NEU	igralo	

Past forms of IGRÁT 'play'

  - Clusivity: a dimension that only makes sense for some feature values.
    E.g. Thulung verbs:

	SG	DU	PL	
1	buŋu	butsi	bui	INCL
		butsuku	buku	EXCL
2	buna	butsi	buni	
3	bu	butsi	buni	

Nonpast forms of BUMU 'be'

Limitations of feature orthogonality II

- Morphomic paradigm organization: systematic syncretisms are not featurally organized. E.g. English verbs:

NONFINITE	
INF	give
PRS.PTCP	giving
PST.PTCP	given

PRESENT		
	SG	PL
1	give	give
2	give	give
3	gives	give

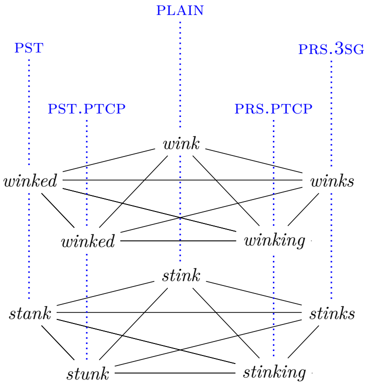
PAST		
	SG	PL
1	gave	gave
2	gave	gave
3	gave	gave

Alternatives

- A general definition should not require orthogonality.

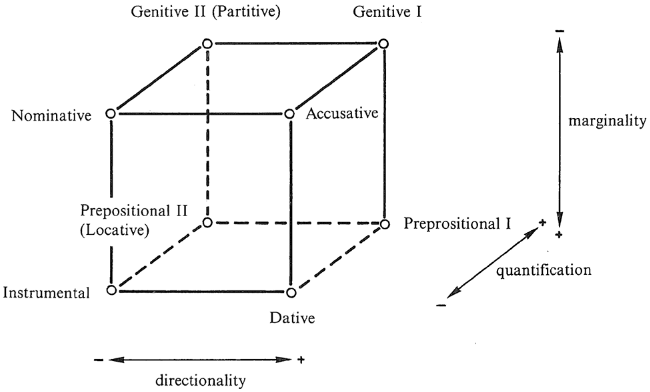
  > *"[...] we define the paradigm of a lexeme L as a complete set of cells for L, where each cell is the pairing of L with a complete and coherent morphosyntactic property set (MPS) for which L is inflectable."* (Stump and Finkel, 2013, p. 9)

- Bonami and Strnadová (2019) go further, building on Štekauer (2014):
  - Paradigms are be defined abstractively in terms of aligned pairwise contrasts
  - Analysis into orthogonal features is a further step of abstraction that is neither necessary nor always insightful.
- Hence the relationship between features and paradigms is a matter of current theoretical interest.



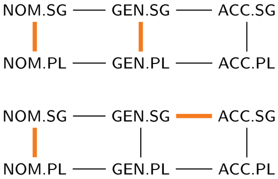
Interesting empirical questions

- Are conventional parallel contrasts really parallel?
  - Benveniste on 1SG vs. 1PL
  - Polite plurals, French *on*, etc.
- Do innovative featural analyses reflect parallel contrasts?
  - Jakobson's (1958) cube



The topic for today

- Can we find empirical evidence to support the idea that some contrasts are parallel, while others are orthogonal?



- Strategy: model contrasts between paradigm cells as contrasts between the corresponding word vectors
  - This should reflect both syntactic and semantic aspects of the relevant contrasts.

Types of contrast

- Given two cells a and b, modelled as sets of *feature* : *value* pairing:
  - $S(a,b)$ denotes the set of feature values specific to a when compared to b, i.e. $S(a,b) \overset{def}{=} \{v \mid f : v \in a \wedge \neg f : v \in b\}$
  - $C(a,b)$ denotes the set of features for which a and b contrast, i.e. $C(a,b) \overset{def}{=} \{f | \exists v \exists w [f : v \in a \wedge f : w \in b \wedge v \neq w]\}$
- Given two pairs of contrasting cells, (a,b) and (a',b'):
  1. (a,b) and (a',b') are parallel iff they contrast in exactly the same way, i.e. $S(a,b) = S(a',b') \wedge S(b,a) = S(b',a')$.

