StimulStat: A lexical database for Russian

Svetlana Alexeeva[1] · Natalia Slioussar[1,2] · Daria Chernova[1]

Published online: 7 December 2017

Abstract In this article, we present *StimulStat* – a lexical database for the Russian language in the form of a web application. The database contains more than 52,000 of the most frequent Russian lemmas and more than 1.7 million word forms derived from them. These lemmas and forms are characterized according to more than 70 properties that were demonstrated to be relevant for psycholinguistic research, including frequency, length, phonological and grammatical properties, orthographic and phonological neighborhood frequency and size, grammatical ambiguity, homonymy and polysemy. Some properties were retrieved from various dictionaries and are presented collectively in a searchable form for the first time, the others were computed specifically for the database. The database can be accessed freely at http://stimul.cognitivestudies.ru.

Keywords Lexical database · Russian · Frequency · Neighborhood · Grammatical properties

Introducing the *StimulStat* database

Experimental studies of language identified a large list of lexical properties that play a role in speech production and comprehension, including lemma and word form frequency and length (e.g., New et al., 2006; Kliegl et al., 2004; Monsell, 1991; Rayner, 1998; Yap & Balota, 2015), the number of syllables (e.g., Ashby & Rayner, 2004; Carreiras & Grainger, 2004; Taft & Forster, 1976), stress pattern (e.g., Arciuli & Cupples, 2006; Colombo, 1992; Schiller, Fikkert, & Levelt, 2004), homonymy and polysemy (e.g., Beretta, Fiorentino, & Poeppel, 2005; Mirman et al., 2010; Rodd et al., 2004), grammatical characteristics, including part of speech, inflectional paradigm, etc. (e.g., Baayen, Dijkstra, & Schreuder, 1997; Taft, 1979), different properties of orthographic and phonological neighborhoods (e.g., Adelman et al., 2013; Andrews, 1997a, b; Perea, 2015), etc. Preparing stimuli for a psycholinguistic experiment usually requires taking many of these characteristics into account at once: selected items should differ with respect to the factors of interest, but should be closely matched with respect to other relevant properties. This task might be very difficult to accomplish without searchable lexical databases that include various characteristics for a large number of words.

Such databases have been created for several languages and are available in the form of a web application or computer software. Among them are the English lexicon project (Balota et al., 2007), eDom (Armstrong, Tokowicz, & Plaut, 2012), N-Watch (Davis, 2005) and MRC database (Coltheart, 1981) for English; DlexDB for German (Heister et al., 2011); Lexique (New, Pallier, Brysbaert, and Ferrand, 2004) for French; EsPal (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) and BuscaPalabras (Davis & Perea, 2005) for Spanish; EHME (Acha, Laka, Landa, & Salaburu, 2014) and E-Hitz (Perea et al., 2006) for Basque; GreekLex (Ktori, van Heuven, & Pitchford, 2008) and GreekLex2 (Kyparissiadis et al., 2017) for Modern Greek; Aralex (Boudelaa & Marslen-Wilson, 2010) for Modern Standard Arabic; the Malay Lexicon Project (Yap, Liow, Jalil, & Faizal, 2010) for Malay; KelemetriK (Erten, Bozsahin, & Zeyrek, 2014) for Turkish;

✉ Svetlana Alexeeva
mail@s-alexeeva.ru

[1] St. Petersburg State University, St. PetersburgGalernya 58/60, 190000, Russia

[2] National Research University Higher School of Economics, Moscow, Russia

the Brazilian Portuguese Lexicon (Estivalet & Meunier, 2015) for Brazilian Portuguese, etc. All these databases are equipped with effective search and filtering tools.

In this article, we present *StimulStat* – the first lexical database of this type created for Russian. Among distinctive features of Russian are its rich inflectional and derivational morphology, a complicated system of inflectional paradigms, a flexible stress, and a Cyrillic alphabet.[1] Russian has a rich lexicographic tradition, with the most recent projects relying on large corpora, primarily on the Russian National Corpus (www.ruscorpora.ru). However, while some dictionaries are available electronically, the others are not, and no existing resource allows combining information from different sources. Moreover, certain characteristics relevant for psycholinguistic research, for example, orthographic neighborhood properties, are not represented in any dictionary or database at all.

The *StimulStat* database addresses these problems, providing an effective tool for conducting psycholinguistic research. It includes more than 52,000 lemmas and more than 1.7 million word forms derived from them. *StimulStat* is available as a web application and allows searching for lemmas and word forms with particular properties, as well as retrieving required properties for a predefined list of items.

Sources used in the database

The *StimulStat* database contains 52,139 lemmas included in the *Frequency dictionary of modern Russian language* (Lyashevskaya & Sharov, 2009). This dictionary is based on the subcorpus of the Russian National Corpus (www.ruscorpora.ru) containing 92 million words. Lemma frequency values and part of speech tags were taken from this dictionary. The Russian National Corpus includes 150 million words; the collection of texts dating from XVIII to XXI century is balanced by genre and style. The 92 million-word subcorpus the dictionary is based on includes texts dating from 1950–2007. It is considered to be the most representative frequency dictionary of modern Russian.

Grammatical characteristics of lemmas were defined using the morphological parser *Pymorphy2* (Korobov, 2015), which relies on the *OpenCorpora* morphological dictionary (Bocharov et al., 2013). The *OpenCorpora* dictionary is distributed under the free content license of Creative Commons. Given this advantage and the considerable size of the dictionary (389,232 lemmas and 5,097,247 word forms),[2] we preferred the *Pymorphy2* parser to *Mystem2* (Segalovich, 2003), another morphological parser widely used for Russian. It was originally developed as an internal commercial tool, and its morphological dictionary is not publicly available.

