Russian Language Journal

Volume 69 | Issue 1 Article 9

2019

Businessmen and Ballerinas Take Different Forms: A Strategic Resource for Acquiring Russian Vocabulary and Morphology

Laura A. Janda

Businessmen and Ballerinas Take Different Forms: A Strategic Resource for Acquiring Russian Vocabulary and Morphology

Laura A. Janda

1. Introduction

Included in the tasks facing a language learner is the acquisition of a lexicon and a grammar. However, when the target language has inflectional morphology, these two parts of the language-learning task intersect in the paradigms of grammatical word forms because each open-class lexeme has a number of forms that allow it to express various combinations of grammatical categories. Among major world languages, Russian is relatively highly inflected, meaning that the challenges of acquiring vocabulary are compounded by the need to master the inflectional morphology. Even a modest basic vocabulary of a few thousand inflected lexemes has over a hundred thousand associated word forms. Recent research (Janda and Tyers 2018, described in more detail below) suggests that there could be an advantage to learning only a handful of high-frequency forms for each lexeme. Section 2 reviews distributional facts about paradigms, their theoretical implications, and the results of a computational experiment that simulates the learning of Russian paradigms either in their entirety or based only on the most frequent word forms. Section 3 presents a free public net-based resource, the Strategic Mastery of Russian Tool (SMARTool), which takes up the challenge of providing strategic input for second-language (L2) learning of Russian vocabulary. The design functions and some pedagogical applications of the SMARTool are detailed. Conclusions are offered in Section 4.

This article is a tribute to Olga Kagan's innovative spirit in the teaching of Russian. I was in the very first class of graduate students that Olga Kagan taught advanced Russian to in the early 1980s. Her steady focus on the practical aspects of teaching and learning Russian based on authentic usage has served as a model to me throughout my career, and is, I believe, also realized in the SMARTool that I present here. For many years, I assumed that mastery of Russian morphology required the ability

to recognize and produce all paradigm forms, but recently I was forced to rethink that assumption, and that process inspired the creation of the SMARTool.

2. Paradigm Model Versus Usage-Based Model Of Russian Word Forms

On the face of it, paradigms seem to be rather straightforward tables listing all the word forms that express the various grammatical categories associated with a given part of speech, as in Zalizniak (1980). These tables can be called the paradigm model of inflectional morphology and probably do not adequately represent the mental grammar of the language. In Russian, nouns express combinations of six cases and two numbers, yielding twelve paradigm slots; adjectives have twenty-eight slots in their paradigms (six cases combined with three genders plus plural, plus four short forms); and verbs have over a hundred paradigm slots (varying depending upon aspect and how one counts the participles). If we follow the paradigm model of morphology, the task of the L2 learner is to master all those tables of word forms.

In its extreme form, the paradigm model was implicit in the traditional grammar and translation method of language teaching, which is now largely obsolete. However, although this focus has diminished considerably in contemporary textbooks, paradigms are by no means gone. For example, the online introductory course *Между нами* (deBenedette et al. 2013) offers declension and conjugation charts under the *Таблицы* menu prominently located right at the top of its homepage, and reference grammars aimed at learners (such as Wade 2011) rely on paradigms to present Russian morphology. While paradigms have been backgrounded, no systematic pedagogical replacement for the paradigm model that would aim at native-like mastery of the morphology has been offered. As Comer (2019, 112) notes with respect to the presentation of vocabulary in *Между нами*, it "does not manage to completely cover the range of morphology that learners need to master to progress to higher levels of proficiency."

When one looks closely, several problems crop up with the paradigm model. There is considerable variation across paradigms, and furthermore, the mathematical facts of the distribution of word forms in natural language cast substantial doubt on the paradigm model. A usage-based model that reflects authentic language usage is offered here as an alternative.

Some details about variation in inflection are described in standard reference works. For example, some Russian nouns have more than twelve forms if we include forms like the second genitive (as in *выпить чаю* 'drink some tea'), second locative (as in *на мосту* 'on the bridge'), second accusative (as in *он пошел в солдаты* 'he joined the ranks of soldiers'), old vocative (as in *господи!* 'lord!'), and new vocative (as in *Саш!* 'Sasha!'). Some nouns have fewer than twelve forms, as in the case of nouns that are singularia tantum (such as *молодежь* 'young people'), are pluralia tantum (such as *ножницы* 'scissors'), or have paradigmatic gaps (such as the genitive plural of *мечта* 'dream'). Similar variations occur for adjectives (particularly with respect to the presence of short forms) and verbs (particularly with respect to certain combinations of aspect with participles and gerunds). Furthermore, both the presence of additional forms and the lack of certain paradigm forms are often variable across speakers and registers.

