Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 560-569
https://doi.org/10.22364/bjmc.2024.12.4.17

ELLE – Estonian Language Learning and Analysis Environment

Kais ALLKIVI, Pille ESLON, Taavi KAMARIK, Karina KERT, Jaagup KIPPAR, Harli KODASMA, Silvia MAINE, Kaisa NORAK

Tallinn University School of Digital Technologies, Tallinn, Estonia

kais@tlu.ee, pille.eslon@tlu.ee, taavi.kamarik@tlu.ee, karina.kert@tlu.ee, jaagup@tlu.ee, harli.kodasma@tlu.ee, silvia.maine@tlu.ee, kaisa.norak@gmail.com

ORCID 0000-0003-3975-5104, ORCID 0009-0003-8546-5076, ORCID 0009-0001-5088-7442, ORCID 0009-0002-0601-8079, ORCID 0009-0005-5949-2366, ORCID 0009-0004-8676-1907, ORCID 0009-0004-2460-6370, ORCID 0000-0002-4806-2660

Abstract. Text corpora provide authentic material for language instruction and an insight into the development of learner language use. This pedagogical potential can be enhanced by accompanying user-friendly text analysis tools designed for researchers, teachers, and learners alike. We introduce the Estonian Language Learning and Analysis Environment (ELLE) that combines a growing corpus of Estonian learner writings (Estonian Interlanguage Corpus – EIC) with various applications for linguistic analysis and automated text evaluation. The toolkit can be employed to analyse EIC and other corpora as well as study materials or users' own texts. ELLE's ongoing implementation follows a prototype which was created using interaction and participatory design methods, involving members of different target groups. The paper outlines the system architecture and presents the functionalities of each tool, highlighting their unique features compared to alternative web applications.

Keywords: computer-assisted language learning, corpus linguistics, data-driven learning, learner corpus, natural language processing, participatory design, text analysis tools

1. Conception and design

Estonian Language Learning and Analysis Environment (ELLE)[1] integrates an electronic collection of Estonian language learner writings with a suite of online tools for text analysis and automated writing evaluation. Tailored for a diverse user base—including non-native and native learners, educators, researchers, and anyone working with texts—it aims to support individual and classroom learning, monitoring student progress, designing instructional materials and curricula as well as improving professional writing.

ELLE is built on the foundation of the Estonian Interlanguage Corpus (EIC) (Eslon, 2014). Developed in collaboration between philologists, educational technologists, and

[1] https://elle.tlu.ee, source code: https://github.com/centre-for-educational-technology/evkk

IT specialists, EIC serves as a dynamic resource of texts written by learners of Estonian as a second language. It comprises examination writings and texts produced in other language learning situations, representing learners with various proficiency levels and language background. For comparative purposes, EIC includes sub-corpora of native Estonian texts, and Russian native and learner texts. Complementing EIC with a virtual learning environment that offers approachable tools for linguistic analysis fosters both immediate and delayed, or direct and indirect pedagogical application (see Granger, 2009; Römer, 2011). While the corpus and toolkit can be used for learner language research, they can also be accessed by learners to uncover language patterns from authentic texts. Additionally, learners can benefit from analysing their own writings and submit them to the corpus, speeding up the collection of new research data.

Such data-driven learning has been inhibited by the technical skills and training time needed by learners and teachers (see Boulton, 2010). These wider target groups have typically not been involved in the development of corpus user interfaces. To help solve these problems, ELLE's prototype was created through design-based research, iteratively defining the needs and preferences of different user profiles (Norak, 2021; Norak and Põldoja, 2021). Language learners were represented by non-native and native high school and university students but also lifelong learners who use Estonian in their work.

Combining a modified Double Diamond model (Santos Ordóñez et al., 2017) and participatory design methods, the design process (summarised in Figure 1) began with background research, competitive analysis, and user interviews. Based on the interviews, 6 user personas and 13 scenarios were developed and discussed in group design sessions. The feedback guided the creation of low-fidelity paper prototypes and high-fidelity digital prototypes, both tested in concurrent think-aloud sessions. All in all, 49 target users participated in the design phases. The final prototype was accompanied by 98 user stories supporting the development of planned software features.