Information about stress position and different properties of inflectional paradigms was taken from the *Grammatical dictionary of Russian language* (Zaliznjak, 1987). This dictionary contains more than 100,000 lemmas and gives a full morphological description for each lemma. Most morphological parsers developed for Russian are based on the electronic version of this dictionary. However, it does not include uninflected parts of speech, i.e., adverbs, prepositions, conjunctions, particles, and interjections, as well as some novel words. Thus, for a number of lemmas included in *StimulStat* we had to define the stress position and inflectional properties individually. This was done by two authors of the project who have a linguistic background.[3] In general, an important part of our project was to align different sources used in the database. This could be done automatically in the majority of cases, but still, a lot of manual checking was required.

The Explanatory dictionary of Modern Russian (Efremova, 2000) was used to extract information about polysemy and homonymy. As is well known, deciding on the number of meanings a word has and differentiating between polysemy and homonymy is often difficult. For example, two meanings of one word can go very far apart in the process of diachronic change, but it is a matter of controversy when the relation between them becomes so obscure that it makes sense to identify two homonymous words. However, since these problems cannot be avoided, we chose Efremova's dictionary as the largest electronically available explanatory dictionary of Russian – it contains more than 120,000 lemmas.

The data on the subjective age of acquisition and imageability were drawn from the database *Verb and action* (Akinina et al., 2014) that contains 375 verbs. The number of verbs is relatively small, but the goal of *StimulStat* is to aggregate all available reliable resources that could be useful to psycholinguists and to allow using them simultaneously. Subjective parameters were shown to be relevant in many experimental studies (e.g., Bates et al., 2003), so it makes sense to include them for as many lemmas as is currently possible.

We used the morphological parser *Pymorphy2* (Korobov, 2015) to generate all forms from the lemmas in the database and to specify their grammatical features. Forms are included in the database as separate items (one can search for forms or for lemmas) and are linked to lemmas. In total, we have 1,700,842 word forms. For 355,935 forms, frequency values were extracted from the database of the corpus-based project *Frequency grammar of Russian* (http://web-corpora.net/freaky_frequency/freq_main.html) (Lyashevskaya, 2013).

[1] There are several ways to transliterate Russian words into the Latin alphabet. Providing examples in this paper, we use the so-called scholarly transliteration system.

[2] The numbers were taken from the project website (http://opencorpora.org/dict.php) on 10 April 2017.

[3] Russian has flexible stress, and in some cases, the norm is unstable, so we need to mention that there were no such controversial cases among the words we had to annotate manually.

This is the only resource that provides frequency information for morphologically disambiguated word forms in Russian.

Finally, we used the *CORPRES* dictionary of phonological variants created at the Laboratory for Experimental Phonetics of St. Petersburg State University (Skrelin et al., 2010). The dictionary is based on the *CORPRES*, which includes 60 h of recorded speech: texts of different genres pronounced by eight speakers. The corpus contains more than 100,000 word forms with two types of phonemic transcription taking some allophonic variation into account: so-called *ideal* transcription (generated automatically based on the existing conventions for standard Russian and then manually checked) and *real* transcription (generated manually and reflecting how a given form was actually pronounced by recorded speakers). A full list of symbols used in transcription and their description can be found here: http://stimul.cognitivestudies.ru/ru_stimul/phoneme_notation/.

For example, the word *leto,* "summer," has one ideal transcription /l' e0 t a4/ and three real transcriptions associated with it in the dictionary: /l' e0 t a4/, /l' e0 t e4/ and /l' e0 t y4/ . Real transcriptions differ in the quality of the post-stressed vowel. In total, there are 9,965 unique pairs of word forms and their ideal phonemic transcriptions and 26,778 unique pairs of word forms and their real phonemic transcriptions.

Russian orthography is relatively transparent, so transcriptions are not provided in the dictionaries, except for small dictionaries intended for beginner L2 learners. However, it is not the case that transcriptions can be easily derived from orthographic representations in a rule-based fashion. Firstly, Russian has a morphologically-based type of orthography: in most cases, different realizations of a morpheme have the same spelling even when they are pronounced differently. Secondly, stress position that influences vowel quality cannot be predicted from the orthographical representation.

We chose to rely on the *CORPRES* dictionary because there is no publicly available spelling-to-sound converter for Russian. It also has an important advantage over the rule-based approach taken, for example, in the *Espal* (Duchon et al., 2013) and the *Malay Lexicon* project (Yap et al., 2010). It reflects not only the pronunciation of isolated words, but also captures phonetic phenomena at word boundaries, like progressive assimilation of voice that is very widespread in Russian. To give an example, the word *vopros,* "question" is pronounced as /v a1 p r o0 s/ (ideal transcription) in isolation. But if it precedes a word starting with a voiced obstruent, like in *vopros zadannyj,* "question asked," the last phoneme would be /z/: /v a1 p r o0 z/. Both transcriptions are represented in the *StimulStat* database.

Available information

The database contains 52,139 lemmas and 1,700,842 word forms derived from them. 451 lemmas are morphologically ambiguous: for example, *dobro* is a noun meaning "good, welfare," a particle "deal, granted," and an adverb "amicably, tenderly." Thus, the number of orthographically unique lemmas is 51,688. The number of orthographically unique forms is only 963,257, due to widespread syncretism: many forms are morphologically ambiguous. For example, *koške* can be a dative singular or a locative[4] singular from the noun *koška,* "cat.".

Frequency information

StimulStat provides information about frequency measured in ipm (instances per million) for all lemmas and for 355,935 word forms, out of them 252,091 orthographically unique ones. We also calculated ln-transformed and lg-transformed frequency values because there is a logarithmic relationship between word frequency and reaction time to this word during lexical access (e.g., Duyck, Desmet, Verbeke, & Brysbaert, 2004; Keuleers et al., 2012; Kinoshita, 2015; Kliegl, Grabner, Rolfs, & Engbert, 2004; Kliegl, Nuthmann, & Engbert, 2006; Monsell, Doyle, & Haggard, 1989; Oldfield & Wingfield, 1965). In addition to that, we computed different frequency measures for ideal and real transcriptions included in the database. Since the number of these transcriptions is relatively small so far, these statistics are mostly useful to determine which phonological variants of high frequency words are more widespread.

