ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2023. Том 16. Выпуск 10 | 2023. Volume 16. Issue 10
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Лемматизация как средство оптимизации макроструктуры электронного словаря

Балканов И. В.

Аннотация. Цель исследования – определить способы оптимизации операции пользовательской лемматизации при работе с электронным переводным словарем. Научная новизна исследования заключается в том, что в статье впервые рассматривается процедура лемматизации, приведения словоформы к начальной форме, выполняемая программным обеспечением электронного словаря с позиций функциональной теории лексикографии. В статье проводится лексикографический анализ действий и вероятных ошибок, в том числе орфографических, пользователя в процессе лемматизации при обращении к корпусу переводного словаря. Автор изучает решения, предлагаемые составителями печатных словарей, и доказывает, что увеличение объема корпуса словаря усложняет процедуру лемматизации и вынуждает составителя полагаться на знания пользователя; анализирует проблемы, возникающие в ситуации восприятия пользователем информации на слух при формировании запроса в поисковой строке электронного словаря; определяет роль контекста в выборе переводного эквивалента при консультации со словарем; предлагает использовать программы голосового ввода и распознавания речи, а также инструменты проверки орфографии текстовых редакторов для удовлетворения коммуникативно- и когнитивно-ориентированных потребностей пользователя. В результате исследования доказано, что автоматизация процесса лемматизации и интеграция в программное обеспечение электронного переводного словаря инструментов анализа контекста (например, применяемых в текстовых редакторах программ проверки орфографии) повышают эффективность консультации с электронным словарем.

EN

Lemmatization as a macrostructure optimization tool of an electronic dictionary

Balkanov I. V.

Abstract. The paper is aimed at identifying the ways of the optimization of the process of user lemmatization in an electronic translation dictionary. The study is novel in that it is the first to analyze the procedure of lemmatization executed by the software of an electronic dictionary from the standpoint of the functional theory of lexicography. The paper provides a lexicographic analysis of a user's operations and possible mistakes, including spelling, in the process of lemmatization when accessing the corpus of a translation dictionary. The author examines the solutions offered by the compilers of printed dictionaries and proves that an increase in the volume of a dictionary corpus complicates the lemmatization procedure and forces the compiler to rely on a user's knowledge; analyzes the problems arising in the lexicographic situation when a user perceives information by ear and then creates a search query for an electronic dictionary; determines the role of context in the selection of a translation equivalent when consulting a dictionary; suggests using voice input, speech recognition and spellchecking as means to satisfy communicative and cognitive needs of a dictionary user. As a result, the study proves that automation of the lemmatization procedure and integration of context analysis tools into the electronic dictionary software (for example, spellcheckers used in text editors) increase the efficiency of a user's consultation with an electronic dictionary.

Введение

Актуальность настоящего исследования обусловлена тем, что существующие в настоящее время способы оптимизации макроструктуры традиционных словарей (например, гнездование) затрудняют работу со словарем и увеличивают время, необходимое на удовлетворение коммуникативно- и когнитивно-ориентированных потребностей пользователя. Лемматизация, или процесс приведения словоформы к ее начальной (словарной)

Научная статья (original research article) | https://doi.org/10.30853/phil20230503

форме, зависит не только от уровня владения пользователем иностранным языком и культуры его работы со словарем, но и от наличия необходимой грамматической информации в традиционном словаре и соответствующего программного обеспечения (решений составителя-разработчика) в электронном словаре (Bogaards, 2003; Lew, 2012; Балканов, 2017b).

Для достижения поставленной цели исследования необходимо последовательно решить следующие задачи:

– во-первых, проанализировать особенности процесса пользовательской лемматизации при обращении к макроструктуре традиционного переводного словаря с точки зрения функциональной теории лексикографии;

– во-вторых, рассмотреть возможности макроструктуры электронного переводного словаря в процессе пользовательской лемматизации;

– в-третьих, определить отношения контекста и переводного эквивалента, возникающие в процессе консультации пользователя со словарем, в том числе электронным;

– в-четвертых, определить способы сокращения времени и повышения точности обработки поискового запроса на основе фонетической информации о лексической единице.

Для достижения цели и решения задач настоящего исследования использованы методы лексикографического сравнительно-сопоставительного и эмпирического анализа. Применение метода лексикографического сравнительно-сопоставительного анализа позволило нам сопоставить мега-, макро- и микроструктуру исследуемых общих и отраслевых переводных (печатных и электронных) словарей английского языка, а использование метода эмпирического анализа – провести наблюдение за процессом консультации пользователей с используемыми в качестве материала исследования печатными и электронными словарями, выявить типовые проблемы и сложности, возникающие в процессе пользовательской лемматизации, и определить способы оптимизации данной операции и макроструктуры электронного переводного словаря.

