Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms

Yu-Ying Chuang[1], Dunstan Brown[2], Harald Baayen[1], Roger Evans[3]

[1]Quantitative Linguistics, Eberhard-Karls University of Tübingen
[2]Language and Linguistic Science, University of York
[3]CSIUS, University of Brighton

Abstract

This study investigates the phenomenon of defectiveness in Russian case and number noun paradigms from the perspective of distributional semantics. We made use of word embeddings, high-dimensional vectors trained from large text corpora, and compared the observed paradigms of nouns that are defective in the genitive plural, as suggested by Zaliznjak (1977), with the observed paradigms for non-defective paradigms. When the embeddings of about 20,000 inflected forms were projected onto a two-dimensional space, clusters of case and number within case were found, suggesting global semantic similarity for words with the same inflectional features. Moreover, defective lexemes were characterized by lower semantic transparency, in that inflected forms of the same lexeme are semantically less similar to each other, and their meanings are also more idiosyncratic. Furthermore, compared to non-defective lexemes, inflected forms from defective lexemes are further away from the idealized average case-number meanings, obtained by averaging over the vectors of all inflected forms of the same case-number combination. As a consequence, the semantics of defective forms are predicted less precisely by a simple model of conceptualization that assumes that the meaning of a given Russian inflected form is approximated well by the sum of pertinent embeddings of the lexeme, case, and number within case. We conclude that semantics, at least the kind captured by word embeddings, also contributes to the defectiveness of Russian noun paradigms.

1 Introduction

Defective lexemes have incomplete paradigms (Matthews 1997, p. 89; Baerman and Corbett 2010, p. 1).[1] That is, speakers have difficulty agreeing on what the form should be for a particular paradigm cell or set of cells. Unlike pluralia tantum or singularia tantum nouns, for example, where lack of a particular sub-paradigm appears to have a semantic basis, for the canonical defective word there appears to be no clear semantic motivation for the gap in the paradigm (Baerman and Corbett, 2010, p. 1). Defectiveness raises interesting questions for linguistic theory, in particular

[1] Brown and Evans thank the Arts and Humanities Research Council (grant AH/T002859/1 'Feast and Famine: Confronting Overabundance and Defectivity in Language'). Chuang and Baayen are indebted to funding from the ERC, project WIDE-742545. Earlier versions of this research were presented at the Feast and Famine project workshop (October 22, 2021), the Ohio State University Linguistics Department colloquium (October 29, 2021), the Stony Brook University Linguistics Department colloquium (November 19, 2021), the Tübingen-York Virtual Workshop on Morphology and Word Embeddings (January 17-18, 2022) and the Surrey Morphology Group (March 1, 2022). We are grateful to attendees at these meetings for their questions, comments and suggestions. We'd like to thank Mark Aronoff, Matthew Baerman, Neil Bermel, Matt Caroll, Grev Corbett, Dagmar Divjak, Nick Evans, Alex Nikolaev and Andrea Sims for help at different stages of the research.

why speakers are unable to agree on an entirely acceptable form when this is otherwise the norm for most words, irrespective of how much of their paradigm can be observed in corpus data.

Our focus here is to consider a small subset of Russian nouns (about 60, listed in Zaliznjak 1977) that have problematic genitive plural forms, as in Table 1.[2] When asked about nouns like *kočergá* 'poker' native speakers typically have difficulty finding a totally acceptable form for the genitive plural. In this study we make a distinction between nouns like *kočergá* that are 'inherently defective' and those that are 'contingently defective'. For the latter it is just a matter of not having observed the form in a corpus yet. Inherent defectives, of course, do not appear to be amenable to corpus-based analysis: in principle they are not observable, and absence of observation of a form, as contingent defectiveness demonstrates, cannot be taken as observation of absence of a form, to paraphrase a more familiar formulation. Furthermore, given the nature of word form distributions, we expect to encounter contingent defectiveness frequently. In contrast, inherent defectiveness appears to be rare and, most importantly, unexpected, because it should be unproblematic that a completely acceptable form could be produced, given the right context. Some researchers frame defectiveness overall in terms of what is observed and provide evidence that absence of forms facilitates learning (Janda and Tyers 2021). While there is evidence for this overall, it leaves the status of inherent defectives as unaddressed. Of course, a corpus of sufficient size may occasionally provide observations of forms listed elsewhere as defective, which our preparatory work indicates to be the case for some of the nouns on the list from Zaliznjak (1977). As Nikolaev (2022) have shown, inherent defectiveness is also dependent to some extent on language users. Our approach here, however, is to take it as given that there is something special about the nouns listed by Zaliznjak ('inherent defectiveness' in our terms) and see if there is anything interesting about their distributional properties, thereby looking at their usage from a different angle.

SG		PL	
NOM	*kočergá*	NOM	*kočergí*
ACC	*kočergú*	ACC	*kočergí*
GEN	*kočergí*	GEN	- -
DAT	*kočergé*	DAT	*kočergám*
PREP	*kočergé*	PREP	*kočergáx*
INS	*kočergój*	INS	*kočergámi*

Table 1: The Russian noun *kočergá* 'poker'. See Sims (2015, 82–95).

There is evidence that defectiveness can be associated with homophony avoidance (Baerman, 2011). In the set of defective nouns in our study there are examples where the defective genitive plural would have the same form as the nominative singular of another lexeme. However, this is far from the case for many of them. Typical explanations for the problematic nature of the genitive plural centre around issues to do with the form side, including assumptions that multiple alternatives cause the difficulty. In particular, the nature and positioning of 'filler' or 'fleeting' vowels and potentially the positioning of stress appear to play a role. The overwhelming majority of the nouns with problematic genitive plural belong to the declension class whose nominative singular ends in *-a*. This is a large productive class.[3] Furthermore, the overwhelming majority also

[2] The list includes nouns that Zaliznjak (1977) has annotated as either having no genitive plural or one that is considered problematic.