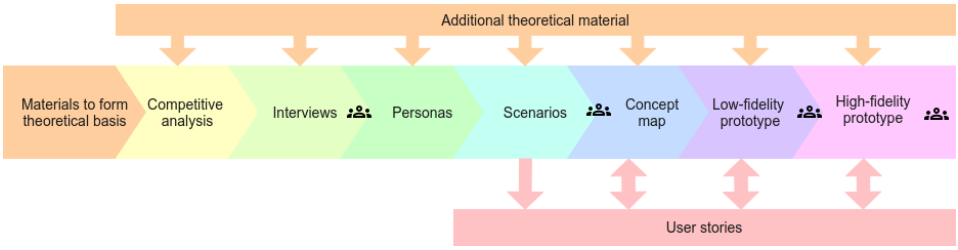

Figure 1. Design process of ELLE

The first version of ELLE was launched in late 2022. Its user interface (UI) is developed on the new technical platform proposed for EIC by Tarvas (2020). Currently, ELLE offers tools for analysing vocabulary and grammatical items used in EIC or any other text material, as well as a writing evaluation tool that assesses the correctness, proficiency level, and complexity of a given text. The environment also features a link collection referencing resources for Estonian language learning and research. Some functionalities are yet to be implemented. These include sharing and creating corpus-based learning materials and interactive exercises, logging in with a user account to save analysis results and get individual feedback and recommendations, and forming study groups to assign tasks to students and assess them.

As there are still few examples of corpus-based language learning tools (e.g., Lärka[2] for Swedish and Write & Improve[3] for English), ELLE can be considered innovative in a broader sense than in Estonian context. The conception of joining corpus research and language learning on one platform may be adopted for other languages.

2. System architecture

ELLE utilises Java and Python back-end applications comprising several modules, a PostgreSQL database, and a React front-end application (UI), as well as other, more minor components. The main architecture can be seen in Figure 2.

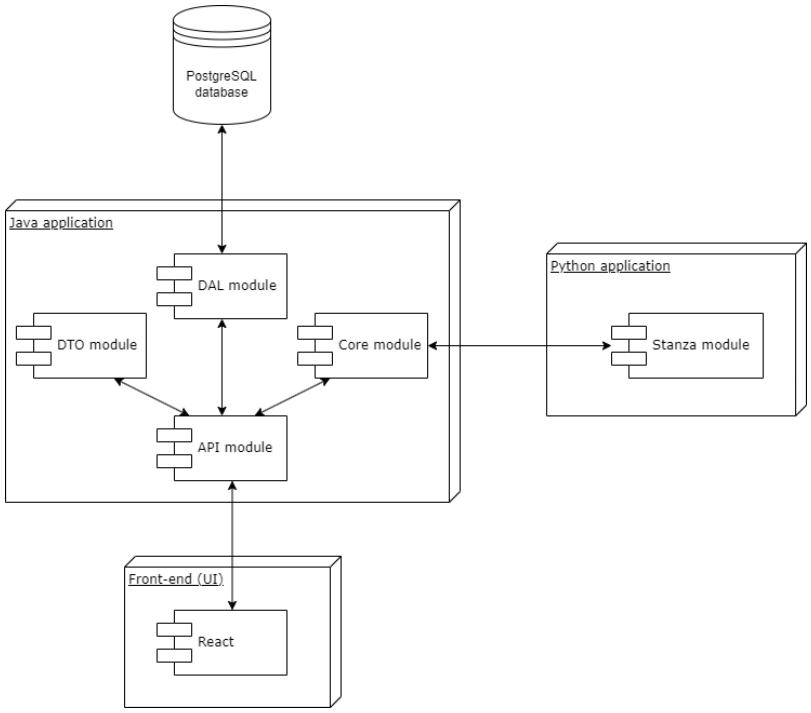

Figure 2. Core architecture of ELLE

While the Java application serves as a mediator responsible for communication between the UI, the database, and all other services, it is itself divided into different modules and takes care for most of the business logic of the tools and applications seen in the UI. Each module has been described in depth by Tarvas (2020) but the key modules to point out are DAL (Data Access Layer) responsible for communication with the database, API (Application Programming Interface) that takes care of requests made by the UI, and

[2] https://spraakbanken.gu.se/larka/, see Alfter et al. (2019)
[3] https://writeandimprove.com, see Yannakoudakis et al. (2018)

Core which links to other servers such as Python applications running various language processing tasks. All these parts are isolated and run as separate Docker containers.

