STIMULSTAT: ИНСТРУМЕНТ ДЛЯ ПОДБОРА ПСИХОЛИНГВИСТИЧЕСКИХ СТИМУЛОВ НА РУССКОМ ЯЗЫКЕ

С. В. Алексеева* (1), Н. А. Слюсарь (2, 1), Д. А. Чернова (1)
mail@s-alexeeva.ru
1 — СпбГУ, Санкт-Петербург; 2 — НИУ ВШЭ, Москва

Аннотация. В статье представлена созданная нами база данных StimulStat, которая включает более 50 000 наиболее частотных слов русского языка и их форм (> 1 700 000 единиц). Эти слова описаны более чем по 70 различным параметрам, играющим роль в лингвистических и психологических исследованиях. База данных может быть использована для подбора стимулов в экспериментальных исследованиях на материале русского языка. База данных включает такие параметры лемм и словоформ, как количество букв и фонем в слове, количество слогов, частотность, фонологическую информацию (идеальную и реальную фонемную транскрипцию, место ударения, слоговую структуру слова), морфологическую информацию по всем частям речи (включая словоизменительный индекс А. А. Зализняка), семантическую информацию (количество значений) и информацию об орфографических и фонологических словах-соседях (квазиомографах или квазиомофонах с заменой, перестановкой, вставкой, удалением букв/звуков). Для обеспечения свободного доступа к базе создан веб-интерфейс (http://stimul.cognitivestudies.ru).

Ключевые слова: база данных, психолингвистика, отбор слов, русский язык

Описанная в статье база данных создана в ходе выполнения проекта РГНФ № 14-04-12034 в 2014 – 2016 гг.

Основная сложность подбора стимульного материаладля психолингвистических исследований обычно заключается в том, что релевантные параметры находятся в разных источниках,которые зачастую не снабжены эффективными механизмами для фильтрации и поиска слов, а некоторые параметры приходится рассчитывать самостоятельно (например, число слогов и место ударения). Чтобы решить эту проблему, для ряда языков были созданы базы данных в виде компьютерных программ или интернет-приложений. Среди них English lexicon project (Balota et al., 2007) и N-Watch (Davis, 2005) для английского языка, DlexDB (Heister et al., 2011) для немецкого, BuscaPalabras (Davis, Perea, 2005) и EsPal (Duchon et al., 2013) для испанского, Aralex (Boudelaa, Marslen-Wilson, 2010) для арабскогои др. Для русского аналогичного ресурса ранее не разрабатывалось.

Нами была создана база StimulStat (http://stimul.cognitivestudies.ru), которая содержит параметры, связанные с частотностью, буквенным и фонемным составом, просодическими особенностями, полисемией и омонимией, грамматическими характеристиками лемм и словоформ, с наличием близких по написанию и произнесению слов и др. Некоторые параметры были взяты из различных источников (преимущество базы заключается в возможности учитывать их одновременно), другие были рассчитаны при создании базы. Пользовательский интерфейс представлен на русском и английском языках.

Выбор источников

В качестве базового источника был взят «Частотный словарь современного русского языка» (Ляшевская, Шаров, 2009), созданный на основе подкорпуса Национального корпуса русского языка (http://www.ruscorpora.ru), включающего 92 млн словоупотреблений. Словарь содержит 52 139 лемм русского языка, охарактеризованных по частотности и частеречной принадлежности. Для получения полной парадигмы слов с грамматическими характеристиками использовался морфологический анализатор Pymorphy2 (https://pymorphy2.readthedocs.org) (Korobov, 2015), опирающийся на словарь проекта OpenCorpora (http://www.opencorpora.org) (Bocharov et al., 2013). С помощью базы можно осуществлять поиск по части речи, числу, роду, падежу, одушевленности, виду, переходности, наклонению, залогу, времени, лицу, личным и неличным формам (для глагола), степени, разрядам (для прилагательных) и др. Всего в базе данных хранится 1 700 842 словоформ, из них более 300 000 словоформам приписана частотность[1].

База включает в себя сведения из «Грамматического словаря русского языка» (Зализняк, 1987), в частности, информацию об ударении и индексы словоизменительных классов, а также различные грамматические характеристики[2]. В нее внесены данные о количестве значений слов из «Нового толково-словообразовательного словаря русского языка» (Ефремова, 2000). Наконец, в базу включена информация о реальной и идеальной фонемной транскрипции (далее РФТ и ИФТ) на основе словаря фонетических вариантов, полученных из корпуса CORPRES (Skrelin et al., 2010). Корпус включает в себя 60 часов начитанной восемью дикторами речи, словарь составлен на основе 105 093 словоупотреблений и включает в себя 9 965 уникальных пар «словоформа в орфографической записи — словоформа в ИФТ» и 26 778 пар «словоформа в орфографической записи — словоформа в РФТ».