Information based on orthographic and phonological representation

First of all, *StimulStat* can provide ideal or real phonological representations paired with a given orthographic representation (only for word forms because phonological representations are associated with word forms). This information and other characteristics relying on phonological representations can be obtained only if the form in question is included in the CORPRES dictionary (Skrelin et al., 2010). We also calculated various parameters for all representations included in the database. When using these parameters in the search, one should specify which representations – orthographic, ideal, or real phonemic – to rely on.

For all lemmas and word forms, *StimulStat* provides information about length (in letters and in phonemes) and so-called uniqueness point. This is the letter/phoneme position reading from left to right that distinguishes a word from all other words (Marslen-Wilson & Tyler, 1980). The uniqueness point was shown to be relevant for psycholinguistic research; for example, this factor affects naming and lexical decision

[4] The tag "prepositional" is used instead of "locative" in some morphological descriptions of Russian.

latencies (e.g., Kwantes & Mewhort, 1999; Lindell, Nicholls, & Castles, 2003).

Among other supplementary parameters are the first and last letter/phoneme of the word, and its reversed orthographic and phonological representation (e.g., *okolom* for *moloko,* "milk"). The reversed representation is useful for experiments dealing with morphology. One cannot directly search *StimulStat* for words with a particular affix because the database does not provide morphological segmentation. However, one can select the pool of words satisfying other relevant parameters (e.g., frequency, length, etc.) and then sort them by their reversed representation. Then the words with the same affixes will be grouped together. For this reason, the *Grammatical dictionary of Russian language* (Zaliznjak, 1987) relies on reversed orthographic representation.

Modern Russian alphabet has 33 letters, but one of them, *ë*, is often substituted for *e* both in books and other printed production and in handwriting. Words with these two letters are pronounced differently, but it is easy for an advanced reader to recover this information in the absolute majority of cases, unless the word is an infrequent proper name etc., in which cases *ë* would be used much more consistently. The database takes this into account: all orthographic parameters including information about neighbors can be computed assuming that *ë* is a separate letter or that it coincides with *e*. Many sources we relied on do not use *ë*, so we had to insert it.

StimulStat has information on syllable structure that can be computed based on the orthographic, ideal and real phonological representation. It is more sensible to rely on phonological representations in this case, but not all words in the database have transcriptions, so all options were realized. *StimulStat* includes word length in syllables (the number of syllables is computed based on the number of vowels), information about syllable boundaries and the CV notation. Information about syllable boundaries is provided in the following form: e.g., *2_4* for *moloko,* "milk," indicating that the boundaries are after the second and the fourth symbol. Syllable boundaries are a matter of controversy in Russian linguistics. We relied on the approach developed by Bondarko (1977), according to which Russian syllables are always open except for terminal syllables ending in a consonant and for non-terminal syllables ending in /j/. It is supported by strong experimental evidence (Bondarko, 1977).

In the CV notation, *V* stands for a vowel, *C* for a consonant, and *F* denotes the letters ь and ъ called *soft* and *hard sign*, which are neither vowels nor consonants. The soft sign signals that the preceding consonant is palatalized, and, if it is followed by a vowel, that /j/ is pronounced between this consonant and this vowel. The hard sign indicates that the consonant is not palatalized despite the following front vowel and that /j/ is pronounced between this consonant and this vowel. Thus, the symbol *F* is used only if the CV notation is computed on the basis of orthographic representation: the soft and hard sign influence phonemic transcription, but do not correspond to any phonemes.

The database also contains information about the main and additional stress position: on which vowel or on which syllable counting from left to right the stress falls. For example, in the word *more,* "sea," the stress falls on the vowel in the first syllable (it is underlined). So the stress position in symbols is *2*, and the stress position in syllables is *1*. For lemmas, it is also indicated whether there is a stress shift in the inflectional paradigm. For example, the word *ruka,* "hand," has it: the stress falls on the ending in nominative singular and on the root in some other forms, like the accusative singular form *ruku*. The word *strana,* "country," has no stress shift, for example, its accusative singular form is *stranu*.

Grammatical information

StimulStat provides information about parts of speech and different grammatical features for lemmas and forms, including gender, number, person, case, animacy, tense, mood, aspect, voice, transitivity, and comparative and superlative degrees. It is also specified whether a given verb form is finite or not, and, in the latter case, whether it is an infinitive, participle or gerund, and whether an adjective or participle form is short or full. These two types of forms have different morphological and syntactic properties in Russian.

Two approaches to parts of speech are represented in the database. The first is adopted in the *Frequency dictionary of modern Russian language* (Lyashevskaya & Sharov, 2009) and the Russian National Corpus (www.ruscorpora.ru). It distinguishes nouns, verbs, adjectives, adverbs, cardinal and ordinal numbers, pronominal nouns, adjectives and adverbs, as well as prepositions, conjunctions, particles, and interjections. The second approach is adopted in the *OpenCorpora* morphological dictionary (Bocharov et al., 2013) and relies primarily on the inflectional characteristics. According to this, ordinal numbers, pronominal adjectives and pronominal adverbs are not separate parts of speech because they do not differ from other adjectives and adverbs with respect to their inflectional properties. At the same time, short forms of adjectives, non-finite verb forms, and comparatives form separate groups.

It is possible to search for various grammatical characteristics separately, and the full list of grammatical features and the full inflectional paradigm can be requested for every item. The database also includes inflectional indices from the *Grammatical dictionary of Russian language* (Zaliznjak, 1987). These indices were introduced to capture different properties of paradigms: inflectional classes, the presence or absence of consonant and vowel alternations, stress shifts, etc. For lemmas, *StimulStat* also provides grammatical features of the citation form, and for forms, the lemma can be found.

