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

Laura A. Janda, UiT The Arctic University of Norway

**Abstract**

Modern electronic corpora and computational techniques make it possible to discover which Russian wordforms are used most often and in which contexts. Furthermore a computational learning experiment suggests that learning of inflectional morphology is best when it is restricted to the highest frequency wordforms. The SMARTool (Strategic Mastery of Russian Tool) builds on these findings to offer a language learning resource that presents a basic vocabulary of 3,000 Russian inflected lexemes (nouns, verbs, and adjectives), along with their highest frequency wordforms illustrated by contexts typical of their use. Filters allow users to select wordforms according to proficiency levels, topics, grammatical categories, and dictionary forms. Users can access audio and English translations for all example sentences.

Keywords: language learning, vocabulary, inflectional morphology, corpus

**1 Introduction**

Included in the tasks facing a language learner are 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 wordforms 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 only a few thousand inflected lexemes has over a hundred thousand associated wordforms. Recent research (Janda & 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 learning of Russian paradigms either in their entirety or based only on the most frequent wordforms. Section 3 presents a free public net-based resource, the SMARTool (Strategic Mastery of Russian Tool), which takes up the challenge of providing strategic input for 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 have served as a model to me throughout my career, and are, I believe, realized also 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” vs. “usage-based model” of Russian wordforms**

On the face of it, paradigms seem to be rather straightforward tables listing all the wordforms that express the various grammatical categories associated with a given part of speech, as in Zaliznjak 1980. These tables can be called the “paradigm model” of inflectional morphology, which probably does not adequately represent the mental grammar of the language. In Russian, nouns express combinations of six cases and two numbers, yielding twelve wordforms, 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 (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 wordforms.

In its extreme form, the “paradigm model” was implicit in the traditional grammar + 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, <http://www.mezhdunami.org/>) 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, in the meantime 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 wordforms in natural language casts considerable 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 also have fewer than twelve forms, as in the case of nouns that are singularia tantum (such as *молодежь* ‘young people’), pluralia tantum (such as *ножницы* ‘scissors’), or have paradigmatic gaps (such as the Genitive Plural of *мечта* ‘dream’). Similar variations occur also 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 wordforms 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 wordforms

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 next most frequent word, *of*, is ½ as frequent as *the*. The third most frequent word, *and*, is 1/3 as frequent as *the*. Fourth comes *a*, which is ¼ 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 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 (non-linguistic) distributional phenomena. Zipf’s Law has a number of surprising entailments, for example that approximately 50% of the unique lexemes in any corpus are hapaxes,[[1]](#footnote-1) and that only 135 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 wordforms 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 wordforms too. I briefly elaborate on each of these facts below.

*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 5-10 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 in the course of a lifetime.

*Zipf’s Law scales up:* Scalability has been tested by Manning and Schütze (1999) and Moreno-Sánchez et al. (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 their native language.

*Zipf’s Law applies to wordforms too:* The Zipfian curve characterizes not just words, but all wordforms as well. This has two implications for paradigms: а) 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 wordforms, and this is borne out by the facts. For any given Russian lexeme of overall high frequency (≥50 per million words), one wordform is most frequent, a couple more might be attested regularly (accounting for >10% of attestations of the lexeme), and the remaining wordforms are rare or unattested (Janda and Tyers 2018). For example, *бизнесмен* ‘businessman’ is attested fifty times in the SynTagRus corpus[[2]](#footnote-2) of just over 1 million words. Sixteen of those attestations (32%) are of the genitive plural *бизнесменов*, ten attestations (20%) are of the nominative plural *бизнесмены*, seven (14%) are of the nominative singular *бизнесмен*, most other wordforms are rare, and three (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 wordforms attested, and recall that the presence of low-frequency lexemes expands proportionately with the size of a corpus.

The implications of Zipfian distribution of wordforms for the representation of full paradigms are even more surprising. Since one wordform in a paradigm will be of highest frequency, with the frequency of other wordforms 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).

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. Let’s take an example. The lexeme *тамада* ‘toastmaster’ has no attestations of dative plural or locative plural forms in the Russian National Corpus (<http://ruscorpora.ru/>; main corpus of 283,431,966 words as of April 2019), and it is likely that many native speakers have never encountered these wordforms. However, all native speakers of Russian can be expected to readily understand the forms *тамадам*, *тамадах* and to produce them in appropriate contexts.