В тех случаях, когда мы не рассчитывали значения параметров сами, а опирались на другие источники, мы не пытались пересмотреть различные решения их составителей.

1 Информация, основанная на данных одномиллионного подкорпуса Национального корпуса русского языка, была любезно предоставлена О. Н. Ляшевской, руководителем проекта «Частотная грамматика русского языка» (http://web-corpora.net/freaky_frequency/freq_main.html) (Ляшевская, 2013).

2 При проставлении ударений в словоформах использовался список словоформ с ударением, созданный А. А. Усачевым на основе «Грамматического словаря русского языка» (http://www.speakrus.ru/dict/#paradigma).

Параметры

Нами были рассчитаны следующие параметры для лемм и форм (как для орфографического представления, так и для РФТ и ИФТ): натуральный и десятичный логарифм частотности (в случае РФТ и ИФТ рассчитана частотность различных вариантов); параметры, связанные с буквенным составом: длина слова в символах, первая и последняя буква, запись слова в обратном порядке (например, *дас* для слова *сад*), отсортированный список всех букв или уникальных букв в слове (например, *клмооо* и *клмо* для *молоко*), позиция однозначной идентификации, или позиция в орфографической записи, начиная с которой слово однозначно распознается (т.н. word uniqueness point); аналогичные параметры для фонемного состава; параметры, связанные с делением на слоги и ударением: количество слогов (по количеству гласных), границы слогов согласно модели Л.В. Бондарко (1977), слоговая структура, место первичного и вторичного ударения, наличие сдвига ударения в словоизменительной парадигме; информация об омонимии и омографии с учетом и без учета частеречной принадлежности; информация о словах-соседях. Кроме того, в базе данных собрана информация об омонимах, относящихся к различным частям речи (например, *о* — предлог и междометие), к одной части речи (*оператор* — одушевленное и неодушевленное существительное), об омонимичных формах одного слова (например, *кошке* — дательный или предложный падеж единственного числа) и об омографах (например, *мука́ / му́ка*).

Экспериментальные исследования лексического доступа показали, что на него влияет наличие близких по написанию и произнесению слов (т.н. соседей) и их тип. Мы рассчитали следующие типы соседства для всех вошедших в базу лемм и форм: соседи с заменой (*ток – сок*) (Coltheart et al., 1977); соседи с перестановкой (*баян – баня*) (Andrews, 1996); соседи с удалением одной буквы (*крот – кот*) и со вставкой одной буквы (*кот – крот*) (Davis et al., 2009); соседи, у которых второе слово в паре полностью включено в первое (*абориген – бор, ген*), и соседи, у которых первое слово в паре полностью включено во второе (*бор – абориген, забор...*) (Bowers et al., 2005); соседи, имеющие общую биграмму (то есть сочетание из двух букв) (*ток – топка*) или триграмму (*порвать – поручень*) (Davis, 2005). Соседи были отдельно выявлены для лемм и форм, в орфографической и в фонематической записи. При определении соседей по фонематической записи соседями считались только такие пары фонетических реализаций, которые соотносятся с разными орфографическими представлениями. Для нахождения словоформ-соседей использовалась только РФТ как более релевантная для исследования восприятия устной речи. Для каждого соседства подсчитано количество слов, суммарная частотность, натуральный и десятичный логарифм частотности, минимальная и максимальнаячастотность; для каждого слова в соседстве известно, сколько слов превышает по частотности данное. Определено, сколько среди словоформ-соседей по РФТ уникальных орфографических словоформ и что это за словоформы.

Веб-интерфейс

Важной задачей проекта было не только создание базы, но и разработка удобного веб-интерфейса, чтобы сделать ее общедоступной. Веб-интерфейс

создавался при помощи веб-фреймворка *Django*. База доступна на сайте http://stimul.cognitivestudies.ru. На сайте есть инструкции и многочисленные комментарии. В базе доступно два режима поиска: можно искать определенный набор параметров для заданного списка лемм или форм на странице «Поиск по слову», а можно — леммы или формы, удовлетворяющие заданному списку параметров, на странице «Отобрать слова». Во втором случае для многих числовых параметров можно задать как точное значение, так и определенный диапазон. Обработав запрос, сайт генерирует файл с результатом, а также отображает первые 50 строк файла на отдельной странице выдачи.

Описательная статистика для лемм и словоформ

База помогает быстро оценить, каковы самые частотные значения описанных выше параметров, а для численных значений получить базовую статистику. Ниже мы приводим средние значения для некоторых параметров из базы. Средняя длина лемм в символах в «Частотном словаре русского языка» (далее «словарь лемм») — 9.1, средняя длина форм в символах в списке словоформ из словаря OpenCorpora (далее «словарь форм») — 10.4. Средняя длина единиц в слогах в «словаре лемм» — 3.5, в «словаре форм» — 4.5. Уникальная точка распознавания в среднем находится в «словаре лемм» на позиции 7.2 символов, в «словаре форм» — 10.5. Средняя позиция ударения в символах в «словаре лемм» — 5.4, в «словаре форм» — 5.5. Средняя позиция ударения в слогах в «словаре лемм» — 2.4, в «словаре форм» — 2.4.

