Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 392–396
Marseille, 11–16 May 2020

Quality Focused Approach to a Learner Corpus Development

Roberts Darģis[1], Ilze Auziņa[2], Kristīne Levāne-Petrova[3], Inga Kaija[4]
Institute of Mathematics and Computer Science, University of Latvia[1,2,3,4], Riga Stradiņš University[4]
Raina bulvaris 29, Riga, LV-1459, Latvia[1,2,3,4], Dzirciema iela 16, Riga, LV-1007, Latvia[4]
{roberts.dargis, ilze.auzina, kristine.levane-petrova}@lumii.lv, inga.kaija@rsu.lv

Abstract

The paper presents quality focused approach to a learner corpus development. The methodology was developed with multiple design considerations put in place to make the annotation process easier and at the same time reduce the amount of mistakes that could be introduced due to inconsistent text correction or carelessness. The approach suggested in this paper consists of multiple parts: comparison of digitized texts by several annotators, text correction, automated morphological analysis, and manual review of annotations. The described approach is used to create Latvian Language Learner corpus (LaVA) which is part of a currently ongoing project *Development of Learner corpus of Latvian: methods, tools and applications*.

Keywords: language acquisition, corpus development, validation

1. Introduction

Learner corpora constitute a new resource for second language acquisition and foreign language teaching specialists. They are particularly useful if they are error-tagged. While error-tagging is problematic in many theoretical aspects, it is probably not controversial anymore that learner corpus with consistently annotated errors can be useful for a number of research questions. Unfortunately, there are multiple reasons why mistakes could be introduced in annotations. Different interpretations of the source text lead to different text corrections. Annotation guidelines do not contain all possible scenarios which also may result in correction differences. Besides, mistakes can be introduced due to carelessness, since error annotation is a highly repetitive task which leads to loss of focus. Multiple design principles are discussed in this article to reduce the risk of these mistakes in every corpus development step. The discussed principles are implemented in the corpus platform which is used to develop the error-tagged Learner corpus of Latvian (LaVA).

The LaVA corpus is developed as a part of an ongoing project *Development of Learner corpus of Latvian: methods, tools and applications*, started in September 2018. The project has several interrelated goals: (1) creation of infrastructure for corpus collection, annotation methodology (both error annotation and morphological annotation); (2) development of an error-tagged Learner Corpus of Latvian (LaVA); and (3) development of corpus-based learning materials and a self-assessment web platform for learners of Latvian.

Latvian is a language with rich morphology and a relatively free word order. Latvian can be generally considered a phonetic language – a language with a relatively simple relationship between orthography and phonology. From the language acquisition perspective, Latvian has several specific properties: short and long vowels and diphthongs, a high degree of inflection, rather free word order. These properties have to be taken into account in error-annotation.

2. Related Work

A learner corpus is a computerized textual database of the language produced by foreign language learners (Leech 1998). Learner corpora have been collected and analyzed for more than 25 years now and their popularity is increasing. There are many learner corpora for English, such as the International Corpus of Learner English (Granger et al., 2009), the Longman Learner's Corpus, or the Cambridge Learner Corpus (Nicholls, 2003). Many have also been created for other languages, e. g., Learner Corpus of Portuguese (del Rio et al., 2016, Mendes et al., 2016), Russian Learner Corpus (Rakhilina, 2016), L1 Learner Corpus for German (Abel et al. 2014), an error-annotated learner corpus of German as a foreign language FALKO (Reznicek et al. 2012; Reznicek et al. 2013), etc. The importance of such empirical data has been widely recognized for studies in the fields of language teaching, language learning and second language acquisition. There are also some noticeable developments in creating learner corpora of the Latvian language: (1) a corpus of the texts collected from the successfully passed tests of the State Language Proficiency Testing which is used to evaluate a person's state language proficiency level. (Darģis et al. 2018); and (2) a publicly available learner corpus of the second Baltic language ESAM (Znotiņa, 2015; Znotiņa, 2017). LaVA is the first freely available Latvian learner corpus.

3. Data collection process

Data is gathered from international students studying in higher education institutions in Latvia and learning Latvian as foreign language in formal courses at these institutions. Language teachers from multiple higher education institutions have agreed to support the corpus creation process by asking their students to write essays on various topics. The handout material contains background questionnaire about language knowledge, as well as the copyright claims, and personal data protection system used in the project. The students are asked not to include any personal information in the essay regardless of the topic.

