Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 392–401
https://doi.org/10.22364/bjmc.2022.10.3.12

Corpus Based Self-Assessment Platform for Latvian Language Learners

Roberts DARĢIS[1], Ilze AUZIŅA[1], Inga KAIJA[1,2],
Kristīne LEVĀNE-PETROVA[1], Kristīne POKRATNIECE[1]

[1] Institute of Mathematics and Computer Science, University of Latvia, 29 Raiņa boulevard, Riga, LV-1459, Latvia,
[2] Rīga Stradiņš University,16 Dzirciema Street, Rīga, LV-1007, Latvia

{roberts.dargis, ilze.auzina, kristine.levane-petrova, kristine.pokratniece}@lumii.lv, inga.kaija@rsu.lv

Abstract. This paper presents a self-assessment platform for Latvian language learners in the breakthrough (A1) and Waystage (A2) levels. The self-assessment platform contains three types of exercises (typing, inflection and gap filling) based on error analysis of the Latvian Language Learner corpus (LaVA). All exercises are automatically generated based on data from multiple corpora. The automatically generated exercises are useful not only for learners outside of classroom or even outside of any formal education setting, but also for educators and authors of learning aids. Currently the self-assessment platform is tailored for language learners at the beginner level, but it can be easily extended for more advanced levels. The self-assessment platform is freely available online (http://uzdevumi.riks.korpuss.lv/en/) and the interface is translated in two language – Latvian and English.

Keywords: Computer-Assisted Language Learning (CALL), Acquisition, Latvian

1 Introduction

The use of learner corpora in language acquisition has been growing steadily over the years, but researchers and teachers also stress the importance of using such corpora in language pedagogy (Granger, 2009).

Various methods of corpus-driven learning, including the use of learner corpora in language pedagogy, have been offered, such as analysing overuse or underuse of certain linguistic features (Granger and Tribble, 2014), automatic grading system development (De Clercq and Van Hoecke, 2020) and the creation of corpus-based exercises (Belz and Vyatkina, 2008), (Mukherjee, 2006). The latter are especially welcome when they are automatically generated and evaluated because that makes them particularly useful

for learners outside of classroom or even outside of any formal education setting. Such automatically generated systems are also beneficial for formal education, for example, to help memorizing grammatical patterns and practice using them in example sentences.

This paper presents a platform that automatically generates corpus-based self-assessment exercises for learners of Latvian in the breakthrough (A1) and Waystage (A2) levels. The paper first introduces the self-assessment platform and then describes the date preparation process.

2 Related Work

Natural language processing (NLP) tools can partially or completely automate a number of exercises related to learning a foreign language and a second language (L2), for example, for automatic sentence selection from different corpora for language learning exercises, e. g. (Smith et al., 2010), (Pilán et al., 2013), (Pilán et al., 2016) and development of learning platforms with automatically generated exercises, e.g. (Volodina et al., 2013), (Boulton, 2016), (Pilán et al., 2018), (Katinskaia et al., 2020). Thus, both the diversity of exercises and support for teachers in the implementation of the curriculum is provided. However, in many cases, the automatic selection of sentences ignores the criteria that determine whether the sentences correspond to the exercise elements of a certain level of language proficiency. When choosing sentences from corpora, there are several additional aspects to consider: (1) whether the sentence is understood in isolation, without the broader context, (2) whether the structure and linguistic complexity of the sentence are appropriate for the appropriate level of language proficiency. Linguistic correspondence to the appropriate level of language proficiency must also be taken into account when it comes to the automatic creation of exercises that offer the acquisition of declinable part-of-speech paradigms.

3 Self-Assessment Platform

First, an error analysis on error-annotated learners corpus has been done to figure out what types of exercises would be useful for language learners. The error analysis is done on Latvian Language Learner corpus (LaVA) (Darģis et al., 2020), which contains error annotated texts written by beginner level (A1 and A2) language learners who are learning Latvian as foreign language at different universities in Latvia. The corpus contains detailed error annotation schema and provides a wide range of statistical analysis, enabling researchers to conduct numerous kinds of quantitative research.

On average every fourth word in the corpus contained an error. The majority of errors are word formation (46%) and spelling errors (45%).

Further breakdown of word formation errors revealed that word normal form is used instead of the form required by the context in 44% of cases, showing the main reason for word formation errors is not the incorrect usage of inflections but the lack of knowledge of word inflections instead.

