DOI 10.37882/2223-2966.2026.02.15

ВЫЧИСЛИТЕЛЬНЫЕ ПОДХОДЫ К АВТОМАТИЗАЦИИ НАПОЛНЕНИЯ МОРФОЛОГИЧЕСКОГО СЛОВАРЯ ДЛЯ ЦЕРКОВНОСЛАВЯНСКОГО ЯЗЫКА

COMPUTATIONAL APPROACHES TO AUTOMATING THE FILLING OF A MORPHOLOGICAL DICTIONARY FOR THE CHURCH SLAVONIC LANGUAGE

M. Klychkov

Summary. This article examines computational approaches to the automated compilation of a morphological dictionary for the Church Slavonic language. The relevance of this topic is due to the insufficient representation of Old Church Slavonic texts (11th–17th centuries) in existing linguistic corpora and the lack of access to full texts, which hinders comprehensive linguistic research. The creation of a specialized corpus of manuscript heritage is proposed, which will allow for the systematization and digitization of Church Slavonic manuscripts and facilitate rapid search and analysis of the morphological and syntactic features of the Old Russian language. The research focus is on the technology of constructing a corpus of handwritten texts, and the development of models and algorithms for the automated generation of a morphological dictionary. The presented methods contribute to the expansion of the capabilities of digital linguistic resources and the development of corpus linguistics in the field of historical Slavic languages.

Keywords: corpus linguistics, Church Slavonic, morphological dictionary, computational methods, linguistic corpus, natural language processing.

Клычков Матвей Дмитриевич
инженер, аспирант, Национальный исследовательский ядерный университет МИФИ
utherluminous@gmail.com

Аннотация. В статье рассматриваются вычислительные подходы к автоматизированному наполнению морфологического словаря для церковнославянского языка. Актуальность темы обусловлена недостаточной представленностью старославянских текстов (XI–XVII вв.) в существующих лингвистических корпусах и отсутствием доступа к полным текстам, что затрудняет проведение комплексных языковых исследований. Предлагается создание специализированного корпуса рукописного наследия, что позволит систематизировать и оцифровать церковнославянские рукописи, обеспечить быстрый поиск и анализ морфологических и синтаксических особенностей древнерусского языка. В качестве объекта исследования выбрана технология построения корпуса рукописных текстов, а предметом — разработка моделей и алгоритмов для автоматизированного формирования морфологического словаря. Представленные методы способствуют расширению возможностей цифровых лингвистических ресурсов и развитию корпусной лингвистики в области исторических славянских языков.

Ключевые слова: корпусная лингвистика, церковнославянский язык, морфологический словарь, вычислительные методы, лингвистический корпус, обработка естественного языка.

Введение

Распознавание древних текстов представляет собой сложную задачу, включающую два ключевых этапа: сегментацию (выделение отдельных символов из изображения текста) и классификацию (определение соответствия каждого символа определённой букве алфавита). Для улучшения качества работы классификации существует необходимость проводить статистическую коррекцию ее результатов. Данный подход может позволить существенно улучшить качество распознавания посредством динамического обучения классификатора.

Создание эффективного морфологического словаря представляет собой основу для качественной статистической коррекции результатов классификации в системах распознавания древних текстов. Современные подходы к построению таких словарей можно разделить на несколько основных направлений, каждое из которых имеет свои преимущества и области применения.

Одним из наиболее распространённых методов является автоматическое извлечение морфологической информации из текстовых корпусов. Этот подход основан на анализе больших массивов текстов для выявления морфологических закономерностей и создания словарных структур [1, 2]. Исследования показывают, что создание частотных словарей на основе литературных произведений позволяет эффективно анализировать структурные и семантические аспекты языка [1]. Лексикографический метод обеспечивает детальное изучение объектов лексикографирования и их функционирования в языке [3,4]. Методы корпусной лингвистики используют специализированные системы для анализа композиционной структуры текстов и извлечения многокомпонентных терминов [5], что особенно важно для древних текстов с их специфической лексикой.