[3] Deriving their counts from Zaliznjak (1977), Brown et al. (1996, p. 57) provide figures on the four key inflection

exhibit a pattern of word prosody where stress falls on the inflection in both the singular and plural, with the exception of the genitive plural itself, where for most nouns belonging to the declension class the stem is the exponent of that case and number combination. The stress pattern is the second most common for Russian nouns (out of eight possibilities). The question of where to position filler vowels means, for instance, that for the example in Table 1 possible forms for the genitive plural include *kočerég*?, *kočérg*?, or *kočeróg**.[4] It should be noted, however, that not all nouns listed as defective present a problem with choice of filler vowel. Furthermore, we should approach with caution an explanation based solely on the avoidance of overabundance (i.e. a choice of possible forms). There are instances of overabundance that may remain stable across centuries (Thornton, 2019); also, historical evidence indicates that defectiveness in first person singular forms of certain non-past Russian verbs may not be the result of synchronic competition between forms so much as lexical specification of a gap where there was once an anomalous alternation (Baerman, 2008), something that Daland et al. (2007) demonstrate can be learned using a multi-agent model with Bayesian learning. Other accounts of defectiveness have focused on the nature of morphological rules. Gorman and Yang (2019) in particular see defectiveness as arising where a number of rules are in competition and none of them can be defined as productive (in terms of Yang's Tolerance Principle, 2016, Chapter 3). However, there are still questions about how we formulate our rules and relate form and paradigmatic meaning in doing so. It seems possible that a variety of factors may conspire to bring about defectiveness. Our aim here is to make a contribution on the meaning side, broadly understood, by looking at the distributional properties of the case and number paradigms of defective nouns. In relation to this, Sims (2015, p. 101) speculates on two ways in which defectiveness can be repaired:

> *I hazard a guess that syncretism may be a natural strategy when there is significant semantic overlap between a problematic paradigm cell and another cell. Semantic closeness may promote the use of one form for both cells. Defectiveness may be a natural strategy when there is a perceived semantic or stylistic incongruity between relevant m-values and a lexeme's meaning ...* (Sims, 2015, p. 101)

Syncretism[5] does not appear to be a viable option to resolve a defective genitive plural in Russian, because — uninflected nouns aside — the genitive plural can be syncretic only for animate nouns, where the form that would otherwise be unique to the genitive plural in non-syncretic nouns is used for the accusative plural as well. So the defective cell itself is the one that would be required for the syncretism, and the directional nature of the syncretism that is observed suggests that it is probably not semantic closeness that brings it about. On the other hand, Sims' conjecture suggests an interesting hypothesis about the distributional properties of Russian defective nouns: While the nature of inherent defectiveness is such that we cannot directly observe semantic or stylistic incongruity for the paradigm cell that is defective, we can do so for some or all of the other cells of lexemes whose paradigms contain a defective genitive plural cell ('defective lexemes'). It is

classes: I (20,690), II (13,611), III (3,929) and IV (5,766). II is the class with nominative singular beginning in *-a*.

[4] According to Zaliznjak (1977) the form should be the first of this set of options, but it is considered problematic. The other options are also problematic, but the second may be possible for some speakers.

[5] Syncretism is where a distinction that is relevant for syntax is not made by the morphology. For instance, the Russian noun meaning 'book' has distinct forms for the nominative and accusative singular, *kniga* (nominative) and *knigu* (accusative), while for the noun 'letter' the form *pis'mo* is used for both case combinations. The latter is considered an instance of syncretism. See Baerman et al. (2005, p. 27-35) for more detailed definitions and Brown and Arkadiev (2018) for a bibliography of key works on syncretism.

possible to observe the distributional properties of the remaining case and number combinations for nouns with defective genitive plurals. The hypothesis is that the remaining paradigm cells of defective nouns are anomalous in the way that they behave distributionally when compared with the majority of nouns. In observing weirdness around the gap, we have some support for assuming that the defective portion itself may involve some oddness in distributional terms. We will go on to show that there is evidence for this claim.

In addressing this hypothesis about the distribution of case and number combinations we use a distributional semantics (see, e.g., Firth, 1968; Landauer and Dumais, 1997; Mikolov et al., 2013) approach, specifically word vectors, to look at case and number in Russian nouns to understand the place of defectives within the wider system. We are, however, mindful of the fact that the method we apply does not distinguish syntactic distribution from semantic information. This is important for at least two reasons: first, canonical defectiveness is not associated with the semantics of the lexeme; second, we should expect there to be baseline patterning related to morpho-syntactic features, because they are themselves defined in distributional terms, as illustrated by Corbett (2012, p. 75-90), including his exposition of how the Moscow set-theoretic school approached their definition (for detailed accounts of the Moscow school approach, see van Helden, 1993; Meyer, 1994). In any event, questions arise about whether we have picked up something interesting in relation to distributional behaviour, broadly understood, that allows the user of the language to establish paradigms for lexemes, thereby suggesting that this process is not entirely felicitous for defectives, or whether the causes of their oddity, although correlated with this pattern, lie elsewhere.

When working with semantic vectors (embeddings) for inflected words, a more general question that needs to be addressed is how to understand these semantic vectors. Within the general framework of realizational morphology, a form such as *kočergáx* is taken to realize the inflectional features [plural] and [dative] for a lexeme that means ‘poker’. Thus, one would expect that the semantic vector calculated for *kočergáx* is a function ϕ of the semantic vectors for plural, dative, and ‘poker’. The Discriminative Lexicon model (Baayen et al., 2019) proposes to implement ϕ using straightforward vector addition, but it is an open question whether this way of formalizing the conceptualization of the meaning of *kočergáx* is correct (for a different approach to semantic compositionality, see Marelli and Baroni, 2015). In order to better understand the distributional semantics of Russian nominal inflection, we will therefore make use of visualization with the t-SNE unsupervised clustering method. This will enable us to assess the factors that structure the distributional space of Russian nouns, forming a baseline against which we can assess the possible semantics of defectiveness.