Решение поставленных задач стало возможно благодаря теоретической базе, представленной трудами, которые посвящены вопросам традиционной лексикографии (Scholfield, 1999; Bogaards, 2003). П. Богаардс изучал влияние пользователя и создаваемой им лексикографической ситуации на процесс создания переводного словаря. Настоящее исследование во многом является продолжением ранее опубликованных работ автора, в том числе по макроструктуре онлайн-словаря (Балканов, 2021) и теоретическим особенностям двуязычной лексикографии (Балканов, 2017b), а также основывается на трудах ученых, занимавшихся вопросами электронной лексикографии (Hartmann, 1998; Sobkowiak, 1999; 2003; Gamper, Knapp, 2000; Svensén, 2009; Heid, 2011; Lew, 2012; Палкова, 2015; Гончарова, Зацман, 2021; Бойко, 2022). Р. Хартман и Б. Свенсен описали особенности макроструктуры электронного словаря в сравнении со словарем традиционным (Hartmann, 1998; Svensén, 2009), а У. Хейд на основе исследований Б. Свенсена и собственных наблюдений предложил алгоритм работы пользователя с электронным словарем (Heid, 2011). Дж. Гэмпер и Дж. Кнапп обосновали необходимость оптимизации структуры электронного словаря для удовлетворения коммуникативно- и когнитивно-ориентированных потребностей пользователя, что, по мнению исследователей, определяет успех лексикографического проекта (Gamper, Knapp, 2000). Р. Лью, рассуждая о способах повышения эффективности и востребованности электронного словаря, впервые поднял проблему пользовательской лемматизации (Lew, 2012) и в совместной работе с Р. Миттоном рассмотрел возможности электронных словарей в ситуации орфографической ошибки, допускаемой пользователем при консультации со словарем (Lew, Mitton, 2011). В. Собковяк рассматривал возможности фонетического ввода, в том числе с помощью символов фонетической транскрипции, запроса пользователя в поисковую строку электронного словаря (Sobkowiak, 1999; 2003).

Практическая значимость исследования заключается в том, что полученные результаты могут быть использованы для составления практических рекомендаций по работе с электронными словарями, в том числе для решения задачи пользовательской лемматизации. Кроме того, выявленные в ходе настоящего исследования удачные решения составителей могут быть использованы для совершенствования структуры разрабатываемых электронных словарей.

Исследование выполнено на материале общих и отраслевых переводных (печатных и электронных) словарей английского языка, в том числе, но не ограничиваясь:

- Большой англо-русский и русско-английский словарь. М.: Центрполиграф, 2009.
- Большой англо-русский словарь: ок. 150 000 слов / под общ. руковод. И. Р. Гальперина. М.: Совет. энцикл., 1972.
- Гарбовский Н. К. Русско-французский словарь военных терминов. М.: Изд-во Московского университета, 2008.
- Мюллер В. К. Англо-русский и русско-английский словарь: 150 000 слов и выражений. М.: Эксмо, 2014.
- Мюллер В. К. Англо-русский словарь. М.: Гос. изд-во нац. и иностр. словарей, 1960.
- Мюллер В. К. Популярный англо-русский русско-английский словарь. М.: АСТ, 2023.
- Нелюбин Л. Л. Иллюстрированный военно-технический словарь. М.: Воениздат, 1968.
- Судзиловский Г. А. Англо-русский военный словарь: ок. 50 000 терминов. М.: Воениздат, 1968.
- Судзиловский Г. А. Англо-русский военный словарь: ок. 70 000 терминов. М.: Воениздат, 1987.
- Таубе А. М. Военный англо-русский словарь. М.: Гос. изд-во нац. и иностр. словарей, 1942.
- Таубе А. М. Военный англо-русский словарь. М.: Гос. изд-во нац. и иностр. словарей, 1949.
- Таубе А. М. Военный англо-русский словарь: ок. 25 000 слов и терминов. М.: Изд-во нац. и иностр. словарей, 1938.
- Электронный словарь Lingvo Live. https://www.lingvolive.com/en-us.
- Электронный словарь Multitran. https://www.multitran.com.

Обсуждение и результаты

Функциональная теория лексикографии называет ряд условий, которые определяют эффективность словаря как справочника (Балканов, 2017a), призванного удовлетворить коммуникативно- и когнитивно-ориентированные потребности пользователя, т. е. потребности, которые возникают в конкретной несобственно лексикографической ситуации (Tarp, 2011). Так, словарь должен обеспечивать необходимую пользователю степень детализации данных и предоставлять доступ к ним в течение приемлемо короткого периода времени. Лексикографическая и несобственно лексикографическая информация, предлагаемая составителем словаря пользователю, должна быть представлена в максимально понятной и удобной для работы форме.