  2. (a,b) and (a',b') are orthogonal iff they do not contrast at all in the same way, i.e. $C(a,b) \cap C(a',b') = \emptyset$.

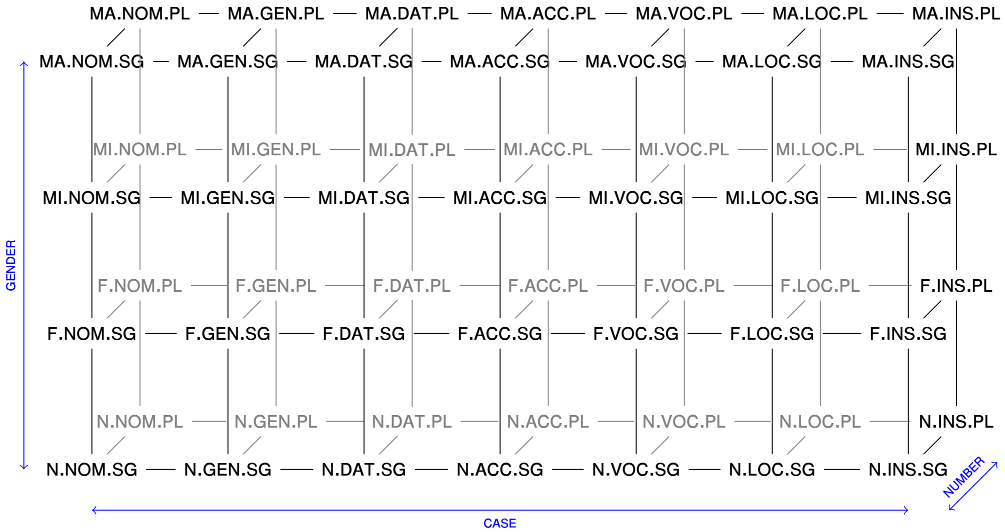
  3. (a,b) and (a',b') form a corner iff $a = a'$ or $a = b'$ or $b = a'$ or $b = b'$.

  4. (a,b) and (a',b') are not comparable iff they contrast in the same features but not the same values, i.e. $C(a,b) = C(a',b') \wedge (S(a,b) \neq S(a',b') \vee S(b,a) \neq S(b',a'))$.

Predictions

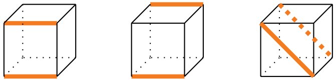
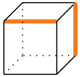
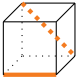
- If two pairs of cells are featurally parallel, the corresponding pairs of vectors will contrast in similar ways.
  - Possibly, they contrast in exactly the same way.
- If two pairs of cells are orthogonal, the corresponding pairs of vectors will contrast in completely different ways.
  - At the very least, they contrast in more different ways than parallel pairs.
- For corner cases, we expect odd behaviors due to sharing a cell: we exclude them from consideration.
- For non comparable cases, we have no prediction: we exclude them from consideration.

Adding dimensions (e.g. Czech adjectives)



Types of contrast in three dimensions

- With more dimensions, new situations arise:

  1. Parallel: 
  2. Orthogonal: 
  3. Neither: 

- Suggests that we need to define a gradient degree of parallelism, the proportion of contrasts shared between two pairs of cells:

$$D(p, p') = \frac{|C(a, b) \cap C(a', b')|}{|C(a, b) \cup C(a', b')|}$$

This will be 1 in case of parallelism, 0 in case of orthogonality, and take intermediate values.

Prediction

- There is a monotonous relation between the degree of parallelism between pairs of cells and the similarity of the corresponding distributional contrasts: the more parallel in terms of feature, the more distributionally parallel.

Outline

Training the model of distributional semantics for Czech

- We train the semantic representations of words by applying **Word2vec** (Mikolov et al., 2013) to **SYN v9 corpus** (Křen et al., 2021).
- SYN v9 corpus
  - large representative corpus of Czech
  - 362M sentences; 4,719M tokens; 7.3M lemmas
  - tagged by MorphoDiTa (accuracy above 95%; Straková et al., 2014)
- Semantic representations (vectors) are trained for combinations of tokens and tags; we rely on the corpus pos-tag annotations.

Existing morphological data resources for Czech

- We use data from **MorfFlexCZ 2.0** (Hajič et al., 2020).
- MorfFlexCZ 2.0
  - inflectional morphological lexicon
  - 125.3M lemma-tag-wordform triples
- Its data has served for a development of MorphoDiTa (tagging SYN v9 corpus).
- We exploit the data when creating samples for our two studies.