The handwritten essays alongside background questionnaires are scanned and uploaded to corpus platform for further processing. Metadata from background questionnaires is manually entered in the corpus platform. Metadata contains: *age, gender, mother tongue(-s), other languages spoken by the author*, and *the length of residence in Latvia*. Authors' names are not included to retain anonymity. The target amount of the corpus is at least 1000 essays.

4. Data annotation

The annotation pipeline is divided in four steps: 1) Digitization; 2) Text Correction; 3) Morphological annotation; 4) Error annotation.
Digitization, text correction and error annotation are done independently by two annotators. Any inconsistencies are reviewed by a third independent annotator to make the final decision.

5. Data digitization

The digitization of essays is a challenging task, as the essays are handwritten. The essays contain a lot of errors as expected from early language learners. Even when it is clear what word student meant to write, sometimes it is difficult to understand how the student spelled that word. Some letters are very similar in certain handwritings (such as *a* and *o*, or *i* and *ī*), and any corpus creators' error can influence the identification and analysis of the students' errors. Unusual markings are also used to spell atypical characters or diacritical marks in Latvian, such as *w*, *ø*, umlaut, acute above the vowel character, etc. In such cases, the graphs and diacritical marks used by the learner are preserved in the transcript, as long as a suitable symbol can be found. (Kaija 2019) Sometimes an essay is written using just capital letters; in that case, capitalization is used only where it is required by orthography rules.

There are just some encoded elements in the transcripts: not understandable essay parts are enclosed in square brackets, for example, *Man patīk iet vai [palaist]*. It can be an understandable word or even phrase that does not make sense in a context, or a row of characters that cannot be recognized at all. Any changes made by the student during the writing process (such as deletions, additions, transposition of segments, etc.) or other elements of the essay are not encoded.

To facilitate the digitization process, the platform offers some technical solutions, for instance, it is possible to zoom the screenshot of the text to be digitized.

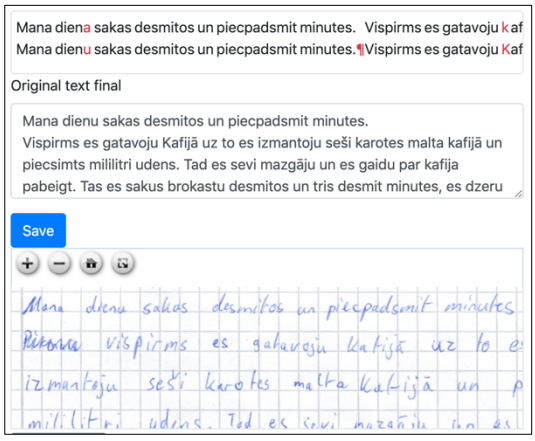

Figure 1: Creating the final version from alignments of two hypotheses

The transcription workflow is as follows: one corpus creator manually digitizes the essay in the corpus platform and saves it *(Original text: First)*, keeping all the students' mistakes and language specifics. Another digitizer performs the same step *(Original text: Second)*. The third corpus creator then creates the final version of this essay *(Original text: Final)* by comparing the two versions and making the final ruling on any inconsistencies. Figure 1 shows this step. The original text typed by the two independent corpus creators *(Original text: First; Original text: Second)* is automatically aligned and mismatches are displayed in red which makes the comparison more efficient. The third corpus creator makes the final decision with the help of the scanned picture. Automatic handwriting recognition was also tried, but it gave no benefit, as more than 50% of the text was recognized incorrectly.

So far, 598 essays (more than half of the planned 1000 essays) containing 524167 characters have been digitized. Each essay has been digitized by a random pair of two out of four corpus creators. Overall character level inter-annotator agreement is 92%. Disagreement could be due to one of two reasons: a) too ambiguous handwriting or b) carelessness mistake. Since no qualitative disagreement analysis has been done for data digitalization, for further digitization analysis the worst case scenario was assumed – all disagreements between corpus creator and final version were considered errors due to carelessness. The character level error rate varies from 0% to 9.5%. The average error rate across all corpus creators is 1.5%. If two corpus creators process the same text, the final error rate should be 0.02%, assuming error distribution for each corpus creator individually is uniform and independent.

