Laura A. Janda* and Francis M. Tyers

Less is more: why all paradigms are defective, and why that is a good thing

https://doi.org/10.1515/cllt-2018-0031

Abstract: Only a fraction of lexemes are encountered in all their paradigm forms in any corpus or even in the lifetime of any speaker. This raises a question as to how it is that native speakers confidently produce and comprehend word forms that they have never witnessed. We present the results of an experiment using a recurrent neural network computational learning model. In particular, we compare the model's production of unencountered forms using two types of training data: full paradigms vs. single word forms for Russian nouns, verbs, and adjectives. In the long run, the model displays better performance when exposed to the more naturalistic training on single word forms, even though the other training data is much larger as it includes full paradigms for each and every word. We discuss why "defective" paradigms may be better for human learners as well.

Keywords: morphology, paradigm, Russian, corpus, computational experiment

1 Introduction

Native speakers of languages with complex inflectional morphology routinely recognize and produce forms that they have never heard or seen (the "Paradigm Cell Filling Problem", cf. Ackerman et al. 2009). How is this possible? We take a learning perspective on this question and present evidence to show that inflectional morphology can be mastered through partially overlapping portions of paradigms in input.

Our data and experiment focus on Russian, a language with moderately complex inflectional morphology for all open-class word classes. In order to orient readers to the behavior of paradigms, we begin with an example from Spanish. We then turn to definition of terms and our theoretical perspective. Section 2 situates Russian with respect to the attestation of word forms in

***Corresponding author: Laura A. Janda,** HSL, UiT Norges arktiske universitet, Tromso, Norway, E-mail: laura.janda@uit.no
Francis M. Tyers, School of Linguistics, Nacional'nyj issledovatel'skij universitet Vyssaa skola ekonomiki, Moskva, Russia, E-mail: ftyers@hse.ru

corpora of languages that vary according to the size of their paradigms. Grammatical profiles of Russian nouns are the topic of Section 3, followed by our experiment on the learning of Russian inflected forms in Section 4.

Table 1 visualizes the attestation of two Spanish verbs, CONTAR 'TELL' (207 attestations) and GUSTAR 'PLEASE' (53 attestations) in the UD Spanish corpus.[1] A grammatical profile (Janda and Lyashevksaya 2011) is the relative frequency distribution of the inflected forms of a lexeme, visualized here by **bold face** to signal very frequent forms (attested over 30 times), plain text for robustly attested (over 10 times) forms, grey text for rarely attested (less than 10 times), and blanks for forms unattested in the corpus. For CONTAR 'TELL', one form, the Third Person Singular Present *cuenta*, is much more frequent than the others, 4 forms are found quite often, 7 forms are found only rarely, and the remaining 18 forms are not attested. The comparison with GUSTAR 'PLEASE' shows that the grammatical profiles of lexemes vary: while Conditional forms are missing for CONTAR 'TELL' in the corpus, the second most frequent form of GUSTAR 'PLEASE' is the Third Person Conditional *gustaría*. But native speakers can produce the "missing" forms even if they have never encountered them, and they can do this also for new and nonce verbs. We expect that if we ask a speaker of Spanish to inflect a verb **trontar*, they should be able to produce the forms.

Table 1 demonstrates the extent to which the paradigm is an artificial construct. Rather than a system in which every lexeme populates the entire space of a full paradigm, each lexeme populates only a portion of that space. Because many of those partial representations of a paradigm overlap, it is possible for native speakers to produce any potential form.

We provide three types of evidence that the inflectional morphology of Russian is based on networks of partial sets of inflected word forms. These partial sets exhibit prototype-periphery effects, differ from lexeme to lexeme, yet overlap enough to make it possible to produce unencountered forms both of known and of newly encountered lexemes. Our evidence comes from: (a) comparison of the percentages of full paradigms attested in corpora of languages with a range of paradigm sizes, showing that attestation of all forms in a paradigm is rare (Section 2); (b) demonstration of the distribution of partial sets of word forms for high-frequency Russian nouns, showing that different nouns have different

1 The UD Spanish corpus (400,000 tokens) is available at: https://github.com/UniversalDependencies/UD_Spanish. This is a "gold standard" (manually disambiguated) corpus, which makes it possible to differentiate among morphologically ambiguous forms (such as the First Person Plural Present and Preterite forms, which are homophonous in Spanish). Larger corpora do not disambiguate homophonous forms and thus contain too much noise to present an accurate grammatical profile, cf. Section 2.

Table 1: Indicative forms of Spanish CONTAR 'TELL' and GUSTAR 'PLEASE' as attested in the UD Spanish corpus. **Boldface** indicates word forms attested over 30 times, plain text indicates word forms attested over 10 times, grey indicates word forms attested fewer than 10 times, blank cells indicate unattested word forms.

	Present		**Preterite**		**Imperfect**		**Conditional**		**Future**	
	CONTAR 'TELL'	**GUSTAR 'PLEASE'**	**CONTAR 'TELL'**	**GUSTAR 'PLEASE'**	CONTAR 'TELL'	GUSTAR 'PLEASE'	**CONTAR 'TELL'**	**GUSTAR 'PLEASE'**	**CONTAR 'TELL'**	**GUSTAR 'PLEASE'**
1sg	cuento		conté						contaré	
2sg										
3sg	**cuenta**	**gusta**	contó	gustó	contaba	gustaba		gustaría		
1pl	contamos		contamos		contábamos					
2pl										
3pl	cuentan	gustan		gustaron	contaban				contarán	

distributions, none have equal frequency across the paradigm, and the partial sets overlap (Section 3); and (c) a computational experiment showing that learning is enhanced by exposure to individual word forms as opposed to full paradigms (Section 4). Collectively, these three types of evidence suggest that all paradigms are defective (meaning that some forms are unattested or extremely rare) to a greater or lesser extent, since all lexemes have some word forms that are attested rarely or not at all, and that inflectional morphology should be modelled in terms of overlapping partial sets of word forms.

1.1 Definitions and theoretical perspectives

Before turning to our evidence, we offer some definitions and situate our investigation in terms of theoretical premises, since with respect to morphology, both the definition of terms and theory vary widely across scholars. We connect these terms to relevant concepts in Cognitive Linguistics.

Word form: We take the word form as the basic unit of morphology (cf. Blevins 2016: 64). A word form is a morphological construction (Booij 2017), and its acknowledgment as the basic level of analysis is in keeping with the cognitivist assertion that the construction is the basic level of linguistic analysis (Diessel 2015; Goldberg 2006). Word forms are inflected forms such as the forms of Spanish CONTAR 'TELL' in Table 1 and the forms of Russian SLOVO 'WORD', presented here in both transliteration (enhanced with stress marks) and phonetic transcription: *slóvo* [slóvə], *slóva* [slóvə], *slóvu* [slóvu], *slóvom* [slóvəm], *slóve* [slóvʲi], *slová* [slʌvá], *slóv* [slóf], *slovám* [slʌvám], *slovámi* [slʌvámʲi], *slováx* [slʌváx]. Transcription reveals that the vowel in the first syllable varies according to stress, and that the following consonant is variously realized as [v], [vʲ], or [f] in these word forms. Although we use transliteration in most places in this article, the truest representation of word forms is their phonological forms (which for Russian can be straightforwardly derived from the transliteration if stress is known).

Lexeme: We take a lexeme to be an abstraction that unifies a set of inflectionally-related word forms (cf. Cruse 1986: Chapter 3). Russian SLOVO 'WORD' is an abstraction over the set of word forms stated above, and Spanish CONTAR 'TELL' is an abstraction over the word forms in Table 1. We state lexemes in small caps in order to distinguish them from word forms. A lexeme can be an abstraction over a partial set, in the case that only one or a few forms are attested. We do not claim that the lexeme necessarily has any psychological reality. In terms of Cognitive Linguistics, a lexeme is a schema, or more precisely, a constructional schema as defined by Langacker (2008: 167).

Lemma: A lemma is the citation word form of a lexeme. The Nominative Singular *slóvo* 'word' is a lemma, as is the Infinitive *contar* 'tell'. The question of whether the lemma has any psychological reality is beyond the scope of this article.

Paradigm: A paradigm is the set of word forms associated with a lexeme and the marking of morphosyntactic features. A full paradigm exhausts all possible morphosyntactic features associated with the given word class and there are typically implicational relationships that hold among the word forms (cf. Wurzel 1984: 116–124 and Wurzel 1989: 112–121 and Bybee 1985: 50–58). From the perspective of Cognitive Linguistics, these relationships form radial category networks with prototypical and peripheral members (Nesset and Janda 2010). For Russian nouns, for example, the full paradigm is normally defined by case and number as in Table 2, where each combination of word forms and case/number features defines a "cell".

Table 2: Full paradigm of Russian SLOVO 'WORD'[2].

	Singular	**Plural**
Nominative	*slóvo*	*slová*
Genitive	*slóva*	*slóv*
Dative	*slóvu*	*slovám*
Accusative	*slóvo*	*slová*
Instrumental	*slóvom*	*slovámi*
Locative	*slóve*	*slovách*

Inflection and Derivation: Our focus is on inflection, which we define as the morphosyntactic marking of a lexeme that serves as the organizational basis for paradigms, including those that show suppletion. Derivation, by contrast, is the extension of a root to a new lexeme, as in the derivation from SLÓVO 'WORD' of words like *slovár'* 'dictionary', *slovésnyj* 'verbal', *blagoslovit'* 'bless', and *slovoobrazovánie* 'word-formation'. However, we recognize no crisp boundary between inflection and derivation since both deploy the same resources and there are hybrid phenomena such as participles, which can be included in the paradigm of a verb or considered deverbal adjectives (cf. detailed arguments against a firm distinction between inflection and derivation in Bybee 1985;

2 In addition to these twelve word forms, a subset of Russian masculine nouns can have additional peripheral case/number forms: an alternate "second" Genitive and/or an alternate "second" Locative case.

