Assessing the featural organisation of paradigms with distributional methods

Olivier Bonami & Lukáš Kyjánek & Marine Wauquier

June 15–17, 2023

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

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

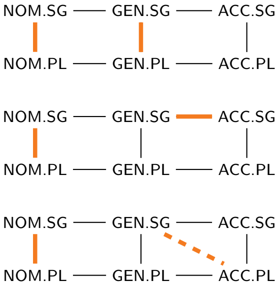
Université Sorbonne Nouvelle
Ecole Normale Supérieure
Centre National de la Recherche Scientifique

Paradigms as systems of orthogonal feature oppositions

- 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 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.
  - Some pairs of forms in paradigms are in direct pairwise contrast, while others are not.

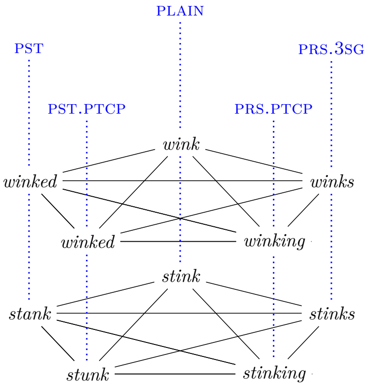


Paradigms as systems of multilateral contrasts

- A general definition should not require orthogonality.

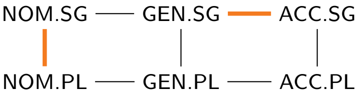
  > *"[...] we define the paradigm of a lexeme L as a complete set of cells for L, where each cell is the pairing of L with a complete and coherent morphosyntactic property set (MPS) for which L is inflectable."* (Stump and Finkel, 2013, p. 9)

- Bonami and Strnadová (2019) go further, building on Štekauer (2014) and Blevins (2016, chap. 5):
  - Paradigms are 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.



The topic for today

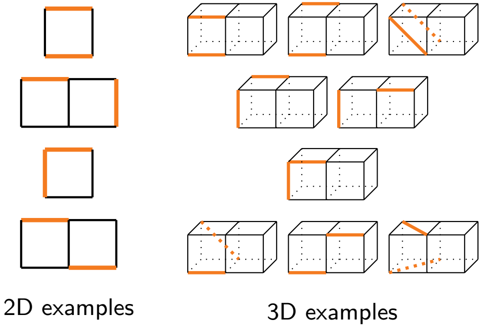
- Can we find empirical evidence that some contrasts are parallel while others are orthogonal?
  - Answering this will contribute to contrasting the two theoretical views of paradigms.



- Strategy:
  - Distributional vectors reflect syntactic and semantic properties of words likewise morpho-syntactic features are combinations of syntactic and morphological features.
  - We hence model contrasts between paradigm cells as between the corresponding distributional vectors.

Outline

- Pairs of cells (a, b) and (a', b') are
  a. parallel if they contrast in exactly the same way
  b. orthogonal if they do not contrast at all in the same way
  c. corner if they share a paradigm cell
  d. neither if they contrast in the same features but not the same values



- We exclude corner and neither types because we
  - either expect odd behaviours due to sharing a cell
  - or have no expectations regarding their behaviours

Existing data resources for Czech

- We trained the distributional vector model (cf. Kyjánek and Bonami, 2022) by applying
  - **Word2vec** (Mikolov et al., 2013)
  - to **SYN v9 corpus** (Křen et al., 2021)
  - trained for combinations of tokens and tags; we rely on the pos-tag annotations
- We use morphological data from **MorfFlexCZ 2.0** (Hajič et al., 2020).
  - inflectional morphological lexicon
  - 125.3M lemma-tag-wordform triples
  - used for the development of MorphoDiTa (tagging SYN v9 corpus)

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

Experiment 1: Classifying contrasting word vectors

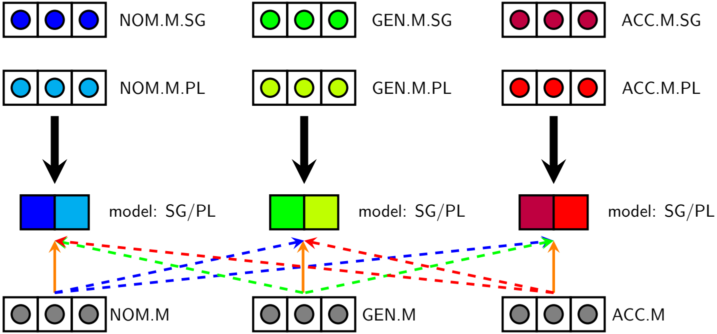
- **Task:** binary prediction of paradigm cell from word vector
- **Data:** 500 word vectors (only words with freq>50 in SYN v9) for each studied paradigm cell were sampled from SYN v9.
  - It resulted in 30 samples for nouns and 30 samples for adjectives; combinations of
    - grammatical cases [NOM, 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----`	*(crayon)*	*100-dim*
`práce>NNFS1-----A----`	*(work)*	*100-dim*
`paměť>NNFS1-----A----`	*(memory)*	*100-dim*
...	...	...

