

AI-assisted second-language teaching and learning in the Zone of Proximal Development

Jue Hou , Anh-Duc Vu , Anisia Katinskaia  & Roman Yangarber 

University of Helsinki, Finland

This paper presents the integration of AI features into the language-teaching platform, Revita. The system is an intelligent online tutor, developed to support learners from lower-intermediate toward advanced levels, in several languages. Target skills currently include grammar, vocabulary, aural comprehension, and pronunciation. Based on authentic texts uploaded by the learners themselves, the system creates a rich variety of exercises that are tailored to the individual learner's level of proficiency. Revita's main guiding principle is *personalization*, motivated by current theories from educational science, notably Vygotsky's concept of the Zone of Proximal Development and dynamic assessment, as well as the principles of diagnostic assessment. The linguistic foundation for the system comes from Construction Grammar, the goal being to build a complete inventory of *constructions* in the target language, as the basis for judging the correctness of the learners' responses to the exercises. Revita is enhanced with AI tools from natural language processing, machine learning, and educational data mining.

Keywords: personalized language teaching, artificial intelligence, computer-aided language learning, construction grammar, grammatical error correction, learning analytics

Email address for correspondence: roman.yangarber@helsinki.fi

© The Author(s) 2025. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Introduction: Personalized language teaching

In recent years, we observe a trend toward flexible online language-learning tools that are accessible anywhere on demand. Despite the growing need for and popularity of such tools, most existing systems do not address the fundamental requirements of language learners and teachers. Many resources exist on the Web—various free and commercial applications—which support *beginners*, some with many millions of users. However, once the learner passes the beginner level and reaches lower-intermediate (LI) level—for example, once she passes beyond A1 on the CEFR scale—resources for progressing toward advanced levels become drastically limited. As our research has shown, no system today provides substantial support for LI learners in multiple languages.

In this paper, we describe Revita, a freely available platform designed to support second language (L2) or foreign language learning and tutoring beyond the beginner level. Revita works with several languages in various stages of development—ranging from languages that are well-developed and are currently deployed in university-level courses, to those still in initial “beta” (β) versions. These include several “majority” languages—Finnish, Russian, Italian, ^{β} English, ^{β} German, ^{β} —and several endangered minority languages (at present all in beta stages).

The primary focus in Revita is on *personalization*: personalized teaching and learning. Personalization is achieved by following Vygotsky’s theories of the Zone of Proximal Development (ZPD) (Vygotsky, 1978). In particular, we pursue these guiding principles:

- A. The intelligent tutor offers the learner practice sessions with exercises, which must be optimally *tailored* to each individual learner’s current level.
- B. During practice, when the learner has problems while working on an exercise, the tutor must provide *graduated feedback* to the learner, which consists of not merely a response saying “correct/incorrect”, but rather of a sequence of hints. The hints *assist* the learner by guiding her toward finding the correct

answer on her own.

These are the main guiding principles in the construction of the intelligent tutor. The relevance of our work to research in Second Language Acquisition (SLA) in general—and to this volume in particular—lies in the fact that these two fundamental principles are directly linked to the current practices in the field of assessment of the learner's proficiency. Personalization in computer-aided language learning (CALL) requires a continual, on-going and detailed assessment of each individual learner's current level of proficiency. Principle (A)—providing personally tailored exercises—involves *diagnostic* assessment: the tutor must have as detailed a model as possible of what the learner knows and does not know at present, in order to identify the learner's Zone of Proximal Development as accurately as possible, and to ensure that the next exercise falls exactly within the learner's ZPD. Principle (B)—providing graduated guidance to the learner—corresponds to the notion of dynamic assessment, where the tutor iteratively assists the learner during practice and, crucially, at each step the tutor provides *minimal* assistance to the learner to help her discover the correct answer on her own. The graduated hints move from offering the most general information toward increasingly more specific information in order to maximize the learner's involvement in the learning process. At the end, once the tutor has run out of hints to offer, the tutor gives the most specific information: revealing the correct answer.

Thus, the principles underlying our work echo the prevalent directions in the field of assessment over the past 20 years: they build on the same theories and practices of diagnostic and dynamic assessment as those covered elsewhere in this Special Issue. We arrived at these same ideas via a parallel development that started from somewhat different points of view, namely, through the drive for personalization, that is technically oriented toward improving the learner's experience. Our approaches converge with those in Dynamic Assessment (Poehner & Leontjev, 2020) and in Diagnostic Assessment (Huhta et al., 2024). This paper describes the concrete details of our implementation of AI features to support dynamic and diagnostic assessment in CALL.

For the content of the practice sessions, Revita allows the learners to select an *arbitrary, authentic* text, which can be uploaded to the platform by *the learners themselves* (or by the teacher), and to automatically generate a rich variety of exercises based on the text. During practice, the system aims to adapt the level of the exercises to each individual user, depending on her current level of proficiency, which it tries to estimate based on her answers to previously completed exercises. The ability of the learner to select content that maximally matches her needs and interests—through the use of authentic material—is intended to boost the learner’s motivation and is another central aspect of personalization.

Revita lies at the intersection of two established areas of research: intelligent tutoring systems (ITS) and CALL. The project seeks to advance intelligent solutions specifically for language teaching and learning. On the other hand, crucially, Revita has the potential for enriching the language teaching process itself, since the platform is used for collecting, mining and analyzing massive learner data.

The paper is organized as follows: Section 2 discusses related prior work; Section 3 introduces the main features of Revita; Section 4 discusses the methodology of incorporating artificial intelligence into language teaching, and how this methodology is supported by current research; and Section 5 concludes the paper and discusses future directions.

Prior work

Interest in computer-assisted language learning (CALL) is growing with the rapid development of language technology. CALL is seen as the “study of applications of the computer in language teaching and learning” (Levy, 1997). Applying ITS to language learning and supporting CALL by intelligent and/or adaptive methodologies, such as expert systems (ES), natural language processing (NLP), and automatic speech recognition (ASR), is the domain of intelligent CALL (sometimes referred to as ICALL). The goal of CALL is to build advanced applications for language learning using NLP and linguistic resources: corpora, lexicons, etc. (Volodina et al., 2014).

The number of academic and commercial tools for language learning is growing rapidly, e.g., popular systems such as Duolingo, Rosetta Stone, Babbel, Busuu, and iTalki. Some CALL systems aim to give learners access to *authentic* materials (White & Reinders, 2010), the opportunity to interact with teachers and native speakers (e.g., the app *Lingoda* is a platform for live video classes) and provide text or audio feedback based on learner needs and knowledge (Bodnar et al., 2017). Modern CALL systems are also mobile, which increases their accessibility (Derakhshan & Khodabakhshzadeh, 2011; Kaceti & Klímová, 2019; Rosell-Aguilar, 2018).