If variations like those listed here were the only challenges to the paradigm model, perhaps they could be swallowed as exceptions and that model could be retained. However, the distributional facts of word forms in an inflected language present much bigger threats to the paradigm model due to the inexorable power of Zipf's Law.

2.1. Zipf's Law and what it means for word forms

In 1949, Zipf discovered that the frequency of any word in a corpus is inversely proportional to its rank. If we take English, for example, the most frequent word is *the*. The second-most frequent word, *of*, is 1/2 as frequent as *the*. The third-most frequent word, *and*, is 1/3 as frequent as *the*. Fourth comes *a*, which is 1/4 as frequent as *the*, and so it goes, ending in a long tail of what are called "hapaxes," words that appear only once. This distributional fact is called "Zipf's Law." Remarkably, Zipf's Law holds true not just for English, but for all other languages that have ever been tested, even including constructed languages (Janda under submission) as well as numerous other (nonlinguistic) distributional phenomena. Zipf's Law has a number of surprising entailments. For example, approximately 50% of the unique lexemes in any corpus are hapaxes[1], and only 135

[1] Baayen (1992, 1993) demonstrates this based on Dutch and English data, and Kuznetsova (2017, 96) shows that more than half of nominal lexemes in the modern subcorpus of the Russian National Corpus appear in only one word form.

vocabulary items are needed to account for half of a corpus of one million English words (cf. the Brown Corpus). The following three facts connected to Zipf's Law are relevant to the discussion of word forms in this article: (1) Exposure to language can be likened to a big corpus, (2) Zipf's Law scales up infinitely, and (3) Zipf's Law applies to word forms too. I briefly elaborate on each of these facts below.

2.1.1. Language exposure as a big corpus

There are many types of language corpora, and even those that are carefully balanced may not perfectly represent the language that a typical native speaker is exposed to, particularly in terms of the way in which language is embedded in other realia. However, a large corpus is a close approximation to the lifetime linguistic input for a native speaker, which is estimated at about five to ten million words per year (cf. Hart and Risley 2003). There is no reason to expect significant deviations between a corpus and native input in the relative frequencies of lexemes, which necessarily follow Zipf's Law. In other words, what we find in terms of Zipfian distributions in large corpora (with millions or billions of words) reflects distributions of what a native speaker is exposed to over the course of a lifetime.

2.1.2. Zipf's Law scales up

Scalability has been tested by Manning and Schütze (1999) and Moreno-Sánchez, Font-Clos, and Corral (2016) with the conclusion that Zipf's Law, along with its entailments, scales up infinitely. This happens because the number of low-frequency items expands at scale as the size of the corpus increases, keeping the relative frequencies stable. This means that the Zipfian distributions remain the same regardless of corpus size, and the entailments hold even for very large corpora, like those that approximate a speaker's exposure to his or her native language.

2.1.3. Zipf's Law applies to word forms too

The Zipfian curve characterizes not just words, but all word forms as well. This has two implications for paradigms: (a) one concerning the distribution of forms within a paradigm and (b) another concerning the representation of entire paradigms. Within the paradigm of any single lexeme, we expect to see large differences in the frequencies of word forms, and this is borne out by the facts. For any given Russian lexeme of overall high frequency (≥50 per million words), one word form is most frequent, a couple more might be attested regularly (accounting

for >10% of attestations of the lexeme), and the remaining word forms are rare or unattested (Janda and Tyers 2018). For example, *бизнесмен* 'businessman' is attested fifty times in the SynTagRus corpus[2] of just over one million words. Sixteen of those attestations (32%) are of the genitive plural *бизнесменов*, ten attestations (20%) are of the nominative plural *бизнесмены*, seven attestations (14%) are of the nominative singular *бизнесмен*, most other word forms are rare, and three word forms (accusative singular, locative singular, and locative plural) are unattested. For some lexemes, the distribution is more extreme: over 90% of attestations of *балерина* 'ballerina' are of the instrumental singular form *балериной*. For low-frequency words, this effect is even more pronounced, usually with only one or two word forms attested – and recall that the presence of low-frequency lexemes expands proportionately with the size of a corpus.