Most ELLE tools rely on Stanza[4], a Python package by Stanford University's Natural Language Processing Group. Based on neural networks, it allows for linguistic analysis of texts, e.g., word and sentence segmentation, lemmatisation, assigning parts of speech (PoS) and grammatical forms. Stanza was chosen to be used due to its higher context sensitivity, more accurate PoS tagging, better adaptability to spelling errors and non-ambiguous output compared to the rule-based tools of the EstNLTK library (Allkivi-Metsoja and Norak, 2021). Stanza has also outperformed the EstBERT and RoBERTa language models in tokenisation and PoS tagging[5]. Other open-source services applied in ELLE are mentioned in the following sections.

3. Text query and analysis

ELLE features five tools that can be used to analyse text corpora as well as single texts. Upon navigating to the page 'Texts & tools', users are prompted to choose language material from EIC or submit their own text(s) before they can apply the following tools: Wordlist, Word in Context, Neighbouring Words, Word Analyser, and Cluster Catcher.

Text query allows to search the EIC collection based on author and text metadata. Depending on the chosen sub-corpus or sub-corpora, the query can be limited by text type, e.g., creative writing, informal/formal letter, or exercise. One sub-corpus contains academic Estonian writings by native and non-native students. Among other features, it is possible to choose the texts' proficiency level (A1–C2), author's age group, level of education, and native language. Users can then select texts from the query result to save them for analysis.

Three interconnected tools—Wordlist, Word in Context, and Neighbouring Words—developed by Kodasma (2023) provide information about word frequency, usage contexts (concordances), and co-occurring words (collocates). Other similar tools are paid software (e.g., Sketch Engine, WordSmith Tools), do not offer Estonian lemmatisation (e.g., AntConc, LancsBox), or can currently be applied on determined corpora without the option of analysing any given text (Estonian Korp[6], collocation search of the balanced subset of Estonian Reference Corpus[7]).

Wordlist compiles a word frequency list based on word forms or base forms (lemmas), grouping the inflected forms of a word. To avoid common words, void of content, from topping the list, users can exclude stop words of their own choice or exploit an existing resource of function words[8]. It is also possible to set a minimum word frequency, barring rare occurrences from the list, or to count capitalised and lowercase words separately. Clicking on the three dots next to each list item enables to view its contexts and collocates, look up its definition in the WordWeb language portal[9], or translate it (we use Google Translate due to simple integration but intend to replace it with open-source software). The wordlist can be sorted alphabetically and downloaded

[4] https://stanfordnlp.github.io/stanza/
[5] See https://github.com/EstSyntax/EstSpaCy
[6] https://korp.keeleressursid.ee/
[7] https://korpused.keeleressursid.ee/clc/
[8] https://datadoi.ee/handle/33/78
[9] https://sonaveeb.ee

as an Excel or CSV file, which applies for all the tools described in this section. Word frequencies can be visualised in the form of a word cloud.

Word in Context views the preceding and following contexts of a word form or all forms of a given base form. Users are required to enter a search word, with an option to adjust case-sensitivity and the number of surrounding words or sentences shown in the table. The tool can also be accessed by choosing a word of interest in the output of Wordlist or Neighbouring Words. The results can be sorted by the left or right context.