Standard Latvian orthography uses 22 unmodified letters of the Latin alphabet (*q*, *w*, *x*, *y* is not used) extend with 10 modified letters (*ā, č, ē, ģ, ķ, ļ, ņ, š, ū, ž*. Macron on vowels (*ā*, *ē*, *ī*, *ū*) is used to show length. Using short vowels instead of long ones

is the most common spelling error (53%). In total incorrect use of modification marks represents 75% of total errors.

Three types of exercises were selected to help learners avoid these errors.

- Typing exercises, where text needs to be retyped from the device screen.
- Word conjugation and declination exercises, where the given word or several words should be written in all word forms (in accordance with the learners language level).
- Gap filling exercises, where learner needs to insert the word in the correct form in the given sentence.

The self-assessment platform is freely available online[3] and the interface is translated in two language - Latvian and English.

3.1 Typing

Rewriting a text helps to better acquire the graphemic system of the Latvian language, both visually paying attention to the sequence of letters and diacritical signs, and repeating this sequence independently. Rewriting is also mentioned as a useful way to learn spelling for learners with dyslexia (Crombie, 2000). Such exercises also make it possible to train the use of diacritical signs typing on a computer, preventing cases where diacritical signs are not used simply because the learner is not technically accustomed to doing so.

In the typing exercise, the language learner has the option to rewrite computer-typed sentences, or handwritten sentences. The computer-typed sentence is randomly selected from a predefined set. The handwritten sentence is a randomly selected image of a sentence, obtained by manually cutting out sentences from the learners' essays.

If a language learner makes a mistake when writing a sentence, it is immediately flagged - the frame around the text remains red. An incorrect letter, a letter without or with an inappropriate diacritical sign, incorrect use of uppercase and lowercase letters, unnecessary or missing space, missing or incorrect punctuation, is considered to be a mistake. When the whole sentence is rewritten correctly, it is highlighted in green (Figure 1). A new sentence might be requested at any time.

3.2 Inflection

In the second group of exercises, the language learners are offered to learn paradigms of declinable part-of-speech: declination of nouns, verbs, adjectives, numerals, and pronouns (Figure 2). The exercises include only the acquisition of word forms corresponding to the characteristics of the linguistic competence given at the level of proficiency in the Latvian language, namely, the established knowledge of grammar (Šalme and Auziņa, 2016).

When choosing a word class / part-of-speech in order to learn the declension of words, additional choice options are offered, for example, for nouns - declensions, for verbs - conjugation, reflexive or non-reflexive verbs. Adjectives and numerals can be inflected together with a noun, because adjectives and declinable numerals agree in

