Article:

Chuang, Yu-Ying, Brown, Dunstan orcid.org/0000-0002-8428-7592, Baayen, Harald et al. (1 more author) (2023) Paradigm gaps are associated with weird "distributional semantics" properties:Russian defective nouns and their case and number paradigms. The Mental Lexicon. ISSN: 1871-1340

https://doi.org/10.1075/ml.22013.chu

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] and Roger Evans[2]
[1] Eberhard-Karls University of Tübingen | [2] University of York

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 nouns. 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 the relationship between defectiveness and semantics, at least the kind captured by word embeddings, is stronger than has been anticipated previously.

Keywords: Russian noun paradigm, defectiveness, distributional semantics, case, number

https://doi.org/10.1075/ml.22013.chu | Published online: 30 May 2023
The Mental Lexicon ISSN 1871-1340 | E-ISSN 1871-1375

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 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.

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

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*

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 may 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. The terms 'defectiveness' and associated adjec-

[1] 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 Tubingen-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, Mae Carroll, Grev Corbett, Marina Chumakina, Dagmar Divjak, Nick Evans, Alex Nikolaev, Andrea Sims, Volya Kapatsinski and Vito Pirrelli for help at different stages of the research.

[2] The list includes nouns that Zaliznjak (1977) has annotated as either having no genitive plural or one that is considered problematic. The full list of defective nouns is provided in the Appendix.

tive 'defective' are usually reserved for inherent defectives only. 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. Sims (2015, p. 26) provides a working definition of defectiveness in which a paradigm cell – such as genitive plural in the example given here – is defective, because of the ungrammaticality that arises when any form of a lexeme associated with that cell is inserted into an otherwise well-formed syntactic structure. But, as Sims (2015, p. 37) later shows, the classification of defectiveness may depend on analytical choices, particularly as regards the relationship between morphology and syntax as defined by morphosyntactic features, and not just on the uncertainty associated with the form, in particular raising the question how we establish paradigm cells. In contrast with the view that defectiveness is about inherent defectives 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 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 and Bermel (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.

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. An important consideration is uncertainty over the shape of the stem, the exponent of the genitive plural, in particular the nature and positioning of 'filler' or 'fleeting' vowels, also known as yers, even though it is possible to make generalisations about the appearance of these (Gouskova and Becker 2013; Becker and Gouskova 2016). 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 exhibit a pattern of word prosody where stress falls on the inflection in both the singular and plural. The noun *kočergá* in Table 1 also exemplifies this property, as each of the inflectional affixes bears the stress (rather than the stem *kočerg-*). This stress pattern is the second most common for Russian nouns (out of eight possibilities.)[4] In this declension class there is normally no overt affix for the genitive plural, and the stem is the exponent of that case and number combination. Filler vowels may be used to break up consonant clusters at the end of the unaffixed stem in the genitive plural. While the stress pattern associated with inflection throughout the noun's paradigm may lead to an expectation of the final syllable of the stem being stressed in the genitive plural (in the absence of an overt affix), the question of the use or position of a filler vowel can create uncertainty. For instance, in Table 1 possible forms for the genitive plural include ?*kočerég*, ?*kočerëg*, ?*kočérg*, or **kočeróg*.[5] 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

[3] Deriving their counts from Zaliznjak (1977), Brown et al. (1996, p. 57) provide figures on the four key inflection 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] For Russian nouns there are four basic patterns that can be defined in terms of position of stress on the stem or inflection: (a) stress fixed on the stem throughout the singular and plural; (b) stress on the inflection throughout the singular and plural; (c) stress on the stem in the singular and on the inflection in the plural; (c) stress on the inflection in the singular and on the stem in plural. A further four patterns constitute deviations from patterns (b), (c) and (d). Each of these deviations brings about stem stress where there would otherwise be inflection-stress in the relevant major stress pattern, and only in the nominative plural or accusative singular. Furthermore, the accusative singular can only deviate (and therefore bear stem stress), if the nominative plural has stem stress. These constraints mean that there are two sub-patterns of pattern (b), one sub-pattern of pattern (c) and one sub-pattern of pattern (d). In all that gives eight possible patterns, four major patterns, and a further four sub-patterns based on those patterns. See Brown et al. (1996) for an overview of these generalisations and an implemented model.

[5] We have given the forms in transliteration, rather than phonological transcription, but stress has also been included, even though this is not usually done for written forms. The grapheme ë is also not often written, instead being represented by e. It indicates that the underlying vowel is /o/ and that the preceding consonant is palatalized. According to the electronic version of Zaliznjak (1977) (Ilola and Mustajoki, 1989) the preferred form should be the first of the set of options given, but it is considered problematic. Švedova (1984, p. 259) indicates that the plural form is *kočerëg*. The third of the listed forms may be possible for some speakers, while the form **kočeróg* does not appear to be acceptable for anyone, even though other nouns with nominative singular ending in the string /rga/ exhibit /o/ as a fleeting vowel in the genitive plural.

