

Оригинальная статья / Original Article

DOI: 10.31857/S1605788024040106

Принципы и методы цифровой лексикографии

© 2024 г. Ю. Ю. Макаров

Научный сотрудник Института русского языка им. В.В. Виноградова РАН,
Россия, 119019, Москва, ул. Волхонка, д. 18/2,
младший научный сотрудник Института языкознания РАН,
Россия, 125009, Москва, Большой Кисловский пер., д. 1, стр. 1,
приглашенный исследователь НИУ ВШЭ, Россия, 115054,
Москва, ул. Старая Басманная, д. 21/4
yurmak@iling-ran.ru

Резюме. Статья описывает принципы и методы цифровой лексикографии, а также то, как они встраиваются в четыре основных этапа создания словарей: подготовку текста, редакционно-издательскую работу, публикацию и постпубликационный период. Обсуждаются преимущества онлайн-систем подготовки словарей (таких, как платформа OnLex), которые значительно упрощают все упомянутые этапы. Вводится классификация подходов к написанию текста словаря (файловый, табличный, основанный на TEI-разметке или базе данных). В заключение подчеркивается важность применения методов цифровой лексикографии, особенно в российской традиции, часто ограничивающейся малофункциональным файловым подходом.

Благодарность. Автор благодарен В.А. Плунгяну за ценные замечания к тексту статьи, а также Д.А. Рыжовой за помощь в поиске мотивации. Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Ключевые слова: лексикография, словари, цифровые методы, системы подготовки словарей, OnLex.

Для цитирования: Макаров Ю.Ю. Принципы и методы цифровой лексикографии // Известия Российской академии наук. Серия литературы и языка. 2024. Т. 83. № 4. С. 102–112. DOI: 10.31857/S1605788024040106

Principles and Methods of Digital Lexicography

© 2024 Yury Makarov

Research Fellow at the V.V. Vinogradov Russian Language Institute
of the Russian Academy of Sciences,
18-2 Volkhonka Str., Moscow, 119019, Russia,
Junior Researcher at the Institute of Linguistics
of the Russian Academy of Sciences,
1 Build. 1 Bolshoy Kislovsky Lane, Moscow, 125009, Russia,
Visiting Scholar at the National Research University “Higher School of Economics”
21/4 Staraya Basmanaya Str., Moscow, 115054, Russia
yurmak@iling-ran.ru

Abstract. The article describes the principles and methods of digital lexicography. It begins by defining the four main stages of the lexicographic process: 1) writing up the dictionary, 2) editing and developing the book layout, 3) publishing, and 4) the post-publication period. The following section focuses on stage 1, comparing the compilation of example corpora for dictionary preparation in the past (using millions of cardboard cards) with modern tools for lexical analysis provided by web corpora like the Russian National Corpus (ruscorpora.ru). The overview of the advancements in finding examples illustrating word usage is followed by an exploration of the ways dictionary writing methods have evolved.

The analysis of computer-based dictionary writing methods starts with a discussion of the two most popular approaches: file-based and tabular. The former involves composing dictionary files with thousands of entries using text editors like Microsoft Word, resulting in poorly structured entries with inconsistent markup. The latter, however, represents each entry as a row with entry zones (forms, meanings, examples, etc.) arranged in separate columns. The section outlines the challenges of these methods, emphasizing their limitations in publishing options and handling complex linguistic data, often employing many-to-one relationships. Alternatives such as Text Encoding Initiative (TEI) formats and database utilization are discussed, highlighting their capacity for structured data representation.

Subsequently, dictionary writing systems (DWS) are introduced, with the OnLex platform serving as a primary example illustrating their functionality. It demonstrates how online editing interfaces streamline lexicographic processes, from data input to publication and feedback collection. By analyzing DWS features, the article emphasizes their efficacy in simplifying the editorial workflow and enhancing user experience.

A critical appraisal of the advantages of online DWS is provided, highlighting their role in addressing key challenges faced by traditional publishing methods. Notable advantages include seamless integration of search functionalities, support for multiple languages, and real-time error reporting mechanisms after publication.

In conclusion, the article advocates for the wider adoption of digital lexicography methods, particularly within the Russian tradition, emphasizing their potential to facilitate every stage of the dictionary creation process.

Acknowledgements. I would like to thank Prof. Vladimir Plungian for his valuable comments on this paper and Dr Daria Ryzhova for helping me find motivation for writing this article. Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

Key words: lexicography, dictionaries, digital humanities, dictionary writing systems, OnLex.

For citation: Makarov, Y. *Printsipy i metody tsifrovoy leksikografii* [Principles and Methods of Digital Lexicography]. *Izvestiâ Rossijskoj akademii nauk. Seriâ literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2024, Vol. 83, No. 4, pp. 102–112. (In Russ.) DOI: 10.31857/S1605788024040106

Этапы лексикографического проекта

Традиционный путь, который проделывает лексикографический проект, состоит из четырех этапов:

1. Подготовка текста словаря.
2. Макетирование и прочая работа с издательством.
3. Публикация книги.
4. Постпубликационный период.

Вкратце охарактеризуем каждый из них.

Этап **подготовки текста словаря** заключается в определении концепции будущего издания, обсуждении словника, сборе корпуса примеров, написании словарных статей, а также их последующем (обычно коллегиальном) редактировании.

Следующий этап, связанный с **редакционно-издательской подготовкой**, обычно касается различных аспектов взаимодействия авторского коллектива с внешними редакторами, корректорами и/или наборщиками. Во многих случаях подготовка макета сопровождается массой проблем: даже если отбросить всевозможные разногласия в стилистической правке, преобразование исходного текста в издательский формат, подходящий для отправки в типографию, занимает долгое время и часто проходит через несколько итераций. Более того, уже сверстаный макет перед утверждением снова приходится вычитывать.