В силу перечисленных выше факторов сторонники функциональной теории лексикографии приходят к обоснованному выводу о том, что основные требования, предъявляемые пользователем к словарю, а следовательно, и условия, определяющие эффективность его структуры, не зависят от платформы (печатной или электронной), на которой создается словарь (Gamper, Knapp, 2000; Bogaards, 2003; Lew, Mitton, 2011; Heid, 2011). Мы разделяем данную точку зрения, но при этом полагаем, что макроструктура, или упорядоченное расположение словарных статей в корпусе словаря, главенствует над микроструктурой, или структурой словарной статьи, на начальном этапе обращения к словарю. Словарная статья – лексикографическое и несобственно лексикографическое описание леммы, или заголовочного слова, – сама по себе бесполезна, если пользователь не может получить к ней доступ в макроструктуре словаря в приемлемый для него промежуток времени (Gamper, Knapp, 2000; Lew, Mitton, 2011).

Как мы уже отмечали ранее (Балканов, 2017b), в теории лексикографии существуют два основных подхода к организации заголовочных слов в макроструктуре словаря – тематический, или семантический, и алфавитный. При этом большинство составителей толковых и переводных словарей в своей работе придерживаются алфавитного подхода, что обусловлено универсальным представлением людей об алфавите как о единой общепринятой последовательности графических символов (букв), существующей в языке. Семантический подход используется преимущественно при составлении учебных и отраслевых словарей (например, Иллюстрированный военно-технический словарь Л. Л. Нелюбина (1968), который по своей природе является отраслевым переводным словарем, предлагает пользователю более 10 000 единиц военной лексики, сгруппированных в 36 тем, и их эквиваленты на шести иностранных языках).

Одна из проблем теории лексикографии, связанная с организацией макроструктуры словаря, – лемматизация, или приведение словоформы, с которой столкнулся пользователь в несобственно лексикографической ситуации, к начальной форме для последующей консультации со словарем (для глагольных частей речи – форма инфинитива, для именных – форма единственного числа именительного падежа). Формируя поисковый запрос на неродном языке, пользователь зачастую не имеет достаточных для выполнения операции лемматизации знаний и навыков, а следовательно, допускает грамматические и орфографические ошибки и не может в полной мере воспользоваться корпусом словаря, организованным по алфавитному принципу (Lew, 2012).

В традиционном словаре поиск начальной формы обычно подразумевает мысленное удаление флексии: корпус печатного словаря не содержит регулярные флективные формы как самостоятельные заголовочные слова. Мы полагаем (Балканов, 2017b), что подобная операция не вызывает сложностей у пользователей, если флективная форма в алфавитном порядке соседствует (или очень близка) к начальной форме искомой лексической единицы. В такой ситуации пользователь печатного словаря может легко сформировать поисковый запрос.

Затруднения вызывают ситуации, связанные с поиском форм-исключений (например, формы множественного числа некоторых имен существительных или второй/третьей формы неправильных глаголов в английском языке). Лексикограф, предвидя подобные затруднения, может включить в макроструктуру печатного словаря формы-исключения как самостоятельные леммы, которые будут расположены в корпусе словаря в алфавитном порядке. По такому принципу построены общие англо-русские словари В. К. Мюллера (например, Англо-русский словарь (Мюллер, 1960)) и И. Р. Гальперина (например, Большой англо-русский словарь (1972)), а также отраслевые (военные) переводные словари А. М. Таубе (например, Военный англо-русский словарь (Таубе, 1938)). В данных словарях леммы-исключения вместо тела словарной статьи в правой части предлагают пользователю отсылку к словарной статье начальной формы (так, заголовочное слово “went” в переводном англо-русском словаре отсылает к словарной статье для леммы “go”).

Как мы уже отмечали ранее (Балканов, 2017b), выбирая алфавитный подход к организации макроструктуры печатного словаря, лексикограф вынужден использовать различные приемы экономии места (например, гнездование, сокращения, пометы), что позволяет сократить объем, занимаемый словарной статьей на страницах словаря, и, следовательно, предложить пользователю большее количество лексических единиц. К сожалению, поиск путей и способов оптимизации места в корпусе печатного словаря затрудняет работу пользователя: формы-исключения теряют самостоятельную позицию в корпусе словаря (например, Англо-русский военный словарь (Судзиловский, 1968)), а специализированные приложения-списки отсутствуют или не дают желаемой информации (например, Популярный англо-русский русско-английский словарь (Мюллер, 2023)).

