

CZECHUP: A WEB SYSTEM FOR PREPARING FOREIGNERS FOR THE CZECH LANGUAGE EXAM WITH ADAPTIVE VOCABULARY AND AI-GENERATED EXERCISES

Polina Komárková¹

¹Department of Economics, Faculty of Business and Economics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/10.11118/ecos2026_016

ABSTRACT

This article deals with the design and implementation of the CzechUP web system that helps foreigners prepare for the mandatory Czech language exam required for permanent residence (level A2) or citizenship (level B1). The work responds to the lack of suitable tools for Russian- and Ukrainian-speaking users, for whom the most popular application Duolingo is available only via English. The system integrates a database of 4,372 words parsed from three official CEFR textbooks, the Majka morphological analyzer (up to 347 forms for a single adjective), exercise generation via the Gemini API, and translation via Seznam Slovník and DeepL. The backend is built in ASP.NET Core over PostgreSQL, the frontend in Next.js with localization into three languages. User testing with five respondents confirmed the usefulness of the grammatical-form filter as a unique feature and estimated the error rate of AI-generated exercises at approximately 10%.

Keywords: Czech language teaching, Czech for foreigners, CEFR, web application, morphological analysis, Gemini API

JEL Code: I21

1 INTRODUCTION

A foreigner seeking permanent residence in the Czech Republic must pass a Czech language exam at level A2; citizenship requires level B1 together with a test of civic knowledge (Národní pedagogický institut 2025). The exam lasts over three hours with a 60% pass threshold in each section — reading, writing, listening, speaking. According to the Czech Statistical Office (2025), the number of foreigners in the country grew by more than 50% after 2022, with Russian- and Ukrainian-speaking residents making up the dominant share. For them, a paradox opens up: the most widespread learning application Duolingo teaches Czech only via English. Users without strong English face a double language barrier before they can start studying. A survey of 78 respondents



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license. <https://creativecommons.org/licenses/by-sa/4.0/deed.en>

from the Brno community (Ukraine 38.5%, Russia 35.9%, Kazakhstan 15.4%) showed that 91% plan to obtain permanent residence or citizenship, but 37% hold no Czech language certificate. The aim of this work was to design and build a functional system that removes three repeatedly mentioned deficiencies of existing tools: the lack of a multilingual interface, missing vocabulary adaptability, and limited coverage of grammatical forms.

1.1 Language Level and Existing Tools

The Common European Framework of Reference for Languages (CEFR 2002) defines six levels from A1 to C2. For Czech, official descriptions exist for levels A1, A2, and B2 (Línek and Vlasáková 2005; Bidlas *et al.* 2005; Adamovičová *et al.* 2005), updated by Cvejnová (2016) and complemented by a description of A1–B2 (Nováková *et al.* 2022). Above B2 the available literature drops sharply. The classical printed source remains the textbook by Holá (2016) for levels A1 to B1. Duolingo is the most used Czech self-study application in the surveyed sample, but users repeatedly point out its limitations: Czech only via English, often unnatural sentences, and missing systematic grammar. Quizlet and Anki provide flashcards but no structured CEFR exercises. Molendová (2024) reached an analogous conclusion for English taught with digital tools: what matters is not the digitization itself, but how well the tool reflects the individual progress of the student.

1.2 Czech Morphology and Content Generation

Czech has rich inflection — a single adjective can have more than 300 forms. The Internet Language Reference of the Czech Language Institute offers no public API and its coverage of adjective forms is incomplete. The Majka morphological analyzer (Šmerk 2007), developed at the Faculty of Informatics of Masaryk University, covers approximately 90% of the CZES corpus and generates up to 347 forms for the word “hezký” with tags in a compact format (K1gFnSc7 = noun, feminine, singular, instrumental). The exercise typology of the system is based on the work of Wu *et al.* (2007), who described seven exercise types for a digital library of language learning exercises.