Experiment 1: Vector classification and its evaluation

- We train gradient boosting trees (Friedman, 2001, Mason et al., 2000) on 1000 unpaired words (500 from both contrasting features).
- Evaluation:
  - Intrinsic (orange arrows) – by means of 10-fold cross-validation on the 1000-word datasets
  - Extrinsic (blue, green, red arrows) – by means of a confusion matrix based on aligned labels, e.g., the model $\mathrm{SG/PL}$ trained on $\mathrm{NOM.M}$ is evaluated on both $\mathrm{GEN.M}$, $\mathrm{ACC.M}$



Experiment 1: Results of the classifications

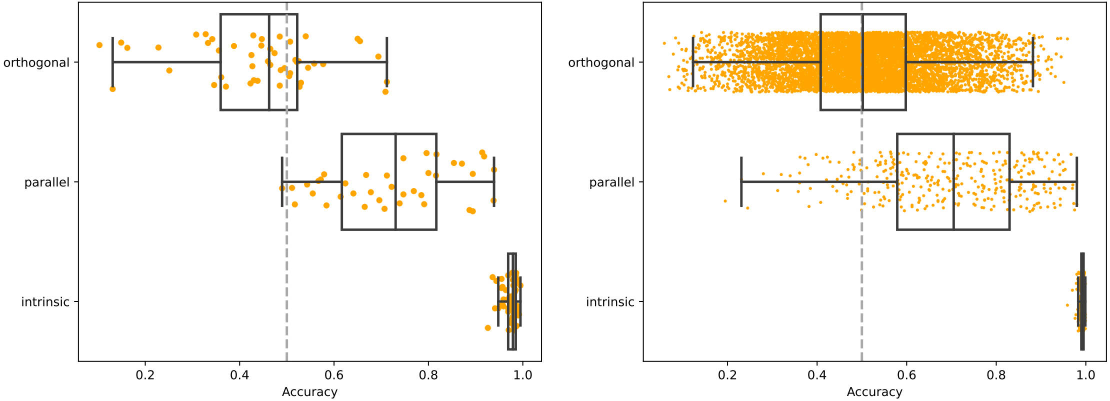


Figure 1: Distribution of accuracy of classifications (Experiment 1) for nouns (left) and adjectives (right). The dashed grey line represents baseline performance at 0.5.

Experiment 2: Predicting relations between word vectors

- **Task**: prediction of a target word vector on the basis of a source word vector
- **Data**: 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).

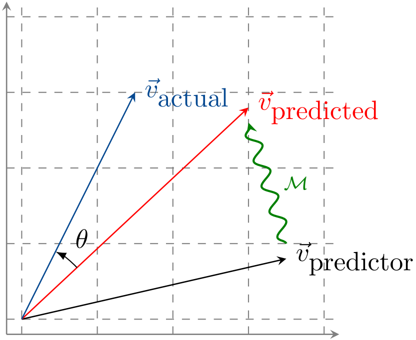
Word A		Word B		Vector A	Vector B
výpůjčka>NNFS1-----A----	*(loan)*	výpůjčky>NNFP1-----A----	*(loans)*	*100-dim*	*100-dim*
hmotnost>NNFS1-----A----	*(weight)*	hmotnosti>NNFP1-----A----	*(weights)*	*100-dim*	*100-dim*
nádrž>NNFS1-----A----	*(tank)*	nádrže>NNFP1-----A----	*(tanks)*	*100-dim*	*100-dim*
líheň>NNFS1-----A----	*(hatchery)*	líhně>NNFP1-----A----	*(hatcheries)*	*100-dim*	*100-dim*
...	...	...	...	...	...

Experiment 2: Vector prediction and its evaluation

- Following Marelli and Baroni (2015), we train one (gradient boosting tree) model per dimension in the target vector and combine the models into the collection $\mathcal{M}$:

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

- The quality of $\mathcal{M}$ in terms of capturing the morpho-syntactic contrast is calculated by 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 contrast of targets is from that of predictors for that particular morpho-syntactic feature.