Название предпоследнего этапа, **публикация книги**, говорит само за себя. Стоит отметить, что

в российской традиции основной акцент делается именно на издании бумажной версии, и лишь с недавнего времени крупные лексикографические проекты стали параллельно с физическими копиями подготавливать онлайн-версии [1]–[3].

Завершающий **постпубликационный период** характеризуется сбором обратной связи, отзывов о словаре, их осмыслением и, возможно, подготовкой нового издания, хотя в силу финансовых ограничений новое издание в большинстве случаев невозможно.

Цель данной статьи – описать, как каждый из перечисленных этапов преобразуется при использовании методов цифровой лексикографии, а также на конкретных примерах проиллюстрировать, как эти методы работают.

Подготовка текста словаря

Время, когда текст словаря подготавливался с использованием чернил и бумаги, осталось позади с приходом компьютерных технологий. Источниковые базы, представленные прежде миллионами примеров на картонных карточках, как правило¹, были заменены цифровыми инструментами наподобие Национального корпуса русского языка (ruscorpora.ru; [4]). Счет словоформ,

¹ Исключение составляют проекты, в картотеках которых собран уникальный материал, не загруженный на платформы типа НКРЯ в цифровом виде, например Словарь русских народных говоров.

составляющих современные онлайн-корпусы, ведется уже не на миллионы, а на миллиарды: на момент написания статьи суммарный объем под-корпусов НКРЯ составляет более двух миллиардов словоформ, а Генеральный интернет-корпус русского языка (webcorpora.ru; [5]; [6]) насчитывает больше 20 миллиардов слов. Помимо объема, важным преимуществом онлайн-формата является возможность гибких поисковых запросов, которые существенно упрощают комплексное изучение лексемы. Несмотря на то, что корпусные технологии являются примером оптимизации лексикографического процесса цифровыми методами, далее они рассматриваться не будут (они в деталях описаны среди прочего в указанных выше источниках).

Файловый подход

Если корпусы пришли на замену картонным карточкам, то что же стало альтернативой чернилам и бумаге? Можно констатировать, что наибольшее распространение в современной российской традиции имеет набор текста с помощью текстовых редакторов наподобие Microsoft Word, или **файловый подход**. Среди его преимуществ можно выделить легкость освоения и широкую доступность, однако нельзя не заметить существенно большее число его недостатков. Начать их рассмотрение стоит с замечания о плохой совместимости с масштабными проектами. При объеме хотя бы в несколько десятков тысяч словарных входов даже мощные компьютеры будут с трудом обрабатывать файлы формата *doc* и *docx*, наиболее распространенные среди пользователей Microsoft Word. Частым решением этой проблемы является разделение текста словаря на фрагменты, содержащиеся в разных файлах, однако стоит отметить, что такой модус работы едва ли надежен (файлы имеют свойство теряться) и удобен (обращение с десятками файлов приводит к путанице). Упомянутые неудобства становятся еще более критичными, если рассмотреть проблему версирования: сбор и хранение версий файлов для различных фрагментов излишне сближает работу лексикографа с работой архивариста.

До сих пор рассматривались скорее проблемы с внешней стороной файлового подхода; обратимся теперь к недостаткам, связанным непосредственно с процессом подготовки текста. Даже если правила оформления различных элементов словарной статьи (заглавное слово, пометы, толкования, речения, идиомы...) были оговорены заранее, в итоговом файле все равно найдутся сотни примеров их нарушения. Отсутствующий курсив, пропущенное

отточие и некорректно расставленные переносы строк, хотя и не всегда критичны с содержательной точки зрения, существенно портят восприятие издания. Среди прочих проблем можно упомянуть плохой доступ к метainформации (кто, когда и что изменял в конкретной статье?), статистике (сколько словарных статей готовы?), невозможность удобного поиска (например, только среди значений или заглавных слов), часто блокируемого диакритиками, а также слабой конвертируемостью в издательские форматы².

Табличный подход

Проблема плохой структурированности менее выражена при **табличном подходе**, который отличается от файлового тем, что словарные статьи представляют собой строки таблицы, а словарные зоны — столбцы. Рисунок 1 иллюстрирует такой тип словаря.

Преимущества табличного подхода проистекают из более строгого контроля за структурой данных (в противном случае таблица просто не будет правильно отображаться): становится доступным поиск по конкретным словарным зонам (путем фильтрации по столбцам). Экспорт файлов формата *csv*, *tsv*, *xml*, *xls*, *xlsx* и проч. в виде онлайн-страницы с возможностью фильтрации по столбцам (ср. рис. 1) и даже для подготовки макета книги в табличном подходе существенно упрощены (видимо, по этой причине издания в табличном формате популярны среди авторов учебных пособий).

Недостатки же проистекают, как это ни парадоксально, из той же строгости организации данных. Если для небольших словарей типа [7] табличный подход может быть достаточным, в более комплексных проектах, где требуется, например, указание массы грамматических форм, перечисление фразеологических единиц, включение обильного иллюстративного материала и т.п., табличное представление вокабулы окажется неоправданно громоздким. Представим, что в словаре регулярно дается пять форм глагола, три формы существительного, а также четыре формы прилагательного. Согласно принципам табличного подхода, каждый тип форм, каждая глосса должна быть вынесена в отдельный столбец. Таблица таким образом увеличивается на 12 столбцов, большая часть которых будет всегда пустой (у глаголов нет именных форм и т.д.). Если же вспомнить о примерах и идиомах, а также прочих

² Верстка макета с помощью Microsoft Word возможна, но не слишком распространена в профессиональной среде.

