Влияние орфографической энтропии на скорость доступа к семантике слов русского языка

Анна В. Сергеева

Российский государственный гуманитарный университет,

Москва, Россия, nyura.shahee@mail.ru

Влияние частотности орфографических ошибок (и, как следствие, множества различных написаний одного и того же слова) на обработку слов было впервые отмечено в рамках исследования [Rahmanian, Kuperman, 2019]. Была предложена гипотеза, предполагающая, что наличие или отсутствие “орфографической стабильности” (т.е. полного доминирования одного орфографического варианта, как правило, орфографически верного, над другими вариантами написания слова) влияет на распознавание и освоение этого слова. Чем чаще некорректная форма встречается в том же значении, что и корректная, тем сложнее становится носителю установить связь между корректной формой и этим значением. Гипотеза о том, что наличие частотных орфографических ошибок затрудняет обработку слова (наравне с факторами частотности и длины слова), была подтверждена в этом исследовании на материале английского языка, а также позже доказана в исследованиях на материале греческого, финского, китайского языков и иврита [Kuperman et al., 2021]. В исследовании [Чернова и др., 2020] эта гипотеза была подтверждена на материале русского языка.

Lexical Quality Hypothesis, гипотеза лексического качества, основывается на том, что качество представления слова в ментальном лексиконе влияет на восприятие, обработку и понимание речи. В свою очередь, представление слова определяется его фонетической формой, орфографической формой и семантикой. Слово распознается на основе этих трех составляющих. По [Perfetti, Hart, 2008] семантическая составляющая включает в себя и грамматическую информацию. Лексическое качество зависит от опыта носителя: более опытный читатель с большей вероятностью способен понять незнакомое слово, так как имеет много высококачественных репрезентаций.

Лексическое качество может снижаться при наличии неоднозначностей в соотнесении между каждой из форм (фонетической, письменной) и смыслом. Если

множество альтернативных написаний порождает связи между каждым из вариантов и одним значением, «размывая» лексическую репрезентацию и усложняя выбор, то можно сделать следующее предположение: время распознавания этого значения может быть больше, чем время, затрачиваемое на распознавание значения слова с одним доминирующим вариантом написания.

В настоящей работе было проведено экспериментальное исследование на основе платформы PC.IBEX, в котором респондентам были предложены слова с разным уровнем орфографической энтропии. Респонденту было необходимо выбрать наиболее подходящий для данного слова синоним из двух вариантов. Для опроса носителей русского языка была составлена анкета из слов-стимулов и их значений. Из материалов исследования [Чернова и др., 2020], где подсчитывались индексы энтропии для определенного набора слов и были указаны частотность и длина этих слов, были взяты 44 стимула для трех разных анкет. Для каждого из них был подобран синоним, а также слово, появляющееся в паре с этим синонимом (были выбраны слова, схожие по звучанию, но со значительно отличающимся лексическим значением). Таким образом, респонденту предъявлялся стимул, для которого из двух вариантов нужно было выбрать слово, соответствующее его семантике. В эксперименте участвовали 46 респондентов в возрасте от 11 до 70 лет. В случайных экспериментальных группах оказалось 18, 12 и 16 человек. После завершения опроса данные анализировались с помощью языка R с использованием метода смешанных линейных моделей.

В результате была построена смешанная линейная модель, учитывающая время ответа на вопрос, орфографическую энтропию, длину стимула, частотность стимула и случайные эффекты. Значимыми оказались частотность ($p=0,026$), т.е. при увеличении частотности уменьшается время ответа; длина ($p=0,033$), т.е. чем длиннее слово, тем больше время ответа; энтропия – на границе значимости ($p=0,058$), более высокая энтропия также может увеличивать время ответа. Между собой факторы не коррелируют, но их влияние зависит друг от друга (для взаимодействия трех факторов $\beta=452,96$, $SE=179,94$, $p=0,014$).

Таким образом, эксперимент показал, что время ответа действительно связано с эффектом энтропии, но зависимость находится на грани значимости. Если раньше влияние этого фактора было продемонстрировано только для времени прочтения, то в этой работе респондентам предлагалось более трудное задание, предполагающее чтение

трех слов и выбора синонима. Для более надежного подтверждения гипотезы необходима более широкая выборка респондентов и большее количество стимулов, однако даже в рамках данного исследования была обнаружена связь времени ответа с уровнем орфографической энтропии. Это свидетельствует о том, что количество и частотность альтернативных орфографических вариантов, вероятно, способны ослаблять связь внешней формы слова с его значением.

Литература:

Чернова, Алексеева, Слюсарь 2020 – *Чернова Д.А., Алексеева С.В., Слюсарь Н.А.* Чему нас учат ошибки: трудности при обработке слов с частотными орфографическими ошибками. Москва, 2020.

Kuperman 2021 – *Kuperman V.* Spelling errors affect reading behavior across languages // Journal of Experimental Psychology. 2021.

PC IBEX – Платформа PC IBEX [Электронный ресурс]. URL: https://doc.pcibex.net (дата обращения: [29.06.2025]).

Perfetti, Hart 2002 – *Perfetti C., Hart L.* The Lexical Quality Hypothesis // Precursors of Functional Literacy. 2002. С. 189–213.

Rahmanian, Kuperman 2019 – *Rahmanian S., Kuperman V.* Spelling errors impede recognition of correctly spelled word forms // Scientific Studies of Reading. 2019. Т. 23. С. 24–36.