Orthographic and phonological neighborhood characteristics

Neighborhood characteristics have not been addressed in any previous work on Russian, so calculating them was an important part of our project. We will describe orthographic neighborhoods first and then will turn to phonological ones. Different properties of orthographic neighborhoods were demonstrated to play a role in a variety of reading tasks, including lexical decision, naming, perceptual identification, and semantic categorization. Several types of orthographic neighbors have been identified:

- Substitution neighbors, or *sns* (e.g., Coltheart et al., 1977). These are words obtained by changing one letter in a given word (in any position) while preserving the other letters, for example, *syn,* "son" – *syr,* "cheese."
- Transposition neighbors, or *tns* (e.g., Andrews, 1996; Perea & Lupker, 2003). These are words that share the same letters, but the positions of two of them are interchanged. These letters can be adjacent, as in *setka,* "net" – *sekta,* "sect," or not, as in *buk,* "beech" – *kub,* "cube."
- Addition and deletion neighbors, or *ans* and *dns* (e.g., Davis, Perea, & Acha, 2009). A deletion neighbor of a word is a letter string that differs from it by deletion of a single letter (in any position), and an addition neighbor is a string with an extra letter in any position. For example, *karta,* "map, card" is an addition neighbor of *kara,* "penalty," and *kara,* "penalty" is a deletion neighbor of *karta,* "map, card."
- Subset and superset neighbors, or *pns* and *wns* (e.g., Bowers, Davis, & Hanley, 2005). A subset (part) neighbor of a given word is a letter string embedded within this word. A superset (whole) neighbor is a letter string that contains the given word. For example, *sort,* "sort," is a superset neighbor of *sor,* "litter," and a subset neighbor of *sortirovat,* "to sort." When subset and superset neighbors were computed for the *StimulStat* database, we did not take words that are shorter than three letters into account.
- Bigram and trigram neighbors, or *bins and trins* (e.g., Davis, 2005). A bigram neighbor of a word is a letter string that shares with it a bigram (two successive letters) in the same position. Trigram neighbors share three successive letters in the same position. For example, *spina,* "back," is a bigram neighbor of *volna,* "wave," whereas *volk,* "wolf,"is its trigram neighbor.

We identified orthographic neighbors for all lemmas and word forms in the database. In addition to that, for every neighborhood, the number of words in it and their summed frequency (also ln-transformed and log-transformed) was calculated. *StimulStat* also provides information about the most frequent and the least frequent word in every neighborhood, and the number of neighbors that are more frequent than the given word. For transposition neighbors, there is a parameter showing whether the transposed letters are adjacent or not.

We calculated the same neighborhood parameters for real phonological representations included in the database. It is important to keep in mind that for many forms, we only have an orthographic representation, so the data set we rely on is smaller in this case. Notably, many word forms have different phonemic realizations that can be classified as phonological neighbors (for example, with a voiced and a voiceless final consonant – its realization in the connected speech depends on the following word). We decided not to count different realizations of one and the same form as neighbors: only realizations of different word forms were taken into account. To give an example, /v a1 p r o0 s/ and /v a1 p r o0 z/ are different realizations of the word *vopros,* "question," so they are not counted as neighbors.

Homonymy, homography, and morphological ambiguity

StimulStat provides various information about lemmas and forms that have the same spelling, but differ in other properties. Firstly, we relied on the *Explanatory dictionary of Modern Russian* (Efremova, 2000) that tags lemmas having homonyms and homographs.[5] However, this dictionary does not differentiate between three following options. Homonyms and homographs may (i) have the same grammatical properties (like *bor,* "pine forest" and *bor,* "(dental) drill"); or (ii) differ in some of them, for example, in animacy, which influences the choice of case endings in Russian (like *operator,* "operator, mechanic, camera man" or "operator, abstract function, statement (in programming)"); or (iii) even belong to different parts of speech (like *zlo*, which can be a noun "evil" or an adverb "in an evil way").

Homonyms of the first type are not differentiated in any other source used in *StimulStat*, but homonyms of the second and third type and all types of homographs can be identified in the pool of lemmas included in the database.[6] *StimulStat* represents these results and the results based on Efremova's dictionary separately. An additional reason to do so is that fact that Efremova's dictionary is relatively conservative, so it does not contain many lemmas included in *StimulStat.* Of

[5] Homographs are words that are orthographically the same, but have different stress, e.g., *zamok,* "castle" – *zamok,* "lock" (the stressed vowel is underlined).

[6] As we explained in the beginning of the paper, the relevant part of the database (the pool of lemmas with grammatical characteristics and stress) relies on several sources and on our own work synchronizing these sources – we had to identify stress position and grammatical characteristics for a number of lemmas. Thus, we cannot identify one particular source we rely on here.

course, the opposite is also true: many archaic, dialectal, and simply infrequent words covered by this dictionary are not included in *StimulStat.*

Word forms that have the same spelling can also coincide or differ with respect to the stress position. Obviously, the crucial parameter for such forms is whether they belong to the same lemma or not. Accordingly, StimulStat allows searching for orthographically identical forms that (i) belong to one lemma and have the same stress (e.g., *koške* is a dative singular or a locative singular from the noun *koška,* "cat"); (ii) belong to one lemma and have different stresses (e.g., *ruki* is a genitive singular and *ruki* is a nominative plural from the noun *ruka,* "hand"); (iii) belong to different lemmas and have the same stress (e.g., *bystro* is a neuter short form from the adjective *bystryj,* "quick" or an adverb "quickly"); (iv) belong to different lemmas and have different stresses (e.g., *tušu* is an accusative singular from the noun *tuša,* "hand" and *tušu* is a first-person singular present tense form from the verb *tušit,* "to extinguish").

Semantic information and subjective parameters

We provide information about polysemy: the number of meanings the word has according to the *Explanatory dictionary of Modern Russian* (Efremova, 2000). Obviously, the dictionary also contains the definition of every meaning, but we did not include this information. We specify whether the word is an abbreviation or a proper name (both in general and in particular a first or a last name, a patronymic or a place name). For 375 verb lemmas, we provide mean values and standard deviations of so called subjective parameters: the age of acquisition and imageability based on (Akinina et al., 2014).