Второе направление включает статистические и машинно-обучаемые подходы к генерации морфологических словарей. Методы неконтролируемого обучения морфологии позволяют автоматически создавать словари и базы морфологических правил из текстовых корпусов без предварительной разметки [6,7]. Статистический подход с использованием парадигм эксплуатирует понятие морфологических парадигм — наборы морфологических категорий, которые могут применяться к однородным множествам слов [7]. Современные исследования демонстрируют эффективность гибридных подходов, сочетающих глубокое обучение с высокоточными морфологическими словарями на этапе вывода [8,9]. Такие методы обеспечивают 50 % снижение ошибок в лематизации и 58 % снижение ошибок в морфологическом теггинге по сравнению с традиционными подходами [8].

Третий подход базируется на ручном создании и экспертных системах. Лексикографические методы предполагают моделирование структуры словарной статьи и создание программных лексикографических продуктов с использованием специализированных языков программирования [5,10]. Исследования в области древнерусских текстов показывают важность создания фреймовых структур словников, содержащих слова в виде ссылок на фреймы букв из соответствующих частей базы знаний [11]. Принципы лексикографического описания включают определение основных принципов при опосредованном переводе и применение методов эквивалентного перевода в процессе обработки словарных статей [12]. Подобные подходы требуют значительных экспертных знаний, но обеспечивают высокую точность для специализированных доменов [13, 10].

Четвёртое направление представлено современными нейросетевыми подходами. Рекурсивные нейронные сети способны создавать представления морфологически сложных слов из их морфем, эффективно обрабатывая редкие и сложные слова [14]. Трансформерные модели демонстрируют высокую эффективность в морфологическом анализе, особенно при использовании контекстуальных эмбеддингов [8,9]. Исследования показывают, что модели на основе векторного квантования (VQ-VAE) с множественными кодовыми книгами обеспечивают интерпретируемое неконтролируемое морфологическое обучение, сравнимое по производительности с контролируемыми моделями [15]. Статистическое машинное обучение для генерации морфологии позволяет создавать модели, способные предсказывать словоформы с точностью до 98,9 % при наличии полных морфологических признаков [16,17]. Графовые полуконтролируемые методы обучения используют морфологические, синтаксические и семантические отношения между словами для автоматического построения широкопокрывающих лексиконов из небольших начальных наборов [18].

Пятое направление включает интегрированные методологии, объединяющие преимущества различных подходов. Гибридные системы сочетают правиловые методы с машинным обучением для достижения высокой точности при сохранении интерпретируемости результатов [19, 20]. Многоуровневые архитектуры используют комбинацию морфологических HMM с контекстно-чувствительными правилами переписывания для индукции базовой морфологии [19]. Современные исследования показывают эффективность систем с множественными кодовыми книгами, где каждая кодовая книга специализируется на различных морфологических признаках, что способствует интерпретируемости модели [15]. Методы автоматического извлечения терминологии с использованием глубокого обучения и статистического анализа обеспечивают создание специализированных словарей для конкретных предметных областей [21, 22, 23]. Подобные подходы особенно перспективны для древних текстов, где требуется сочетание лингвистической экспертизы с вычислительными методами.

Материалы и методы исследования

Выполненные работы включали в себя исследовательскую и инженерную части.

В исследовательскую часть входила задача определить современные методы по обогащению словарей новыми словоформами. Для этого были проанализированы современные подходы, описанные ранее. Учитывая имеющийся набор данных в виде небольшого количества размеченных текстов и словаря расширяемого и обогащаемого филологами, было выдвинуто решение сконцентрироваться на гибридном dictionary-based подходе.

В качестве основного инструмента морфологического анализа был выбран pymorphy2. Данный выбор, во-первых, обусловлен тем, что pymorphy2 использует словари OpenCorpora, что обеспечивает возможность к приведению словаря slavcorpora к этому формату. Во-вторых, библиотека предоставляет эвристические алгоритмы для анализа неизвестных слов, что критически важно при работе с постоянно пополняемым словарем. В-третьих, pymorphy2 обладает возможностями генерации полных парадигм склонения, что позволяет автоматически получать все словоформы для новых лексем, добавляемых филологами. Открытый исходный код библиотеки обеспечивает возможность модификации и адаптации под старославянский язык.

