УДК 81’23 Вестник СПбГУ. Язык и литература. 2023. Т. 20. Вып. 4

Климова Маргарита Андреевна

Национальный исследовательский университет «Высшая школа экономики»,
Россия, 603155, Нижний Новгород, ул. Большая Печерская, 25/12
mfokina@hse.ru

Виклова Анна Валерьевна

Российская академия народного хозяйства и государственной службы
при Президенте Российской Федерации,
Россия, 119571, Москва, пр. Вернадского, 82
viklova-av@ranepa.ru

Оверникова Дарья Алексеевна

Национальный исследовательский университет «Высшая школа экономики»,
Россия, 101000, Москва, ул. Мясницкая, 20
daria.a.overnikova@gmail.com

Обработка слов с частотными орфографическими ошибками (исследование на базе учебного корпуса английского языка)*

Для цитирования: Климова М. А., Виклова А. В., Оверникова Д. А. Обработка слов с частотными орфографическими ошибками (исследование на базе учебного корпуса английского языка). *Вестник Санкт-Петербургского университета. Язык и литература.* 2023, 20 (4): 824–837. https://doi.org/10.21638/spbu09.2023.409

Статья посвящена экспериментальному исследованию влияния частотности орфографических ошибок в слове на качество его репрезентации в ментальном лексиконе. Гипотеза о том, что слова, в которых часто совершаются ошибки правописания, затрудняют восприятие при чтении, даже если написаны правильно, ранее была подтверждена для носителей русского и английского языков. Данная работа нацелена на проверку гипотезы на материале учебного корпуса REALEC (Russian Error-Annotated Learner English Corpus), включающего тексты русскоязычных студентов, изучающих английский язык. Из корпуса были отобраны слова, наиболее часто пишущиеся неверно, которые стали материалом для эксперимента на распознавание верных и неверных написаний. При интерпретации результатов оценивалось влияние на успешность распознавания таких факторов, как частотность ошибок, частотность слова в учебном корпусе, энтропия (мера, отражающая количество усилий, затрачиваемых на выбор между вариантами написания) и тип ошибки. Согласно полученным данным, значимыми оказались факторы энтропии и частотности слова в корпусе, что соответствует результатам предыдущих исследований. Значимость продемонстрировал и конкретный

* Исследование выполнено в рамках исследовательского проекта Национального исследовательского университета «Высшая школа экономики» «Разработка моделей усвоения второго языка в рамках существующих теорий с помощью методик автоматизации экспериментов на основе материалов корпуса REALEC и других учебных корпусов».

тип ошибки — замена буквы. Особая роль данного типа проявляется в затруднениях, которые ошибки замены вызвали у респондентов при восприятии, и соответствует результатам предшествующих исследований производства письменной речи изучающих английский язык, согласно которым данный тип неверных написаний является самым частотным. Меньшая значимость фактора частотности неверного написания по сравнению с исследованиями на базе родного языка может объясняться различиями в языковой среде: так, изучающие язык реже сталкиваются с неверными написаниями.

Ключевые слова: обработка слов, орфографические ошибки, ментальный лексикон, учебный корпус.

Введение

Исследование посвящено проблеме влияния частотности орфографических ошибок в слове на качество его репрезентации в ментальном лексиконе. Предполагается, что слова, в которых часто совершаются ошибки правописания, затрудняют восприятие при чтении, даже если написаны правильно. Данная гипотеза уже была подтверждена для носителей языка на материале английского [Rahmanian, Kuperman 2019] и русского [Чернова и др. 2020а].

Цель данного исследования — проверить гипотезу на материале английского как второго языка, изучаемого русскоязычными студентами. Для достижения цели проведен эксперимент на успешность распознавания респондентами правильных и неправильных написаний слов с частотными ошибками. Материалом для эксперимента послужили слова с частотными орфографическими ошибками из учебного корпуса REALEC, включающего студенческие работы на английском языке. При интерпретации результатов оценивалось влияние на успешность распознавания верных и неверных написаний таких факторов, как частотность ошибок, частотность слова в учебном корпусе, энтропия и тип ошибки. В фокусе данной работы следующие исследовательские вопросы:

1. В каких словах наиболее часто совершаются орфографические ошибки (на материале учебного корпуса REALEC)?
2. Влияют ли на успешность распознавания верных и неверных написаний русскоязычными студентами, изучающими английский язык, следующие факторы:
   а) частотность неверного написания слова в корпусе;
   б) общая частотность слова в корпусе;
   в) количество вариантов неверного написания слова в корпусе;
   г) тип орфографической ошибки?

Гипотеза лексического качества и экспериментальные исследования орфографической обработки слова

В исследованиях отражения слова в ментальном лексиконе важную роль играет гипотеза лексического качества, Lexical Quality Hypothesis [Perfetti 1985; Perfetti, Hart 2001; Perfetti 2007]. Согласно ей, качество лексической репрезентации обеспечивается правильностью (precision) и гибкостью, взаимосвязанностью (flexibility)

следующих компонентов: орфографического, фонологического, грамматического и семантического [Perfetti 2007]. Орфографический компонент занимает особое место, и его оценка через измерение орфографических навыков является значимым способом оценки качества лексической репрезентации [Andrews et al. 2020]. Инструмент для измерения устойчивости графического представления слова в памяти для английского языка был разработан С. Эндрюс и коллегами. Тест включает задания как на производство письменной речи (диктант), так и на чтение слов с целью определить правильные и неправильные написания (spelling recognition test) [Andrews et al. 2020]. Среди российских исследований, посвященных изучению орфографической обработки слов при чтении, необходимо отметить работы Н. А. Слюсарь, С. В. Алексеевой и Д. А. Черновой [Алексеева, Слюсарь 2017; Чернова и др. 2020б; Чернова 2022].