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 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. There are a number of different ways in which meaning could play a role in defectiveness, or its repair. Viewed from a paradigm-based perspective each cell covers part of the space of meaning associated with the lexeme as a whole. For many lexemes the partition of this space may be fairly consistent. However, where the role of a particular cell in the paradigm is uncertain this may affect how the space is partitioned across other cells in the paradigm. What is harder to judge is the causal direction related to this partitioning. For defective nouns either the semantics of the observable forms are such that they cause the uncertainty associated with the missing form, or alternatively the remaining forms fit into a system where their semantics are distorted by the need to compensate for lack of a viable realization. In relation to the repair of defectiveness, there are hypotheses about semantic proximity between paradigm cells facilitating the avoidance of defectiveness by sharing an unproblematic realization across cells (i.e. syncretism).[6] This is a possibility discussed by Sims (2015, p. 101). 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

[6] 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.

genitive plural cell ('defective lexemes'). It is 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. The fine-grained representations afforded by word vectors make them well-suited to some advanced semantic substitution tasks, but our starting point is that fundamentally they are distributional models which are linguistically holistic, not just semantic. Syntactic and morphological features are also distributional, as illustrated by Corbett (2012, p. 75–90), including his exposition of how the Moscow set-theoretic school approached their definition (van Helden, 1993; Meyer, 1994), and so are also accessible in principle in such models.

This is important because on the one hand canonical defectiveness is not associated with the semantics of a word or lexeme (Baerman and Corbett, 2010, p. 1), but on the other hand actual defectiveness may sometimes depend on some semantic or collocational interactions. Viewing 'distributional semantics' as a holistic model allows us to explore the distributional space for evidence of defectiveness without pre-judging where we might find it. Whatever we might find, the question remains whether we have picked up something interesting in relation to distributional behaviour of 'lexeme paradigms', which accounts for missing forms within them, 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 *kočergámi* is taken to realize the inflectional features [plural] and [instrumental] for a lexeme that means 'poker'. Thus, one would expect that the semantic vector calculated for *kočergámi* is a function ϕ of the semantic vectors for plural, instrumental, 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 *kočergámi* 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 visualisation 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 NoSketch 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.[7] 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).[8] Whereas the algorithm underlying `word2vec` treats words (strings of letters bounded by space characters) as elementary units, the algorithm underlying `fast-`

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

[8] 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. Fasttext embeddings are trained on Wikipedia and Common Crawl: the former consists of about 823 million word tokens, and the latter 102 billion. `Word2vec` embeddings are trained on Wikipedia only. For both embeddings, window size of training are set to five, and both are of 300 dimensions. More details of model setup can be found in Grave et al. (2018) and Yamada et al. (2020).

text 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. As fasttext makes use of subword strings, it cannot be ruled out that it picks up on form similarity in addition to distributional similarity. By way of example, since both *walking* and *eating* have the sublexical string *ing*, it can be argued that the similarity found between the embeddings of *walking* and *eating* is due in part to the sharing of the letter string *ing* (also found in words like *king, sing*, and *ring)*, and not due to just the sharing of the PROGRESSIVE meaning. Although we make use mainly of fasttext, we have also used word2vec to replicate critical findings.[9]

For 27,033 forms, a fasttext vector was available. For visualisation 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.[10]

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.[11] t-SNE is a dimension reduction technique, which we used to project the original 300 dimensions onto a two-dimensional plane. Figure 1 visualises this 2D plane for fasttext embeddings. First consider the righthand 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 towards the periphery. The large overlap between red and black

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

[10] 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.

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

clusters is a straightforward consequence of the syncretism of many nominative and accusative forms – the colour choice corresponds to the most frequent case for each form.

The lefthand 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 righthand panel, where all paradigm sizes are included, the data on the left are restricted to those nouns that have at least 12 paradigm members whose embeddings are available. For these larger paradigms, 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.[12]

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 analysis 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.

It should be kept in mind that word embeddings capture not only lexical semantics but also similarities in the use of syntactic constructions. What is picked up by the t-SNE analyses depends on which kind of similarity dominates in the dataset. If the number of observations for a given lexeme is not too small compared to the number of observations for a given case-number combination (mean tokens per lexeme: 14.82; mean tokens per case-number: 55.58), then the t-SNE reveals clustering by lexemes, indicating that the lexical semantics provide stronger evidence for grouping than the syntactic information provided by case and number (as shown in the left panel of Figure 1). Conversely, when the number of lexemes is large (3055 in our dataset), and many lexemes have partially filled paradigms, then there is abundant information about case and number (mean

[12] To observe this we had incrementally worked through different paradigm sizes (number of available embeddings); we noticed a switch in the t-SNE analysis 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.

tokens per case-number: 1587.83), and only a few observations about any given lexeme (mean inflectional variants per paradigm: 6.24). What Figure 1 clarifies is that in this situation, syntactic similarity outweighs lexical similarity.