Following the accepted practice in ITS, we view the system as containing three principal components: (1) the Domain model, (2) the Student model, and (3) the Instruction model (Katinskaia et al., 2018). This is designed to mirror the characteristics that a good (human) teacher should possess: a good teacher (1) knows the domain; (2) knows the student: what she knows vs. does not know; and (3) knows how to teach: what exercises would be most appropriate or productive to offer to the student next to optimize the learning process. These three characteristics correspond to the three principal models in the ITS frameworks. The Domain model (1) captures the information about the subject to be learned—in our case, the language. The Student model (2) captures the knowledge the system has about each individual student at any time, based on the history of the student's past performance on exercises already completed. The Instruction model (3) encodes the system's knowledge about which learning paths would produce optimal results, given the student's current state. In selecting the best exercises, the Instruction model is based on Vygotsky's theory, in particular, the concept of the Zone of Proximal Development (Vygotsky, 1978, 2012). This builds on the idea that the exercises must match the learner's current level of proficiency. If the exercises are too easy for the student too often, the student will become bored; if the exercises are too difficult too often, the student will become frustrated—in either case, the student will quit learning. Thus, it is of primary importance to predict for each exercise and each student, how likely the student is to succeed with this exercise. We aim to select those for which the probability is near 0.5 (for the

majority of the time; occasional deviations are, of course, admissible).

In developing CALL, pedagogical goals rather than technological means should be the primary focus (Gray, 2008). It has been shown that using CALL systems for education improves learner motivation and attitudes, increases options for self-study (Golonka et al., 2014), and improves retention of various learning concepts, as well as communication between students and teachers, academic self-efficacy (Bandura, 1977; Rachels & Rockinson-Szapkiw, 2018), and overall language skills (Yeh & Lai, 2019; Zhang & Zou, 2022).

Despite decades of research, CALL still has a number of serious limitations. Apart from platforms where teachers interact with students in video class, most existing CALL systems follow the “canned” approach: the available sets of exercises are *static*, pre-made, and therefore limited. This restricts personalization: static, pre-made exercises can be rearranged into different individual programmes, but a limited selection of exercises means that once the student has covered the available exercises, there is no opportunity for further learning.

We build intelligence into the system to perform a number of tasks similarly to how a good teacher would perform them. One type of intelligence needed is the ability to generate exercises *dynamically*, based on any content that the learner prefers or needs to study. This results in an unlimited supply of exercises. Another type of intelligence is needed when selecting from among the many generated exercises those that optimally fit the current level of each individual learner, which requires accurate assessment of the learner’s current skills. This builds on the vast body of work in Dynamic Assessment, (*inter alia* Leontjev et al., 2025; Poehner & Leontjev, 2020), to guide the learner’s practice sessions, as well as in Diagnostic Assessment (Huhta et al., 2024). We describe how these aspects of intelligence are achieved in the following sections.

System overview

Main principles

The approach adopted for our AI-based language-teaching platform, Revita, rests on the following main principles, which we have introduced in the preceding sections:

1. Practice should be based on *authentic content*. By authentic we mean a text that is not artificial and is not written expressly for learning purposes. The users (teachers and learners) can upload any text that interests them, e.g., from the Internet using a URL, or from an uploaded file, and use it directly as learning content.
2. Exercises are *automatically generated* based on any authentic text chosen by the user.
3. Exercises are *personalized* to match the learner's current skill levels, so that each new exercise is selected to be a challenge that the learner is prepared to meet.
4. *Graduated feedback*: rather than providing only a “right/wrong” response, the tutor *gradually guides* the learner toward finding the correct answer on her own, by providing a series of *hints*, which begin as general hints about the context and then give more and more specific information about the correct answer.
5. *Continual assessment* of skills allows Revita to select exercises that are optimally personalized for each learner based on current performance.

The first principle is the bedrock of Revita's philosophy: in the story-based approach, all learning activities are based on authentic texts, which should be *inherently interesting* for the learner to read, thus stimulating motivation and hence longer practice sessions. A few sample texts are pre-loaded into the system's “public” library for each language; also, several stories are recommended daily as

“*Story of the day*”, crawled from selected websites containing, e.g., current news in the language of choice. But the main idea is to allow the learners themselves (or their teachers) choose texts and upload them into their private libraries.

Linguistic constructs

At the core of Revita’s Domain model, the description of the domain to be learned, is the inventory of linguistic constructs. Constructs are linguistic phenomena or rules that vary in their level of specificity, with some being quite specific and some quite general. Ultimately, the goal in Revita is to include all *constructions*, as conceived of in Construction Grammar (CG) (Fillmore, 1988; Janda et al., 2020). CG treats a wide range of linguistic phenomena —grammatical constructs, multi-word expressions, collocations, idioms, etc.— within a unified formalism. A construction is defined as any association of form and meaning, whose form or meaning *cannot* be inferred from the form or meaning of its individual parts (Goldberg, 2006). We follow the fundamental view that L2 teaching and learning consists in teaching and learning constructions (Boas, 2022; Herbst, 2016): the L2 learner has or has not mastered the language to the extent that she has or has not mastered any of the constructions needed to operate within the language. A construction must have some fixed elements and may have one or more variable elements.

One important example of a construct is *verb government*. For example, (in Finnish) in *tutustua Pekkaan* (“to get acquainted with Pekka”), the verb *tutustua* (“to become acquainted”) requires its argument to be in the illative case: *Pekkaan* (literally: “into Pekka”), while in *tykätä Pekasta* (“to like Pekka”), the verb *tykätä* (“to like”) requires its argument to be in the elative case: *Pekasta* (literally: “from Pekka”).

In government, the lemma of the verb and the case of the argument never change, while everything else can vary: the lemma and number of the argument, the mood and tense of the verb, etc. The key point is that the case of the argument is part of the construction, it is part of the lexical *knowledge* of the speaker or learner and

cannot be deduced from the form or the meaning of the verb.

Another example is the “substitute relative clause” construction, which has the following components: (1) a **verb** with the semantics of conveying information (that is conveyed in the relative clause); (2) a **subject** (of the relative clause) in the genitive case; and (3) a **participle**, also in the genitive singular, which acts as the substitute for the relative clause and semantically is its verb.¹² This construction substitutes for a relative clause (which would be introduced by “that”):

Maija kertoi vanhempien asuvan kaupungissa.

(“Maija **told** [that her parents live in the city]”).

(literally: “Maija **told of her parents**_{Genitive} **living**_{Genitive} in the city.”)

In Revita, *constructs* form the core of the Domain model (Katinskaia et al., 2023). When customizing the system for a new language, we must resort to expertise in language pedagogy for creating the inventory of constructs which is to be mastered by the learners. Currently, Finnish and Russian have the most developed systems of constructs, each with over 200 of them (starting with the most frequent ones that are important for learning). The number of constructs can be potentially much larger. The Russian constructs evolved from the extensive grammatical inventory covered in placement tests for L2 learners developed at the University of Helsinki (Kopotev, 2012). Finnish constructs are based on inventories of grammatical topics developed by experts in teaching Finnish as L2. Inventories of constructs, known as *constructicons*, are being developed for a number of languages, and this is an area of active research in applied linguistics (Endresen et al., 2022; Janda et al., 2020), with language teaching considered their primary application.