Technical specifications and the web interface

We used Python scripts to extract and compute all parameters mentioned above. The output of the scripts were several lists, including two main lists: one for lemmas and another one for word forms. To make the database available as a web application, we imported these lists with linguistic parameters to a PostgreSQL database. The web interface was created using Django web application library.

The website http://stimul.cognitivestudies.ru has four pages (in English and in Russian). The title page contains a description of the database, a user manual, and references to all external sources used in the project. Another page contains additional materials from an independent project: information about frequencies of different grammatical features and inflectional affixes in Russian nouns. The other two pages are for searching the database.

Firstly, it is possible to look for lemmas and word forms with certain characteristics. For all numeric parameters, =, < and > signs are available, so one can search for exact values or for a particular range. Secondly, *StimulStat* can supply selected characteristics for a predefined list of lemmas or forms. Lemmas or forms can be typed into a search field or uploaded as a list in a *.txt or *.csv file (in utf-8 encoding). The output will appear on a separate web page and can be downloaded as a *.csv file.

An overview and cross-linguistic comparisons

For some of the parameters included in *StimulStat*, we computed average values and the range of possible values. The results are presented in Table 1, except for orthographic neighborhood characteristics, which will be discussed below. Calculations were done separately for all lemmas in the database, for word forms with frequency values and for all word forms generated from the lemmas included in the database. In addition to that, when a certain parameter, like the average lemma length in letters, is calculated, lemma frequency can be taken into account. The average length of all lemmas in *StimulStat* is 9.1, but more frequent lemmas tend to be shorter, and the average length corrected for frequency is 5.5. In the first case, we rely on the number of words that consist of one, two, three and more letters. In the second case, we rely on the summed frequencies of these words.

Several papers discussing lexical databases created for other languages also report average values of different parameters, but a cross-linguistic comparison is often complicated by various differences in database sources. The CLEARPOND project (Marian et al., 2012) aims to overcome this problem. It relies on the databases for five languages: English, French, German, Dutch, and Spanish, which are based on movie subtitle corpora. The databases are of the same size: they contain the most frequent 27,751 word forms encountered in the corpus of the relevant language. To arrive at this number, the authors took word form frequencies for every corpus and excluded the forms whose frequency was lower than 0.34 ipm. The list of remaining forms was the shortest in the English corpus: it contained 27,751 items, so this number was taken as a threshold for all five languages.

The CLEARPOND project reports the following average form frequency values: 32.6 ipm for Dutch, 32.7 ipm for English, 30.9 ipm for French, 33.7 ipm for German, and 33.9 ipm for Spanish. The values presented for Russian in Table 1 are dramatically different, but the size of the *StimulStat* database is much bigger. For the sake of comparison, we recalculated the values of several parameters for 27,751 most frequent word forms. The resulting average form frequency is 29.4 ipm (*SD* = 379.5, range: 3.2–38,107.4). This is very close to the values reported by Marian et al., (2012), especially taking into account that CLEARPOND databases are based on movie subtitles, while the *Frequency grammar of*

Table 1 The properties of lemmas and word forms in the *StimulStat* database

	Lemmas			Word forms that have frequency information			All word forms		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Length in letters	9.1 (5.5)[a]	3.2 (3.3)[a]	1–34	9.2 (5.4)[a]	2.8 (3.2)[a]	1–31	10.4	2.9	1–34
Length in syllables[b]	3.5 (2.1)[a]	1.4 (1.3)[a]	0–15	3.9 (2.3)[a]	1.3 (1.4)[a]	0–13	4.5	1.5	0–15
Frequency	18.5	291.9	0.4–35,801.8	3.8	126.2	0.2–38,107.4	N/A	N/A	N/A
Uniqueness point[b]	7.2	2.6	2–21	N/A	N/A	N/A	10.5	2.8	2–32
Stress position in letters[c]	5.4 (3.2)[a]	2.4 (2.0)[a]	1–21	5.2 (3.6)[a]	2.3 (2.0)[a]	1–17	5.5	2.2	1–21
Stress position in syllables[c]	2.4 (1.5)[a]	1.1 (0.8)[a]	1–9	2.3 (1.7)[a]	1.0 (0.8)[a]	1–8	2.4	1.1	1–9

[a] If word frequencies are taken into account

[b] We decided not to calculate uniqueness point values for the subset of forms that have frequency values

[c] Some Russian words (prepositions and particles) consist of a single consonant and therefore contain no syllables. They were not taken into account when stress position was calculated

Russian project (Lyashevskaya, 2013) *StimulStat* relies on is based mainly on fiction and newspaper texts. The average frequency of 40,481 forms in the English Lexicon Project (Balota et al., 2007) is 29.7. This project relies on frequency measures from Kučera and Francis' frequency list (Kučera & Francis, 1967), which was based on fiction and newspaper texts.

Information about average lemma frequency is available for Greek: 33.9 ipm (Ktori, van Heuven, & Pitchford, 2008). The figure in Table 1 is much lower: 18.5 ipm. However, the GreekLex database is considerably smaller than *StimulStat*, it contains 35,304 lemmas. If only 35,304 most frequent lemmas in *StimulStat* are taken into account, the average lemma frequency equals 26.8 ipm (*SD* = 352.9, range: 1–35,801.8).

Now let us turn to the average form length in symbols. The following values are reported in the CLEARPOND project: 8.4 for Dutch, 7.3 for English, 7.9 for French, 8.3 for German, 7.9 for Spanish. The values in Table 1 are larger both for forms with frequency information and for all forms. However, if only 27,751 most frequent forms are taken, the average length is 7.6 (*SD* = 2.5, range: 1–24). The average form length in the English Lexicon Project (Balota et al., 2007) is 8.0. Thus, the popular belief that words tend to be longer in Russian than in Germanic and Romance languages is not supported.