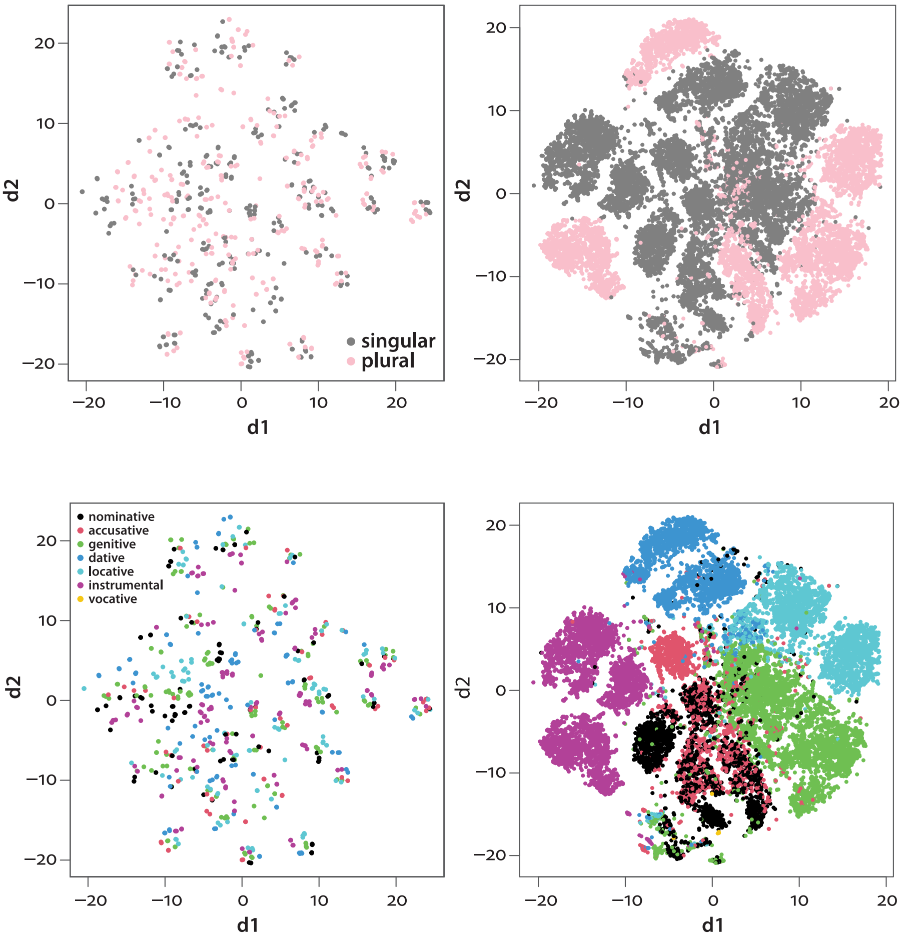

Figure 1. t-SNE clusters of Russian noun word form vectors classified by observed paradigm size and morphosyntactic feature. For lexemes with more than twelve inflectional variants whose embeddings are available (column 1), word forms cluster into lexeme groups. When all noun forms are included (column 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.)

To further explore this hypothesised 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 more correlated with their lexeme vectors than with their case-number vectors. Below, we discuss this finding in more detail, with specific attention to the specific behavior of defective nouns.

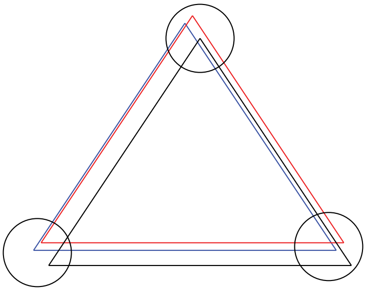

Figure 2. Circles highlight lexeme clusters. Within each lexeme cluster, the relative positions of the inflected variants are the same. This gives rise to high-level similarities that are highlighted by triangles (i.e., a given case-number combination)

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:

$$\overrightarrow{\textsc{plural}\vert\textsc{case}} = \overrightarrow{\textsc{singular}\vert\textsc{case}} + \underbrace{(\overrightarrow{\textsc{plural}\vert\textsc{case}} - \overrightarrow{\textsc{singular}\vert\textsc{case}})}_{\text{shift vector}}.$$

The pipe sign stands for "conditional on". For detailed discussion of shift vectors, see ShafaeiBajestan, et al. (this volume). The idea is straightforward. A shift vector starts at the point in space where the singular is located, and "moves" this point to where the corresponding plural is located. 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, *k* a case, and *ν* a number. Then

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

In all likelihood, the case vectors $\vec{\kappa}$ and the number shift vectors $\overrightarrow{\nu|\kappa}$ are relatively small, resulting in constellations of inflected forms that form clusters around their lexeme vectors $\vec{\lambda}$, as illustrated by the points within the circles 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.

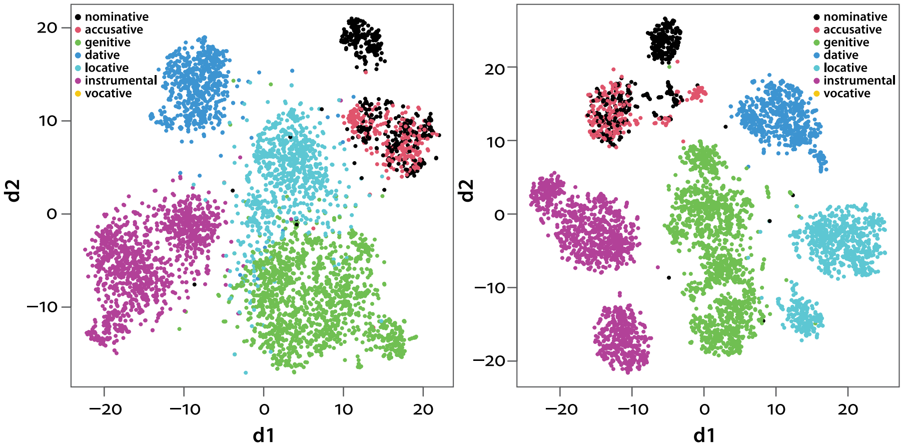

Figure 3. t-SNE clustering of shift vectors for number when case is held constant (left: `word2vec`, right: `fasttext`). Each point ($N=5212$) represents the difference between the singular and plural form vectors of a lexeme. Shift vectors cluster by case, providing further evidence that number is conceptualized 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[13] and 3,070 nondefective 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.