Spencer 2016; and Booij 2017: 243 acknowledgement that constructional schemas are relevant for both derivation and inflection).

Defectiveness: We take a broad view of defectiveness, including any situation in which a word form (representing a specific combination of morphosyntactic features) of a lexeme is rare or unattested. This definition is in keeping with the cognitivist observation that language phenomena tend to be scalar rather than categorical. We postulate a continuum between equiprobability of word forms, which would be found if all possible word forms of a lexeme were attested in equal numbers, and the extreme defectiveness found in inflectional paradigmatic gaps. There may be some characteristics of paradigmatic gaps that make them special (see Albright 2003; Sims 2006; for discussion of the influence of variation in forms and inferences from paradigm structure, and Baerman 2011; for the role of homophony), however speakers can usually fill paradigmatic gaps both when asked to do so in experiments (cf. Sims 2006; Pertsova and Kuznetsova 2015) and spontaneously (as evidenced in corpora and internet citations).

Of major concern are the complexity of paradigms and how it is possible for speakers to produce word forms that they have never encountered (Ackerman et al. 2009). The complexity of paradigms can be measured by means of conditional entropy (Ackerman and Malouf 2016; Blevins 2016: Chapter 7), a numerical measure of how unexpected a word form is given one or more other cells in the paradigm. The average conditional entropy of any language is typically fairly low (Ackerman and Malouf 2016). From the perspective of Russian, if I know that there is a Nominative Plural form *slová* 'words', how many word forms might be possible candidates for the Nominative Singular? Assuming a perfect mastery of Russian morphological patterns, the answer is three: *slóvo* (assuming a neuter noun with shifting stress), **slovó* (assuming a neuter noun with fixed end stress like *veščestvó* 'substance'), and **slóv* (assuming a masculine noun like *dóm* 'house', which has a Nominative Plural *domá*). So in this case there is a one in three chance of correctly predicting the Nominative Singular from the Nominative Plural. And for many other predictions (like predicting any of the other Singular forms from the Nominative Singular *slóvo*), there is only one possible candidate.

In predicting the Nominative Singular given the Nominative Plural *slová*, the correct answer also selects the most likely option, since neuter nouns with shifting stress are more common than both neuter nouns with fixed end stress and masculine nouns with the stressed *-á* Nominative Plural ending. However, for the lexeme SLOVO 'WORD', by far the most frequent word form (34.4% in the SynTagRus corpus described below in Section 2) is actually the Dative Plural *slovám*, which figures in the common construction *po slovám* + X-Genitive 'according to what X says', so in this instance it would make most sense to

make predictions from that form, which is somewhat less predictive, since it leaves open the possibility that this could be a feminine noun (since nouns of all three genders have Dative Plural forms in *-am*). The next most frequent word forms of SLOVO 'WORD' are the Genitive Singular (11.3%) and Nominative Singular (10.07%), and the remainder are infrequent.

Recognizing and producing word forms is an essential skill that language learners must master. Language pedagogy has traditionally relied on presentation of full paradigms, and most computational experiments modelling the learning of inflectional morphology use full paradigms for training (but note a recent pioneering work that departs from this tradition: Malouf 2017).

We show that all Russian paradigms are defective to a greater or lesser degree and that defectiveness is strategic, providing enough cues and overlap to make it possible to learn the implicational relationships between word forms without swamping the learner with word forms that they are unlikely to ever see, hear or need to produce.

2 The relationship between attestations of full paradigms and paradigm size

Zipf's (1949) law observes that the frequency of any word is inversely proportional to its frequency rank (a power law). This means that there are a few words of high frequency, then the curve declines sharply, ending with a long tail of hapaxes (words that appear only once), and hapaxes typically account for around 50% of unique lexemes in a corpus.[3] Zipf's law also applies to word forms, and as a result, the number of lexemes that appear in all their forms (their full paradigm) is small, and this number quickly drops toward zero as the size of the paradigm expands.

Table 3 reports data from several languages that differ according to the size of their noun paradigm. Only data from "gold standard" (manually annotated) corpora can be used for this purpose, since the noise in data from larger (automatically annotated) corpora is so great as to make it impossible to accurately determine what paradigm forms are attested.[4] Both the total number of

3 Cf. Baayen (1992, 1993) on the frequency of hapaxes. Kuznetsova (2017: 96) shows that for texts in the modern subcorpus of the Russian National Corpus (110 million words) "more than half of the nominal lexemes that appear in a text appear in only one word form".

4 Gold-standard corpora are essential for this comparison, which relies on fully disambiguated data. Morphological ambiguity is a long-standing and still largely intractable problem for corpus linguistics, because larger automatically tagged corpora cannot disambiguate

unique noun lexemes and the number of noun lexemes that appear in all forms in the full noun paradigm for each language has been tallied up, and the latter divided by the former to arrive at the percentage of lexemes that appear in the full set of paradigm forms.

Table 3 puts the position of Russian in terms of the size of its noun paradigm and the proportion of noun lexemes attested in all word forms into perspective.

Table 3: Relationship between paradigm size and number of full paradigms for nouns.

Language & corpus name	Corpus size	Paradigm size	Total lexemes	Lexemes with full paradigm	% Lexemes with full paradigm
English Web Treebank	254,830	2	6,369	1,524	23.92%
Norwegian Dependency Treebank	311,277	4	12,587	393	3.12%
Russian SynTagRus	**1,069,561**	**12**	**21,945**	**13**	**0.06%**
Czech Prague Dependency Treebank	1,509,242	14	17,904	3	0.02%
Estonian ArborEst	234,351	28	14,075	0	0%

English has the simplest morphological system with two word forms (singular and plural as in *window, windows*) for nouns, but only about 24% of nouns appear in both forms in a corpus. Norwegian has both definiteness and number, yielding four forms[5] (singular indefinite *vindu* 'window', singular definite *vinduet* 'the window', plural indefinite *vinduer* 'windows', and plural definite *vinduene* 'the windows'). In Norwegian, the proportion of nouns that we encounter in their full paradigm of four forms is 3%. Russian has six grammatical cases in singular and plural, yielding 12 word forms, and for some nouns there are as

homophonous forms. In Russian, fully 45% of words in running text are morphologically ambiguous. For example, the Russian form *stali*, can be the Past tense Plural form of the verb *stat'* 'become', or any of five forms (Genitive, Dative, and Locative Singular, or Nominative and Accusative Plural) of the noun *stal'* 'steel'.

5 In both English and Norwegian, some have argued that the Genitive *-'s/s* is an inflectional ending and that would then double the size of the paradigms in those two languages. However, in both languages this interpretation is dubious because the Genitive *-'s/s* behaves like a phrasal clitic, as in *The King of Denmark's problems/Kongen av Danmarks problemer*, where *-'s/s* is not attached directly to the noun *King/Kongen*, but to the end of the phrase. Cf. Payne and Huddleston (2002) for further discussion.

many as 14 word forms due to marginal cases (the second Genitive and second Locative). But less than 1% of Russian lexemes appear in 12 or more word forms. Czech has seven cases and two numbers for all noun paradigms, so a total of 14 word forms, and even fewer lexemes appear in all word forms in a corpus (cf. similar results reported for Czech in Malouf 2017). The Estonian paradigm is twice as large as the Czech one, and here the number of noun lexemes that appear in all word forms drops to 0% (in a vastly larger corpus a few noun lexemes might be attested in all word forms, but still the number will be very close to zero). We can take this comparison even further to languages with truly large noun paradigms. North Saami has 130 cells in its noun paradigm (Nickel and Sammallahti 2011), but a manual analysis of over 0.66M words (cf. Janda and Antonsen 2016) reveals not only that no noun lexeme is attested in all its word forms: in addition, 36 of the word forms are never attested at all for any lexeme, and nine more are attested only once. North Saami noun paradigms pale in comparison with the paradigms of some languages that linguists claim to have thousands or even millions of forms (cf. the claim that the Archi language has over 1.5 million verb forms: Kibrik 2001; Corbett 2015). However, claims of truly enormous paradigms have to be considered with caution since most involve multiplication via various combinations of grammatical markers that are both semantically transparent and compositional, as in agglutinative languages (cf. Comrie and Polinsky 1998).

Obviously in a larger corpus, a larger number of words would appear in all paradigm forms, but the percentage of fully-attested paradigms would not increase because those would be overwhelmed by the vastly larger number of additional hapaxes and lexemes attested in only a handful of forms. Since Zipf's law scales up,[6] one could hypothesize that a speaker's total exposure to her/his native language is like a very large corpus with the same properties. This means that 76% of English noun lexemes and 97% of Norwegian noun lexemes will never be encountered in their full paradigms by native speakers of those languages. Native speakers of Russian and Czech will be exposed to full paradigms for fewer than 1% of their noun lexemes. An Estonian speaker will encounter all the word forms of a noun lexeme only very rarely, if at all. And a native speaker of North Saami will probably never come across any examples for some of the word forms in the noun paradigm of that language, much less all forms of any single lexeme (cf. similar observations in Malouf 2016).