² Note, that (2) may be singular or plural but (3) must be singular, which may seem unnatural to the learner, since normally subject and verb agree in case and number. Yet in this construct, the verb must be singular. This is a matter of knowledge of the construction and cannot be guessed or inferred without such knowledge.

2 / 5

2. Aurinkoenergia tulevaisuuden kaupungeissa.

Highlight exercise difficulty

Energiakriisin lähestyessä kaikki keinot on otettu käyttöön. Eri ratkaisuja sähköntuotantoon ja lämmitykseen on kehitetty. Esimerkiksi asiantuntijat ennustavat, että kesäisin ja lämmitystä talvisin.

Myös Suomessa kaupungit ovat halukas ottaa aurinkopaneeliin. Esimerkiksi Hanko ja Helsinki ryhtyneensä lisäämään aurinkopaneeli rakennuksiinsa ollakseen valmiina energiakriisiin iskiessä.

Show me a hint!

- This is an object of active positive verb.
- This is the object of "lisätä".
- Use plural.
- Use another case.

aurinkopaneeli → English

Translate into ⇒ English

aurinkopaneeli ✓ ?

- solar panel
- solar cell panel
- solar energy panel
- solar-cell array
- solar collector

paneeli

- panelling
- panel
- paneling
- panels

aurinko

- sun

Check Answers

Next snippet ↓

Figure 1. Practice mode based on a story.

Exercise generation based on constructs

The main learning activity in Revita is the *Practice mode*. Practice is based on stories from private or public libraries. Revita splits the story into small snippets, each corresponding to about two sentences or a short paragraph. The system offers the learner one snippet at a time for practice.

In each snippet, the system selects a certain word or phrase in the text, hides it from the learner, and replaces it with an exercise. Revita can offer several types of exercises:

1. *Cloze* (fill-in-the-blank): the learner is presented with a text box, with a cue about the hidden word or phrase inside the box. The cue is typically the *lemma* of the hidden word, i.e., the “dictionary form”: nominative singular forms for nouns, infinitive forms for verbs, etc. The correct word form is known (from the original text) but is hidden from the learner. The learner is expected to answer with the correct form of the word, based on the cue and the surrounding context. For example, in Figure 1, the lemma **aurinkopaneeli** (“solar panel”) is given as a cue. The learner is expected to answer with the correct form, which is partitive plural, “*aurinkopaneeleja*” (as required by the preceding verb). Cloze exercises also include analytic verb forms, for example, the perfect tense form *olen ottanut* (“I have

taken”), will be introduced by the cue which is the lemma of the head verb *ottaa* (“to take”).

2. *Multiple-choice*: Revita gives the learner several options, including the correct answer and distractors. Generating good distractors is an open problem in NLP; the challenge is that they must be incorrect alternatives for the surrounding context, and yet they must not be too easily recognizable as being incorrect.
3. *Listening exercises*: Revita uses state-of-the-art text-to-speech (TTS) models, to play a short spoken fragment, and the learner is expected to enter the forms she heard. This type of exercise is useful for developing aural comprehension. The fragment is played *with its surrounding context*, so part of the challenge is to segment the audio signal into parts and identify the parts that form the answer in the exercise.
4. *Pronunciation exercises*: Revita displays the original words in the text, and the learner is expected to pronounce the words correctly. The learner’s audio signal is analyzed by a speech-to-text (STT) model, trained specifically on learner speech (Kurimo et al., 2023).

Figure 1 is an example of what the learner sees in Practice mode. The second snippet (of five snippets, as indicated by the top progress bar) is active, containing three exercises: two clozes— halukas and aurinkopaneeli —and one multiple-choice (drop-down menu). The current exercise is *aurinkopaneeli*, and the hint box shows the hints / feedback that the learner has requested and received so far for completing the cloze exercise (see the following Subsection).

Above it is the first snippet, which shows previously studied text with previously given answers: green marks correct answers; blue indicates incorrect answers. Additional decorations (circles, underlining, etc.) indicate to the learner various connections and dependencies among the words and phrases in the sentence to help with the understanding of the relationships in the context.

Graduated feedback

Graduated feedback follows the foundational didactic principles of Dynamic Assessment in second language teaching (Poehner, 2008). The learner can request help while attempting to answer an exercise. Requesting help indicates that the learner has not yet mastered some relevant constructs sufficiently to answer the exercise independently; it is likely to lie within the learner's ZPD.

Revita contains a *feedback* component, which is part of the Instruction model. Feedback implements Dynamic Assessment. It is part of the assessment system in the sense that all hints given to the learner are saved as part of the learner's history (Student Model)---they inform the selection of future exercises by adjusting Revita's notion of what the learner has/has not mastered. It is dynamic in the sense that (1) it is dependent on the learner's (incorrect) answer, the context of the exercise and all related constructs; and (2) it should be instructive: using the feedback, the intelligent tutor gradually guides the learner toward the correct answer, rather than simply telling the learner that a response is "right/wrong." Therefore, the sequence of feedback hints is *ordered*, so they become gradually more specific, starting from very general hints, referring to the surrounding context or word paradigms, and then on to specific grammatical features that are required to be present in the correct answer. In the green box in Figure 1, four of the available hints are already "used up" or revealed, and one is remaining.³

Feedback given to the learner can be stratified into the following categories, presented here in order, from the most general to the most specific:

- *Validity*: Feedback that indicates that the answer was spelled incorrectly. Revita analyzes the learner's answer by using morphological analyzers. If

³ An additional challenge is that different classes of high-stakes learners require different kinds of feedback. L2 learners who are university students, and who have a linguistics background, are well prepared to receive feedback in the form of technical linguistic terms. However, a large proportion of our learners are migrants, who need to learn the language for day-to-day functioning—employment, professional examinations, etc. For them complex linguistic terminology is a heavy unnecessary burden, which must be avoided in the teaching process.

the learner's answer is rejected by the analyzer, the intelligent tutor gives the feedback that the answer is not valid. At present, Revita does not attempt to distinguish a "typo"—an accidental misspelling on the part of the learner—from an answer that is incorrect due to a more fundamental lack of knowledge or understanding of some underlying grammatical concept. This distinction is usually too difficult to establish, even for the human teacher.