The implications of Zipfian distribution of word forms for the representation of full paradigms are even more surprising. Since one word form in a paradigm will be of highest frequency, with the frequency of other word forms dropping off along the Zipfian curve, and since most unique lexemes are not of high frequency (recall that half of the unique lexemes in a corpus are hapaxes), the rate of fully attested paradigms declines sharply as the number of paradigm slots increases. For example, the SynTagRus corpus contains attestations of 21,945 unique Russian nominal lexemes; however, only thirteen of these lexemes are attested in all twelve forms of the nominal paradigm, equivalent to only 0.06% (Janda and Tyers 2018, 8). This statistic, in combination with the above observations about language exposure and the scalability of Zipfian distributions, means that a native speaker of Russian encounters all twelve paradigm forms of less than 0.1% of nouns that they are exposed to in the course of a lifetime. Conversely, for 99.9% of Russian nouns, the full paradigm is never realized. Since they have larger paradigms, the portion of adjectives and verbs that are attested in all paradigm forms is vanishingly small, for all practical purposes zero. These implications for paradigms are not limited to Russian but have been observed across languages and appear to be universal (cf. Malouf 2016).

[2] The SynTagRus corpus is available at http://www.ruscorpora.ru/instruction-syntax.html. SynTagRus is the only human-corrected corpus of Russian containing comprehensive morphological annotation that disambiguates syncretic word forms. For more about this corpus, see Diachenko et al. (2015).

Some readers are no doubt experiencing a degree of discomfort with these facts, particularly native speakers who have the intuition that the full paradigms are cognitively real. Oddly enough, the intuition that full paradigms are cognitively real is not necessarily incompatible with the data on Zipfian distributions. This paradox is addressed in relation to the Paradigm Cell Filling Problem in the next subsection.

2.2. The Paradigm Cell Filling Problem

Acknowledging the Zipfian implications for paradigms, Ackerman et al. (2009) express a linguistic conundrum they term the Paradigm Cell Filling Problem, namely the fact that native speakers of languages with complex inflectional morphology routinely recognize and produce forms that they have never been exposed to. For example, the lexeme *тамада* 'toastmaster' has no attestations of dative plural or locative plural forms in the Russian National Corpus (http://ruscorpora.ru/; the main corpus contains 283,431,966 words as of April 2019), and it is likely that many native speakers have never encountered these word forms. However, all native speakers of Russian can be expected to readily understand the forms *тамадам* and *тамадах* and to produce them in appropriate contexts.

In Janda and Tyers (2018), we provide statistical evidence that the word forms in the paradigm of an inflected part of speech (in other words, nouns, adjectives, or verbs) can be modeled as a multidimensional space. The entire space is the full paradigm. For Russian nouns, for example, the space is defined in terms of case and number and the distribution of word forms. Each nominal lexeme populates some part of that space. Taking our examples from above, *бизнесмен* 'businessman' most strongly populates the genitive plural, nominative plural, and nominative singular parts of the space, while *балерина* 'ballerina' most strongly populates the instrumental singular part of the space. Other nouns populate other parts of the space, with many nouns overlapping in their contributions to the space. In aggregate, the attestations of word forms for nouns populate the entire space, creating the sense that it is a whole, and making it easy for native speakers to triangulate from attested word forms to fill in gaps. This solves the Paradigm Cell Filling Problem and also explains the intuitions of native speakers. But what might the Zipfian distribution of word forms mean for the acquisition of inflectional morphology? This question is addressed in a learning experiment.

2.3. Results from a computational learning experiment

In Janda and Tyers (2018), we present a computational simulation of the learning of Russian inflectional morphology for all open-class inflected parts of speech: nouns, verbs, and adjectives. This experiment is based on data from the SynTagRus corpus. The dataset contains the single most frequent word form for each of 5,500 unique lexemes that appear at least fifty times in that corpus. The experiment had both a learning task and a production task. The experiment was run in two versions: the full-paradigm version, in which the learning task was to learn the entire paradigm of each lexeme, and the highest-frequency-word-form version, in which the learning task was to learn just the single highest frequency word form and the lemma (dictionary) form. The production task was the same for both versions, namely, given the lemma form of a previously unseen lexeme and the parse set for that lexeme's most frequent word form, to predict the word form. For example, given the lemma *жизнь* 'life' and the parse set "genitive singular," the production task would be to predict the form *жизни*.

The experiment was run in parallel in the two versions (full paradigm vs. single form), in fifty-four successive iterations. In both versions a computer simulated learning of Russian morphology. In the first iteration, the training set was based on the 1–100 most frequent word forms in SynTagRus, and the production set consisted of the 101–200 most frequent word forms of unique, unseen lexemes (i.e., lexemes that did not appear in the training set). The full-paradigm model learned the entire paradigms for 100 words, while the single-form model learned only the single most frequent form and the lemma form. Both models then predicted the 101–200 most frequent word forms given only the lemma and the parse set for each. In the second iteration, the training set was based on the 1–200 most frequent word forms (and their paradigms for the full paradigm model), and the production task was based on the 201–300 most frequent word forms of unique unseen lexemes. This procedure was repeated through fifty-four iterations, each time adding the data from the production task of the previous iteration into the training data for the successive iteration. Thus the size of the training set increased across the two models, but at different rates, such that the full-paradigm model learned over 200,000 word forms, while the single-form model learned only 5,400 word forms plus the associated lemmas.

