STIMULSTAT: БАЗА ДАННЫХ, ОХВАТЫВАЮЩАЯ РАЗЛИЧНЫЕ ХАРАКТЕРИСТИКИ СЛОВ РУССКОГО ЯЗЫКА, ВАЖНЫЕ ДЛЯ ЛИНГВИСТИЧЕСКИХ И ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Алексеева С. В. (mail@s-alexeeva.ru)[1],
Слюсарь Н. А. (slioussar@gmail.com)[1, 2],
Чернова Д. А. (chernovadasha@yandex.ru)[1]

[1]СПбГУ, Санкт-Петербург, Россия
[2]НИУ ВШЭ, Москва

Ключевые слова: база данных, русский язык, характеристики лемм, орфографические соседи

STIMULSTAT: A LEXICAL DATABASE FOR LINGUISTIC AND PSYCHOLOGICAL RESEARCH ON RUSSIAN LANGUAGE

Alexeeva S. V. (mail@s-alexeeva.ru)[1],
Slioussar N. A. (slioussar@gmail.com)[1, 2],
Chernova D. A. (chernovadasha@yandex.ru)[1]

[1]St. Petersburg State University, St. Petersburg, Russia
[2]HSE, Moscow

Experimental studies conducted by linguists, psychologists and other researchers identified a large list of word properties that play a role for speech production and comprehension. They include lemma and form frequency, word length, the number of syllables, stress pattern, frequency of letters and syllables the word consists of, whether the word has homonyms, homographs or orthographical neighbors, whether it has multiple senses, whether it is archaic or colloquial etc. Of course, various grammatical characteristics (part of speech, inflectional paradigm etc.) are also important. Taking these properties into account in new studies became an important problem. For several languages, databases with search tools were designed to solve this problem. In this paper, we present *StimulStat*, the first such database for Russian language. We are still working on it, but a preliminary version

is already functional and is available on the web. It includes more than 50,000 most frequent Russian words characterized according to 97 properties. The database can be used for stimulus selection in experimental studies of Russian and provides a lot of information that may be relevant in other linguistic domains because one can easily calculate how words with many combinations of different characteristics are distributed, which patterns are widespread and which are infrequent.

Keywords: lexical database, Russian language, lemma properties, orthographic neighbors

1. Введение

В результате многочисленных экспериментальных исследований на материале различных языков, которые проводятся лингвистами, психологами, а также представителями других наук, был установлен целый ряд факторов, которые влияют на порождение и восприятие речи. Некоторые из этих факторов характеризуют отдельные слова, другие — синтаксические конструкции, третьи относятся к уровню дискурса и текста в целом и т. д. В данной работе речь пойдет о характеристиках слов. Среди характеристик, играющих важную роль для лингвистических и психологических исследований, можно назвать частотность (как отдельных форм слова, так и леммы в целом), длину, слоговую структуру, место ударения и акцентную парадигму, то, имеет ли слово омонимы, омографы или орфографических соседей (то есть сходные по написанию слова), сколько у слова значений, а также многое другое.

При планировании новых исследований часто возникает необходимость учесть целый ряд таких факторов, что превращает подбор стимульного материала в чрезвычайно трудоемкий процесс. Для нескольких языков существуют базы данных в виде компьютерных программ или интернет-приложений, которые позволяют отбирать слова, учитывая различные факторы, а также получать сведения о количестве слов с теми или иными характеристиками. Среди них *English lexicon project* (http://elexicon.wustl.edu) (Balota et al. 2007), *N-Watch* (Davis 2005) и *MRC database* (Colheart 1981) для английского языка, *DlexDB* (http://www.dlexdb.de) (Heister et al. 2011) для немецкого, *CELEX Lexical Database* (http://wwwlands2.let.kun.nl/members/software/celex.html) (Baayen, Piepenbrock, van Rijn 1995) для голландского, английского и немецкого, *Lexique* (http://www.lexique.org) (New et al. 2001; New et al. 2004) для французского, *BuscaPalabras* (Davis, Perea 2005) для испанского и *E-Hitz* (Perea et al. 2006) для баскского.