- *Construct*: Feedback that is attached to a construct. For example, for the construct "Complex pronoun" (such as analytic pronouns in Russian), the feedback may be: "*This is a complex pronoun.*" These hints are generated at the stage when the construct is linked to the text, while the text is processed initially.
- *Phrasal agreement*: Feedback giving hints about the surrounding context of the exercise. For example, in a Finnish noun phrase, the feedback may be: "*This word should agree with other words in this phrase.*" In addition to this textual hint, Revita will give visual hints by circling the entire relevant phrase in red, to indicate the phrase structure. For example (in German), a hint "*Use past perfect tense*" will be attached to the phrasal verb, e.g., "*wäre gekommen.*"
- *Construction*: Feedback that indicates government or more complex syntactic constructions. For example, for an object of a verb that governs the partitive case, the feedback sequence may be: 1) "*This is the object of the verb 'xyz'.*", reminding the learner to try to think of what this verb requires of its arguments, and then 2) "*The verb 'xyz' requires its object to be in partitive case.*", revealing the information about the case of the argument required by the verb, but not the actual answer. Similarly to phrasal agreement, Revita gives visual hints and underlines the government relations or constructions in the text.
- *Morphological features*: Feedback that gives the more specific hints about the grammatical features required in the answer. First, the system uses

morphological analyzers and syntactic parsers to obtain the grammatical features of the *learner's answer*. Then it compares these features to those of the correct answer. More than one feature may be incorrect in the learner's answer. Revita gives hints about the incorrect features in order of increasing specificity, according to a language-specific hierarchy.. Figure 1 shows how the hints become more specific as the learner proceeds, until she finds the correct answer (or exceeds the maximum number of attempts).

- *Language-specific feature*: While all of the preceding types of feedback are applicable across many languages, this category of feedback is specific to one particular language. For example, Finnish has complex morphological processes such as consonant gradation and vowel harmony. Sometimes the learner may know what is expected in the answer, but may make a mistake in consonant gradation. In such situations, Revita will give feedback about the consonant gradation, rather than simply telling the learner that the answer is *invalid*. Note, this is one example of “validity” errors (presented above), where the system *does* attempt to give the learner the benefit of the doubt and to interpret possible spelling errors as the gaps in the understanding of particular linguistic concepts.⁴

Revita implements this variety of feedback types and makes the feedback responsive to the learner's answer and the context of the exercise; this is achieved using a language-specific feedback hierarchy, implemented as a set of rules. This potentially limits the “friendliness” of the feedback and risks repetitive feedback messages. To address this problem, we supplement the rule-based feedback system with a chat-bot powered by large language models (LLMs) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). The learner can ask *arbitrary* questions related to the exercise. The chat-bot returns a personalized answer and

⁴ Examples: in Finnish, the lemma *vetää* (“to pull”), when inflected in the first person present: *vedän* (“I pull”) exhibits the consonant “d”, which is the weak grade of “t”. If the learner answers **vetän*, we can with a high degree of certainty attribute this to a mistake in consonant gradation, rather than a random typo.

explanations for each question, and RAG ensures that the answer is relevant and of high quality. The RAG chat-bot combines the benefits of both worlds: the quality of the messages from our feedback hierarchy, and the fluency and flexibility from advanced LLMs.

Intelligent components

In this section we introduce some of the challenges involving artificial intelligence (AI) in the language teaching process, and how these challenges are addressed by current research. From the preceding sections, it should be clear that the generation of a wide variety of high-quality exercises and feedback based on arbitrary text builds upon an array of analytic NLP tools, ranging from morphological analysis and parsing to generation of distractors, translation, etc. These NLP tools are not perfect; they have their own challenges, and each corresponds to its own area of research in NLP. Thus, improvements in the accuracy and quality of these analytic NLP tools translate directly into improvements in the quality of training that Revita can deliver. The range of NLP tools employed in the Revita workflows is quite wide. In this section we briefly review our work on the areas of grammatical error correction; LLM probing, in particular for government and verbal aspect; and learner analytics and learner modeling.

Grammatical error correction

In the “standard” way of generating learning exercises based on arbitrary texts, we assume that the correct answer is the form that was found in the original text, and any other form that the learner enters is an incorrect answer. However, in a substantial proportion of instances, this assumption is false: in fact, more than one answer is possible given the cue (lemma) and the context. One simple example is the number feature: consider the sentence *Muumikirjoista on tehty lukuisia näytelmiä ja ooppera* (“Moomin books were made into famous shows and opera.”) Here the expected answer is “an opera” (singular), but without world knowledge (i.e., that in fact only one opera was made), the plural is also a grammatically acceptable answer in the immediate context. These are called “alternative-correct”

answers, and according to our measures, they arise in about 15% of all exercises, depending on the language (Katinskaia et al., 2019).

This requires the system to be able to determine whether the learner's answer—when it differs from the *expected* answer—is also grammatically acceptable given the context. This is the domain of Grammatical Error Detection and Grammatical Error Correction (GED/GEC). Our work on this subject investigates how well LLMs are suited for the task, and how their shortcomings can be mitigated (Katinskaia & Yangarber, 2021, 2023, 2024a).

By analyzing the corpus of learner answers that Revita collects (Katinskaia et al., 2020; Katinskaia et al., 2022), we discover that the problem is especially important because the more advanced learners are more likely to enter alternative-correct answers, and they are the learners who require more precise feedback.

We compare the performance of pre-trained LLMs with fine-tuned LLMs on the task of GEC. Because fine-tuning requires large amounts of training data, we also devise methods for generating *synthetic* learner data to supplement the data collected from real learners. Current results show that at present fine-tuned models perform better, but there is still quite a lot of room for improvement, since performance is still on the order of 80% accuracy.

Probing LLMs for government

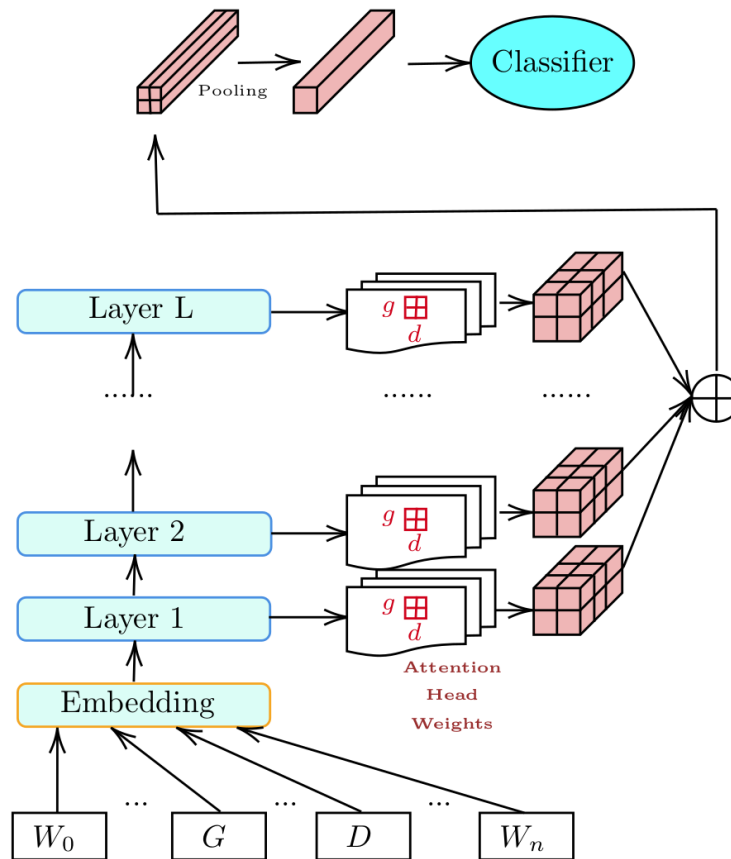


Figure 2. Weights of transformer attention heads used as input to government classifier

