КОГНИТИВНАЯ НАУКА

В МОСКВЕ

НОВЫЕ ИССЛЕДОВАНИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ 2023

Под ред. Е.В. Печенковой, М.В. Фаликман, А.Я. Койфман

УДК 159.9
ББК 88.25
К57

Когнитивная наука в Москве: новые исследования. Материалы конференции 21 – 22 июня 2023 г. Под ред. Е.В. Печенковой, М.В. Фаликман, А.Я. Койфман. – М.: ООО «Буки Веди», Московский институт психоанализа. 2023 г. — 604 стр.

ISBN 978-5-4465-3880-5

УДК 159.9
ББК 88.25

ISBN 978-5-4465-3880-5

МЕТОДИКА ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ: ВЗГЛЯД НА ПОНИМАНИЕ ТЕКСТА НА РОДНОМ И НЕРОДНОМ ЯЗЫКЕ

В. И. Зубов*, А. А. Коновалова
v.zubov@spbu.ru
СПбГУ, Санкт-Петербург

Аннотация. Работа посвящена оценке компетенции понимания текста на родном и неродном языках с помощью вопросно-ответной методики и методики выделения ключевых слов. Участники, носители русского языка, читали два текста на родном и неродном (английском) языке. Оценивалась успешность выделения ключевых слов и правильность ответов на вопросы по содержанию текста. Обнаружена значимая корреляция между истинными наборами ключевых слов для одного и того же текста на разных языках. По результатам эксперимента успешность выделения ключевых слов оказалась связана не с языковой компетенцией читающих, а с содержанием текста, в то время как на правильность ответов на вопросы значимое влияние оказывают языковые компетенции. Сделано предположение о том, что навык выделения ключевых слов в меньшей степени лингвоспецифический, чем навык отвечать на вопросы по содержанию текста, и методика выделения ключевых слов может использоваться в дальнейших исследованиях понимания текста на родном и неродном языке.

Ключевые слова: чтение, понимание, методика выделения ключевых слов, неродной язык, русский язык, английский язык

Исследование поддержано грантом СПбГУ № ID94034584 «Механизмы чтения и интерпретации текста на родном и неродном языках: междисциплинарное экспериментальное исследование с использованием методов регистрации движения глаз, визуальной аналитики и технологий виртуальной реальности».

Введение

Понимание текста — сложный многоуровневый процесс, на который влияет множество факторов, таких как уровень языковой компетенции читателя (беглость чтения, морфологическое осознание и др.), мотивация читателя, когнитивные способности читателя (рабочая память, способность контролировать процесс чтения и другие: Lervåg, Melby-Lervåg, 2022). Традиционной методикой для оценки понимания текста выступает вопросно-ответная методика. Несмотря на распространенность ее использования, она не лишена недостатков, поскольку умение отвечать на вопросы по тексту связано не только с компетенцией понимания текста, но и с другими компетенциями (например, языковой компетенцией, уровнем интеллекта, энциклопедическими знания-

ми и др.). Так, существуют работы, показывающие, что на значительную часть вопросов некоторых стандартизированных тестов можно ответить правильно даже без чтения текста (Cain, 2022).

В отечественной психолингвистике при изучении процессов восприятия и понимания текста используется методика выделения ключевых слов (КС; напр., Петрова и др., 2017). Методика описана в работе Л. Н. Мурзина и А. С. Штерн (1991). Группе участников предлагается выписать ключевые слова из текста. Каждый из участников будет извлекать свой собственный набор ключевых слов. «Какие-то слова будут общими, а какие-то — различными, что обусловлено, с одной стороны, одинаковым пониманием текста, а с другой — индивидуальными различиями в понимании как содержания текста, так и задачи индексирования» (Мурзин, Штерн, 1991, с. 74). Самые частотные КС отражают «общее в восприятии (понимании) текста — „инвариантный смысл“ (Л.А. Черняховская)» (цит. по Мурзин, Штерн 1991, с. 77).

В зарубежных исследованиях методика выделения ключевых слов в первую очередь используется в смежных с лингвистикой областях: интеллектуальном анализе текста, поиске информации, обработке естественного языка и др. Наборы ключевых слов используются в качестве метаданных для обогащения содержания документов, облегчения классификации, систематизации, индексации и резюмирования текстовых данных для более удобного поиска и настроек рекомендаций для читателя (Firoozeh et al., 2020). Однако существуют и работы, в которых методика выделения КС используется для исследований в области понимания текста (напр., de Bruin et al., 2011; Engelen et al., 2018). Так, в работе (Engelen et al., 2018) было показано, что успешность выделения ключевых слов и количество правильных ответов на вопросы по содержанию текстов не были связаны, однако было сделано предположение о том, что ключевые слова все же отражают уровень понимания текста и требуются дальнейшие исследования в этой области.