[3] http://uzdevumi.riks.korpuss.lv/en/

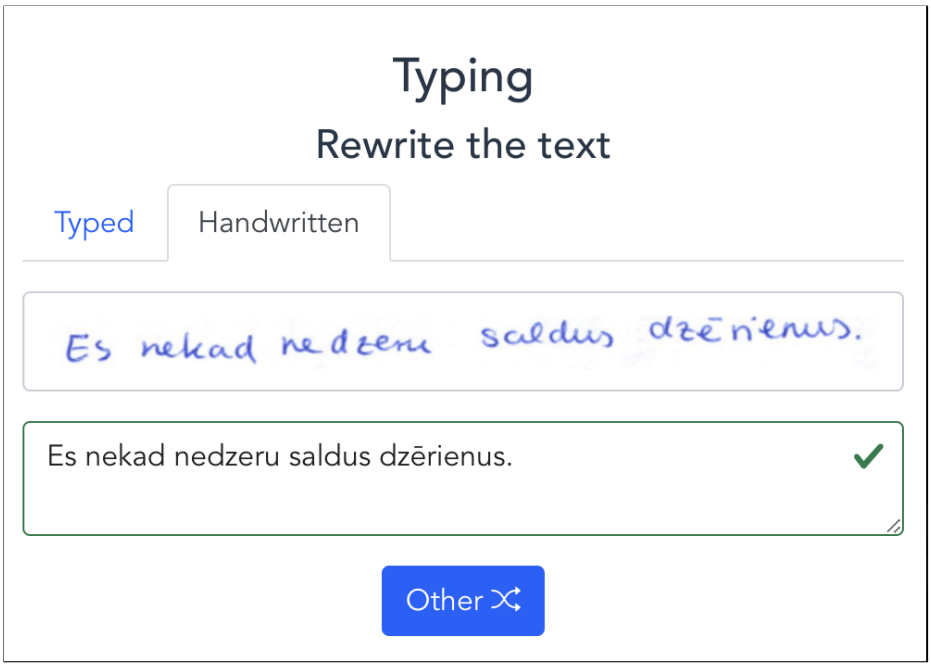

Fig. 1. Screenshot of the writing exercis for a handwritten sentence

gender, number, and case with the noun to which they are syntactically linked, for example, *liela māja* 'big house' (sg.nom.fem.), *lielā mājā* 'in big house' (sg.loc.fem.).

Depending on the choice made, one word from the vocabulary found in the LaVA[4] corpus is offered at random (Section 4.2). Only those grammatical forms that correspond to the level of language proficiency are shown in the exercises (Section 4.1).

When completing the exercise, several options are offered: (1) check the entered word forms (*Verify*), (2) view the correct answers (*Show Answers*), (3) choose other words (*New Sample*). You can view the correct answers at any time and hide them again.

3.3 Gap Filling

Gap filling exercises that are used to learn vocabulary and grammar are very important in language learning. This exercise type is also used in the LaVA exercise set. Similar to inflection exercises, the learners can choose a part-of-speech and corresponding grammatical categories. After making a selection, 10 sentences are randomly selected from a pre-prepared set of sentences (Section 4.3), that contain words according to the selection criteria (Figure 3).

[4] http://lava.korpuss.lv

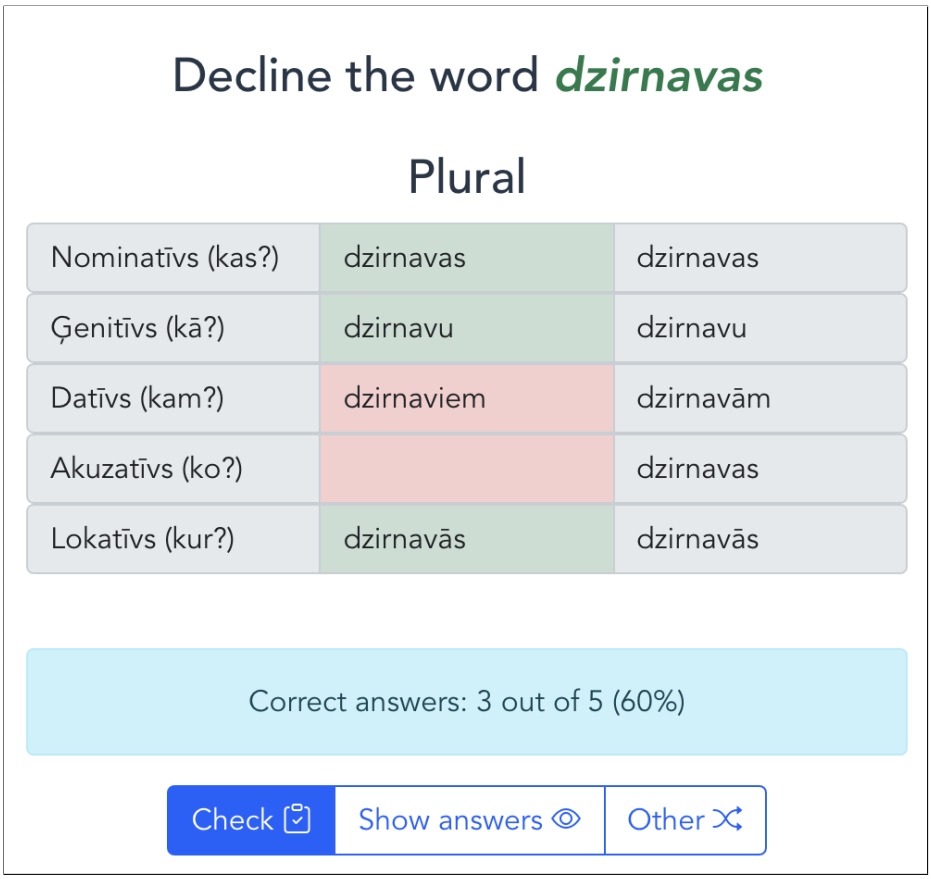

Fig. 2. Screenshot of the inflection exercise for the word *dzirnavas (mill)*

Exactly one word must be inserted or entered in each sentence. The word to be inserted is given in the basic form, with additional grammatical information.