Вопрос зависимости качества орфографической репрезентации слова от частотности ошибок в его написании рассматривается в работе С. Рахманьян и В. Купермана [Rahmanian, Kuperman 2019]. Исследователи выдвинули и экспериментально подтвердили на материале английского языка гипотезу о том, что не только написание, но и обработка слова оказываются затрудненными у слов с частотными орфографическими ошибками, даже в случае если они написаны верно. Это происходит под влиянием фактора энтропии — меры неопределенности, отражающей количество усилий, затрачиваемых на выбор между вариантами написания. Отдельные результаты исследования оказались противоречивыми: фактор энтропии не обнаружил связи с уровнем начитанности и грамотности респондентов; что касается взаимодействия с частотностью слова, в одном эксперименте эффект энтропии оказался сильнее для более частотных слов, в другом — для менее частотных.

Для русского языка эксперимент был продолжен и дополнен Д. А. Черновой, С. В. Алексеевой и Н. А. Слюсарь [Чернова и др. 2020а]. Одна из его частей была направлена на оценку успешности распознавания правильных и неправильных написаний слов с частотными ошибками. Результаты подтвердили гипотезу о влиянии факторов общей частотности слова в корпусе и частотности его ошибочных написаний: чем чаще слово встречается в ошибочных вариантах, тем сложнее понять, что оно написано верно; чем частотнее слово, тем легче распознать верное написание.

Таким образом, исследования [Rahmanian, Kuperman 2019; Чернова и др. 2020а] продемонстрировали, что слова, в которых чаще допускаются орфографические ошибки, вызывают повышенные трудности при восприятии. В отличие от работы С. Рахманьян и В. Купермана, в эксперименте российских исследователей испытуемым предъявлялись не только правильные, но и ошибочные написания. В обоих случаях гипотеза была подтверждена для носителей языка (английского или русского).

Данное исследование повторяет дизайн эксперимента Д. А. Черновой, С. В. Алексеевой и Н. А. Слюсарь и осуществляется в отношении второго языка, а именно английского, изучаемого русскоязычными студентами. Успешность распознавания орфографических ошибок рассматривается с точки зрения влияния следующих факторов: частотность ошибок, частотность слова в учебном корпусе, энтропия и тип ошибки.

Орфографические ошибки в рамках усвоения второго языка: привлечение корпусных данных

Орфографические ошибки во втором языке изучаются не только в педагогических целях, но и в аспекте усовершенствования алгоритмов их автоматического распознавания и исправления, учитывающих особенности родного языка изучающих [Leacock et al. 2015]. Для решения ключевых задач — классификации орфографических ошибок изучающих язык и сравнения с данными носителей — активно привлекаются учебные корпуса, содержащие более или менее подробную разметку ошибок правописания. Что касается английского языка как иностранного (второго), соответствующие корпуса и исследования представлены для изучающих с японским [Okada 2005], малайзийским [Botley et al. 2007] и другими родными языками.

Значимыми для данной области являются исследования Вивиан Кук [Cook 1997; 2014]. Для сопоставления особенностей правописания в английском как родном и втором языке исследовательница использовала следующую классификацию орфографических ошибок: добавление буквы (*untill* вместо *until*), пропуск буквы (*occuring* вместо *occurring*), замена буквы (*definate* вместо *definite*), транспозиция (взаимная замена) двух соседних букв (*freind* вместо *friend*), замена графемы, то есть несколько взаимосвязанных изменений в буквенном составе (*thort* вместо *thought*), другие ошибки (*fought* вместо *thought*) [Cook 1997]. В последующих работах эта классификация была дополнена: в частности, в [Bestgen, Granger 2011] были добавлены классы, связанные с употреблением апострофа (*womans* вместо *woman's*) и разграничением слов с помощью пробела (*airpollution* вместо *air pollution*). Неизменными остаются четыре класса, отражающие формальные различия в буквенном составе между верными и ошибочными написаниями: добавление, пропуск, замена буквы, транспозиция [Okada 2005]. Данные классы универсальны, так как используются в исследованиях, решающих широкий круг задач на материале учебных корпусов с разными родными языками. По этой причине представленная классификация была выбрана для настоящего исследования. Необходимо отметить, что в нем не рассматривается влияние родного языка как отдельный класс или фактор, хотя существуют классификации, выделяющие допущенные под его влиянием ошибки в самостоятельную разновидность [Botley et al. 2007]. В качестве основания для классификации некоторыми исследователями также рассматривается отсутствие или существование ошибочного варианта в качестве слова анализируемого языка, в связи с чем выделяются non-word errors (*businees* вместо *business*) и real-word errors (*they* вместо *then*) [Flor, Futagi 2012]. В учебном корпусе REALEC, использованном как источник материала для эксперимента в настоящем исследовании, ошибочные варианты, представляющие собой существующие слова, маркируются с помощью комбинации орфографического (*Spelling*) и лексического (*Choice of lexical item*) тегов. Тем не менее данная работа концентрируется исключительно на второй группе — словах, чьи ошибочные варианты написания не представляют собой существующие слова (в противном случае невозможно было бы использовать их в качестве стимулов вне контекста).