Что касается русского языка, то часть необходимых сведений можно найти в различных базах данных и электронных версиях словарей, но эти ресурсы не объединены и зачастую не снабжены эффективными механизмами для фильтрации и поиска слов по заданным характеристикам. Некоторые параметры, которые есть в упомянутых выше ресурсах, созданных для других языков, для русского языка не представлены нигде. Чтобы хотя бы отчасти восполнить этот пробел, мы работаем над созданием базы данных SimulStat, которая

включает более 50 000 наиболее частотных слов русского языка и их форм. Все эти слова описаны по 97 различным параметрам, касающимся, в числе прочего, (1) частотности лемм, форм, а также входящих в них слогов и букв, (2) омонимов, омографов, орфографических соседей, (3) различных грамматических характеристик.

Заметим сразу, что в тех случаях, когда мы не рассчитывали значения параметров сами, а опирались на существующие словари, мы не пытались оценить и тем более пересмотреть различные решения их составителей, которые могут показаться спорными. Многие спорные вопросы такого рода могли бы стать темой для больших самостоятельных исследований, далеко выходящих за рамки нашей работы. Поэтому в таких случаях нашей целью было лишь включить данные в базу, чтобы соответствующие параметры можно было использовать при поиске слов, сами по себе или в сочетании с другими. Так как информация о сторонних источниках, на которые опирается база, доступна пользователям, они сами могут решать, достаточно ли тот или иной источник подходит для целей их исследования и с какими подводными камнями они могут столкнуться.

На данный момент практически завершена работа с леммами, результаты которой и будут представлены в этой статье. Работа со словоформами еще продолжается. Предварительная версия базы данных доступна в интернете под названием *ru_stimul* на сайте Лаборатории когнитивных исследований СПбГУ (http://stimul.cognitivestudies.ru, доступ предоставляется по требованию). Сейчас там можно посмотреть значения параметров для каждого из слов, включенных в базу, а также для групп слов, и отобрать слова по заданной комбинации параметров. Кроме того, нами были сделаны первичные расчеты, позволяющие оценить количество слов с различными представленными в базе характеристиками. В данный момент мы работаем над созданием пользовательского веб-интерфейса, который существенно упростит использование базы. Проект осуществляется при поддержке гранта РГНФ №14-04-12034.

2. Описание базы

В качестве основы для базы данных был взят список лемм из «Частотного словаря современного русского языка» (Ляшевская, Шаров 2009), созданного на основе подкорпуса Национального корпуса русского языка (http://www.ruscorpora.ru), включающего 92 миллиона словоупотреблений. Эти леммы описаны по 84 параметрам. Группы слов, близкие по написанию (которые также называются квазиомографами или орфографическими соседями) охарактеризованы еще по 13 параметрам. В качестве примеров квазиомографов можно привести следующие: *сетка-секта* («соседи» с перестановкой), *сок-сук* («соседи» с заменой).

База данных реализована в виде пяти основных таблиц, связанных между собой:

- таблица уникальных лемм (таблица «Леммы») с параметрами, которые независимы от морфологической принадлежности лексической единицы (длина, слоговая структура, количество слогов и др.);

- таблицы с информацией об орфографических соседях (два варианта: с буквой *ё* — таблица «Соседи», без буквы *ё* — таблица «Соседи без ё») (таблицы «Соседи» и «Соседи без ё», связаны с таблицей «Леммы»);
- таблица лемм с информацией о морфологических характеристиках, частотности в зависимости от части речи, а также месте ударения, количестве значений и др. (таблица «Морфология», связана с таблицей «Леммы»);
- таблица с информацией об омонимах (таблица «Омонимы», связана с таблицей «Морфология»).

Для создания таблицы «Леммы» из электронной версии «Частотного словаря современного русского языка» (Ляшевская, Шаров 2009) были выделены уникальные лексические единицы, независимо от частей речи. Этот этап был необходим, поскольку при определении орфографических соседей не имеет значения, какой частью речи является данное слово. Если производить поиск соседей в базе, где леммы могут дублироваться (за счет того, что относятся к разным грамматическим классам), в итоге в списке появятся дубликаты. Таким образом, в таблицу «Леммы» вошло 51 688 лексических единиц, выделенных из 52 139 единиц, содержащихся в частотном словаре. 419 из этих 51 688 единиц представляют пары или тройки омонимов, относящихся к разным частям речи: например, *добро* может быть существительным, наречием или частицей.