At each iteration, the predictions on the production task were measured for both models, in terms of both overall accuracy (number of correct predictions out of 100) and severity of errors measured in Levenshtein distance (i.e., the number of letters needed to change to arrive at the correct form). In terms of overall accuracy, both models failed completely on the first two iterations. For the next eight iterations, the full paradigm model did better than the single forms model, but both models were still quite poor, with 40% or fewer correct predictions. On iterations eleven through fifteen, the performance of the two models was similar, at about 45%–62% correct. Thereafter, for the remaining thirty-eight iterations, the single-form model outperformed the full-paradigm model every time. The learning curve of the full-paradigm model flattened out in the 60%–70% range, while the single-form model performed in the 80%–95% range. In terms of average Levenshtein distance, when errors were made, in the first six iterations the full-paradigm model made less severe errors than the single-form model, but both models performed rather poorly (average edit distance of >3 letters). In the seventh iteration, the scores were nearly identical. After that, for all remaining iterations except one (iteration thirty-five), the single-form model made less severe errors when it did make errors (average edit distance in the range of 1–2.5).

In summary, our computational learning experiment shows that, after exposure to about 1,000 lexemes, learning that focuses only on the most frequent word forms consistently outperforms learning based on full paradigms both in terms of the accuracy of predictions of word forms of previously unseen lexemes and in terms of the severity of errors. Learning full paradigms does not appear to be the most effective way to acquire Russian inflectional morphology — it might simply overpopulate the search domain to the point that producing word forms gets harder rather than easier.[3]

2.4. What these facts mean for L2 acquisition of Russian

We can summarize the contents of the previous three subsections as follows. The distribution of word forms according to Zipf's Law means

[3] It is not possible in the scope of this article to address the inevitable differences between the human mind and a computational model. However, it seems reasonable that one should not expect the human mind to outperform a computer in terms of the memorization required by the full paradigm model.

that only a fraction of word forms of any given lexeme are encountered frequently, while the majority of word forms are encountered rarely, and many word forms may never be encountered. Different lexemes have different patterns of attested word forms, and overlapping patterns populate the conceptual space of the paradigm. Despite the usage-based facts of distribution, native speakers easily recognize and produce even rare and unattested word forms. Evidence from a computational learning experiment suggests that when learning focuses only on the most frequent word forms, the ability to produce specific word forms for new lexemes is better, both in terms of overall accuracy and severity of errors.

In light of these facts, asking L2 students to memorize and produce entire paradigms for all lexemes when learning Russian vocabulary is probably ill-advised. It makes more sense to utilize existing quantitative data on the distribution of Russian word forms to inform teaching in a strategic fashion. Corpus data can guide the design of teaching tools by showing us both the frequency distribution for Russian word forms and the contexts in which they most typically appear. In the next section, I describe a resource inspired by the research outlined above.

3. Design Of The Smartool

The SMARTool is a free resource publicly available at http://uit-no.github.io/smartool/. In this section, I detail the design of the SMARTool, including the selection of vocabulary and word forms, the presentation of contexts of use, and additional features, such as audio, translations, and filters.

Among technological resources for second-language learning, corpora have not been used to their full potential largely because they are devised by and for corpus linguists rather than for L2 learners and rate low in terms of user-friendliness, particularly for students at lower levels (Golonka et al. 2014, 78; Chun, Kern, and Smith 2016, 72). The SMARTool is a purposeful technological resource that bridges the gap between the facts of Russian morphology that can be gleaned from a corpus and the needs and abilities of L2 learners at various levels of proficiency, including that of the novice.

3.1. Vocabulary selection

The initial goal of the SMARTool is to represent word forms of 3,000 Russian lexemes, distributed across the first four Common European

Framework of Reference for Languages (CEFR) levels[4] and their ACTFL (American Council on the Teaching of Foreign Languages) and Russian equivalents (ТЭУ = Тест элементарного уровня, ТБУ = Тест базового уровня, ТРКИ = Тестирование по русскому языку как иностранному), as displayed in Table 1.