Что касается сопоставления орфографических ошибок в английском языке как в родном и иностранном, оно производится исследователями преимущественно

в аспекте порождения, а не восприятия речи. Ученые отмечают, что изучающие язык и носители существенно не отличаются в отношении популярности того или иного типа ошибок [Cook 1997; Okada 2005]. Такеши Окада, сравнивавший ошибки четырех базовых типов в английском носителей и изучающих-японцев, пришел к выводу, что в обоих случаях совпадает не только порядок типов по мере убывания популярности (замена, пропуск, добавление буквы, транспозиция), но и их количественное соотношение [Okada 2005]. Согласно исследованию [Flor et al. 2015], в английском языке и носители, и изучающие чаще совершают орфографические ошибки в нечастотных словах. Таким образом, в отношении их природы скорее подтверждается гипотеза, основанная на знании (knowledge-based hypothesis), чем гипотеза, основанная на возможности (opportunity hypothesis): возникновение ошибок не является случайным, а объясняется недостаточным знакомством со словом ввиду его низкой частотности.

В данном исследовании результаты упомянутых научных работ, касающиеся распределения типов ошибок и влияния фактора частотности, будут сопоставлены с итогами эксперимента в отношении обработки слов при восприятии.

Материал и методы

Настоящее исследование проводилось на материале учебного корпуса REALEC (Russian Error-Annotated Learner English Corpus, http://realec.org/index.xhtml#/exam/), содержащего эссе двух жанров — описание графического материала и аргументативное эссе, — написанные русскоязычными студентами в рамках экзамена по английскому языку на втором году обучения в Национальном исследовательском университете «Высшая школа экономики». На сегодняшний день REALEC включает свыше 6000 экзаменационных текстов, общее количество слов в которых достигает 1,5 миллионов. Корпус снабжен иерархической системой, включающей 98 тегов для различных типов ошибок, в том числе пунктуационных, орфографических, грамматических, лексических и дискурсных. Эти теги используются аннотаторами-экспертами, также носителями русского языка, при ручной разметке текстов; кроме того, тексты в корпусе подвергаются автоматической частеречной разметке (POS tagging).

Корпус REALEC позволяет производить отбор материала по конкретному тегу, и эта возможность уже использовалась при проведении исследований. В качестве примера можно привести работу М. А. Климовой и коллег [Klimova et al. 2021], в которой рассматриваются так называемые confusables — трудноразличимые для учащихся слова, — извлеченные из корпуса по двум лексическим тегам (*Choice of a lexical item* и *Words often confused*). Что касается настоящего исследования, материалом для него послужили слова, отобранные по тегу *Spelling*, специально предназначенному для орфографических ошибок.

На начальном этапе исследования мы проанализировали список наиболее частотных слов, помеченных тегом *Spelling* и извлеченных из корпуса с помощью программы на языке Python. Из данного списка мы исключили очевидные опечатки, омофоны и различные формы одного и того же слова, в том числе варианты написания, характерные для британского или американского английского языка. При отборе материала для эксперимента основным критерием была частотность оши-

бочного написания; кроме того, мы стремились включить в опрос различные типы ошибок, и по этой причине для некоторых слов был отобран второй или третий вариант в списке наиболее частотных.

В результате вслед за [Чернова и др. 2020а] мы составили список из 44 слов, демонстрирующих четыре типа ошибок:

— добавление буквы (*atheletes* вместо *athletes*, *schould* вместо *should*, *ammount вместо amount*, *usefull* вместо *useful*);
— пропуск буквы (*goverment* вместо *government*, *commited* вместо *committed*, *noticable* вместо *noticeable*, *exept* вместо *except*);
— замена буквы (*persentage* вместо *percentage*, *convinient* вместо *convenient*, *definetely* вместо *definitely*, *tendancy* вместо *tendency*);
— транспозиция (*beleive* вместо *believe, buisness* вместо *business*).

Далее мы составили два протокола, в одном из которых половина слов (22 слова) была написана верно, а другая половина (также 22 слова) содержала орфографические ошибки (то, какие слова будут представлены в верном написании, а какие — в ошибочном, решалось по принципу случайного отбора). Во втором протоколе был представлен тот же самый список слов, но верные и ошибочные написания были распределены противоположным образом: те слова, которые в первом протоколе были написаны верно, во втором были даны с ошибками, и наоборот.

Участниками эксперимента стали 63 студента второго курса, изучающих лингвистику в Национальном исследовательском университете «Высшая школа экономики» и владеющих английским языком на среднем и продвинутом уровне. Их задачей было определить, верно или неверно написаны слова, предъявленные им в случайном порядке в виде анонимных гугл-форм. 28 человек получили первый протокол, а 35 — второй. При выполнении задания участники не были ограничены во времени, но для чистоты эксперимента мы просили их не пользоваться словарями и интернет-ресурсами.