Due to the composition of our dataset, it does not make sense to include lemma frequency as a predictor of the log-odds of defectivity: The vast majority

[13] Nine of the defective lexemes do not have any inflectional variants found in our full dataset, and were therefore excluded from all the analyses.

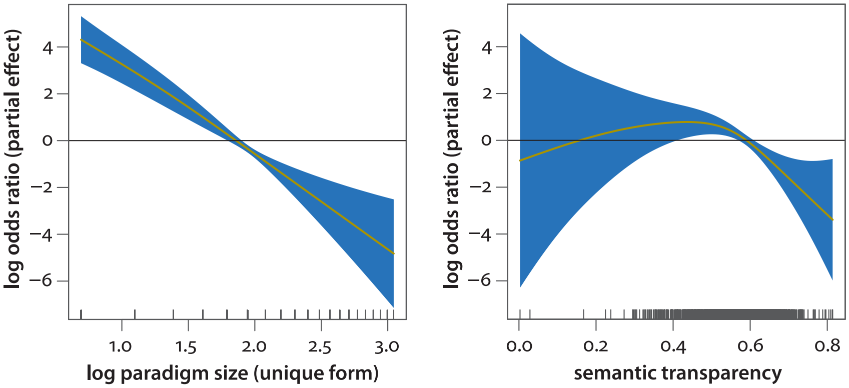

Figure 4. GAM plots showing the log odds of defectiveness against (log) paradigm size (left) and against semantic transparency measure (right). These plots indicate that defectiveness decreases with larger paradigm size and greater semantic transparency

of defective nouns has a lemma frequency that is much lower than the lemma frequencies of non-defective nouns. Below, we investigate the consequences of differences of lemma frequency from a different perspective. Here, we offer two general observations. First, frequent words with large paradigms offer more opportunities for semantic/syntactic drift and greater opacity within the paradigm. Given that defectives pervasively have small paradigms, we would have expected them to have higher semantic transparency, which is exactly opposite to what we have observed (cf. right panel, Figure 4). Second, embeddings are engineered to be highly independent of frequency of use. For the current data, neither the L1 norm (city block distance) nor the L2 norm (euclidean distance) of the `fasttext` vectors correlates with frequency ($r = -.02$ and $-.01$ respectively).[14] This suggests to us that it is highly unlikely that results obtained with embeddings are just mirroring back effects of frequency.

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.

[14] With regard to `word2vec` vectors, a nonlinear trend was observed. While the norms of mid-to-high frequency words are negatively correlated with frequency (i.e., higher frequency, shorter norms), in the low-to-mid frequency range (where most of our defective nouns are), the norms are relatively independent of frequency.

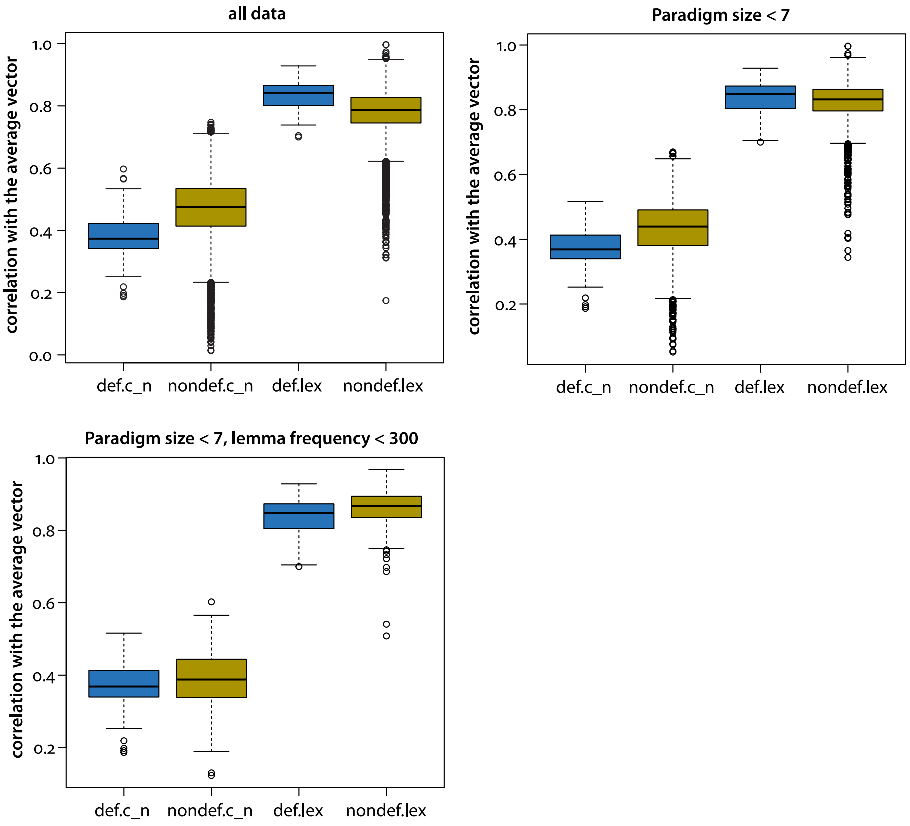

Figure 5. Distribution of correlations (angle) between actual form vectors and case-number and lexeme vectors, partitioned into defective and non-defective groups. The top left panel shows results based on the full dataset, the top right panel shows results of lexemes with paradigm size smaller than 7. In the bottom panel, not only paradigm size but also lemma frequency is controlled for, where we plot the results of lexemes with paradigm size smaller than 7 and lemma frequency lower than 300

