

Здорова Н.С.,<sup>1,2</sup> Гриднева Е.М.,<sup>1</sup> Иваненко А.А.,<sup>1</sup>  
Грабовская М.А.,<sup>1</sup> Макарова П.С.,<sup>1</sup> Ладинская Н.С.<sup>1</sup>

<sup>1</sup>Национальный исследовательский университет «Высшая школа экономики»,  
г. Москва, Россия

<sup>2</sup>Институт языкознания РАН, г. Москва, Россия

## База русских идиом: нормирование психолингвистических параметров (предварительные результаты)

*Исследование выполнено при поддержке конкурса проектов студентов и аспирантов факультета гуманитарных наук ВШЭ 2021-2022 гг.*

**Аннотация.** Настоящее исследование посвящено созданию Базы Русских Идиом (БРИ), которая рассматривается как необходимый инструмент для контролирования экспериментального материала в (психо)лингвистических исследованиях русской идиоматики. БРИ включает следующие объективные, то есть внешне доступные для описания параметры 376 идиом: частотность, синтаксическую структуру и длину. Кроме того, БРИ содержит субъективные рейтинги таких психолингвистических параметров, как встречаемость, знакомость и буквальность, подсчитанные на основании опроса 130 взрослых носителей русского языка. В данной работе представлены предварительные результаты нормирования перечисленных параметров, сбор данных продолжается. Также ведётся работа над нормированием параметра предсказуемости идиомы, который наравне с другими параметрами будет включён в финальную версию БРИ.

**Ключевые слова:** идиоматика, фразеология, русский язык, экспериментальный материал, контролирование стимулов.

### Введение

Под идиомой понимается выражение, значение которого не выводится из значения составляющих его компонентов и которое функционирует в языке и речи как самостоятельная единица (Abel, 2003; Hornby et al., 1999; Irujo, 1986). Основная трудность исследования идиом заключается в большом разнообразии идиоматических выражений по структуре, семантике, длине и их функционированию в речи (изменяемость, дополняемость и пр.). Разработанные для русского языка классификации идиом В.В. Виноградова (Виноградов, 1977) и Н.М. Шанского (Шанский, 1985) опираются частично на семантический, а частично на морфологический принцип. Однако сами описания классификаций охватывают лишь малую часть идиоматических выражений русского языка и, вместе с тем, не содержат строгих формальных параметров, опираясь на которые можно с лёгкостью отнести не упомянутые в классификации идиомы в ту или иную категорию. Более того, психолингвистические исследования языковой обработки и, в частности, обработки идиом показывают важность таких психолингвистических параметров, как частотность, буквальность (или фигуративность), известность носителям (Cronk et al., 1993; Glass, 1983). Подобное нормирование и описание идиом существует для английского и немецкого языков (Beck & Weber, 2016, 2020; Libben & Titone, 2008), однако идиомы русского языка ранее не были описаны с данной точки зрения, в то время как вопросы изучения русских идиом остаются актуальными (Gridneva, 2018; Gridneva et al., 2022; Kopotev et al., 2020).

Таким образом, настоящее исследование ставит своей целью закрыть эту лауну путём создания нормированной Базы Русских Идиом (БРИ) с приведением их психолингвистических параметров. Параметр встречаемости означает то, как часто носитель языка, на его субъективный взгляд, встречает в жизни ту или иную идиому.

Предположительно, встречаемость коррелирует с частотностью. Однако мы не ограничивали поиск по частотности временем создания текстов и/или жанрами, в связи с чем параметр встречаемости - это скорее субъективная мера частотности среди носителей современного русского языка. Знакомость идиомы означает то, насколько легко участник может объяснить значение идиомы. Наконец, буквальность означает степень, с которой можно использовать идиому в её прямом, то есть буквальном значении. В данной работе представлены промежуточные результаты создания базы: подсчёт частотности и длины, описание синтаксических структур, а также предварительные результаты опроса взрослых носителей русского языка для нормирования встречаемости, знакомости и буквальности.