Table 1. Distribution of SMARTool lexemes across L2 acquisition levels

CEFR Level	ACTFL Equivalent	Russian Equivalent	SMARTool number of lexemes
A1 "Beginner"	Novice Low-Mid	ТЭУ	500
A2 "Elementary"	Novice High	ТБУ	500
B1 "Intermediate"	Intermediate Low-Mid	ТРКИ-1 I Сертификационный уровень	1,000
B2 "Upper Intermediate"	Intermediate High-Advanced Low	ТРКИ-2 Второй уровень	1,000

This distribution of lexemes is designed to provide a basic vocabulary for the first four semesters of Russian study for L2 learners. Since the architecture supporting the SMARTool is now in place, it will be possible to expand the vocabulary at these levels and also to add vocabulary at the C1 "Advanced"/ Advanced Mid-High/ ТРКИ-3 and C2 "Mastery"/ Superior/ ТРКИ-4 levels in the future.

Of course it would have been possible to simply harvest the highest-frequency lexemes from a corpus or frequency dictionary. However, the vocabulary needed by an L2 learner cannot be derived that simply, since there are numerous topics that are more specific to the experience and expectations of L2 speakers (cf. Comer [2019, 96] for a comparison of the needs of learners with frequency dictionaries). Lexemes were selected from a merged list of vocabulary from five Russian language textbooks (Hertz et al. 2001, Chernyshov 2004, Robin,

[4] For more on CEFR levels as established by the Council of Europe, see http://www.coe.int/ en/web/language-policy/home.

Shatalina, and Evans-Romaine 2012, deBenedette et al. 2013, Bondar' and Lutin 2013) plus the *Лексический минимум по русскому языку как иностранному* (Andriushchina et al. 2014–2015) for the corresponding levels. A panel of experienced teachers of Russian from three universities in Russia and Europe collaborated on the selection of lexemes (see SMARTool team members listed in the Acknowledgements).

Because the goal of the SMARTool is to provide input for acquisition of inflectional morphology, only open-class inflected lexemes are targeted in the SMARTool: nouns, verbs, and adjectives. Closed-class lexemes, such as pronouns, and uninflected lexemes, such as prepositions, are not represented. The SMARTool aims for a distributional balance across nouns, verbs, and adjectives that reflects the overall distribution of these parts of speech in Russian.[5] In most cases, both the perfective and imperfective partners of verb pairs are represented (provided that both are of reasonably high frequency). Supplying missing aspectual partner verbs expanded the number of verb lexemes.

3.2. Identification of high-frequency word forms

The next task was to identify the highest-frequency word forms associated with each lexeme. One challenge in this task was the presence of syncretism in Russian paradigms. For example, the form *радости* could potentially be any of five word forms of радость 'joy': the genitive singular, dative singular, locative singular, nominative plural, or accusative plural. Even the disambiguated subcorpus ("*снятник*") of the Russian National Corpus is not adequate for this task, since it has not been thoroughly corrected manually. The only substantial corpus of Russian that has 100% manually corrected disambiguation is SynTagRus, which belongs to the class of "gold standard" corpora with reliable morphological tagging (which is why SynTagRus is cited also in Section 2 above). According to SynTagRus, *радости* is most often the genitive singular form, which is the second-most-common form of this word, after *радость* as the nominative singular and before *радостью* as the instrumental singular.

[5] Endresen et al. (2016) report the following figures on attestations of parts of speech from the disambiguated subcorpus ("*снятник*") of the Russian National Corpus: 1,707,312 attestations of nouns, 1,007,526 attestations of verbs, and 784,340 attestations of adjectives. Given these figures, the distribution among open-class inflected lexemes is approximately 49% nouns, 29% verbs, and 22% adjectives.

The selected lexemes were queried in the SynTagRus corpus to determine the frequency distributions of their word forms, also known as "grammatical profiles" (cf. Janda and Tyers 2018). Like *бизнесмен* 'businessman' and *балерина* 'ballerina' cited above in Section 2, each lexeme has a unique grammatical profile with a small subset of word forms that occur often, while the rest of the forms are rare or even unattested. For each lexeme, we selected the three most common word forms. However, if over 90% of attestations for a given lexeme were accounted for by only one or two forms, then only those forms were selected. For example, for *бизнесмен* 'businessman', the three most common forms were selected: the genitive plural *бизнесменов*, the nominative plural *бизнесмены*, and the nominative singular *бизнесмен*. For *сентябрь* 'September' two word forms account for over 90% of attestations: the genitive singular *сентября* and the locative singular *сентябре*, so only those two forms are represented in the SMARTool. And since over 90% of attestations of *балерина* 'ballerina' are the instrumental singular form *балериной*, only that form is selected for the SMARTool. In total over 9,000 word forms are represented in the SMARTool.[6]