Затем при помощи программы «Yo» (http://vgiv.narod.ru/yo/yo.html) мы в полуавтоматическом режиме проставили в таблице «Леммы» букву *ё* в 1597 словах (в частотном словаре используется запись без *ё*). Так как наличие/отсутствие в орфографической записи буквы *ё* влияет на значение некоторых важных для нас параметров (например, если слово *лёд* написано с буквой *ё*, оно не является орфографическим соседом для слова *лес*, в то время как без нее является), нам пришлось не только создавать два варианта таблицы с орфографическими соседями, но и иметь два варианта для пяти параметров внутри основной таблицы. После этого были написаны скрипты для проставления в таблице «Леммы» значений различных параметров. Многие параметры перечислены в третьем разделе.

В таблице «Морфология» собрана информация о леммах, которая зависит от части речи. Туда вошла вся грамматическая информация, а также сведения о месте ударения, количестве значений слова и многое другое. Для ее создания мы взяли список лемм из «Частотного словаря современного русского языка», не удаляя информацию об их частеречной принадлежности. Для выделения ключевых грамматических характеристик мы использовали морфологический анализатор «Pymorphy2» (https://pymorphy2.readthedocs.org/en/latest/), который опирается на словарь проекта «OpenCorpora» (http://www.opencorpora.org) (Bocharov et al. 2013). В анализатор мы подавали лемму и часть речи. Если такой комбинации в анализаторе не оказывалось, мы помещали в таблицу «Морфология» только информацию о части речи. На данный момент в таблице 1972 таких лемм, и мы планируем вернуться к ним в дальнейшем. Таким образом, 97 % слов в базе получили полный грамматический разбор, большинство слов без грамматической информации представляют собой аббревиатуры и имена

собственные, которые редко используются в качестве стимулов в психолингвистических экспериментах.

Если анализатор выдавал несколько разборов (например, два для леммы *оператор*: одушевленное и неодушевленное существительное или для леммы *женить*: глагол совершенного и несовершенного вида), то обе эти леммы (омонимы на уровне парадигмы) помещались в таблицу «Морфология», а также в таблицу «Омонимы». При этом в настоящий момент в базе данных нет информации о частотности каждого из омонимов, в ней представлена только суммарная частотность обеих лемм.[1] Затем мы в полуавтоматическом режиме добавили в таблицу «Морфология» информацию из «Грамматического словаря русского языка» (Зализняк 1977). В этом словаре представлено более 100 000 слов. В таблицу «Морфология» были включены сведения об ударении, особенностях словоизменения и др. Если в словаре содержалось несколько омографов (например, *мУка* и *мукА*), мы добавляли эти варианты в таблицу, а потом при помощи поля *rel_lemma_morph_zal_subordinate* связывали их с уже имеющейся в таблице леммой. 8731 лемма, содержащаяся в нашей базе данных, не представлена в «Грамматическом словаре русского языка». Для этих лемм информация об ударении, а также часть грамматических характеристик были проставлены вручную.

Кроме того, мы дополнили таблицу «Морфология» информацией о полисемии, включив туда количество значений слов из «Нового толково-словообразовательного словаря русского языка» (Ефремова 2000).[2] В этом словаре около 140 000 слов. Более 350 глаголов в таблице «Леммы» также снабжены информацией о субъективном возрасте усвоения, представимости содержания и знакомости действия, взятой из базы данных «Глагол и действие» (http://www.neuroling.ru) (Akinina et al. 2014).[3]

Кроме основных таблиц «Леммы» и «Морфология», мы создали таблицы «Соседи» и «Соседи без ё», которые содержат информацию о квазиомографах (орфографических соседях), т. е. близких по написанию словах, и таблицу «Омонимы», содержащую информацию об омонимах. Таблица «Соседи» составлена с учетом буквы *ё*, таблица «Соседи без ё» — без учета. В таблице «Омонимы»

---

[1] В дальнейшем мы планируем вернуться к данному вопросу. Для определения частотности каждого варианта можно использовать метод К. Лилока и коллег (Lealock et al. 2000), разработанный для оценки частотности отдельных значений внутри многозначного слова. Этот метод подразумевает следующую процедуру: необходимо случайным образом выбрать из представительного корпуса двести или более словоупотреблений, разметить их вручную, вычислить долю каждого варианта и на основе этого и суммарной частотности вычислить частотность каждого значения.

[2] Анонимные рецензенты отмечают, что, возможно, выбор этого словаря был не оптимальным, и при создании итогового варианта базы мы рассмотрим вопрос о включении других источников.