### Метод

Для создания БРИ были отобраны 376 идиом из словаря русской идиоматики (Баранов, 2009), которые не содержали пометы *жарг.*, *сниж.*, *журн.*, *книжн.*, *совет.*. Таким образом, БРИ состоит из идиом, употребление которых не ограничено определенной сферой. Частотность идиом была подсчитана на основе двух корпусов: Национального корпуса русского языка (НКРЯ, [www.ruscorpora.ru](http://www.ruscorpora.ru)) и SketchEngine (подкорпус ruTenTen11, <https://www.sketchengine.eu/>) и считалась как общее количество вхождений в корпусе, а также как количество вхождений на миллион. Далее идиомы были размечены по структуре. Наконец, для каждой идиомы была подсчитана длина в словах и символах.

Для сбора нормативных значений встречаемости, знакомости и буквальности был проведён опрос взрослых носителей русского языка ( $N = 130$ , 103 женщины, 24 мужчины и 3 с ответом “иное”, диапазон возрастов 19-75,  $M_{age} = 31.19$ ,  $SD_{age} = 11.13$ ). Для удобства проведения опроса идиомы были разделены на четыре экспериментальных листа по 94 идиомы в каждом. Листы были сбалансированы по структуре идиом. Распределение участников по экспериментальным листам было следующим: 34 человека на 1 лист, 33 на 2 лист, 37 на 3 лист и 26 на 4 лист соответственно). Среднее количество лет образования составило 16.42 лет (в диапазоне 9-21,  $SD = 2.16$ ), 46% участников имели лингвистическое образование и 96.9% владели как минимум одним иностранным языком (из них подавляющее большинство владело английским языком — 91.5%). По месту проживания 0.8% участников пришлось на село, 2.3% - на посёлок городского типа, 25.4% проживали в городе с населением меньше одного миллиона и 71.5% участников проживали в городе-миллионнике.

Опрос проводился онлайн на платформе Google forms. Встречаемость, знакомость и буквальность идиом оценивались участниками по 7-балльной шкале от 1 до 7. Инструкции для оценки каждого параметра приведены ниже.

Встречаемость: “Для каждого выражения в этом разделе вам нужно оценить, как часто вы видели или слышали выражение, независимо от того, знаете ли вы, что оно означает. Поставьте свои оценки по шкале от 1 до 7, где 1 - никогда не видел(а)/слышал(а), 7 - часто вижу/слышу”.

Знакомость: “Оцените те же выражения по шкале от 1 до 7, в зависимости от того, насколько хорошо вы знаете их значение. Оценка 1 будет означать, что вы абсолютно не представляете, что означает это выражение, 4 — что вы вполне уверены в том, что оно означает, а 7 — что вы на 100% уверены в значении и можете легко передать значение своими словами.”

Буквальность: “Для каждого выражения мы просим вас сделать суждение о буквальности. Хотя все идиомы имеют образную интерпретацию, только некоторые из них имеют правильно сформированное буквальное значение. Например, идиома «белая ворона» в переносном смысле означает «отличающегося от общества человека», а буквально — ворону белого цвета. Однако идиома «чесать языками», которая в переносном смысле означает “болтать, сплетничать”, не имеет столь же ясного буквального значения (если вообще имеет) по сравнению с «белой вороной» (например, неясно, что означает буквально чесать языком).

Ваша задача при оценке этих выражений состоит в том, чтобы решить, возможно ли буквальное понимание, и если да, то насколько оно возможно по 7-балльной шкале. Оценка 1 будет означать, что идиома определенно не имеет какой-либо возможной буквальной интерпретации (чесать языками). Оценка 7 будет означать, что идиома определенно имеет четкую и хорошо построенную буквальную интерпретацию (белая ворона), которая вполне адекватна. Промежуточные значения шкалы должны отражать ваши суждения о возможности появления этих выражений в прямом значении.”