2 SYSTEM ARCHITECTURE

From the foreigner survey and the analysis of existing applications, five groups of functional requirements were derived: account management (registration, JWT authentication), word learning (filtering by CEFR, topic, and tags, manual additions), grammar rules with user notes, exercises with automatic evaluation, and offline study via PDF export.

The backend is implemented in ASP.NET Core over PostgreSQL with Entity Framework Core (Code First) following the four-layer Clean Architecture model (Martin 2018). Three projects separate the entity layer (CzechUp.EF with 19 entities), the service layer (CzechUp.Services with DTOs and business logic), and the controller layer (CzechUp.WebApi). The frontend is built in Next.js with the App Router and TypeScript; localization is handled by next-intl for three languages (Czech, Russian, English), indexed in URL prefixes `/cs/`, `/ru/`, `/en/`. Styling uses Tailwind CSS, client-side PDF export uses jsPDF.

The database contains 25 tables and 19 entities in third normal form. The tables are split into two groups: **General*** for static data (official vocabulary, default topics, grammar rules) and **User*** for dynamic data of a specific user (personal dictionary, exercise results, notes). For a newly

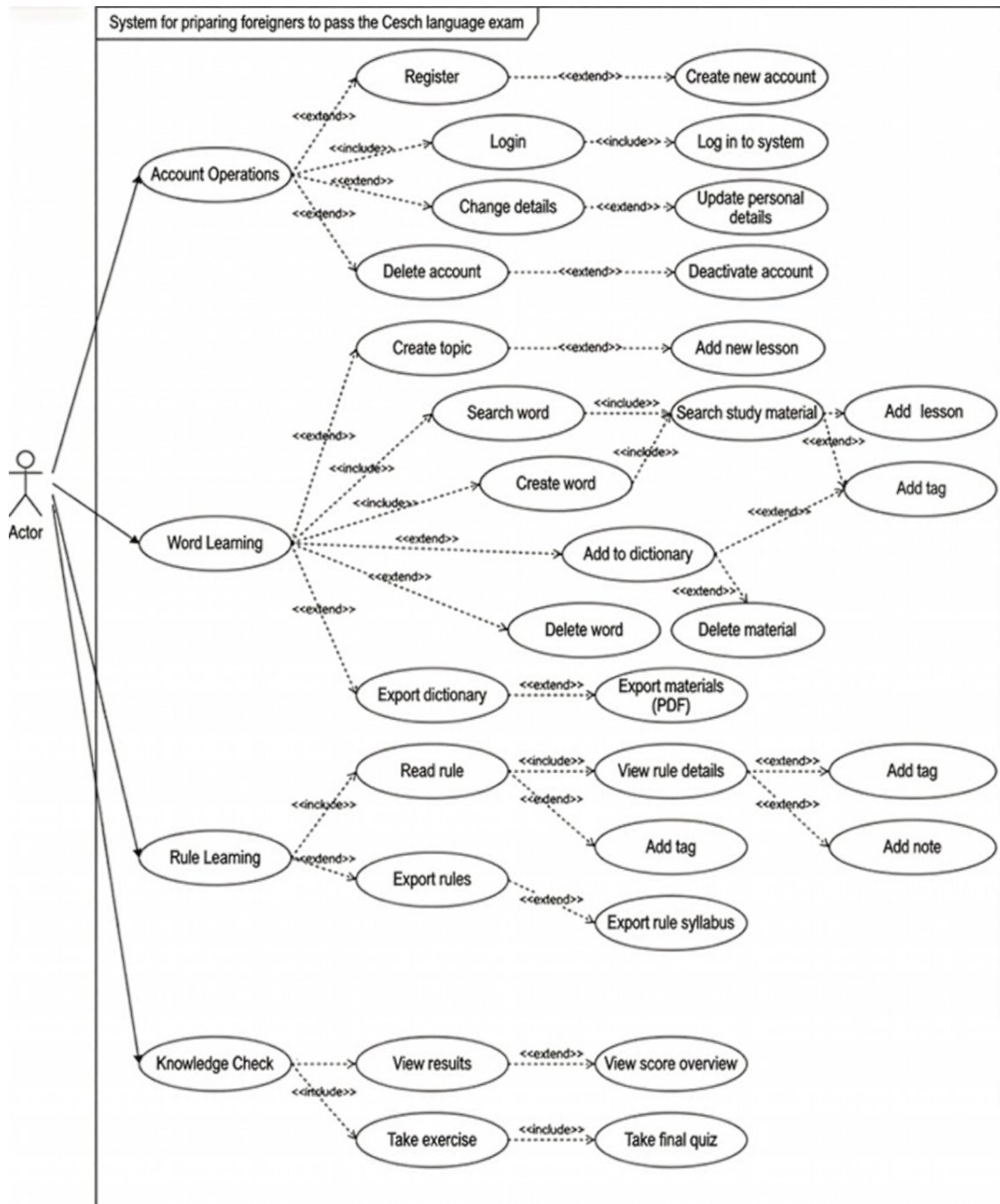


Fig. 1: USE CASE diagram of the CzechUP system (adopted from Tselishcheva 2025).

registered user, data from **General*** tables are copied into **User*** tables, so the user can edit the default content without affecting others. This split requires more storage but dramatically simplifies user data isolation.

For authentication, the system uses JSON Web Tokens with a 1-hour validity, and passwords are hashed with PBKDF2 with a random salt. The [Authorize] attribute protects most endpoints; data isolation works by every CRUD endpoint operating only on records whose UserGuid matches the identifier in the token.

3 IMPLEMENTATION AND RESULTS

To populate the default database, a console application WordHandler was written that parses three different formats of the official CEFR Czech reference textbooks. The merging into the shared GeneralOriginalWords table is summarized in Tab. 1.

Tab. 1 Number of words from three CEFR reference sources after merging

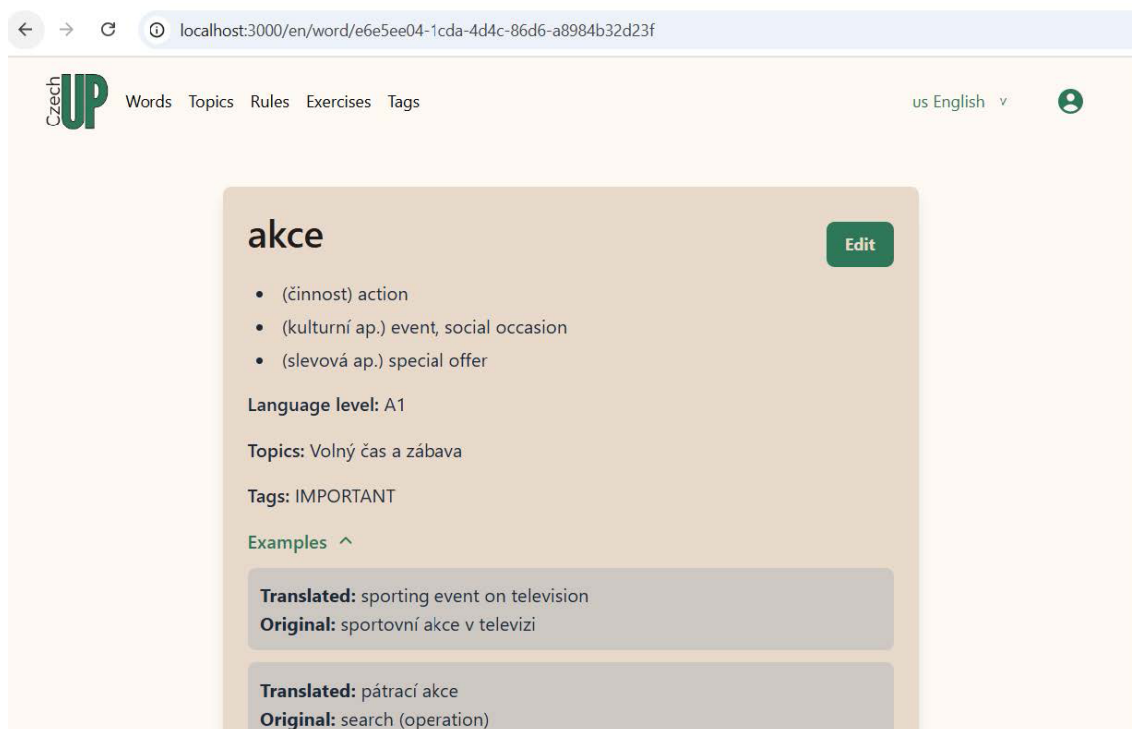
Level	Source 1	Source 2	Source 3	Total
A1	1094	123	156	1373
A2	1213	88	110	1411
B1	1128	0	0	1128
B2	353	0	107	460
Σ	3788	211	373	4372