3.3. Identification of typical contexts

The next task in building the SMARTool was to determine, for every single word form, what grammatical and lexical contexts were most typical. In other words, what grammatical constructions and lexical collocations motivate each word form. For a few items, the answer to this question was trivial, as in the case of *сентябрь* 'September', for which the genitive singular *сентября* and the locative singular *сентябре* are motivated by typical constructions involving months, as in *первого сентября* 'on the first of September' and *в сентябре* 'in September'. But for the majority of word forms, this was a labor-intensive task, entailing some research, such as queries in the Russian National Corpus, in the Collocations Colligations Corpora (http://cococo.cosyco.ru/), and in the Russian Constructicon (https: //spraakbanken.gu.se/karp/#?mode=konstruktikon-rus). For example, a typical context for the genitive plural *бизнесменов* involves the collocation *защищать интересы бизнесменов* 'protect the interests of

[6] As mentioned above, the goal of providing both perfective and imperfective partner verbs somewhat expanded the number of verbs, and this compensated for the reduction in forms due to highly skewed grammatical profiles for words like балерина 'ballerina' and сентябрь 'September'.

businessmen', whereas a typical context for the instrumental singular *балериной* is *мечтать стать балериной* 'dream of becoming a ballerina'.

After typical contexts have been determined, we provide an example sentence showing the use of each word form, as in these examples:

Новый закон защищает интересы бизнесменов.
'The new law protects the interests of businessmen.'

Бизнесмен должен быть честным.
'A businessman has to be honest.'

Российские бизнесмены протестуют против повышения налогов.
'Russian businessmen are protesting against a tax increase.'

Первого сентября начинается учебный год.
'The academic year starts on the first of September.'

В сентябре начинают опадать листья.
'In September the leaves begin to fall.'

Анна Павлова с детства мечтала стать балериной.
'As a child, Anna Pavlova dreamed of becoming a ballerina.'

The example sentences are inspired by corpus examples but are adjusted to take into account the needs of learners at various levels. At the time this article was written (April–June 2019), all of the most frequent word forms had been identified for all lexemes at all four CEFR levels (A1, A2, B1, and B2), and example sentences had been supplied for all word forms at the A1 and A2 levels and for most of the word forms at the B1 level, and all of those items are currently available through the web interface with all of the features described in the next subsection. Work is ongoing and is expected to be completed through the B2 level in 2019.

3.4. Using the SMARTool: Additional features

The SMARTool interface provides access to the word forms and sentences. In each sentence, the relevant word form is highlighted in blue to make

it easy to spot. After the end of the sentence, there is a parse of the word form. For example, for *бизнесменов* the parse is given as "(Gen.Plur)." Next to the parse is a "?" that the user can mouse over to get the full name of the parse, if needed. In this case, it would be "Genitive Plural." After the parse, there is a speaker button that activates an audio rendering of the sentence. This audio rendering can be accessed in either a male voice or a female voice by making the appropriate selection above the sentence. Audio is provided via a text-to-speech synthesizer. While this solution may not always provide ideal renderings of intonation contours, it is very effective at delivering accurate placement of stress and accompanying vowel reduction, which are important for learners.[7] There is additionally a "Show translation" button that the user can click on to get the English translations of the sentences.

To use the SMARTool, one first needs to select the appropriate CEFR level. Thereafter it is possible to filter items in three different ways: search by topic, search by analysis, and search by dictionary. Alternatively, the user may choose "All Levels," in which case vocabulary from all levels is available through the filters.

3.4.1. Search by topic

The lexemes in the SMARTool are categorized according to eighteen topics inspired by the textbooks consulted: *внутренний мир* 'mental experience', *время* 'time', *еда* 'food', *животные/растения* 'animals/plants', *жильё* 'home', *здоровье* 'health', *люди* 'people', *магазин* 'shopping', *мера* 'measurement', *общение* 'communication', *одежда* 'clothing', *описание* 'description', *погода* 'weather', *политика* 'politics', *путешествие* 'travel', *свободное время* 'leisure', *транспорт* 'transportation', and *учёба/работа* 'study/work'. When the user selects "Search by topic," the menu of topics opens up, giving both the Russian and the English names for each topic. A given lexeme can appear with multiple topics; for example, *бизнесмен* 'businessman' is categorized with both *люди* 'people' and *учёба/работа* 'study/work'. When the user selects one of the topics, lexemes are represented one by one with sentences illustrating the use of their word forms. For example, if one selects Level A1 and the topic *люди* 'people',

[7] An alternative solution might have been to insert stress marks in the Russian example sentences. However, recent research shows that L2 learners of Russian derive very little, if any, benefit from stress marks; they just ignore them (Hayes-Harb and Hacking 2015). The only stress information given graphically in the SMARTool is the dieresis over ë as in *лётчик* 'pilot'.

the second word that appears is *бизнесмен* 'businessman', with the three Russian sentences using that word given in the examples cited above. When searching by topic, the user can move on to the next lexeme by clicking on the right-arrow (→) button and return to the previous lexeme by clicking on the left-arrow (←) button.