Поскольку набор КС к тексту является его смысловым ядром, «инвариантом», можно предположить, что для текстов с одинаковым содержанием, но написанном на разных языках, читатели будут выделять схожие или даже одинаковые наборы КС. Настоящее исследование представляет собой попытку применить методику выделения ключевых слов к исследованию компетенции понимания текста на родном и неродном языке, а также сравнить успешность этой методики с вопросно-ответной методикой.

Гипотезы. 1) Истинные наборы ключевых слов для текстов на разных языках будут совпадать, поскольку инвариантный смысл текста нелингвоспецифичен; 2) успешность выделения ключевых слов будет связана с успешностью ответов на вопросы по содержанию текста, поскольку оба задания нацелены на оценку компетенции понимания текста.

Методы

Материал исследования. Материалом исследования стали две пары прозаических текстов научно-популярного стиля: один посвящен значению и происхождению жеста шака («Шака»), второй повествует о древнеримском

боге Янусе («Янус»). Тексты и ряд вопросов к ним были заимствованы из исследования (Kuperman et al., 2022). К четырем разработанным в исследовании (Kuperman et al., 2022) вопросам были добавлены еще четыре, нацеленные на оценку понимания разных уровней глубины. Таким образом, для каждого текста было подготовлено восемь вопросов. Стимульные тексты и результаты исследования доступны по ссылке: https://osf.io/34d5w/.

Некоторые параметры текстов отражены в табл. 1. Читабельность текстов на русском языке была проверена с помощью сервиса https://readability.io, на английском языке — с помощью https://readable.com).

Таблица 1. Параметры текстов, использованных в качестве материала первого этапа исследования

Текст	Количество слов	Количество предложений	Формула Flesch-Kincaid	Формула SMOG
Янус	149	9	8.98	9.73
Janus	182	10	10.40	12.70
Шака	142	7	14.51	13.42
Shaka	183	6	14.00	15.90

Участники исследования. В эксперименте приняли участие 112 носителей русского языка (92 женщины и 20 мужчин в возрасте от 18 до 59 лет, средний возраст — 23 года). Тексты «Янус» и «Shaka» читали 56 человек, тексты «Janus» и «Шака» читали также 56 человек. Все участники владеют английским языком на уровне B2 – C2 (B2 – 66 участников, C1 – 38 участников, C2 – 7 участников; один участник не указал свой уровень владения английским языком). Участники сами определяли уровень владения языком исходя из своих компетенций. Кроме того, участники сообщили о количестве лет, в течение которых они изучали английский язык ($M = 13.14$).

Методика и ход эксперимента. Каждому участнику предлагалось прочитать два текста на разные темы (один на русском языке (L1), другой — на английском (L2)), выписать из текстов по 10 ключевых слов (КС) и ответить на вопросы по содержанию. КС было определено как «наиболее важное с точки зрения содержания текста слово». Для каждого текста было восемь вопросов по содержанию: два вопроса с выбором нескольких вариантов ответа, два вопроса с развернутым ответом и четыре вопроса с выбором ответа «да» или «нет». Во время ответа на вопросы участники не могли возвращаться к тексту. Участники могли записывать ответы на любом языке. Все инструкции и вопросы были написаны на том же языке, на котором был написан текст. Порядок предъявления текстов был случайным.

Обработка КС, которые выписали участники, проводилась следующим образом. Различные формы слова сводились к форме, наиболее часто встреча-

ющейся в ответах (*ворота* (10), *ворот* (7) → *ворота* (17)), словосочетания, не являющиеся устойчивыми, делились на слова, входящие в них (*двоякая природа* → *двоякая* & *природа*), союзы, предлоги, артикли и вспомогательные глаголы удалялись (*to worship* → *worship*), имена собственные и устойчивые сочетания не делились и приводились во всех ответах участников к полному варианту (*Aloha spirit*, *Aloha* → *Aloha spirit*). В соответствии с инструкцией подсчета КС, описанной в (Мурзин, Штерн, 1991), для каждого КС, встретившегося в ответах участников, была рассчитана абсолютная (*m*) и относительная (*p*) частота встречаемости: $p = m/n$, где n — количество участников. Группа слов, получивших наибольшую относительную частоту встречаемости, составила истинный набор КС, в который были включены все КС, которые выписали как минимум 20% участников. Для текста «Янус» на русском и английском языках в истинный набор КС вошли по 18 слов, для текста «Шака» на русском языке — 17 слов, на английском языке — 21 слово. Далее, для каждого участника рассчитывался показатель успешности выделения КС для каждого текста. Если в индивидуальном наборе КС конкретного участника встречалось слово из истинного набора КС, ключевому слову присваивался 1 балл. Максимальное количество баллов, которое можно было получить за успешность выделения ключевых слов за один текст, — 10. Если участники выписывали более 10 слов из истинного набора КС, они не получали дополнительных баллов. Каждый правильный ответ на вопросы по содержанию текстов оценивался одним баллом. При оценке вопросов, подразумевающих развернутый ответ, использовались ключи со всеми возможными вариантами ответа. Максимальное количество баллов, которое можно было получить за успешность ответов на вопросы по одному тексту, — 8.