6 Cf. Manning and Schütze (1999). Moreno-Sánchez et al. (2016) conducted a large-scale test of Zipf's Law on English texts, and while they report some irregularities, they also find that the pure power-law form of Zipf's Law holds up well.

The vast majority of lexemes in a language with complex inflectional morphology are normally encountered only in some of their word forms. This does not mean that the word forms that are unattested in the corpora in Tables 1 and 3 do not exist (cf. Piperski's 2015 argument that lack of attestation in a corpus cannot be taken to imply non-existence). In a larger or different corpus, some of these word forms will be encountered. However, the majority of word forms missing from a given corpus will be missing or very rare even in another or larger corpus. And, because a different or larger corpus will also have proportionally just as many hapaxes and low-frequency lexemes, the percentage of lexemes that will be attested in only a portion of their paradigms will remain approximately the same. It is necessary to scale up only by two orders of magnitude in order to approximate the input available to L1 learners, who are probably exposed to between 5 and 10 million words per year.[7]

These observations concerning the skewed distribution of attested word forms is consistent with Sinclair's (1991: 109–115) "idiom principle", according to which we should not expect cells of a paradigm to be evenly attested. Whereas the "open-choice" principle, allowing virtually any word or word form to occur in slots, is applied in the guidelines of grammars, in authentic text the majority of slots are filled according to the idiom principle, meaning that there is only one or a very limited number of available choices, and these include choices about grammatical categories such as those that define paradigms.

3 Overlapping partial paradigms and their distribution for Russian nouns

As the data in Table 3 show, only a fraction of a percent of Russian noun lexemes appear in all the word forms of their paradigm, and this proportion will not change substantially no matter how large the sample is. This suggests that nearly all noun lexemes occur only in some subset of potential word forms. In this section, we examine what this means in more detail. Our aim is to show that different noun lexemes are associated with different sets of word forms, in aggregate creating a lexicon containing networks of word forms, which overlap to varying degrees in terms of the case and number values they express.

Linguists have long recognized that some lexemes have "defective paradigms" either due to a restriction on number yielding singularia tantum like

7 This estimate is based on Hart and Risley's (2003) longitudinal study of L1 learners of English.

BEDNOST' 'POVERTY' and pluralia tantum like NOŽNICY 'SCISSORS', or due to a more specific restriction on a single word form, often called a "paradigm gap", as we see in words like MEČTA 'DREAM' that lack a Genitive Plural form. From the perspective we offer in this article, virtually all Russian nouns have "defective paradigms" to some extent because only a few word forms are normally associated with any given lexeme. Or, to put it differently, "defectiveness" is the norm and is a matter of degree, with lexemes that show absolute restrictions merely at one extreme end of the spectrum. Even the lexemes at the other end of the spectrum, namely those few noun lexemes that really do occur in all possible case and number word forms, do not represent all of those word forms equally, since some word forms are much more common than others. Furthermore, the supposed restrictions are not always absolute. Websites dedicated to eradicating grammatical errors indicate that Russians often fail to observe tantum noun restrictions, and examples of Genitive Plural forms of lexemes that supposedly lack such word forms are not hard to come by.[8]

Each lexeme has its own signature grammatical profile: the relative frequency distribution of word forms that are associated with it. A grammatical profile typically points to one word form that is most frequent (most prototypical for that lexeme) and a few that are not uncommon, while most possible word forms are very infrequent or unattested. From the perspective of a usage-based approach, a grammatical profile provides an approximation of the prototypicality of the word forms of a lexeme.[9] Sections 3.1 and 3.2 show what this means in terms of concrete lexemes and their grammatical profiles.[10]

3.1 Grammatical profiles in tables

As stated in Section 1, a grammatical profile is the relative frequency distribution of the word forms of a lexeme as attested in a corpus. We demonstrate the grammatical profiles of Russian noun lexemes based on data from SynTagRus, a

8 For example, gramota.ru lists examples of plural forms of singularia tantum nouns such as *podderžka* 'approval' (http://www.gramota.ru/biblio/research/rubric_370/rubric_388/), and the Russian National Corpus (ruscorpora.ru) lists 24 examples of *mečt*, 22 of which are Genitive Plural forms of *mečta* 'dream', despite claims of a paradigmatic gap for that cell in grammars and dictionaries.

9 Relative frequency is not a direct measure of prototypicality, but the two often coincide. We use relative frequency as a proxy for prototypicality.

10 The data and the statistical code for our analyses are publicly archived at https://doi.org/10.18710/VDWPZS.

deeply annotated (preprocessed and then manually corrected) corpus of 1,069,561 tokens, which is relatively error-free in terms of morphological tagging. Because we wish to examine the relative frequency of word forms, we restrict our sample to high-frequency lexemes, in this case with a frequency of 50 or more in SynTagRus.[11] This is important for at least two reasons. The first reason is the large number of hapaxes mentioned in Section 2: if we do not set a frequency threshold, half of our lexemes will be hapaxes that appear in only one word form. The second reason is that even after we eliminate the hapaxes, there are many nouns that appear in only a handful of forms and here we still have too small a sample to say anything reliable about a frequency distribution. If we have three attestations of a lexeme and all of them happen to be Genitive Plural word forms, does that really mean that this lexeme appears only in the Genitive Plural, or is this just a fluke due to the fact that we have so few datapoints for this lexeme? The inclusion of only high-frequency lexemes skews the view of the phenomenon that we are examining, and this needs to be kept in mind. By excluding hapaxes and other low-frequency lexemes, we are removing from this dataset the lexemes that show the least amount of overlap in the attestation of word forms. In the high-frequency data, overlap of partial sets of word forms is maximized. However, it is also the high-frequency lexemes, and in particular their most high-frequency word forms, that are most salient from a usage-based perspective.

We sample all the lexemes with a frequency ≥50 in SynTagRus that represent five paradigm types: masculine inanimate (312 lexemes), masculine animate (95 lexemes), neuter inanimate (238 lexemes), feminine inanimate II (ending in *-a/-ja*, 261 lexemes), and feminine inanimate III (ending in -', 75 lexemes). This grouping gives us a fairly large set of lexemes (982) that are relatively evenly divided across types.

Tables 4 and 5 give a visual presentation of the grammatical profiles of sample nouns. The sample in 4 is of nouns with exactly (or nearly exactly) the same frequency, whereas Table 4 presents a sample of nouns that are strongly attracted to case/number combinations that are relatively unusual for each type. The purpose of Tables 4 and 5 is to give the reader a sense of the kinds of similarities and differences encountered across lexemes of the five types.

Table 4 displays examples of lexemes from each group with a total raw frequency of 100 (or 97 in the case of PAMJAT' 'MEMORY'). The rows in Table 4 show the case/number combinations in the Russian noun paradigm. Table 4 visualizes the grammatical profiles by giving the most frequent word forms (over 20% of grammatical profile) in boldface, word forms of moderate frequency

11 This threshold was selected because it yielded a relatively large number of nouns from the SynTagRus corpus, although in principle another threshold could have been chosen.

Table 4: Visualization of grammatical profiles of high-frequency noun lexemes (100 per million words) representing five declension classes in Russian showing good coverage of paradigms. N = Nominative, G = Genitive, D = Dative, A = Accusative, I = Instrumental, L = Locative, sg = Singular, pl = Plural. **Boldface** indicates word forms that account for over 20% of the lexeme's grammatical profile, plain text indicates word forms that account for between 10% and 20%, grey indicates word forms that account for under 10%, blank cells indicate unattested word forms.

	Masculine inanimate	**Masculine animate**	**Neuter inanimate**	**Feminine inanimate II**	**Feminine inanimate III**
	'FEAR'	**'SOLDIER'**	**'DEPARTMENT'**	**'CONCEPT'**	**'MEMORY'**
Nsg	**strax**	**soldat**	otdelenie	**koncepcija**	pamjat'
Gsg	**straxa**	soldata	**otdelenija**	**koncepcii**	**pamjati**
Dsg	straxu	soldatu	otdeleniju	koncepcii	pamjati
Asg	strax	soldata	otdelenie	koncepciju	pamjat'
Isg	straxom	soldatom	otdeleniem	koncepciej	pamjat'ju
Lsg	straxe		**otdelenii**	koncepcii	pamjati
Npl	straxi	**soldaty**	otdelenija		
Gpl	straxov	**soldat**	otdelenij	koncepcij	
Dpl		soldatam			
Apl	straxi	soldat	otdelenija	koncepcii	
Ipl	straxami		otdelenijami	koncepcijami	
Lpl	straxax	soldatax	otdelenijax		

Table 5: Visualization of grammatical profiles of high-frequency Russian noun lexemes showing different coverage of paradigms.

	Masculine inanimate	**Masculine animate**	**Neuter inanimate**	**Feminine inanimate II**	**Feminine inanimate III**
	'BACKGROUND'	**'CHAMPION'**	**'EXTENT'**	**'FRAME'**	**'DIFFICULTY'**
Nsg	fon	čempion			trudnost'
Gsg	fona	**čempiona**			trudnosti
Dsg		čempionu			trudnosti
Asg		čempiona			trudnost'
Isg		**čempionom**			trudnost'ju
Lsg	**fone**		**protjaženii**		
Npl		čempiony		ramki	**trudnosti**
Gpl		čempionov		ramok	**trudnostej**
Dpl		čempionam			
Apl		čempionov		ramki	**trudnosti**
Ipl		čempionami		ramkami	trudnostjami
Lpl				**ramkax**	trudnostjax

(over 10%) in plain text, rare forms in grey text, and blanks for forms unattested in SynTagRus.

