Assessing the featural organisation of paradigms with distributional methods

Olivier Bonami
Université Paris Cité,
Laboratoire de linguistique formelle,
CNRS
olivier.bonami@u-paris.fr

Lukáš Kyjánek
Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
kyjanek@ufal.mff.cuni.cz

Marine Wauquier
Université Sorbonne Nouvelle,
Laboratoire Lattice,
CNRS
marine.wauquier@sorbonne-nouvelle.fr

Abstract

In this paper, we apply distributional methods to Czech data to compare the predictions of two views of inflectional paradigms, as systems of orthogonal morphosyntactic feature oppositions, or as systems of multilateral contrasts between pairs of morphologically related words, not necessarily reducible to orthogonal features.

We define two predictive tasks that probe what it means for two pairs of paradigm cells to contrast in the same features: in the first, we train a classifier to discriminate between two paradigm cells; in the second, we train a family of models to predict the vector of the word in one cell from that of the word in another cell. By varying the choice of training and test data, we show that (i) a model trained on data that contrast in a manner orthogonal to its test data performs on average at chance level, while (ii) a model trained on data that contrast in a manner parallel to its test data performs on average better than chance but still worse than a model trained on the same pair of cell used for testing. This is incompatible with the predictions of a reductive view of paradigms as systems of feature contrasts.

1 Introduction

The notion of an inflectional paradigm is an invaluable tool for linguistic description and has played an increasing role in linguistic theory in the last few decades. Explicit reference to paradigm structure has been claimed to be necessary to account for phenomena as diverse as patterns of syncretism (Zwicky, 1985; Stump, 1993; Baerman et al., 2005), competition between synthetic and periphrastic expression of morphosyntactic categories (Ackerman and Stump, 2004; Kiparsky, 2005; Bonami, 2015), and universal constraints on the shape of inflection systems (Carstairs-McCarthy, 1994; Ackerman and Malouf, 2013). While many of these claims have been met with scepticism by some (see e.g. papers collected in Bachrach and Nevins 2008), there is general agreement that some form of paradigmatic organisation plays a role in morphology, if only through the existence of collections of pairs of expressions that differ by contrasting in the same morphosyntactic features. Hence although morphologists may differ in how they think of paradigms, they will agree that there is something in common between the way *man* relates to *men* and *dog* relates to *dogs*. That something in common is what we will call a paradigmatic relation.

That being said, there is variation in the literature regarding the way paradigms are defined, and differences between these formulations are seldom discussed. A common position, ultimately grounded in Jakobson (1958) and cogently articulated by Wunderlich and Fabri (1995, p. 266), holds that “A paradigm is an n-dimensional space whose dimensions are the attributes (or features) used for the classification of word forms”. In other words, paradigms can be reduced to a system of orthogonal contrasts in morphosyntactic feature values.[1] This claim is appealing when we look at some very-well behaved inflection systems. Consider the paradigm of an Italian adjective in Table 1. Every cell in that paradigm can be defined as the combination of a number and a gender value. If this holds in general, it suggests that paradigm structure is entirely

[1] Note that we follow Matthews (1991) in calling ‘morphosyntactic’ whatever features are relevant to the organisation of inflectional paradigms. Some of these will be semantically relevant, others not. Our usage departs from that of Corbett (2012), who would call some of the features we discuss here ‘morphosemantic’.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 310-320.
Amherst, Massachusetts, June 15-17, 2023

	MAS	FEM
SG	buono	buona
PL	buoni	buone

Table 1: Paradigm of Italian BUONO ‘good’.

			IND		IMP
			PRS	PST	
FINITE	SG	1	eat	ate	—
		2	eat	ate	eat
		3	eats	ate	—
	PL	1	eat	ate	—
		2	eat	ate	eat
		3	eat	ate	—
NFIN	PART		eating	eaten	
	INF		eat		

Table 2: Paradigm of English EAT as a system of orthogonal oppositions. Periphrastic forms ignored.

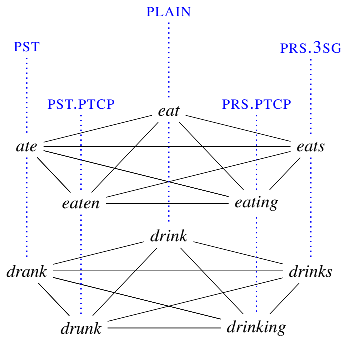

Figure 1: English verbal paradigms seen as a system of basic contrasts in content.

derivative of a system of feature oppositions.

This view of paradigms becomes less appealing as soon as we move away from well-behaved declension systems. In conjugation systems, it often is the case that orthogonal feature oppositions are unhelpful. English conjugation provides an extreme example of that situation. Table 2 is our best attempt at presenting the paradigm of an English verb as a system of orthogonal oppositions. Multiple problems arise: some feature oppositions are neutralised (no tense distinction in the imperative or infinitive), and some paradigm cells are non-existent (no 1st or 3rd person imperatives). Most importantly, there is a disconnect between the shape of the paradigm as motivated by feature oppositions and the inventory of forms filling that paradigm: with the exception of BE, no lexeme uses more than 5 distinct forms to fill 17 cells, and arbitrary collections of cells exhibit systematic syncretism — e.g. all non-3rd present form, imperative forms, and the bare infinitive.