2 Data

We extracted 504,506 unique word forms and their associated lemmas from the *Araneum Russicum Russicum Maius* corpus (Benko, 2014), using functionality provided by No Sketch Engine (`https://nlp.fi.muni.cz/trac/noske`). The corpus data were further tidied to remove non-cyrillic items. We took the first 10,000 most frequent word forms and used the associated lemmas (lexemes) for these word forms to search the full dataset of 504,506 for further word forms associated with those lemmas. This step allowed us to increase the number of forms observed for the paradigms of the lemmas. Noun lexemes that are listed as having a problematic or non-existent genitive plural in Zaliznjak (1977) were searched for separately in the set of 504,506 word forms and matched with the list from Zaliznjak (1977). The word forms were then matched with two sets of pre-compiled embeddings. The intersection of the corpus forms and the pre-compiled embeddings yielded 27,033 word forms.[6] The association of lexemes and word forms is based on the dataset from the *Araneum Russicum Russicum Maius* corpus, as the pre-compiled embeddings we used did not contain lemma information. We extracted all available embeddings from the two pre-compiled sets, one based on word2vec (Mikolov et al., 2013), and the other based on fasttext (Bojanowski et al., 2017).[7] Whereas the algorithm underlying word2vec treats words (strings of letters bounded by space characters) as elementary units, the algorithm underlying fasttext also works with substrings of words. Especially for languages with complex inflectional systems, this has been found to be an important innovation that avoids problems of data sparsity (see Nikolaev et al., this volume, for the case of Finnish). As fasttext makes use of subword strings, it cannot be ruled out that it picks up on form similarity in addition to semantic similarity. Although we make use mainly of fasttext, we have also used word2vec to replicate critical findings.[8]

For 27,033 forms, a fasttext vector was available. For visualization with t-SNE, duplicate embeddings are not allowed. As syncretic forms have identical embeddings, we associated a form with its most frequent function, basing this on the frequency counts in the *Araneum Russicum Russicum Maius* corpus. This left us with 19,791 forms, among which 19,062 forms also have word2vec vectors available.[9]

3 Visual exploration of the distributional space of Russian nouns

As a first step, we visually explored the distributional space of Russian nouns using t-SNE (Van der Maaten and Hinton, 2008), applied to both fasttext and word2vec embeddings.[10] The t-SNE conducts dimension reduction, projecting the original 300 dimensions onto a two-dimensional plane. Figure 1 visualizes this 2D plane for fasttext embeddings. First consider the right two panels, which contain all the singular and plural forms in our dataset. The upper right panel color-codes for number (gray: singular; pink: plural), and the lower right panel codes for case (black: nominative; red: accusative; green: genitive; light blue: locative; dark blue: dative; purple: instrumental; yellow: vocative). Considered jointly, these two panels show that, surprisingly, words cluster by case, and that within case, they cluster by number, with plural clusters typically occurring at the periphery. The large overlap between red and black clusters is a straightforward consequence of the syncretism of many nominative and accusative forms. It is noteworthy that of all other cases, the genitives are located closest to the origin, indicating that genitives are relatively difficult to differentiate based on their distributional statistics.

The left panels of Figure 1 illustrate the very different clusters that emerge when we consider frequency and paradigm size associated with lexemes. In contrast with the right panel, where all paradigm sizes are included, the data are restricted to those nouns that have at least 12 paradigm

[6] These 27,033 forms correspond to 7,807,999 tokens in the *Araneum Russicum Russicum Maius* corpus.

[7] The fasttext vectors were downloaded from https://fasttext.cc/docs/en/crawl-vectors.html, and the word2vec vectors from https://wikipedia2vec.github.io/wikipedia2vec/pretrained.

[8] The results obtained with word2vec embeddings are provided in the supplementary material (availabel at https://osf.io/gqudb).

[9] The 27,033 forms are unique pairings of word-form and function. The 19,791/19,062 forms, for which fasttext and word2vec vectors were available respectively, are unique forms *sensu strictu*, which we have associated with the most frequent function of the form in question.

[10] For these data, results are robust with respect to small changes in the parameters of the t-SNE algorithm.

members.[11] At that point instead of clustering by case and number, we now observe clustering by lexeme. (As words with smaller paradigms are included in the analysis, the clustering by lexeme morphs into clustering by case and number.)

The two different ways in which inflected forms cluster are highly informative. First, the panels on the left indicate that inflected variants of lexemes are likely to form tight clusters in distributional space. However, the t-SNE clustering technique only sees this when the distributional space is not saturated with lexemes that have many ‘contingently defective’ paradigm cells. In the presence of the many lexemes that have small paradigms (about two-thirds of the lexemes in our dataset have paradigm size smaller than seven), the t-SNE highlights the structure originating from case as well as from number within case. Thus, the distributional space of Russian nouns appears to be structured both by local clustering of inflected variants around their lexemes, and by large-scale similarities originating from case and number.

To further explore this hypothesized local clustering, we calculated, for each lexeme, an average vector by averaging the vectors of its inflectional variants. We likewise obtained, again by averaging, vectors for each case and number combination. We then calculated the correlations between the individual word vectors to the vectors of their lexeme, their corresponding case-number vector. These calculations revealed that word forms tend to be much closer to their lexeme vectors than to their case-number vectors. Below, we discuss this finding in more detail, with specific attention to the specific behavior of defective nouns.

[11] To observe this we had incrementally worked through different paradigm sizes; we noticed a switch in the t-SNE point at 10 different word forms or higher. This is partly also due to the fact that more than 97% of the lexemes in our dataset have paradigm size smaller than 10. When the number of lexemes for different paradigm sizes is controlled for, the boundary shift from lexeme clustering to case-number clustering changes from 10 to 9.