For example, STRAX 'FEAR', occurs most often (34%) in the Genitive Singular, followed by the Nominative Singular (24%). This lexeme is less common in the Accusative Singular (13%) and Instrumental Singular (10%), and occurs only rarely (1–5%) in the Genitive Plural, Locative Singular, Nominative Plural, Dative Singular, Accusative Plural, Instrumental Plural and Locative Plural forms (listed in order of decreasing frequency). STRAX 'FEAR' is not attested in the Dative Plural. Table 5 gives other representatives of the same five groups of noun lexemes, showing that the grammatical profiles of individual high-frequency lexemes can be very different and even nonoverlapping.

As we see in Tables 4 and 5, it is typical even for high-frequency lexemes to appear predominantly in three or fewer word forms, and to be rare or unattested in the rest. All of the nouns in Tables 4 and 5 are unattested in at least one case/number form, and some lexemes are unattested in most word forms. There is furthermore no single case/number word form that is attested for all 10 nouns in Tables 4 and 5, and in some instances the rate of "missing forms" is quite high. For example, over 50% of feminine II lexemes are unattested in the Dative Plural form.

Some lexemes have strong preferences for a single word form. The most extreme is PROTJAŽENIE 'EXTENT', attested 69 times in SynTagRus, every time (100%) in the Locative Singular in the construction *na protjaženii* + Genitive 'in the course of'. The next three strongest preferences involve days of the week that occur almost exclusively in the Accusative Singular due to the high-frequency construction *v* + Accusative, as in *v ponedel'nik* 'on Monday': PONEDEL'NIK 'MONDAY' (94.40%), VOSKRESEN'E 'SUNDAY' (94.17%), and PJATNICA 'FRIDAY' (91.49%).

Because Zipf's Law scales up, the grammatical profiles of lexemes like those visualized in Tables 4 and 5 will not change substantially, no matter how big our sample size is. And we must keep in mind that we are looking only at the highest frequency lexemes here – if we took all lexemes, we would find that the majority lack attestations of most case/number word forms.

In order to better grasp the grammatical profiles of Russian nouns it is helpful to visualize them in terms of graphs. Graphs make it possible to see how in aggregate a collection of nouns can populate the "space" of case/number combinations, even though each noun covers only a portion of that space.

3.2 Grammatical profiles in a graph

We use the technique of correspondence analysis to depict how partial sets of word forms overlap. Correspondence analysis of grammatical profiles makes it

possible to map the mathematical distances between lexemes based on the partial sets of word forms attested and their relative frequency. In a correspondence analysis plot, lexemes that are close to each other have similar, highly overlapping sets of attested word forms, while lexemes that are far apart on a plot have dissimilar sets of attested word forms with little or no overlap. Correspondence analysis also plots the relationships among the case/number values for nouns.

We illustrate with the data on the 95 masculine animate lexemes that are attested fifty or more times in SynTagRus. Table 4 visualizes the grammatical profile of one of these lexemes: SOLDAT 'SOLDIER'. This grammatical profile is a row of numbers (a vector), listing the relative frequency distribution of this lexeme across all possible case/number word forms. The entire dataset for masculine animate lexemes is a matrix of 95 such rows, with each row representing a single lexeme, and each column one of the case/number word forms. Thus we have a matrix that is 95 (row vectors) × 12 (column vectors). The task of correspondence analysis is to measure the mathematical distances between the row vectors and the column vectors, showing which of them are closest together (most similar), which are farthest apart (most different), and where all the others fit in. This is done by calculating a multidimensional space defined by "Factors" that are mathematical constructs. Factor 1 is the mathematical dimension that accounts for the largest amount of variance in the data, followed by Factor 2, etc. We can then obtain a plot of the two most important dimensions, showing where the items associated with the rows and the items associated with the columns land along those two dimensions. Since our rows are lexemes and our columns are word forms and our data show the grammatical profiles (relative frequency distributions) of the lexemes, the plot will show the positions of the 95 nouns relative to each other and to the case/number word forms, based on their grammatical profiles. Figure 1 displays the plot for the masculine animate lexemes.

Figure 1 displays the relative positions of both the row vectors – the lexemes printed in black – and the column vectors – the case/number values printed in red. The legends show that Factor 1 is plotted on the x-axis and that it accounts for 53.3% of the variation in the data (a very strong factor), while Factor 2 is plotted on the y-axis and accounts for only 9.7% of the variation in the data. Thus collectively these two Factors account for 63% of the variation in the masculine inanimate data, while the 37% remaining variation is accounted for by successively weaker Factors (all weaker than Factor 2) that are not depicted. Together Factors 1 and 2 divide the data into four groups, arranged as quadrants, with the top right quadrant having positive values for both Factor 1 and Factor 2, the bottom right with positive values for Factor 1 but negative for

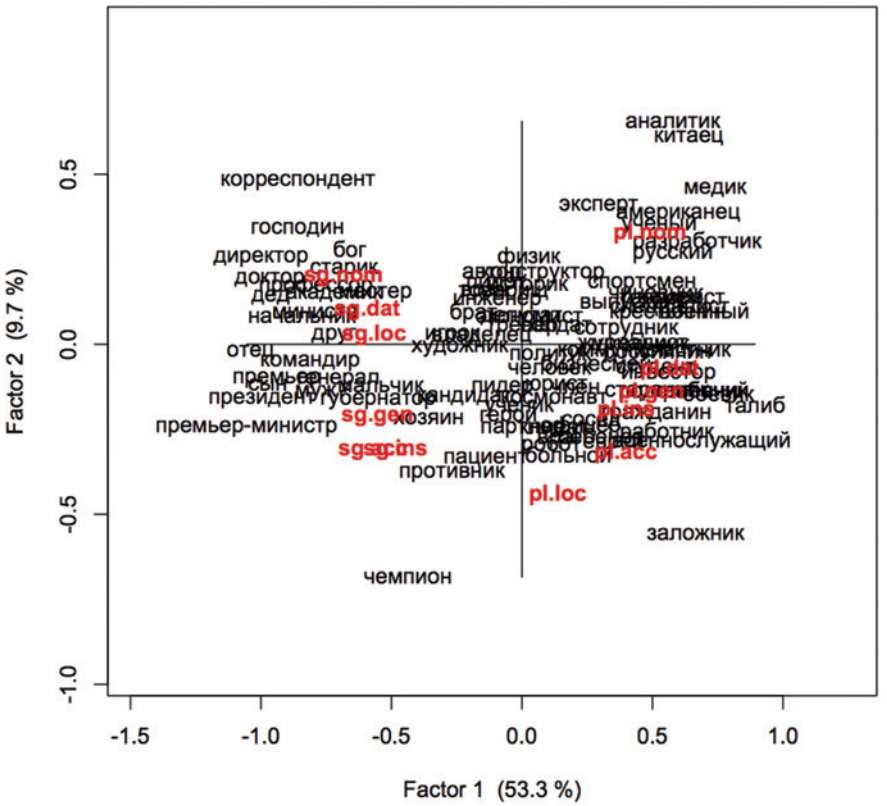

Figure 1: Correspondence analysis for masculine animate lexemes.

Factor 2, etc. Some of the lexemes are very close to each other, which can make them hard to read. SOLDAT 'SOLDIER', for example, has a Factor 1 value of 0.131 and a Factor 2 value of 0.049, which places it very near the origin (where both Factors = 0, shown by crosshairs in the graph) in the upper right quadrant, but it is hard to see because there are other nouns such as TRENER 'TRAINER' and ÈKONOMIST 'ECONOMIST' nearby.

In this plot Factor 1 can be interpreted as Number, with negative values assigned to lexemes that are more attracted to singular forms, and positive values assigned to lexemes more attracted to plural forms. Factor 2 is associated with case, for Singular separating the Nominative, Dative, and Locative from the Genitive, Accusative, and Instrumental, and for plural separating the Nominative from all other cases.

Table 6: Grammatical profiles of the four lexemes in the extreme corners of Figure 1. **Boldface** indicates word forms that account for over 20% of the lexeme's grammatical profile, plain text indicates word forms that account for between 10% and 20%, grey indicates word forms that account for under 10%, blank cells indicate unattested word forms.

	'ANALYST'	'HOSTAGE'	'CHAMPION'	'CORRESPONDENT'
Nsg	analitik	založnik	čempion	**korrespondent**
Gsg	analitika		**čempiona**	korrespondenta
Dsg			čempionu	korrespondentu
Asg		založnika	čempiona	korrespondenta
Isg		založnikom	**čempionom**	korrespondentom
Lsg				
Npl	**analitiki**	založniki	čempiony	korrespondenty
Gpl	analitikov	**založnikov**	čempionov	korrespondentov
Dpl	analitikam	založnikam	čempionam	korrespondentam
Apl	analitikov	založnikov	čempionov	korrespondentov
Ipl	analitikami	založnikami	čempionami	korrespondentami
Lpl		založnikax		

Figure 1 happens to have one lexeme that is most extreme in each of the quadrants, and the complete grammatical profiles of those four lexemes are presented in Table 6, in clockwise order.