Neighbouring Words allows to find words that co-occur more often than expected by chance. Users must enter a search word or choose it in the Wordlist tool. The search can be based on word forms or base forms. The collocation window, i.e., the number of preceding and following words considered neighbouring words, can be 1–5. Statistical methods provided for detecting collocates include LogDice, T-score, and MI-score that yielded different results in comparative experiments (Kodasma, 2023). The highest scoring collocates are shown at the top of the list which can also be sorted by the words' overall frequency or the number of co-occurrences with the search term. Users can view the contexts, definition, and translation of each word. The results can be visualised as an interactive graph displaying the importance, distance, and position of the collocates.

The other two tools combine grammatical and lexical analysis. **Word Analyser** can be applied for exploring texts word by word and viewing various text statistics. Clicking on a text word reveals its syllables, base form, PoS, and grammatical form (see Figure 3), helping the user learn to better understand language structures. Although this learner-oriented feature is available for texts with up to 1,000 words, different frequency tables can also be compiled for larger language material. The first table shows syllable frequencies, their positions in words (beginning, middle, end), and words containing them. It is possible to filter syllables by position, e.g., to analyse word endings. The second table features the base forms of words together with the different word forms represented in text along with their frequencies. Thirdly, grammatical forms are presented with their frequency and occurrences. The latter two tables can be filtered by PoS to focus on a specific word class, e.g., nouns or verbs. Word Analyser was developed throughout two student projects in 2022. A new syllabification solution (Kukke, 2024) has later been implemented, improving response time and accuracy.

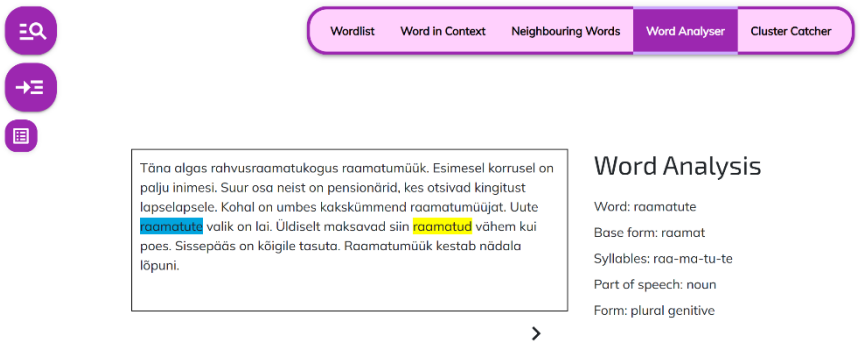

Figure 3. Exploring the use of the word *raamat* 'book' in an A2-level reading task text

Cluster Catcher searches for similar n-grams, i.e., word sequences based on the words' PoS, grammatical form, and/or syntactic function. N-gram groups, or clusters, are presented along with frequency and all examples found in text. This helps to detect common grammatical patterns and related lexical choices. The tool was first developed by Ots (2012) as a desktop application, based on a word order error detector prototype (Matsak et al., 2010), and renewed by Liiva (2022) to be integrated with ELLE. For syntactic analysis, Cluster Catcher uses the VISLCG3Parser from the EstNLTK toolkit[10]. Users can specify n-gram length (1–5 words), the type of analysis (PoS-based, morphological, syntactic), and whether punctuation marks are considered as n-gram components. It is also possible to search for patterns containing a specific grammatical unit. E.g., by defining the properties of the first word, one can analyse its following context. While Sketch Engine allows to find PoS n-grams, Cluster Catcher is the only known tool for extracting more advanced grammatical n-grams.

Word Analyser and Cluster Catcher have undergone initial classroom testing with 12th grade students, each tool in a different high school. After instruction, students solved a sample online exercise and filled in a semi-structured feedback questionnaire.

With Word Analyser, the task was to choose the correct word from a word pair to complete a sentence and explain the choice by detecting the base and grammatical form of each word. The overall impression of the application was positive. Useful features highlighted were the option to insert a search term in any form without necessarily knowing its base form and displaying diverse information about words in one place, which makes it easier to form mental associations and remember irregularities. Students claimed that Word Analyser was easy to use but made some suggestions that have been taken into account in improving the UI (e.g., adding a loading animation).