In Janda and Tyers 2018 we provide statistical evidence that the wordforms in the paradigm of an inflected part of speech (in other words nouns or 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 wordforms. 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, and other nouns populate other parts of the space, with many nouns overlapping in their contributions to the space. In aggregate, the attestations of wordforms 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 wordforms 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 wordforms 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, representing the single most frequent wordform 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 wordform version, in which the learning task was to learn just the single highest frequency wordform and the lemma (dictionary) form. The production task was the same for both versions, namely: given the lemma (dictionary) form of a previously unseen lexeme and the parse set for that lexeme’s most frequent wordform, to predict the wordform. In other words, 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 the first iteration the training set was based on the 1-100 most frequent wordforms in SynTagRus, and the production set consisted of the 101-200 most frequent wordforms 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 wordforms 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 wordforms (and their paradigms for the full paradigm model), and the production task was based on the 201-300 most frequent wordforms 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 wordforms, while the single form model learned only 5,400 wordforms 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 does better than the single forms model, but both models are still quite poor, with 40% or fewer correct predictions. On iterations eleven through fifteen, the performance of the two models is similar, at about 45%-62% correct. Thereafter, for the remaining thirty-eight iterations, the single form model outperforms the full paradigm model every single time. The learning curve of the full paradigm model flattens out in the 60%-70% range, while the single form model performs in the 80%-95% range. In terms of average Levenshtein distance, when errors are made, in the first six iterations the full paradigm model makes less severe errors than the single form model, but both models perform rather poorly (average edit distance of >3 letters). In the seventh iteration the scores are nearly identical. After that, for all remaining iterations except one (iteration thirty-five), the single form model makes less severe errors when it does 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 wordforms consistently outperforms learning based on full paradigms both in terms of the accuracy of predictions of wordforms of previously unseen lexemes and in terms of the severity of errors. Learning full paradigms does not look like the most effective way to acquire Russian inflectional morphology – it might simply overpopulate the search domain to the point that producing wordforms gets harder rather than easier.[[3]](#footnote-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 wordforms according to Zipf’s Law means that only a fraction of wordforms of any given lexeme are encountered frequently, while the majority of wordforms are encountered rarely, and many wordforms may never be encountered. Different lexemes have different patterns of attested wordforms, 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 wordforms. Evidence from a computational learning experiment suggests that when learning focuses only on the most frequent wordforms, the ability to produce specific wordforms 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 wordforms 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 wordforms 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: Strategic Mastery of Russian Tool**

The SMARTool (Strategic Mastery of Russian Tool) 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 wordforms, the presentation of contexts of use, and additional features of the resource such as audio, translations, and filters.

Among technological resources for second language learning, the potential of corpora has not been fully realized largely because corpora 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 et al. 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 wordforms of 3,000 Russian lexemes, distributed across the first four CEFR (Common European Framework of Reference for Languages) Levels[[4]](#footnote-4) as displayed in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| CEFR Level | ACTFL Equivalent | Russian Equivalent | SMARTool number of lexemes |
| A1 “Beginner” | Novice Low-Mid | ТЭУ Элементарный уровень | 500 |
| A2 “Elementary” | Novice High | ТБУ Базовый уровень | 500 |
| B1 “Intermediate” | Intermediate Low-Mid | ТРКИ-1 I Cертификационный уровень | 1,000 |
| B2 “Upper Intermediate” | Intermediate High-Advanced Low | ТРКИ-2 | 1,000 |

*Table 1. Distribution of SMARTool lexemes across L2 acquisition levels*

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, Černyšov 2004, Robin et al. 2012, deBenedette et al. 2013, Bondar’ and Lutin 2013) plus the *Лексический минимум по русскому языку как иностранному* (Andrjušina 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]](#footnote-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 wordforms

The next task was to identify the highest-frequency wordforms associated with each lexeme. One challenge in this task is the presence of syncretism in Russian paradigms. For example, the form *радости* could potentially be any of five wordforms 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 to 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.