ANALITIK 'ANALYST' is in the top right corner of the quadrant where Nominative Plural dominates. This lexeme has 59 attestations in our dataset, 34 of which (57.63%) are Nominative Plural forms, which is the highest percentage of Nominative Plural for any lexeme of this type. ANALITIK 'ANALYST' is mostly averse to the Singular, with only a few attestations for Nominative Singular (6) and Genitive Singular (3) and none for any other Singular forms. ZALOŽNIK 'HOSTAGE', by contrast, is found most in the Genitive Plural (34 attestations, 50.75% of total), and this lexeme also avoids the Singular. The portion of Accusative Plural (16.42%) is higher for this lexeme than for any other in this group. ČEMPION 'CHAMPION' is distinguished from other masculine animate lexemes by its large share of Instrumental Singular (25.68%), which exceeds that of any other lexemes of this type. For KORRESPONDENT 'CORRESPONDENT', the numbers for both the Nominative Singular (54.12%) and Dative Singular (17.65%) are very high, though neither are the highest for this type. The highest proportion of Nominative Singular is found with DIREKTOR 'DIRECTOR' at 61.87%, and the highest proportion of Dative Singular is found with BOG 'GOD' at 21.54%.

The differences in the grammatical profiles of the four lexemes in the corners of Figure 1 are motivated by the grammatical constructions that they

typically occur in. ANALITIK 'ANALYST' is often found in the construction *analitiki otmečajut, čto* 'analysts point out that' where the Nominative Plural word form fills the role of the subject. ZALOŽNIK 'HOSTAGE' appears most often as the Genitive plural complement of *zaxvat* 'seizure', *spasenie* 'rescue', and *rasstrel* 'execution'. When verbs *stat'* 'become' and *byt'* 'be' are used depictively to describe temporary states, they govern the Instrumental case, which is a typical context for ČEMPION 'CHAMPION'. The lexeme KORRESPONDENT 'CORRESPONDENT' is strongly associated with two constructions, one that identifies the correspondent with respect to a news outlet named in the Genitive, as in *korrespondent Izvestij* 'a correspondent for Izvestija [a Russian newspaper]' and another that involves verbs of communication, with the correspondent as the recipient of the message, as in *skazat'/soobščit' korrespondentu* 'tell/inform the correspondent'.

The four lexemes in Table 6 give us some perspective on the partial overlap in sets of word forms. With regard to their grammatical profiles, each of these nouns has a different center of gravity, represented in boldface in Table 6. KORRESPONDENT 'CORRESPONDENT', for example, provides coverage for Dative Singular that is missing for ANALITIK 'ANALYST' and ZALOŽNIK 'HOSTAGE, and rare for ČEMPION 'CHAMPION'. Note, however, that one needs to look at more lexemes in order to find attestations of all of the potential forms, since, for example, none of the lexemes in Table 6 is attested in the Locative Singular, which is the rarest word form for masculine animate lexemes. The lexeme with the highest proportion of Locative Singular attestations in this group is POLITIK 'POLITICIAN' with only 3.31%.[12]

Correspondence analysis of the remaining groups of nouns in our sample (masculine inanimate, neuter inanimate, feminine inanimate II, and feminine inanimate III) yielded parallel results.

3.3 What grammatical profiles tell us about Russian nominal paradigms

We arrive at a model of Russian nominal morphology consisting of collections of grammatical profiles, such that each lexeme is at least partially "defective" due to unattested or rare word forms, but the entire "space" of the case/number values is populated by lexemes that differ according to their centers of gravity in

12 Note that because gold standard corpus data correctly connects each word form to the corresponding lemma, error induced by morphological ambiguity is eliminated. For example, this statement does not involve any misidentification of the homonymous Locative Singular form of the lexeme POLITIKA 'POLITICS, POLICY'.

that space. Is it possible to learn Russian morphology based on a network of partially overlapping sets of word forms? In other words, can one fill in the "holes" left by this system based on the partially overlapping collection of word forms? Or does one need to rely on full paradigms? These questions bring us back to the Paradigm Cell Filling Problem mentioned in Section 1. Section 4 details a computational experiment in which we address these questions.

4 Learning Russian morphology based on full paradigms vs. single word forms

We present a computational learning experiment that addresses the Paradigm Cell Filling Problem from the perspective of a model of overlapping partial sets of word forms. Our experiment differs from other morphological generation experiments in that it (1) takes into account the frequencies of word forms, and (2) compares the efficacy of learning from training on full paradigms with training on single word forms. Our results show that, while training on full paradigms gives greater gains early in the process (when the number of training items is small and accuracy is low), learning from training on single word forms quickly overtakes full paradigms, and single word forms ultimately facilitate more accurate predictions. Before describing our experiment, we situate it relative to previous achievements in morphological generation.

Among the primary motives for development of morphological generation models in computational linguistics are the data sparsity problems caused by languages with rich inflectional morphology. Traditionally, the most reliable way to solve these problems is by building two-level finite-state transducer models for each language. However, building such models can be an extremely labor-intensive enterprise, involving the crafting of hundreds or thousands of language-specific linguistic rules, and finite-state transducers have their own limitations: they overgenerate, meaning that they can become unwieldy with information that is never or almost never needed, and they cannot comfortably handle all types of morphological phenomena (a particular weak spot is reduplication).

The Cotterell et al. (2016) and (Cotterell et al. 2017) Shared Tasks were designed to discover new ways to handle inflectional morphology. The Sigmorphon challenge was taken up by nine teams of computational linguists in Europe and North America to create models for morphological generation that would learn from input and be applicable cross-linguistically. In 2016, 10 languages provided morphological challenges for the task; in 2017 the challenge was expanded to 52 languages. While the approaches of the teams differed

(see Cotterell et al. 2016 for a summary), the set up for all Sigmorphon submissions was similar. They worked from the perspective of full paradigms and the task was "reinflection": morphological analysis of a given word form and then generation of another word form of the same lexeme. Typically this involved supervised training on a subset of word forms of a few hundred given lexemes (for example, the word forms that constitute 90% or 60% of paradigms) and then producing the remaining (10% or 40%) word forms. Recurrent neural networks were found to give the best results, in particular the submission of Kann and Schütze (2016a-b). However, both submissions to the Sigmorphon Shared Tasks and other "reinflection" models place a wide variety of restrictions on the types of input data. For example, Faruqui et al. (2016) ran a model that handles only one part of speech at a time, and Aharoni et al. (2016) worked on only one paradigm per part of speech, while Malouf (2016) modelled only noun inflection. Most recently, Malouf (2017) modelled production of word forms based on partial paradigms, making that study more similar to our own.

While our approach is informed by and shares key components of previous achievements, our goal is different, since we use frequency-ordered input and aim to compare learning from exposure to full paradigms with learning from exposure to single word forms.

4.1 Our experimental set-up

Our experiment includes noun, verb, and adjective word forms presented for training and testing in decreasing order of their relative frequency, starting from the most frequent word form. Training is performed according to two models: a full-paradigm model in which training includes exposure to all word forms in the paradigm of each lexeme, and a single-form model in which training gives exposure only to individual word forms supplied with a lemma and tagset. The testing task for both models is the same: the production of a word form of a previously unencountered lexeme given only the lemma and tagset.

Note that the inclusion of all three open-class inflected parts of speech considerably complicates the task with, in addition to the 12 (or as many as 14) possible word forms for nouns, 28 word forms for adjectives, and numbers of possible word forms on the order of one hundred for verbs (varying somewhat from verb to verb).

The SynTagRus corpus provides the measurement of frequency of word forms used in our experiment. All of the inflected word forms in SynTagRus were ordered according to their frequency and supplied with their lemma, part of speech, and their tagset from SynTagRus. A sample of the top 25 most frequent word forms is presented in Table 7.

Table 7: The top 25 most frequent word forms in the SynTagRus corpus with their tagsets (Imp = Imperfective, Ind = Indicative, Sing = Singular, 3 = third person, Pres = Present, Fin = finite, Act = Active, Inan = inanimate, Gen = Genitive, Masc = masculine, Plur = Plural, Loc = Locative, Acc = Accusative, Neut = neuter, Fem = feminine, Anim = animate, Nom = Nominative, Ins = Instrumental, Pos = positive, Perf = Perfective).

Frequency &word form	Lemma	Part of speech	Tagset of word form
1447 možet	moč'	VERB	Aspect = Imp|Mood = Ind|Number = Sing|Person = 3|Tense = Pres|VerbForm = Fin|Voice = Act
1286 goda	god	NOUN	Animacy = Inan|Case = Gen|Gender = Masc|Number = Sing
999 let	god	NOUN	Animacy = Inan|Case = Gen|Gender = Masc|Number = Plur
832 godu	god	NOUN	Animacy = Inan|Case = Loc|Gender = Masc|Number = Sing
813 vremja	vremja	NOUN	Animacy = Inan|Case = Acc|Gender = Neut|Number = Sing
678 rossii	rossija	NOUN	Animacy = Inan|Case = Gen|Gender = Fem|Number = Sing
571 mogut	moč'	VERB	Aspect = Imp|Mood = Ind|Number = Plur|Person = 3|Tense = Pres|VerbForm = Fin|Voice = Act
571 ljudi	čelovek	NOUN	Animacy = Anim|Case = Nom|Gender = Masc|Number = Plur
543 rossii	rossija	NOUN	Animacy = Inan|Case = Loc|Gender = Fem|Number = Sing
436 javljaetsja	javljat'sja	VERB	Aspect = Imp|Mood = Ind|Number = Sing|Person = 3|Tense = Pres|VerbForm = Fin|Voice = Act
416 slučae	slučaj	NOUN	Animacy = Inan|Case = Loc|Gender = Masc|Number = Sing
411 ljudej	čelovek	NOUN	Animacy = Anim|Case = Gen|Gender = Masc|Number = Plur
403 strany	strana	NOUN	Animacy = Inan|Case = Gen|Gender = Fem|Number = Sing
400 žizni	žizn'	NOUN	Animacy = Inan|Case = Gen|Gender = Fem|Number = Sing
392 čelovek	čelovek	NOUN	Animacy = Anim|Case = Nom|Gender = Masc|Number = Sing
377 obrazom	obraz	NOUN	Animacy = Inan|Case = Ins|Gender = Masc|Number = Sing
375 razvitija	razvitie	NOUN	Animacy = Inan|Case = Gen|Gender = Neut|Number = Sing