We mentioned earlier that the Domain model needs to know about many constructions and, in particular, government is one important type of construction. Originally government relations were entered as rules into the system by consulting external resources, such as government dictionaries, which are available in print for some languages. This allowed us to add approximately 1200 relations for Finnish, 2000 for Russian, and 500 for Italian. At the same time, everyone is aware that modern neural LLMs are capable of generating extremely fluent language, which means that these LLMs certainly must encode knowledge about many government relations—otherwise, they would make mistakes in government and produce obvious disfluencies. We investigated, for several languages: (1) where inside the LLMs this knowledge is stored, and (2) whether finding the source of this knowledge would enable us to substantially expand the base of government

relations available to the Domain model (Hou et al., 2024). The technique is “probing” the attention heads of transformer LLMs. The probing architecture is illustrated in Figure 2. The following account is rather technical and readers without background knowledge of the architecture of transformer neural networks may prefer to skip to the final paragraph of this section.

We want to train a classifier that, for any pair of words (G,D), will learn to predict with high accuracy whether G governs its dependent D or not. The classifier observes the weights in the attention heads of a BERT model in each transformer layer. Each attention head in each layer assigns some weight to the connection between G and D. If we can train a classifier that can predict the presence of government from these weights, that means that we have localized where inside the LLM the information about government is stored, and we will be able to use that information to extract many more government relations.

In Figure 2, at the bottom we start with embeddings for all the input words, including the governor verb (G) and its dependent argument (D). The dependency relation is established using a dependency parser. Then on each layer l , for each attention head h , we extract its weights for (G,D), and concatenate these weights into a vector. The vector will contain $L \cdot A$ elements, where L is the number of transformer layers and A is the number of attention heads per layer.⁵ We use the vector as a one-dimensional representation for the governor-argument pair, which encodes the syntactic relations between the governor and the argument. This vector finally becomes the input to the probing classifier.

⁵ In the default BERT-base settings, $L = A = 12$.

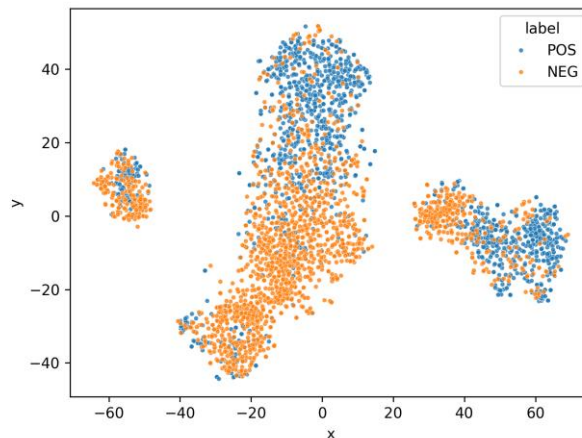


Figure 3. t-SNE visualization of positive vs. negative instances of government (Finnish)

We collect a large data set (for training and testing) of sentences with pairs of words that have a government relation between them (positive instances) vs. pairs that are not government (negative). To examine the difference between the attention vectors of the positive vs. negative instances, we use t-SNE dimensionality reduction implemented in the ScikitLearn package with default parameters (van der Maaten and Hinton 2008). Figure 3 shows how t-SNE projects the high-dimensional attention vectors of the probing classifiers onto the 2-D plane. We plot all instances from the test data, for Finnish data. We can observe some strikingly well-defined clusters of the positive vs. the negative instances. This separability helps explain the very high accuracy — 83–89% — achieved by all probes on government prediction for all languages tested. Additional details are in (Hou et al., 2024).⁶

⁶ We tested four different types of probing classifiers: logistic regression, multi-layer perceptron (MLP) with one fully-connected layer, MLP with two fully-connected layers, and Random Forest.

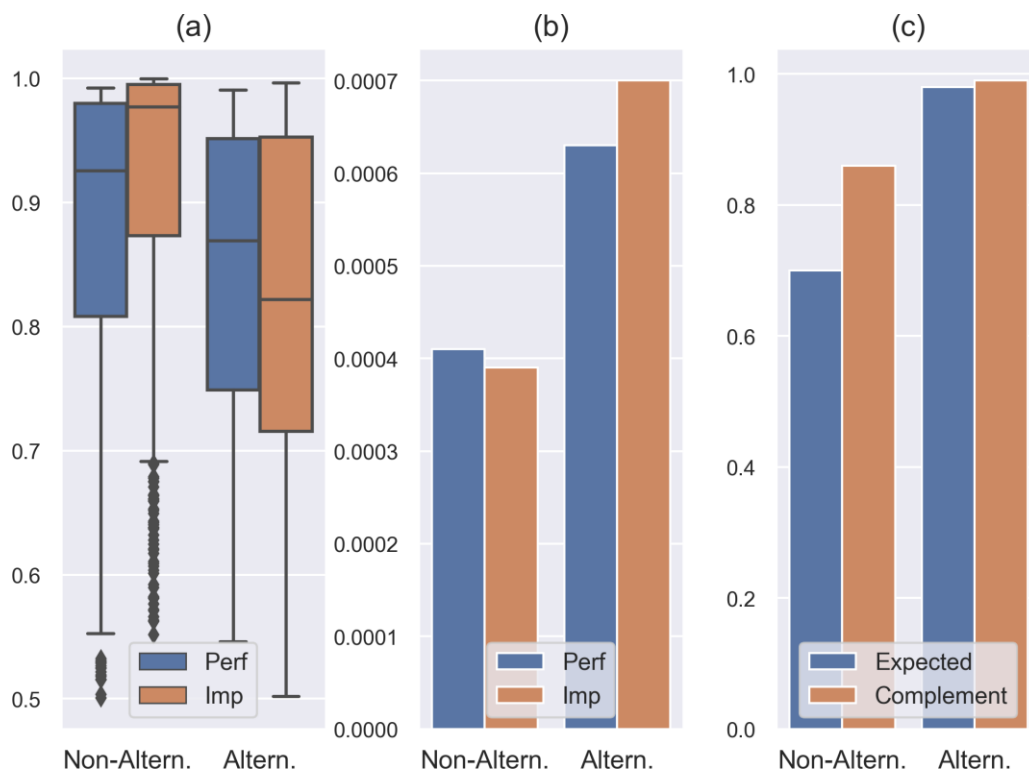


Figure 4. (a): Scores assigned to perfective and imperfective classes. (b): Variance of scores assigned to perfective and imperfective classes. (c): Percentage of contexts where target verb does not contain hints with semantics of bounded vs. unbounded action. *Altern*: alternative contexts, *Non-Altern*: non-alternative. *Expected*: contexts where model prefers expected aspect; *Complement*: contexts where it prefers the other (complementary) aspect. Model is BERT-large, fine-tuned final layer.