The top 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. A two-way ANOVA supported this observation, with a significant interaction effect between def (defective/non-defective) and type (lexeme/case-number). Post-hoc analyses with the TukeyHSD test confirmed significant differences between defectives and non-defectives for both within lexeme and within case-number comparisons, though with opposite signs (both with adjusted $p < .0001$). 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 top 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. This pattern of results is again confirmed by ANOVA. Post-hoc analyses with the TukeyHSD test showed that while the defective and non-defective difference remains significant for case-number (adjusted $p < .0001$), the within-lexeme difference is not supported. A further control that we made is to also limit non-defective nouns' lemma frequency. Given that the vast majority of defective nouns have lemma frequency lower than 300, we therefore also only included non-defective nouns within this frequency range. As shown in the bottom panel, defective forms tend to have lower correlations, for both the case-number and lexeme comparisons. A two-way ANOVA reports significant main effects of both def and type, but no interaction. Post-hoc tests showed that for both case-number and lexeme comparisons, defectives have lower correlations than non-defectives ($p < .05$). Taken together, defectives always have lower correlation with average case-number vectors. Viewing correlation as a measure of cohesion, we can thus conclude that defective forms are less cohesive morpho-syntactically than non-defectives. On the other hand, results of lexeme comparisons depend crucially on which lexemes are included for comparison. Compared to non-defectives that are of similar frequency range and paradigm size, defectives are less close to their average lexeme vectors. This observation indicates that lack of semantic closeness with other paradigm cells is associated with the defectiveness we observe here. Below, we will discuss the consequences of differences in frequency in further detail.

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 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 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 (lower row, Figure 6) or not (upper row). Statistical analyses also led to very similar conclusions as suggested by the overall results presented in Figure 5. For the full dataset (upper row, Figure 6), defectives have lower correlations than the other three non-defective groups for the case-number comparison (all adjusted $p < .005$), and have higher correlations for the lexeme comparison (all adjusted

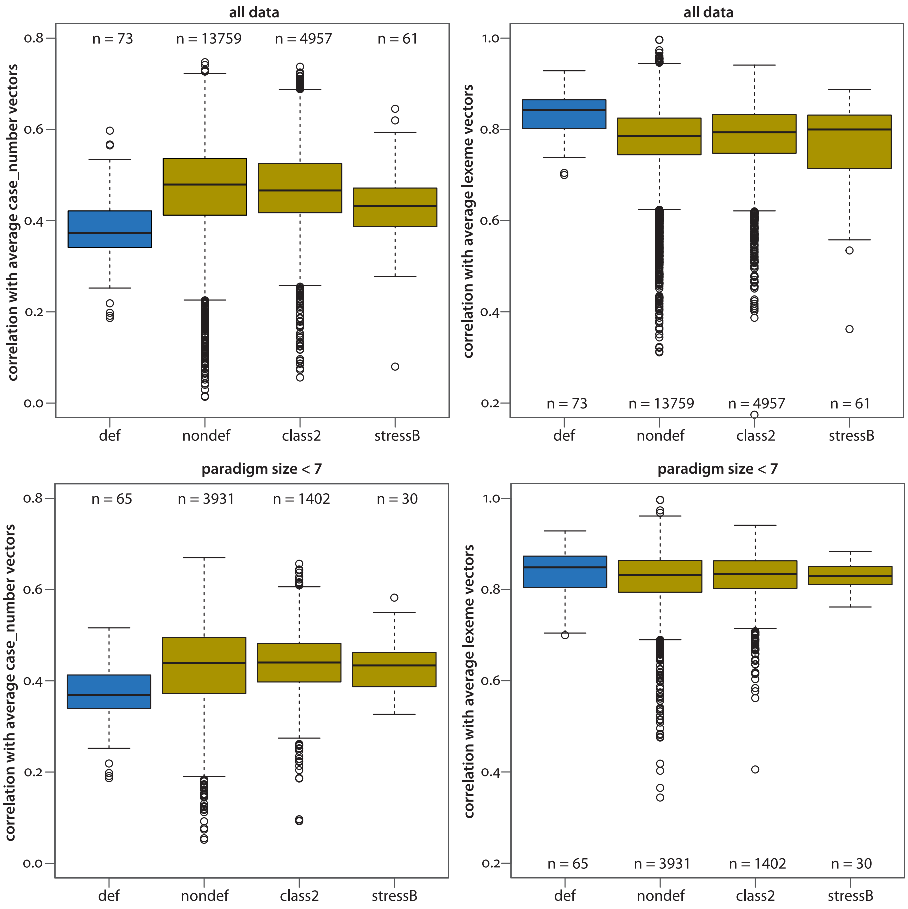

Figure 6. Boxplots of correlations between actual form vectors and morphosyntactic (left column) and lexeme (right column) vectors, partitioned into defective, non-defective, classII, and classII+stressB groups. The plots on the upper row include all nouns in the dataset, while those on the lower row are based on nouns with small paradigms (size < 7)

$p < .0001$). For the smaller dataset (lower row), whereas the differences for case-number are still present (all adjusted $p < .0005$), those for lexeme have disappeared. 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.