	id	Russian	English	orthography	transcription	pos
	All	All	All	All	All	All
⊖	1	ЗЕМЛЯ (ПОВЕРХНОСТЬ)	GROUND	гӀунщи	ʒunʃi	n
forms: гӀуншил; гӀунщибади, гӀунщалди additional forms: context: Твёрдая поверхность земли; отличать от земли как материала и от почвы. meaning: Животные обитают на земле, рыбы — в воде. – Рух1ч1агольирди гӀумругирахе гӀуншила, ччуг1иби- лӀинл1и.						
⊕	2	ЗЕМЛЯ (ПОЧВА)	EARTH	гӀунщи	ʒunʃi	n
⊕	3	ПЫЛЬ	DUST	хӀури	ħuri	n
⊕	4	ГРЯЗЬ	MUD	хӀари	ħari	n
⊕	5	ПЕСОК	SAND	сали	sali	n

Рис. 1. Словарь тукитинского диалекта каратинского языка (lingconlab.ru/TukitaDict/; [7])

потенциальных зонах, то внешний облик максимально подробного словаря-таблицы оказывается за пределами возможностей восприятия³. Помимо этого, одна таблица априори не может успешно отображать many-to-one («многие к одному») отношения, которыми пронизан типичный словарь. Чтобы проиллюстрировать этот тип отношений, достаточно вспомнить о том, как заглавная лексема ('one') соотносится с ее значениями ('many') и как, в свою очередь, каждое значение ('one') соотносится с примерами и предложениями ('many'). В таблице как на рис. 1 отображение подобных связей вынуждено сведено к примитивному перечислению в одной ячейке, о проблематичности которого см. выше. Иными словами, табличный подход плохо совместим с полисемией, вариативностью и прочими атрибутами любого естественного языка.

Язык разметки XML и стандарт TEI

Стоит специально остановиться на определении XML (англ. *extensible markup language* 'расширяемый язык разметки') — метаязыка, с помощью которого делаются указания о том, как должен интерпретироваться тот или иной фрагмент цифрового текста (поэтому XML и называется языком разметки). Важнейший принцип организации XML — иерархичность; более того, метки, или тэги,

³ Следует отметить, что перечисление всех форм, примеров и прочих единиц в одной ячейке приведет к существенным потерям в организации данных, а также в возможностях поиска (так, если все именные формы перечислены через запятую, то нельзя выполнить поиск только среди, например, форм множественного числа).

используемые для разметки, могут быть вложены друг в друг, как показано на рис. 2а.

TEI (tei-c.org; англ. *text encoding initiative* 'инициатива по кодированию текста') представляет собой стандарт для создания, обмена и анализа текстовой информации в цифровой форме. В частности, TEI предоставляет рекомендации по разметки разнообразных текстовых документов с использованием XML. Среди сотен модулей (т.е. наборов тэгов и правил их использования) имеется и схема описания словарей, см. рис. 2а. Помимо этого, существует проект TEI-Lex-0⁴ (рис. 2б), на большом количестве примеров показывающий, как словарные статьи разных форматов могут быть представлены с помощью более продвинутого набора тэгов (при этом не нарушающих стандарт TEI).

TEI широко используется в сфере digital humanities (~ цифровых гуманитарных наук) для создания электронных архивов, текстовых корпусов и других проектов, требующих структурированного представления текста. Одним из преимуществ этого стандарта является представление каждого элемента словарной статьи как отдельной сущности (имеется свой набор тэгов для каждой словарной зоны). Если разные словарные проекты реализуются в соответствии с TEI, лексикографические процессы существенно упрощаются: достаточно лишь один раз разработать программное обеспечение, позволяющее записывать словарные статьи в формате TEI через удобный интерфейс,

⁴ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>


```

<body>
  <div>
    <head>English-French</head>
    <entry>
      <form>
        <orth>cat</orth>
      </form>
    <!-- ... -->
  </entry>
</div>
<div>
  <head>French-English</head>
  <entry>
    <form>
      <orth>chat</orth>
    </form>
  <!-- ... -->
</entry>
</div>
</body>

```

```

<entry xml:lang="en" xml:id="CH.gol"
  xml:base=".../TEILex0.examples/examples.stripped.xml">
  <form type="lemma">
    <orth>go</orth>
    <pron>gə</pron>
  </form>
  <lbl rend="sup">1</lbl>
  <gramGrp>
    <gram type="pos">vi</gram>
  </gramGrp>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <gram type="participle">prp</gram>
    </gramGrp>
    <orth>gō'ing</orth>
  </form>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <gram type="participle">pap</gram>
    </gramGrp>
    <orth>gone</orth>
    <pron>gon</pron>
    <note>(see separate entries)</note>
  </form>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <gram type="participle">pat</gram>
    </gramGrp>
    <orth>went</orth>
    <note>(supplied from <xr type="related">
      <ref type="entry">wend</ref>
    </xr></note>
  </form>
  <pc></pc>
  <form type="inflected">
    <gramGrp>
      <gram type="person">3rd</gram>
      <gram type="tense">pers</gram>
      <gram type="number">sing</gram>
      <gram type="tense">pres</gram>
      <gram type="mood">indicative</gram>
    </gramGrp>
    <orth>goes</orth>
  </form>
  <pc></pc>
  <!-- ... -->
</entry>
</Chambers (2011)>

```

go¹ /gə/ vi (prp gō'ing; pap gone /gon/ (see separate entries); pat went (supplied from wend); 3rd pers sing pres indicative goes) used

dariah-eric.github.io

Рис. 2. Примеры а) TEI-разметки (слева) и б) TEI-Lex-0-разметки (справа) фрагмента словарной статьи

а также автоматически конвертировать TEI-словарь в издательский макет. Пример проекта, использующего TEI при создании и публикации словаря, описан в [2]; [8].

