САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2021»

1–3 июля 2021 г., Санкт-Петербург

СКИФИЯ
принт
Санкт-Петербург
2021

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Т78 **Труды международной конференции «Корпусная лингвистика–2021».** — СПб.: Издательство Скифия-принт, 2021. — 396 с.

ISSN 2412-9623
ISBN 978-5-98620-557-1

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика-2021» 1–3 июля 2021 г. в Санкт-Петербурге.

Создание и использование корпусов текстов является одним из приоритетных направлений в современной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

ББК 81.1

ISSN 2412-9623
ISBN 978-5-98620-557-1

В. И. Зубов, Е. И. Риехакайнен

V. I. Zubov, E. I. Riekhakaynen

ВМЕСТЕ ИЛИ ВРОЗЬ: НЕОДНОСЛОВНЫЕ ЕДИНИЦЫ В КОРПУСАХ И В МЕНТАЛЬНОМ ЛЕКСИКОНЕ НОСИТЕЛЯ РУССКОГО ЯЗЫКА[1]

TOGETHER OR NOT: MULTIWORD UNITS IN CORPORA AND IN THE MENTAL LEXICON OF A RUSSIAN SPEAKER

Аннотация. В статье описывается первый этап исследования, направленного на определение статуса неоднословных единиц в ментальном лексиконе носителя русского языка. На материале справочников по русской орфографии и текстов Национального корпуса русского языка начиная с 1956 года составлен список из наиболее употребительных двусловных сочетаний, имеющих пары среди однословных единиц. Этот список (201 пара) может использоваться в практике преподавания русского языка. В качестве наиболее вероятных кандидатов на вхождение в ментальный лексикон носителя русского языка выделены семь неоднословных сочетаний, которые являются высокочастотными или сопоставимы по частотности со своими однословными вариантами. Следующим этапом исследования станет анализ фонетических реализаций отобранных единиц в естественной русской устной речи.
Ключевые слова. Неоднословные единицы, русский язык, ментальный лексикон, Национальный корпус русского языка, орфография.

Abstract. The paper describes the first stage of a study of multiword units in the mental lexicon of a Russian speaker. Based on the guides on Russian spelling and the texts from the Russian National Corpus since 1956, we compiled a list of 201 most common two-word combinations that have pairs among one-word units. Seven two-word units, which are high-frequency or comparable in frequency with their one-word counterparts, were considered the most likely candidates for entering the mental lexicon of a Russian speaker. The next step will be the analysis of the phonetic realizations of these units in natural Russian speech.
Keywords. Multiword units, Russian, mental lexicon, Russian National Corpus, orthography.

1. Предпосылки исследования

Для решения одной из наиболее актуальных задач современной психолингвистики — описания структуры ментального лексикона (внутреннего словаря) носителя языка — необходимо среди прочего ответить на вопрос о единицах ментального лексикона. Представляется, что методы корпусной лингвистики и существующие на данный момент корпусы устных и письменных текстов могут быть использованы для поиска ответа на этот вопрос. В статье будет описан начальный этап исследования, посвященного определению статуса некоторых неоднословных единиц в ментальном лексиконе носителя русского языка.

[1] Исследование выполнено при поддержке гранта РФФИ № 19-012-00629.

В работах по порождению и восприятию речи активно обсуждается, могут ли самостоятельными единицами ментального лексикона являться сочетания нескольких слов. В качестве кандидатов в такие единицы предлагаются, например, коллокации, т. е. сочетания знаменательных слов, которые характеризуются частичной невыводимостью [Ellis, Simpson-Vlach 2009 и др.], или т.н. составные слова / эквиваленты слова, которые представляют собой сочетания нескольких орфографических слов, но при этом функционально приближаются к словам и часто произносятся как единое целое (например, *потому_что, то_есть* и др.) (см. обзор русскоязычных работ в [Мустайоки, Копотев 2004]).

На наш взгляд, еще одной группой неоднословных единиц, которые хранятся в ментальном лексиконе носителя языка как самостоятельные единицы, могут быть сочетания двух и более орфографических слов, имеющие пары среди однословных единиц (и *так — итак*, *на счет — насчет*), т. е. отличающиеся от последних только наличием пробела. Можно предполагать, что фонетическая близость подобных сочетаний к отдельным словам будет способствовать их объединению в одну единицу с соответствующими однословными единицами на уровне перцептивного словаря, т. е. того уровня ментального лексикона, на котором представлен фонетический облик слов (подробнее см. в [Венцов 2007]). Для проверки этого предположения необходимо проведение комплексного исследования, соединяющего в себе методы корпусной лингвистики и психолингвистики, а именно нужно выяснить, действительно ли пары, различающиеся только наличием/отсутствием пробела, всегда реализуются одинаково в естественной устной речи, и если это так, как происходит выбор необходимой интерпретации в процессе восприятия речи.