Учитывая факт того, что количество установленных образцов словоизменения (парадигм) для старославянского языка может быть недостаточным для качественной работы pymorphy2, представляется целесообразным использование современных много-

язычных инструментов для автоматического анализа дополнительных текстов и последующего обогащения словаря. В контексте pymorphy2 парадигма представляет собой полный набор словоформ одной лексемы — например, все падежные формы существительного или все временные формы глагола, объединенные общей основой и грамматическими характеристиками [24].

Stanza как решение для предварительного анализа представляет особый интерес благодаря своей языковоагностической архитектуре [25]. Инструмент показывает конкурентоспособную производительность и использует полностью нейронный пайплайн, который включает лемматизацию, морфологический анализ частей речи и универсальных морфологических признаков. Ключевой особенностью является способность системы обрабатывать сырой текст без предварительной токенизации, что очень важно при работе с историческими текстами старославянского языка.

Также внимания заслуживает использование триграммной модели для морфологической дисамбигуации поверх результатов pymorphy2[26]. Как показывают исследования, точность определения грамматического падежа с учетом контекста возрастает с 82 % до 94 % при использовании триграммной модели. Модель анализирует последовательность из трех морфологических тегов для выбора наиболее вероятной интерпретации каждого слова в предложении.

Интеграция этих технологий позволит создать многоуровневую систему обогащения словаря: первичный анализ дополнительных текстов с помощью Stanza/UDPipe, извлечение новых парадигм словоизменения, их валидация и последующая интеграция в словарь pymorphy2.

Были проанализированы и проведены эксперименты со следующими технологиями:

1. Pymorphy2
2. Stanza
3. UDPipe

Pymorphy2 — морфологический анализатор, реализованный на Python для русского языка. Анализатор принимает слово и на основе его морфологии производит серию гипотез классификации относительно части речи, рода, числа, падежа и других характеристик [24].

Данная технология может представлять собой интерес не только для готовых словарей, но и для тех, которые требуют расширения, так как особое внимание в данной технологии уделяется обработке слов, отсутствующих в словаре.

Автор данной технологии разработал комплексную систему правил для анализа неизвестных слов:

1. Удаление общих префиксов
2. Анализ слов, заканчивающихся на другие словарные слова
3. Сопоставление окончаний
4. Обработка слов с дефисами
5. Специальные правила для различных типов токенов

Это решение обусловлено пониманием того, что простое увеличение размера словаря не решает проблему эффективно из-за закона Ципфа и связанного с ним закона убывающей отдачи.

Автор статьи вводит такое понятие как парадигма слова, согласно статье [1] — это модель флексии лексемы. Она состоит из триплетов (prefix, suffix, tag) для каждой словоформы в лексеме, где каждая словоформа i может быть представлена как: prefix + stem + suffix, при этом основа (stem) остается одинаковой для всех слов в лексеме.

```
ёж      NOUN,anim,masc sing,nomn
ежа     NOUN,anim,masc sing,gent
ежу     NOUN,anim,masc sing,datv
...
ежами   NOUN,anim,masc plur,ablt
ежах    NOUN,anim,masc plur,loct
```

Рис 1. Пример лексемы [24]

Основываясь на статье о pymorphy [24], можно сделать следующее предположение: имея достаточное количество «парадигм» слов, мы можем получить хорошее качество работы анализатора даже на неполном наборе данных. Под набором данных понимаются леммы и соответствующие им словоформы.

Для того чтобы апробировать данную гипотезу был взят словарь OpenCorpora(тут какая то ссылка тоже), проанализирован и реализован следующий алгоритм стратифицированного деления словаря по морфологическим признакам.

Математическое описание алгоритма стратификации

1. Исходные данные

Пусть $L = l_1, l_2, ..., l_n$ — множество всех лемм в словаре OpenCorpora, где $|L| = N$.