The paper on the GreekLex database (Ktori, van Heuven, & Pitchford, 2008) reports 9.0 and 5.7 as the average lemma length (the second figure is corrected for frequency). The figures in Table 1 are 9.1 and 5.5, but if we select a subcorpus of the same size as GreekLex, they will be smaller: 8.7 (*SD* = 3.0, range: 1–31) and 5.4 (*SD* = 3.3, range: 1–31).

The average form length in syllables can be compared to the data presented in the paper on the Malay Lexicon Project

Table 2 The number of orthographic neighborhoods (N) and words that comprise them for lemmas and word forms included in the *StimulStat* database

Neighbor- hood type	Orthographically unique lemmas (51,688)		Orthographically unique word forms (963,257)	
	Number of Ns	Number of words[a]	Number of Ns	Number of words[a]
sns	12,280	15,241 (29.5 %)	704,072	819,965 (85.1 %)
tns	642	1,130 (2.2 %)	17,657	30,718 (3.2 %)
ans	2,409	5,077 (9.8 %)	311,694	648,223 (67.3 %)
dns	3,346	5,077 (9.8 %)	510,193	648,223(67.3 %)
pns	44,592	47,380 (91.7 %)	N/A[b]	N/A[b]
wns	14,175	47,380 (91.7 %)	N/A[b]	N/A[b]
bins	6,636	51,645 (99.9 %)	8,950	963,227 (99.9 %)
trins	25,610	51,175 (99.0 %)	51,516	962,835 (99.9 %)

[a] The percentage of words in the database that have neighbors of a certain type is indicated in parentheses

[b] The information about bigram and trigram neighbors (*bins* and *trins*) for all word forms is not stored in the database, unlike in the other cases, such neighbors are calculated online for every query about individual forms

(Yap et al., 2010). In addition to Malay, this paper discusses four other languages, but the statistics reported for them are derived from the databases of different sizes:[7] 3.0 for Malay (corpus size: 9,592), 2.5 for French (corpus size: 38,335), 2.5 for English (corpus size: 38,477), 3.4 for German (corpus size: 50,658), and 3.5 for Dutch (corpus size: 117,867).

The figure in Table 1 is larger (3.9), but if we select subcorpora of the same sizes, the values will be 2.8, 3.2, 3.2, 3.3, and 3.5, respectively. Thus, the average form length in syllables in Russian is similar to German and Dutch, while English and French tend to have less syllables per form. Presumably, this is due to the fact that in both languages, diphthongs and letters that are not pronounced, i.e. do not correspond to any phonemes, are quite frequent. The figure in Malay is only slightly larger than in Russian, and it is difficult to speculate about the reasons. Apparently, the average form length is slightly larger in Malay, and it also tends to have open syllables.

Average values for the uniqueness point are not discussed for other databases, so no cross-linguistic comparison is possible. As for the stress position, average values can be found only for the GreekLex database (Ktori, van Heuven, & Pitchford, 2008). However, they cannot be directly compared to our data because the stress position is calculated from the end of the word, not from the beginning – this makes sense because in Greek, the stress can fall only on one of the three final syllables.

Now let us turn to the characteristics of orthographic neighborhoods, which we analyzed in more detail because they have not been discussed for Russian before. Table 2 presents the number of orthographic neighborhoods of different types identified in *StimulStat*, as well as the number and the percentage of lemmas and forms in the database that are included in these neighborhoods. Tables 3 and 4 show how many neighbors of a certain type a lemma or a form has on average.

Table 3 provides the numbers for all orthographically unique lemmas and forms included in *StimulStat*. This demonstrates how widespread a certain type of neighbors is. Table 4 presents similar calculations, but only for lemmas and forms that have neighbors of the relevant type: for example, we calculated how many lemmas or forms are included in every substitution neighborhood. This shows the average size of neighborhoods of different types.

The data on substitution neighbors in Table 3 can be compared to the results obtained in the CLEARPOND project. As before, to have a valid comparison, we recalculated the values for a subcorpus including 27,751 most frequent word forms in *StimulStat*. The average number of substitution neighbors per form is 3.1 in Table 1, and for this subcorpus, it is 1.6 (*SD* = 1.9, range: 0–17). It is similar to 1.5 reported for Spanish, higher than in Dutch, French, and German (all about 1 on average), but smaller than in English (about 2 on average).

Table 3 The number of orthographic neighbors of different types per lemma or word form in *StimulStat* (all lemmas and forms in the database are taken into account)

	Lemmas			Word forms		
	Mean	SD	Range	Mean	SD	Range
sns	0.8	2.0	0–28	3.1	2.5	0–41
tns	0.02	0.2	0–2	0.03	0.2	0–4
ans	0.07	0.5	0–29	0.6	1.2	0–25
dns	0.07	0.3	0–3	0.6	0.7	0–5
pns	2.5	2.0	0–17	N/A[a]	N/A[a]	N/A[a]
wns	2.5	32.4	0–3,463	N/A[a]	N/A[a]	N/A[a]
bins	433.1	551.9	0–3,098	8,520.4	12,444.4	0–82,759
trins	102.6	197.6	0–1,178	1,974.1	4,018.9	0–30,395

[a] The information is not available because the relevant parameters are calculated online for every query and are not stored in the database

Notably, in the database of the English lexicon project (Balota et al., 2007), which contains 40,481 forms, the average number of substitution neighbors per form is 1.2. In *StimulStat*, the average number of neighbors always decreased when a smaller subcorpus was taken. This points at interesting cross-linguistic differences beyond easily comparable average numbers. Exploring them is beyond the scope of this paper, so we can only provide an informal observation. As far as we can judge, the absolute majority of neighbors in English have different roots. Many neighbors in Russian have different affixes, while the root is the same. For example, prefixes are very widespread, and many prefixes differ by one letter: *za-* and *na-*, *do-* and *po-*, *po-* and *pod-*, *v-* and *vy-* etc.[8]

Some cross-linguistic differences in neighborhood properties have already been explored, primarily for transposition neighbors, and demonstrated to be psycholinguistically relevant. For example, Frost (2012, 2015) reviewed priming effects across writing systems to conclude that they crucially depend on the frequency of such neighbors, which can be very different in different languages.[9] A new model of reading was suggested based on these findings.