Отказ от приема гнездования, наиболее распространенного способа экономии места в корпусе печатного словаря, позволяет представить каждую включенную в словарь лексическую единицу в виде самостоятельной леммы, что значительно сокращает временные затраты пользователя при консультации со словарем.

Однако подобная организация макроструктуры ограничивает общий объем корпуса и количество доступных пользователю лексических единиц (Scholfield, 1999; Svensén, 2009). В первой половине XX века, когда объем отраслевой (например, военной) лексики, подлежащей включению в корпус отраслевого словаря, не превышал 10 000 – 30 000 единиц, разработчики отраслевых переводных словарей (например, А. М. Таубе, составитель основных англо-русских, немецко-русских и французско-русских военных переводных словарей первой половины XX века) зачастую отказывались от приема гнездования, включали в словарь дополнительную лексикографическую и несобственно лексикографическую информацию, что делало отраслевой переводной словарь независимым от общего двуязычного словаря (например, Военный англо-русский словарь (Таубе, 1938)). Так, включенные в корпус военного переводного словаря А. М. Таубе в качестве самостоятельных лемм формы-исключения отсылали пользователя к словарной статье, содержащей начальную словоформу искомой лексической единицы для получения лексикографической и несобственно лексикографической информации о ней (Военный англо-русский словарь (Таубе, 1942)). Составители общих и отраслевых переводных словарей включали в мегаструктуру словаря приложения-списки, содержащие формы-исключения, что давало пользователю возможность лемматизировать словоформу, с которой он столкнулся, получить доступ к искомой словарной статье и удовлетворить свои коммуникативно- и когнитивно-ориентированные потребности (например, Англо-русский словарь (Мюллер, 1960)).

Составители традиционных переводных словарей, в том числе отраслевых (например, Г. А. Судзиловский при работе над военным англо-русским словарем 1987 г. или Н. К. Гарбовский при составлении русско-французского словаря военных терминов 2008 г.), второй половины XX – начала XXI в., в отличие от своих предшественников, вынуждены были использовать приемы экономии места, отказываться от включения в словарь грамматической, фонетической и несобственно лексикографической информации, создавать системы помет и сокращений. Данные меры были обусловлены среди прочего развитием отраслевой, в том числе военной, терминологии и, следовательно, ростом числа лексических единиц, подлежащих включению в корпус словаря. Применение приема гнездования позволяло выстроить макроструктуру словаря в алфавитном порядке по принципу «словарного гнезда»: заголовочное слово с помощью специального знака (например, «|» или «||») делится на словообразовательные аффиксы и основу, которая в теле словарной статьи и других словарных статьях данного гнезда заменяется знаком тильды (~). Прием гнездования предоставляет составителям традиционных переводных и толковых словарей ряд возможностей: экономия места позволяет увеличить объем корпуса словаря, включить в словарные статьи примеры употребления лексических единиц, предложить пользователю несобственно-лексикографическую информацию для удовлетворения его когнитивно-ориентированных потребностей.

Таким образом, исследование перечисленных во Введении печатных словарей первой и второй половины XX и начала XXI в. позволяет нам выделить следующие особенности процесса пользовательской лемматизации при работе с печатными словарями:

1) наличие в макроструктуре словаря форм-исключений позволяет пользователю, не выполняя лемматизацию, перейти к искомой словарной статье и удовлетворить поисковый запрос;

2) включение в мегаструктуру словаря приложений-списков и грамматической информации позволяет пользователю самостоятельно привести искомую форму к начальной и удовлетворить поисковый запрос;

3) увеличение объема корпуса переводного словаря делает словарь зависимым от справочников и знаний пользователя в области грамматики иностранного языка, так как ведет к сокращению объема грамматической информации в мега- и макроструктуре словаря.

В начале XXI в. теоретическая лексикография столкнулась с обратной тенденцией: разработчики электронных словарей, в отличие от составителей печатных словарей, не нуждаются в поиске способов экономии места (Heid, 2011; Lew, 2012; Гончарова, Зацман, 2021). Интерфейс программного обеспечения, использованного при создании электронного словаря, предлагает пользователю доступ к макроструктуре словаря посредством поисковой строки (Heid, 2011). Автоматизация процесса лемматизации облегчает работу пользователя с электронным словарем и делает словарь менее зависимым от уровня владения пользователем иностранным языком (Lew, 2012). Более того, в настоящее время процесс обращения к электронному словарю представляет собой сочетание типовых для современного программного обеспечения действий («копировать лексическую единицу» – «вставить [скопированный фрагмент] в поисковую строку словаря»), в то время как любые дополнительные операции с данной единицей в окне поисковой строки приводят к увеличению временных затрат пользователя, что делает словарь менее привлекательным (Lew, 2012; Балканов, 2021).