3.4.2. Search by analysis

Every word form in the SMARTool is tagged with a parse of the grammatical categories that it expresses. For nouns, this includes case and number, while adjectives can also express gender. The parse of verbs always includes aspect and can include person, number, tense, infinitive, imperative, gerund, and longer parses for participles (including their adjectival attributes). When using the "Search by analysis" function, the user views a menu listing the parse options. The user then chooses one item from the menu and gets an inventory of just the sentences with word forms with the chosen attributes. For example, if in Level B1 the user selects "Ins.Sing" for instrumental singular forms, in addition to the sentence with *балериной* 'ballerina', given above, the user receives sentences with other high-frequency instrumental singular forms, such as *кровью* 'blood', *лётчиком* 'pilot', *картошкой* 'potatoes', *гимнастикой* 'gymnastics', etc. Each sentence has all of the options for getting the English translation, audio rendering, and full description of the parse that are described under the "Search by topic" function described above. The "Search by analysis" function has already been found to have important pedagogical uses, since it allows users (including instructors) to instantly locate examples of lexemes that are frequently found in the given paradigm form. This can be useful, for example, when reviewing the meanings of the Russian grammatical cases and the use and form of difficult parts of the verbal paradigm, such as imperatives, participles, and gerunds.

3.4.3. Search by dictionary

When the user selects "Search by dictionary," a menu with the dictionary form of every lexeme at the given CEFR level appears. Lexemes are listed in Russian alphabetical order, and each lexeme is accompanied by an English equivalent. When the user selects an item from the menu, the three (or two or one) sentences illustrating the highest-frequency word forms of that lexeme appear with all the features (options to access audio, translation, and parse explanation) described above.

4. Conclusion

It is certainly the case that the authors of Russian textbooks have always tried to represent the word forms that L2 learners are most likely to encounter. However, today it is possible to realize this goal in a more precise manner by taking advantage of existing data on the authentic use of Russian word forms.

The SMARTool takes a usage-based approach to modeling Russian inflectional morphology. Inspired by research on the distribution and simulated learning of Russian word forms, the SMARTool strategically focuses the acquisition of a basic Russian vocabulary on the highest-frequency word forms and the contexts that motivate their use. In so doing, the SMARTool reduces the task of learning a basic vocabulary of about 3,000 lexemes by over 90%. While learning the entire paradigms of that many lexemes would entail mastery of over 100,000 word forms, with the SMARTool only about 9,000 word forms are needed. The SMARTool provides a variety of search options to support both lexical and grammatical approaches to the learning of vocabulary and morphology. Because the SMARTool is an online resource, it can be continually updated and expanded and can also be custom-tailored to excerpt specific vocabulary, for example, in connection with given lessons.

Acknowledgments

The SMARTool has been supported by the author's employer, UiT The Arctic University of Norway, and by grant number CPRU-2017/10027 from DIKU, the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (https://diku.no/en/about-diku).

The following individuals constitute the SMARTool team:

Radovan Bast (UiT The Arctic University of Norway): Design and Programming

Laura A. Janda (UiT The Arctic University of Norway): Background research, Concept, Design, Vocabulary selection, Editing of content

Tore Nesset (UiT The Arctic University of Norway): Concept, Design, Vocabulary selection

Svetlana Sokolova (UiT The Arctic University of Norway): Concept, Design, Vocabulary selection

James McDonald (UiT The Arctic University of Norway): Editing of content
Mikhail Kopotev (University of Helsinki): Design, Vocabulary selection
Francis M. Tyers (Indiana University): Background research, Concept, Design, Vocabulary selection
Ekaterina Rakhilina (Higher School of Economics in Moscow): Concept
Olga Lyashevskaya (Higher School of Economics in Moscow): Concept, Design, Vocabulary selection
Valentina Zhukova (Higher School of Economics in Moscow): Content
Evgeniia Sudarikova (Higher School of Economics in Moscow): Content

References

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

Andriushchina, N. P., G. A. Bitekhtina, L. P. Klobukova, L. N. Noreiko, I. V. Odintsova, eds. 2014–2015. *Leksicheskii mimimum po russkomu iazyku kak inostrannomu*. Levels A1–B2. St. Petersburg: Zlatoust.