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

(1) Classifying contrasting word vectors

- **Data:** combinations of two samples of unpaired words for the studied inflectional contrasts
- **Task:** binary classification of a target word on the basis of its vector
- **Evaluation:**
  - **intrinsic** assesses discriminative power of a given feature for classifying word vectors
  - **extrinsic** assesses stability of classifying word vectors in a different context

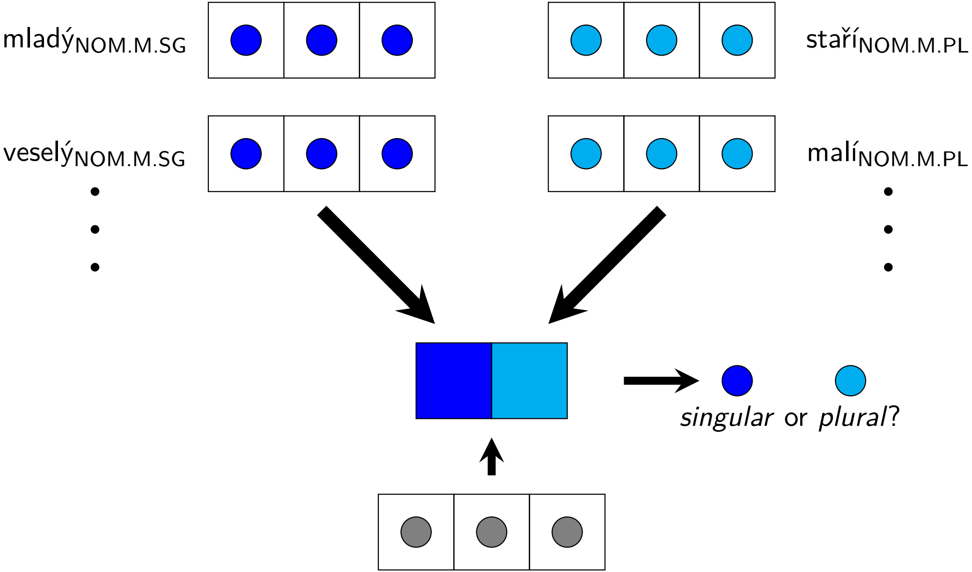
Sampling research data for classification study

- 500 word vectors (only words with freq>50 in SYN v9) for each studied inflectional category were sampled from SYN v9.
- It resulted in 30 samples for nouns and 30 samples for adjectives; combinations of gram.
  - cases [NOM, GEN, ACC],
  - numbers [SG, PL], and
  - genders [MASC.ANIM, MASC.INANIM, FEM, NEUT] (only for adjectives).

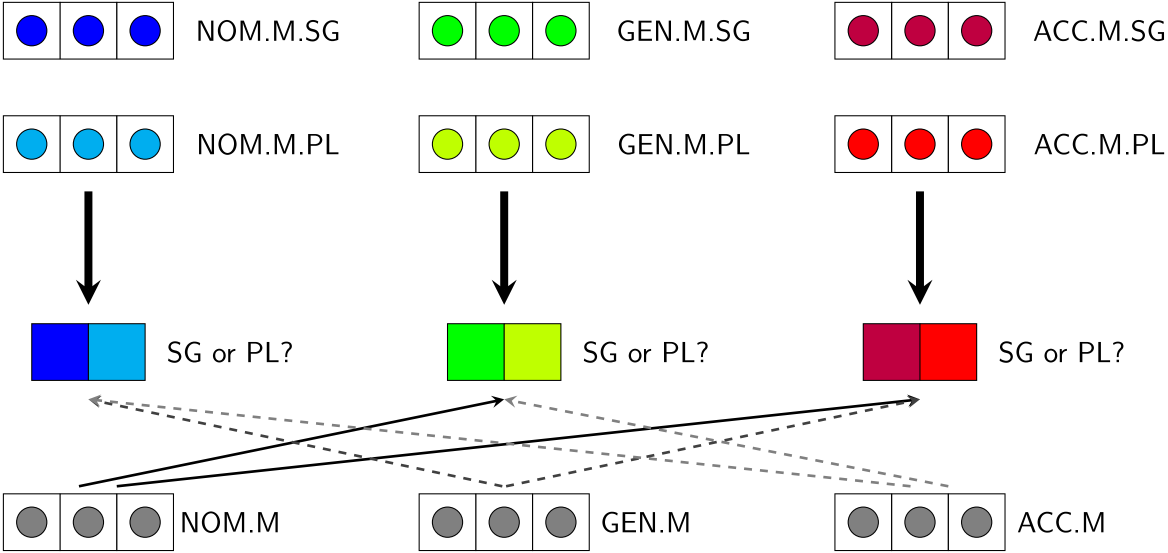
Example for the category *'NFS1'* (NOUN.FEM.SG.NOM).