(*continued*)

Table 7: (*continued*)

Frequency &word form	Lemma	Part of speech	Tagset of word form
367 èkonomiki	èkonomika	NOUN	Animacy = Inan|Case = Gen|Gender = Fem|Number = Sing
366 čeloveka	čelovek	NOUN	Animacy = Anim|Case = Gen|Gender = Masc|Number = Sing
360 mnogie	mnogie	ADJ	Case = Nom|Degree = Pos|Number = Plur
351 vlasti	vlast'	NOUN	Animacy = Inan|Case = Gen|Gender = Fem|Number = Sing
350 delo	delo	NOUN	Animacy = Inan|Case = Nom|Gender = Neut|Number = Sing
349 drugix	drugoj	ADJ	Case = Gen|Degree = Pos|Number = Plur
347 skazal	skazat'	VERB	Aspect = Perf|Gender = Masc|Mood = Ind|Number = Sing|Tense = Past|VerbForm = Fin|Voice = Act
343 raz	raz	NOUN	Animacy = Inan|Case = Acc|Gender = Masc|Number = Sing

Reading from the top of Table 7, for example, we find the most frequent word form is *možet*, which appears 1,447 times and is the third person Singular Indicative Present Tense finite form of the Imperfective active verb that has *moč'* 'be able' as its lemma. Table 7 shows that all three inflected parts of speech are included, as are lexemes with irregular and suppletive paradigms, such as ČELOVEK 'PERSON' and GOD 'YEAR'.

We generated full paradigms for all lemmas in the list of frequency-ordered word forms.[13] These full paradigms served as the basis for training on full paradigms, while the frequency-ordered list of word forms served as the basis for training on single forms. In both cases, for each word form, the input was as represented in Table 7: a word form, plus the lemma, part of speech, and tagset.

The idea was to model learning from the word forms a learner was most likely to encounter, and see how well the learning model could, on the basis of those forms, produce the next most likely word forms, and then to progressively iterate this process, mimicking how a learner might gradually build up a vocabulary of word forms as well as an ability to produce the word forms that

13 The experiments were performed using version 1.4 of the SynTagRus corpus converted to Universal Dependencies (Nivre et al. 2016) and UDAR (Reynolds 2016), a morphological analyser/generator for Russian. Because the tagsets for the SynTagRus were not compatible with those of UDAR, we performed a conversion via a simple longest set overlap algorithm. Of 6837 lemmas, we discarded 911 (13%) for which the full paradigm could not be generated by UDAR.

the learner is most likely to need next. We batched the data in sets of 100 for the purpose of successive iterations, such that in the first iteration the learning model was trained on the word form, lemma, and tagsets of the 100 most frequent word forms (those ranked 1–100), and then tested on (asked to produce) the next 100 most frequent word forms (those ranked 101–200). In the second iteration, training was performed on the 200 most frequent word forms (those ranked 1–200); while testing took the next 100 most frequent word forms (those ranked 201–300). This process was ultimately repeated in fifty-four consecutive iterations, each time adding the testing data from the previous round to the training data and using for testing the next 100 most frequent word forms. In the last round training involved the 5,400 most frequent word forms (for the single-form model) and the full paradigms of the corresponding lexemes (for the full-paradigm model) and tested the production of word forms ranked 5,401–5,500 in terms of frequency, at which point the data resources of SynTagRus were exhausted. In each iteration, the full-paradigm model got trained on all the word forms in the paradigm associated with each word form in the training dataset, while the single-form model got trained only on the specific word forms in the training dataset.

Figure 2 compares the quantity of training data provided to the two models, showing the gap between the large number of word forms that the full-paradigm model received training on vs. the word forms that the single-form model was trained on.

In order to level the playing field for the comparison between learning based on full paradigms and learning based on single word forms, we had to ensure that the full-paradigm model did not gain an unfair advantage from seeing word forms that it would then be tested on (an advantage that could not be gained in the single-form model). To illustrate the problem, notice in Table 7 that the word form *žizni*, which is the Genitive Singular form of the lexeme ŽIZN' 'LIFE', is the 14th most frequent word form in SynTagRus. When we come to this word form, the full-paradigm model will be trained on all of the word forms of the lexeme ŽIZN' 'LIFE', while the single-form model will be trained only on the Genitive Singular form *žizni*. The Accusative Singular form of the same lexeme, which is *žizn'*, is ranked as the 193rd most frequent word form, so if it was not removed, it would be a candidate for the testing data in the first iteration. However, from the point of view of the single-form model, it would be unfair to use Accusative Singular *žizn'* as a test item because the full-paradigm model was trained on Accusative Singular *žizn'* (and all inflected forms of that lexeme), but the single-form model was not (all it got was the Genitive Singular form *žizni*). This is the case for every lexeme that has more than one word form in the dataset: once the most frequent word form of that lexeme has been included in the training data,

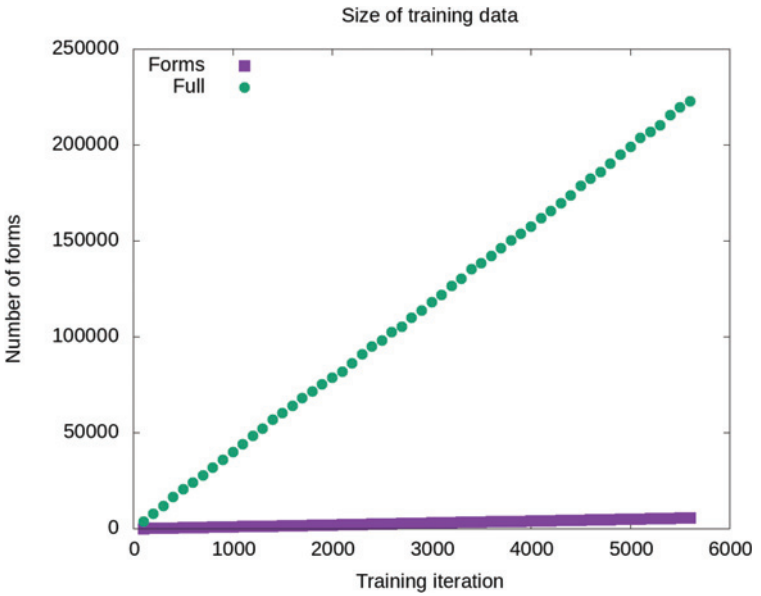

Figure 2: Size of training data for single-form model (purple squares) and full-paradigm model (green circles) measured on y-axis in terms of number of forms learned in training. X-axis is the number of forms that training is based on, in batches of 100, from 100 to 5400.

the full-paradigm model will be trained on all of the lexeme's word forms, but the single-form model will not. So when the second most frequent word form of the same lexeme comes up as a candidate for testing, it has to be eliminated because, like Accusative Singular *žizn'*, it would not measure the production abilities of the full-paradigm model on unencountered forms. All of the lexemes are identical across the two models, making them parallel: both models receive data from the same set of training lexemes (with single forms for the single-form model, but full paradigms for the full-paradigm model), and are tested on the same forms, where the task is to produce, given a lemma and a tagset, a word form from a lexeme that has not previously been encountered by either model. In both models, the same lexeme never appears in both training and testing.

We used a sequence-to-sequence character LSTM (long short-term memory) architecture modelled on MED (morphological encoder-decoder, cf. Kann and Schütze 2016a-b), the 2016 system of the team that won the ACL SIGMORPHON shared task on morphological generation both in 2016 and 2017 (Cotterell et al. 2016; Cotterell et al. 2017). Our implementation is based on Theano (Theano Development Team 2016) and Blocks (Merriënboer et al. 2015) and is freely available for testing online. Each training cycle consisted of several "epochs",

each of which went through the entire set of training data in random order. In each epoch, the training data is divided into "minibatches" for the learning model to use to update the weights in its network. It cycles through all the minibatches thirty times in each epoch.

4.2 Results of the experiment

Given the huge advantage for the full-paradigm model in terms of training data shown in Figure 2, it would be reasonable to expect the full-paradigm model to outperform the single-form model. However, that is not the case. Figure 3 is a visualization of the results of our computational experiment, comparing the results for training on single forms with training on full paradigms through 54 iterations of training and testing.