6. Text Correction

Nowadays many language learner corpora are error-annotated. In order to annotate errors, the text of the learner must be reconstructed (target hypothesis must be set). It means that a 'correct' version of the utterances is assumed. Adding a fine-worded target hypothesis is necessary for intelligible analysis and all kinds of further annotation of learner corpus (Reznicek et al. 2013). However, it is often hard to agree on one target hypothesis. The ambiguity of target hypothesis is discussed in several studies (Lüdeling 2008; Reznicek et al. 2103; etc.). They experimentally prove that there is a possibly infinite number of target hypotheses for a learner utterance, depending on the linguistic level that is corrected (orthography, grammar, lexical, etc.).

It is useful to describe the principles of text correction (detailed guidelines) in an annotation manual to avoid possibly conflicting target hypotheses. Such manual has been developed for the LaVA corpus.

Target hypotheses for the whole essay are written based on the annotation manual to reduce the possibility of ambiguous target hypothesis as much as possible. The main principle behind LaVA annotation manual is that the target hypothesis should stay as close to the learner utterance structure as possible when orthography, word formation and derivation, punctuation, as well as lexical and some syntactic features are corrected. The word order in the utterance is not changed. The language style also is not modified.

Text correction workflow is similar to the digitization workflow process. Two corpus creators independently write their target hypothesis for the whole essay. Both of

the target hypotheses are automatically aligned and mismatched sections are reviewed by the third corpus creator. Multiple features were introduced in the corpus development platform to achieve higher quality in text correction.

One such quality feature is based on the phenomenon that people's minds tend to guess words instead of reading them if the text is known. This can cause carelessness mistakes in text correction. There are 4 linguists working on the corpus creation. To minimize the guessing phenomenon, text correction is done by the other two corpus creators, who didn't work on the first step of essay digitization. Although one of the corpus creators has already seen the text in text digitization review step, usually only the mismatched places are reviewed; therefore, due to the team size, choosing the text correctors this way is still better than any other option.

To reduce the amount of incorrectly spelled words that might have been missed by both correctors, the review window besides mismatched segments also highlights words that are not in the dictionary. Although highlighting could be also added to the text correction step, it is deliberately omitted because corpus creators might pay more attention to the highlighted places, paying less attention to the surrounding text which could introduce more errors due to carelessness.

Annotation guidelines cannot describe all possible correction scenarios because it is impossible to predict all learners' mistakes. Incomplete annotation guidelines could lead to different text correction. Once every three months word level corrections are grouped and reviewed (Figure 2) to find cases which are common enough to be included in the annotation guidelines. The first word shows the original word, the arrow points to the corrected word and the number of occurrences.

viņs (he with a typo)	
→ *unchanged*	1
→ viņš (he)	81
→ viņam (for him)	12
universitāte (university)	
→ *unchanged*	61
→ universitātē (in university)	44
patīk (like)	
→ *unchanged*	34
→ garšo (like the taste)	23
ābols (apple)	
→ *unchanged*	4
→ āboli (apples)	8

Figure 2: Review of grouped corrections

From the first group for the word "viņs" we can see that one misspelled occurrence has been left unchanged and that needs to be fixed.

Looking at the second group for the word "universitāte", one can compare the number of times the word has been unchanged with the number of times it has been changed to a different – but similar – form "universitātē". This could be because students often confuse both forms. This error could be easily missed by the corpus creators because words are very similar and both are in dictionary. It might be worth it to take another look at the unchanged cases just to make sure that none of the errors has been missed.

Students often write about things they like. In two sentences (a) "I like cats" and (b) "I like pizza" the word "like" translates different in Latvian. Although both can be translated to "patīk", usually the word "garšo" is used about food. If this scenario is not described in the annotation guidelines, some corpus creators might correct this and some might leave the original word. The review process helps to find such cases, allowing to update annotation guidelines and to standardize previous occurrences.

The last group presents a similar question: whether to correct cases where students write a singular form instead of a plural form, although the singular form in not strictly incorrect. For example, in sentence "I like (an) apple", student probably meant to say he likes apples in general – in which case a plural form of "apple" should be used. These cases should be normalized regardless of the final decision about which approach to use.

After the list of isolated edits is reviewed and suspicious ones are marked, and the annotation guidelines are updated where necessary, the concordances for the suspicious cases are reviewed. Each concordance contains a link to the final version of the corrected text to provide easy access to text if it needs corrections.