Каждая лемма l_i характеризуется:

- $POS(l_i)$ — часть речи леммы
- $P(l_i)$ — морфологическая парадигма леммы

2. Первый уровень стратификации: по частям речи

Разбиваем словарь на страты по частям речи:

$$L = U_{j=1}^{m} L_j, \text{ где } L_j = l_i \in L : POS(l_i) = POS_j$$

$$|L_j| = n_j, \sum_{j=1}^{m} n_j = N$$

3. Второй уровень стратификации: по парадигмам внутри частей речи

Внутри каждой части речи разбиваем на подстраты по парадигмам:

$$L_j = U_{k=1}^{Kj} L_{jk}, \text{ где } L_{jk} = l_i \in L_j : P(l_i) = P_{jk}$$

$$|L_{jk}| = n_{jk}, \sum n_{jk} = n_j$$

Формулы для создания подмножества размером p%

4. Целевой размер подмножества

$$S = \left\lfloor N \times \frac{p}{100} \right\rfloor$$

5. Размер выборки из каждой части речи

Пропорциональное распределение по частям речи:

$$s_j = max\left(1, \left\lfloor S \times \frac{n_j}{N} \right\rfloor\right)$$

где: s_j — количество лемм для выборки из части речи j max (1, ...) гарантирует минимум 1 лемму из каждой части речи.

6. Размер выборки из каждой парадигмы

Внутри каждой части речи j распределяем пропорционально по парадигмам:

$$s_{jk} = max\left(1, \left\lfloor s_j \times \frac{n_{jk}}{n_j} \right\rfloor\right)$$

где: s_{jk} — количество лемм для выборки из парадигмы k части речи j

7. Корректировка размеров

Поскольку округления могут дать неточный размер, применяем корректировку:

$s'_{jk} = min(s_{jk}, n_{jk})$ (не можем взять больше, чем есть)

Если $\sum s'_{jk} < s_j$, добираем случайно из оставшихся лемм части речи j

Если $\sum\sum s'_{jk} > S$, обрезаем случайно до размера S

Свойства у данного алгоритма будут следующие:

Доля части речи в подмножестве ≈ Доля части речи в исходном словаре:

$$P(POS = j \vee lemma \in Subarr) \approx$$
$$\approx P(POS = j \vee lemma \in L)$$

Каждая продуктивная парадигма представлена минимум 1 леммой:

$$\forall j,k : n_{jk} > 0 \rightarrow s'_{jk} \geq 1$$

Внутри каждой страты выборка случайна, что обеспечивает несмещенность:

E[качество_анализа|Subsarr]=E[качество_анализа|L]

Этот алгоритм гарантирует, что морфологическая структура подмножества максимально близка к структуре исходного словаря, что критически важно для объективной оценки влияния размера словаря на качество морфологического анализа.

Результаты и обсуждения

Таким образом словарь opencorpora был разделен на 9 частей в соотношении 10 %, 20 %, 30 % …, 90 % от объема словаря. Для каждого подмножества словаря была составлена выборка из 100,000 слов из всего словаря из тех слов, которые не попадали в подмножества.

Были получены следующие результаты представленные на рис 2:

Исходя из графика с метриками точности, можно сделать вывод, что для составленного набора данных под каждое стратифицированное подмножество словаря качество его морфологического анализа по мере расширения растет, но темпы прироста — небольшие. Общий процент точного определения грамматических характеристик находится в пределах от 89% до 97% в соответствии с распределением.

Исходя из этого гипотеза может считаться успешно апробированой в рамках данного эксперимента.

Благодаря данному заключению можно сделать вывод, что нам действительно подходит данная технология. В связи с чем можно спроектировать систему наполнения словаря следующим образом (Рис. 3).

Концепт решения представляет собой комплексную систему для автоматизированного анализа древнерусских текстов и интерактивного расширения морфологи-