The observation of such discrepancies naturally leads one to revise their expectations as to the paradigmatic organisation. Spencer (2013), Boyé and Schalchli (2016), and Stump (2016) make slightly different proposals for distinguishing different notions of paradigms. Bonami and Strnadová (2019), building among others on Štekauer (2015) and Blevins (2016, chap. 5), take another route illustrated for English verbs in Figure 1. Under this view, contrasts in content between sets of pairs of words, materialised in the figure by vertical alignments across morphological families, are the primitive notion from which paradigms are defined. Analysis of such paradigms in terms of orthogonal features is a further step that may be more or less useful and insightful depending on the system under examination. Crucially, paradigms (horizontal planes in Figure 1) and paradigm cells (vertically aligned collections of words) exist independently of such a featural analysis.

In this paper, we explore empirically the predictions of the two basic conceptualisations of paradigm structure outlined above. Focusing on cases where a feature-based definition of paradigms seems warranted as in Table 1, we ask to what extent the featural composition of the paradigm can be trusted. For example, is the contrast between masculine singular and plural really the same as the contrast between feminine singular and feminine plural? To answer that question, we explore contrasts between pairs of words (nouns or adjectives) in Czech using distributional vectors familiar from distributional semantics. Note that distributional vectors typically capture both syntactic and semantic contrasts between words. While this is sometimes an embarrassment when disentangling the two is important, it is fine for our purposes, as paradigmatic contrasts may be semantically potent or not.

Section 2 provides a precise definition of what it means for two pairs of cells to encode contrasts that are parallel, orthogonal or neither. We then use this definition to lay out predictions on the expected structure of the distributional vector space under the assumption that paradigms are defined by features. In Section 3 we present two experiments testing these predictions: in the first experiment, we train classifiers to discriminate between vectors of words from two paradigm cells, while in the

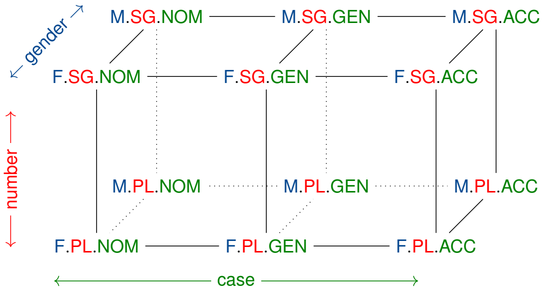

Figure 2: Illustrative organisation of a paradigm as a system of orthogonal featural contrasts. In this example, we have three features, namely case, number and gender, represented as three geometric dimensions. Paradigm cells are represented as points in 3D space combining a particular value for each feature.

second experiment, we train a model to predict the vector of a word in one paradigm cell from that of the word in another paradigm cell. In both cases, we compare the quality of prediction of models trained on data from the same pair of cells, from a parallel pair of cells, or from an orthogonal pair of cells. Section 4 discusses the implications of our findings for morphological theory, and Section 5 outlines avenues for future work.

This paper presents a terminological difficulty, as the term 'feature' has different meanings in the context of descriptive and theoretical morphology and in the context of computational linguistics and machine learning. To alleviate that difficulty, we refrained from using the term at all when discussing machine learning, talking of *predictors* or *variables* instead; and we prefixed *feature* with *morphosyntactic* wherever there was potential for ambiguity.

2 Predictions

In this section, we define ways of comparing how inflected forms of the same lexeme differ in meaning and use this to derive predictions of the claim that paradigms reduce to featural contrasts.

For the sake of exploring the featural organisation of paradigms, we assume that each cell in a paradigm can meaningfully be mapped to a morphosyntactic description which we formalise as a functional relation between a set of features $\mathcal{F}$ and a set of values $\mathcal{V}$, where no two features can map to the same value.[2] Given two paradigm

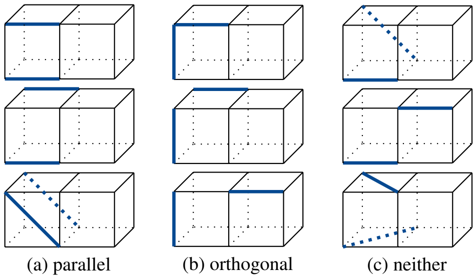

Figure 3: Types of relations between pairs of cells.

cells a and b, we note $S(a,b) \overset{def}{=} \{v \mid f : v \in a \wedge \neg f : v \in b\}$ the set of feature values specific to a when compared to b. We then say that two pairs of contrasting cells (a,b) and (a',b') are **parallel** if $S(a,b) = S(a',b')$ and $S(b,a) = S(b',a')$. We likewise note $C(a,b) \overset{def}{=} \{f | \exists v \exists w[f : v \in a \wedge f : w \in b \wedge v \neq w]\}$ the set of features along which a and b contrast, and then call two pairs of cells **orthogonal** if they do not share any contrast, i.e. $C(a,b) \cap C(a',b') = \emptyset$.