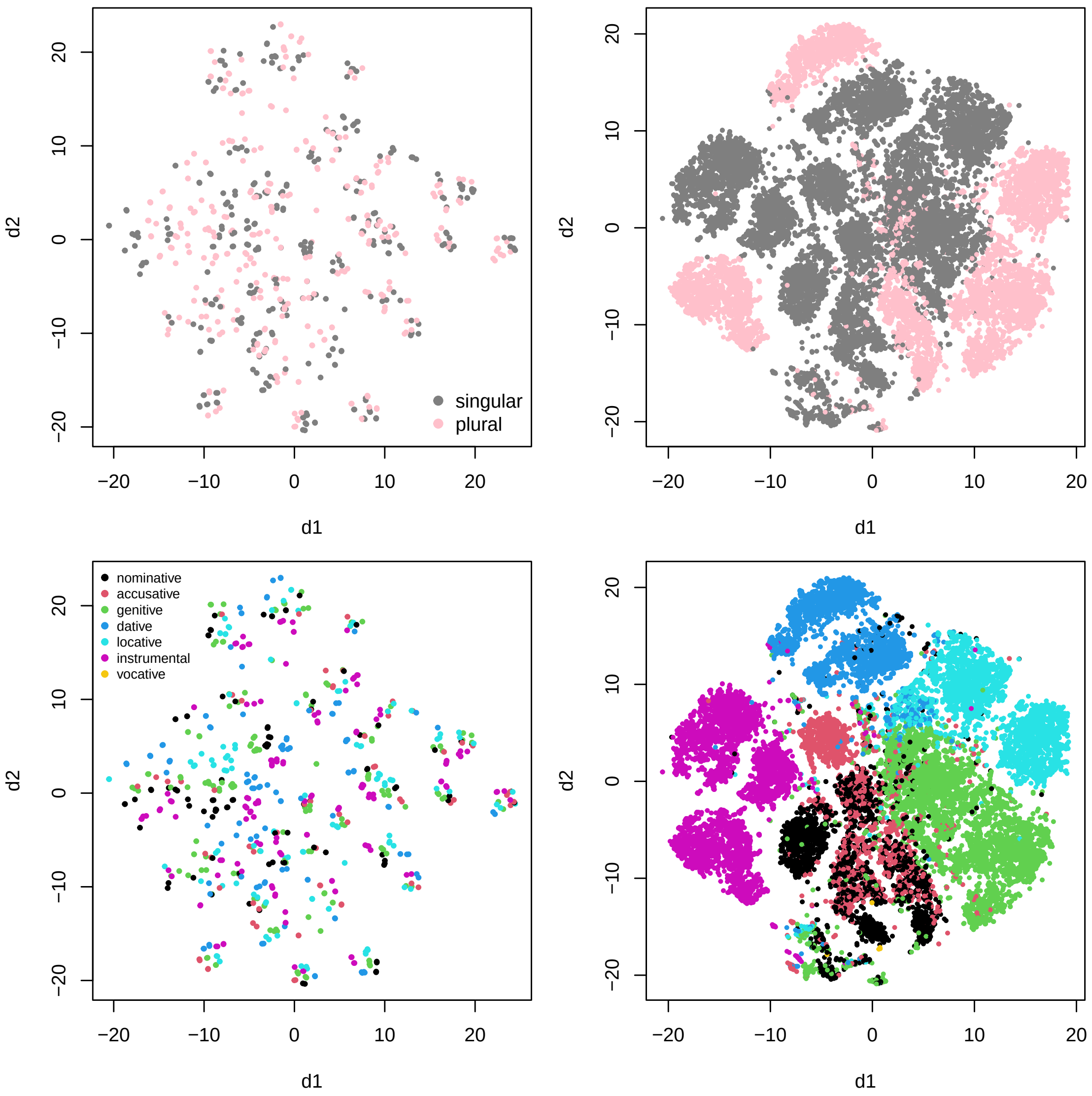
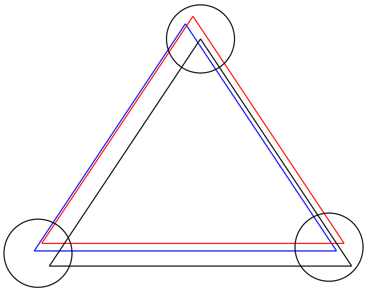


Figure 1: t-SNE clusters of Russian noun word form vectors classified by observed paradigm size and morphosyntactic feature. For nouns with size 12 or greater (column 1) forms cluster into lexeme groups. When all noun forms are included (column 2) they cluster into morphosyntactic feature groups. Colouring forms according to number (top row) or case (bottom row) feature values shows this effect in each feature independently. (An interactive plot for left panels is available here, and an interactive plot for the right panels is available here.

Figure 2: Micro-clusters in combination with identical within-cluster shifts, highlighted by triangles.



In order to further explore the hypothesized global structure provided by case and number, we calculated, for each case separately, the shift vector from the singular to the plural:

$$\overrightarrow{\text{PLURAL}|\text{CASE}} = \overrightarrow{\text{SINGULAR}|\text{CASE}} + \underbrace{(\overrightarrow{\text{PLURAL}|\text{CASE}} - \overrightarrow{\text{SINGULAR}|\text{CASE}})}_{\text{shift vector}}.$$

In other words, shift vectors create plural vectors out of the corresponding singular vectors by straightforward vector addition. (For studies using vector addition to model derivation, see the review in Boleda, 2020). What we expect, given Figure 1, is that the shift vectors for Russian nouns cluster by case. Figure 3 shows that this is indeed the case, independently of whether `word2vec` vectors are used or `fasttext` vectors.