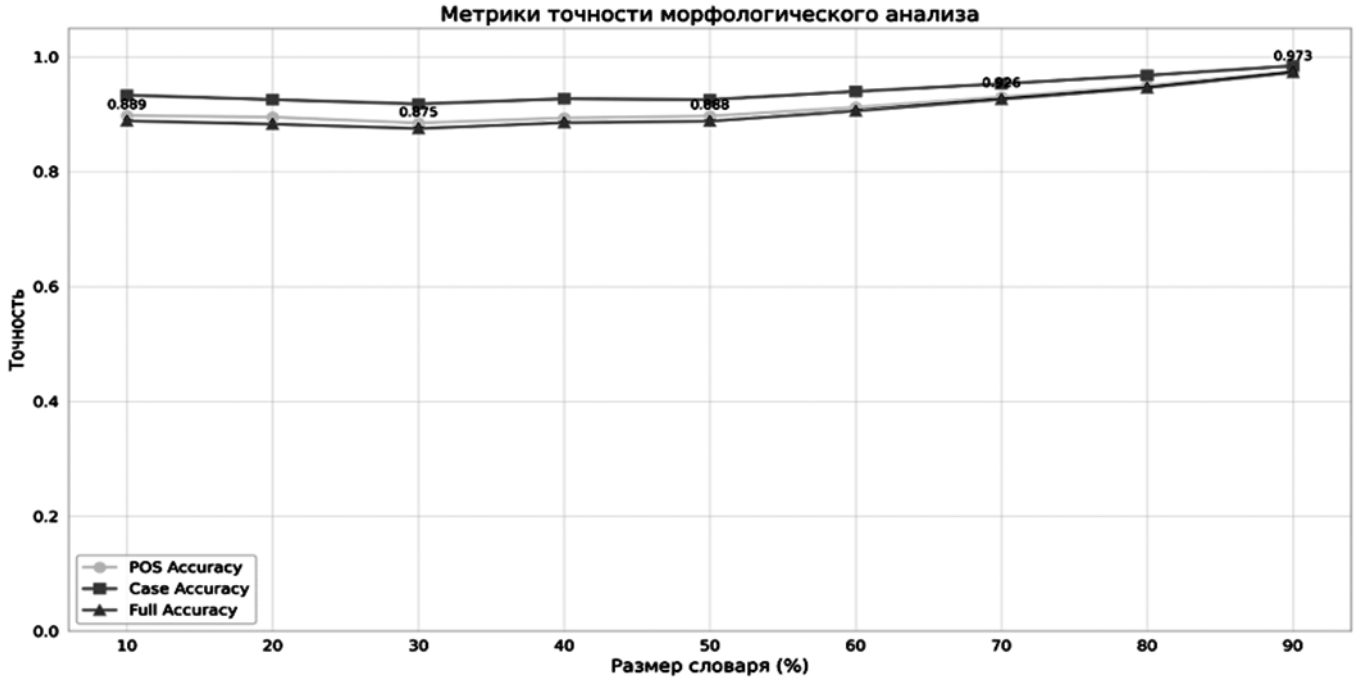

Рис. 2. Метрики точности морфологического анализа (POS: Доля слов, для которых часть речи определена правильно, Case: Доля слов, для которых падеж определен правильно, Full: Доля слов, для которых все основные характеристики определены правильно одновременно)

ческого словаря PyMorphy2 с участием экспертов-филологов.

Система реализует непрерывный цикл улучшения: каждое валидированное экспертом слово немедленно включается в словарь, что повышает качество анализа последующих неизвестных слов. Этот подход обеспечивает адаптивное развитие морфологического анализатора.

Данное решение обеспечивает эффективный и научно обоснованный подход к расширению морфологических ресурсов для древнерусского языка, сочетая возможности современных NLP-технологий с экспертизой специалистов-филологов.

На текущий момент для реализации данного концепта проделаны следующие этапы работы:

1. Реализованы программные средства, переводящие словарь slavcorpora в формат OpenCorpora для последующей компиляции в словарь pymorphy.
2. Написано расширение для пайморфи для старославянского языка, которое учитывает грамматические особенности данного языка (добавлена форма глагола супин, двойственное число, новые времена итд.). Расширен анализатор для старославянского языка.
3. Реализован прототип веб интерфейса для филологов. А именно: все шаблоны для частей речи, шаблон для предзаполненных грамматических характеристик, редактор, включающий в себя раскладку со старославянской клавиатурой.

Прототип интерфейса для филологов предславлен на Рис. 4.

Заключение

В ходе проведенного исследования была разработана и реализована комплексная методология изучения влияния размера словаря на качество морфологического анализа в системе pymorphy2 на материале словаря OpenCorpora.

Спроектирована комплексная система для расширения и адаптации морфологического анализатора PyMorphy2 под задачи обработки древнерусских текстов. Проделанная работа охватывает широкий спектр технических и лингвистических аспектов, от низкоуровневой настройки морфологического движка до создания пользовательского интерфейса для экспертной валидации.

Полученные результаты и разработанные методики закладывают фундамент для создания комплексных систем распознавания и анализа древних рукописных текстов, что имеет важное значение для сохранения культурного наследия и научных исследований.