Word	**Vector**
`pastelka>NNFS1-----A----`	*100-dim vector*
`tichost>NNFS1-----A----`	*100-dim vector*
`meduňka>NNFS1-----A----`	*100-dim vector*
...	...
`práce>NNFS1-----A----`	*100-dim vector*
`letargie>NNFS1-----A----`	*100-dim vector*
`paměť>NNFS1-----A----`	*100-dim vector*

Intrinsic classification task



Extrinsic prediction task

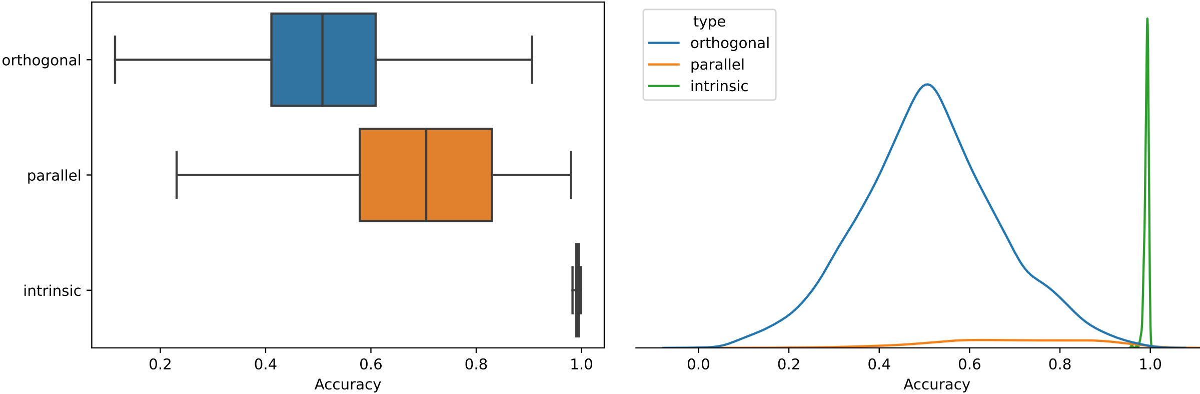


Method

- We train classifiers with gradient boosting (Friedman, 2001, Mason et al., 2000) applied on decision trees
  - 500 estimators, learning rate of 0.01, max depth of 2, random state of 0, and 'deviance' as the loss function
  - 1000 unpaired words (500 by condition)
- Intrinsic classification is evaluated by means of 10-fold cross validation on the 1000-word dataset
- Extrinsic classification is by means of a confusion matrix based on aligned labels (eg. SG for both masculine and feminine nominative adjectives)