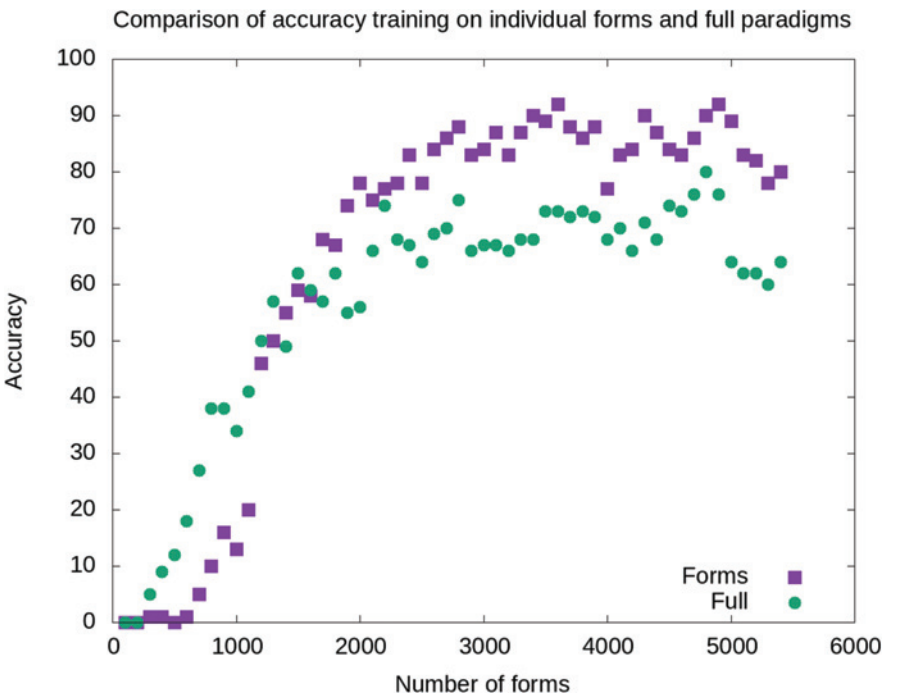

Figure 3: Results of computational experiment. X-axis is the number of forms that training is based on, in batches of 100, from 100 to 5400. Y-axis is the percent accuracy in production of tested forms, expressed as the number correctly produced. Purple squares show results for training on single forms; green circles show results for training on full paradigms.

As in Figure 2, in Figure 3, each iteration for the parallel models is marked with purple squares for the single-form model, and with green circles for the full-paradigm model. The numbers of word forms/paradigms used as training data appear along the x-axis. The y-axis represents the accuracy of the two models in producing correct forms in percentages. At the origin at the bottom left we see that in the first two iterations, when the models were trained based on the first 100 and then the first 200 most frequent word forms, both the single-form model and the full-paradigm model failed completely, with 0% of tested forms correctly produced. For the next eight iterations, the full-paradigm model outperforms the single-form model, but accuracy for both models is low (about 40% or less). For the next six iterations, the two models are roughly equal in their performance (45%–62%). But for the remaining 38 iterations of the experiment, the single-form model consistently outperforms the full-paradigm model on every single iteration and the single-form model is the only model that ever scores above 80%.

Figure 4 shows the average edit distance (Levenshtein 1965/1966) of the errors made by the single-form model (purple squares) vs. the full-paradigm model (green circles). Edit distance is measured as the number of letters one needs to change in order to convert an error into a correct form. In the first seven iterations, the full-paradigm model consistently shows a smaller average edit distance. However, after that, for all remaining iterations except for iteration 35, the average edit distance for the single-form model is lower than that of the full-paradigm model. Whereas Figure 3 shows us that the single-form model consistently produces a higher percentage of correct forms, Figure 4 tells us that even when the single-form model makes errors, those errors are smaller than the errors made by the full-paradigm model.

4.3 What the computational experiment tells us about Russian nominal paradigms

Our experimental results indicate that learning might actually be better, at least in the long run, when training is restricted only to the forms that are most frequent, rather than requiring learning to encompass entire paradigms. It appears likely that training on full paradigms overpopulates the search domain with a multitude of word forms that one is unlikely to be exposed to or need to produce, and which can, after training on 1800 or more word forms, be predicted anyway.

There are, of course, many caveats to the interpretation of this experiment. For example, the SynTagRus corpus is primarily a written source, whereas L1 learning involves primarily child-directed speech. However, existing child-directed speech corpora are insufficient to the task of such an experiment, and for

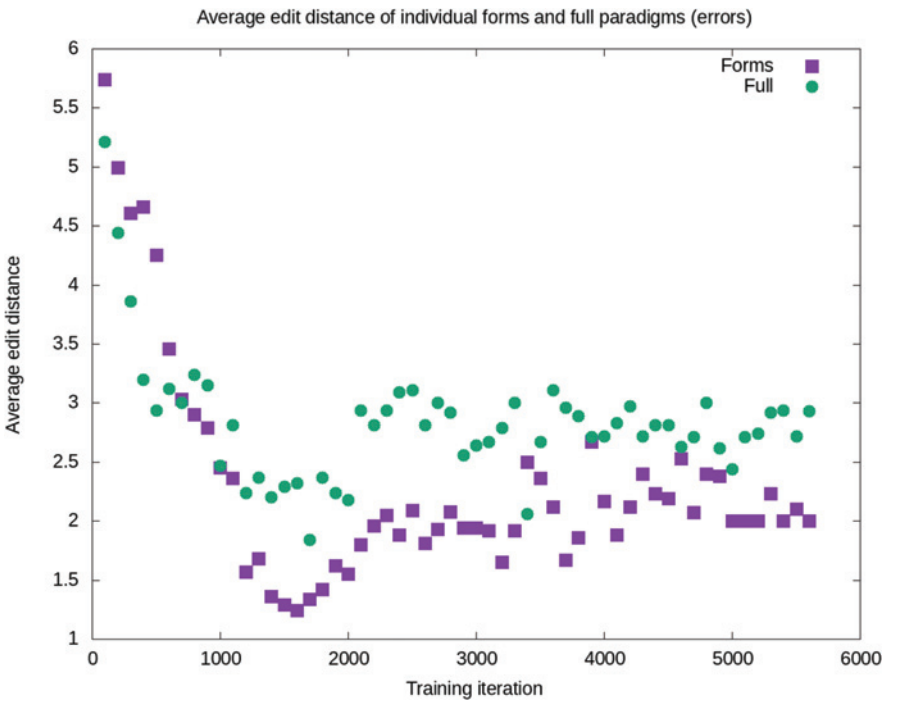

Figure 4: Average edit distance for errors made in testing. X-axis is the number of forms that training is based on, in batches of 100, from 100 to 5400. Y-axis is the average edit distance for forms produced incorrectly. Forms produced correctly are not taken into account in this graph. Purple squares show results for training on single forms; green circles show results for training on full paradigms.

this reason we use SynTagRus, while acknowledging that it is merely an approximation of average exposure to word forms. Human beings also learn language in contexts that are relatively rich both linguistically and experientially. In our experiment, the tagsets serve as a proxy for context in training, albeit of course a very limited context. Furthermore, although the ordering of the training and testing data in our experiment is matched to frequency as measured in a corpus, in real life L1 learners aren't exposed only to the highest frequency word forms precisely in their order of frequency, but to a variety of word forms of different frequencies. However, the frequency ordering does reflect the likelihood that a learner will encounter and have to produce word forms.

With respect to L2 learning, the difference in results for the single-form vs. full-paradigm models can be compared along the scale of lexemes that a learner of Russian is expected to master. Andrjušina's (2006: 4) "lexical minimum" stipulates

the following scale: beginning learners should acquire 780 lexemes, a basic vocabulary is 1,300 lexemes, 2,300 is the minimum for certification at level one (for education in Russia), while levels two and three require 6,000 and 12,000 lexemes, respectively. The beginning and basic levels correspond roughly to the first two semester college courses in Russian language as it is taught outside of Russia. In other words, already by the end of the first semester, the L2 learner should be at the level of mastery simulated in approximately the eighth iteration of our experiment, by which point the severity of errors is less in the single-form model. Soon thereafter, corresponding to a point early in second-semester Russian, overall accuracy of word form production is consistently better for the single-form model.

5 Conclusions

This article presents three kinds of evidence for inflectional morphology as a system of partial sets of word forms expressing partially overlapping combinations of morphosyntactic features. The overlap is sufficient to enable the production of unencountered word forms without overloading the system by filling in all of the "gaps" left by the partial sets of word forms. We show that the proportion of full paradigms experienced by native speakers is small, and quickly vanishes as the number of word forms in a full paradigm expands. All lexemes have "defective" paradigms to some extent since some word forms are rarely or never encountered. A lexeme usually has 1–3 word forms that are most prototypical for that lexeme and are motivated by the grammatical constructions and collocations that are typical for that lexeme. We show what the patterns of overlapping partial sets of word forms look like for five types of Russian nouns. We also show that in an experiment with Russian word forms from all open-class inflected parts of speech ranked according to frequency (a proxy for the likelihood that a speaker would need to produce them), a computational model trained on single word forms outperforms one trained on full paradigms. It seems that learning may be enhanced by focusing only on the word forms most likely to be encountered rather than taking entire paradigms as input. This result is consistent with a usage-based model of language in which memorization and the learning of patterns coexist. High-frequency forms are likely stored and may also be used as the basis for abstracting schemas (in this case lexemes and the patterns among their word forms).

While our study focused on Russian, it is likely that other languages with inflectional morphology pattern similarly. For example, already in the 1980s Karlsson (1985; 1986) observed that the probability of various word forms differed across the cells of Finnish paradigms, although he attributed this to

non-lexeme-specific trends for nouns in general, and Arppe (2006) took this research further, investigating the patterns for Finnish verbs. Russian is a good point of reference because among languages that are well-documented and have a gold standard corpus (like SynTagRus), Russian is morphologically relatively complex, in terms of the number of word forms in its paradigms, the number of inflectional classes, and the proportion of irregular and suppletive word forms.