For purposes of illustration, we will use the example laid out in Figure 2 of a system with two binary features (number and gender) and one ternary feature (case), and represent visually each feature as a geometric dimension. The definitions of parallelism and orthogonality are illustrated in Figure 3. Note that we can find parallel pairs of contrasts where the contrasting cells have no feature in common (bottom left). Note also that our notion of parallelism does not extend to situations where the two contrasts involve the same features but different values (middle right): in that situation, contrasts are neither parallel nor orthogonal.

Given these definitions, we can now derive our predictions. Let us assume that we have satisfactory representations of the content of inflected words in a language (combining semantic and syntactic information). Let us also assume that paradigmatic relations are fully reducible to some correct description in terms of feature contrasts. Then, given two pairs of words (v,w) and (v',w') filling cells (a,b) and (a',b') of some paradigm:

- If (a,b) and (a',b') are parallel, then the content of v and w should differ in exactly the same way as the content of v' and w' differ. Hence if we define a predictive task which relies on capturing the relationship between

[2] We follow Stump and Finkel (2013) in assuming that the list of paradigm cells can be a proper subset of the set of all such functional relations, leaving room for the description of systems such as that exemplified with English conjugation above. The requirement that no two feature map to the same values is purely motivated by mathematical elegance, and could easily be dropped.

cells a and b, it should be immaterial whether we train our system on data from cells a and b (what we call *intrinsic prediction*) or cells a' and b' (what we call *extrinsic prediction*).

- If (a, b) and (a', b') are orthogonal, then the contrast between the content of v and w is unrelated to the contrast between the content content of v' and w'. Hence if we define a predictive task which relies on capturing the relationship between cells a and b and train our system on data from cells a' and b', we should witness dramatically poor performance, at the chance level.

In Section 3, we test these predictions on data from Czech nouns and adjectives. Czech nouns inflect for 2 numbers (singular, plural) and 7 cases (nominative, genitive, dative, accusative, vocative, locative, instrumental), leading to a 2-dimensional system with 14 cells, while adjectives also inflect for 4 genders (masculine animate, masculine inanimate, feminine, neuter) and 3 grades (positive, comparative, superlative), leading to a 4-dimensional system with 168 cells. In the interest of tractability, we restrict attention to the positive grade of adjectives and the three main structural cases (nominative, genitive, accusative). This leads for nouns to 6 cells in 2 dimensions, and for adjectives to 24 cells in 3 dimensions — see Tables 3 and 4 for examples. We also leave out from consideration orthogonal contrasts forming a corner, as in the top example of column (b) in Figure 3, as sharing of a cell between the two pairs is likely to affect performance.

3 Experiments

3.1 Data

We use distributional representations of Czech word vectors from the vector spaces provided by Kyjánek and Bonami (2022). These models were trained by applying word2vec (Mikolov et al., 2013) to the SYN v9 corpus (Křen et al., 2021), a large corpus of contemporary edited text compiled, lemmatised and tagged by the Czech National Corpus team (4,719M tokens; 7.3M lemmas; 362M sentences). Vectors were trained on the concatenation of tokens and POS tags, and hence in effect represent a form filling a particular paradigm cell. For instance FEM.NOM.SG and NEU.NOM.PL *malá* from Table 3 get separate representations. This is crucial for our purposes:

POSITIVE GRADE

		MA	MI	FEM	NEU
SG	NOM	**malý**	**malý**	**malá**	**malé**
	GEN	**malého**	**malého**	**malé**	**malého**
	DAT	malému	malému	malé	malému
	ACC	**malého**	**malý**	**malou**	**malé**
	VOC	malý	malý	malá	malé
	LOC	malém	malém	malé	malém
	INS	malým	malým	malou	malým
PL	NOM	**malí**	**malé**	**malé**	**malá**
	GEN	**malých**	**malých**	**malých**	**malých**
	DAT	malým	malým	malým	malým
	ACC	**malé**	**malé**	**malé**	**malá**
	VOC	malí	malé	malé	malá
	LOC	malých	malých	malých	malých
	INS	malými	malými	malými	malými

COMPARATIVE GRADE

		MA	MI	FEM	NEU
SG	NOM	menší	menší	menší	menší
	GEN	menšího	menšího	menší	menšího
	DAT	menšímu	menšímu	menší	menšímu
	ACC	menšího	menší	menší	menší
	VOC	menší	menší	menší	menší
	LOC	menším	menším	menší	menším
	INS	menším	menším	menší	menším
PL	NOM	menší	menší	menší	menší
	GEN	menších	menších	menších	menších
	DAT	menším	menším	menším	menším
	ACC	menší	menší	menší	menší
	VOC	menší	menší	menší	menší
	LOC	menších	menších	menších	menších
	INS	menšími	menšími	menšími	menšími