База данных как основа словаря

Перед тем как перейти к рассмотрению систем подготовки словарей, стоит ввести понятие базы данных (БД). БД — это некоторая информация, которая хранится в соответствии со строгой схемой. В отличие от таблицы в табличном подходе (см. выше), где словарная статья соответствует одной строке, при использовании БД вокабула представима в виде нескольких таблиц с прописанными отношениями. БД, в отличие от табличного подхода, поддерживает many-to-one отношения между элементами словарной статьи. Предположим, что в некотором словаре имеется заглавная лексема, а также множество грамматических форм и множество значений, привязанных к заглавной лексеме; при этом у каждого значения есть множество речений и примеров. Выше было показано, что табличный подход не может эффективно отобразить подобную структуру. В БД же каждое из множеств представимо в виде отдельной таблицы, причем в этой таблице у каждой лексемы, формы, значения или примера / речения есть идентификационный номер (ID). Связи же между этими

множествами могут быть прописаны в отдельных таблицах, сопоставляющих, например, одному и тому же ID заглавной лексемы множество различных ID форм и значений.

По сравнению с файловым и табличным подходами БД могут хранить огромные массивы данных, а системы управления базами данных (СУБД) — эффективно управлять информацией в БД, см. рис. 3.

Несмотря на то, что использование БД в качестве основы словаря позволяет структурированно хранить данные с поддержкой различных типов отношений (в том числе many-to-one), а также легко конвертировать эти данные в другие форматы, обращение с (СУБД) предполагает наличие ряда технических навыков (как минимум владение языком SQL от англ. *structured query language* ‘язык структурированных запросов’), что едва ли можно требовать от лексикографа. БД плохо совместима с медиафайлами (иллюстрации, аудио-приложения и проч.) и не может заменить собой публикацию словаря в интернете или в виде физического издания.

Системы подготовки словарей

Рассмотренный выше набор инструментов (TEI-разметка, БД) нельзя назвать дружелюбным

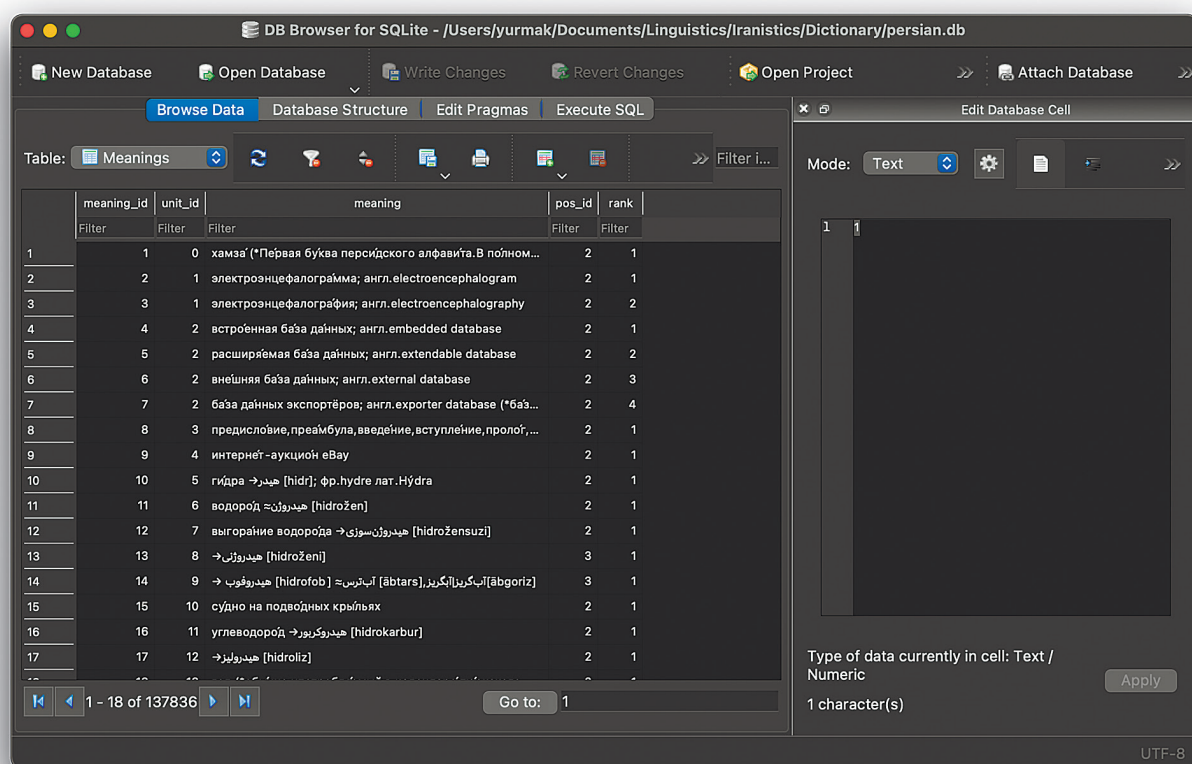


Рис. 3. Снимок экрана СУБД SQLite, показывающий таблицу значений онлайн-словаря персидского языка *iranica.space* [1], [9]; *meaning_id* – ID значения, *unit_id* – ID заглавной лексемы, *meaning* – текст значения, *pos_id* и *rank* – технические индексы

по отношению к лексикографу, от которого несправедливо было бы требовать владения массой технических компетенций в дополнение к лингвистическим и филологическим. Именно поэтому с 1990-х годов разрабатываются системы, упрощающие взаимодействие авторов словарей с цифровыми методами лексикографии, ср. [10]. Эти **системы подготовки словарей** (СПС; англ. *dictionary writing systems*) можно разделить на коммерческие (доступные для широкой аудитории; например, IDM DPS⁵ или TLex⁶) и внутренние (созданные для конкретных проектов и недоступные для прочих); дополнительно стоит различать СПС, ориентированные на работу в интернете (онлайн-СПС), и офлайн-СПС, требующие установки программ на компьютер. Общими составляющими СПС является некоторая система организации и хранения данных (таблица, БД...), редакторский интерфейс, а также модули управления проектом, экспорта данных в издательских форматах и иногда веб-модуль.