[3] Очевидно, что подавляющее большинство представленных в базе слов останется неразмеченными по этому параметру, в связи с чем может возникнуть вопрос, целесообразно ли включать его в базу. Нам кажется, что ответ, безусловно, положительный, так как одна из функций базы — это консолидация информации из различных источников, чтобы их можно было использовать одновременно.

каждый омоним занимает одну строчку, а связь между омонимами осуществляется при помощи поля *rel_lemma_stat_main_omos*, где слова, которые являются омонимами, имеют один и тот уже уникальный идентификатор. Омонимы связываются с таблицей «Морфология» через поле *rel_lemma_stat_morph*.

Заполнение базы данных осуществлялось при помощи скриптов на языке Python, сама база использует технологию СУБД PostgreSQL. Предварительная версия базы выложена на сайте Лаборатории когнитивных исследований СПбГУ (http://stimul.cognitivestudies.ru, доступ предоставляется по требованию), на котором сейчас действует версия интерфейса для администратора (phpPgAdmin). Мы работаем над созданием пользовательского интерфейса.

3. Краткий обзор содержащейся в базе информации

3.1. Некоторые сведения о леммах в таблицах «Леммы» и «Морфология»

В таблице «Леммы» содержится информация о 25 характеристиках лемм. В их числе:

- частотность согласно «Частотному словарю современного русского языка» (Ляшевская, Шаров 2009), а также натуральный и десятичный логарифм частотности;
- длина слова в символах, количество слогов (по количеству гласных), границы слогов согласно модели Л. В. Бондарко (Бондарко 1977)[4], слоговая структура — запись, в которой любой гласной соответствует символ *V*, согласной — *C*, буквам ь и ъ — *F* (например, *мышь* — CVCF);
- позиция в орфографической записи, начиная с которой слово однозначно распознается, т. е. других слов с таким же начальным сегментом нет (т. н. word uniqueness point);
- запись слова в обратном порядке (например, *дас* для слова *сад*);
- информация, касающаяся входящих в слово букв: первая и последняя буква в слове, отсортированный список всех букв и уникальных букв, содержащихся в слове (например, *клмооо* и *клмо* для слова *молоко*) и др.

Часть параметров рассчитана с учетом и без учета буквы *ё*.

В таблице «Морфология» содержится информация о том, какой по счету букве и какому по счету слогу в слове соответствует позиция первичного и вторичного ударения, и есть ли в словоизменительной парадигме данного слова сдвиг ударения (в русском языке различаются разные типы смещения ударения, и более детальную информацию такого рода можно получить, анализируя пометы, связанные с акцентными парадигмами, которые также включены

[4] Мы предполагаем в дальнейшем включить в базу и другие подходы.

в базу). Кроме того, там содержится информация о более чем сорока грамматических характеристиках слов, в частности:

- часть речи (согласно системе, использующейся в «Частотном словаре современного русского языка» (Ляшевская, Шаров 2009), и системе в словаре «OpenCorpora» (http://www.opencorpora.org) (Bocharov et al. 2013)). При переходе от нотации «Частотного словаря» к системе тегов словаря «OpenCorpora» мы автоматически произвели «схлопывание» некоторых грамматических классов, которые выделяются в «Частотном словаре» как отдельные части речи, но являются частью других грамматических классов в словаре OpenCorpora: местоименные и порядковые прилагательные были соотнесены с прилагательными, местоименные наречия — с наречиями, а тег «имя собственное» был заменен на тег «существительное» .
- грамматические категории и другие характеристики (род, одушевленность, вид, переходность, является ли существительное именем собственным, является ли прилагательное или наречие местоименным и т. д.);
- различные особенности словоизменения согласно «Грамматическому словарю русского языка», то есть номер словоизменительного типа и некоторые другие пометы (Зализняк 1977).

Кроме того, из «Грамматического словаря русского языка» были взяты и некоторые другие сведения (например, есть ли у слова варианты, как в случае с *номеровать-нумеровать* или *зал-зала*, не является ли слово разговорным, архаичным и т. д.). Также в таблице «Морфология» содержится информация о том, сколько у слова различных значений (причем омонимы и омографы учитываются как вместе, так и раздельно).