For automatic translation, two sources were combined: the Seznam Slovník API as the primary source for individual words (Czech plus 10 languages) and the DeepL API as a fallback for missing words and for whole sentences. Testing showed that Seznam is more accurate for individual words while DeepL is stronger on sentences; for single words DeepL has approximately 10% inaccuracy — typically picking one of several possible translations without context. For multi-word expressions, Majka supplies the morphological forms; five typical structures were identified (verb + noun, adjective + noun, etc.) with rules for which words to inflect.

Exercise generation uses the Gemini API from Google (2025), available in a free tier sufficient for occasional generation. Three exercises were planned per word per level — more than 13,000 exercises in total, which manual authoring could not produce. Five exercise types were implemented: completing a word form in a sentence, sentence construction from words, completing key words in a text (all via AI), choosing the correct translation from 4–8 options, and writing the Czech translation (without AI, drawn from the database). Finding a prompt that returns consistent JSON required more than 30 iterations — the models occasionally changed the output structure, added comments outside the JSON, or used different field names.

The Word Forms section is a unique feature of the application. For nouns the filter offers two criteria (number, case), for adjectives five (case, number, negation, gender, degree of comparison), and for verbs six. For “přijatelný” in dative, plural, second degree, the application returns “nepřijatelnějším / přijatelnějším”. For “kočka” Majka generates 13 forms, for “hezky” up to 347.

The application was tested by five users: three Russian-speaking foreigners at levels A1, B2, and C1 (Russian localization), one English-speaking foreigner with no Czech, and one native Czech speaker with C1 English (who evaluated translation quality). Positive feedback highlighted the breadth of functionality, translation quality, and especially the grammatical-form filter, which is missing from other tools and serves as a key aid for advanced foreigners when writing in Czech. All respondents stated they would use the application after release. Negative feedback comprised three points: three of five users noted that the UI does not feel modern enough and requires a graphic redesign; two suggested an instructional video, since some functions are not immediately obvious; and the native speaker estimated that approximately 1 in 10 generated exercises contains a minor error (Tselishcheva, 2025).



Tab. 2 Detail of the word “akce” with translations, examples, and the Word Forms section (adopted from Tselishcheva 2025)

For regression stability, three groups of xUnit tests were written: external API tests (URL validity, data structure, translation quality), endpoint tests with mock objects via the Moq library, and integration tests with an in-memory database via `Microsoft.EntityFrameworkCore.InMemory`.

4 DISCUSSION

The feature that users named the most valuable — the grammatical-form filter — emerged from a technical decision that originally had a different purpose. When it became clear that the Internet Language Reference offered no API while Majka did, the complete list of forms became a by-product. This is a general lesson: some of the most valuable features are not designed during planning, but emerge from solving technical problems with other goals. For AI-generated content the second lesson holds: the ~10% error rate means that AI alone is not enough as a final source of content for language teaching — it requires either human review (expensive), a second LLM as a validator, or a user error-reporting mechanism. The main limitations of the system remain the missing audio for pronunciation and listening, suboptimal UI design, and the status of external APIs, which complicates public deployment without commercial agreements.