Options *Verify*, *Show answers* and *New samples* function the same as options in the inflection exercises.

4 Data Preparation

The key of exercise generation is data. The first step is to define which language skill learners should know at beginner level (A1 and A2). These definitions will be used to filter data with appropriate complexity for exercise generation.

Fig. 3. Screenshot of the gap filling exercise

4.1 Definition of Language Skills

According to ELP, at A1 level learner can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. A learner can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. At A2 level he/she can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). A learner can communicate in simple and routine exercises requiring a simple and direct exchange of information on familiar and routine matters. He/she can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

More detailed description that is implemented in the text filtration was developed based on grammatical forms and constructions that are typically thought at that level for Latvian:

- nouns of 1st–6th declension;
- active voice verbs in indicative mood;
- adjectives with indefinite endings;
- pronouns
  - personal pronouns in nominative, genitive, dative, accusative, and locative;
  - demonstrative, possessive, interrogative, indefinite, and definite pronouns in nominative, genitive, dative, accusative, and locative;
  - relative, reflexive, and negative pronouns are not included;
- adverbs;
- prepositive prepositions used with independent words in singular:
  - *aiz*, *virs*, *zem*, *pie*, *no*, *ārpus*, *pirms*, *pēc*, *kopš*, *bez* used with singular genitive forms;
  - *līdz* used with singular dative forms;
  - *ap*, *gar*, *pa*, *caur*, *pret*, *starp*, *pār*, *ar*, *par* used with singular accusative forms;
  - *uz* used with singular accusative or singular genitive forms;
- simple conjunctions *un*, *bet*, *vai*, *ka*, *jo*, *tāpēc ka*, *ja*, *kā*, *lai*;
- cardinal and ordinal numerals (simple or compound) in all cases, except for vocative;
- interjections;
- simple particles *vai* (only as an interrogative particle), *jā*, *nē*, *varbūt*, *arī*, *diemžēl*, *kā* (only as a comparison), *nekā*;
- abbreviations are not included.

4.2 Vocabulary

When generating exercises, it is important to only use vocabulary that the learners are supposed to know. A learner corpus can provide the most precise information about the vocabulary used by learners.

The Latvian learner corpus LaVa was used to extract information about the lemmas used by the learners in their texts. They may include specific words that do not match beginner level on language acquisition because a certain author may have found such a word in a dictionary for the needs of their text. In order to avoid including such words in the word list, only words found in at least three texts were included.

The size of the learner corpus is not sufficient to have representative statistical data about the forms used of each word. Thus, the words in the word list were conjugated automatically by the open-source IMCS UL morphological tagger (Paikens et al., 2013) and filtered based on the level description. The list of word forms was further used to generate exercises and select example sentences.

4.3 Sentences

One of the most complicated parts in exercise generation is finding appropriate sentences. Sentences must be diverse and must correspond to the level of language proficiency – simple enough that the learner can understand them and complex enough that the learner can test his/her abilities. Moreover, they must be understood in isolation.

A lot of sentences are required to automatically generate diverse exercises, so manual creation is not feasible. Corpora offer a wide selection of conveniently retrievable examples.

Using examples from corpora has been proven to positively influence the development of learners' linguistic abilities. Simplest solution would be taking sentences from learner corpora. Unfortunately, learner corpora for Latvian are not large enough to yield sufficient coverage of different words in different context. Learners in the beginner level might experience difficulties understanding all the sentences from any other general corpus, so carefully designed selection criteria are required to filter out sentences with the appropriate complexity.

Criteria developed by experienced educators based on the A1–A2 level description were used to select additional sentences from The Balanced Corpus of Modern Latvian (LVK2018) (Dargis et al., 2020).

There are two factors to sentence complexity: vocabulary and syntactical structure. To make the complexity of a sentence's vocabulary appropriate, each sentence should contain words only from the vocabulary described in Section 4.2.

Validating the complexity of a sentence's syntactical structure requires checking multiple criteria:

- one sentence must not consist of more than three independent clauses;
- all clauses must include grammatical centre – preferably, both a subject and a predicate;
- each clause should have no more than two adverbs (however, more than two objects and modifiers can be included);
- there should be no more than two particles in each sentence.