SUPERLATIVE GRADE

		MA	MI	FEM	NEU
SG	NOM	nejmenší	nejmenší	nejmenší	nejmenší
	GEN	nejmenšího	nejmenšího	nejmenší	nejmenšího
	DAT	nejmenšímu	nejmenšímu	nejmenší	nejmenšímu
	ACC	nejmenšího	nejmenší	nejmenší	nejmenší
	VOC	nejmenší	nejmenší	nejmenší	nejmenší
	LOC	nejmenším	nejmenším	nejmenší	nejmenším
	INS	nejmenším	nejmenším	nejmenší	nejmenším
PL	NOM	nejmenší	nejmenší	nejmenší	nejmenší
	GEN	nejmenších	nejmenších	nejmenších	nejmenších
	DAT	nejmenším	nejmenším	nejmenším	nejmenším
	ACC	nejmenší	nejmenší	nejmenší	nejmenší
	VOC	nejmenší	nejmenší	nejmenší	nejmenší
	LOC	nejmenších	nejmenších	nejmenších	nejmenších
	INS	nejmenšími	nejmenšími	nejmenšími	nejmenšími

Table 3: Paradigm of Czech MALÝ ‘small’. Cells used in the experiments are highlighted in boldface.

	SG	PL
NOM	**holka**	**holky**
GEN	**holky**	**holek**
DAT	holce	holkám
ACC	**holku**	**holky**
VOC	holko	holky
LOC	holce	holkách
INS	holkou	holkami

	SG	PL
NOM	**cíl**	**cíle**
GEN	**cíle**	**cílů**
DAT	cíli	cílům
ACC	**cíl**	**cíle**
VOC	cíli	cíle
LOC	cíli	cílech
INS	cílem	cíli

Table 4: Paradigms of two Czech nouns: feminine HOLKA ‘girl’ and masculine inanimate CÍL ‘goal’. Cells used in the experiments are highlighted in boldface.

since syncretism is rampant in Czech inflection, distributional representations of raw strings would be useless to make comparisons across paradigm cells. We used the tagging distributed with the corpus, which was obtained automatically using the MorphoDiTa tool (with a reported accuracy over 95%, Straková et al., 2014). In our experiments, we use a 100-dimensional vector space trained as a continuous bag of words (CBOW) model.[3] We also used the inflectional morphological dictionary MorfFlexCZ 2.0 (Hajič et al., 2020), which contains 125.3M triplets of word form and its respective lemma and tag, to sample vectors of tokens with relevant morphosyntactic categories. Note that MorfFlexCZ and the SYN corpus share the same tagset.

For the first experiment, we sampled 500 random word vectors for each paradigm cell under investigation, allowing us to have combined datasets for classification of size 1000. We included only word vectors for words that occurred at least 50 times in the SYN v9 corpus. This led to 24 datasets for adjectives corresponding to the 24 paradigm cells highlighted in Table 3. For nouns we created separate datasets for each of the genders, leading again to 24 ($= 4$ genders $\times$ 6 paradigm cells) datasets.

For the second experiment, we needed datasets consisting of ordered pairs of vectors for forms of the same lexeme for two particular cells in the paradigm. We used MorfFlexCZ to identify relevant pairs and randomly sampled datasets of 1,000 pairs; again, we included only vectors with a frequency of 50 or more. For adjectives, with 24 paradigm cells under examination, we ended up with $24 \times 23 = 552$ datasets. For nouns, we again created separate datasets for each gender. With 6 paradigm cells under examination, this led to $4 \times 6 \times 5 = 120$ datasets.

3.2 Experiment 1

In our first experiment, we want to assess how hard it is to discriminate two paradigm cells when trained on data from the same or other cells. To this end, we train classifiers to discriminate between two paradigm cells and apply it to data from the same pairs of cells, parallel pairs of cells, and orthogonal pairs of cells.

More specifically, we design two-step experiments. First, we conduct *intrinsic classification*, meaning that we train a classifier to discriminate a given contrast realised by a pair of paradigm cells, and we apply it to words inducing the same contrast. An example of this would be training to discriminate FEM.SG.ACC and FEM.PL.ACC forms of adjectives, and testing the classifier on the forms of other lexemes in the same two cells. Second, we investigate the interoperability of the morphosyntactic feature by means of an *extrinsic classification* task. An example of this would be training to discriminate FEM.SG.ACC and FEM.PL.ACC forms of adjectives, and testing the performance of the classifier on its ability to discriminate words in two other cells, e.g. FEM.SG.GEN and FEM.PL.GEN. We hypothesise a classifier trained to discriminate the contrast between two cells should also be able to discriminate between two other cells provided the two pairs of cells are parallel.

Concretely, for each relevant predictor pair of cells (a, b), we train a classifier to discriminate vectors of words in cell a from vectors of words in cell b. We used gradient boosting (Friedman, 2001a; Mason et al., 2000) applied to decision trees as our classification method. Predictors are the 100 dimensions of the vectors, and boosting trees parameters are set to 500 estimators, a learning rate of 0.01, a max depth of 2, a random state of 0, and the deviance loss function. In total, we trained 60 classifiers for nouns, to be used in 60 and 86 intrinsic and extrinsic classification tasks respectively; and 276 classifiers for adjectives, used in 276 intrinsic and 7824 extrinsic classifications tasks. The much higher number of tasks for adjectives is due to their larger paradigm size due to gender agreement, cf. Tables 3 and 4.