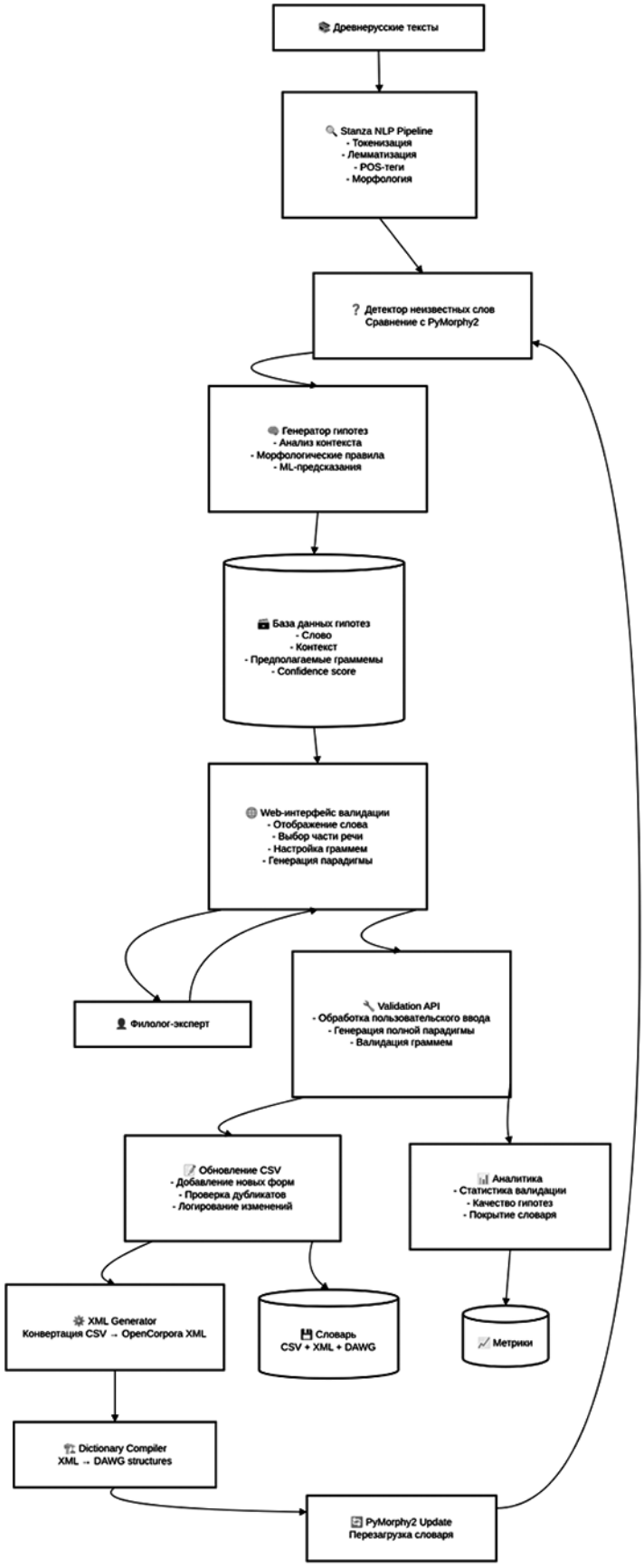

Рис. 3. Система интерактивного расширения морфологического словаря для древнерусского язык