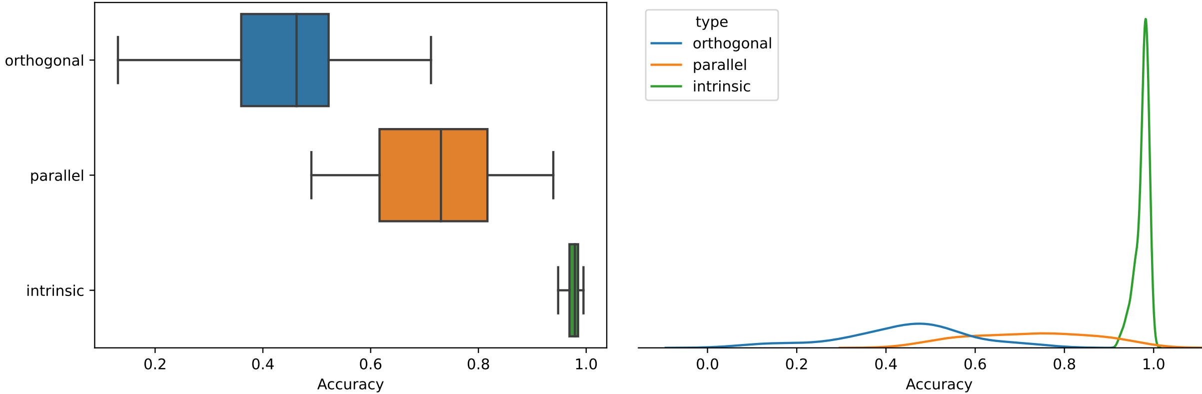
Classification results I

- Distribution of classification of contrasts for adjectives, by type



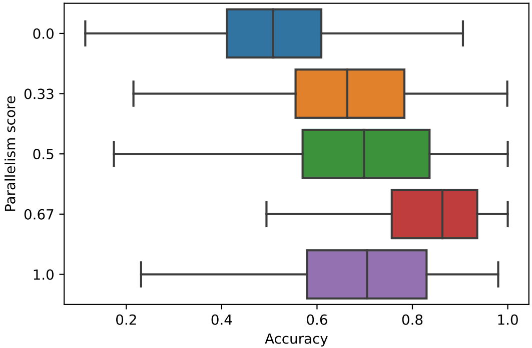
Classification results II

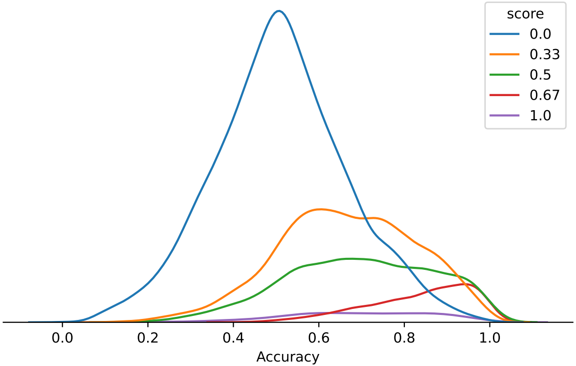
- Distribution of classification of contrasts for nouns, by type



Classification results III

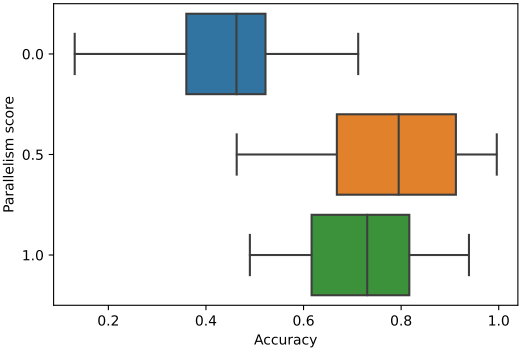
- Distribution of classification of contrasts for adjectives, by parallelism score

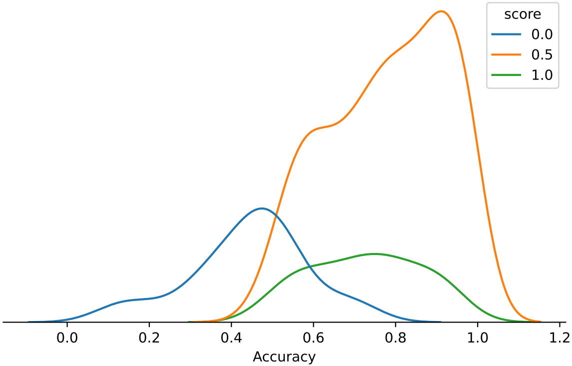




Classification results IV

- Distribution of classification of contrasts for nouns, by parallelism score





(2) Predicting relations between word vectors

- **Data**: samples of pairs of word vectors for the studied inflectional contrasts
- **Task**: to predict a target word vector on the basis of a source word vector
- **Evaluation**:
  - **intrinsic** assesses discriminative power for predicting word vectors
    - 10-fold cross-validation
    - prediction of the same contrast as for the one on which the model was trained
  - **extrinsic** assesses stability of predicting word vectors in different context
    - prediction of different contrasts than the one on which the model was trained

Sampling research data for prediction study

- 1000 pairs of word vectors (only words with freq>50 in SYN v9) for each studied inflectional contrast were sampled from SYN v9 (linked by lemmas from MorfFlexCZ).
- It resulted in 60 samples for nouns and 276 for adjectives; combinations of gram.
  - cases [NOM, GEN, ACC],
  - numbers [SG, PL], and
  - genders [MASC.ANIM, MASC.INANIM, FEM, NEUT] (only for adjectives).

