

УДК 004.89

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПОПОЛНЕНИЯ СЛОВАРНОГО ЗАПАСА НА БАЗЕ ИНТЕГРАЦИИ СИСТЕМ ОБРАБОТКИ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

Павлов Михаил Дмитриевич, студент, направление подготовки 09.03.03 Прикладная информатика, Российский государственный гуманитарный университет, Москва, Россия
e-mail: pmd.fizicon@gmail.com

Научный руководитель: **Султанов Наиль Закиевич**, доктор технических наук, профессор, профессор кафедры информационных технологий и систем, Российский государственный гуманитарный университет, Москва, Россия
e-mail: sultanovnz@mail.ru

Аннотация. В условиях роста популярности цифровых инструментов для изучения иностранных языков сохраняется проблема высоких трудозатрат на ручное создание карточек для интервального повторения (SRS). В статье представлена автоматизированная система, интегрирующая методы обработки естественного языка (NLP) для генерации грамматических форм, лемматизации, определения части речи и автоматического перевода. Цель работы – снижение ручного труда учащихся и преподавателей при подготовке карточек. Используются микросервисная архитектура (Next.js, PostgreSQL, Docker), библиотека spaCy с UniMorph и LibreTranslate. Результатом является веб-приложение, которое по одному введённому слову формирует полноценную карточку с морфологическими формами, переводом и настраиваемым алгоритмом SRS. Новизна подхода заключается в комбинации морфологической генерации, семантической дизамбигуации и интеграции с SRS в едином конвейере. Практическая значимость состоит в сокращении времени подготовки карточек в 3–5 раз и повышении качества лексического материала. Дальнейшие исследования направлены на расширение языковой поддержки и внедрение адаптивных кривых забывания.

Ключевые слова: обработка естественного языка, интервальное повторение, SRS, автоматизация изучения лексики, spaCy, LibreTranslate, морфологический анализ, Next.js, Docker, карточки для запоминания.

Благодарности. Автор выражает благодарность научному руководителю д.т.н., профессору Султанову Н.З. за ценные рекомендации при проектировании архитектуры системы.

Для цитирования: Павлов М. Д. Автоматизация процесса пополнения словарного запаса на базе интеграции систем обработки естественных языков // Шаг в науку. – 2026. – № 3. – С. 15–18.

AUTOMATION OF VOCABULARY REPLENISHMENT BASED ON INTEGRATION OF NATURAL LANGUAGE PROCESSING SYSTEMS

Pavlov Mikhail Dmitrievich, student, training program 09.03.03 Applied Informatics, Russian state university for the humanities, Moscow, Russia
e-mail: pmd.fizicon@gmail.com

Research advisor: **Sultanov Nail Zakievich**, Doctor of Technical Sciences, Professor, Professor of the Department of Information Technologies and Systems, Russian state university for the humanities, Moscow, Russia
e-mail: sultanovnz@mail.ru

Abstract. With the growing popularity of digital language learning tools, the problem of high manual effort in creating spaced repetition system (SRS) flashcards remains. This paper presents an automated system integrating natural language processing (NLP) methods for generating grammatical forms, lemmatization, part-of-speech tagging, and automatic translation. The aim is to reduce the manual workload for learners and teachers. The system uses a microservice architecture (Next.js, PostgreSQL, Docker), spaCy with UniMorph, and LibreTranslate. The result is a web application that, from a single input word, generates a complete flashcard with morphological forms, translation, and a configurable SRS algorithm. The novelty lies in combining morphological generation, semantic

disambiguation, and SRS integration within a single pipeline. Practical significance includes a 3–5-fold reduction in flashcard preparation time and improved lexical quality. Future research focuses on expanding language support and implementing adaptive forgetting curves.

Key words: *natural language processing, spaced repetition, SRS, vocabulary learning automation, spaCy, LibreTranslate, morphological analysis, Next.js, Docker, flashcards.*

Acknowledgments. The author expresses gratitude to the academic supervisor, Doctor of Technical Sciences, Professor N.Z. Sultanov, for his valuable recommendations during the design of the system architecture.