With Cluster Catcher, students analysed the stylistic features of three text genres: opinion piece, research article, and prose fiction. In comparison, they looked at common linguistic patterns of their own argumentative writing. According to student responses, the tool helped to detect genre differences fast and efficiently, while it was lacking clear instructions and examples that would help to better understand and interpret the results. We have therefore updated the UI texts and produced video tutorials for all tools.

Both student groups reported that using a hands-on online tool in the language classroom was an interesting and engaging alternative to usual study activities, and they would occasionally continue to use Word Analyser or Cluster Catcher in their studies.

4. Writing evaluation

ELLE also integrates the Writing Evaluator, which offers automated feedback on user submissions by evaluating four linguistic dimensions. The UI that connects several back-end services has been developed by Kamarik (2024). Firstly, the tool assesses textual **correctness** using two distinct correction models created in collaboration between the University of Tartu and Tallinn University. The first model, a statistical spell-checker, identifies spelling errors relying on word contexts (Allkivi-Metsoja and Kippar, 2023). The second is a neural machine translation based grammatical error correction model detecting sentence-level errors (Luhtaru et al., 2024). Writing Evaluator highlights potential linguistic errors and provides the likeliest correction candidate that can be

[10] https://github.com/estnltk/estnltk

accepted or rejected (see Figure 4). Corrections are categorised into five types: lexical (replaced, missing, or unnecessary word), punctuation (incorrect, missing, or unnecessary punctuation), compounding, and word order corrections as well as a category for multiple co-occurring changes, labelled 'other correction'.

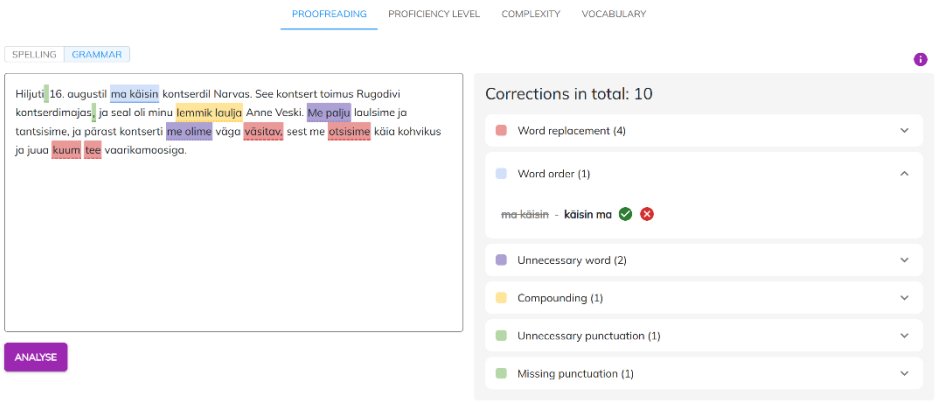

Figure 4. Analysing the correctness of an A2-level proficiency examination writing

Secondly, supervised machine learning models are employed to estimate a second language learner text's probable **proficiency level** on a scale from A2 to C1, which are nationally tested levels. In addition to the overall proficiency score, separate scores are provided for general text complexity (related to text, sentence, and word length), grammatical complexity (related to PoS and grammatical form frequency, and the diversity of grammatical forms), and lexical complexity (related to the diversity, range, and abstractness of vocabulary). The development of the assessment models has been informed by comparative linguistic analysis of A2–C1-level examination writings (Allkivi-Metsoja, 2021, 2022).

Thirdly, text **complexity** is rated on a scale from easy to difficult. This assessment uses the LIX, SMOG, and Flesch-Kincaid Grade Level indices derived from word and sentence length measures. The tool also analyses the noun-to-verb ratio to gauge thematic density. Furthermore, it shows the count of sentences, words, syllables, polysyllabic words (containing at least 3 syllables), long words (containing 7 or more characters), and nouns. To facilitate text editing and simplification, the text box highlights long words, long sentences (containing more than 17 words), and nouns that may impact readability. Since text complexity indicates both learner proficiency and reader comprehension, this tool is effective for evaluating language learners and analysing the difficulty of educational or other text material.