Considered jointly, these observations make it possible to specify a model for the conceptualization of Russian inflected nouns. Let λ denote a lexeme, κ a case, and ν a number. Then

$$\phi(\lambda, \kappa, \nu) = \overrightarrow{\lambda} + \overrightarrow{\kappa} + \overrightarrow{\nu \mid \kappa} \tag{1}$$

(see Nikolaev et al, this volume, for a detailed application of this modeling approach to Finnish). In all likelihood, the case vectors $\overrightarrow{\kappa}$ and the number shift vectors $\overrightarrow{\nu \mid \kappa}$ are relatively small, resulting in constellations of inflected forms that form clusters around their lexeme vectors $\overrightarrow{\lambda}$, as illustrated in Figure 2. When given the full dataset, the t-SNE algorithm detects the high-level clusters based on case, and number within case, because number and case move inflected meanings consistently in different directions, across very large numbers of observations that only partially fill paradigm cells. When the number of observations for case and number are balanced, it is lexeme-based clusters that emerge in the t-SNE map. Both structures are there, but the t-SNE, which is designed to find groups based on geometrical patterning, cannot extract macro-structure and micro-structure at the same time, and will zoom in on the structure that is most pervasively present.

Equation (1) has the important property that it does not require the meaning of a particular inflected form to be derived from that of another inflected form. In the spirit of realizational morphology, the conceptualization process is built on the semantics of the lexeme and the inflectional features that are to be realized. What Equation (1) adds to standard realizational accounts is an interaction of case and number: number is realized differently depending on case (see for analogous results for English, Shafaei-Bajestan et al., this volume, and for Finnish, Nikolaev et al., this volume).

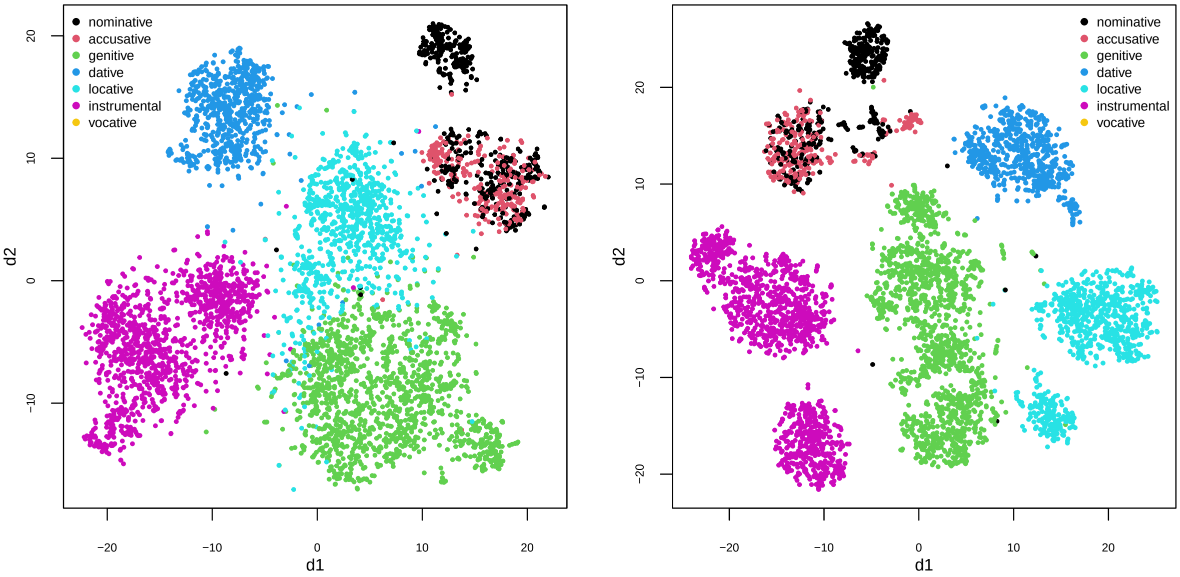


Figure 3: t-SNE clustering of shift vectors for number when case is held constant (left: `fasttext`, right: `word2vec`). Each point (N=5212) represents the difference between the singular and plural form vectors of a lexeme. Shift vectors cluster by case, providing further evidence that number is conceptualized differently for each case.

4 Defectiveness in distributional space

Now that we have an understanding of the structure of the distributional space of Russian nouns, we return to the question of whether defective nouns are defective in part because of their distributional semantics. Consider again Figure 3. The closer shift vectors are to the origin, the less clear their contribution to the inflected word's semantics will be. Do defective nouns suffer from this kind of semantic indeterminacy?

4.1 Semantic transparency and defectiveness

Are defective nouns characterized by lower semantic transparency, compared to non-defective nouns? We operationalized the concept of semantic transparency by first calculating, for a given lexeme, all pairwise correlations of its inflectional embeddings, and then taking the average. This results in a measure of the semantic affinity of the inflected forms of a given lexeme. In terms of the geometry of Figure 2, greater transparency amounts to more concentrated lexeme clusters.

We addressed this question for a dataset containing 47 defective lexemes and 3,070 non-defective ones. For each lexeme, we calculated its unique paradigm size, i.e., the number of unique inflected forms found in the full dataset of 504,506 word forms extracted from the *Araneum Russicum Russicum Maius* corpus, as well as its within-paradigm semantic transparency. To investigate whether we can predict defectiveness with these measures, we fitted a Generalized Additive Model (GAM, Wood, 2017) to the log odds ratio of defectiveness with paradigm size and semantic transparency as predictors. The left panel of Figure 4 shows that as paradigm size increases, the probability of being a defective lexeme decreases, suggesting that defective lexemes tend to have smaller paradigm size. The effect of semantic transparency is presented in the right panel. As there is little data at the lower end (indicated by rugs at the bottom), we do not see a significant effect of semantic transparency within the range of 0 and 0.4. However, from the mid to high transparency, we see a downward trend, suggesting that defectiveness is less likely to be characterized by high semantic transparency. Taken together, the current results indicate that defective lexemes have fewer inflectional variants, which are also semantically less coherent than those of non-defective lexemes.