The average number of addition and deletion neighbors in Table 1 is 0.6. For the 27,751 form subcorpus, the numbers are 0.3 (*SD* = 0.9, range: 0–15) and 0.3 (*SD* = 0.5, range: 0–4) respectively. These numbers are similar to the ones reported for Dutch, German, and Spanish in the CLEARPOND project (0.4 for both neighbor types). In English and French, the

[7] Form length in symbols is also analyzed in this paper, but for this reason we preferred to rely on the CLEARPOND project.

[8] Every prefix has a variety of meanings, so it is difficult to provide translations. For example, *za-* can be inchoative, resultant, has a variety of spatial uses and several other meanings.

[9] Frost did not rely on large lexical databases, focusing on the comparison between several European languages and Semitic languages.

Table 4 The size of orthographic neighborhoods of different types per lemma or word form in *StimulStat* (for every neighborhood type, only lemmas and forms that have neighbors of this type are taken into account)

	Lemmas			Word forms		
	Mean	SD	Range	Mean	SD	Range
sns	3.9	3.2	2–29	4.8	2.4	2–42
tns	2.1	0.3	2–3	2.1	0.3	2–5
ans	2.5	1.7	2–30	2.9	1.3	2–26
dns	2.1	0.3	2–4	2.2	0.4	2–6
pns	3.9	3.2	2–18	N/A[a]	N/A[a]	N/A[a]
wns	10.3	61.3	2–3,464	N/A[a]	N/A[a]	N/A[a]
bins	62.9	153.2	2–3,099	1014.2	2759.5	2–82760
trins	13.7	36.0	2–1179	157.4	535.1	2–30,396

[a] The information is not available because the relevant parameters are calculated online for every query and are not stored in the database

numbers are slightly higher (0.5 in English and 0.6 in French for both neighbor types).

The results for transposition, addition, and deletion lemma neighbors are available for the GreekLex database (Ktori, van Heuven, & Pitchford, 2008). To have a valid comparison, we calculated the relevant values for a subcorpus including 35,304 lemmas, as in the GreekLex project. The proportion of lemmas that have at least one transposition neighbor is 2.0 % in this subcorpus, whereas in GreekLex it is only 0.6 %. For addition and deletion neighbors, the figures are 4.6 % and 6.4 % for Russian and 8.0 % and 9.7 % for Greek, respectively.

Other results presented in Tables 3 and 4 cannot be subjected to cross-linguistic comparisons because the relevant data are not available for other languages.

Conclusions

The *StimulStat* database presented in this paper may be useful for linguists, psychologists and other scientists conducting experimental research on Russian. It is the first lexical database of this type created for Russian. It contains more than 52,000 lemmas and more than 1.7 million word forms and features more parameters than most databases created for other languages, including frequency, length, stress, syllabic structure, phonemic transcription, uniqueness point, as well as other parameters related to orthographic and phonological representations, various grammatical properties, orthographic and phonological neighborhood characteristics, homonymy, polysemy and subjective parameters: subjective age of acquisition and imageability. We took some parameters from various sources and computed the others ourselves.

StimulStat is freely available as a web application, so users do not need to buy and install any specialized software. In the future, we plan to add ideal phonological transcription for all forms included in the database, to recalculate all relevant statistics and to include the option to search for homophonous forms, i.e. the forms that have the same phonological representations, but different spellings. We also plan to develop a tool for generating nonce words with certain properties and for computing required properties for the list of nonce words uploaded by the user.

Acknowledgements This project is funded by the grant 14-04-12034 from the Russian Foundation for Humanities.

References

Acha, J., Laka, I., Landa, J., & Salaburu, P. (2014). EHME: A new word database for research in Basque language. *The Spanish Journal of Psychology*, *17,* E79. https://doi.org/10.1017/sjp.2014.79.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1037. https://doi.org/10.1037/a0031829.

Akinina, Y., Malyutina, S., Ivanova, M., Iskra, E., Mannova, E., & Dragoy, O. (2014). Russian normative data for 375 action pictures and verbs. *Behavior Research Methods*, *47*(3), 691–707. https://doi.org/10.3758/s13428-014-0492-9.

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, *35*(6), 775–800. https://doi.org/10.1006/jmla.1996.0040.

Andrews, S. (1997a). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. https://doi.org/10.3758/BF03214334.

Andrews, S. (1997b). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. https://doi.org/10.3758/BF03214334.

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *The Quarterly Journal of Experimental Psychology*, *59*(05), 920–948. https://doi.org/10.1080/02724980443000782.

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, *44*(4), 1015–1027. https://doi.org/10.3758/s13428-012-0199-8.

Ashby, J., & Rayner, K. (2004). Representing syllable information during silent reading: Evidence from eye movements. *Language & Cognitive Processes*, *19*(3), 391–426. https://doi.org/10.1080/01690960344000233.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*(1), 94–117.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., … Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*, 344–380. https://doi.org/10.3758/BF03196494.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, *24*(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006.

Bocharov, V. V., Alexeeva, S. V., Granovsky, D. V., Protopopova, E. V., Stepanova, M. E., & Surikov, A. V. (2013). Crowdsourcing morphological annotation. In V. P. Selegey (Ed.), *Computational Linguistics and Intellectual Technologies*. Vol. 12 (pp. 109–114). Moscow: RGGU. [In Russian].

Bondarko, L. V. (1977). *Zvukovoj stroj sovremennogo russkogo jazyka [Sound system of the modern Russian language].* Moscow: Prosveschenie. [In Russian].

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, *42*(2), 481–487. https://doi.org/10.3758/BRM.42.2.481.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, *52*(1), 131–143. https://doi.org/10.1016/j.jml.2004.09.003.

Carreiras, M., & Grainger, J. (2004). Sublexical representations and the "front end" of visual word recognition. *Language & Cognitive Processes*, *19*(3), 321–331. https://doi.org/10.1080/01690960344000288.

