

DESIGN AND EVALUATION OF A SOFTWARE TOOL FOR FOREIGN LANGUAGE LEARNING

LUKOV D, KHOSHABA O (pzmag2022@gmail.com)

Vinnitsia National Technical University (Ukraine)

Abstract. This paper addresses the design and evaluation of a software tool that supports foreign-language learning through multimedia content, adaptive practice, and learner analytics. The practical motivation is twofold: first, many CALL (computer-assisted language learning) systems underutilise empirically grounded design principles from educational psychology and usability engineering; second, despite widespread interest in gamification, spaced practice, and pronunciation feedback, many tools lack transparent quality models and rigorous evaluation protocols aligned with educational outcomes.

Keywords: computer-assisted language learning, adaptivity, spaced repetition, pronunciation feedback, CEFR, usability, learning analytics.

Problem statement. Learners and instructors require software that effectively supports vocabulary growth, pronunciation practice, and comprehension while remaining usable and trustworthy. Typical gaps include (i) insufficient alignment with evidence-based multimedia learning principles [2], (ii) inconsistent vocabulary modelling and feedback [4], (iii) ad-hoc heuristic design without systematic usability assessment [5], and (iv) lack of clearly defined quality attributes for reliability, security, and maintainability in educational deployments [3].

Objectives and tasks. We pursue four objectives: O1 — design an integrated, modular prototype covering vocabulary, pronunciation, and comprehension; O2 — embed evidence-based instructional principles (multimedia learning, spaced practice) and user-centred heuristics; O3 — operationalise product-quality requirements with ISO/IEC 25010:2023; O4 — establish a two-stage evaluation protocol with quantitative acceptance thresholds and learning outcomes. The solved tasks include: T1 — requirements analysis mapped to SQuARE product-quality characteristics [3]; T2 — definition of a content and item model for lexical units (form, lemma, IPA, audio, sense relations) and tasks (recall, recognition, pronunciation); T3 — implementation of an adaptive scheduling policy for spaced practice grounded in the spacing effect [8]; T4 — design of pronunciation feedback with phone-level alignment and confidence scoring; T5 — instrumentation for usability analytics and learning telemetry; T6 — design of a formative usability study and a 4–6-week pilot aligned to CEFR (Common European Framework of Reference) descriptors [9].

Design and implementation. The system adopts a modular architecture to support ablation-friendly evaluation. A content service stores lexical items with metadata (orthography, lemma, part of speech, IPA transcription, syllabification, audio tokens, difficulty history), and task templates for responsive exercises (e.g., picture-word matching, cloze, minimal pairs). A practice engine schedules items using an item-level memory model that estimates forgetting rates from response latency and accuracy; initial intervals (e.g., 1, 2, 6, 20 days) are adapted via Bayesian updates over item- and learner parameters, operationalising distributed practice [8]. A pronunciation module offers per-phone feedback (alignment, Goodness-of-Pronunciation-like scores) with actionable cues (stress placement, common substitutions) and optional phonological minimal-pair drills. The UI follows multimedia learning principles (contiguity, signalling, coherence) to limit extraneous cognitive load [2], with optional micro-gamification (points, streaks) constrained by pedagogical goals and evidence from the literature [1].

Quality model and non-functional requirements. We adopt ISO/IEC 25010:2023 to elicit and verify product quality characteristics (functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, portability) with scenario-level measures, e.g., time-to-first-byte for audio, uptime targets, P95 latency under classroom Wi-Fi, and

error-budget policies for speech services [3]. Usability is operationalised via Nielsen–Norman’s ten heuristics (visibility, match to real world, user control, consistency, error prevention, recognition over recall, flexibility and efficiency, aesthetic and minimalist design, error recovery, help & documentation) as a review checklist during design sprints and as a basis for formative heuristic evaluation [5].

Evaluation protocol. We predefine a mixed-methods evaluation with two stages. Stage A (formative) combines heuristic evaluation and think-aloud usability sessions; Stage B (summative) is a 4–6-week classroom pilot aligned with CEFR descriptors [9]. Primary quantitative metrics and acceptance thresholds are: A1 — System Usability Scale (SUS) ≥ 80 ; A2 — time-on-task reduction $\geq 20\%$ relative to a baseline interface; A3 — task-accuracy improvement of +10–15 percentage points on targeted items; A4 — learning gains with standardised effect size $d \approx 0.4$ – 0.6 on CEFR-aligned pre/post measures. Secondary metrics include retention curves for scheduled items, pronunciation error rates by phone class, and engagement continuity (session frequency, adherence to schedule). Where applicable, outcomes are stratified by prior proficiency and task type.