Таким образом, мы делаем вывод о том, что грамотно спроектированный электронный словарь может самостоятельно сводить флективную форму к начальной. Тем не менее автоматизированный переход от флективной формы к лемме – это лишь первое условие эффективности макроструктуры современного электронного словаря.

Отметим, что в процессе лемматизации пользователь, в силу типовых ошибок, возникающих в коммуникативно-ориентированных ситуациях (например, при восприятии информации на слух в процессе общения или при просмотре аудиовизуальных произведений на неродном языке без субтитров), неизбежно столкнется с ситуацией ошибочного ввода словоформы при формировании запроса в поисковой строке (Sobkowiak, 1999; Lew, Mitton, 2011). Таким образом, возникает несобственно лексикографическая ситуация, когда пользователь обращается к словарю как орфографическому справочнику, желая уточнить написание слова (Балканов, 2017а). Необходимость приведения словоформы поискового запроса к начальной форме в ее правильном написании, в том числе в ситуации, когда пользователь допускает орфографическую ошибку при формировании запроса, становится предметом исследования теоретической электронной лексикографии (Sobkowiak, 1999; Lew, Mitton, 2011; Lew, 2012).

Отметим, что современное программное обеспечение (например, текстовый редактор Microsoft Word) имеет средства проверки орфографии, которые либо автоматически исправляют ошибки и опечатки пользователей, либо привлекают внимание пользователя к ситуации, в которой могла быть допущена орфографическая ошибка и предлагают варианты написания лексической единицы. Мы полагаем, что подобная практика может быть применена в электронной лексикографии: словарь укажет пользователю на возможную ошибку в написании словоформы, автоматически выполнит операцию лемматизации и предложит ряд словарных статей, в заголовочных словах которых программное обеспечение опознает искомую лемму, и, следовательно, удовлетворит коммуникативно- и/или когнитивно-ориентированные потребности пользователя.

При этом необходимо учитывать тот факт, что текстовые редакторы и программы проверки орфографии, как правило, оптимизированы для пользователей, пишущих на своем родном языке, в то время как большая часть поисковых запросов в словарях выполняется не носителями языка (Sobkowiak, 1999). Кроме того, электронный словарь, в отличие от текстового редактора, не имеет доступа к контексту, что затрудняет процесс проверки орфографии, ведь именно контекст служит источником дополнительных подсказок: эквивалентность текста и адекватность перевода зависят от контекста (Бойко, 2022). Так, без контекста невозможно определить ошибки, связанные с явлением омофонии. Например, на поисковый запрос “peil” (фон. [peil]) (ошибочно транскрибированное пользователем написание английского слова “pale” (также фон. [peil]) («бледнеть», «меркнуть», «терять значение»)) электронный словарь ABBYY Lingvo (ABBYY Lingvo: европейский словарь. Версия 1.12.2, 2021) предлагает 20 вариантов написания (pail, peel, pell, peril, pal и т. д.), ни один из которых не удовлетворяет поисковый запрос. Переводной онлайн-словарь Multitran предлагает словарную статью леммы “PEILS” (аббревиатура “PACOM Executive Intelligence Summary”). Электронный толковый и переводной онлайн-словарь Lingvo Live переадресует нас на словарную статью “peel”, не предлагая альтернативных вариантов написания. Толковые онлайн-словари английского языка (Cambridge. https://dictionary.cambridge.org; Longman. https://www.ldoceonline.com; Macmillan. http://web.archive.org/web/20230307231405/https://www.macmillandictionary.com/; Collins. https://www.collinsdictionary.com; Oxford. https://www.oxfordlearnersdictionaries.com) также с задачей не справляются, формируя списки лемм, которые не удовлетворяют наш запрос. В то же время текстовый редактор Microsoft Word при проверке орфографии в предложении “The threats that you are telling about peil amid the risks taken by the researchers” («Угрозы, о которых ты говоришь, ничто по сравнению с риском, на который идут эти ученые» (здесь и далее – перевод наш. – *И. Б.*)) первым вариантом предлагает нам искомый глагол “pale”.