For intrinsic classification tasks, we performed 10-fold cross-validation, and report aggregated accuracy across the 10 folds. For extrinsic classification, there was no avoidable risk of over-fitting, as the training and test datasets are inherently disjoint.[4] Note that, since our samples are balanced, chance performance is at 0.5. We use this as our baseline for evaluation. Figure 4 summarises our results.

Classifiers for both nouns and adjectives achieve very high performance at intrinsic classification,

[3] We also experimented with models trained by the skip-gram method or having 400-dimensional vectors, but this led to no qualitative difference in the results.

[4] As a reviewer notes, the test data is included in the training corpus for the vector space, and hence can in principle have some influence on the results. There is no way of avoiding that potential problem with the methods used here, as we do need vectors from the same space for test items for evaluation purposes.

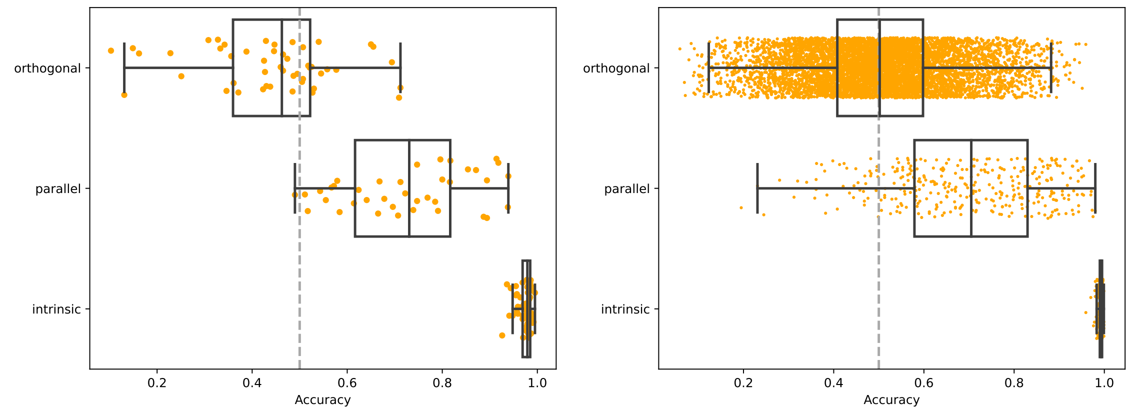

Figure 4: Distribution of accuracy of classifications (Experiment 1) for nouns (left) and adjectives (right). The dashed grey line represents baseline performance at 0.5.

with a median accuracy of 0.98 and 0.99 respectively and a standard deviation of 0.02 and 0.005 respectively. Performances are significantly lower for extrinsic classification, although the use of classifiers for parallel contrasts still leads to above-chance level performance for a vast majority of models, with a median accuracy of 0.72 for nouns and 0.71 for adjectives. On the other hand, extrinsic classification for orthogonal contrasts barely achieves chance-level performances. Median accuracy is at 0.46 for nouns and 0.51 for adjectives. There is a lot of variation around this median, which is not surprising given the high number of models we trained, but the distribution of accuracy across orthogonal classifiers is clearly symmetric and centred on 0.5, suggesting that any structure that individual classifiers pick out is due to lucky sampling.

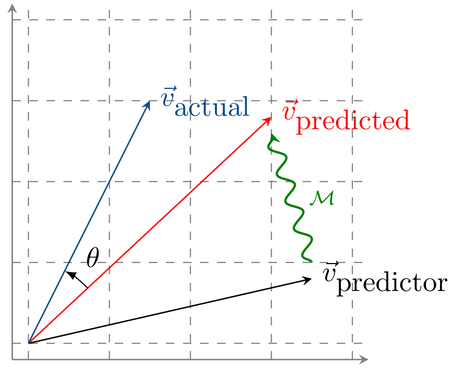
$\vec{v}_{\text{actual}}$
$\vec{v}_{\text{predicted}}$
$\mathcal{M}$
θ
$\vec{v}_{\text{predictor}}$

Figure 5: Evaluation of vector prediction. Performance of model $\mathcal{M}$ is assessed by the cosine of the angle θ between the actual vector for the target word and the vector predicted by $\mathcal{M}$ for the that based on the predictor vector of a related word.

3.3 Experiment 2

In our second experiment, we predict the vector of a word in the target paradigm cell ($\vec{v}_{\text{predicted}}$) from that of the word in another paradigm cell ($\vec{v}_{\text{predictor}}$), and evaluate the quality of our prediction by comparing it to the actual vector $\vec{v}_{\text{actual}}$. This is represented graphically in Figure 5, where $\mathcal{M}$ denotes the model deriving the prediction.