Далее мы попытались определить, в какой мере на успешность распознавания испытуемыми верных и неверных написаний предъявленных им слов повлияли такие факторы, как частотность ошибок, частотность слова в учебном корпусе, энтропия и тип ошибки. В качестве метода анализа значимости факторов мы использовали логистическую регрессию со смешанными эффектами по испытуемому и по стимулу [R Core Team 2013; Bates et al. 2015]. В случае небинарных факторов мы также использовали метод множественных сравнений с поправкой Холма [Hothorn et al. 2015].

Распределение неправильных написаний и типов ошибок в REALEC

Чтобы ответить на вопросы о том, в каких словах в корпусе REALEC чаще всего встречаются орфографические ошибки и какие типы ошибок наиболее широко представлены в корпусе, был проведен анализ слов из корпуса REALEC, отмеченных тегом *Spelling*. Мы составили список уникальных слов, помеченных аннотаторами как исправление орфографической ошибки, после чего для каждого из этих слов было подсчитано количество случаев, когда его потребовалось использовать в качестве исправления (т. е. число вхождений ошибочных написаний в корпус).

Из числа ошибочных вхождений были исключены варианты, очевидно являющиеся опечатками или помеченные тегом *Typo* в корпусе.

Наибольшее количество случаев ошибочного написания в корпусе было выявлено для слова percentage, встретившегося в ошибочном написании 278 раз (в правильном варианте оно встретилось в корпусе 2596 раз). Такое количество вхождений в корпус объясняется тем, что в REALEC собраны тексты, написанные в рамках экзамена по английскому языку, одним из заданий в котором является описание графика, зачастую требующее упоминания процентных соотношений. При этом слово *percentage* содержит одновременно звук /s/, который в английском языке может передаваться как с помощью буквы *c*, так и с помощью буквы *s*, и звук /ɛ/, который носителям русского языка может быть затруднительно отличить от других встречающихся в английском гласных переднего ряда, которые могут передаваться на письме буквой *a*, в результате чего самыми популярными ошибочными написаниями становятся *persentage*, *percantage* и *persantage*. Тем не менее по количеству уникальных неправильных написаний слово *percentage*, у которого таких вариантов 36, уступает *countries* (42) и *government* (41), занявшим по количеству ошибочных вхождений второе и третье место соответственно. Частота ошибочных вхождений слова *countries*, возможно, отчасти связана с его высокой частотностью в корпусе вообще — при значении относительной частотности 4452,5 ipm *countries* занимает второе место после *should* (4459,8 ipm) среди слов, отобранных для эксперимента. Тем не менее затруднения, возникающие у студентов, также могут объясняться наличием в слове гласного звука /ʌ/, передающегося двумя буквами (что ведет к наиболее частотному ошибочному варианту *contries*), а также необходимостью трансформации графической формы окончания при образовании множественного числа. Частотность ошибочных вариантов слова *government*, бесспорно, также связана с наличием в нем кластера согласных *rn, в* котором звук /r/ не произносится. Тем не менее вариант *goverment* (74 вхождения) оказался гораздо частотнее варианта *govenment* (5 вхождений), что, возможно, связано с особенностями произношения этого слова русскоязычными студентами. Буква *r* зачастую пропускалась и в слове *furthermore* (ошибочный вариант — *futhermore*). Ошибки пропуска буквы наблюдались также в словах, содержащих удвоенные согласные (*commited вместо committed*, *oportunity* вместо *opportunity*) или букву *e* при отсутствии гласного звука (*diffrent* вместо *different, temprature* вместо *temperature*). Стоит отметить, что формы слов с удвоенными согласными и непроизносимыми *e* также часто встречались среди написаний, созданных в результате ошибки вставки, — среди этих ошибок *ammount* вместо *amount*, *employement* вместо *employment*, *comparisson* вместо *comparison*.

Хотя наиболее частотным по количеству неверных вхождений в корпусе было слово с ошибкой замены буквы (*persentage* вместо *percentage*), самым частотным типом ошибки среди 44 неверных написаний, отобранных для эксперимента, оказался пропуск буквы — ошибок такого типа в списке 17 (*nowdays* вместо *nowadays*, *futhermore* вместо *furthermore* и т. д.). Ошибок замены буквы оказалось 14 (*convinient* вместо *convenient*, *trand* вместо *trend* и т. д.), добавления буквы — 11 (*drammatically* вместо *dramatically*, *unemployement* вместо *unemployment* и т. д.), а ошибок транспозиции в список вошло всего 2: *buisness* вместо *business* и *beleive* вместо *believe*. В рамках списка отобранных написаний это распределение не соответствует распределе-

нию [Okada 2005], в котором ошибки замены были наиболее распространенными, однако распределение ошибок среди 44 наиболее часто встречающихся в неверном написании слов может не соответствовать распределению ошибок во всем корпусе.