Результаты

Во-первых, были сопоставлены истинные наборы КС на разных языках для обоих текстов. Для этого сопоставлялась абсолютная частота встречаемости одинаковых КС в текстах на разных языках (*жест* (55) — *gesture* (45) и т. п.) — и была обнаружена заметная корреляция между истинными наборами КС в обоих парах текстов: «Янус»/«Janus» $r = .605$, $p = .005$, «Шака»/«Shaka» $r = .651$, $p = .001$.

Далее, для оценки влияния ряда факторов на успешность выделения КС и ответов на вопросы по содержанию текста на L2 использовалась линейная регрессия. В качестве зависимых переменных выступили успешность выделения КС и количество правильных ответов на вопросы на L2, в качестве предикторов — успешность выделения КС и ответов на вопросы на L1, уровень языка и текст (см. табл. 2).

При выделении КС на L2 значимое влияние на успешность выполнения задания оказывали успешность выделения КС на L1, а также тип текста (для текста «Шака» лучше выделяли КС). Для предсказания ответов на вопросы на L2 значимыми оказались количество верных ответов на вопросы на L1, а также уровень языка участников. Участники с уровнем C1 лучше справлялись с заданием, чем участники с уровнем B2. По всей видимости, в группу C2 по-

пали участники с очень разным уровнем L2, поэтому значимого эффекта для этого фактора обнаружено не было. Также можно заметить, что успешность выделения КС на L2 и количество правильных ответов на вопросы на L2 оказываются в некоторой степени взаимосвязаны ($p = .047$).

Таблица 2. Факторы, влияющие на успешность выделения КС и ответов на вопросы на L2

Предиктор	Оценка параметра	95%-ные доверительные интервалы		*p*
Успешность выделения КС на L2				
Свободный коэффициент	2.431	0.115	4.748	**.040**
Успешность выделения КС на L1	0.293	0.104	0.478	**.003**
Ответы на вопросы на L1	0.209	−0.052	0.470	.116
Ответы на вопросы на L2	0.218	0.003	0.432	**.047**
Уровень языка (C1)	−0.005	−0.652	0.643	.988
Уровень языка (C2)	0.288	−0.841	1.417	.614
Текст (Шака)	1.095	0.541	1.649	**< .001**
Ответы на вопросы на L2				
Свободный коэффициент	2.380	0.310	4.451	**.025**
Ответы на вопросы на L1	0.444	0.223	0.664	**< .001**
Успешность выделения КС на L1	−0.102	−0.278	0.074	.254
Успешность выделения КС на L2	0.175	0.002	0.348	**.047**
Уровень языка (C1)	1.199	0.668	1.730	**< .001**
Уровень языка (C2)	0.568	−0.439	1.576	.266
Текст (Шака)	0.115	−0.419	0.648	.671

Обсуждение и вывод

Исходя из полученных результатов, можно предположить, что навык выделения КС в меньшей степени лингвоспецифический навык в отличие от навыка

отвечать на вопросы по содержанию текста. Истинные наборы КС для текстов на разных языках будут совпадать и представлять некоторый инвариант текста. По результатам эксперимента успешность выделения КС оказалась связана не с языковой компетенцией читающих, а с содержанием текста, в то время как на правильность ответов на вопросы значимое влияние оказывают языковые компетенции. В то же время успешность выделения КС в некоторой степени отражает понимание конкретного текста, поскольку связана с ответами на вопросы по содержанию. Поскольку традиционной методикой измерения компетенции понимания текста считаются ответы на вопросы по содержанию, существует большое количество эмпирических данных, описывающих вклад в понимание прочитанного ряда факторов, связанных, например, с морфологическим осознанием, словарным запасом, невербальным интеллектом, памятью, вниманием, метакогнитивными навыками и т. д. (напр., Li, Kirby, 2014).