Probing LLMs for verbal aspect

We further pursue related questions, about how LLMs encode important grammatical categories, and how this knowledge can be leveraged in L2 teaching. (Katinskaia & Yangarber, 2024b). One such category is verbal *aspect* in Russian, which is one of the most difficult ones for L2 learners; non-native speakers make errors in the choice of aspect (perfective vs. imperfective), even when they reach very advanced levels of expertise (Bar-Shalom & Zaretsky, 2008).

Revita generates multiple-choice aspect exercises for learners, with the expected answer being the one given in the original text. In certain contexts, both aspects may be suitable (possibly with a slight difference in meaning). In such contexts—*alternative-correct* contexts—it is crucial not to mislead the learner by giving

feedback that the alternative answer that was not found in text is *incorrect*.⁷

We collected a corpus of instances of perfective and imperfective verbs, in alternative and non-alternative contexts.⁸ We investigate these questions: Do transformer LLMs encode the category of aspect, and if so, how? Does the encoding of aspect in these models correspond to linguistic theory? Is aspect encoded in alternative contexts differently from non-alternative contexts—the latter having special importance for teaching. We probe several transformer models, similarly to the probing approach above, to investigate these questions.

Our probing methods confirm that BERT and RoBERTa models do encode aspect, and most of the encoding is situated in their *final* layers, where other kinds of higher-level semantic information are currently known to be encoded. We perform “*causal probing*”, by applying “interventions” in the sentence semantics—modifying the “boundedness” feature of the verb—with results consistent with linguistic theory: imperfective verbs typically describe unbounded actions, whereas perfective verbs describe bounded actions. When we fine-tune only the final layers of BERT for aspect prediction, prediction performance improves.

In the alternative-correct contexts, we observe differences in model behavior. We find that both pre-trained and fine-tuned transformer models exhibit *high uncertainty* regarding aspect preference in alternative contexts, where both aspect forms are acceptable. Alternative contexts are also more sensitive to causal intervention in the semantics of boundedness. Such contexts usually lack explicit hints about the action’s boundedness, which makes both humans and language models uncertain about the correct choice of aspect.

The main results are summarized in Figure 4. We run experiments on two types of contexts: where only one aspect form is valid (non-alternative) vs. those where both

⁷ Another option may be to avoid such contexts and not offer exercises in alternative-correct contexts at all. However, this still requires us to have a model that will identify such contexts reliably, so they can be avoided.

⁸ Released to the research community at: github.com/RevitaAI/AspectProbing

forms are valid (alternative context). We find that, in line with grammatical theory, information about the boundedness of the action is encoded in BERT's context representation and affects the choice of aspect.

Using this finding, we fine-tune the last layers of BERT for aspect prediction, which leads to more effective and faster tuning. We found that the fine-tuned BERT can detect contexts with multiple valid aspects by the high uncertainty of its predictions. In Plot (a), we see that the mean of the prediction accuracy (the bars in the middle of the boxes) are much higher for non-alternative contexts, for both perfective and imperfective aspect (the two bars on the left), than for alternative contexts (the two bars on the right). A closer look at alternative contexts shows that indeed they do not have enough cues that can help the model (or a learner) to decide which aspect to use. Plot (b) in Figure 4 shows that the model has much lower confidence for alternative contexts: in contexts where both aspects are acceptable, LLMs also cannot make a preference for predicting a particular aspect.

Checking how many target contexts have no cue words in the context, we see from Plot (c) that almost 100% of verbs in alternative contexts have no cues (the two bars on the right). Looking closer at non-alternative contexts, more of the errors made by BERT occur in contexts without cues. Therefore, these contexts can be avoided when the AI tutor generates exercises. The experiments confirm that probing yields not only theoretical insights into the nature of the phenomenon of aspect, but it also has important practical applications for language learning.

Learner modeling and exercise sampling

Revita keeps track of all learner answers and every requested hint requested during each exercise, while allowing the learner to attempt each exercise multiple times. The collected information about learner performance can be used to model learner proficiency and the difficulty of the exercises and linguistic grammatical topics (Vu et al., 2025). To model learner skills and exercise difficulty, we adapt a form of Item Response Theory (IRT) (Embretson & Reise, 2013; van der Linden & Hambleton, 2013). Originally developed for psychometric testing, IRT has become

widely used in educational assessment (Klinkenberg et al., 2011). We briefly describe the main challenges in applying IRT in our language teaching setting.

IRT is an information-theoretic framework used to estimate the probability that a subject will be able to respond to a test item correctly, given the subject's ability and the item difficulty. In our adapted setting, the subject is the learner, and the *Item* refers to a grammatical topic that the learner has encountered during exercises. Traditional uses of IRT usually have a clear definition of an *item*, and a clear credit standard: the classic example is a *test question*, e.g., in mathematics, which is unambiguous and *dichotomous*: there is a clear judgement of whether the answer is correct or wrong.

An important challenge in our application setting is that language constructs cannot be directly treated in the same way as test items in other learning domains. It is challenging to determine the credit and penalty for constructs based on an answer that the student supplies to an exercise, because it is the exercises that are answered correctly or incorrectly, and the links from exercises to constructs are *many-to-many*. To enable treating constructs as items in IRT, for each exercise and each attempt, Revita analyzes the answer and the requested hints, and calculates *credits and penalties* for all constructs that are attached to the exercise. We need to distinguish here between selected-response language exercises, like multiple-choice, (which are scored dichotomously), and supplied-response exercises, like gap-filling, where partial correctness can occur. Partial correctness of answers is taken into account, e.g., if the answer used the correct tense but wrong number, only number will be penalized, and tense will receive credit.

To pre-process (“clean”) the learner data, we consider the minimum number of exercises that student S has completed. We denote min_{exer} as the minimum number of exercises. The number of unique students after filtering used in this analysis was 1,639. These students have performed a total of over 470K exercise attempts. We currently model over 200 grammatical constructs in total (for Russian). For the modeling and assessment, we group them into a smaller number of learning topics

that correspond to instructional goals. A topic is marked as correct only if all of its underlying constructs are answered correctly. Approximately 80 topics are used as items whose difficulty we estimate in IRT. In addition, we incorporate the probability of a lucky guess dynamically, based on the type of exercise. For multiple-choice questions, this probability is tied to the number of distractors. For gap-filling exercises, the probability of guessing is relatively lower. Based on the learner data for these students and grammatical topics, the IRT model automatically learns proficiency scores for all students, and difficulty scores for all topics by optimising a global probabilistic objective.