The selected lexemes were queried in the SynTagRus corpus to determine the frequency distributions of their wordforms, 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 wordforms that occur often, while the rest are rare or even unattested. For each lexeme, we selected the three most common wordforms. However, if over 90% of attestations for a given lexeme were accounted for by only two forms or by only one form, 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 wordforms 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 wordforms are represented in the SMARTool.[[6]](#footnote-6)

3.3 Identification of typical contexts

The next task in building the SMARTool was to determine, for every single wordform, what grammatical and lexical contexts were most typical. In other words, what grammatical constructions and lexical collocations motivate each wordform. For a few items the answer to this question is trivial, as in the case of *сентябрь* ‘September’, where 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 wordforms, this is 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 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 wordform, 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 when this article was written (April-June 2019), all of the most frequent wordforms had been identified for all lexemes at all four CEFR Levels (A1, A2, B1, B2), and example sentences had been supplied for all wordforms at the A1 and A2 levels, plus most of the wordforms 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 wordforms and sentences. In each sentence, the relevant wordform is highlighted in blue to make it easy to spot. After the end of the sentence, there is a parse of the wordform. 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 “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]](#footnote-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.

*Search by topic:* The lexemes in the SMARTool are categorized according to seventeen 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’, *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. А 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 wordforms. For example, if one selects Level A1 and the topic *люди* ‘people’, the second word that appears is *бизнесмен* ‘businessman’ with the three Russian sentences given in the examples cited here 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.

*Search by analysis:* Every wordform 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 additionally 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 wordforms 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* 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 “Search by topic” here 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.

*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 wordforms of that lexeme appear, with all the features (options to access audio, translation, parse explanation) as described above.

**4 Conclusion**

It is certainly the case that the authors of textbooks of Russian have always tried to represent the wordforms 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 wordforms.

The SMARTool takes a usage-based approach to modelling Russian inflectional morphology. Inspired by research on the distribution and simulated learning of Russian wordforms, the SMARTool strategically focuses the acquisition of a basic vocabulary of Russian on the highest frequency wordforms 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 wordforms, with the SMARTool only about 9,000 wordforms 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 continuously 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**

Andrjušina, N. P., G. A. Bitextina, L. P. Klobukova, L. N. Norejko, I. V. Odincova (eds.). 2014-2015. *Лексический минимум по русскому языку как иностранному.* Levels A1-B2. St. Petersburg: Zlatoust.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Gert E. Booij and 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 and J. Van Marle (eds.), *Yearbook of Morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

Bondar’, N. I. and S. A. Lutin. 2013. *Как спросить? Как сказать?*, 2nd ed. Moscow: Russkij jazyk.

Černyšov, Stanislav. 2004. *Поехали!* 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, no. 1, 92–114.

deBenedette, Lynne, William J. Comer, Alla Smyslova, Jonathan Perkins. 2013. *Между нами (Between You and Me): An Interactive Introduction to Russian*. <http://www.mezhdunami.org/> (accessed 1 April 2018).

Djačenko, P. V., L. L. Iomdin, A. V. Lazurskij, L. G. Mitjušin, O. Ju. Podlesskaja, V. G. Sizov, T. I. Frolova, L. L. Cinman. 2015. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус). *Сборник «Национальный корпус русского языка: 10 лет проекту». Труды Института русского языка им. В.В. Виноградова.* Вып. 6., 272–299.

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, 103-132. DOI 10.1007/s11185-016-9160-2.

Golonka, Ewa M., Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, 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 2003. 4–9.

Hayes-Harb, Rachel, 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, 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), 33pp. 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 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.

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

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

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.

Robin, Richard, Galina Shatalina, Karen Evans-Romaine. 2012. *Голоса* (Vols. 1-2), 5th ed. New York: Pearson.

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

Zaliznjak, A. A. 1980. *Грамматический словарь русского языка*. Moscow: Russkij jazyk.

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

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. [↑](#footnote-ref-1)
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 wordforms. For more about this corpus, see Djačenko et al. 2015. [↑](#footnote-ref-2)
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. [↑](#footnote-ref-3)
4. For more on CEFR Levels as established by the Council of Europe, see <http://www.coe.int/en/web/language-policy/home>. [↑](#footnote-ref-4)
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. [↑](#footnote-ref-5)
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’. [↑](#footnote-ref-6)
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 diaresis over ё as in *лётчик* ‘pilot’. [↑](#footnote-ref-7)