One question that remains is whether the difference observed between defective and non-defective nouns is due to differences in the frequency of use of the inflected variants. As shown in the left panel of Figure 7, in our dataset, defectives are predominantly low frequency words, as compared to non-defectives. When fitting a GAM to our correlation measures with frequency as predictor, for only non-defective nouns, we observed an opposite effect of frequency on the morphosyntactic (case-number) vector and the lexeme vector. As shown in the middle panel of Figure 7, while the case-number correlation increases with frequency (middle panel), the lexeme correlation decreases with frequency (right panel). To understand the opposite effects of frequency on the two correlation measures, we provide a graphical illustration in the left panel of Figure 8. Suppose that we have a high frequency lexeme with five inflectional forms (Hcn$_1$, Hcn$_2$, Hcn$_3$, Hcn$_4$, Hcn$_5$), and a low frequency lexeme with only three inflectional forms (Lcn$_1$, Lcn$_3$, Lcn$_4$). Given that high frequency word forms are likely to be polysemous and tend to have case-number specific idiosyncratic meanings, these inflectional forms are therefore farther apart from each other in the semantic space. On the contrary, low frequency word forms, usually with smaller paradigm size, have less chance to develop idiosyncratic meanings, so they are closer to each other in the semantic space (compared the blue triangle to the red pentagon in the figure). This leads to the consequence that when we calculate the average lexeme vector, the low frequency lexeme vector $(\vec{L})$ is closer to all the vectors of its inflectional forms, whereas the high frequency lexeme vector $(\vec{H})$ is at a greater distance from those of its inflectional forms. For instance, as indicated in the plot, the angle between $\vec{H}$ and $\overrightarrow{H_{cn1}}$ will be wider than that between $\vec{L}$ and $\overrightarrow{L_{cn1}}$, resulting in a lower correlation for the high frequency lexeme and a higher correlation for lower frequency lexeme. This explains the negative trend that we observed for lemma frequency and lexeme correlation.

With regard to the case-number correlation, it turns out that the average case-number vectors tend to be closer to their corresponding high frequency forms. In other words, high-frequency inflected words are more similar to their respective theoretical (or prototypical) case-number meaning than low frequency inflected words. To understand why high frequency words dominate the prototypical case-number meaning, we took the dative singular forms of the top 10% high-frequency lexemes and those of the 10% least frequent lexemes in our data. We then calculated all pairwise correlations among the high-frequency dative singulars and among the low-frequency ones. The right panel of Figure 8 presents boxplots of the pairwise correlations for high and low frequency forms. As can be seen, the dative singulars of high frequency lexemes are altogether more similar to each other than the dative singulars of low frequency lexemes. This suggests that when we calculate the average dative singular vector (by averaging over all dative singular forms), this average vector will be more drawn to the centroid of high frequency forms. The result is illustrated in the graphical representation in the left panel of Figure 8: The average vector for the first inflected form $(\overrightarrow{CN_1})$ is closer to the corresponding inflected form of the high frequency lexeme $(\overrightarrow{H_{cn_1}})$ than to that of the low-frequency lexeme $(\overrightarrow{L_{cn_1}})$. The smaller angle between $\overrightarrow{CN_1}$ and $\overrightarrow{H_{cn_1}}$ corresponds to a larger correlation, and the bigger angle between $\overrightarrow{CN_1}$ and $\overrightarrow{L_{cn_1}}$ corresponds to a smaller correlation.

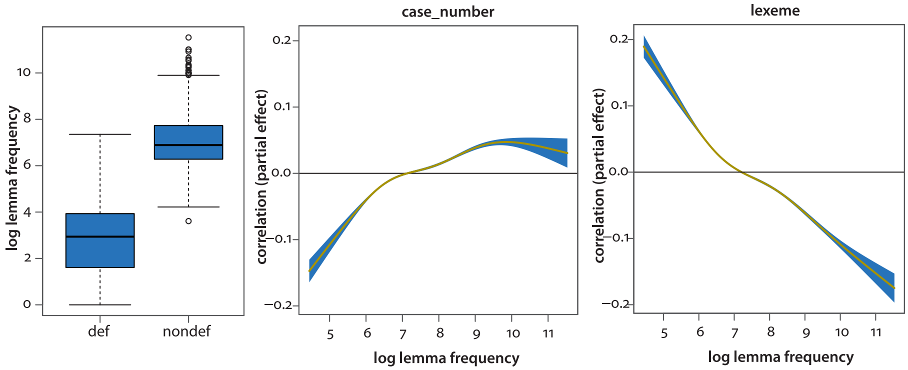

Figure 7. Left: Boxplots of log lemma frequency for defective and non-defective nouns. Middle: The partial effect of log lemma frequency on the case-number correlation measure for non-defective nouns. Right: The partial effect of log lemma frequency on the lexeme correlation measure for non-defective nouns

In the current dataset, defective nouns are mostly very low frequency words. Given the effect of frequency on our correlation measures, we would expect defectives to have lower case-number correlations but higher lexeme correlations. Defectives indeed have consistently lower case-number correlations than non-defectives (cf. Figure 5). The lexeme correlation, however, exhibits exactly the opposite trend. As can be seen in the bottom panel of Figure 5, when the defectives are matched, to the extent that this is possible for our dataset, with respect to paradigm size and frequency, the defectives (which are of lower frequency even for this subset of the data) emerge with lower correlations, rather than the higher correlations that the trend for non-defectives shown in the right panel of Figure 7 predicts. This pattern of result corresponds well with the finding of defectives being less semantically transparent reported in Section 4.1 – because the inflected forms of defective paradigms are semantically less coherent, individually they are therefore less similar to their respective aggregate lexeme meanings.