Multiple ways of constructing the model $\mathcal{M}$ are found in the literature. A simple approach relies on adding to the predictor vector the offset vector relating two words standing in the same relation (Mikolov et al., 2013) or averaging over such offset vectors (Drozd et al., 2016; Mickus et al., 2019). Marelli and Baroni (2015) propose instead to use a linear transformation to predict the target vector — that is, they predict the value of each dimension of the target vector using a linear combination of the values of all dimensions in the predictor vector. They argue that this should allow capturing at least some aspects of affix polysemy. Bonami and Naranjo (2023) use a variant of this approach using principal component analysis to reduce the number of independent variables in the linear models.

In this paper we follow closely the methodology of Marelli and Baroni (2015), using Gradient Boosting Tree regression models (Friedman, 2001b) instead of linear models.[5] For each morphological contrast, we train 100 models per pairwise combination of paradigm cells as there are 100 vector dimensions in the input vector space models. In to-

[5] We also tested linear regression models, but the gradient boosting tree method achieved better evaluation results.

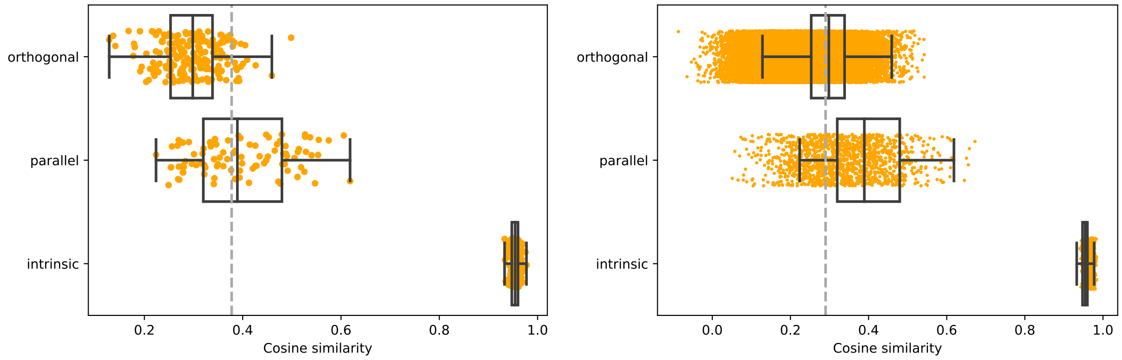

Figure 6: Distribution of quality (cosine similarity) of vector predictions (Experiment 2) for nouns (left) and adjectives (right). Grey lines indicate the average cosine similarity between members of the same lemma.

tal, we trained $100 \times (120+552) = 67,200$ models ($\times 10$ because of cross-validation) to predict all vector dimensions of words from the paradigm cells under analysis. We then evaluate the performance of our models in both intrinsic and extrinsic predictions, using the average cosine similarity between predicted and actual vector ($\cos(\vec{v}_{\text{predicted}}, \vec{v}_{\text{actual}})$) as our measure of quality. While the evaluation of the intrinsic predictions assesses discriminating power for predicting word vectors, i.e., the prediction of the same contrasts as the one on which the model was trained, the evaluation of the extrinsic predictions assesses the stability of predicting word vectors in different contexts, i.e., the prediction of contrasts different from the one used for training the model.

Results are presented in Figure 6. We get very high scores for intrinsic prediction, ranging between 0.92 and 0.98. Cross-validated models have barely lower performance (median difference 0.012, max. 0.02), indicating that there is no over-fitting to speak of. The extrinsic predictions achieved vastly lower cosine similarities than their intrinsic counterparts, with a gap of more than 25% between the best-performing extrinsic prediction and the worst-performing intrinsic prediction. As in Experiment 1, results for both orthogonal and parallel prediction are quite spread out, but there is a clear central tendency to have higher performance for parallel prediction than for orthogonal prediction.

We contextualise the results of trained models in two ways. Our first approach is to compute the average pairwise cosine similarity between vectors of words belonging to the same lemma, for the paradigm cells of interest, and for each part of speech. This gives us an indication of what would be the performance of a model that perfectly captured the fact that the target vector conveys the right lexical semantics, but does not capture anything about the contribution of morphosyntactic features. These are materialised by grey lines in Figure 6. It is most relevant to compare that number to the performance of intrinsic models: here we see very clearly that these models do capture much more than just the lexical semantics associated with belonging to the same lemma.

For orthogonal and parallel prediction, this comparison is hard to interpret, given the high variability of the quality of prediction across tasks of the same type. We suspect that this variability is due at least in part to the fact that some test sets are inherently easier or harder to predict due to the structure of the vector space. We hence develop a baseline that is directly sensitive to the test set, and we compare the results of our cross-validated models to those from the baseline. The simplest baseline would be to create a predicted vector from random numbers; however, sampling random numbers might lead to vectors that are out of the vector space model. Therefore, we instead pick random word vectors from the vector space model and use them as predicted word vectors. To mitigate knowledge that such randomly picked word vectors might encode, we pick randomly 20 word vectors for each pair of word vectors and calculate the average of cosine similarities between the actual vector $\vec{v}_{\text{actual}}$ and individual randomly picked word vectors ($\vec{v}_{\text{predicted}_1}, \ldots, \vec{v}_{\text{predicted}_{20}}$). The resulting cosine similarity for a given contrast is computed as the average of the averages achieved by individual pairs of word vectors.