Статистический анализ результатов эксперимента

Участники эксперимента в среднем правильно ответили на 37,44 из 44 вопросов, что означает, что в среднем примерно на 85 % вопросов был дан верный ответ. Для случаев, в которых слово было дано в верном написании, среднее количество неправильных ответов составляло 5,67 %, тогда как для слов, данных в неверном написании, этот показатель возрастал до 23,12 %. Анализ показал, что неверным написанием, наиболее часто ошибочно определенным как верное, было *commited* вместо *committed*: как верное его опознали 71,43 % из 35 опрошенных, получивших его в ошибочном варианте. Среди верных написаний, наиболее часто ложно определенных как неверные, на первом месте оказалось *preferred* (35,71 % из 28 опрошенных, получивших его в верном варианте), причем ошибочное написание *prefered* также вызвало затруднения: как верное его распознали 57,17 % из 35 опрошенных, получивших его в ошибочном варианте. *Becouse* (вместо *because*) оказался единственным ошибочным вариантом, ни разу не идентифицированным как верный, тогда как среди правильных написаний неверных ответов не имели 19. Как и во всем списке, среди этих слов чаще всего встречались ошибки пропуска и замены (8 и 7 слов соответственно), тогда как ошибки вставки и транспозиции (3 и 1 соответственно) встречались реже. Также стоит отметить, что в числе слов, верное написание которых правильно распознали все студенты, оказались три слова с наибольшей частотностью ошибочных вхождений: *percentage, countries* и *government*.

Хотя в словах *preferred* и *committed*, вызвавших наибольшие затруднения у студентов в верном и неверном написании соответственно, содержится ошибка на пропуск буквы, в среднем наибольшие затруднения у студентов вызвали слова с ошибкой замены: ошибки в этих словах в среднем неверно идентифицировали 27,96 %, тогда как для ошибок пропуска этот показатель составил 20,34 %, а для ошибок вставки — 24,29 %. Этот факт позволяет предположить, что тип ошибки может влиять на вероятность правильного или неправильного ответа. Также стоит отметить, что высокая доля ошибок в словах с заменой буквы соответствует данным работы [Okada 2005], в которой было обнаружено, что этот тип ошибки наиболее популярен как в английском носителей, так и в английском людей с родным японским языком. Следовательно, можно предположить, что слова с потенциальными ошибками замены вызывают затруднения как при производстве англоязычной речи, так при и ее восприятии. Тем не менее, хотя согласно данным [Okada 2005] ошибки пропуска более распространены, чем ошибки добавления, в нашем исследовании они чаще были идентифицированы корректно. Возможно, это связано с большей средней частотностью отобранных слов с ошибкой пропуска: для этих слов средняя частотность составила 1019,2 ipm, тогда как для слов с ошибкой добавления — 975,8 ipm.

Для дальнейшего анализа результатов и определения влияния различных факторов на правильность ответа была использована логистическая регрессия со смешанными эффектами по испытуемому и по стимулу; для построения использова-

лась библиотека lme4 [Bates et al. 2015]. Случайными факторами являлись участники и слова из списка, тогда как фиксированными факторами были относительная частотность слова в целом и его неправильных вхождений в корпусе REALEC (соотношение числа вхождений и общего числа слов в корпусе), тип ошибки, количество неверных вариантов слова в REALEC и энтропия. Как и в работе [Rahmanian, Kuperman 2019], энтропия рассчитывалась на основе относительной частотности каждого варианта написания слова и использовалась как мера неопределенности, вызываемой наличием у слова большого количества вариантов написания с относительно высокой частотностью. При расчетах использовалась формула:

$$H = -\sum_{i=1}^{n} p_i \log(p_i),$$

где H — энтропия, n — количество вариантов слова, i — один вариант слова, p — относительная частотность (вероятность) i в n.

На практике это означало, что чем чаще слово встречалось в корпусе в правильном написании, тем ниже у него была энтропия: к примеру, 98 % вхождений слова *health* в корпус имели правильное написание, и вариантов написания было всего 7, в результате чего его энтропия была низкой (0,12), тогда как слово *definitely* встречалось в правильном написании лишь в 72 % случаев и имело 20 вариантов написания, что приводило к высокой энтропии (1,2). Также высокая энтропия оказалась у слов *illegal* (1,07), *necessary* (0,95), *tendency* (0,93), *diseases* (0,90), а низкая — у слов *because* (0,17), *amount* (0,17), *different* (0,16) и *should* (0,11). Как правило, при частотности больше 500 ipm энтропия слова была низкой (меньше 0,5), а при частотности меньше 500 ipm — высокой, однако из этого правила встречались некоторые исключения: к примеру, хотя у слова *percentage* частотность была более 2000 ipm, его энтропия составила 0,52. Среднее значение энтропии для слов, отобранных для эксперимента, оказалось равно 0,50, медианное значение — 0,45.