С точки зрения теории электронной лексикографии для решения задачи лемматизации лексических единиц, в написании которых при формировании запроса допущена орфографическая ошибка, значительный интерес представляет сравнительно-сопоставительный анализ онлайн-словарей английского языка, проведенный Р. Лью и Р. Миттоном (Lew, Mitton, 2011). Учеными были сформированы 202 поисковых запроса, «содержащие наиболее распространенные орфографические ошибки, допускаемые пользователями, изучающими английский язык» (Lew, Mitton, 2011, p. 173). Для каждого запроса отмечалась позиция искомой лексической единицы в списке лемм, составленном электронным словарем в процессе лемматизации при обработке поискового запроса. Исследователями были проанализированы следующие наиболее популярные по числу обращений на момент проведения исследования онлайн-словари английского языка:

1. Advanced Learners Dictionary (толковый онлайн-словарь издательства Merriam Webster для пользователей, изучающих английский язык (рассчитан на пользователей, владеющих иностранным языком на уровне до C1/C2 по общеевропейской шкале CEFR), бесплатная версия).
2. Cambridge Advanced Learner’s English Dictionary (толковый онлайн-словарь издательства Cambridge Dictionaries для пользователей, изучающих английский язык (рассчитан на пользователей, владеющих иностранным языком на уровне до C1/C2 по общеевропейской шкале CEFR), бесплатная версия).
3. Dictionary of Contemporary English (толковый онлайн-словарь современного английского языка издательства Longman (рассчитан на все категории пользователей), бесплатная версия).
4. Dictionary of Contemporary English (толковый онлайн-словарь современного английского языка издательства Longman (рассчитан на все категории пользователей), платная версия, доступ через аккаунт пользователя).
5. Google Dictionary (толковый онлайн-словарь поисковой платформы Google, построенный на базе современного словаря американского английского языка (Oxford Advanced American Dictionary) издательства Oxford University Press (рассчитан на все категории пользователей), бесплатная версия).
6. Macmillan English Dictionary Online (толковый онлайн-словарь английского языка издательства Macmillan (рассчитан на все категории пользователей), бесплатная версия).
7. Oxford Advanced Learners’ Dictionary (толковый онлайн-словарь издательства Oxford University Press для пользователей, изучающих английский язык (рассчитан на пользователей, владеющих иностранным языком на уровне до C1/C2 по общеевропейской шкале CEFR), бесплатная версия).

Результаты исследования показали, что толковые онлайн-словари английского языка издательств Longman и Merriam Webster способны определить и исправить более половины ошибок. Эффективность остальных словарей не превышала 20%. После этого ученые предложили программе проверки орфографии текстового редактора Microsoft Word исправить допущенные пользователями ошибки в контексте. Результат работы программы проверки орфографии, 97%, доказал значимость контекста при формировании пользователем поискового запроса.

Данное исследование подтверждает наше предположение о том, что задача лемматизации словоформ, в написании которых допущены орфографические ошибки, представляет собой существенный вызов для разработчиков электронных словарей и требует отдельного исследования. В качестве решения данной проблемы мы считаем возможным интегрировать в электронный словарь программу проверки орфографии в контексте: при формировании поискового запроса пользователь сможет не только ввести в поисковую строку лексическую единицу в ее отличной от начальной форме, но и, в случае отсутствия результата или неудовлетворенности результатом, уточнить свой запрос, добавив в отдельное поле фрагмент текста, в котором он столкнулся с данной единицей. Это значительно увеличит время работы со словарем, но позволит более точно удовлетворить поисковый запрос, в том числе в ситуации орфографической ошибки, допущенной пользователем при лемматизации.

В теории электронной лексикографии существуют другие варианты решения проблемы орфографической ошибки при формировании поискового запроса. Так, в своих исследованиях В. Собковяк предлагает повысить эффективность электронного словаря с помощью включения в поисковую строку средств поиска по фонетическим символам (Sobkowiak, 1999; 2003). Мы полагаем, что данный способ не получит широкого распространения в практике электронной лексикографии, поскольку требует от пользователя уверенных знаний системы фонетических символов и принципов транскрибирования неродного языка, что не является необходимым условием удовлетворения коммуникативно-ориентированных потребностей среднестатистического пользователя, так как электронные словари, в отличие от словарей печатных, предлагают аудиозаписи произношения заголовочных слов (в том числе на разных вариантах языка (например, американский и британский английский)) (Lew, 2012). Отметим, что те немногие пользователи, которые демонстрируют достаточный для формирования фонетического запроса уровень знаний, умений и навыков работы с фонетической информацией, обладают уверенными знаниями о системе иностранного языка в целом и могут использовать другие способы поиска информации для удовлетворения своих потребностей при работе со словарем (Lew, Mitton, 2011; Lew, 2012). Конечно, это не означает, что фонетический путь доступа к информации не должен предлагаться в качестве альтернативы в электронном словаре, особенно если его можно обеспечить без особых дополнительных затрат со стороны разработчика, что, например, и было реализовано при создании учебного электронного словаря английского языка Macmillan English Dictionary 2002 года и дополнения SoundSearch на CD-диске к нему (http://macmillandictionaries.com/MED-Magazine/November2002/02-CD-ROM-soundsearch-US.htm). В настоящее время данный словарь является единственным толковым словарем английского языка, предлагающим пользователю возможность формирования запроса по фонетическим символам.