4.3 Defectiveness and predicted semantic vectors

Figure 3 revealed an interaction between case and number: the plural shift vectors cluster 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, $\vec{\varepsilon}$, representing a word form's semantic idiosyncracies (as well as measurement error in the embeddings):

$$\phi(\lambda, \kappa, \nu) = \vec{\lambda} + \vec{\kappa} + \overrightarrow{\nu|\kappa} + \vec{\varepsilon} \tag{2}$$

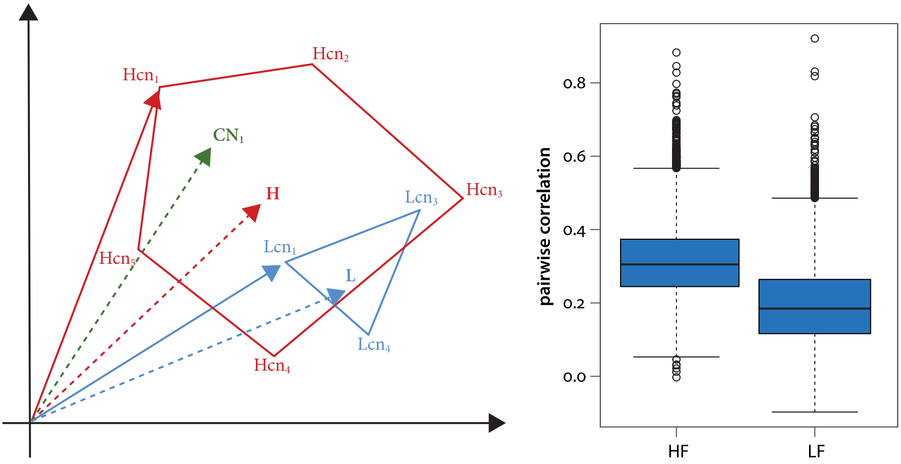

Figure 8. Left: Graphical illustration of the relative positioning of the embeddings of high and low frequency lexemes (see text for explanation). Right: Boxplots of pairwise correlations among the embeddings of dative singular forms from high and low frequency lexemes

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 ($\vec{\varepsilon}$, 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 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 9. As expected, defective forms indeed have larger error vectors (two-sample Wilcoxon tests, $W = 1148240$, $p < 0.0001$).[15] Taken together, the results suggested that the meanings of nouns from defective paradigms deviate from their theoretically predicted meaning to a larger extent as compared to non-defective forms.[16]

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, as given by Zaliznjak (1977), have inflected variants that are less transparent, that are further away from the prototypical vectors for case and number, and have semantic vectors that can be predicted less accurately.

In our dataset, the defective nouns are in general of smaller paradigm size and of lower token frequency. In our study we have taken these measures into account, and observed that the semantic differences that we observed between defective and non-defective nouns are not confounded with frequency or paradigm size. The size of a paradigm as gauged with counts of occurrences in a corpus is a measure that does not do full justice to the true complexities of paradigms. However, even though paradigm size is a crude measure, it has been found to be predictive for lexical processing (e.g., Lõo et al., 2018b, a). More distribution-sensitive measures such as inflectional entropy (del Prado Martin et al., 2004) may provide tighter control than is possible with the simple paradigm size measure. Nevertheless, the quantitative trends we have observed in this study contribute new facts about the semantics and syntax of defective nouns that we hope will lead 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 vectors representing lexemes, case, and number, we need to condition the vector of number on case. The same clustering structure can also be observed in nouns of another highly inflectional language, Finnish (Niko-

[15] Although there are more outliers for non-defective nouns, the proportions of outliers for defectives and non-defectives are not significantly different, according to a proportions test ($p = .29$).

[16] Note that for this set of analyses, 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 (see also Nikolaev et al., this volume). Both results can be found in the supplementary material.