Логистические регрессии показали, что энтропия является значимым фактором, причем чем ниже у слова энтропия, тем вероятнее, что участник верно ответит на вопрос (b = –2,3173, SE = 0,5320, p < 0,001, где b — коэффициент регрессии *B*, SE — стандартная ошибка для коэффициента регрессии *B,* p — P-значение). Значимость энтропии в нашем эксперименте соответствует выводам С. Рахманьян и В. Купермана об отрицательном влиянии фактора энтропии на обработку слова, что позволяет сделать вывод о том, что этому эффекту подвержены как носители английского языка, так и русскоязычные люди, изучающие английский язык. Также значимым оказался фактор относительной частотности в корпусе (b = 3,8582, SE = 1,6665, p < 0,05): чем чаще слово встречается в REALEC, тем больше вероятность, что участник верно распознает наличие или отсутствие в нем ошибки. Влияние частотности ошибочного написания в корпусе, однако, было обнаружено лишь на уровне тенденции (b = –52,4340, SE = 30,5319, p < 0,1). Более низкая значимость этого показателя по сравнению с работой [Чернова и др. 2020a], где p = 0,02, может быть связана с разницей языковой среды. Если в исследовании носителей языка участники чаще сталкивались с неправильными вариантами представленных слов в повседневной жизни, то изучающие иностранный язык, как правило, чаще взаимодействуют с языковым материалом, в котором лексика представлена в верном написании: как в учебниках, так и в литературе на иностранном языке.

В модели, построенной только с учетом фактора типа ошибки, а также при попарном сравнении типов ошибок значимых различий обнаружено не было. Тем не менее также была построена логистическая модель, учитывающая потенциальное взаимодействие типа ошибки и энтропии. В целом оказалось, что эти два фактора уменьшают отрицательное влияние друг друга на правильность ответа. В этой модели оказались значимы как тип ошибки «замена буквы» сам по себе ($b = -1{,}5166$, $SE = 0{,}7541$, $p < 0{,}05$), так и его взаимодействие с энтропией ($b = 2{,}9609$, $SE = 1{,}3132$, $p < 0{,}05$). В сочетании с данными о частотности затруднений в определении правильного или неправильного написания слов с этим типом ошибки это в очередной раз указывает на то, что данный тип ошибки вызывает особенные затруднения при восприятии письменной иностранной речи.

Заключение

Настоящее исследование было посвящено влиянию частотности орфографических ошибок в слове на качество его репрезентации в ментальном лексиконе. С помощью эксперимента на материале учебного корпуса английского языка REALEC (Russian Error-Annotated Learner English Corpus) была протестирована следующая гипотеза, ранее подтвержденная для родного английского и русского языков: слова, в которых часто совершаются ошибки правописания, вызывают трудности при восприятии, даже если написаны правильно. В качестве материала для исследования были отобраны слова с наиболее частотными ошибочными написаниями; кроме того, мы стремились представить четыре типа ошибок, традиционно рассматриваемые при исследовании учебных корпусов: добавление буквы, пропуск буквы, замена буквы и транспозиция.

Результаты проведенного нами эксперимента подтвердили для второго языка вывод, сделанный Д. А. Черновой, С. В. Алексеевой и Н. А. Слюсарь [Чернова и др. 2020a] о значимости фактора частотности слова в корпусе: чем чаще слово встречается в корпусе REALEC, тем больше вероятность, что испытуемый верно распознает наличие или отсутствие в нем ошибки. Однако влияние частотности ошибочного написания в корпусе оказалось менее значимым фактором по сравнению с результатом, полученным в [Чернова и др. 2020a]. Подобное отличие учебного корпуса от корпуса носителей могло бы стать темой отдельного исследования.

Результаты эксперимента также позволяют предположить, что отдельные типы ошибок могут влиять на вероятность правильного или неправильного их опознания. Как выяснилось, в целом наибольшие затруднения у студентов вызвали слова с ошибкой замены буквы. Интересно, что, согласно выводам Такеши Окада, данный тип ошибок оказывается наиболее распространенным и при производстве англоязычной речи как среди носителей, так и среди изучающих [Okada 2005]. Что касается восприятия, в нашем эксперименте он продемонстрировал свою значимость как сам по себе, так и при взаимодействии с энтропией.

Как показало наше исследование, чем ниже у слова энтропия, тем вероятнее, что испытуемый верно ответит на вопрос. Это подтверждает вывод, сделанный в работе С. Рахманьян и В. Купермана о влиянии фактора энтропии на обработку слова [Rahmanian, Kuperman 2019]: очевидно, этому эффекту подвержены как но-

сители английского языка, так и русскоговорящие студенты, изучающие английский язык.

Что касается перспективы дальнейших исследований, представляется целесообразным продолжить изучение фактора частотности орфографических ошибок и его роли при овладении вторым языком. Кроме того, интересно было бы исследовать влияние различных факторов, не учитывавшихся в проведенном нами эксперименте, — например уровня владения иностранным языком. Так, на материале второго языка можно было бы протестировать вывод об отсутствии корреляции энтропии с уровнем начитанности и грамотности респондентов, сделанный С. Рахманьян и В. Куперманом в отношении носителей языка. Еще одной интересной перспективой может стать исследование восприятия слов в контексте; наконец, корпус REALEC предоставляет возможность исследовать влияние на различные типы ошибок, в том числе и орфографические, такого сложного и многогранного явления, как интерференция родного языка.

Литература

Алексеева, Слюсарь 2017 — Алексеева С. В., Слюсарь Н. А. Орфографические соседи в русском языке: База данных и эксперимент, направленный на изучение морфологической декомпозиции. *Вопросы психолингвистики*. 2017, 32 (2): 12–27.

Чернова и др. 2020а — Чернова Д. А., Алексеева С. В., Слюсарь Н. А. Чему нас учат ошибки: трудности при обработке слов с частотными орфографическими ошибками. *Компьютерная лингвистика и интеллектуальные технологии*. 2020, (19): 147–159.