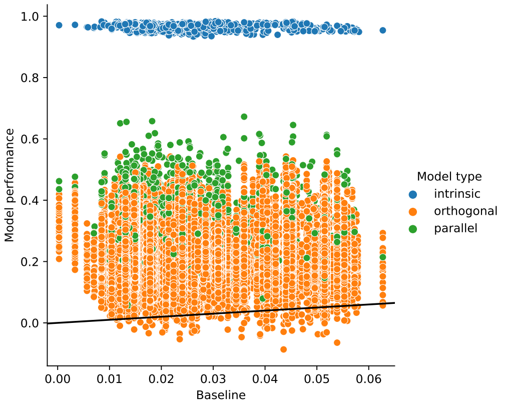

Figure 7: Comparison of our models to the baseline. The black line stands for equal values on the x and y axis.

Figure 7 shows pairwise comparisons between baseline and model performance. The clear conclusion is that both intrinsic and parallel prediction clearly outperform the baseline. A few orthogonal models perform at the baseline level, but most still clearly beat the baseline. To put these results in perspective, it is important to remember that, while orthogonal models are trained on irrelevant morphosyntactic contrasts, they still see pairs of forms of the same lexeme. To the extent that the vectors disentangle lexical semantics from morphosyntactic features, they should still be able to predict lexical semantics correctly — by not changing the values of the relevant dimensions. It is hence expected that performance should be above baseline on average; the fact that it is not always suggested that lexical semantics and morphosyntactic features are not clearly separated by the vectors.

4 Discussion

Our two experiments lead to similar results that we discuss in the following paragraphs.

First, intrinsic prediction works very well: classifiers learning to discriminate two paradigm cells on the basis of the corresponding word vectors reach very high accuracy, even under cross-validation; and a model learning to deduce the vector in one cell from the vector in another cell makes predictions that are very close to the actual vectors, and go well beyond capturing the fact that words belonging to the same lemma tend to be similar. Together, these indicate that the word vectors we use do capture the relevant syntactic and semantic differences between paradigm cells with a high degree of accuracy.

Second, orthogonal prediction leads to poor performance: training a model on a contrast orthogonal to that found in the test data is, unsurprisingly, a bad idea. This is most clearly established for the classification task of Experiment 1, where we see that most models have a performance close to the baseline, while a few models got lucky or unlucky, in a symmetric fashion. In the vector prediction task of Experiment 2, performance is still on average much better than the random baseline, due to the fact that 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.

The third and most important result is that found in the situation of parallel prediction, where a model is trained on one pair of cells implementing a feature contrast and tested on a different pair of cells implementing the same feature contrast. Here we find that, in both experiments, performance is measurably higher (on average) than with orthogonal models, but markedly lower than in intrinsic prediction. This last result is in direct contradiction to the predictions laid out in Section 2. If contrasts between paradigm cells were fully reducible to contrasts in feature values, then parallel pairs of cells should contrast in exactly the same way, and hence parallel prediction and intrinsic prediction should lead to comparable performance.

These results lead to a nuanced view of the role of morphosyntactic features in the analysis of inflectional paradigms. First, paradigm structure is not fully reducible to a system of orthogonal feature contrasts, *pace* Wunderlich and Fabri (1995) and many others. Paradigm cells have irreducible distributional properties that cannot be deduced from their featural analysis. Note that this is compatible with the view articulated by Bonami and Strnadová (2019), where each paradigm cell is characterised by the full set of its contrasts with all other cells. Second, morphosyntactic features do capture relevant similarities between pairs of cells: if they did not, parallel predictions should fare no better than orthogonal predictions.

Of course, one may dispute the extent to which these results are relevant to the featural analysis of paradigms. Our results are compatible with a situation where distributional vectors are influenced by morphosyntactic features, which are nicely organised in orthogonal dimensions, plus some other factors, which are not. We see no empirical way of dismissing such an analysis. However, we submit that it does not affect our conclusion: whatever the relevant factors are, it remains that paradigm cells have properties that are not reducible to orthogonal features.

Let us finish by noting that the nuanced conclusion (features capture some but not all paradigm structure) is most congruent with what Blevins (2006) calls an abstractive model of morphology. Under this view, surface words and the surface relations they entertain are the basic primitive, and objects such as stems and affixes are abstractions that may (but need not) be defined out of words and their relations. Arguably, morphosyntactic features can also be seen as such useful abstractions, that do not *define* paradigmatic relations but highlight some of their properties.

5 Outlook

We end by discussing areas of future research based on the results presented in this paper.

First, this paper did not explore what it is exactly that makes contrasts across parallel pairs of paradigm cells different; for instance, we did not look into whether some feature contrasts are easier or harder to predict, or more or less parallel across pairs of cells. We leave such questions for future research. We also leave for the future detailed analysis of particular parallel contrasts: we could e.g. examine distributional similarities and differences for a set of nouns in the NOM.SG, ACC.SG, NOM.PL and ACC.PL, and see whether these explain the performance models on this particular set of contrast.