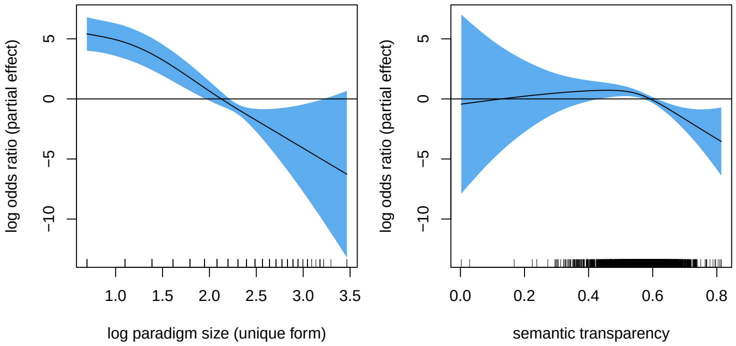


Figure 4: GAM plots showing the log odds of defectiveness against (log) paradigm size (left) and against semantic transparency measure (right). These plots indicate that defectiveness decreases with larger paradigm size and greater semantic transparency.

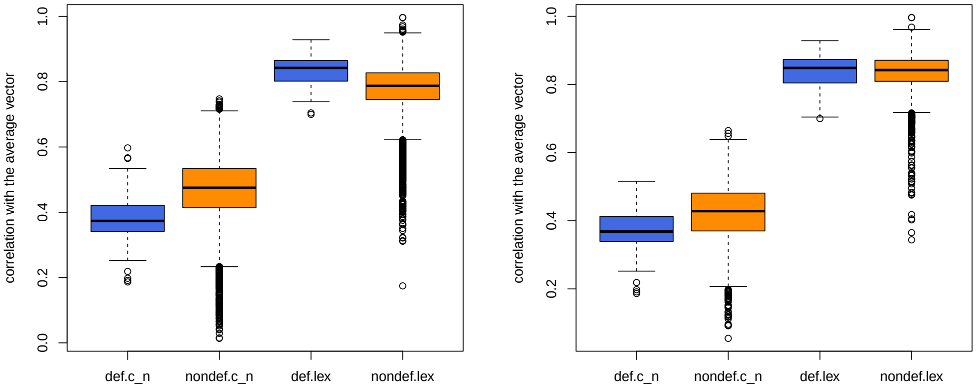


Figure 5: Distribution of correlations (angle) between actual form vectors and morphosyntactic and lexeme vectors, partitioned into defective and non-defective groups. The left panel shows results based on the full dataset, whereas the right panel shows results of lexemes with paradigm size equal to 7 or smaller.

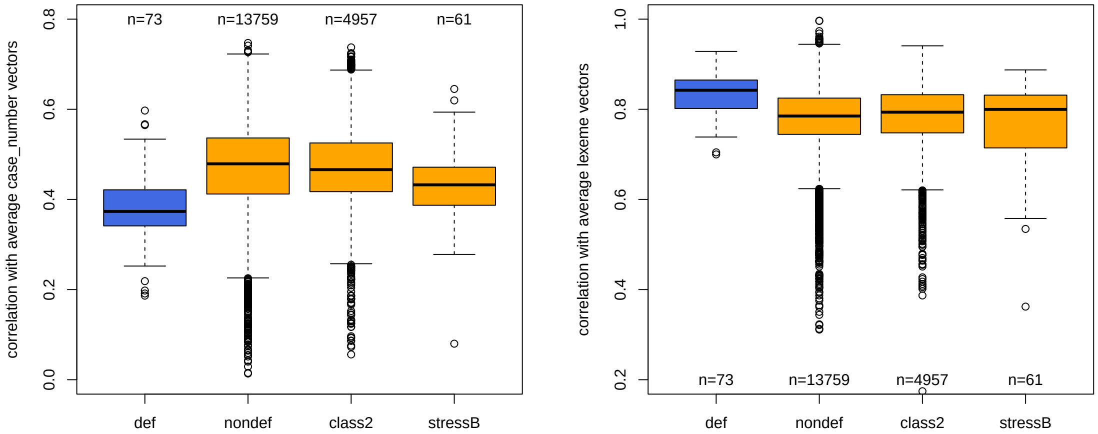


Figure 6: Boxplots of correlations between actual form vectors and morphosyntactic (left) and lexeme (right) vectors, partitioned into defective, non-defective, classII, and classII+stressB groups, for nouns with large paradigms.

4.2 The distributional geometry of defectiveness

We have seen that defective nouns have semantically less transparent paradigms. In what follows, we examine how defective and non-defective nouns pattern with respect to the embeddings of the lexeme, as well as the vectors obtained by averaging over all vectors sharing a given case-number combination.

The left panel of Figure 5 shows the correlations between each inflected form and its respective case-number vector and lexeme vector. Higher correlations indicate higher similarity. Compared to non-defective nouns, inflected forms of defective paradigms are generally less similar to the average case-number vectors, but more similar to the average lexeme vectors. However, since defective nouns usually have smaller paradigm size (cf. Figure 4), this pattern of results might be due to a confound between defectiveness and paradigm size. We addressed this issue by only considering word forms from smaller paradigms. The right panel of Figure 5 shows that once paradigm size is controlled for, the difference of lexeme correlation between defective and non-defective nouns disappears, while the difference in the case-number correlations is still present.[12] Viewing correlation as a measure of cohesion, we can thus conclude that defective forms are less cohesive morpho-syntactically than non-defectives.

As we noted in Section 1, the overwhelming majority of non-defective nouns belong to declension II and have a stress pattern (pattern B in Zaliznjak 1977) where stress falls on the inflection throughout the paradigm when this is possible. In Figures 6 and 7 it can be seen that non-defective declension II nouns behave like non-defectives overall in showing a greater correlation with the average case-number vector when compared with the defective nouns. A similar pattern

[12] It follows that nouns with smaller paradigms are semantically more cohesive than larger paradigms. This observation dovetails well with Shen & Baayen, this volume.