Исследования на материале русского языка, в которых сопоставляются именно пары «одно графическое слово — два графических слова», остаются единичными. Насколько мы можем судить, на данный момент в русскоязычной традиции отсутствует даже устоявшийся термин для обозначения подобных пар единиц. В [Ягунова 2008] такие пары выделяются в особые фонетические слова, имеющие омоним. В поэзии для близкого явления используется термин «пантограмма», когда строки полностью совпадают по буквенному составу, но различаются расстановкой пробелов (*ночей — но чей*; *задело — за дело*) [Бубнов 2002]. Если принимать во внимание орфографическую близость таких пар, то их можно отнести к т. н. орфографическим со-

седям — близким по написанию словам, которые отличаются только одним графическим элементом, а именно наличием/отсутствием пробела. Однако обычно к соседям относятся слова с перестановкой, добавлением или удалением букв; обсуждаемые же единицы на данный момент не включены в базу данных орфографических соседей в русском языке [Алексеева, Слюсарь 2017]. Таким образом, первым этапом нашего исследования стало формирование списка слов, которые станут объектом изучения.

2. Принципы формирования списка

В качестве источников для формирования первоначального списка неоднословных единиц, которые могут претендовать на самостоятельность в ментальном лексиконе, мы воспользовались справочниками по русской орфографии [Лопатин 2009; Розенталь 2011], где перечислены слова, слитное/раздельное написание которых вызывает сложности у носителей языка (*вбок — в бок; помногу — по многу*). В первую очередь нас интересовали служебные части речи, а также наречия и сочетания самостоятельных слов с предлогами. Можно предположить, что возникновению затруднений в выборе корректного написания способствует как раз то, что сочетание двух орфографических слов, совпадающее по буквенному составу с одним словом, претендует на некоторую степень устойчивости в ментальном лексиконе носителей языка. На данном этапе мы сосредоточились только на слитном или раздельном написании, поэтому слова с дефисным написанием не учитывались.

Слова из орфографических справочников были дополнительно проверены по Орфографическому академическому ресурсу «АКАДЕМОС» (http://orfo.ruslang.ru/). Так, пары типа в *общем — вобщем, по-суху — по суху в* список включены не были, потому что нормативным является единственный вариант написания. Для пар же типа *вперегиб — в перегиб, внакладку — в накладку* слитное написание является нормативным, а также существуют слова типа *перегиб*, *накладка*, которые могут быть употреблены с предшествующим предлогом. Всего в список была включена 271 пара, имеющая слитный или раздельный варианты написания. Часть из них однозначно по-разному реализуется в устной речи, поскольку единицы, образующие пару, имеют разные ударения (*впервые — в первые; втихую — в тихую*) — таких пар 47, они были исключены из дальнейшего анализа. Все оставшиеся в списке

единицы были проверены на наличие в основном подкорпусе Национального корпуса русского языка (далее — НКРЯ) начиная с 1956 года, то есть с года принятия Правил русской орфографии и пунктуации. Для части низкочастотных слов мы не встретили ни одного словоупотребления, например: *вперегиб — в перегиб*, *вперегонку — в перегонку*, *вприхватку — в прихватку*, *изнизу — из низу*; в некоторых парах одно из слов было представлено в корпусе, другое — нет: *навырез* (0)[2] — *на вырез* (3), *наудалую* (5) — *на удалую* (0), *навыкат* (29) — *на выкат* (0). Всего таких пар было 23, и они также были исключены из анализа.

Оставшаяся 201 пара вошла в основной список единиц, реализация которых в устной речи представляет интерес для дальнейшего исследования. Хотя частеречный анализ единиц не входил в задачи нашей работы на данном этапе, можно отметить, что в группу слов со слитным написанием вошли предлоги (*ввиду*, *насчет*), союзы (*зато*, *тоже*), наречия (*впустую*, *наутро*), вводные слова (*например*), а также слова, которые могут являться несколькими частями речи (*навстречу* — предлог, наречие, *оттого* — союз, наречие). В группе слов с раздельным написанием подавляющее большинство — сочетания самостоятельных слов с предлогами *в*, *до*, *за*, *из*, *на*, *от*, *под*, *по*, *при*, *с*, а также местоимения с частицами или союзами и *так*, *так же*, *что бы*. Слова со слитным написанием встречаются в корпусе значительно чаще слов с раздельным написанием (среднее значение частотности — 64,52 ipm и 6,59 ipm соответственно). Отметим, однако, что подобное соотношение справедливо для 161 пары (более частотными являются, например, *чтобы*, *потом*, *сейчас*, *причем*, *например* и др.), для остальных 40 раздельный вариант представлен чаще (*с ходу, с плеча, в начале*, *от того* и др.).