So far 586 essays (more than half of planned 1000 essays) containing 104505 tokens have been corrected. Each essay has been corrected by a random pair of two out of four corpus creators. 28% of tokens were changed by at least one corpus creator. Overall token level inter-annotator agreement is 93%. Only 2% of corrections were done by both of the creators. The remaining 5% of edits were done by only one of the corpus creators. Assuming 5% is the average error rate due to carelessness or unnecessary corrections for one corpus creator, the final error rate using two corpus creators should be 0.25%.

Doing qualitative error analysis it was confirmed that most of the disagreements were due to carelessness. Other disagreements were 1) in matters of punctuation (for example, no comma was inserted between parts of a compound sentence before *un* 'and'), 2) in places where standard language norms allow more than one correct word form (for example, *astoņpadsmit gadi* (Pl.Nom.) / *gadu* (Pl.Gen.) 'eighteen years', 3) in cases where the word form was correct but semantically another form is necessary (for example, *Man garšo auglis* (Sg.Nom.) 'I like fruit' → *Man garšo augļi* (Pl.Nom.) 'I like fruits'.

7. Morphological annotation

The learner texts are morphologically annotated by IMCS morphological tagger. (Paikens 2007; Paikens et al., 2013; Paikens 2016) Morphological annotation is a challenging task, even more so for languages (such us Latvian, Czech) with rich inflection, derivation, agreement, and rather free word order. (Rosen et al. 2014) Automatic morphological analysis provides additional information about lemma, POS and morphological categories for word forms. Two or more errors may be present in one word form. Morphological tags help automatically determine the error type, for example if the error is at the root of the noun, it is

most likely a spelling error, the mistaken ending is a word formation error.

Morphologically annotated text can be used to perform simple quantitative analysis to determine in which part of speech (such as nouns, pronouns, verbs, adjectives), inflection, tense, person, etc., learners make most mistakes.

Morphological annotations are added to the original and the corrected text. It is done only by one corpus creator because morphological annotation is a straightforward process. There is an *unclear* option for morphological features which can't be determined due to an incorrect spelling in the original text by a language learner which does not represent any valid word form.

Morphological annotations are typically added one by one for each word in each text. Annotation process is organized differently in this corpus creation stage. Instead of annotating every text word by word, all word forms are grouped from all the texts. Annotation is done word form by word form for all the occurrences in the texts. It was done this way because most of the word forms have only one morphological tag in all occurrences, so adding annotation to one word form in all occurrences is faster than adding annotation word by word.

In the morphological annotation interface (Figure 3), all the occurrences for each wordform are shown in a list of concordances. The automatically generated annotations are already filled in. If necessary, annotations can be edited for each case individually, or all the cases can be annotated at once with a batch annotation operation if there is just one correct annotation or if there is one more common morphological annotation. Adding morphological annotations to every word in the corpus gives another opportunity to review the text corrections. It is crucial to have the opportunity to edit annotations from previous level if a mistake is discovered, so every concordance has a link to the full text where it can be edited.

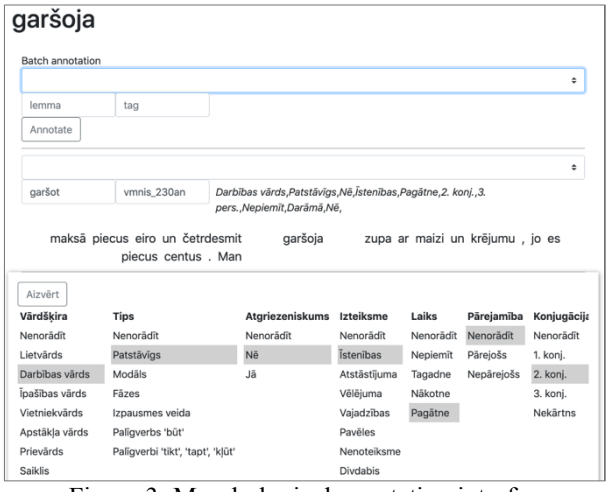

Figure 3: Morphological annotation interface

8. Error annotation

An updated version of the error annotation system from a previous project is used (Darģis et al. 2018). The system generates suggestions for error annotations which are verified by two corpus creators. Segments where annotations between the two corpus creators differ are reviewed by the third corpus creator. Automatically generated suggestions are a lot more precise than those generated in previous project due to manually verified morphological annotations. The annotation suggestions could be removed for the same reason as dictionary suggestions are not shown for the first two corpus creators in text corrections step, but it was decided to show them from the start because it significantly reduces the time needed to create error annotations.