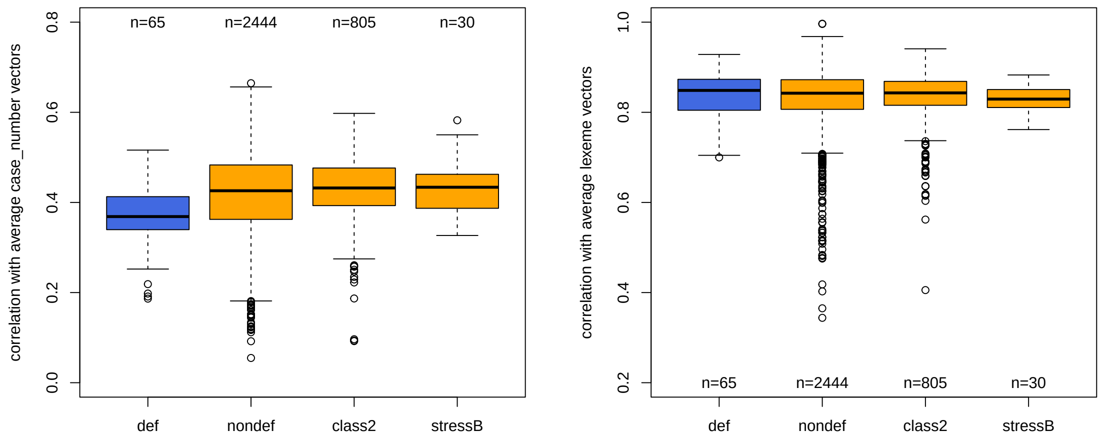


Figure 7: Boxplots of correlations between actual form vectors and morphosyntactic (left) and lexeme (right) vectors, partitioned into defective, non-defective, classII, and classII+stressB groups, for nouns with small paradigms.

is observed for non-defective stress pattern B nouns belonging to declension II. In relation to the average lexeme vector, non-defectives of these classes behave similarly to non-defectives overall, irrespective of whether paradigm size is controlled for (Figure 7) or not (Figure 6). This shows that the distributional behaviour of the majority of defectives cannot be associated with the declension class of which they are a subset, nor is there much support for the idea that their anomalous distributional behaviour can be attributed to the stress pattern to which they belong.[13]

4.3 Defectiveness and predicted semantic vectors

Figure 3 revealed an interaction between case and number: the plural clusters are positioned differently depending on case. Above, we proposed a decompositional model in which the meaning of a Russian noun is the sum of its lexeme, case, and case-conditional number meaning (Equation 1). To complete this model specification, we need to extend it with an error vector, $\overrightarrow{\epsilon}$, representing a word form's semantic idiosyncracies (as well as measurement error in the embeddings):

$$\phi(\lambda, \kappa, \nu) = \overrightarrow{\lambda} + \overrightarrow{\kappa} + \overrightarrow{\nu \mid \kappa} + \overrightarrow{\epsilon}. \quad (2)$$

As defective nouns are in general semantically more idiosyncratic, we hypothesize that this model will fit non-defective nouns better than defective ones. In other words, if we reconstruct the meaning of Russian nouns using the proposed model (2), we should find that the error ($\overrightarrow{\epsilon}$, the difference between predicted and observed embeddings) is larger for defective inflected forms, assuming the same amount of measurement error. Similar to the analyses presented in the preceding section, we

[13] We thank Matthew Baerman and Greville Corbett for suggesting this possibility for consideration. This finding does not rule out the role of the form side, declension and stress pattern, in contributing to defectiveness in Russian nouns. It probably shows that they cannot be used on their own to account for it.

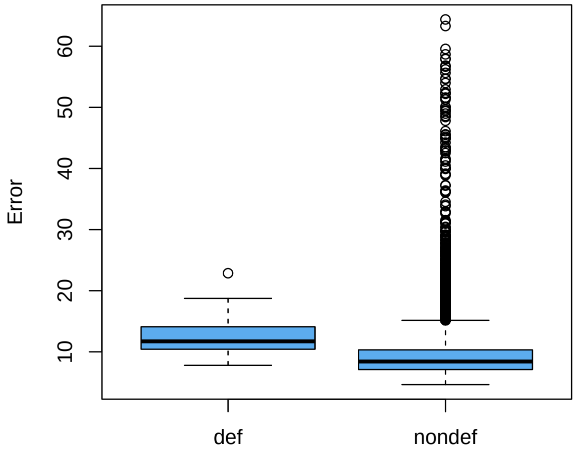


Figure 8: The degree of which the reconstructed semantic vector deviates from the empirical embedding, gauged by the L1-norm of the error vector, for defective and non-defective nouns.

first calculated the average vectors for every lexeme and case, and number shift vectors conditional on case, and reconstructed a predicted embedding for every word in the dataset. We then subtracted the predicted embedding from the empirical one to obtain the error vector. To gauge the degree of deviation from observed vector, we took the L1-norm (the sum of absolute values) of the error vectors. The distribution of the L1-norm for defective and non-defective inflected forms is shown in Figure 8. As expected, defective forms indeed have larger error vectors (two-sample Wilcoxon tests, $W = 1148240, p < 0.0001$), suggesting that their meanings deviate from their theoretically predicted meaning to a larger extent as compared to non-defective forms.[14]

5 Concluding remarks

This study reports on an investigation into possible semantic factors co-determining defectiveness in Russian noun paradigms. Using distributional semantics, we have shown that, compared to non-defective nouns, defective nouns have inflected variants that are less transparent, that are further away from the vectors for case and number, and have semantic vectors that can be predicted less accurately. Of course, it cannot be concluded that the trends we have observed provide a causal explanation of the oddity of defective nouns. But we hope that the quantitative trends we have observed will contribute to a better understanding of the many constraints that together give rise to defectiveness in the Russian noun system.