Другим способом фонетического доступа к корпусу электронного словаря является прямое использование аудиоканала, когда специальное программное обеспечение переводит голосовой запрос в текстовый и предлагает лемму или несколько лемм, способных удовлетворить коммуникативно- и когнитивно-ориентированные потребности пользователя (Тюменцев, Мелешенко, 2021; Балканов, 2021).

Программы распознавания речи создаются многими производителями программного обеспечения (Apple, Microsoft, Facebook (признана экстремистской организацией, ее деятельность запрещена на территории РФ), Yandex, Google), встраиваются в текстовые редакторы, поисковые системы, программы обмена сообщениями, доступны пользователям различных платформ (персональных компьютеров, ноутбуков, смартфонов). Программы голосового ввода способны обучаться и подстраиваться под конкретного пользователя, что значительно повышает их эффективность (Тюменцев, Мелешенко, 2021). Мы полагаем, что интеграция данных программ в электронные словари может значительно расширить возможности последних, повысить качество и сократить время консультации пользователя со словарем.

Более того, программы распознавания речи могут быть научены работе с контекстом. Например, пользователь может сначала произносить словоформу поискового запроса, а после нажатия на специальную клавишу – фразу, в которой данная форма была использована. Таким образом, программное обеспечение словаря сможет предложить пользователю список лемм, соотнесенных с контекстом. Мы полагаем, что реализация подобных предложений возможна только в рамках крупного лексикографического проекта. Во-первых, программы распознавания речи разрабатываются крупными компаниями, которые за их использование в лексикографических инструментах могут потребовать вознаграждение. Во-вторых, только масштабный проект позволит сделать словарь обучаемым и научит машину понимать особенности речи как конкретного пользователя, так и определенной этнической общности. К сожалению, отсутствие словарей, предлагающих пользователю возможность голосового ввода при формировании поискового запроса, не позволяет нам оценить эффективность такого способа доступа к макроструктуре.

Заключение

В результате настоящего исследования нами были сделаны следующие выводы:

1. При работе с печатным переводным словарем пользователь с опорой на ранее полученные знания самостоятельно выполняет приведение искомой словоформы к начальной (как правило, путем отбрасывания флексии). Традиционные переводные словари позволяют пользователю лемматизировать искомую неправильную форму (форму-исключение) как за счет грамматической информации, помещенной в структуру словарной статьи (указание на форму-исключение с последующей отсылкой к словарной статье, содержащей начальную форму искомой единицы), так и путем обращения к приложениям, помещенным в раздел предваряющих или завершающих текстов словаря. При этом увеличение объема корпуса вынуждает составителей

искать способы экономии места, в том числе за счет сокращения объема грамматической информации, что возлагает задачу лемматизации на самого пользователя и заставляет его опираться на собственные знания грамматики иностранного языка. Это усложняет процесс работы со словарем и вынуждает пользователя обращаться за помощью к дополнительным ресурсам (учебным пособиям и грамматическим справочникам).

2. Электронный переводной словарь способен самостоятельно приводить флективные формы и формы-исключения к лемме, что сокращает время, требуемое на решение коммуникативно- и когнитивно-ориентированных задач пользователя, а также открывает доступ к словарю пользователям, начинающим изучать иностранный язык. Тем не менее автоматизированная лемматизация является лишь одним из условий эффективности современного электронного словаря.

3. Анализ контекста определяет правильность выбора пользователем переводного эквивалента. Электронный словарь должен не только предлагать пользователю набор переводных эквивалентов к искомой лексической единице, но и соотносить данные эквиваленты с ситуацией (контекстом), которая подтолкнула пользователя к консультации со словарем.

4. Электронный словарь, в отличие от печатного словаря, может со временем как интегрировать в свою структуру программы анализа контекста, работающие по принципу инструментов проверки орфографии текстовых редакторов, так и научиться воспринимать и переводить в текст речь пользователя, что значительно сократит время и повысит точность обработки поискового запроса.