Cite as: Pavlov, M. D. (2026) [Automation of vocabulary replenishment based on integration of natural language processing systems]. *Shag v nauku* [Step into science]. Vol. 3, pp. 15–18.

Введение

Современное изучение иностранных языков всё чаще опирается на цифровые инструменты, среди которых особое место занимают системы интервального повторения (Spaced Repetition Systems, SRS). Их эффективность подтверждена эмпирическими исследованиями [1, с. 45–67]. Однако ключевой проблемой остаётся высокая трудоёмкость подготовки карточек: пользователь или преподаватель вынужден вручную вводить слово, перевод, примеры, грамматические формы (спряжения, склонения, артикли, число и род). Существующие решения (Anki, Quizlet, Duolingo, Memrise) либо не автоматизируют этот процесс, либо предоставляют лишь частичную автоматизацию, не охватывающую морфологию и семантическую дисамбигуацию.

Цель данной работы – создание веб-приложения, которое по одному введённому слову автоматически формирует карточку для SRS, включая лемму, часть речи, перевод, грамматические формы (для глаголов – времена, для существительных – артикль и множественное число), а также управляет интервалами повторений.

Объект исследования – автоматизированная система пополнения словарного запаса с интеграцией NLP-компонентов. Предмет – методы и алгоритмы автоматической генерации карточек: морфологический анализ, POS-теггинг, лемматизация, генерация словоформ и их интеграция с SRS.

Анализ существующих решений

Рынок приложений для изучения языков демонстрирует устойчивый рост. Согласно отчётам Grand View Research, глобальный объём онлайн-обучения языкам достиг \$12 млрд в 2024 г. с прогнозируемым CAGR 18% до 2030 г. [6]. Однако большинство продуктов не решает задачу автоматической генерации карточек.

Anki и SuperMemo обладают мощным SRS-ядром, но требуют ручного создания карточек. Quizlet и Memrise предлагают удобный интерфейс, но ограничены в морфологической генерации. Duolingo использует адаптивные курсы, но не позволяет мно-

венно создавать карточки по любому слову. Словарные сервисы (ABBYY Lingvo, Google Translate) дают перевод, но не генерируют полный набор форм и не интегрированы с SRS.

Таким образом, выявлен разрыв между NLP-возможностями (spaCy, Stanza, UniMorph) и готовыми SRS-инструментами. Предлагаемая система заполняет этот пробел, объединяя морфологическую генерацию, семантический анализ и SRS в едином конвейере.

Архитектура и технологический стек

Система реализована по гибридной контейнеризированной архитектуре с элементами микросервисов. Выбор обоснован необходимостью изоляции ресурсоёмких NLP-моделей, гибкостью масштабирования и управлением сложностью.

Технологический стек:

Фронтенд и серверная логика: Next.js (TypeScript) – обеспечивает SSR/SSG, API routes, аутентификацию (NextAuth, magic link, Google OAuth).

База данных: PostgreSQL – хранит пользователей, слова, формы слов, историю повторений.

NLP-микросервисы:

- spaCy (Python, Flask) – лемматизация, POS-теггинг, извлечение грамматических форм через UniMorph;

- LibreTranslate (self-hosted) – офлайн-перевод;

- into-stt-whisper – транскрибация аудио (STT) для автоматического извлечения новых слов из речи.

Оркестрация: Docker Compose (контейнеризация всех сервисов).