Чернова и др. 2020б — Чернова Д. А., Слюсарь Н. А., Алексеева С. В. Особенности орфографической обработки падежных форм русских существительных в контексте предложения. *Вестник Томского государственного университета*. 2020, (454): 45–54.

Чернова 2022 — Чернова Д. А. Фонологическая и графическая репрезентации слова в ментальном лексиконе: восприятие омофонов при чтении. *Вестник Санкт-Петербургского университета. Язык и литература*. 2022, 19 (1): 181–194.

Andrews et al. 2020 — Andrews S., Veldre A., Clarke I. E. Measuring lexical quality: The role of spelling ability. *Behavior Research Methods*. 2020, 52 (6): 2257–2282.

Bates et al. 2015 — Bates D., Maechler M., Bolker B., Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015, 67 (1): 1–48.

Bestgen, Granger 2011 — Bestgen Y., Granger S. Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*. 2011, 21 (2–3): 235–252.

Botley et al. 2007 — Botley S., Hakim F., Dillah D. Investigating Spelling Errors in a Malaysian Learner Corpus. *Malaysian Journal of ELT Research*. 2007, (3): 74–93.

Cook 1997 — Cook V. J. L2 Users and English Spelling. *Journal of Multilingual and Multicultural Development*. 1997, 18 (6): 474–488.

Cook 2014 — Cook V. J. *The English writing system*. London; New York: Routledge, 2014.

Flor, Futagi 2012 — Flor M., Futagi Y. On using context for automatic correction of non-word misspellings in student essays. In: *Proceedings of the seventh workshop on building educational applications using NLP*. 2012. P. 105–115.

Flor et al. 2015 — Flor M., Futagi Y., Lopez M., Mulholland M. Patterns of misspellings in L2 and L1 English: A view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies*. 2015, (6): 107–132.

Hothorn et al. 2015 — Hothorn T., Bretz F., Ag P., Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal*. 2015, 50 (3): 346–363.

Klimova et al. 2021 — Klimova M. A., Smilga V. K., Overnikova D. A. Using an Error-Annotated Learner Corpus (REALEC) in DDL Lessons. В сб.: *Труды международной конференции «Корпусная лингвистика — 2021»*. Захаров В. П. (ред.). СПб.: Изд-во С.-Петерб. ун-та, 2021. С. 112–121.

Leacock et al. 2015 — Leacock C., Chodorow M., Tetreault J. Automatic grammar and spell-checking for language learners. In: *The Cambridge Handbook of Learner Corpus Research*. Granger S., Gilquin G., Meunier F. (eds). Cambridge: Cambridge University Press, 2015. P. 267–286.

Okada 2005 — Okada T. A Corpus-based Study of Spelling Errors of Japanese EFL Writers with Reference to Errors Occurring in Word-initial and Word-final Positions. In: *Second Language Writing Systems*. Cook V., Bassetti B. (Eds). Clevedon; Buffalo; Toronto: Multilingual Matters, 2005. P. 164–183.

Perfetti 1985 — Perfetti C. A. *Reading ability*. Oxford: Oxford University Press, 1985.

Perfetti 2007 — Perfetti C. A. Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*. 2007, 11 (4): 357–383.

Perfetti, Hart 2001 — Perfetti C. A., Hart L. The lexical basis of comprehension skill. In: *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*. Gorfein D. S. (ed.). Washington, DC: American Psychological Association, 2001. P. 67–86.

Rahmanian, Kuperman 2019 — Rahmanian S., Kuperman V. Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*. 2019, 23 (1): 24–36.

R Core Team 2013 — R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, 2013. URL: http://www.R-project.org/ (дата обращения: 22.07.2022).

Статья поступила в редакцию 30 июля 2022 г.
Рекомендована в печать 16 июня 2023 г.

Margarita A. Klimova

HSE University,
25/12, ul. Bolshaya Pecherskaya, Nizhniy Novgorod, 603155, Russia
mfokina@hse.ru

Anna V. Viklova

Russian Presidential Academy of National Economy and Public Administration,
82/1, pr. Vernadskogo, Moscow, 119571, Russia
viklova-av@ranepa.ru

Daria A. Overnikova

HSE University,
20, ul. Myasnitskaya, Moscow, 101000, Russia
daria.a.overnikova@gmail.com

Processing frequently misspelled words (A study based on an English learner corpus)*

For citation: Klimova M. A., Viklova A. V., Overnikova D. A. Processing frequently misspelled words (A study based on an English learner corpus). *Vestnik of Saint Petersburg University. Language and Literature.* 2023, 20 (4): 824–837. https://doi.org/10.21638/spbu09.2023.409 (In Russian)

The article presents an experimental study of the influence of the frequency of spelling errors in a word on its representation in mental lexicon. The hypothesis that frequently misspelled words cause difficulties in reading even if they are written correctly has been proved for native speakers of Russian and English. This paper aims to check the hypothesis on the basis of the learner corpus REALEC (Russian Error-Annotated Learner English Corpus), comprising texts by Russian L1 learners of English. The most frequently misspelled words collected from the corpus were used for the experiment that consisted in recognising correct and incorrect