3.2. Сведения об орфографических соседях в таблицах «Соседи» и «Соседи без ё» и об омонимах в таблице «Омонимы»

Экспериментальные исследования лексического доступа при чтении с использованием различных методик (принятие лексического решения, называние и др.) показали, что на этот процесс влияет наличие у слова орфографических соседей и их тип. В таблицы II и «Соседи без ё» мы включили следующие типы орфографических соседей:

- Sns — «соседи с заменой» (*ток — сок*) (ср. Coltheart et al. 1977). Всего в базе 12 280 групп таких соседей, в которые входят 15 242 леммы в различных комбинациях.
- Tns — «соседи с перестановкой» (*баян — баня*) (ср. Andrews 1996; Perea, Lupker 2004). Всего в базе 642 групп таких соседей, в них входят 1130 лемм.
- Dns — «соседи с удалением одной буквы» (*крот — кот*) и ans — «соседи с вставкой одной буквы» (*кот — крот*) (ср. Davis, Perea, Acha, 2009). В базе 2409 групп соседей с удалением одной буквы и 3346 групп с вставкой одной буквы, в которые входят 5077 леммы.

- Pns/ppns — «соседи, у которых второе слово в паре (ppns) полностью включено в первое» (*абориген — бор, ген*) и wns/wwns — «соседи, у которых первое слово в паре (wwns) полностью включено во второе (wwns)» (*бор — абориген, забор…*) (ср. Bowers, Davis, Hanley 2005). В данном случае мы использовали слова длиной не менее трех букв. Иначе получилось бы, что однословные и двусловные союзы, предлоги и пр. включены в качестве соседей в огромное количество слов. В базе 43 900 групп соседей первого типа и 13 993 групп второго типа, в которые входят 47 105 леммы.
- Bins — «соседи, имеющие в той или иной позиции общую биграмму (т. е. сочетание из двух букв)» (*ток — топка*) и trins — «соседи, имеющие в той или иной позиции обшую триграмму (т. е. сочетание из трех букв)» (*порвать — поручень*) (ср. Davis 2005). Почти все слова в базе данных входят в группы соседей первого типа (51 658) и второго типа (51 175).

Слова-соседи учитываются в таблицах II и «Соседи без ё» следующим образом. В колонке *lemma_stat_neigh_rel* обозначен тип квазиомографа, в колонке *lemma_stat_neigh_position* — с какой или в какой позиции квазиомографы отличаются друг от друга (а для квазиомографов, имеющих общую биграмму или триграмму, позиции, где эти биграммы или триграммы находятся). В колонке *lemma_stat_neigh_count* указано количество слов (вместе с данным), которые похожи друг на друга (т. е. входят в данную группу соседей). Кроме того, в таблицах дается информация об общей частотности группы соседей, к которой принадлежит данное слово, о частотности самого частотного и самого редкого члена группы, о том, сколько слов в группе имеют частотность выше, чем у данного, а также некоторые другие вспомогательные сведения, необходимые для того, чтобы связать слова-соседи внутри группы.

В таблице «Омонимы», а также в других таблицах там, где это необходимо, содержится информация об омонимах и омографах. В базу входят 295 групп омографов, 419 групп омонимов, относящихся к разным частям речи (в 395 группах два слова, в 20 — три, в четырех — четыре) и 1534 групп омонимов, относящихся к одной и той же части речи, в которые входит 3126 лемм. Всего в базе 56 270 лемм с учетом омонимов и омографов и 51 688 лемм без их учета. Учет омонимов и омографов в таблице организован в целом так же, как учет слов-соседей, описанных выше.

3.3. Распределение различных характеристик слов в базе данных

В Таблице 1 представлена информация о некоторых перечисленных выше параметрах с указанием того, какие их значения оказались в базе наиболее распространенными (с учетом и без учета частотности лемм, обладающих соответствующими параметрами). В Таблице 2 представлены минимальные, максимальные и средние значения различных параметров. Это дает некоторое представление о том, какого рода информацию можно извлечь из базы.