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 987–1003. https://doi.org/10.1037/0096-1523.18.4.987.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805.

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). New York: Academic Press.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70. https://doi.org/10.3758/BF03206399.

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*(4), 665–671. https://doi.org/10.3758/BF03192738.

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1550–1570. https://doi.org/10.1037/a0014253.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*(4), 1246–1258. https://doi.org/10.3758/s13428-013-0326-1.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595.

Efremova, T. (2000). *Novyj slovar' russkogo jazyka. Tolkovo-slovoobrazovatel'nyj [The new explanatory dictionary of Russian language].* Moscow: Russkij jazyk. [In Russian].

Erten, B., Bozsahin, C., & Zeyrek, D. (2014). Turkish resources for visual word recognition. In *The LREC 2014 Proceedings* (pp. 2106–2110). Retrieved from http://users.metu.edu.tr/bozsahin/LREC2014-final-copy.pdf.

Estivalet, G. L., & Meunier, F. (2015). The Brazilian Portuguese Lexicon: An instrument for psycholinguistic research. *PLoS ONE*, *10*(12), e0144016. https://doi.org/10.1371/journal.pone.0144016.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*, 263–279.

Frost, R. (2015). Cross-linguistic perspectives on letter-order processing: Empirical findings and theoretical considerations. In *The Oxford Handbook of Reading* (pp. 88–98). Oxford: Oxford University Press.

Heister, J., Würzner, K. -M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB — eine lexikalische Datenbank für die psychologische und linguistische Forschung. [dlexDB — a lexical database for psychological and linguistic research]. *Psychologische Rundschau*, *62*(1), 10–20. [In German]. https://doi.org/10.1026/0033-3042/a000029.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4.

Kinoshita, S. (2015). Visual word recognition in the Bayesian reader framework. In *The Oxford Handbook of Reading* (pp. 63–75). Oxford: Oxford University Press.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284. https://doi.org/10.1080/09541440340000213.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135*, 12–35. https://doi.org/10.1037/0096-3445.135.1.12.

Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In Khachay, M. Y., Konstantinova, N., Panchenko, A., Ignatov, D. I., & Labunets, V. G. (Eds.), *Analysis of images, social networks and texts* (pp. 320–332). Berlin: Springer.

Ktori, M., van Heuven, W. J., & Pitchford, N. J. (2008). GreekLex: A lexical database of Modern Greek. *Behavior Research Methods*, *40*(3), 773–783. https://doi.org/10.3758/BRM.40.3.773.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Kwantes, P. J., & Mewhort, D. J. K. (1999). Evidence for sequential processing in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 376–381. https://doi.org/10.1037/0096-1523.25.2.376.

Kyparissiadis, A., van Heuven, W. J., Pitchford, N. J., & Ledgeway, T. (2017). GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLoS ONE*, *12*(2), e0172493. https://doi.org/10.1371/journal.pone.0172493.

Lindell, A. K., Nicholls, M. E., & Castles, A. E. (2003). The effect of orthographic uniqueness and deviation points on lexical decisions: Evidence from unilateral and bilateral-redundant presentations. *The Quarterly Journal of Experimental Psychology: Section A*, *56*(2), 287–307. https://doi.org/10.1080/02724980244000341.

Lyashevskaya, O. (2013). Chastotnyj leksiko-grammaticheskij slovar': Prospect proekta [Lexico-grammatical frequency dictionary: A preliminary design]. In V. P. Selegey (Ed.), *Computational Linguistics and Intellectual Technologies*. Vol. 12 (pp. 478–489). Moscow: RGGU. [In Russian].

Lyashevskaya, O., & Sharov, S. (2009). *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo korpusa russkogo jazyka) [The frequency dictionary of modern Russian language based on Russian National Corpus].* Moscow: Azbukovnik. [In Russian].

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, *7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. https://doi.org/10.1016/0010-0277(80)90015-3.

Mirman, D., Strauss, T. J., Dixon, J. A., & Magnuson, J. S. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, *34*(1), 161–173. https://doi.org/10.1111/j.1551-6709.2009.01069.x.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes* *in reading: Visual word recognition* (pp. 148–197). Hillsdale: Erlbaum.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*(1), 43–71.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review*, *13*(1), 45–52. https://doi.org/10.3758/BF03193811.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524. https://doi.org/10.3758/BF03195598.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Perea, M. (2015). Neighborhood effects in visual word recognition and reading. In *The Oxford Handbook of Reading* (p. 76). Oxford: Oxford University Press.

Perea, M., & Lupker, S. J. (2003). Does judge activate COURT? Transposed-letter similarity effects in masked associative priming. *Memory & Cognition*, *31*(6), 829–841. https://doi.org/10.3758/BF03196438.

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, *38*(4), 610–615. https://doi.org/10.3758/BF03193893.

Rayner, K. (1998). Eye movements in reading and information processing: 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. https://doi.org/10.1016/j.cogsci.2003.08.002.

Schiller, N. O., Fikkert, P., & Levelt, C. C. (2004). Stress priming in picture naming: An SOA study. *Brain and Language*, *90*(1), 231–240. https://doi.org/10.1016/S0093-934X(03)00436-X.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proceedings of the MLMTA 2003* (pp. 273–280). Las Vegas: CSREA Press.

Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., & Evdokimova, V. (2010). A fully annotated corpus of Russian speech. In *The LREC 2010 Proceedings* (pp. 109–112). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.

Taft, M., & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*(6), 607–620.

Yap, M. J., & Balota, D. A. (2015). Visual word recognition. In *The Oxford Handbook of Reading* (pp. 26–43). Oxford: Oxford University Press.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992–1003. https://doi.org/10.3758/BRM.42.4.992.

Zaliznjak, A. A. (1987). *Grammaticheskij slovar' russkogo jazyka [Grammatical dictionary of the Russian Language].* 3rd ed. Moscow: Russkij jazyk. [In Russian].