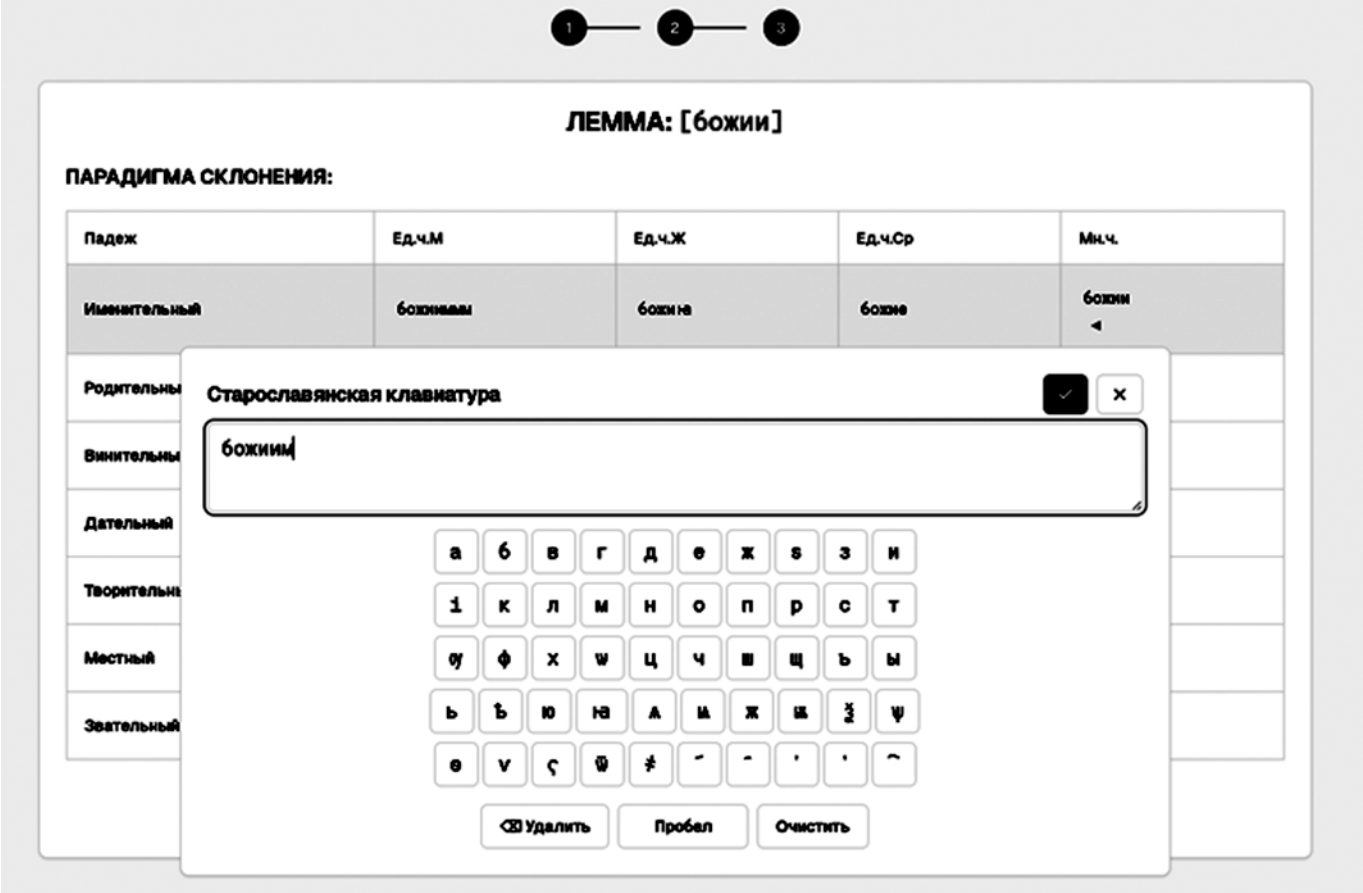

Рис. 4. Прототип интерфейса для филологов

ЛИТЕРАТУРА

1. Жанабекова А., Пірманова К. Частотный словарь языка Абая Кунанбайулы: разработка методологии и ключевые результаты. Вестник КазНПУ имени Абая. Серия: Филологические науки, 2023. Т. 85, № 3. С. 62–70.
2. Cole M., Rayson P., Mariani J. LexiDB: Patterns & Methods for Corpus Linguistic Database Management. // Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille. 11-16 May 2020. С. 3128–3134.
3. Архив журнала Вестник ТГПУ. [Электронный ресурс] // URL: http://vestnik.tspu.edu.ru/archive.html?year=2021&issue=3&article_id=8061 (дата обращения 15.01.2026).
4. Архив журнала Вестник ТГПУ. [Электронный ресурс] // URL: http://vestnik.tspu.edu.ru/archive.html?year=2020&issue=4&article_id=7786 (дата обращения 15.01.2026).
5. Бутенко Ю.И., Солошенко К.А. Лексический тренажер по иностранному языку для студентов технических специальностей МГТУ им. Н.Э. Баумана. // Экономика. Информатика. 2024. Т. 51, № 1. С. 189–200.
6. Portal ACM. [Электронный ресурс] // URL: http://portal.acm.org/citation.cfm?doid=1118647.1118648 (дата обращения 15.01.2026).
7. Unsupervised morphology learning with statistical paradigms. Papers with Code. [Электронный ресурс] // URL: https://paperswithcode.com/paper/unsupervised-morphology-learning-with (дата обращения 15.01.2026).
8. Straka M., Straková J. Open-Source Web Service with Morphological Dictionary–Supplemented Deep Learning for Morphosyntactic Analysis of Czech //International Conference on Text, Speech, and Dialogue. Cham: Springer Nature Switzerland, 2024. С. 279–290.
9. Supriya M. et al. Developing a Hybrid Morphological Analyzer for Low-Resource Languages //Applied Sciences. 2025. Т. 15. №. 10. С. 5682.
10. Беляева Л.Н. Потенциал автоматизированной лексикографии и прикладная лингвистика //Известия Российского государственного педагогического университета им. АИ Герцена. — 2010. — №. 134. С. 70–79.
11. Зеленцов И.А. Автоматизированная система распознавания древнерусских скорописных текстов. Дипломный проект. // М.: 2008. 36 с.
12. Филология БГУ. [Электронный ресурс] // URL: http://journals.bsu.ru/journals/filologiya/?issue=363&article=3659&rus (дата обращения 15.01.2026).
13. Pentheroudakis J., Vanderwende L. Automatically Identifying Morphological Relations in= Machine-Readable Dictionaries //arXiv preprint cmp-lg/9411014. 1994. С. 1–19.