Таблица 1. Три самых частотных значения некоторых параметров в базе StimulStat

	Без учета частотности лемм	**С учетом частотности лемм**
Слоговая структура[1]	CVCVC, CVCVCVC, CVCCVC	CV, V, C
Первая буква	*п, с, в*	*п, с, в*
Последняя буква	*й, ь, а*	*ь, й, а*
Часть речи	существительное, глагол, прилагательное	существительное, глагол, прилагательное
Род	мужской, женский, средний	мужской, женский, средний
Вид глагола	совершенный, несовершенный	несовершенный, совершенный

Таблица 2. Минимальные, максимальные и средние значения некоторых параметров в базе StimulStat

	Min	**Max**	**Среднее**	**Стандартное отклонение**	**Медиана**
Длина	1	34	9,1 (5,5)[a]	3,1 (3,3)[a]	9 (5)[a]
Длина слога	0[b]	15	3,5 (2,1)[a]	1,4 (1,3)[a]	3 (2)[a]
Частотность (ipm)	0,4	35 801,8	18,5	291,2	1,9
Натуральный логарифм частотности	0,3	10,5	1,4	1,2	1,1
Десятичный логарифм частотности	0,2	4,6	0,6	0,5	0,5
Позиция в орфографической записи, начиная с которой слово однозначно распознается	1	21	7,1	2,5	7
Фонология					
Ударение (букв.)	0[b]	30	5,5 (3,2)[a]	2,6 (2,1)[a]	5 (3)[a]
Ударение_побочн (букв)	0	15	0,08 (0,01)[a]	0,6 (0,2)[a]	0 (0)[a]
Ударение (слог)	0[b]	13	2,4 (1,5)[a]	1,1 (0,9)[a]	2 (1)[a]
Ударение_побочн (слог)	0	6	0,04 (0,01)[a]	0,3 (0,1)[a]	0 (0)[a]
Кол-во значений	1	38	2,1 (5,8)[a]	1,5 (7,1)[a]	2 (3)[a]
Омонимы и омографы					
Размер группы омографов	2	2	2	0	2
Размер группы функциональных омонимов на уровне частей речи	2	4	2,1	0,3	2
Размер группы функциональных омонимов на уровне парадигмы	2	4	2	0,2	2

	Min	Max	Среднее	Стандартное отклонение	Медиана
Орфографические соседи					
Размер группы соседей sns[c]	2	29	3,9	3,2	3
Частотность группы соседей sns[c]	0,8	57 281,8	327,5	2126,9	21,9
Размер группы соседей tns[c]	2	3	2,1	0,3	2
Частотность группы соседей tns[c]	0,8	17 774	183,5	1075,6	13,4
Размер группы соседей ans[c]	2	30	2,5	1,73	2
Частотность группы соседей ans[c]	0,8	43 759,7	267,4	1985,4	14,5
Размер группы соседей dns[c]	2	4	2,1	0,3	2
Частотность группы соседей dns[c]	0,8	67 180,3	858,8	4086,7	16,3
Размер группы соседей pns[c]	2	18	3,9	1,9	4
Частотность группы соседей pns[c]	0,8	16 049,0	419,8	976,2	64,1
Размер группы соседей wns[c]	2	3464	10,3	61,7	2
Частотность группы соседей wns[c]	0,8	39 669,0	121,5	692,9	13,4
Размер группы соседей bins[c]	2	8054	2719,0	1519,5	2482,0
Частотность группы соседей bins[c]	1,2	104 670,2	31 400,8	18 594,7	27 877,1
Размер группы соседей trins[c]	2	2892	529,1	495,2	344,0
Частотность группы соседей trins[c]	0,9	32 638,9	5165,7	4998,2	3350,0

[a]Число за скобками не учитывает частотность лемм, число в скобках — учитывает.
[b]Ноль означает, что слово не содержит гласных, как, например, предлог *в*.
[c]Только для непустых групп соседей.

Данная база данных была использована в ряде исследований, проведенных в Лаборатории когнитивных исследований СПбГУ. Например, в работе А. М. Фроловой (2014) изучался принцип возможного слова при сегментации устной речи на русском языке. Для данного исследования были подобраны существительные, прилагательные и глаголы с ударением на второй слог в диапазоне натурального логарифма частот от 1,8 до 4,5, длиной от пяти до семи букв и начинающиеся на буквы *в, р, н, л*.

4. Заключение

В данной статье представлена база данных StimulStat, охватывающая важнейшие психолингвистические характеристики для основного лексического фонда русского языка. В настоящий момент в базу включено более 50 000 лемм, охарактеризованных по более чем 90 параметрам. Аналогов для русского языка не существует, а все немногочисленные подобные базы, созданные на материале других языков, учитывают значительно меньшее число различных параметров.

База данных может быть полезна при подборе стимулов для экспериментальных исследований порождения и восприятия речи, проводящихся лингвистами, психологами и другими учеными — для этого она была задумана. Однако, как нам кажется, она представляет интерес и для других лингвистических исследований, так как впервые позволяет оценить распределение различных характеристик среди наиболее частотных слов русского языка, а также, скажем, оценить, насколько они коррелируют друг с другом.