⁵ <https://dps.cw.idm.fr/>

⁶ <https://tshwanedje.com/tshwanelex/>

Проиллюстрируем работу СПС на примере редакторского интерфейса **платформы OnLex**⁷, с 2020 г. разрабатываемой автором данной статьи. OnLex представляет собой онлайн-СПС, т.е. все лексикографическими процессами можно управлять, используя браузер на любом устройстве. На рис. 4 представлен фрагмент страницы, используемой для редактирования словарных входов, а именно поля для оригинального текста статьи (скопированного из источника без изменений), значений заглавной лексемы, а также для идиом и сложных глаголов. Во все эти поля возможен ввод текста в произвольном формате, в то время как в случае со списком помет (например, грамматические пометы F, M и CAUS, т.е. мужской и женский роды, а также каузатив) и выбором части речи необходимо сделать выбор из предварительно заданных вариантов. Такой

⁷ На платформе OnLex в данный момент функционируют три словаря: памирских (*ramiri.online*; [11]), персидского (*iranica.space*; [1]; [9]; [12]; [13]) и осетинского (*ossetica.space*; [2]; [8]) языков. На стадии разработки находятся словари сербского языка [3], Академического толкового словаря русского языка [14]; [15]; [16], а также другие проекты.

Оригинальный текст статьи

ш. ДОД 1 отец; ш. ам тāt лўвам, ам дод (мы) и "тат" (отец) говорим и "дод"; брв. ху дод йах, ху холāk-ат мак ўйн-та виц лўвām сестру своего отца, жену своего дяди по матери и отцу мы называем "виц" (тетя); ш. дōд-ат нāн отец и мать, родители; ш. хāби-нди ар чāлак мā-ниѳ, бе дōд-ат нāн сāви в ночное время не садись у очага, (не то) останешься без родителей (народная примета); ш. ху дод гўхт хйдов сильно мучиться, терзаться; страдать; ш. wўз-ум ви қати ху дод гўхт хўўч я с ним измучился; ш. дод дивйхtow сильно измучить, истерзать, извести; ш. йид хйч му дōд-и му-рд дивйхт этот бык извел меня. Ср. тāt, пид, падар.

Значения заглавной лексемы

Добавить значение

Ранг	Пометы	Значение
✗ 1 1	ПОМЕТЫ сущ. ▾ нет F M CAUS	отец

Устойчивые сочетания и идиомы

Добавить идиому или устойчивое сочетание

№	Формы	Значения												
✗ 1	<table><thead><tr><th>№</th><th>Пометы</th><th>Орфографические варианты</th></tr></thead><tbody><tr><td> ✗ 1 </td><td> ш. </td><td> Латиница dod-at nān Кириллица дōд-ат нāн </td></tr></tbody></table> Добавить форму	№	Пометы	Орфографические варианты	✗ 1	ш.	Латиница dod-at nān Кириллица дōд-ат нāн	<table><thead><tr><th>Ранг</th><th>Пометы</th><th>Значение</th></tr></thead><tbody><tr><td> ✗ 1 1 </td><td> ПОМЕТЫ не ре ▾ нет F M CAUS </td><td> отец и мать, родители </td></tr></tbody></table> Добавить значение	Ранг	Пометы	Значение	✗ 1 1	ПОМЕТЫ не ре ▾ нет F M CAUS	отец и мать, родители
№	Пометы	Орфографические варианты												
✗ 1	ш.	Латиница dod-at nān Кириллица дōд-ат нāн												
Ранг	Пометы	Значение												
✗ 1 1	ПОМЕТЫ не ре ▾ нет F M CAUS	отец и мать, родители												

Сложные глаголы

Добавить сложный глагол

№	Формы	Значения												
✗ 1	<table><thead><tr><th>№</th><th>Пометы</th><th>Орфографические варианты</th></tr></thead><tbody><tr><td> ✗ 1 </td><td> ш. </td><td> Латиница xu dod gūxt xidow Кириллица ху дод гўхт хйдов </td></tr></tbody></table> Добавить форму	№	Пометы	Орфографические варианты	✗ 1	ш.	Латиница xu dod gūxt xidow Кириллица ху дод гўхт хйдов	<table><thead><tr><th>Ранг</th><th>Пометы</th><th>Значение</th></tr></thead><tbody><tr><td> ✗ 1 1 </td><td> ПОМЕТЫ нет F M CAUS </td><td> сильно мучиться, терзаться; страдать </td></tr></tbody></table> Добавить значение	Ранг	Пометы	Значение	✗ 1 1	ПОМЕТЫ нет F M CAUS	сильно мучиться, терзаться; страдать
№	Пометы	Орфографические варианты												
✗ 1	ш.	Латиница xu dod gūxt xidow Кириллица ху дод гўхт хйдов												
Ранг	Пометы	Значение												
✗ 1 1	ПОМЕТЫ нет F M CAUS	сильно мучиться, терзаться; страдать												

Рис. 4. Фрагмент редакторского интерфейса платформы OnLex, пример с сайта *pamiri.online* [11]

подход к регулярно повторяющимся элементам статьи позволяет минимизировать расхождения и опечатки (ср. невозможность выбрать помету CAUS для существительных, доступную лишь для глаголов), а также централизованно изменять соответствующие пометы во всех местах их употребления (например, при принятии решения о переименовании). Также стоит обратить внимание на возможность добавления неограниченного количества форм и значений (достаточно нажать на подходящую зеленую кнопку), которые впоследствии можно отсортировать с помощью

системы рангов (чем выше ранг, тем раньше элемент будет отображен на экране). Наконец, отметим поддержку в рассматриваемом словаре нескольких орфографий (шугнанский, относящийся к памирским языкам, не имеет кодифицированной системы письма) – платформа OnLex может быть настроена под нужды конкретного проекта, даже весьма специфичные.