Настоящее исследование представляет результаты использования методики выделения КС при чтении на L1 и L2, которые, на наш взгляд, позволяют сделать предположение о перспективности использования этой методики для измерения компетенции понимания текста. В будущих исследованиях также стоит учесть ряд ограничений текущей работы: 1) участники исследования субъективно определяли уровень L2, 2) участники обладали довольно высоким уровнем L2, 3) были использованы только два текста и одна пара языков L1 – L2, 4) не был учтен порядок записи КС, однако он может оказаться значимым.

Литература

Мурзин Л.Н., Штерн А.С. Текст и его восприятие. Свердловск: Изд-во Урал. ун-та, 1991.

Петрова Т.Е., Риехакайнен Е.И., Кузнецова А.С., Мараев А.В., Шаталов М.А. Выделение ключевых слов в вербальных и невербальных паттернах // Социо- и психолингвистические исследования. 2017. № 5. С. 149 – 156.

de Bruin A.B.H., Thiede K.W., Camp G., Redford J. Generating keywords improves metacomprehension and self-regulation in elementary and middle school children // Journal of Experimental Child Psychology. 2011. Vol. 109. No. 3. P. 294 – 310. https://doi.org/10.1016/j.jecp.2011.02.005

Cain K. Children's reading comprehension difficulties // The science of reading: A handbook / M.J. Snowling, C. Hulme, K. Nation (Eds.). Wiley-Blackwell, 2022. P. 298 – 322. https://doi.org/10.1002/9781119705116.ch14

Engelen J.A.A., Camp G., van de Pol J., de Bruin A.B.H. Teachers' monitoring of students' text comprehension: Can students' keywords and summaries improve teachers' judgment accuracy? // Metacognition and Learning. 2018. Vol. 13. No. 3. P. 287 – 307. https://doi.org/10.1007/s11409-018-9187-4

Firoozeh N., Nazarenko A., Alizon F., Daille B. Keyword extraction: Issues and methods // Natural Language Engineering. 2020. Vol. 26. No. 3. P. 259 – 291. https://doi.org/10.1017/s1351324919000457

Kuperman V., Siegelman N., Schroeder S., Acartürk C., Alexeeva S., Amenta S., Bertram R., Bonandrini R., Brysbaert M., Chernova D., Da Fonseca S.M., Dirix N., Duyck W., Fella A., Frost R., Gattei C.A., Kalaitzi A., Lõo K., Marelli M., Nisbet K., Papadopoulos T., Protopapas A., Savo S., Shalom D.E., Sliousar N., Stein R., Sui L., Taboh A., Tønnesen V., Usal K.A. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus // Studies

in Second Language Acquisition. 2022. Vol. 45. No. 1. P. 3 – 37. https://doi.org/10.1017/s0272263121000954

Lervåg A., Melby-Lervåg M. Modeling the development of reading comprehension // The science of reading: A handbook / M.J. Snowling, C. Hulme, K. Nation (Eds.). Hoboken, NJ: Wiley-Blackwell, 2022. P. 280 – 297. https://doi.org/10.1002/9781119705116.ch13

Li M., Kirby J.R. Unexpected poor comprehenders among adolescent ESL students // Scientific Studies of Reading. 2014. Vol. 18. No. 2. P. 75 – 93. https://doi.org/10.1080/10888438.2013.775130

KEYWORD EXTRACTION METHOD: A PERSPECTIVE ON TEXT COMPREHENSION IN L1 AND L2

V. I. Zubov*, A. A. Konovalova
v.zubov@spbu.ru
Saint Petersburg University, St. Petersburg

Abstract. This study aimed to evaluate the competence of text comprehension in native and non-native languages using a question-answer technique and a keyword extraction method. Native Russian speakers participated in the study and read two texts in their native and non-native (English) languages. The study assessed the success of extracting keywords and the accuracy of answers to content-related questions. A significant overlap was found between the true sets of keywords for the same text in different languages. The results of the experiment indicate that keyword extraction is related to the content of the text rather than the language competence of the readers, while the number of correct answers to questions is significantly influenced by language competencies. We suggest that keyword extraction is less linguistically specific than the skill of answering content-based questions. The keyword extraction method can be employed in further research on text comprehension in native and non-native languages.

Keywords: reading, text comprehension, keyword extraction method, L2, Russian, English

Research supported by the grant ID 94034584 “Text processing in L1 and L2: Experimental study with eye-tracking, visual analytics and virtual reality technologies” from St. Petersburg University.