### Предварительные результаты

Средняя частотность идиом на миллион по НКРЯ составляет 0.74 (диапазон 0.003-13.73, SD=1.53), по SketchEngine — 0.29 (диапазон 0.0001-13.09, SD=0.83). Пример среднечастотной идиомы — *на вес золота* (с частотностью 0.74 по НКРЯ и 0.68 по SketchEngine); низкочастотных идиом: *кричать как полоумный*, *обагрить руки кровью*, *рёва-корова* (с частотностью ниже 0.00 по НКРЯ и SketchEngine); и высокочастотной идиомы — *рано или поздно* (с частотностью 13.73 по НКРЯ и 13.09 по SketchEngine).

По синтаксической структуре идиомы были распределены на 22 категории от 8 до 64 вхождений в каждой (M = 17.09, SD = 13.98). Категории выделялись по типу вершины идиомы. Например, *деваться некуда* — глагольная группа, *белая ворона* — именная группа, *вверх тормашками* — адвербиальная, *до упаду* - предложная группа и т.д. Крупные категории были разбиты на более мелкие: например, глагольная группа включает 12 подкатегорий в зависимости от типа дополнения (прямое или косвенное), от падежа и предлога. Подробное описание каждой синтаксической категории с примерами и количеством вхождений размещено в открытом доступе на платформе [OSF](#).

Длина идиом, подсчитанная в словах, варьирует от 1 до 5 слов (M = 2.75, SD = 0.79), в символах (включая буквы, пробелы и знаки пунктуации) — от 6 до 30 символов (M = 14.05, SD = 4.03). Короткие идиомы, состоящие из одного слова, в базе встречаются 2 раза: *по-чёрному*, *худо-бедно*. Примерами идиом средней длины (2-3 слова) могут служить *дух захватывает*, *зарубить на носу*. Примерами длинных идиом являются *и на старуху бывает проруха*, *дать обе руки на отсечение*.

Средняя встречаемость идиом в базе составляет 5.17 (SD = 0.88), пример средневстречаемой идиомы — *вариться в собственном соку* (5.17). Диапазон встречаемости — от 2.24 (*пыль столбом*) до 6.91 (*разорвать в клочья*). Средняя знакомость составила 6.41 (SD = 0.51), пример такой идиомы — *отслужить своё* и охватывает диапазон от 3.71 (*пыль столбом*) до 6.97 (*пропащая душа*). Средняя буквальность составила 4.33 (SD = 1.4), пример идиомы с близким рейтингом — *идти на поводу* (4.38). Диапазон буквальности — от 1.58 с идиомой *на седьмом небе* и до 6.8 с идиомой *стоять горой*. Описательные характеристики всех перечисленных параметров сведены в Таблицу 1. База русских идиом на этапе разработки к сентябрю 2022 г. доступна в открытом доступе на платформе [OSF](#).

**Таблица 1. Описательные характеристики частотности (число вхождений на миллион по двум корпусам), длины (в символах и словах), встречаемости, знакомости и буквальности.**

	НКРЯ _ipm	SE_ipm	Len_sym	Len_w	Встречаемость	Знакомость	Буквальность
Mean	0.78	0.33	14.94	2.75	5.15	6.39	4.31
Median	0.25	0.08	14.50	3.00	5.24	6.58	4.34
SD	1.67	1.06	4.12	0.80	0.92	0.61	1.41
Min	0.00	0.00	4.05	0.79	0.88	0.51	1.40
Max	13.73	13.09	30	5	6.91	6.97	6.80

*Примечание: НКРЯ\_ірт — частотность идиомы на миллион в Национальном корпусе русского языка, SE\_ірт — частотность идиомы на миллион в корпусе Sketch Engine, Len\_сyt — длина идиомы в символах, Len\_w — длина идиомы в словах.*

### **Выводы и заключение**

Настоящее исследование ставило своей целью создание базы русских идиом (БРИ) с описанием психолингвистических параметров, важных при контроле экспериментального материала в исследованиях русской идиоматики. В данной работе были описаны предварительные результаты создания базы из 376 идиом, а именно подсчитаны такие параметры идиом, как частотность по двум корпусам и длина идиом в словах и символах. Также все идиомы были размечены по синтаксическим категориям, выделенным по синтаксической вершине. На основании опроса взрослых носителей русского были подсчитаны предварительные нормативные значения для параметров встречаемости, знакомости и буквальности.