На рис. 5 представлено то, как статья с рис. 4 отображается на сайте. Как видно, помимо соответствующих зон уже упомянутых элементов

ID: 4976	Формы и значения	Примеры	История	Создать задачу	Изменить орфографию
WORD	dod	Карамшоев	шугнанский	Ш.	
См. также		Значение	1. СУЩ. М. отец		
♦ tāt			○ Ш., <i>ām māt lūwām, ām dōd</i>		
♦ pid			'мы и "тат" (отец) говорим, и "дод"		
♦ padar			○ ВРВ., <i>ху dōd ūax, ху холāk-ат мак ўйн-та виц лўвām</i>		
♦ dodej			'сестру своего отца, жену своего дяди по матери и отцу мы называем "виц" (тетя)'		
Таксономия		Устойчивые сочетания и идиомы	Ш., <i>dōd-at nān</i>		
KINSHIP			1. отец и мать, родители		
			■ Ш., ПРИМЕТА <i>хābī-нди ар чāлак мā-ниθ, бе dōd-ат nān cāwī</i>		
			'в ночное время не садись у очага, (не то) останешься без родителей'		
		Сложные глаголы	Ш., <i>xu dōd gūxt xīdow</i>		
			1. ГЛ., сильно мучиться, терзаться; страдать		
			■ Ш., <i>иўз-ум ии қати ху dōd gūxt хўўч</i>		
			'я с ним измучился'		
			Ш., <i>dōd divīxtow</i>		
			1. ГЛ., сильно измучить, истерзать, извести		
			■ Ш., <i>ўид хўч му dōd-и му-рд divīxt</i>		
			'этот бык извел меня'		
		Оригинальный текст статьи			

Рис. 5. Статья *dod* из Шугнанско-русского словаря Д. Карамшоева, представленного на сайте *pamiri.online* [11]

редакторского интерфейса, в словаре *pamiri.online* имеется возможность делать гиперссылки на другие статьи, присваивать таксономические метки (в данном случае KINSHIP, т.е. термин родства) и добавлять примеры. Функционал платформы, однако, этим не ограничивается, и эта тема заслуживает подробного рассмотрения в отдельной статье. Сейчас же лишь упомянем то, что среди прочих доступных функций имеются: поисковый модуль (поддержка разных зон поиска, регулярных выражений, режим игнорирования диакритик, (не)полнословный поиск), система обратной связи для пользователей, система управления редакторскими задачами для команды проекта, медиаприложения (аудио, видео, изображения, карты), поддержка нескольких языков, система управления контентом и т.д.

В завершение раздела нужно еще раз подчеркнуть, что одна из главных задач СПС — минимизировать техническую нагрузку на лексикографа (в случае с интерфейсом OnLex она сводится, грубо говоря, к заполнению полей и нажатию на кнопки выпадающих меню).

Макетирование, публикация и постпубликационный период

При использовании файлового, а также табличного подходов этапы, наступающие после подготовки текста, следуют классическому сценарию: привлекается издательство, предоставляющее услуги верстальщика, редактора, корректора и

дистрибьютера выпущенной книги. Сфокусируемся на том, как эти этапы упрощаются при использовании СПС.

При наличии веб-модуля (как на платформе OnLex) весь издательский процесс сводится к нажатию кнопки «Опубликовать» в интерфейсе редактора. Несмотря на то, что достоинства бумажной книги хорошо известны, нельзя игнорировать слабые стороны этого традиционного формата, которые реже обсуждаются в русскоязычной традиции. Приведем некоторые из недостатков:

1. Поиск обычно возможен только среди заглавных слов.
2. Поиск может занимать много времени, что особенно заметно при необходимости быстро переключаться между статьями.
3. Не поддерживаются аудио- и видеоприложения.
4. Необходимо экономить место, а значит, использовать множество сокращений и часто урезать содержательную часть.
5. Во многих случаях отсутствует возможность поддерживать два языка (физические издания, эквивалентно предоставляющие информацию на нескольких языках, крайне редки).
6. Читатели не могут (или могут, но по неудобным каналам) передать свои отзывы о книге. При пользовании словарем замечания возникают не в результате постраничного чтения, а при постоянном переключении между

случайными вокабулами. Возникшая в моменте необходимость сообщить об ошибке не может быть легко удовлетворена: писать письмо про каждую статью расточительно, а собирать все ошибки в один файл больше напоминает обязанности редактора, а никак не добрую волю читателя.

7. После публикации книгу нельзя изменить (только переиздать с исправлениями).

В онлайн-СПС типа OnLex проблемы 1 и 2 полностью решены поисковым модулем: поскольку в БД каждый элемент вокабулы представлен изолированно, возможен поиск именно среди всех представителей некоторого класса (например, только среди значений или форм — отдельных таблиц в БД). Высокая скорость поиска нужных статей определяется, во-первых, возможностью перейти к упомянутой лексеме по нажатию гиперссылки, а во-вторых, тем, что поиск выполняется не читателем, а специальным алгоритмом. Недостатки 3, 4 и 5 отсутствуют в онлайн-формате. Помимо полного отказа от сокращений, можно сопоставить каждому из них «всплывающую подсказку» с расшифровкой. Что касается поддержки нескольких языков, то в онлайн-словаре имеется возможность предусмотреть параллельный ввод данных для нескольких версий вокабулы (аналогично нескольким орфографиям на рис. 4). Наконец, пункт 6 устраняется добавлением на страницу каждой словарной статьи поля «Сообщить об ошибке», нажав на которое пользователь может моментально отправить замечание, касающееся конкретной вокабулы. Редакторы сразу получают это сообщение и могут внести исправления, опубликовав новую версию нажатием одной кнопки, тем самым разрешая проблему 7.