Одна из проблем, с которой мы столкнулись при подсчете частотности, — это орфографические ошибки в корпусе. Например, для единицы *на долго* из 25 случаев употребления встретилось всего пять с верным написанием (например, *похожим* ***на долго*** *показываемую фигу*), остальные 20 — орфографические ошибки (*там* ***на долго*** *зависать нельзя*), для единицы *на вынос* из 53 случаев употребления лишь 18 не содержат ошибки, остальные 35 — это некорректное написание наречия *навынос*. Подобные случаи, с одной стороны, свидетельствуют о сложности выбора верной орфографической формы для таких единиц, с другой стороны, могут быть косвенным подтверждением

[2] В скобках указано количество вхождений в подкорпус объемом 163 271 973 с/у.

того, что близкие по написанию слова представляют собой одну единицу перцептивного словаря. Отметим также, что для корректного определения частотности по основному подкорпусу НКРЯ необходимо исключить подобные ошибки, что весьма трудоемко в случае высокочастотных слов вроде *зато*, *притом* и т. п., поэтому решено было вычислить частотность и по подкорпусу НКРЯ со снятой омонимией, также начиная с 1956 года (объем корпуса — 4 759 671 с/у), где, как можно ожидать, все ошибки были устранены в ходе разрешения грамматической неоднозначности. Кроме того, мы указали для каждой единицы частотность в текстах устного подкорпуса НКРЯ за тот же период (целиком — 12 538 108 с/у и только со снятой омонимией — 205 994 с/у), что позволит сопоставить частоту употребления той или иной единицы в устной и письменной речи. Получившийся список, на наш взгляд, имеет самостоятельную практическую ценность: он может использоваться в практике преподавания русского языка, так как наглядно показывает, на какие пары сложных для написания единиц из представленных в справочниках по русской орфографии стоит обращать внимание при обучении русской орфографии в первую очередь в силу их высокой частотности. С перечнем единиц можно ознакомиться по ссылке: https://osf.io/bsy2t/.

3. Неоднословные сочетания — кандидаты в единицы ментального лексикона

Поскольку частотность является одним из ключевых факторов, влияющих на структуру ментального лексикона (см. обзор в [Риехакайнен 2016: 51–56]), мы предположили, что прежде всего на статус самостоятельных единиц ментального лексикона должны претендовать: 1) высокочастотные неоднословные единицы; 2) неоднословные единицы, которые сопоставимы по частотности с их однословными «соседями».

1) Высокочастотными мы считали двусловные сочетания из нашего списка, ipm которых хотя бы по одному из корпусов выше 100. Таких сочетаний оказалось четыре: *так же*, *и так*, *то же*, *в начале*, причем последнее имеет необходимую частотность только в основном подкорпусе. При этом соотношения по частотности внутри пар с этими сочетаниями оказались различными и в ряде случаев зависят от типа корпуса. Так, неоднословное сочетание частотнее однословного в парах *итак — и так* и *вначале — в начале*, но во второй паре

единицы в устной речи более близки по частотности, чем в письменной. В паре *тоже — то же* во всех подкорпусах намного частотнее однословный вариант. В паре *также — так же* в письменных текстах преобладает однословный вариант. В устном подкорпусе оба варианта практически не различаются по частотности, но если анализировать только устные тексты со снятой омонимией, неоднословный вариант будет иметь более высокую частотность.

2) Мы считали, что однословная и неоднословная единицы в паре имеют сопоставимую частотность, если их показатели ipm различались не более чем в два раза. Таких пар оказалось 24, но большинство из них являются редкими: только в трех парах (*зато — за то, оттого — от того* и *притом — при том*) вариант с раздельным написанием имеет ipm больше 10 хотя бы одном устном и письменном подкорпусе. Таким образом, мы включили в итоговый список неоднословных единиц, которые могут быть представлены в ментальном лексиконе носителя языка в качестве целостных единиц, семь сочетаний (см. табл. 1).

Таблица 1. **Список единиц для дальнейшего фонетического анализа**

Единица	ОП, ipm	УП, ipm	ОС, ipm	УС, ipm
также	499,52	190,06	451,92	179,62
так же	172,35	189,74	171,65	199,03
итак	64,80	98,26	55,68	58,25
и так	167,57	581,51	211,57	422,34
тоже	819,57	2003,57	959,52	1995,20
то же	159,68	229,38	144,34	237,87
вначале	33,10	35,01	29,83	29,13
в начале	101,35	71,30	72,06	38,84
зато	130,67	74,33	132,36	53,40
за то	66,90	89,25	75,43	77,67
оттого	50,33	21,77	52,10	9,71
от того	53,12	74,25	50,84	101,94
притом	17,59	15,39	19,12	19,42
при том	14,75	14,44	12,82	29,13

ОП, УП — основной и устный подкорпусы; ОС, УС — основной и устный подкорпусы со снятой омонимией.