We have also shown that when empirical embeddings for Russian nouns are decomposed into

[14] Note that for this set of analysis, we did not replicate the results with `word2vec` embeddings. This is likely due to the fact that the case and number clustering structures are less clear-cut for the `word2vec` embeddings, according to the t-SNE analyses. Both results can be found in the supplementary material.

vectors representing lexemes, case, and number, we need to condition the vector of number on case (see also Shafaei-Bajestan, this volume, for English noun plurals, and Nikolaev et al. for Finnish nouns). This finding clarifies that the way in which the Discriminative Lexicon model (Baayen et al., 2019) approximates inflection, namely by simple vector addition, is not precise enough. Likewise, our results also suggest that realizational theories of morphology need to reflect on how the observed case-conditioned semantics of plurality is best accounted for.

A further contribution of our study is the insight it offers into two kinds of similarities that are brought to light by our t-SNE analyses, depending on the input supplied to this unsupervised clustering method. When the t-SNE is supplied with only complete or nearly complete paradigms, it finds clusters based on lexemes. When supplied with data that are not screened for paradigm size, the t-SNE finds clusters based on case, and number within case. This is perhaps unsurprising, as a vast majority of nouns have paradigms with many paradigm cells that are not attested in the corpora that we consulted. As a consequence, across all inflected words, given the high incidence of contingent defectiveness, case, and number within case, appear to provide robust and pervasive structure to the semantic space of Russian nouns. Importantly, Russian inflected nouns that have different lexemes but share case, number, or both, are also similar in meaning (compare, e.g., English, *on the table* with *on the mountain*) even though this similarity does not hinge on the similarity of the lexemes. The consequences for lexical processing of the 'global' semantic similarity that is grounded in case and number, and the 'local' semantic similarity that is grounded in individual lexemes, is a topic that we think is worth further empirical investigation.

More in general, it is an open question how, across languages from very different language families, the realization of multiple morpho-syntactic features and their interactions are best understood. An attempt to model the more complex inflectional system of Finnish nouns is presented in Nikolaev et al. (this volume), but research should be directed not only to nominal inflection, but also verbal inflection and compounding (for compounding in Mandarin Chinese, see Shen Tian & Baayen, this volume).

In the research presented here, we have assumed that word embeddings are a valid tool for investigating inflectional semantics. Fortunately, our central results do not depend on whether `fasttext` or `word2vec` vectors are used. Nevertheless, it is not clear to us what exactly is captured by word embeddings, and to what extent the embeddings for Russian nouns are reflecting distributional structure that goes beyond lexical semantics and the semantics of case and number. It is possible that current Russian embeddings are picking up subtle distributional information that has escaped our attention, but that actually is crucial for Russian speakers to establish paradigms for lexemes. However, whatever the precise nature of this hidden distributional information might be, given the present results, it is unlikely to be entirely felicitous for defectives. We conclude that investigating in further detail possible semantic factors that co-determine defectiveness is a profitable enterprise.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*.

Baerman, M. (2008). Historical observations on defectiveness: the first singular non-past. *Russian Linguistics*, 32(1):81–97.

Baerman, M. (2011). Defectiveness and homophony avoidance. *Journal of Linguistics*, 47(1):1–29.

Baerman, M., Brown, D., and Corbett, G. G. (2005). *The Syntax-Morphology Interface: A Study of Syncretism.* Cambridge Studies in Linguistics. Cambridge University Press.

Baerman, M. and Corbett, G. G. (2010). Introduction: Defectiveness: Typology and diachrony. In Baerman, M., Corbett, G. G., and Brown, D., editors, *Defective Paradigms: Missing forms and what they tell us*, pages 1–18. Cambridge University Press.

Benko, V. (2014). Compatible sketch grammars for comparable corpora. In Abel, A., Vettori, C., and Ralli, N., editors, *Proceedings of the 16th EURALEX International Congress*, pages 417–430, Bolzano, Italy. EURAC research.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annu. Rev. Linguist.*, 6:1–22.

Brown, D. and Arkadiev, P. (2018). *Syncretism (second edition).* Oxford University Press. Oxford Bibliographies in Linguistics. New York: Oxford University Press.

Brown, D., Corbett, G. G., Fraser, N. M., Hippisley, A., and Timberlake, A. (1996). Russian noun stress and network morphology. *Linguistics*, 34:53–107.

Corbett, G. (2012). *Features.* Cambridge Textbooks in Linguistics. Cambridge University Press.

Daland, R., Sims, A. D., and Pierrehumbert, J. (2007). Much ado about nothing: A social network model of Russian paradigmatic gaps. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 936–943, Prague, Czech Republic. Association for Computational Linguistics.

Firth, J. R. (1968). *Selected papers of J. R. Firth, 1952–59.* Indiana University Press.

Gorman, K. and Yang, C. (2019). When nobody wins. In Rainer, F., Gardani, F., Dressler, W. U., and Luschützky, H. C., editors, *Competition in Inflection and Word-Formation*, pages 169–193. Springer International Publishing, Cham.

Janda, A. L. and Tyers, M. F. (2021). Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*, 17(1):109–141.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

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

Matthews, P. H. (1997). *The concise Oxford dictionary of linguistics.* Oxford University Press.

Meyer, P. (1994). Grammatical categories and the methodology of linguistics: Review article on van helden, w. andries: 1993, 'concept formation between morphology and syntax'. *Russian Linguistics*, 18:341–377.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nikolaev, Alexander; Bermel, N. (2022). Explaining uncertainty and defectivity of inflectional paradigms. *Cognitive Linguistics*. in press.

Sims, A. D. (2015). *Inflectional Defectiveness*. Cambridge University Press.

Thornton, A. M. (2019). *Oxford Research Encyclopedia of Linguistics*, chapter Overabundance in morphology. Oxford University Press.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

van Helden, W. A. (1993). *Case and gender: Concept formation between morphology and syntax*, volume II volumes of *Studies in Slavic and general linguistics*. Rodopi, 20 edition.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Yang, C. (2016). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press.

Zaliznjak, A. A. (1977). *Grammatičeskij slovar' russkogo jazyka*. Russkij jazyk, Moscow.