14. Luong M.T., Socher R., Manning C.D. Better word representations with recursive neural networks for morphology //Proceedings of the seventeenth conference on computational natural language learning. 2013. С. 104–113.

15. El Kholy A., Habash N. Rich morphology generation using statistical machine translation //INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference. 2012. С. 90–94.
16. Dušek O., Jurcícek F. Robust multilingual statistical morphological generation models //51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop. 2013. С. 158–164.
17. Faruqui M., McDonald R., Soricut R. Morpho-syntactic Lexicon Generation Using Graph-based Semi-supervised Learning. //Transactions of the Association for Computational Linguistics. 2016. Vol. 4. С. 1–16.
18. Kawtrakul A., Thumkanon C.A statistical approach to thai morphological analyzer //Fifth Workshop on Very Large Corpora. 1997. С. 290–296.
19. Lee S. Mastering Morphological Analysis. Number Analytics Blog, May 27, 2025. [Электронный ресурс] // URL: https://www.numberanalytics.com/blog/mastering-morphological-analysis (дата обращения 15.01.2026).
20. Rozhkov Y. Linguocognitive Approach to Extracting Terms from a Corpus of Veterinary Texts //International Journal of Philology. — 2023. — Т. 4. — №. 27. — С. 46–55.
21. IEEE Conference Publication. [Электронный ресурс] // URL: https://ieeexplore.ieee.org/document/10286646(дата обращения 15.01.2026).
22. Monakhov S., Turchanenko V., Fedyukova E., Cherdakov D. New method of automated terminology extraction: case study of russian-language textbooks // Proceedings of the Future Technologies Conference. Cham: Springer International Publishing, 2021. С. 363–373
23. Smith N. Morphological analysis of historical languages //Bulletin of the Institute of Classical Studies. 2016. Т. 59. №. 2. С. 89–102.
24. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages //International conference on analysis of images, social networks, and texts. Cham: Springer International Publishing, 2015. С. 320–332.
25. Peng Qi* Yuhao Zhang* Yuhui Zhang Jason Bolton Christopher D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, Stanford University Stanford, CA 94305 [Электронный ресурс] // URL: https://aclanthology.org/2020.acl-demos.14.pdf(дата обращения 15.01.2026).
26. Vissio N.C., Zakharov V.A Disambiguator for Pymorphy2 Morphological Analyzer //IMS. 2021. С. 81–88.

Журнал «Современная наука: актуальные проблемы теории и практики»