На основе сделанных выводов и проведенного сравнительно-сопоставительного лексикографического анализа мы можем определить следующие способы оптимизации операции пользовательской лемматизации и макроструктуры электронного переводного словаря:

– включение в структуру словарной статьи грамматической информации о заголовочном слове;
– автоматизация процесса приведения искомой словоформы (правильной и неправильной) к начальной;
– интеграция в программное обеспечение электронного словаря инструментов проверки орфографии и анализа контекста, а также программ распознавания речи пользователя.

Перспективы дальнейшего исследования мы видим в изучении типовых ошибок, обусловленных родным языком пользователя переводного словаря, что позволит лексикографам продвинуться в решении задачи оптимизации операции лемматизации. Отдельного исследования заслуживают ситуации, когда текст поискового запроса представляет собой сочетание нескольких лексических единиц (например, устойчивое словосочетание или фразеологический оборот). Кроме того, мы предлагаем провести сопоставление макроструктуры печатного и электронного переводного словаря с учетом коммуникативно- и когнитивно-ориентированных потребностей пользователя в ситуации, когда корпус словаря построен по семантическому (тематическому) принципу.

Источники | References

1. Балканов И. В. К вопросу организации макроструктуры онлайн-словаря // Военно-филологический журнал. 2021. № 3.
2. Балканов И. В. Словарь как справочник, текст и коммуникативная система // Филологические науки. Вопросы теории и практики. 2017a. № 3-1 (69).
3. Балканов И. В. Теоретические основы двуязычной лексикографии (на материале военных переводных словарей): дисс. … к. филол. н. М., 2017b.
4. Бойко Б. Л. Военная лексика и военная терминология в словарях и текстах различных функциональных стилей // Военно-филологический журнал. 2022. № 3.
5. Гончарова А. А., Зацман И. М. Принципы структуризации статей в электронных словарях // Информатика и ее применения. 2021. Т. 15. № 2.
6. Палкова А. В. Основные понятия электронной лексикографии // Вестник Тверского государственного университета. Серия «Филология». 2015. № 4.
7. Тюменцев Е. А., Мелешенко Т. В. Разработка адаптируемого под пользователя редактора форм ввода данных // Математическое и компьютерное моделирование: сб. мат. IX междунар. науч. конф. Омск, 2021.
8. Bogaards P. Uses and Users of Dictionaries // A Practical Guide to Lexicography, Terminology and Lexicography Research and Practice / ed. by P. van Sterkenburg. Amsterdam – Philadephia: John Benjamins, 2003.
9. Gamper J., Knapp J. Towards an Adaptive Learners' Dictionary // Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2000). Berlin – N. Y.: Springer, 2000.
10. Hartmann R. R. K. Dictionary of Lexicography. L.: Routledge, 1998.
11. Heid U. Electronic Dictionaries as Tools: Towards an Assessment of Usability // E-Lexicography: The Internet, Digital Initiatives and Lexicography / ed. by P. A. Fuertes-Olivera, H. Bergenholtz. L. – N. Y.: Continuum, 2011.
12. Lew R. How Can We Make Electronic Dictionaries More Effective // Electronic Lexicography / ed. by S. Granger, M. Paquot. Oxford: Oxford University Press, 2012.
13. Lew R., Mitton R. Not the Word I Wanted? How Online English Learners' Dictionaries Deal with Misspelled Words // Electronic Lexicography in the 21st Century: New Applications for New Users / ed. by K. Kosem. Ljubljana: Trojina, Institute for Applied Slovene Studies, 2011.
14. Scholfield P. Dictionary Use in Reception // International Journal of Lexicography. 1999. Vol. 12 (1).
15. Sobkowiak W. Pronunciation in EFL Machine-Readable Dictionaries. Poznań: Motivex, 1999.

16. Sobkowiak W. Pronunciation in Macmillan English Dictionary for Advanced Learners on CD-ROM // International Journal of Lexicography. 2003. Vol. 16 (4).
17. Svensén B. A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press, 2009.
18. Tarp S. Lexicography for the Third Millennium: Cognitive-Oriented Specialised Dictionaries for Learners // Ibérica. 2011. Vol. 21 (1).

Информация об авторах | Author information

RU

Балканов Илья Владимирович[1], к. филол. н., доц.
[1] Московский государственный институт международных отношений (МГИМО);
Военный университет им. князя Александра Невского Минобороны России, г. Москва

EN

Balkanov Ilya Vladimirovich[1], PhD
[1] Moscow State University of International Relations (MGIMO);
Alexander Nevsky Military University of the Russian Defense Ministry, Moscow

[1] *i-balkanov@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 04.08.2023; опубликовано online (published online): 28.09.2023.

Ключевые слова (keywords): электронная лексикография; корпус словаря; заголовочное слово; пользователь словаря; electronic lexicography; dictionary corpus; headword; dictionary user.