Given our findings, one has to question the wisdom of making L2 learners memorize full paradigms. Today most textbooks build up the paradigm gradually, one subparadigm or one morphosyntactic feature combination at a time (for example, teaching the Past tense subparadigm of all verbs or the Locative Singular form for all declension classes of nouns), but one goal is still to memorize full paradigms for a series of lexemes that represent the various inflectional classes. While efficiency is certainly a concern (L2 learners typically do not get anywhere near the amount of one-on-one input that L1 learners get), there remains the problem that most of the information in the full paradigm of any given lexeme is not very useful to the L2 learner either. One can envision a learning environment in which dictionaries present entries headed by the most common word forms, along with the constructions that motivate those word forms, in addition to gathering such information under the entry for the lemma. An electronic dictionary of this type would not only allow the user to search for any word form, but also alert the user to the relative frequency of each word form for its lexeme. Morphological drills would target the handful of word forms most likely to be encountered for each lexeme and present those in the context of their most prototypical constructions. Reading enhancement tools would likewise code word forms according to whether they should be memorized or merely recognized, and these could be gauged to the proficiency level by scaling up gradually to less frequent lexemes and word forms. In this way, we could avoid overburdening students with the full-paradigm memorization task that we have shown to be less efficient in the long run.

References

Ackerman, Farrell, James P Blevins & Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P Blevins & Juliette Blevins (eds.), *Analogy in Grammar: Form and Acquisition*, 54–82. Oxford: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2016. Implicative relations in word-based morphological systems. In Andrew Hippisley & Gregory Stump (eds.), *Cambridge Handbook of Morphology*, 297–328. Cambridge: Cambridge University Press.

Aharoni, Roee, Yoav Goldberg & Yonatan Belnikov. 2016. Improving sequence to sequence learning for morphological inflection generation: The BIU-MIT Systems for the SIGMORPHON 2016 shared task for morphological reinflection. *In Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics*, Phonology, *and* Morphology *(SIGMORPHON at ACL)* 2016. DOI: 10.18653/v1/W16-2007

Albright, Adam. 2003. A quantitative study of Spanish paradigm gaps. In G. Garding & M. Tsujimura (eds.), *West Coast Conference on Formal Linguistics 22 proceedings*. Somerville, MA: Cascadilla Press, 1–14. http://web.mit.edu/albright/www/papers/Albright-WCCFL22.pdf

Andrjušina, N. P. 2006. *Leksičeskij minimum po russkomu jazyku kak inostrannomu. Bazovyj uroven'*. Obščee vladenie. Moscow/St. Petersburg: TsMO MGU/Zlatoust.

Arppe, Antti. 2006. Frequency considerations in morphology, revisited - Finnish verbs differ, too. In M. Suominen, A. Arppe, A. Airola, O. Heinämäki, M. Miestamo, U. Määttä, J. Niemi, K. K. Pitkänen, K. Sinnemäki & Kaius (eds.), *A Man of Measure. Festschrift in Honour of Fred Karlsson in his 60th Birthday*, Special Supplement to SKY Journal of Linguistics. vol. 19/2006. 175–189. Turku: Linguistic Association of Finland. http://www.ling.helsinki.fi/sky/julkaisut/SKY2006_1/1.3.1.ARPPE.pdf.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Gert E Booij & J. Van Marle (eds.), *Yearbook of Morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald. 1993. On frequency, transparency, and productivity. In Gert E Booij & J. Van Marle (eds.), *Yearbook of Morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

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

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

Booij, Gert. 2017. The construction of words In Barbara Dancygier (ed.), *The Cambridge Handbook of Cognitive Linguistics*, Chapter 15. Cambridge: Cambridge University Press.

Bybee, Joan L. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.

Comrie, Bernard & Maria Polinsky. 1998. The Great Dagestanian Case Hoax. In Anna Siewierska & Jae Jung Song (eds.), *Case, Typology, and Grammar*, 95–114. Amsterdam: John Benjamins.

Corbett, Greville G. 2015. Morphosyntactic complexity: A typology of lexical splits. *Language*. 91. 145–193. 10.1353/lan.2015.0003.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Gėraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner & Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages.In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 1–30.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner & Mans Hulden. 2016. The SIGMORPHON 2016 shared task— Morphological reinflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 10–22.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Diessel, Holger. 2015. Usage-based construction grammar In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, Chapter 14. Berlin: De Gruyter Mouton.

Faruqui, Manaal, Yulia Tsvetkov, Graham Neubig & Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *San Diego, California, USA*, *June 12 - June 17, 2016*. https://arxiv.org/abs/1512.06110

Goldberg, Adele. 2006. *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Hart, Betty & Todd R Risley. 2003. *The early catastrophe. The 30 million word gap by age 3. American Educator Spring 2003*. 4–9.

Janda, Laura A & Lene Antonsen. 2016. The ongoing eclipse of possessive suffixes in North Saami: A case study in reduction of morphological complexity. *Diachronica*. 33(3). 330–366. http://dx.doi.org/10.1075/dia.33.3.02jan.

Janda, Laura A & Olga Lyashevksaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. *Cognitive Linguistics*. 22(4) 719–763.

Kann, Katharina & Hinrich Schütze. 2016a. Single-model encoder-decoder with explicit morphological representation for reinflection. The Association for Computational Linguistics.In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 555–560.

Kann, Katharina & Hinrich Schütze. 2016b. MED: The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection.In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 62–70.

Karlsson, Fred. 1985. Paradigms and word forms. *Studia gramatyczne VII. Ossolineum*, 135–154.

Karlsson, Fred. 1986. Frequency considerations in morphology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*. 39. 19–28.

Kibrik, Andrei E. 2001. Archi (Caucasian—Daghestanian) In Andrew Spencer & Arnold M Zwicky (eds.), *The Handbook of Morphology*, Chapter 23. Hoboken, NJ: Wiley-Blackwell.

Kuznetsova, Julia. 2017. The ratio of unique word forms as a measure of creativity. In Anastasia Makarova, Stephen M. Dickey & Dagmar Divjak (eds.), *Each Venture a New Beginning: Studies in Honor of Laura A. Janda*, 85–97. Bloomington, In Slavica Publishers.

Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Levenshtein, Vladimir I. 1965/1966. Dvojnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Doklady Akademii Nauk SSSR*. 163(4). 845–848. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710.

Malouf, Robert. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistic Papers*. 6. 122–129.

Malouf, Robert. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*. 27. 431–458. 10.1007/s11525-017-9307-x.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Merriënboer, Bart van, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski & Yoshua Bengio. 2015. *Blocks and fuel: Frameworks for deep learning*. *arXiv preprint arXiv:1506.00619 [cs.LG]*.

Moreno-Sánchez, Isabel, Francesc Font-Clos & Álvaro Corral. 2016. Large-scale analysis of Zipf's Law in English texts. *PLoS One*. 11(1). e0147073. 10.1371/journal.pone.0147073.

Nesset, Tore & Laura A Janda. 2010. Paradigm structure: Evidence from Russian suffix shift. *Cognitive Linguistics*. 21(4) 699–725.

Nickel, Klaus P & Pekka Sammallahti. 2011. *Nordsamisk grammatikk*. Karasjok: Davvi Girji.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Christopher D Jan Hajic, Ryan McDonald Manning, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Paris: European Language Resources Association (ELRA).
http://www.lrec-conf.org/proceedings/lrec2016/summaries/348.html

Payne, John & Rodney Huddleston. 2002. Nouns and noun phrases. In Rodney Huddleston & Geoffrey Pullum (eds.), *The Cambridge Grammar of the English Language*, 479–481. Cambridge/New York: Cambridge University Press.

Pertsova, Katya & Julia Kuznetsova. 2015. Experimental evidence for lexical conservatism in Russian: Defective verbs revisited. In Yohei Oseki, Masha Esipova & Stephanie Harves (eds.), *Proceedings of the 24th Meeting of Formal Approaches to Slavic Linguistics*. Ann Arbor, Michigan: Michigan Slavic Publications. https://nyu.edu/projects/fasl24/proceedings/pertsova_kuznetsova_fasl24.pdf

Piperski, Alexander. Ch. 2015. To be or not to be: Corpora as indicators of (non-)existence. In V. P. Selegej (ed.), *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015)* 14(1),515–522.

Reynolds, Robert J. 2016. *Russian natural language processing for computer-assisted language learning*. Doctoral Dissertation, UiT The Arctic University of Norway.

Sims, Andrea D. 2006. *Minding the Gaps: Inflectional Defectiveness in a Paradigmatic Theory*. PhD Dissertation, Ohio State University.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spencer, Andrew. 2016. Two morphologies or one? Inflection versus word-formation. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge Handbook of Morphology*, 27–49. Cambridge: Cambridge University Press.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv:1605.02688v1.

Wurzel, Wolfgang U. 1984. *Flexionsmorphologie und Natürlichkeit.* Berlin: Akademie-Verlag.

Wurzel, Wolfgang U. 1989. *Inflectional Morphology and Naturalness*. Dordrecht. Boston and London: Kluwer Academic Publishers.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley.

Bionotes

Laura A. Janda

Laura A. Janda (born 1957, Ph.D., UCLA, 1984) is Professor of Russian Linguistics at UiT the Arctic University of Norway. Her special areas of interest are the complex factors associated with the grammatical categories of case and aspect and how these can be investigated using corpus data and experiments.

Francis M. Tyers

Francis M. Tyers (born 1983, Ph.D., Universitat d'Alacant, 2013) is Assistant Professor of Linguistics at Higher School of Economics in Moscow. He is passionate about language technology for lesser-resourced languages and has co-organised workshops on machine translation in a number of countries including Russia and Finland.