We also have over 50 of these students who have done over 100 exercises each, and who received independent assessment on the CEFR scale from their teachers. This batch of students is used for a validation of the quality of the assessment that Revita provides based on exercises and constructs. The results from IRT modeling of student ability are shown in Figure 5. The figure shows a very good correlation between the teacher's estimate and the IRT estimate of student ability. The correlation coefficient is $\rho = 0.62$, which indicates a good positive correlation, and good agreement between the teachers' assessment and IRT estimates. This suggests that we can estimate student ability using IRT and linguistic topics.

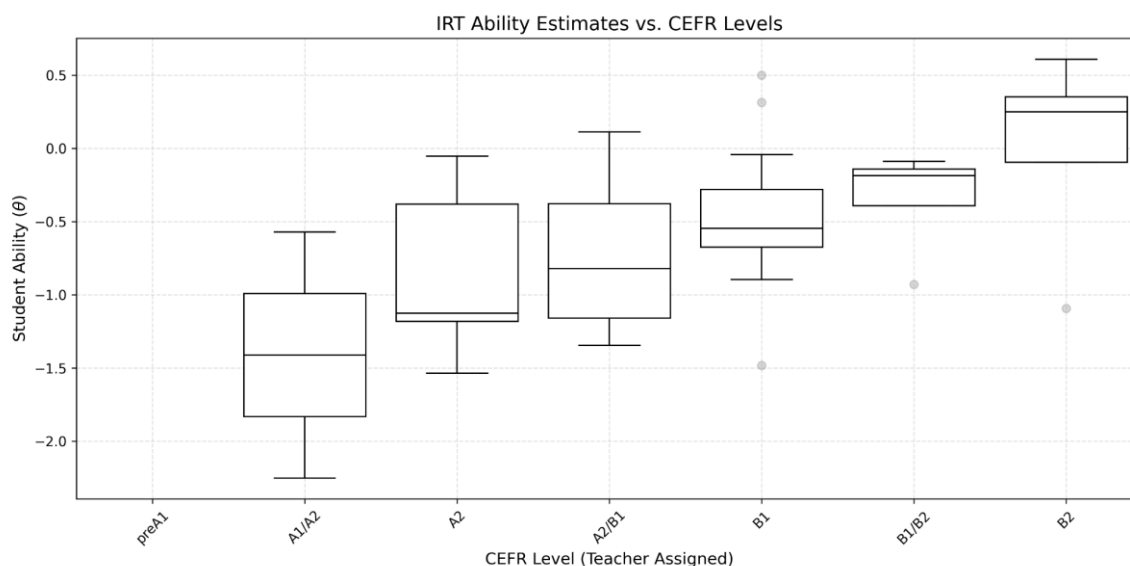


Figure 5. Estimation of student ability via IRT, based on learner data. X-axis: grades given by teachers. Y-axis: ability estimated by IRT.

Conclusions and future work

This paper presents an in-depth overview of the features of the Revita language learning system. A core component of Revita is the Domain model, embodied in a system of linguistic constructs. This system of constructs underlies and supports all aspects of the learning experience in Revita. It supports the generation of a wide variety of exercises based on arbitrary texts, and generation of feedback and hints for the learner. The Student model stores the history of all activities of the learners—all exercises done—in order to support continual assessment of learner performance. It also supports the modelling of learner skills more accurately to provide informative progress analytics, and to offer exercises that are most appropriate for the learner's current level.

The research on probing LLMs for knowledge about linguistic constructions has so far focused exclusively on government relations. In future work we plan to probe LLMs for knowledge about broader kinds of constructions in order to offer richer exercises to the learner. We have presented results from pilot studies with Finnish and Russian L2 learners using the Domain Model. In future work, we plan to improve the Domain Model by adding more information about the *interactions* and dependencies among the constructs, which will enable the creation of more intelligent learning paths. We also plan to add new types of activities, e.g., pronunciation exercises, and additional languages.

Author disclosures

The authors declare no conflicts of interest.


This work was supported in part by the Research Council of Finland, Project “*Know-AI: What do language models know and when do they know it*” (Grant 359285), and by BusinessFinland: Agency for Technology and Innovation, Project “*Easy Language for accessible workplace communication*” (Grant 4173/31/2024).

All authors contributed to the conceptualization, formal analysis, validation and writing. A. Vu, J. Hou and A. Katinskaia contributed to development of software

and visualization. R. Yangarber contributed to supervision.

ORCID iDS

Jue Hou  <https://orcid.org/0000-0001-9404-2022>

Anh-Duc Vu  <https://orcid.org/0009-0005-1186-0510>

Anisia Katinskaia  <https://orcid.org/0000-0003-4137-6760>

Roman Yangarber  <https://orcid.org/0000-0001-5264-9870>

References

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215. <https://doi.org/10.1037/0033-295x.84.2.191>
- Bar-Shalom, E. G., & Zaretsky, E. (2008). Selective attrition in Russian-English bilingual children: Preservation of grammatical aspect. *International Journal of Bilingualism*, 12(4), 281–302. <https://doi.org/10.1177/1367006908098572>
- Boas, H. C. (Ed.) (2022). *Directions for pedagogical construction grammar: Learning and teaching (with) constructions*. De Gruyter. <https://doi.org/10.1515/9783110746723>
- Bodnar, S., Cucchiarini, C., Penning de Vries, B., Strik, H., & van Hout, R. (2017). Learner affect in computerised L2 oral grammar practice with corrective feedback. *Computer Assisted Language Learning*, 30(3-4), 223–246. <https://doi.org/10.1080/09588221.2017.1302964>
- Derakhshan, A., & Khodabakhshzadeh, H. (2011). Why CALL why not MALL: An in-depth review of text-message vocabulary learning. *Theory and Practice in Language Studies*, 1(9), 1150–1159. <https://doi.org/10.4304/tpls.1.9.1150-1159>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*.