Example for the contrast *'NF(PS)1'* (NOUN.FEM.SG.NOM ∼ NOUN.FEM.PL.NOM).

Word A	**Word B**	**Vector A**	**Vector B**
`výpůjčka>NNFS1-----A----`	`výpůjčky>NNFP1-----A----`	*100-dim vector*	*100-dim vector*
`hmotnost>NNFS1-----A----`	`hmotnosti>NNFP1-----A----`	*100-dim vector*	*100-dim vector*
`nádrž>NNFS1-----A----`	`nádrže>NNFP1-----A----`	*100-dim vector*	*100-dim vector*
...	...	...	...
`rosa>NNFS1-----A----`	`rosy>NNFP1-----A----`	*100-dim vector*	*100-dim vector*
`dojnice>NNFS1-----A----`	`dojnice>NNFP1-----A----`	*100-dim vector*	*100-dim vector*
`líheň>NNFS1-----A----`	`líhně>NNFP1-----A----`	*100-dim vector*	*100-dim vector*

Predicting vectors

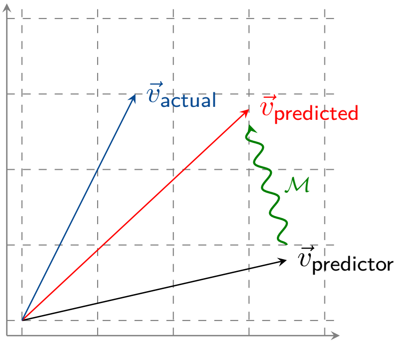
- Following Marelli and Baroni (2015), we train one linear model per dimension in the target vector: each model predicts one dimension in the target from all dimensions in the predictor.

$$
\begin{aligned}
\texttt{target\_val\_1} &\sim \texttt{pred\_val\_1} + \texttt{pred\_val\_2} + \cdots + \texttt{pred\_val\_100} \\
\texttt{target\_val\_2} &\sim \texttt{pred\_val\_1} + \texttt{pred\_val\_2} + \cdots + \texttt{pred\_val\_100} \\
&\vdots \\
\texttt{target\_val\_100} &\sim \texttt{pred\_val\_1} + \texttt{pred\_val\_2} + \cdots + \texttt{pred\_val\_100}
\end{aligned}
$$

Evaluating prediction accuracy

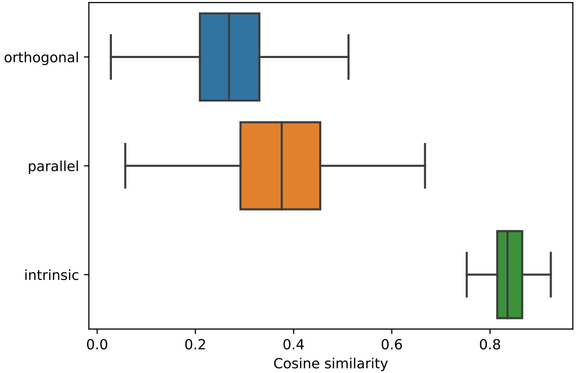
- We then measure how good the model collection $\mathcal{M}$ is at capturing the semantics of the morphological relation by examining the cosine between the predicted and the actual target vector.

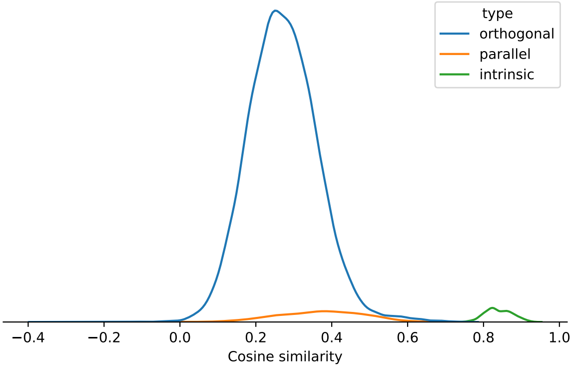
- The average value of $\cos(\vec{v}_{\text{predicted}}, \vec{v}_{\text{actual}})$ is indicative of how predictable the meaning of targets is from that of predictors for that particular morphological relation.