Если же издание книги необходимо, то большинство СПС предоставляют возможность экспорта словарных статей из БД по заданным правилам композиции и оформления (например: заглавное слово полужирным, пометы курсивом, значения на новой строке и т.п.).

Заключение

В статье были описаны методы и принципы цифровой лексикографии. В частности, были введены понятия файлового и табличного подходов к составлению словарей, описаны лексикографические форматы TEI, основные особенности баз данных, а также рассмотрены функции и цели систем подготовки словарей, проиллюстрированные платформой OnLex. Было показано, чем полезно наличие онлайн-версии, во многих аспектах представляющей собой более удобный способ

представления словаря — и для пользователя, и для составителя. Преимущества использования систем подготовки словарей не должны игнорироваться лексикографическими проектами нашего времени, и хочется верить, что в ближайшие годы в русскоязычной словарной традиции частота использования продвинутых методов цифровой лексикографии будет стремительно возрастать.

СПИСОК ЛИТЕРАТУРЫ

1. *Беляев О.И., Макаров Ю.Ю., Новокшанов Д.А., Сеницына Ю.В., Хомченкова И.А.* Онлайн-словари иранских языков // 1-ая Международная научно-образовательная конференция «Пейсиковские чтения: проблемы современного академического востоковедения»: материалы конференции / ред. А. А. Маслов. М.: ИСАА МГУ имени М.В. Ломоносова, 2023. С. 7–11. [Электронный ресурс]: <https://elibrary.ru/item.asp?id=58073241&pff=1>
2. *Беляев О.И., Хомченкова И.А., Сеницына Ю.В., Дьячков В.В., Бызова А.А., Бадеев А.О., Алексеев Д.А., Макаров Ю.Ю.* Историко-этимологический словарь осетинского языка В.И. Абаева: проблемы создания цифровой двуязычной версии // Вестн. Моск. ун-та. Серия 9. Филология. 2024. Вып. 2. С. 75–86. <http://dx.doi.org/10.55959/MSU0130-0075-9-2024-47-02-4>
3. *Dragičević R., Makarov Y., Ryzhova D., Shapich Y., Yakushkina E.* A new bilingual Serbian–Russian dictionary // *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress* / ред. K. Despot, I. Brač, A. Ostroški Anić. Zagreb: Institute for the Croatian Language, 2024. Pp. 93–100.
4. *Плунгян В.А.* Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // *Русский язык в научном освещении*. 2008. Т. 16. Вып. 2. С. 7–20.
5. *Беликов В.И., Копылов Н.Ю., Пиперски А.Ч., Сеlegeй В.П., Шаров С.А.* Корпус как язык: от масштабируемости к дифференциальной полноте // *Компьютерная лингвистика и интеллектуальные технологии*. 2013. Т. 1. Вып. 12. С. 19.
6. *Piperski A., Belikov V., Kopylov N., Morozov E., Selegey V., Monakhov S.* Big and diverse is beautiful: A large corpus of Russian to study linguistic variation // *Proceedings of the 8th Web as Corpus Workshop (WAC-8) @ Corpus Linguistics 2013* / ред. S. Evert, E. Stemle, P. Rayson. 2013. С. 24–28.
7. *Magomedgazhieva P., Daniel M.* Dictionary of Tukita (v2.0.0). Linguistic Convergence Laboratory, HSE University, Moscow, 2023. <https://doi.org/10.5281/zenodo.7803955>
8. *Belyaev O., Khomchenkova I., Sinitsyna J., Dyachkov V.* Digitizing print dictionaries using TEI: The Abaev Dictionary Project // *Proceedings of the Seventh International Workshop on Computational Linguistics*