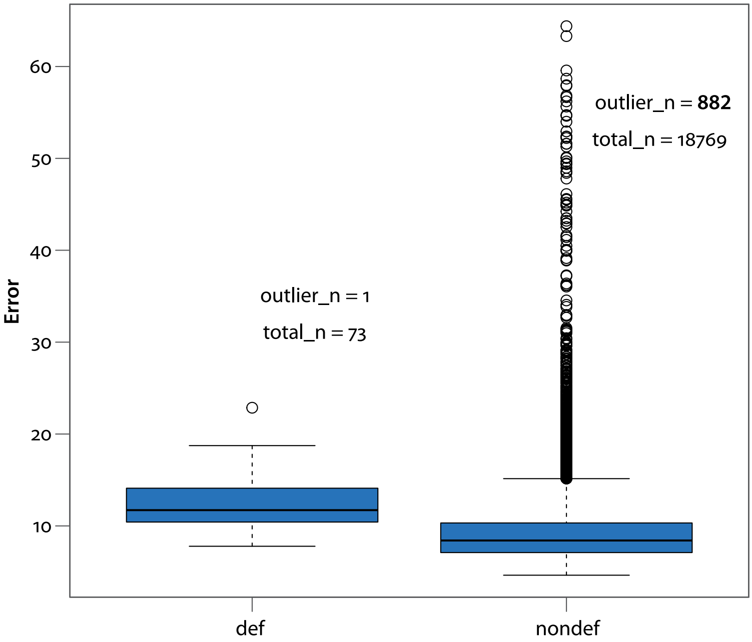

Figure 9. The degree to which the reconstructed semantic vector deviates from the empirical embedding, gauged by the L1-norm of the error vector, for defective and non-defective nouns. The numbers of words (total_n) and outliers (outlier_n) in each category are provided as well

laev et al., this volume), as well as in nouns of English (Shafaei-Bajestan et al., this volume). While for Finnish and Russian, it seems to be the case that the use of plurals is different primarily across different syntactic constructions, as indicated by case, for English, the conditioning of plurality is by semantic class. This finding of the conditional realization of number 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 embedding 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 generally, 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 and Baayen, this volume).

The differences between defective and non-defective nouns that are brought to light by our analyses may in part arise due to the reallocation of morpho-syntactic functions to other paradigm slots that is necessitated by defectiveness. In the absence of acceptable genitive plural forms, speakers have to use other forms or syntactic expressions in compensation. These alternatives will require the use of other case endings and possibly a change of number. As a consequence, the other paradigm members are overloaded with morpho-syntactic functions that are atypical for non-defective lexemes. In other words, our conjecture is that defectiveness is not just a cell in a paradigm being unavailable, but that in fact the non-availability of this cell leads to changes in use of the other paradigm members. It is these changes in distributional patterns that would then be detected by our analyses of Russian embeddings. If so, this would imply a causal link from form to meaning, in which case defectiveness unavoidably comes with subtle changes in meaning. On the other hand, we cannot rule out that defective nouns have their own specific semantics, and that it is these idiosyncratic semantics that drive defectiveness. Perhaps, from a distributional semantics perspective, positing a causal directionality seems suspiciously like a chicken and egg problem that might be resolved using diachronic data, but that resists resolution given synchronic language use.

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. Possibly, the embeddings for Russian nouns are reflecting distributional structure that goes beyond lexical semantics and the semantics of case and number. Current Russian embeddings possibly 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.

Funding

Brown and Evans thank the Arts and Humanities Research Council (grant AH/T002859/1 'Feast and Famine: Confronting Overabundance and Detectivity in Language'). Chuang and Baayen are indebted to funding from the ERC, project WIDE-742545.

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.

Becker, M. and Gouskova, M. (2016). Source-Oriented Generalizations as Grammar Inference in Russian Vowel Deletion. *Linguistic Inquiry*, *47*(3):391–425.

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.

del Prado Martin, F. M., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1):1–18.

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 Luschutzky, H. C., editors, *Competition in Inflection and Word-Formation*, pages 169–193. Springer International Publishing, Cham.

Gouskova, M. and Becker, M. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory*, *31*(3):735–765.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ilola, E. and Mustajoki, A. (1989). *Report on Russian Morphology as it appears in Zaliznyak's Grammatical Dictionary*. Helsinki University Press, Helsinki. Type: Book.

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.

Lõo, K., Järvikivi, J., and Baayen, R. H. (2018a). Whole-word frequency and inflectional paradigm size facilitate estonian case-inflected noun processing. *Cognition*, *175*:20–25.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., and Baayen, R. H. (2018b). Production of estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology*, *28*(1):71–97.

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, A. and 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.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

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.

Švedova, N.J., editor (1984). *Slovar' russkogo jazyka (S. I. Ožegov)*. Russkij jazyk, Moscow, 16th edition.

Appendix

List of defective nouns from Zaliznjak (1977):

баба-яга, балда, башка, беремя, брюзга, гомоза, дуда, егоза, зуда, казна, кайма, камка, киса, корма, кочерга, краса, кума, майя, мга, мгла, мечта, мзда, мольба, мулла, пакля, пенька, полумгла, полутьма, раба, радиоэхо, райя, сайга, сума, тамга, тахта, темя, тетива, треска, тьма, тьма-тьмущая, фата, фита, хвала, хна, хула, хурма, цевье, чадра, чалма, чека, чета, чоха, чуха, эхо, юла, яга

Address for correspondence

Yu-Ying Chuang
Seminar für Sprachwissenschaft/Quantitative Linguistics
Eberhard Karls University
Keplerstrasse 2
72074 Tübingen
Germany

yu-ying.chuang@uni-tuebingen.de
https://orcid.org/0000-0002-2733-2748

Co-author information

Dunstan Brown
Language & Linguistic Science
University of York
dunstan.brown@york.ac.uk

Roger Evans
Language & Linguistic Science
University of York
roger.evans@york.ac.uk

Harald Baayen
Seminar für Sprachwissenschaft/Quantitative Linguistics
Eberhard Karls University
harald.baayen@uni-tuebingen.de

Publication history

Date received: 30 June 2022
Date accepted: 16 February 2023
Published online: 30 May 2023