* The study was carried out within the framework of the research project of the HSE University "Development of models of learning a second language in the framework of existing theories using methods for automating experiments based on the materials of the REALEC corpus and other educational buildings".

spellings. We analysed the influence of the following factors: the frequency of spelling errors in a word, its frequency in the corpus, entropy (the measure reflecting the amount of effort needed to choose between the variants of spelling), type of error. The results demonstrate the significance of entropy and frequency in the corpus, which corresponds to the findings of previous studies. A particular error type, substitution, has been found to be significant. Its special role corresponds both to the greatest difficulties this error type caused during the experiment and the results of previous research into written speech production of L2 English speakers, according to which substitution was considered the most frequent error type. The lower significance of the frequency of errors factor in comparison with the corresponding studies of L1 English can be explained by differences in language environments, in which learners of English are less exposed to incorrect spellings.

Keywords: word processing, spelling errors, mental lexicon, learner corpus.

References

Алексеева, Слюсарь 2017 — Alexeeva S. V., Slioussar N. A. Orthographic neighbours: A database on Russian language and experimental studies of morphological decomposition. *Voprosy psikholingvistiki*. 2017, 32 (2): 12–27. (In Russian)

Чернова и др. 2020а — Chernova D. A., Alexeeva S. V., Slioussar N. A. What do we learn from mistakes: Processing difficulties with frequently misspelled words. *Komp'iuternaia lingvistika i intellektual'nye tekhnologii*. 2020, (19): 147–159. (In Russian)

Чернова и др. 2020б — Chernova D. A., Slioussar N. A., Alexeeva S. V. Orthographic processing of Russian case forms in sentential context. *Vestnik Tomskogo gosudarstvennogo universiteta*. 2020, (454): 45–54. (In Russian)

Чернова 2022 — Chernova D. A. Phonological and graphic representations of words in mental lexicon: Homophone processing while reading. *Vestnik of Saint Petersburg University. Language and Literature*. 2022, 19 (1): 181–194. (In Russian)

Andrews et al. 2020 — Andrews S., Veldre A., Clarke I. E. Measuring lexical quality: The role of spelling ability. *Behavior Research Methods*. 2020, 52 (6): 2257–2282.

Bates et al. 2015 — Bates D., Maechler M., Bolker B., Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015, 67 (1): 1–48.

Bestgen, Granger 2011 — Bestgen Y., Granger S. Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*. 2011, 21 (2–3): 235–252.

Botley et al. 2007 — Botley S., Hakim F., Dillah D. Investigating Spelling Errors in a Malaysian Learner Corpus. *Malaysian Journal of ELT Research*. 2007, (3): 74–93.

Cook 1997 — Cook V. J. L2 Users and English Spelling. *Journal of Multilingual and Multicultural Development*. 1997, 18 (6): 474–488.

Cook 2014 — Cook V. J. *The English writing system*. London; New York: Routledge, 2014.

Flor, Futagi 2012 — Flor M., Futagi Y. On using context for automatic correction of non-word misspellings in student essays. In: *Proceedings of the seventh workshop on building educational applications using NLP.* 2012. P. 105–115.

Flor et al. 2015 — Flor M., Futagi Y., Lopez M., Mulholland M. Patterns of misspellings in L2 and L1 English: A view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies*. 2015, (6): 107–132.

Hothorn et al. 2015 — Hothorn T., Bretz F., Ag P., Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal*. 2015, 50 (3): 346–363.

Klimova et al. 2021 — Klimova M. A., Smilga V. K., Overnikova D. A. Using an Error-Annotated Learner Corpus (REALEC) in DDL Lessons. In: *Trudy mezhdunarodnoi konferentsii "Korpusnaia lingvistika — 2021"*. Zakharov V. P. (ed.). St. Petersburg: St. Petersburg University Press, 2021. P. 112–121.

Leacock et al. 2015 — Leacock C., Chodorow M., Tetreault J. Automatic grammar and spell-checking for language learners. In: *The Cambridge Handbook of Learner Corpus Research*. Granger S., Gilquin G., Meunier F. (eds). Cambridge: Cambridge University Press, 2015. P. 267–286.

Okada 2005 — Okada T. A Corpus-based Study of Spelling Errors of Japanese EFL Writers with Reference to Errors Occurring in Word-initial and Word-final Positions. In: *Second Language Writing Systems*. Cook V., Bassetti B. (eds). Clevedon; Buffalo; Toronto: Multilingual Matters, 2005. P. 164–183.

Perfetti 1985 — Perfetti C. A. *Reading ability*. Oxford: Oxford University Press, 1985.

Perfetti 2007 — Perfetti C. A. Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*. 2007, 11 (4): 357–383.

Perfetti, Hart 2001 — Perfetti C. A., Hart L. The lexical basis of comprehension skill. In: *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*. Gorfein D. S. (ed.). Washington: American Psychological Association, 2001. P. 67–86.

Rahmanian, Kuperman 2019 — Rahmanian S., Kuperman V. Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*. 2019, 23 (1): 24–36.

R Core Team 2013 — R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, 2013. Available at: http://www.R-project.org/ (accessed: 22.07.2022).

Received: July 30, 2022
Accepted: June 16, 2023