Error annotation methodology is also described in the corpus annotation guidelines. Once every three months error annotations for the same text corrections are grouped, reviewed, and guidelines are updated if necessary – similar to the process in text correction.

9. Conclusion

This paper describes a quality focused approach to corpus development, focusing on annotation validation methodologies. The described methodology is used to develop a platform which is used to create the error-tagged Learner corpus of Latvian (LaVA). When finished, LaVA corpus will be the largest Latvian language learner corpus. It will also be freely available.

Inter-annotator agreement analysis on the work done so far revealed that main cause of disagreements is carelessness, showing that quality assurance procedures have a curtailing role in learner corpus development.

To further reduce mistakes in digitization, experiments with optical character recognition could be carried out either by adapting some exiting system to achieve better accuracy or by using it to align the digitalized text with the image to find any inconsistencies. An isolated location of the character in question could also speed up the review process.

To further reduce mistakes in text correction process, a more sophisticated context analysis system could be used instead of a plain dictionary to find not only incorrectly spelled words but also words that seem suspicious in the given context (comparable to the correct usage of *your* and *you're* in English).

10. Acknowledgements

The work reported in this paper are 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 also a part of the Latvian State Research Programme "Latvian Language" (No. VPP-IZM-2018/2-0002) subproject "Acquisition of Latvian Language" that is being implemented at IMCS UL.

11. Bibliographical References

Abel, A., Glaznieks, A., Nicolas, L., and Stemle, E. (2014). Koko: An L1 learner corpus for german. In Proceedings of LREC 2014, pp. 2414–2421.

Darģis, R., Auziņa, I., and Levāne-Petrova, K. (2018). The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

del Río, I., Antunes, S., Mendes, A. and Janssen, M. (2016). Towards error annotation in a learner corpus of Portuguese. In Proceedings of the 5th NLP4CALL and 1st NLP4LA workshop in Sixth Swedish Language Technology Conference (SLTC). Umeå University, Sweden, 17-18 November.

Granger, S., E. Dagneaux, F. Meunier and M. Paquot. Eds. (2009). International Corpus of Learner English. Version 2. UCL: Presses Universitaires de Louvain.

Laarmann-Quante, R., Ortmann, K., Ehlert, A., Vogel, M., and Dipper, S. (2017). Annotating Orthographic Target Hypotheses in a German L1 Learner Corpus. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Copenhagen, Denmark, pp. 444–456. Online: https://www.aclweb.org/anthology/W17-5051.pdf

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. Fortgeschrittene Lernervarietäten Advanced Learner Varieties. Corpus Linguistics and Research into Second Language Acquisition: Korpuslinguistik und Zweitsprachenerwerbsforschung. pp. 119–140. Online: https://doi.org/10.1515/9783484970342.2.119

Mendes, A., Antunes, S., Janssen, M., and Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. LREC 2016.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson and T. McEnery (Eds.). Proceedings of the Corpus Linguistics 2003 Conference. Lancaster University, pp. 572–581.

Paikens, P. (2007). Lexicon-based morphological analysis of Latvian language. Proceedings of the 3rd Baltic Conference on Human Language Technologies (Kaunas, October 2007). Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 235–240.

Paikens, P., Rituma, L, Pretkalniņa, L. (2013). Morphological analysis with limited resources: Latvian example. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA). Linköping: Linköping University Electronic Press, 267–277.

Paikens, P. (2016). Deep neural learning approaches for Latvian morphological tagging. Human Language Technologies – The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016. 289. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 136–143.

Rakhilina, E., Vyrenkova, A., Mustakimova, E., Ladygina, A., and Smirnov, I. (2016). Building a learner corpus for Russian. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016. http://aclweb.org/anthology/W16-65

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In: Automatic Treatment and Analysis of Learner Corpus Data. Edited by Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson, pp. 101–124.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., and Torsten, A. (2012). Das Falko-Handbuch Korpusaufbau und Annotationen. Version 2.01. Humboldt-Universität zu Berlin.

Rosen, A., Hana, J., Štindlová, B., Škodová, S., and Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. In: Language Resources and Evaluation. March 2014, Volume 48, Issue 1, pp 65–92.