- Psychology Press. <https://doi.org/10.4324/9781410605269>
- Endresen, A., Zhukova, V., Bjørgve, E., Demidova, D., Kalanova, N., Butenko, Z., Lonshakov, G., & Lavén, D. H. (2022). Construxercise! implementation of a construction-based approach to language pedagogy. *Russian Language Journal*, 72(1/2), 47–70. <https://doi.org/10.70163/0036-0252.1283>
- Fillmore, C. J. (1988). The mechanisms of construction grammar. *Annual Meeting of the Berkeley Linguistics Society*, 35–55.
- Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199268511.001.0001>
- Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1), 70–105. <https://doi.org/10.1080/09588221.2012.700315>
- Gray, L. (2008). Effective practice with e-portfolios. *Higher Education Funding Council for England, JISC, Bristol*.
- Herbst, T. (2016). Foreign language learning is construction learning—what else? Moving towards pedagogical construction grammar. In S. de Knop & G. Gilquin (Eds.), *Applied Construction Grammar* (pp. 56–96). De Gruyter.
<https://doi.org/10.1515/9783110458268-003>
- Hou, J., Katinskaia, A., Kotilainen, L., Trangcasanchai, S., Vu, A.-D., & Yangarber, R. (2024). What do transformers know about government? *Proceedings of LREC-COLING: the Joint International Conference on Computational Linguistics, and Language Resources and Evaluation*. 17459–17472.
<http://www.lrec-conf.org/proceedings/lrec-coling-2024/index.html>
- Huhta, A., Harsch, C., Leontjev, D., & Nieminen, L. (2024). *The diagnosis of writing in a second or foreign language*. Routledge.
- Janda, L. A., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics*, 34(1), 161–173.

<https://doi.org/10.1075/bjl.00043.jan>

- Kacetyl, J., & Klímová, B. (2019). Use of smartphone applications in English language learning—a challenge for foreign language education. *Education Sciences*, 9(3), 179. <https://doi.org/10.3390/educsci9030179>
- Katinskaia, A., Hou, J., Vu, A.-D., & Yangarber, R. (2023). Linguistic constructs represent the domain model in intelligent language tutoring. *Proceedings of EACL: 17th Conference of the European Chapter of the Association for Computational Linguistics*. <https://aclanthology.org/events/eacl-2023/#2023eacl-main>
- Katinskaia, A., Ivanova, S., & Yangarber, R. (2019). Multiple admissibility: Judging grammaticality using unlabeled data in language learning. *Proceedings of BSNLP: the 7th Workshop on Balto-Slavic Natural Language Processing, ACL: 56th Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/w19-3702>
- Katinskaia, A., Ivanova, S., & Yangarber, R. (2020). Toward a paradigm shift in collection of learner corpora. *Proceedings of LREC: 12th Language Resources and Evaluation Conference*, 386–391. <http://www.lrec-conf.org/proceedings/lrec2020/index.html>
- Katinskaia, A., Lebedeva, M., Hou, J., & Yangarber, R. (2022). Semi-automatically annotated learner corpus for Russian. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 832–839. <http://www.lrec-conf.org/proceedings/lrec2022/index.html>
- Katinskaia, A., Nouri, J., & Yangarber, R. (2018). Revita: A language-learning platform at the intersection of ITS and CALL. *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*. <http://www.lrec-conf.org/proceedings/lrec2018/index.html>
- Katinskaia, A., & Yangarber, R. (2021). Assessing grammatical correctness in language learning. *Proceedings of BEA: 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 135–146.
- Katinskaia, A., & Yangarber, R. (2023). Grammatical error correction for

- sentence-level assessment in language learning. *BEA: 18th Workshop on Innovative Use of NLP for Building Educational Applications*, 488–502. <https://doi.org/10.18653/v1/2023.bea-1.41>
- Katinskaia, A., & Yangarber, R. (2024a). GPT-3.5 for grammatical error correction. *Proceedings of LREC-COLING: the Joint International Conference on Computational Linguistics, and Language Resources and Evaluation*, 7831–7843. <http://www.lrec-conf.org/proceedings/lrec-coling-2024/index.html>
- Katinskaia, A., & Yangarber, R. (2024b). Probing the category of verbal aspect in transformer language models. *NAACL: Findings of the Association for Computational Linguistics*, 3347–3366. <https://doi.org/10.18653/v1/2024.findings-naacl.212>
- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on-the-fly ability and difficulty estimation. *Computers & Education*, 57(2), 1813–1824. <https://doi.org/10.1016/j.compedu.2011.02.003>
- Kopotev, M. (2012). Karttu: Results of language testing in schools and universities. *Formation and assessment of communicative competency of bilinguals in bilingual education*, 312–339.
- Kurimo, M., Getman, Y., Voskoboinik, E., Al-Ghezi, R., Kallio, H., Kuronen, M., von Zansen, A., Hilden, R., Kronholm, S., Huhta, A., & Linden, K. (2023). New data, benchmark and baseline for L2 speaking assessment for low-resource languages. *Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 166–170. <https://doi.org/10.21437/slate.2023-32>
- Leontjev, D., Poehner, M. E., & Huhta, A. (Eds.) (2025). *Dynamic and diagnostic language assessment: Learning across frameworks to support second/foreign language education*. De Gruyter. <https://doi.org/10.1515/9783111233918>
- Levy, M. (1997). *Computer-assisted language learning: Context and*

- conceptualization*. Oxford University Press.
<https://doi.org/10.1093/oso/9780198236320.001.0001>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
<https://doi.org/10.48550/arXiv.2005.11401>
- Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Springer.
- Poehner, M. E., & Leontjev, D. (2020). To correct or to cooperate: Mediation processes and L2 development. *Language Teaching Research*, 24(3), 295–316. <https://doi.org/10.1177/1362168818783212>
- Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning*, 31(1-2), 72–89.
<https://doi.org/10.1080/09588221.2017.1382536>
- Rosell-Aguilar, F. (2018). Autonomous language learning through a mobile application: A user evaluation of the busuu app. *Computer Assisted Language Learning*, 31(8), 854–881.
<https://doi.org/10.1080/09588221.2018.1456465>
- van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Nature. <https://doi.org/10.1007/978-1-4757-2691-6>
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.
<http://jmlr.org/papers/v9/vandermaaten08a.html>
- Volodina, E., Pilán, I., Borin, L., & Tiedemann, T. L. (2014). A flexible language learning platform based on language resources and Web services. *LREC: Ninth International Conference on Language Resources and Evaluation* 973–3978. <http://www.lrec-conf.org/proceedings/lrec2014/index.html>

- Vu, A.-D., Hou, J., Katinskaia, A., Sheu, C.-F., & Yangarber, R. (2025). A Bayesian approach to inferring prerequisite structures and topic difficulty in language learning. *BEA: Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
<https://doi.org/10.18653/v1/2025.bea-1.53>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, Ed.). Harvard University Press.
<https://doi.org/10.2307/j.ctvjf9vz4>
- Vygotsky, L. S. (2012). *Thought and language* (Revised edn). MIT Press.
- White, C., & Reinders, H. (2010). *The theory and practice of technology in materials development and task design*. Cambridge University Press.
- Yeh, H.-C., & Lai, W.-Y. (2019). Speaking progress and meaning negotiation processes in synchronous online tutoring. *System*, 81, 179–191.
<https://doi.org/10.1016/j.system.2019.01.001>
- Zhang, R., & Zou, D. (2022). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, 35(4), 696–742.
<https://doi.org/10.1080/09588221.2020.1744666>