Automatically selected sentences are reviewed manually, leaving out those which are not understood outside a context.

5 Conclusion

This paper presented a corpus based platform for language learners. The platform can also be useful for educators and authors of learning aids when developing written exercises or exams. In our opinion corpus based exercises provide more natural and diverse learning experience. The platform development principles described in this paper can be directly applied to other languages.

Currently the main target audience for the platform are adults who are learning Latvian as foreign language in the Breakthrough (A1) and Waystage (A2) levels. Future work includes extending the platform for more advanced levels. The platform can be

easily adjusted for native speakers learning Latvian in a school, by removing the limitations for the vocabulary and inflections. One of the future research direction could be defining language skill requirements for all language levels which would allow to add complexity scale to the self-assessment platform based on learner's level. After that, another interesting research direction would be developing a test exercises that would determine the language proficiency level of the learner. The determined level could be use to automatically adjust the complexity of exercise.

Teachers and students – both those studying in Latvian higher education institutions and those studying Latvian abroad – have been introduced to the self-assessment exercise platform. The platform and exercises are currently being tested and the results of the testing will be available soon.

6 Acknowledgements

The work reported in this paper is a part of the project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) that is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The project is financed by Latvian Council of Science.

This work is also a part of the National Research Programme *Digital Resources of the Humanities* project *Digital Resources for Humanities: Integration and Development* (No. VPP-IZM-DH-2020/1-0001) and has received financial support from the Latvian Language Agency through the grant agreement No. 4.6/2019-029.

7 Bibliographical References

References

Belz, J. A., Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction, *Language Learning and Technology* **12**, 33–52.

Boulton, A. (2016). Integrating corpus tools and techniques in ESP courses, *ASp. la revue du GERAS* **69**, 113–137.

Crombie, M. A. (2000). Dyslexia and the learning of a foreign language in school: where are we going?, *Dyslexia* **6**(2), 112–123.

Dargis, R., Levane-Petrova, K., Poikans, I. (2020). Lessons learned from creating a balanced corpus from online data, *Human Language Technologies – The Baltic Perspective*, Vol. 328, IOS Press, pp. 127–134.
https://ebooks.iospress.nl/volumearticle/55535

Darģis, R., Auziņa, I., Levāne-Petrova, K., Kaija, I. (2020). Detailed error annotation for morphologically rich languages: Latvian use case, *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 241–244.

De Clercq, O., Van Hoecke, S. (2020). An exploratory study into automated précis grading, *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 397–404.
https://www.aclweb.org/anthology/2020.lrec-1.50

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching, *Corpora and language teaching* **33**, 13–32.

Granger, S., Tribble, C. (2014). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning, *Learner English on computer*, Routledge, pp. 199–209.

Katinskaia, A., Ivanova, S., Yangarber, R. (2020). Toward a paradigm shift in collection of learner corpora, *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 386–391.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art–and beyond, *Corpus technology and language pedagogy: New resources, new tools, new methods* pp. 5–24.

Paikens, P., Rituma, L., Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, Oslo, Norway, pp. 267–277.
http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA24.pdf

Pilán, I., David, A., Borin, L., Tiedemann, T. L., Volodina, E. (2018). From language learning platform to infrastructure for research on language learning, *CLARIN Annual Conference 2018*, p. 53.

Pilán, I., Vajjala, S., Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity, *arXiv preprint arXiv:1603.08868* .

Pilán, I., Volodina, E., Johansson, R. (2013). Automatic selection of suitable sentences for language learning exercises, *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*, Research-publishing. net Dublin, pp. 218–225.

Šalme, A., Auziņa, I. (2016). *Latviešu valodas prasmes līmeņi: Pamātlīmenis A1, A2, Vidējais līmenis B1, B2*, Latviešu valodas aģentūra.

Smith, S., Avinesh, P., Kilgarriff, A. (2010). Gap-fill tests for language learners: Corpus-driven item generation, *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, Macmillan Publishers, pp. 1–6.

Volodina, E., Pijetlovic, D., Pilán, I., Kokkinakis, S. J. (2013). Towards a gold standard for swedish cefr-based icall, *Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning. NEALT Proceedings Series*, Vol. 17.

Received August 19, 2022 , accepted August 24, 2022