Participants and materials. Stage A samples $n \approx 12$ – 18 adult learners (A2–B1) to saturate usability issues; Stage B samples $n \approx 60$ learners across two course sections for four to six weeks (two 45-minute sessions per week). Materials include a 1,200–1,800-item lexical set covering high-utility lemmas, topical clusters, and minimal-pair lists balanced by functional load; item difficulty is seeded via corpus frequency and iteratively updated from learner responses [4], [9]. Ethical procedures include informed consent, anonymisation of telemetry, and the right to withdraw without penalty.

Measures and analysis. SUS is computed per standard; time-on-task and error rates are logged across canonical tasks (add item to review, complete pronunciation drill, complete recall set). Learning gains are measured with CEFR-referenced tasks (listening/reading micro-tests and controlled-production prompts) scored with analytic rubrics. For Stage B, we apply pre-registered analyses: mixed-effects models on item-level accuracy and latency; standardised mean-difference (Hedges g) for pre/post; and survival-curve-based modelling for retention. Piloting includes ablations (e.g., disabling signalling cues or removing per-phone feedback) to estimate contributions of submodules.

Preliminary results (Stage A). In a formative study ($n=14$), the prototype achieved SUS 82.1 (SD 6.4), a 22% median reduction in time-on-task across three core tasks, and a 12.7 percentage-point improvement in immediate item accuracy compared with a wireframe baseline. Qualitative themes included appreciation of signalling and contiguity, requests for slower audio in minimal-pair drills, and the need for clearer progress affordances. These results meet A1 – A3 and justify the Stage B pilot; A4 will be evaluated over longer exposure.

Discussion. The design operationalises well-supported principles: multimedia learning reduces extraneous load [2]; spaced practice enhances durable retention [8]; and structured heuristics improve usability [5]. Incorporating ISO/IEC 25010:2023 embeds verifiable quality requirements beyond pedagogy, which is essential for institutional adoption [3]. Gamification is employed judiciously per the meta-analytic evidence that effects depend on meaningful alignment with learning processes rather than superficial badges [1]. The architecture’s modularity enables targeted improvements and fair testing of subcomponents.

Limitations and risks. External validity remains constrained by the pilot’s duration and sampling; ASR robustness may degrade for certain L1 accents and low-bitrate channels; content coverage skews towards general-purpose lexis; and teacher-facing analytics require further validation for decision-making. Future work includes: robust accent adaptation, improved offline caching for connectivity-limited classrooms, and expansion to task-based speaking practice with calibrated descriptors aligned to CEFR mediation scales [9].

Conclusions. We present a prototype and evaluation plan for a language-learning tool grounded in multimedia learning, spaced practice, and usability engineering. The system meets predefined usability and efficiency thresholds in formative tests and establishes transparent quality and outcome measures for a classroom pilot. The approach demonstrates how educational software

can be engineered with auditable links from pedagogy to product quality and measured learning gains, addressing persistent gaps in CALL practice [1]–[5], [8], [9].

References:

- [1] M. Sailer and L. Homner, “The Gamification of Learning: A Meta-Analysis,” *Educational Psychology Review*, vol. 32, no. 1, pp. 77–112, Mar. 2020, doi: 10.1007/s10648-019-09498-w.
- [2] R. E. Mayer, *Multimedia Learning*, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2021.
- [3] ISO/IEC 25010:2023, *Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—Product quality model*. [Online]. Available: [\[https://www.iso.org/standard/78176.html\]](https://www.iso.org/standard/78176.html)(<https://www.iso.org/standard/78176.html>). [Accessed: 16 Oct. 2025].
- [4] I. S. P. Nation, *Learning Vocabulary in Another Language*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2013.
- [5] Nielsen Norman Group, “10 Usability Heuristics for User Interface Design.” [Online]. Available: [\[https://www.nngroup.com/articles/ten-usability-heuristics/\]](https://www.nngroup.com/articles/ten-usability-heuristics/)(<https://www.nngroup.com/articles/ten-usability-heuristics/>). [Accessed: 16 Oct. 2025].
- [6] C. A. Chapelle, *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research*. Cambridge, U.K.: Cambridge University Press, 2001. doi: 10.1017/CBO9781139524681.
- [7] P. A. Kirschner, J. Sweller, and R. E. Clark, “Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching,” *Educational Psychologist*, vol. 41, no. 2, pp. 75–86, 2006, doi: 10.1207/s15326985ep4102_1.
- [8] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, “Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis,” *Psychological Bulletin*, vol. 132, no. 3, pp. 354–380, 2006, doi: 10.1037/0033-2909.132.3.354.
- [9] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*, 2020. [Online]. Available: [\[https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions\]](https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions)(<https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions>). [Accessed: 16 Oct. 2025].