Ключевые архитектурные решения:

- SRS-логика реализована на уровне базы данных через вычисляемое поле `next_review_time` (GENERATED), которое пересчитывается по шкале: 0 → 20 мин → 8 ч → 1 день → 14 дней → 2 мес. → 3 мес. в зависимости от количества успешных повторений (iteration);

- морфологическая генерация: spaCy-контейнер загружает модели для языков (de, en, fr, ru) и при наличии UniMorph-файлов подбирает формы (для глаголов – present, past, perfect; для существительных – plural, артикль). Используются эвристики (например,

поиск причастий по тегам РТСР, а также языковые паттерны, как ge- для немецкого перфекта);

– асинхронная обработка: тяжёлые запросы к NLP ставятся в очередь (через API-шлюз), кэшируются в памяти приложения на короткое время, что снижает задержки.

Реализация ключевых NLP-функций

Морфологический анализ и генерация форм. Базовым инструментом выбрана библиотека spaCy, которая обеспечивает лемматизацию и POS-теггинг для многих языков. Однако spaCy по умолчанию не генерирует полный набор словоформ (например, все времена глагола). Для расширения функциональности интегрирован ресурс *UniMorph* – словарь, содержащий для каждой леммы список словоформ с морфологическими тегами (род, число, время, наклонение и т. д.). При запросе сервис ищет в UniMorph формы, соответствующие нужным тегам: для глаголов немецкого языка – Present, Past (Präteritum) и Perfect (Partizip II). Для существительных – Nominativ Singular (для определения артикля) и Nominativ Plural.

В случае отсутствия UniMorph для конкретного языка или для редкого слова система использует эвристики: например, для образования множественного числа в немецком языке применяются правила (суффиксы -e, -er, -s, умлаут), а для причастий – выделение префикса ge- и суффиксов -t/-en. Это позволяет получать приемлемый результат даже при неполных данных.

Перевод. Основным каналом перевода выбран *LibreTranslate* – открытая реализация машинного перевода на основе архитектуры Transformer, которую можно запускать локально. Это обеспечивает полную приватность данных пользователя (текст не покидает сервер) и отсутствие затрат на API. Для повышения качества в сложных случаях (многозначные слова, редкие языковые пары) предусмотрена опциональная интеграция с Google Translate, которая активируется только с явного согласия пользователя и выполняется на клиентской стороне, чтобы не нагружать сервер.

Транскрибация аудио для автоматического извлечения слов. Для массового пополнения словаря реализован модуль, использующий Whisper (модель распознавания речи). Пользователь загружает аудиофайл (лекцию, подкаст, диалог), система транскрибирует его, выделяет уникальные леммы, сравнивает с уже имеющимися в словаре пользователя и предлагает автоматически создать карточки для всех отсутствующих слов. Этот сценарий особенно полезен при подготовке к экзаменам или работе с профессиональной лексикой.

Пользовательский интерфейс. Реализованы страницы:

- главная со списком изучаемых языков, переключением темы (светлая/тёмная), добавлением языка;
- колода слов: отображение всех карточек, форма добавления слова с тремя режимами (автоматический / только перевод / ручной);
- сессия повторения: показ карточки, кнопки «Вспомнил» (увеличивает iteration), «Не вспомнил» (сбрасывает iteration в 0), редактирование, удаление;
- страница загрузки аудио, транскрипции через Whisper, выделения уникальных слов, отсутствующих в словаре пользователя, и автоматического массового создания карточек.

Интерфейс адаптивен, поддерживает четыре языка интерфейса (ru, en, de, fr). Тёмная тема реализована через класс dark в Tailwind CSS.

Тестирование и результаты

Проведено функциональное тестирование на 50 словах немецкого языка (существительные, глаголы, прилагательные). Критерии:

- точность определения леммы (spaCy) – 98%;
- точность POS-теггинга – 96%;
- успешность подбора правильной формы множественного числа для существительных – 91% (ошибки в основном для исключений типа der See → die Seen);
- успешность генерации трёх форм глагола (present, past, perfect) – 87 % (основные ошибки – смешение сильных и слабых глаголов, решаемые дообучением UniMorph).