Vector prediction results I

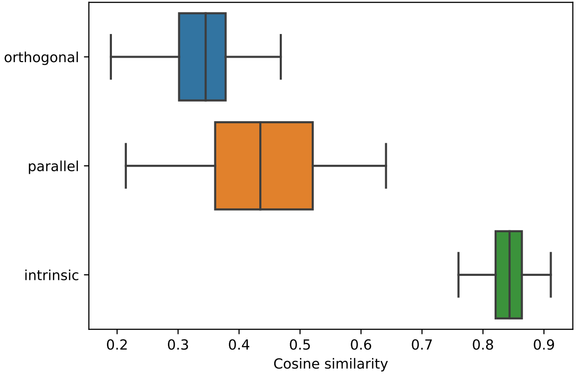
- Distribution of quality of prediction for adjectives, by type

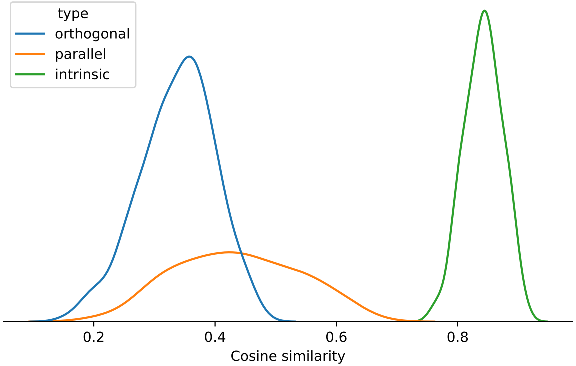




Vector prediction results II

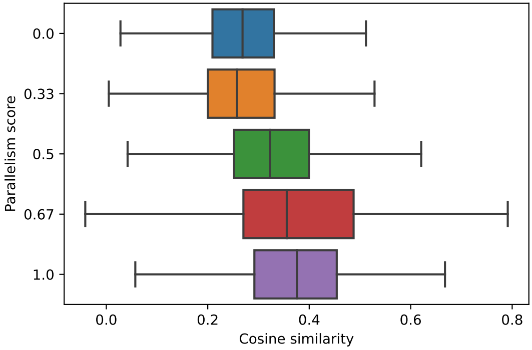
- Distribution of quality of prediction for nouns, by type

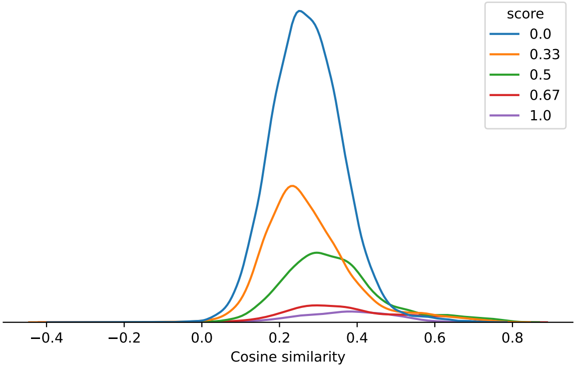




Vector prediction results III

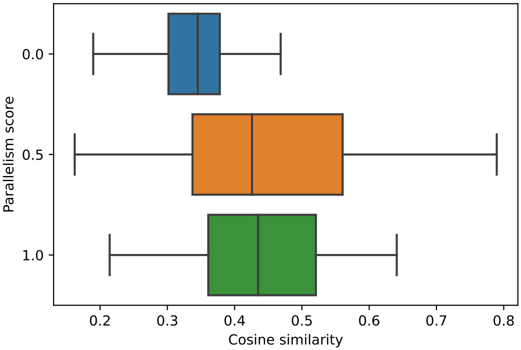
- Distribution of quality of prediction for adjectives, by parallelism score