Lastly, Writing Evaluator provides feedback on **vocabulary**, reflecting both the user's eloquence and the text's clarity. It assesses lexical diversity using metrics such as the root type-token ratio, MTLD, and HDD. Vocabulary range, or lexical sophistication, indicates the percentage of relatively rare words that are not among the 5,000 most frequent words (considering the base form) in the balanced sub-corpus of the Estonian Reference Corpus. Noun abstractness is rated on a scale of 1–3. Both the word frequency

and noun abstractness are analysed with components of a speed-reading tool[11] developed at the University of Tartu. Lexical density, i.e., the proportion of content words, is calculated by filtering out function words based on the list of Estonian stop words. To assist users, the text box highlights word repetitions, rare words, abstract nouns (level 3), and content words. Again, this can be beneficial not only for assessing writing skill development but also for refining and simplifying professional or educational texts.

While writing assistant tools that suggest corrections and/or measure text complexity have been developed for English and some other bigger languages (see (Norak, 2021) and (Kamarik, 2024) for competitive analysis), ELLE's Writing Evaluator entails novel options for analysing and improving Estonian language texts. It provides more versatile feedback compared to other similar web applications, including the Filosoft speller[12], TartuNLP grammar correction demo[13], and the text evaluation module of the WordWeb Teacher's Tools[14]. The latter outputs some complexity scores and assigns an associated proficiency level to each individual word, helping teachers to choose texts for reading comprehension tasks. ELLE's proficiency assessment considers the entire text and relies on the learners' language use, making it more suitable for evaluating writing production.

Writing Evaluator has been tested with 10th grade students in one high school. The task was to get familiar with text complexity parameters and simplify a media text, e.g., by using shorter sentences and more frequent vocabulary to make it comprehensible for B-level learners. Students formed small groups for the exercise. Their evaluations were favourable, highlighting the tool's potential for correcting written homework and phrasing complex concepts in a more accessible way. Additionally, students suggested to provide the tool as a browser extension. At the time of testing, Writing Evaluator only offered spelling correction and was lacking some functionalities of the complexity and vocabulary view. Thus, additional user tests are needed.

5. Perspectives

Catering for the needs of different user groups, ELLE features unique tools for Estonian language analysis—from receiving automated feedback on personal writing to exploring linguistic patterns together with the teacher or as an individual task. In addition to learner language research, ELLE enables linguists and other scholars to analyse text content and style. With current server resources, we recommend analysing corpora that do not exceed 1 million words. The work on query optimisation is ongoing. Our open-source tools could also be integrated to other text analysis platforms.

ELLE has been introduced to teachers, researchers, and university students at workshops and seminars. Some applications have been tested with mostly Estonian-speaking high school students. However, we plan on a more extensive and systematic testing of the whole environment with both non-native and native target users.

Further developments currently under way include: a statistics page summarising the metadata and linguistic data of EIC, also involving error statistics based on the error-annotated corpus subset; advancing the Writing Evaluator to provide more detailed

[11] https://kiirlugemine.keeleressursid.ee
[12] https://www.filosoft.ee/html_speller_et/
[13] https://grammar.tartunlp.ai
[14] https://sonaveeb.ee/teacher-tools/#/rating

explanations and suggestions by using generative artificial intelligence, and to be able to score native learner texts, e.g., 9th and 12th grade exam writings; adding the option to log in as a learner or teacher/researcher, so the users can save their language material for repeated analysis and track their activities. In the longer term, we plan to compile corpus-based instructional materials, and to allow users to upload and create their own learning resources, enhancing the learning environment aspect of ELLE.

Acknowledgements

This work has been funded by the Tallinn University Research Fund (2019–2024). The spelling and grammatical error correction tools were developed in collaboration with the University of Tartu (2021–2023), funded by the "Estonian Language Technology 2018–2027" programme of the Ministry of Education and Research. We thank all students who have contributed to the development and testing of ELLE through bachelor's and master's theses, individual and group internships, and course work. We are also grateful to Ilja Šmorgun and Jekaterina Trainis who advised us in early discussions about accommodating EIC to the needs of language teachers and learners.