5 CONCLUSIONS

The work demonstrated that commonly available APIs and open-source tools can be used to build a prototype language learning application that addresses a concrete pain point of foreigners in the Czech Republic — the Czech language exam for permanent residence and citizenship. Multi-source integration works, but at the cost of more complex tooling — merging three official CEFR textbooks (4,372 words) required three different parsers, and the combination of Seznam Slovník + DeepL + Majka was necessary because no single API handles both Czech translation and morphology. AI-generated content for language teaching needs a validation layer — the Gemini API accelerated exercise preparation, but ~10% of outputs contain errors a student would treat as authoritative. A feature that does not seem important during planning (a filter over 347 adjective forms) can be what users see as the deciding factor. Future expansions may include adding audio recordings via TTS, a validation layer for AI-generated exercises, an administrator interface for bulk corrections, and extending the localization to Ukrainian and Vietnamese.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-financed from the Operational Programme Research, Development and Education.

REFERENCES

- Adamovičová, A. *et al.* (2005). *Czech as a foreign language. Level B2 according to the Common European Framework of Reference for Languages [in Czech]*. Praha: Tauris.
- Bidas, V., Confortiová, H., Turzíková, M. (2005). *Czech as a foreign language. Level A2 according to the Common European Framework of Reference for Languages [in Czech]*. Praha: Tauris.
- CEFR. (2002). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Czech edition)*. Olomouc: Univerzita Palackého. ISBN 80-244-0404-4.
- Cvejnová, J. (2016). *Reference Description of Czech for the Language Exam for Permanent Residence in the Czech Republic — Levels A1, A2 [in Czech]*. Praha: Národní ústav pro vzdělávání. ISBN 978-80-7481-164-7.
- Czech Statistical Office. (2025). Number of foreigners, demographic events [in Czech]. CSÚ. <https://csu.gov.cz/pocet-cizincu-demograficke-udalosti>
- Google. (2025). *Google AI for Developers (Gemini API)*. <https://ai.google.dev/>
- Holá, L. (2016). *Česky krok za krokem 1 = Czech step by step 1: A1–A2*. 1st ed. Praha: Akropolis. ISBN 978-80-7470-129-0.
- Línek, J., Vlasáková, K. (2005). *Czech as a foreign language. Level A1 according to the Common European Framework of Reference for Languages [in Czech]*. Olomouc: Univerzita Palackého.
- MARTIN, R. C. (2018). *Clean Architecture: A Craftsman's Guide to Software Structure and Design*. Boston: Prentice Hall. ISBN 978-0-13-449416-6.
- Molendová, K. (2024). *Vocabulary Teaching in English with Digital Technologies [in Czech]*. Master's thesis. Plzeň: University of West Bohemia.
- National Pedagogical Institute of The Czech Republic (2025). Czech language exam for permanent residence and citizenship [in Czech]. *npi*. <https://cestina-pro-cizince.cz/>
- Nováková, R. *et al.* (2022). *Descriptions of Reference Levels A1–B2 for Czech Sign Language [in Czech]*. Praha: Národní pedagogický institut České republiky. ISBN 978-80-7578-102-4.

- Šmerk, P. (2007). Fast Morphological Analysis of Czech. In: Sojka, P., Horák, A. (eds.) *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. Brno: Masaryk University, pp. 13–16. ISBN 978-80-210-5048-8.
- Tselishcheva, P. (2025). CzechUP: a web system for preparing foreigners for the Czech language exam [in Czech]. Brno: Mendel University in Brno. Master's thesis, head: FARANA, R.
- Wu, S., Witten, I. H., Edwards, A., Nichols, D., Aquino, R. (2007). A digital library of language learning exercises. *International Journal of Emerging Technologies in Learning*. 2(1).

Contact information

Polina Komárková: e-mail: polia.tselisheva@mail.ru