- of Uralic Languages, Syktyvkar, Russia (Online): Association for Computational Linguistics. Сен. 2021. С. 57–64. [Электронный ресурс]: <https://aclanthology.org/2021.iwclul-1.7>
9. Иванов В.Б. Большой персидско-русский словарь. Т. 1. М.: Наука, 2020.
 10. Abel A. Dictionary writing systems and beyond // Electronic Lexicography / ред. S. Granger, M. Paquot. Oxford University Press, 2012. С. 83–106. <https://doi.org/10.1093/acprof:oso/9780199654864.003.0005>
 11. Makarov Y., Melenchenko M., Novokshanov D. Digital Resources for the Shughni Language // Proceedings of The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association. Июнь 2022. С. 61–64. [Электронный ресурс]: <https://aclanthology.org/2022.eurali-1.9>
 12. Иванов В.Б. Большой персидско-русский словарь. Т. 2. М.: Фонд Ибн Сины, 2023.
 13. Иванов В.Б. Большой персидско-русский словарь. Т. 3. М.: ООО «Садра», 2024.
 14. Крысин Л.П. (отв. ред.) Академический толковый словарь русского языка. Т. 1: А – ВИЛЯТЬ. М.: Издательский дом ЯСК, 2016.
 15. Крысин Л.П. (отв. ред.) Академический толковый словарь русского языка. Т. 2: ВИНА – ГЯУР. М.: Издательский дом ЯСК, 2016.
 16. Цумарев А.Э., Шестакова Л.Л., Нечаева И.В., Кулева А.С., Грунченко О.М. «Академический толковый словарь русского языка»: традиционное и новое // Известия Российской академии наук. Серия литературы и языка. 2017. Т. 76. № 5. С. 5–21.
 3. Dragičević, R., Makarov, Y., Ryzhova, D., Shapich, Y., Yakushkina, E. A new bilingual Serbian–Russian dictionary. (Eds.) K. Despot, I. Brač, A. Ostroški Anić. Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress. Zagreb: Institute for the Croatian Language, 2024, pp. 93–100.
 4. Plungian, V. A. *Korpus kak instrument i kak ideologija: o nekotorykh urokakh sovremennoj korpusnoj lingvistiki* [A Corpus as a Research Tool and Ideology: Some Lessons from Modern Corpus Linguistics]. *Russkij jazyk v nauchnom osveshchenii* [Russian Language and Linguistic Theory]. 2008, No. 16(2), pp. 7–20. (In Russ.)
 5. Belikov, V.I., Kopylov, N.Ju., Piperski, A.Ch., Selegey, V.P., Sharoff, S.A. *Korpus kak yazyk: ot masshtabiruemosti k differencialnoj polnote* [Corpus as Language: From Scalability to Register Variation]. *Kompiuternaia lingvistika i intellektualnye tekhnologii* [Computational Linguistics and Intelligent Technologies]. 2013, No. 12(1), p. 19. (In Russ.)
 6. Piperski, A., Belikov, V., Kopylov, N., Morozov, E., Selegey, V., Monakhov, S. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. (Eds.) S. Evert, E. Stemle, P. Rayson. Proceedings of the 8th Web as Corpus Workshop (WAC-8) @ Corpus Linguistics 2013. 2013. P. 24–28.
 7. Magomedgazhieva, P., Daniel, M. Dictionary of Tukita (v2.0.0). Linguistic Convergence Laboratory, HSE University, Moscow, 2023. <https://doi.org/10.5281/zenodo.7803955>
 8. Belyaev, O., Khomchenkova, I., Sinitsyna, J., Djachkov, V. Digitizing print dictionaries using TEI: The Abaev Dictionary Project. Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages, Syktyvkar, Russia (Online): Association for Computational Linguistics. 2021. P. 57–64. URL: <https://aclanthology.org/2021.iwclul-1.7>
 9. Ivanov, V.B. *Bolshoj persidsko-russkij slovar* [Persian-Russian Dictionary]. Vol. 1. Moscow: Nauka Publ., 2020. (In Russ.)
 10. Abel, A. Dictionary writing systems and beyond. Electronic Lexicography. (Eds.) S. Granger, M. Paquot. Oxford University Press, 2012. P. 83–106. <https://doi.org/10.1093/acprof:oso/9780199654864.003.0005>
 11. Makarov, Y., Melenchenko, M., Novokshanov, D. Digital Resources for the Shughni Language. Proceedings of The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association, 2022. P. 61–64. URL: <https://aclanthology.org/2022.eurali-1.9>

REFERENCES

1. Belyaev, O.I., Makarov, Y., Novokshanov, D.A., Sinitsyna, Ju.V., Khomchenkova, I.A. *Onlajn-slovari iranskikh jazykov* [Online Dictionaries of Iranian Languages]. *1-aja Mezhdunarodnaja nauchno-obrazovatel'naja konferentsija "Pejsikovskie chtenija: problemy sovremennogo akademicheskogo vostokovedenija": materialy konferentsii* [1st International Scientific and Educational Conference "Peisikov Readings: Problems of Modern Academic Oriental Studies": Conference Materials]. Ed. A.A. Maslov. Moscow: ISAA MGU imeni M.V. Lomonosova Publ., 2023, pp. 7–11. URL: <https://elibrary.ru/item.asp?id=58073241&pf=1> (In Russ.)
2. Belyaev, O.I., Khomchenkova, I.A., Sinitsyna, J.V., Dyachkov, V.V., Byzova, A.A., Badeev, A.O., Alekseev, D.A., Makarov, Y. *Istoriko-etimologicheskij slovar osetinskogo jazyka V.I. Abaeva: problemy sozdanija tsifrovaj dvujazyčnoj versii* [V.I. Abaev's Historical-Etymological Dictionary: Issues in the Development of

12. Ivanov, V.B. *Bolshoj persidsko-russkij slovar* [Persian-Russian Dictionary]. Vol. 2. Moscow: Fond Ibn Siny Publ., 2023. (In Russ.)
13. Ivanov, V.B. *Bolshoj persidsko-russkij slovar* [Persian-Russian Dictionary]. Vol. 3. Moscow: ООО “Sadra” Publ., 2024. (In Russ.)
14. Krysin, L.P. (ed.) *Akademicheskij tolkovyj slovar russkogo jazyka. Tom 1: A – VILIAT’* [Academic Explanatory Dictionary of Russian. Vol. 1]. Moscow: Izdatelskij dom IASK Publ., 2016. (In Russ.)
15. Krysin, L.P. (ed.) *Akademicheskij tolkovyj slovar russkogo jazyka. Tom 2: VINA – GIAUR* [Academic Explanatory dictionary of Russian. Vol. 2]. Moscow: Izdatelskij dom IASK Publ., 2016. (In Russ.)
16. Tsumarev, A.E., Shestakova, L.L., Nechaeva, I.V., Kuleva, A.S., Grunchenko, O.M. “*Akademicheskij tolkovyj slovar russkogo jazyka*”: *traditsionnoe i novoe* [“Academic Explanatory Dictionary of the Russian Language”: the Traditional and the New]. *Izvestiâ Rossijskoj akademii nauk. Seriâ literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2017, Vol. 76, No. 5, pp. 5–21. (In Russ.)

Дата поступления материала в редакцию: 7 мая 2024 г.

Статья поступила после рецензирования и доработки: 5 июня 2024 г.

Статья принята к публикации: 15 июня 2024 г.

Дата публикации: 31 августа 2024 г.

Received by Editor on May 7, 2024

Revised on June 5, 2024

Accepted on June 15, 2024

Date of publication: August 31, 2024