В дальнейшем мы планируем закончить разработку пользовательского интерфейса, который заметно облегчит работу с базой, завершить работу над таблицей словоформ, а также, возможно, добавить информацию о звучании слов (в виде транскрипции). Мы также разрабатываем алгоритмы по созданию квазислов для лингвистических экспериментов с учетом заданных параметров и по описанию списков квазислов, загруженных пользователем, который позволял бы подсчитывать их длину, количество слогов, число орфографических соседей среди реальных слов и т. д.

References

1. *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.
2. *Andrews, S.* (1996), Lexical retrieval and selection processes: Effects of transposed-letter confusability. Journal of Memory and Language, 35, pp. 775–800.
3. *Baayen, R. H., Piepenbrock, R., &van Rijn, H.* (1995), The CELEX Lexical Database.
4. *Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R.* (2007), The English Lexicon Project. Behavior Research Methods, 39, pp. 445–459.
5. *Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V.* (2013), Crowdsourcing morphological annotation. Computational Linguistics and Intellectual Technologies: DIALOG 2013, vol. (12) 19. Russian State Humanitarian University, Moscow, pp. 109–114.
6. *Bondarko L. V.* (1977), Sound system of modern Russian [Zvukovoy stroy sovremennogo russkogo yazyka], Moscow .
7. *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, pp. 131–143.

8. *Coltheart, M.* (1981), The MRC Psycholinguistic Database, Quarterly Journal of Experimental Psychology,33A, pp. 497–505.
9. *Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D.* (1977). Access to the internal lexicon. In S. Dornic (Ed.), Attention and per- formance VI (pp. 535–555). New York: Academic Press.
10. *Davis, C. J.* (2005), N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. Behavior Research Methods,37, 65–70.
11. *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, pp. 665–671.
12. *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 & Performance, 35, pp. 1550–1570.
13. *Efremova T. F.* (2000), New dictionary of Russian. Explanatory and word-formative. [Novyy slovar russkogo yazyka. Tolkovo-slovoobrazovatelnyy], Moscow.
14. *Frolova A. M.* (2014), Aspects of speech segmentation: an experimental study with reference to Russian (master thesis) [Osobennosti segmentatsii ustnoy rechi: eksperimentalnoye issledovaniye na material russkogo yazyka (magisterskaya dissertatsiya)], St.Petersburg State University (http://phil.spbu.ru/ucheba-1/zaschita-magisterskih-rabot-v-2014-g/kafedra-obschego-yazykoznaniya).
15. *Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al.* (2011), dlexDB — eine lexikalische Daten- bank für die psychologische und linguistische Forschung. Psychologische Rundschau, 62(1), pp. 10–20.
16. *Lealock C., Chodorov M., Miller G. A. A.* (2000), Topical/Local Classifier for Word Sense Identification // Computers and the Humanities vol. 34, Special Issue on SENSEVAL. Netherlands, Kluwer Academic Publishers, pp. 115–120.
17. *Lyashevskaya O. N., Sharov S. A.* (2009), Frequency dictionary of modern Russian (based on Russian National Corpus) [Chastotnyj slovar sovremennogo russkogo yazyka (na materialakh Natsionalnogo korpusa russkogo yazyka)], Azbukovnik, Moscow.
18. *New, B., Pallier, C., Brysbaert, M., & Ferrand, L.* (2004), Lexique 2: A new French lexical database. Behavior Research Methods, Instruments & Computers, 36, pp. 516–524.
19. *New, B., Pallier, C., Ferrand, L., & Matos, R.* (2001), Une base de données lexicales du français contemporain sur internet: LEXIQUE [A french lexical database on internet: LEXIQUE], L'Année Psychologique 101, pp. 447–462.
20. *Perea, M. and Lupker, S. J.* (2004), Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. Journal of Memory and Language, 51, pp. 231–246.
21. *Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M.* (2006), E-Hitz: A word frequency list and a program for deriv- ing psycholinguistic statistics in an agglutinative language (Basque). Behavior Research Methods, 38, pp. 610–615.
22. *Release 2* [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.
23. *Zaliznyak A. A.* (1977), Grammar dictionary of Russian. Inflection. [Grammaticheskiy slovar russkogo yazyka. Slovoizmenenije], Moscow.