Baayen, R. Harald. 1992. "Quantitative Aspects of Morphological Productivity." *In Yearbook of Morphology* 1991, edited by Gert E. Booij and J. Van Marle, 109–49. Dordrecht: Kluwer Academic Publishers.

———. 1993. "On Frequency, Transparency, and Productivity." *In Yearbook of Morphology* 1992, edited by Gert E. Booij and J. Van Marle, 181–208. Dordrecht: Kluwer Academic Publishers.

Bondar', Nataliia I., and Sergei A. Lutin. 2013. *Kak sprosit'? Kak skazat'?* 2nd ed. Moscow: Russkii iazyk.

Chernyshov, Stanislav. 2004. *Poexali!* St. Petersburg: Zlatoust.

Chun, Dorothy, Richard Kern, and Bryan Smith. 2016. "Technology in Language Use, Language Teaching, and Language Learning." *The Modern Language Journal* 100: 64–80. doi:10.1111/modl.12302.

Comer, William. 2019. "Measured words: Quantifying Vocabulary Exposure in Beginning Russian." *Slavic and East European Journal* 60 (1): 92–114.

deBenedette, Lynne, William J. Comer, Alla Smyslova, and Jonathan

Perkins. 2013. *Между нами (Between You and Me): An Interactive Introduction to Russian*. Accessed April 1, 2018. http://www.mezhdunami.org/.

Diachenko, Pavel V., Leonid L. Iomdin, A. V. Lazurskii, L. G. Mitiushin, Olga Iu. Podlesskaia, Victor G. Sizov, T. I. Frolova, and L. L. Tsinman. 2015. "Sovremennoe sostoianie gluboko annotirovannogo korpusa tekstov russkogo iazyka (SinTagRus)." *Sbornik «Natsional'nyi korpus russkogo iazyka: 10 let proektu». Trudy Instituta russkogo iayzka im. V.V. Vinogradova*.Vyp. 6: 272–99.

Endresen, Anna, Laura A. Janda, Robert Reynolds, and Francis M. Tyers. 2016. "Who Needs Particles? A Challenge to the Classification of Particles as a Part of Speech in Russian." *Russian Linguistics* 40 (2): 103–32. doi:10.1007/s11185-016-9160-2.

Golonka, Ewa M., Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, and Suzanne Freynik. 2014. "Technologies for Foreign Language Learning: A Review of Technology Types and Their Effectiveness." *Computer Assisted Language Learning* 27 (1): 70–105. doi:10.1080/09588221.2012.700315.

Hart, Betty, and Todd R. Risley. 2003. "The Early Catastrophe: The 30 Million Word Gap by Age 3." *American Educator Spring*: 4–9.

Hayes-Harb, Rachel, and Jane Hacking. 2015. "The Influence of Written Stress Marks on Native English Speakers' Acquisition of Russian Lexical Stress Contrasts." *Slavic and East European Journal* 59 (1): 91–109.

Hertz, Birgitte, Hanne Leervad, Henrik Lærkes, Henrik Møller, and Peter Schousboe. 2001. *Свидание в Петербурге. Møde i Petersborg*. Copenhagen: Gyldendal.

Janda, Laura A., and Francis M. Tyers. 2018. "Less is More: Why All Paradigms Are Defective, and Why that Is a Good Thing." *Corpus Linguistics and Linguistic Theory* 14 (2): 33. doi.org/10.1515/cllt-2018-0031.

Janda, Laura A. (Under submission). "Yggur and the Power of Language: A Linguistic Invention Embedded in a Czech Novel."

Kuznetsova, Julia. 2017. "The Ratio of Unique Word Forms as a Measure of Creativity." *In Each Venture a New Beginning: Studies in Honor of Laura A. Janda*, edited by Anastasia Makarova, Stephen M. Dickey, and Dagmar Divjak, 85–97. Bloomington, IN: Slavica Publishers.

Malouf, Robert. 2016. "Generating Morphological Paradigms with a Recurrent Neural Network." *San Diego Linguistic Papers* 6: 122–29.

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

Moreno-Sánchez, Isabel, Francesc Font-Clos, and Álvaro Corral. 2016. "Large-scale Analysis of Zipf's Law in English Texts." *PLoS One* 11 (1). e0147073. doi:10.1371/journal. pone.0147073.

Robin, Richard, Galina Shatalina, and Karen Evans-Romaine. 2012. *Golosa*. Vols 1–2. 5th ed. New York: Pearson.

Wade, Terence. 2011. *A Comprehensive Russian Grammar*. 3rd ed. Oxford: Wiley-Blackwell.

Zalizniak, A. A. 1980. *Grammaticheskii slovar' russkogo iazyka*. Moscow: Russkii iazyk.

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