БРИ сопоставима с аналогичными базами идиом на английском и немецком языках (Beck & Weber, 2016, 2020; Libben & Titone, 2008), однако её отличает наличие разметки по синтаксическим категориям. По аналогии с зарубежными базами в БРИ войдут также нормативные данные по параметру предсказуемости идиомы. Предсказуемость покажет процент респондентов, которые смогли закончить выражение как существующую идиому. Кроме того, будет учтено, с какого именно слова в идиоматическом выражении наибольшее число участников заканчивало её корректно. Сбор данных продолжается и будет закончен при достижении 100 участников на каждый экспериментальный лист. Такое распределение позволит обоснованно назвать данные базы нормативными и репрезентативными. Стоит отметить, что БРИ имеет широкое практическое применение и может быть ресурсом как для психолингвистических экспериментов, так и для исследований в области корпусной и компьютерной лингвистики.

### **Литература**

1. Abel, B. English Idioms in the First Language and Second Language Lexicon: A Dual Representation Approach // *Second Language Research*. 2003. Vol. 19. P. 329–358. <https://doi.org/10.1191/0267658303sr226oa>
2. Beck, S. D., & Weber, A. Bilingual and Monolingual Idiom Processing Is Cut from the Same Cloth: The Role of the L1 in Literal and Figurative Meaning Activation // *Frontiers in Psychology*. 2016. Vol. 7. P. 1350. <https://doi.org/10.3389/fpsyg.2016.01350>
3. Beck, S. D., & Weber, A. Context and Literality in Idiom Processing: Evidence from Self-Paced Reading // *Journal of Psycholinguistic Research*. 2020. Vol. 49(5). P. 837–863. <https://doi.org/10.1007/s10936-020-09719-2>
4. Cronk, B. C., Lima, S. D., & Schweigert, W. A. Idioms in sentences: Effects of frequency, literalness, and familiarity // *Journal of Psycholinguistic Research*. 1993. Vol. 22 (1). P. 59–82. <https://doi.org/10.1007/BF01068157>
5. Glass, A. L. The comprehension of idioms // *Journal of Psycholinguistic Research*. 1983. Vol. 12(4). P. 429–442. <https://doi.org/10.1007/BF01067624>
6. Gridneva, E. Lexical Knowledge in Heritage Russian: Linguistic and Cultural Aspects // *Актуальные проблемы и перспективы русистики*. Universitat de Barcelona, 2018. P. 571–579.
7. Gridneva, E., Grabovskaya, M., Ivanenko, A., Zdorova, N., & Ladinskaya, N. The Processing of Russian Idioms in Heritage and L2 Russian Speakers. 2022 (in press).
8. Hornby, A. S., Crowther, J., Kavanagh, K., & Ashby, M. *Oxford advanced learner's dictionary of current English*. Oxford University Press. 1999
9. Irujo, S. A piece of cake: Earning and teaching idioms // *Elt Journal*. 1986. Vol. 40. P. 236–243. <https://doi.org/10.1093/elt/40.3.236>
10. Kopotev, M., Kisselev, O., & Polinsky, M. Collocations and near-native competence: Lexical strategies of heritage speakers of Russian // *International Journal of Bilingualism*. 2020. 1367006920921594. <https://doi.org/10.1177/1367006920921594>
11. Libben, M. R., & Titone, D. A. The multidetermined nature of idiom processing // *Memory & Cognition*. 2008. Vol. 36(6). P. 1103–1121. <https://doi.org/10.3758/MC.36.6.1103>

12. Баранов, А. Н. Словарь-тезаурус современной русской идиоматики. 2009
13. Виноградов, В. В. Об основных типах фразеологических единиц в русском языке // Избранные труды. Лексикология и лексикография. 1977. Р. 140–161.
14. Шанский, Н. М. Фразеология современного русского языка: Учебное пособие для студентов высших учебных заведений, обучающихся по специальности 'Русский язык и литература' (3rd ed.). Москва - Высшая школа. 1985