Заключение

Созданная на материале целого ряда различных словарей и других электронных источников база данных StimulStat позволяет подбирать слова и формы по многочисленным параметрам, а также получать значения различных параметров для заданного списка слов и форм русского языка. Подобные базы созданы для целого ряда языков, однако для русского такой ресурс разработан впервые.

Литература

Бондарко Л. В. Звуковой строй современного русского языка. М., 1977.

Ефремова Т. Ф. Новый словарь русского языка. Толково-словообразовательный. М., 2000.

Зализняк А. А. Грамматический словарь русского языка. М., 1987.

Ляшевская О. Н. Частотный лексико-грамматический словарь: проспект проекта // Computational Linguistics and Intellectual Technologies. 2013. Т. 12. С. 478 – 489.

Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики. М.: 2009.

Andrews S. Lexical retrieval and selection processes: Effects of transposed-letter confusability // Journal of Memory and Language. 1996. Vol. 35. No. 6. P. 775 – 800. doi:10.1006/jmla.1996.0040

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. The English Lexicon Project // Behavior Research Methods. 2007. Vol. 39. No. 3. P. 445 – 459. doi:10.3758/bf03193014

Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V. Crowdsourcing morphological annotation // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 года). В 2-х т. / В. П. Селегей (Ed.). РГГУ, 2013. P. 109–114. URL: http://www.dialog-21.ru/media/1227/bocharovvv.pdf.

Boudelaa S., Marslen-Wilson W. D. Aralex: A lexical database for Modern Standard Arabic // Behavior Research Methods. 2010. Vol. 42. No. 2. P. 481–487. doi:10.3758/brm.42.2.481

Bowers J. S., Davis C. J., Hanley D. A. Automatic semantic activation of embedded words: Is there a "hat" in "that"? // Journal of Memory and Language. 2005. Vol. 52. No. 1. P. 131–143. doi:10.1016/j.jml.2004.09.003

Coltheart M., Davelaar E., Jonasson J. T., Besner D. Access to the internal lexicon // Attention and performance VI / S. Dornic (Ed.). New York, 1977. P. 535–555.

Davis C. J. N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics // Behavior Research Methods. 2005. Vol. 37. No. 1. P. 65–70. doi:10.3758/bf03206399

Davis C. J., Perea M. BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish // Behavior Research Methods. 2005. Vol. 37. No. 4. P. 665–671. doi:10.3758/bf03192738

Davis C. J., Perea M., Acha J. 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. 2009. Vol. 35. No. 5. P. 1550–1570. doi:10.1037/a0014253

Duchon A., Perea M., Sebastián-Gallés N., Martí A., Carreiras M. EsPal: One-stop shopping for Spanish word properties // Behavior Research Methods. 2013. Vol. 45. No. 4. P. 1246–1258. doi:10.3758/s13428-013-0326-1

Heister J., Würzner K.-M., Bubenzer J., Pohl E., Hanneforth T., Geyken A. dlexDB — eine lexikalische Datenbank für die psychologische und linguistische Forschung // Psychologische Rundschau. 2011. Vol. 62. No. 1. P. 10–20. doi:10.1026/0033-3042/a000029

Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts / M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V. Labunets (Eds.). Springer International Publishing, 2015. P. 320–332. doi:10.1007/978-3-319-26123-2_31

Skrelin P., Volskaya N., Kocharov D., Evgrafova K., Glotova O., Evdokimova V. CORPRES — Corpus of Russian professionally read speech // Proceedings of the 13th International Conference on Text, Speech and Dialogue. Berlin, 2010. P. 392–399.

StimulStat: A Tool for Selecting Psycholinguistic Stimuli in Russian

Alexeeva S.* (1), Slioussar N. (2, 1), Chernova D. (1)
mail@s-alexeeva.ru
1 — St. Petersburg State University, St. Petersburg;
2 — Higher School of Economics, Moscow

Abstract. The article presents the StimulStat database which includes more than 50,000 of the most frequent Russian words and their forms (> 1,700,000 units). These words are described in more than 70 different parameters that play an important role in linguistic and psychological research. The database can be used for stimuli selection in experimental studies of Russian. The database includes such parameters of lemmas and word forms as the number of letters and phonemes in the word, the number of syllables, frequency, phono-

logical information (ideal and real phonemic transcription, stress position, syllabic structure of the word), morphological information for all parts of speech (including the inflectional Zaliznyak index), semantic information (the number of meanings) and information about orthographical and phonological neighbors (substitution neighbors, transposition neighbors, addition/deletion neighbors). To provide free access to the database, a web interface (http://stimul.cognitivestudies.ru) has been created.

Keywords: psycholinguistic database, Russian