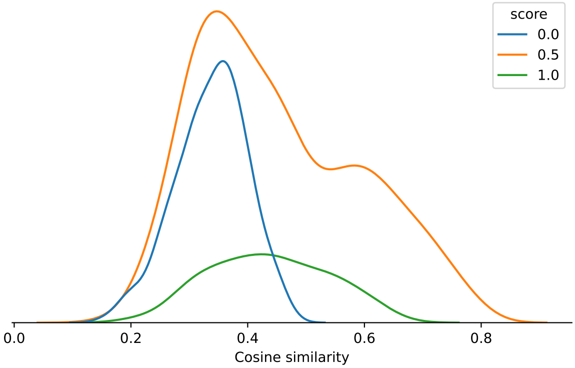




Vector prediction results IV

- Distribution of quality of prediction for nouns, by parallelism score





Conclusion

- High perfomance of cross-validated intrinsic prediction, with both methods.
  - Shows that distributional semantics captures contrasts between paradigm cells.
- While orthogonal contrasts lead to chance-level performance in extrinsic prediction, parallel contrasts lead to performance above chance level.
  - Shows that parallel contrasts in features capture some degree of parallelism in terms of actual content, as measured by distributional methods.
  - Hence the analysis of paradigms in terms of orthogonal features does capture interesting aspects of paradigm structure.
- Parallel contrasts in extrinsic prediction still lead to much poorer performance than intrinsic prediction.
  - Shows that the difference between two paradigm cells is not reducible to the featural description of those paradigm cells.
  - Hence, paradigm cells have properties that are not reducible to their description in terms of features.
  - Calls into question the reducibility of paradigmatic organisation in terms of orthogonal features, à la Wunderlich and Fabri (1995), and supports the view of paradigm organisation defended by Bonami and Strnadová (2019).

Future work

- The same methodology can be applied to more complicated paradigms such as to verbs.
- Future challenges:
  - Are number contrasts the same in the context of person (in the present) vs. gender (in the past)?
  - PAST tense of PERF verbs *vs.* PAST tense of IMPF verbs
  - FUT tense of PERF verbs *vs.* PRES tense of IMPF verbs
  - technical issue of auxiliaries in PAST and FUT tenses when training word vectors

Inflectional paradigm of the perfective verb *'udělat'* (*'to complete'*) and the imperfective verb *'dělat'* (*'to do'*).

<table>
<tr><th></th><th>PERS</th><th>PRES.SG</th><th>PRES.PL</th><th>PAST.SG</th><th>PAST.PL</th><th>FUT.SG</th><th>FUT.PL</th></tr>
<tr><td rowspan="3">PERF</td><td>1.</td><td>–</td><td>–</td><td>udělal-[∅|a|o] (jsem)</td><td>udělal-[i|y|a] (jsme)</td><td>udělá-m</td><td>udělá-me</td></tr>
<tr><td>2.</td><td>–</td><td>–</td><td>udělal-[∅|a|o] (jsi)</td><td>udělal-[i|y|a] (jste)</td><td>udělá-š</td><td>udělá-te</td></tr>
<tr><td>3.</td><td>–</td><td>–</td><td>udělal-[∅|a|o]</td><td>udělal-[i|y|a]</td><td>udělá-∅</td><td>uděla-jí</td></tr>
<tr><td rowspan="3">IMPF</td><td>1.</td><td>dělá-m</td><td>dělá-me</td><td>dělal-[∅|a|o] (jsem)</td><td>dělal-[i|y|a] (jsme)</td><td>(budu) dělat</td><td>(budeme) dělat</td></tr>
<tr><td>2.</td><td>dělá-š</td><td>dělá-te</td><td>dělal-[∅|a|o] (jsi)</td><td>dělal-[i|y|a] (jste)</td><td>(budeš) dělat</td><td>(budete) dělat</td></tr>
<tr><td>3.</td><td>dělá-∅</td><td>děla-jí</td><td>dělal-[∅|a|o]</td><td>dělal-[i|y|a]</td><td>(bude) dělat</td><td>(budou) dělat</td></tr>
</table>

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

Olivier Bonami and Jana Strnadová. Paradigm structure and predictability in derivational morphology. *Morphology*, 29(2):167–197, 2019.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

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.

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.

Marco Marelli and Marco Baroni. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3):485, 2015.

Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

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

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P14/P14-5003.pdf.

Gregory T. Stump and Raphael Finkel. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge, 2013.

Pavol Štekauer. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, pages 354–369. Oxford University Press, Oxford, 2014.

Dieter Wunderlich and Ray Fabri. Minimalist Morphology: an approach to inflection. *Zeitschrift für Sprachwissenschaft*, 14(2):236–294, 1995.