References

Alfter, D., Borin, L., Pilán, I., Tiedemann, T. L., Volodina, E. (2019). Lärka: From language learning platform to infrastructure for research on language learning, *in* Skadina, I., Eskevich, M. (eds.), *Selected papers from the CLARIN Annual Conference 2018*, Linköping Electronic Conference Proceedings **159**, pp. 1–14.

Allkivi-Metsoja, K. (2021). Written Estonian at the levels A2–C1: Comparative automated analysis (Estonian), *Lähivõrdlusi. Lähivertailuja* **31**, 13–59. DOI: 10.5128/LV31.01.

Allkivi-Metsoja, K. (2022). Use of nominals in Estonian A2–C1-level exam writings (Estonian), *Estonian Papers in Applied Linguistics* **18**, 33–53. DOI: 10.5128/533.

Allkivi-Metsoja, K., Kippar, J. (2023). Spelling correction for Estonian learner language, *in* Alumäe, T., Fishel, M. (eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics*, NEALT Proceedings Series **52**, pp. 782–788.

Allkivi-Metsoja, K., Norak, K. (2021). *PoS tagging and lemmatization of Estonian learner language*, Workshop on Parts of Speech (18 May), available at https://youtu.be/EiYwcUQe7ls.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning* **60**(3), 534–572. DOI: 10.1111/j.1467-9922.2010.00566.x.

Eslon, P. (2014). Estonian Interlanguage Corpus (Estonian), *Keel ja Kirjandus* **6**, 436–451.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation, *in* Aijmer, K. (ed.), *Corpora and language teaching*, John Benjamins, pp. 13–32. DOI: 10.1075/scl.33.04gra.

Kamarik, T. (2024). *User interface of an automatic evaluation tool for Estonian texts* (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Kodasma, H. (2023). *Analysis of word frequency, context and collocates in the ELLE language learning environment* (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Kukke, R. (2024). *New syllabification solution for ELLE's tools* (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Liiva, K. (2022). The Clusterfinder web application and its development principles (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Luhtaru, A., Korotkova, E., Fishel, M. (2024). No error left behind: Multilingual grammatical error correction with pe-trained translation models, *in* Graham, Y., Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 1209–1222.

Matsak, E., Eslon, P., Kippar, J. (2010). The development of the prototype for an automatic word order error detector for the Estonian language (Estonian), *in* Eslon, P., Õim, K. (eds.), *Korpusuuring ja meetodid*, Philologia Estonica Tallinnensis 12, pp. 59−100.

Norak, K. (2021). Design of a virtual learning environment prototype for a learner language corpus (Estonian), Master's thesis, Tallinn University, Estonia.

Norak, K., Põldoja, H. (2021). Designing a virtual learning environment based on a learner language corpus, *in* Zhou, W., Mu, Y. (eds.), *Advances in Web-Based-Learning – ICWL 2021*, Lecture Notes in Computer Science **13103**, Springer, Cham. DOI: 10.1007/978-3-030-90785-3_4.

Ots, S. (2012). *Software for morphosyntactic cluster extraction from Estonian texts* (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Römer, U. (2011). Corpus research applications in second language teaching, *Annual Review of Applied Linguistics* **31**, 205–225. DOI: 10.1017/S0267190511000055.

Santos Ordóñez, A., González Lema, C., Puga, M. F. M., Párraga Lema, C., Vega, M. F. C. (2017). Design thinking as a methodology for solving problems: Contributions from academia to society, *in* Larrondo Petrie, M. M., Alvarez, H. (eds.), *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*. DOI: 10.18687/LACCEI2017.1.1.256.

Tarvas, M. (2020). *Estonian Interlanguage Corpus: Implementation and service integration of the new software platform* (Estonian), Bachelor's thesis, Tallinn University, Estonia.

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., Nicholls, D. (2018). Developing an automated writing placement system for ESL learners, *Applied Measurement in Education* **31**(3), 251–267. DOI: 10.1080/08957347.2018.1464447.

Received November 29, 2024, accepted December 4, 2024