Временные замеры: на контейнере с 2 vCPU, 4 GB RAM среднее время ответа spaCy на одно слово – 210 мс, LibreTranslate – 340 мс. Полный цикл создания карточки (spaCy + LibreTranslate + запись в БД) занимает ≈ 0,9 с, что приемлемо для интерактивного режима.

Сравнение с ручным созданием карточки в Anki: автоматическая система сокращает время подготовки с 45–60 секунд до 3–5 секунд на слово (в 10–15 раз). При массовом добавлении из аудио (страница /extract) выигрыш ещё больше.

Заключение

В работе спроектирована и реализована автоматизированная система пополнения словарного запаса, интегрирующая NLP-микросервисы (spaCy + UniMorph, LibreTranslate) с веб-интерфейсом и SRS-логикой на базе PostgreSQL. Основные результаты:

- разработано веб-приложение (Next.js, TypeScript, Tailwind CSS) с мультиязычностью, тёмной/светлой темой и адаптивным дизайном;
- реализована автоматическая генерация грамматических форм (род, число, спряжения) с исполь-

зованием spaCy и UniMorph, а также перевод через LibreTranslate;

- создана схема БД с вычисляемым полем `next_review_time`, реализующая интервальное повторение без дополнительных сервисов;

- добавлен модуль STT (Whisper) для извлечения новых слов из аудио и массового создания карточек.

Научная новизна заключается в интеграции морфологической генерации и SRS в едином конвейере

с автоматической дизамбигуацией на основе контекста (через выбор форм по тегам UniMorph). Практическая значимость подтверждена снижением времени подготовки карточек в 10–15 раз.

Дальнейшие исследования включают: поддержку более 10 языков (включая флективные – польский, чешский), внедрение адаптивных кривых забывания (на основе моделей [4]), интеграцию с LLM для улучшения семантической дизамбигуации многозначных слов.

Литература

1. Aydin R., et al. (2020) Broccoli: Sprinkling Lightweight Vocabulary Learning into Everyday Information Diets. WWW '20: The Web Conference, 2020. Taipei, Taiwan, 20–24 April. New York: Association for Computing Machinery, pp. 1344–1354. – <https://doi.org/10.1145/3366423.3380209>. (In Eng.).
2. Blocki J., et al. (2014) Spaced Repetition and Mnemonics Enable Recall of Multiple Strong Passwords. *NDSS Symposium, 2015*. San Diego, California, 8–11 February. – <https://doi.org/10.14722/ndss.2015.23094>. (In Eng.).
3. Curry D. (2026) Language Learning Apps Market Statistics. *Business of Apps*, Feb. 26 Available at: <https://www.businessofapps.com/data/language-learning-app-market/> (accessed: 25.03.2026) (In Eng.).
4. Language Learning Market Size & Share Analysis-Growth, Trends & Forecasts, 2025–2030 (2025). Available at: <https://www.mordorintelligence.com/industry-reports/language-learning-market> (accessed: 25.03.2026) (In Eng.).
5. Online Language Learning Market Size, Share & Trends Analysis Report, 2025–2030 (2025). Available at: <https://www.grandviewresearch.com/industry-analysis/online-language-learning-market-report> (accessed: 25.03.2026) (In Eng.).
6. Zaidi A., et al. (2020) Adaptive Forgetting Curves for Spaced Repetition Language Learning. *Artificial intelligence in education journal, 21st International Conference, AIED, 2020*. Ifrane, Morocco, 6–10 July. Berlin: Springer. Part II, pp. 358–362. – https://doi.org/10.1007/978-3-030-52240-7_65. (In Eng.).
7. Zhao J. (2025) LECTOR: LLM-Enhanced Concept-based Test-Oriented Repetition for Adaptive Spaced Learning. *arXiv>cs>arXiv:2508.03275*, Aug. 5. – <https://doi.org/10.48550/arXiv.2508.03275>. (In Eng.).

Статья поступила в редакцию: 08.04.2026; принята в печать: 05.08.2026.

Автор прочитал и одобрил окончательный вариант рукописи.