Experiment 2: Results of the predictions

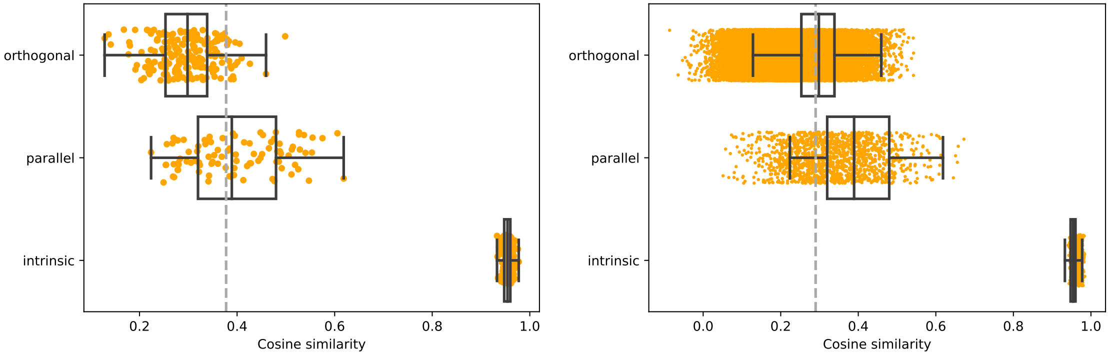


Figure 2: Distribution of quality (cosine similarity) of vector predictions (Experiment 2) for nouns (left) and adjectives (right). Grey lines indicate the average cosine similarity between members of the same lemma.

Experiment 2: Baseline

- 20 random word vectors are picked from the vector space model and used as predicted word vectors.
- The average of cosine similarities between the actual vector v_{actual} and individual randomly picked word vectors $(v_{\text{predicted}_1}, \ldots, v_{\text{predicted}_{20}})$ are calculated.
- The resulting cosine similarity for a given contrast is computed as the average of the averages achieved by individual pairs of word vectors.

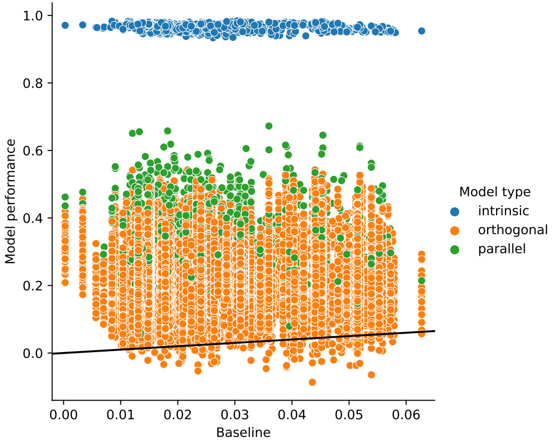


Figure 3: Our models *vs.* the baseline. The black line stands for equal values on the x and y axis.

Discussion & Conclusion

- **Intrinsic** predictions achieve high performance (even under cross-validation)
  - word vectors capture the relevant syntactic and semantic differences between paradigm cells
- **Orthogonal** types of contrasts lead to poor performance (unsurprisingly)
  - from Experiment 1: most models have a performance close to the baseline – were lucky or unlucky, in a symmetric fashion
  - from Experiment 2: performance is still on average much better than the random baseline (orthogonal models, unlike the baseline, have the capacity to accurately predict some aspects of distributions that are due to being forms of the same lexeme)
- **Parallel** types of contrasts achieve measurably higher performance than orthogonal predictions but are markedly lower than intrinsic predictions.
  - it is in direct contradiction to the predictions that contrast between paradigm cells are fully reducible to contrasts in feature values
  - it 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

1. What exactly makes contrasts across parallel pairs of paradigm cells different?
2. A different use of the same methodology would explore situations where the literature is disputed.
   - Are number contrasts the same in the context of person (in present) *vs.* gender (in 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
3. How do different morpho-syntactic features differ in their degree of parallelism across contrasts?

Acknowledgement

Thank you for your attention.

We thank **Sacha Beniamine**, **Mae Caroll**, **Timothee Mickus**, and **Erich Round** for comments on early versions of our work.

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);
- the public grant overseen by the French National Research Agency (ANR) as part of the 'Investissements d'Avenir' program (reference: ANR-10-LABX-0083).

It has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (`https://lindat.cz`), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References I

James P. Blevins. *Word and Paradigm Morphology*. Oxford University Press, Oxford, 2016.

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.

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

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.

Appendix: Inflectional paradigm of Czech verbs

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

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