Second, we focused in this paper on cases where the assumption of orthogonality of features was maximally convincing. A different use of the same methodology would be to explore situations where the literature is disputed as to what the feature contrasts actually are and attempt to settle the dispute by assessing how fruitful a feature analysis is in terms of capturing distributional parallelism or orthogonality. Obvious targets include Jakobson (1958)'s three-dimensional analysis of the Russian case systems, as well as many later proposals inspired by it; or the vexed question of the independence of person and number (see e.g. Siewierska 2004).

Third and finally, we have not explored whether and how different morphosyntactic features differ in their degree of parallelism across contrasts. We have reasons to believe that they could. Much recent literature has highlighted the multidimensional and gradient nature of the distinction between inflection and derivation (Booij, 1996; Bauer, 2004; Corbett, 2010; Spencer, 2013); in particular, semantically potent inherent morphosyntactic features, such as the number of nouns, are more derivation-like that purely morphosyntactic and contextual features, such as grammatical case. Previous research has shown that inflectional and derivational morphological relations as a whole difference in the predictability of their distributional consequences (Bonami and Paperno, 2018), and found some distributional reflexes for the existence of a gradient (Copot et al., 2022). Degree of parallelism might be another relevant distributional property: we may expect, for instance, there to be less parallelism of the number feature across cases than of cases across the number feature.

Acknowledgements

We thank Sacha Beniamine, Mae Caroll, Maria Copot, Timothee Mickus, and Erich Round for comments on early versions of this paper.

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), as well as a 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

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language*, 89:429–464.

Farrell Ackerman and Gregory T. Stump. 2004. Paradigms and periphrastic expression. In Louisa Sadler and Andrew Spencer, editors, *Projecting Morphology*, pages 111–157. CSLI Publications, Stanford, CA.

Asaf Bachrach and Andrew Nevins, editors. 2008. *Inflectional Identity*. Oxford University Press, Oxford.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2005. *The Syntax-Morphology Interface: A*

Study of Syncretism. Cambridge University Press, Cambridge.

Laurie Bauer. 2004. 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.

James P. Blevins. 2006. Word-based morphology. *Journal of Linguistics*, 42:531–573.

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

Olivier Bonami. 2015. Periphrasis as collocation. *Morphology*, 25:63–110.

Olivier Bonami and Matías Guzmán Naranjo. 2023. Distributional evidence for derivational paradigms. In Sven Kotowski and Ingo Plag, editors, *The Semantics of Derivational Morphology: Theory, Methods, Evidence*, pages 197–235. de Gruyter.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17:173–195.

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

Geert Booij. 1996. 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.

Gilles Boyé and Gauvain Schalchli. 2016. The status of paradigms. In Andrew Hippisley and Gregory Stump, editors, *The Cambridge Handbook of Morphology*, pages 206–234. Cambridge University Press.

Andrew Carstairs-McCarthy. 1994. Inflection classes, gender, and the principle of contrast. *Language*, 70:737–788.

Maria Copot, Timothee Mickus, and Olivier Bonami. 2022. Idiosyncratic frequency as a measure of derivation vs. inflection. *Journal of Language Modelling*, 10(2).

Greville G. Corbett. 2010. Canonical derivational morphology. *Word Structure*, 3(2):141–155.

Greville G. Corbett. 2012. *Features*. Cambridge University Press, Cambridge.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Jerome H. Friedman. 2001b. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

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

Roman Jakobson. 1958. Morfologicheskie nabljudenija nad slavjanskim skloneniem (sostav russkix padeznyx form). In *Amer- ican contributions to the fourth international congress of Slavists*. Mouton. Reprinted in English translation in Jakobson (1971).

Roman Jakobson. 1971. *Selected Writings II*. Mouton, The Hague.

Paul Kiparsky. 2005. Blocking and periphrasis in inflectional paradigms. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 2004*, pages 113–135. Springer, Dordrecht.

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. 2021. SYN v9: large corpus of written czech. 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. 2022. Package of word embeddings of czech from a large corpus. 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. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3):485–515.

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

P. H. Matthews. 1991. *Morphology*, 2nd edition. Cambridge University Press, Cambridge.

Timothee Mickus, Olivier Bonami, and Denis Paperno. 2019. Distributional Effects of Gender Contrasts Across Categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, volume 2, pages 174–184.

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

Anna Siewierska. 2004. *Person*. Cambridge University Press, Cambridge.

Andrew Spencer. 2013. *Lexical Relatedness*. Oxford University Press.

Pavol Štekauer. 2015. 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.

Jana Straková, Milan Straka, and Jan Hajič. 2014. 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. Association for Computational Linguistics.

Gregory T. Stump. 1993. On rules of referral. *Language*, 69:449–479.

Gregory T. Stump. 2016. *Inflectional paradigms*. Cambridge University Press, Cambridge.

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

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

Arnold M. Zwicky. 1985. How to describe inflection. In *Proceedings of Berkeley Linguistics Society 11*, pages 372–386.