Следующим этапом исследования станет сопоставление того, как двусловные сочетания и соответствующие им однословные единицы реализуются в устной речи. Если результаты фонетического анализа покажут, что однословные и неоднословные единицы в каждой паре имеют одинаковые варианты произнесения, то в дальнейшем в психолингвистических экспериментах можно будет проверить гипотезу о влиянии соотношения по частотности внутри каждой пары на то, каким образом носитель языка интерпретирует такие единицы при восприятии их на слух.

Литература

1. *Алексеева С. В., Слюсарь Н. А.* (2017), Орфографические соседи в русском языке: база данных и эксперимент, направленный на изучение морфологической декомпозиции. Вопросы психолингвистики. № 32, с. 12–27.
2. *Бубнов А. В.* (2002), Палиндромия: от перевертня до пантограммы. Новое литературное обозрение. № 57, с. 295–312.
3. *Венцов А. В.* (2007), Восприятие устной речи и ментальный лексикон. Русская языковая личность: Материалы шестой выездной школы-семинара. Череповец, с. 63–69.
4. *Лопатин В. В.* (ред.) (2009), Правила русской орфографии и пунктуации. Полный академический справочник. М.
5. *Мустайоки А., Копотев М.* (2004), К вопросу о статусе эквивалентов слова типа *потому что*, в *зависимости от*, к *сожалению.* Вопросы языкознания. № 3, с. 88–107.
6. *Риехакайнен Е. И.* (2016), Восприятие русской устной речи: контекст + частотность. СПб.
7. *Розенталь Д. Э.* (2011), Русский язык. Орфография и пунктуация. М.
8. *Ягунова Е. В.* (2008), Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь.
9. *Ellis N. C., Simpson-Vlach R.* (2009), Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. In: Corpus Linguistics and Linguistic Theory. Vol. 5, pp. 61–78.

References

1. *Alexeeva S. V., Slioussar N. A.* (2017), Orfograficheskie sosedi v russkom yazyke: baza dannyh i eksperiment, napravlennyj na izuchenie morfologicheskoj dekompozicii [Orthographic Neighbors in Russian: a Database and an Experiment Aimed at Studying Morphological Decomposition]. In: Voprosy Psiholingvistiki [Journal of Psycholinguistics]. No. 32, pp. 12–27.
2. *Bubnov A. V.* (2002), Palindromiya: ot Perevertnya do Pantogrammy [Palindromia:

From Shifter to Pantogram]. In: Novoe Literaturnoe Obozrenie [New Literary Observer]. No. 57, pp. 295–312.
3. *Ellis N. C., Simpson-Vlach R.* (2009), Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. In: Corpus Linguistics and Linguistic Theory. No. 5, pp. 61–78.
4. *Lopatin V. V.* (ed.) (2009), Pravila russkoj orfografii i punktuacii. Polnyj akademicheskij spravochnik [Rules of Russian spelling and punctuation. The Complete Academic Guide]. Moscow.
5. *Mustajoki A., Kopotev M.* (2004), K voprosu o statuse ekvivalentov slova tipa *potomu chto*, *v zavisimosti ot*, *k sozhaleniyu* [On the Status of Word-Equivalents of the Type *potomu chto*, *v zavisimosti ot*, *k sozhaleniyu*]. In: Voprosy Yazykoznaniya [Topics in the study of language]. No. 3, pp. 88–107.
6. *Riekhakajnen E. I.* (2016), Vospriyatie russkoj ustnoj rechi: kontekst + chastotnost' [Perception of Russian Oral Speech: Context + Frequency]. Saint Petersburg.
7. *Rozental' D. E.* (2011), Russkij yazyk. Orfografiya i punktuaciya [Russian. Orthography and Punctuation]. Moscow.
8. *Ventsov A. V.* (2007), Vospriyatie Ustnoj Rechi i Mental'nyj Leksikon [Oral Speech Perception and Mental Lexicon]. In: Russkaya Yazykovaya Lichnost': Materialy Shestoj Vyezdnoj Shkoly-Seminara [Russian Language Personality: Materials of the sixth field school-seminar]. Cherepovets, pp. 63–69.
9. *Yagunova E. V.* (2008), Variativnost' Strategij Vospriyatiya Zvuchashchego Teksta (Eksperimental'noe issledovanie na materiale russkoyazychnyh tekstov raznyh funkcional'nyh stilej) [Strategies Variability of the Oral Text Perception (An Experimental Study of Russian-Language Texts of Different Functional Styles)]. Perm'.

Зубов Владислав Иванович
Санкт-Петербургский государственный университет (Россия)
Zubov Vladislav
Saint Petersburg State University (Russia)
E-mail: v.zubov@spbu.ru

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru