The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners

Roberts Darģis[1], Ilze Auziņa[2], Kristīne Levāne-Petrova[3]
Faculty of Computing, University of Latvia[1], Institute of Mathematics and Computer Science, University of Latvia[2,3]
Raina bulvaris 19, Riga, LV-1459, Latvia[1], Raina bulvaris 29, Riga, LV-1459, Latvia[2,3]
{ roberts.dargis, ilze.auzina, kristine.levane-petrova }@lumii.lv

Abstract

This article presents a different method for creation of error annotated corpora. The approach suggested in this paper consists of multiple parts - text correction, automated morphological analysis, automated text alignment and error annotation. Error annotation can easily be semi-automated with a rule-based system, similar to the one used in this paper. The text correction can also be semi-automated using a rule-based system or even machine learning. The use of the text correction, word, and letter alignment enables more in-depth analysis of errors types, providing opportunities for quantitative research. The proposed method has been approbated in the development of the corpus of the Latvian language learners. Spelling, punctuation, grammatical, syntactic and lexical errors are annotated in the corpus. Text that is not understandable is marked as unclear for additional analysis. The method can easily be adapted for the development of error corpora in any other languages with relatively free word order. The highest gain from this method will be for highly inflected languages with rich morphology.

Keywords: learner corpus, error annotation, word alignment

1. Introduction

The purpose of this article is to describe the error-annotating methodology and the tool that is used to annotate *The Corpus of the Latvian Language Learners* (Latvian as L2 and foreign). As the Sylviane Granger admits, the learner corpora constitute a new resource for second language acquisition and foreign language teaching specialists, especially if they are error-tagged. (Granger, 2003).

Appropriately designed learner corpus and consistently annotated errors can provide answers to global questions such as: what is the most frequent type of error, how the native language influence the error type. As the developed corpus includes the texts of different levels of language acquisition the corpus can provide an answer to very specific questions, for example, are mistakes related to noun endings more frequent for B2 or C1 level? Based on the data from the corpus, also different workbooks might be developed for people learning a second language.

Latvian is a language with rich morphology and a relatively free word order. Latvian in general can be 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, highly inflected language, rather free word order. These properties have to be taken into account in error-annotation.

There are many learner corpora for English and last decades learner corpus have been created for other languages as well, for example, French, Swedish, Norwegian, Dutch, Spanish and German (Granger, 2008), and their range is expanding.

Currently, *The Corpus of the Latvian Language Learners* is being created. The corpus contains the successfully passed tests of the State Language Proficiency Testing (Certification) that is used to evaluate a person's (henceforth – Applicant's) state language proficiency level. For every language proficiency level (A1, A2, B1, B2, C1, C2) 150 tests have been used that makes in total 900 tests. If the State Language Proficiency Examination is passed successfully, the Applicant receives the state language proficiency certificate, that is required for employment requirements and acquisition of a permanent residence permit. The methodology and tool described in this paper are used to create this corpus.

At this moment, there is no other Latvian learner corpus. One more learner corpus of Latvian is being developed (www.esamkorpuss.lv) by PhD student Inga Znotiņa. The corpus "Esam" is a learner corpus that consists of the texts that have been written by university students, learners of the second Baltic language; namely, Latvian for students of Lithuanian background, and Lithuanian for students of Latvian background. (Znotiņa, 2015; Znotiņa, 2017).

The paper is further structured as follows: section 2 describes the creation stages of the corpus, section 3 gives an introduction to the error annotation guidelines, section 4 describes the automated processing of the data, section 5 explains the computing of the statistics of the annotated errors. The paper is concluded in section 6.

2. The Creation Stages of the Corpus

There are several stages of creating the corpus:

1. Data digitalization;
2. Text correction;
3. Automated morphological annotation, including tokenization, part-of-speech tagging, lemmatization;
4. Original and corrected text alignment;
5. Automated error annotation and manual revision.

First, texts are digitalized by manually transcribing hand-written test. The transcriptions fully correspond to the original document (test). Sometimes handwriting deciphering causes difficulties.

After data digitization, the texts are manually corrected. The texts are overwritten according to the norms of the Latvian language. All spelling, grammatical, lexical and punctuation errors are corrected. If there is a redundant word in the sentence, it is deleted, while the released word is written in the sentence (syntactical error). To be able to align words, inadequate word order is not changed, but it will be annotated. If some portion of the text is unclear, it is left unchanged, and it will be annotated.

Further, the data is automatically processed. Original and corrected text is tokenized, morphologically annotated and aligned. From the alignments, initial error annotations are generated and prepared for manual revision.

3. Criteria of Error Annotation

Learner corpora are usually error annotated, that is, spelling (orthographic), lexical, and grammatical errors in the corpus have been annotated with the help of a standardized system of error tags (Granger, 2003).

The texts are error annotated using an error taxonomy created for the Latvian language (Table 1). Similar error taxonomy is used in the learner corpus of the second Baltic language "Esam" (Znotiņa, 2015). This error taxonomy can be used for other inflected languages with free word order.

Type	Subtype
Spelling errors	Upper / lower case letter
	Diacritics
	Separately / together spelled words
	Missing letters
	Redundant letters
	Other spelling errors
Punctuation errors	Missing punctuation
	Redundant punctuation
	Incorrect punctuation
Grammatical errors	Incorrect word form (such as inflection, gender, number, definite/indefinite ending, tense, person)
	Derivation
	Morphophonemic consonant alternation
Syntactic errors	Word order
	Redundant word
	Missing word
Lexical errors	Meaning
	Compliance
	Readability
	Collocation
Unclear text	

Table 1: Error types

Most of the spelling and grammatical errors are tied to a single token, but there are some constructions, that consists of multiple words, for example, analytical forms. In these cases, it is necessary to be able to annotate multiword expression as a single token.

If in some segment the word order is incorrect, it is not changed, because it will make automatic alignment a lot more difficult and sometimes even impossible. Other errors are still annotated in these segments, and the text segment is marked as one with wrong word order.

On the contrary to the English language, in Latvian, punctuation is very important. The punctuation is based on the grammatical principles, and the different use of punctuation marks often completely change the meaning of the sentence.

Occasionally the spelling errors may overlap with grammatical errors. Error annotation system, therefore, should allow annotating several types of errors (usually grammatical and spelling errors) for one wordform simultaneously.

There are ambiguous errors, for example, one missing diacritic can change the grammatical meaning, but the misuse of diacritics is a common error in Latvian learners' texts as well. In these cases, both error types (grammatical and spelling) are annotated.

4. The Automated Processing of the Data

The automated processing consists of three steps:

1. Tokenization and morphological analysis;
2. Text alignment (including token and letter level analysis);
3. Automatic error annotation.

Each of this step is described in more details in the following subsections.

4.1 Morphological Analysis

First, the original text and corrected text are tokenized and automated morphological annotations are generated. Morphological annotation consists of a morphological tag (including part of speech), lemma and stem. In most cases, only the morphological information from the corrected text is used. Although the morphological annotation is done for the original text as well, this information is often inaccurate because of the many grammatical errors. Morphological information from the original text is used only when there is no corresponding word in the corrected text, i.e., the word was redundant in the original text, and it was deleted in the corrected text.

For Latvian the morphological annotator developed by Paikens was used (Paikens, 2013).

4.2 Text Alignment

The tokens are aligned, using word level alignment into one-to-one relationships, where each token in the corrected text has one or none aligned tokens in the original text and vice versa. The alignment is found by using a similar approach to the one used in word error rate calculations in speech recognition. The token relationships are found by computing the alignment with the lowest edit distance. The edit distance is calculated as follows:

- The cost of deleting a token is 1.
- The cost of inserting a token is 1.
- The cost of substituting a token is the relative edit distance between tokens.

The relative edit distance is obtained by computing the edit distance between tokens and dividing it by the length of the longest token, so the value is between zero and one. If the cost of the substitution were 1, the same as in speech recognition tasks, in segments with insertions/deletions and many spelling errors, there would be multiple alignments for the same cost, because there would be no way how to tell which token is the inserted/deleted one.

After token level alignment, the next step is letter level alignment for the substituted tokens. The letter level alignments are used to generate automatical error annotations and to improve user experience in manual error labeling by emphasizing the differences in two tokens. A significant portion of spelling errors is an incorrect use of diacritical marks or letter case, ignoring them when computing letter alignment helps to get the correct alignment especially when if there are some missing or redundant letters.

4.3 Automatic Error Annotation

Automatic error annotations, which later will be manually edited using annotation revision interface (Figure 1), are generated by a rule-based system from the alignments and morphological annotations.

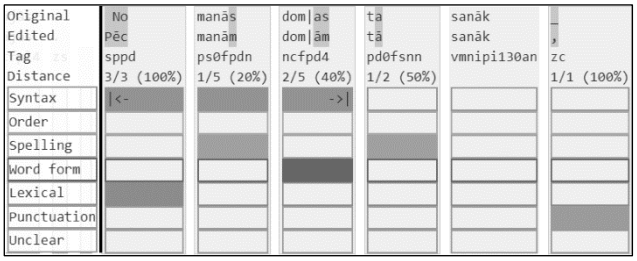

Figure 1: Error annotation revision interface

The order of the rules is important because after the first applicable rule is found, the evaluation of the rules is stopped.

The rules go as follows:

- If both tokens (the original and the corrected one) matches, there is no error.
- If the token consists only of punctuation marks, it is punctuation error.
- If one of the tokens is missing (it was a redundant or missing word), it is a syntax error.
- If the relative edit distance between tokens is greater than 0.8, it is considered that the word is most likely replaced with a different word and it is a lexical error.
- If none of the rules above applied, it can be one or both of two error types – spelling or grammatical error.

Letter level alignments and morphological information are used to determine if it is spelling or grammatical error. It is annotated as a grammatical error if the differences between two tokens are at the ending of a word. Otherwise, it is a grammatical error. The token contains grammatical and spelling errors if the differences are at the beginning of the word and the ending of the word. For the words in the corrected text, the boundary between the beginning and end of the word is obtained from automatical morphological annotations. For the words in the original text, the boundary is projected from corrected text using letter level alignments.

5. The Analysis of Annotated Errors

The analysis of any data could be divided into two types – quantitative and qualitative analysis. In quantitative analysis, the data is grouped by some feature, for example, by the misspelled letter. For this analysis, it is necessary to know what kind of feature meaningful statistics could be obtained and how to get this feature from the data automatically. If meaningful features are not known, or it is not possible to extract them automatically, qualitative analysis is an option where one tries to identify the features or extract them manually.

The categories used in error annotation tool include all of the error types. The possibilities for automatic error subtype determination and other meaningful feature extraction differs for each error type. The available options for quantitative analysis of the error corpus from the annotations suggested in this paper will be discussed in this section.

5.1 The Analysis of Spelling Errors

For spelling errors, it is possible to do a quantitative analysis of subtypes from words with spelling errors using letter level alignments. The subtype analysis is done only for words that contain only spelling errors because if the words contain grammatical errors as well, it is hard to automatically differentiate which inconsistencies in letter level alignment are due to grammatical errors and which due to spelling errors. In many cases, it is also hard to manually differentiate between grammatical and spelling errors.

5.2 The Analysis of Punctuation Errors

Punctuation errors are the simplest error type. With quantitative analysis, it would be possible to show which punctuation errors are the most frequent. More complex quantitative error analysis could be added as well, for example, investigating in what context commas are missing or redundant most frequently. Commas are important in languages with free word order.

5.3 The Analysis of Grammatical Errors

The simplest quantitative analysis of grammatical errors could be done from the morphological annotations of corrected text to determine in which part of speech (such as nouns, pronouns, verbs, adjectives), inflection, tense, person, etc., learners make most mistakes.

5.4 The Analysis of Syntactic Errors

For syntactic errors, the relative percentage of text segments with syntactic errors in different language levels (A1 to C2) can be quantitatively analyzed.

In the error annotations, syntactic errors are annotated in two subtypes – word order errors and other syntactic errors. Word order errors are annotated separately because these errors are not corrected. The main reason is that the alignment approach used currently assumes the order in both texts are the same.

For word order errors, no other quantitative analysis is possible because the corrected text is not available. In other syntactic errors, the correct text is available, so more detailed quantitative analysis is possible for this subtype.

5.5 The Analysis of Lexical Errors

For lexical errors, a meaningful quantitative analysis is not straightforward. Because of the spelling and grammatical errors, the original words cannot be grouped directly. To work around this problem original words could be grouped based on similarity. If the differences between some group of words look like spelling errors (for example, different use of diacritics), these words could be considered to be the same and grouping them would provide more meaningful quantitative analysis. Further research is required to make better conclusions about the best approach for the analysis of this error type.

5.6 The Analysis of Unclear Text

Segments that can't be understood are annotated as unclear text. Similar to the word order errors, the relative percentage of text segments with unclear text in different language levels (A1 to C2) can be quantitatively analyzed.

6. Inter-Annotator Agreement

To evaluate inter-annotator agreement 20 documents containing 1942 tokens were annotated by two users. The annotation was done in two steps. First, the text was rewritten by each user individually. Then, each user annotates errors on their rewritten version of text.

Comparing the texts rewritten by each user, 92.7% of tokens matched (1800 out of 1942 tokens). Error level inter-annotator agreement was calculated only on matched tokens. The number of tokens annotated with different error classes only by User A, User B or equally by both of the users are shown in Table 2. The inter-annotator agreement was measured with Cohen's kappa coefficient (κ) (Cohen, 1960). The value is within the interval [−1, 1], where κ = 1 means perfect agreement, κ = 0 agreement equal to chance, and κ = −1 "perfect" disagreement.

Error Type	User A	User B	Both	κ
Spelling	9	23	219	0.85
Grammatical	13	9	70	0.83
Lexical	19	3	8	0.40
Punctuation	1	1	74	0.98
Unclear text	17	23	2	0.04
Word order	13	0	0	0.00
Syntactical	0	6	3	0.49

Table 2: Inter-annotator agreement

7. Corpus Statistics

The corpus contains 142684 tokens from 1496 documents. On average 22.2% tokens contained errors. The distribution of different error types in the corpus is given in the Table 3. Percentages relative to tokes with errors sums up to more than 100% because one token can contain multiple errors.

Error Type	Count	Percentage from tokens with errors	Percentage from total tokens
Spelling	14956	47.13%	10.48%
Grammatical	8075	25.45%	5.66%
Punctuation	5857	18.46%	4.10%
Lexical	1756	5.53%	1.23%
Word order	1703	5.37%	1.19%
Unclear text	1546	4.87%	1.08%
Syntactical	1321	4.16%	0.93%

Table 3: The distribution of different error types

To evaluate how good the naive error prediction system works, the number of tokens marked with different error types only by user, only by system or both was calculated (Table 4). Correctness was chosen as a measurement of the system's performance. Correctness is the percentage of the unchanged tags from the total number of tokens that contained any type of error. The error prediction system was developed to speed up the annotation process, it wasn't meant to be 100% correct. Examining the statistics it can be can concluded that error prediction system predicts a spelling error when it is actually a grammatical error. This is something that could be improved. The system's current version will never predict a lexical error. The time spent on building a system that predicts lexical errors might not be worth it because It is hard to predict this kind of errors and the inter-annotator agreement for lexical errors were also significantly lower than for other error types. Lexical errors are also less common than Spelling, grammatical or punctuation errors.

Error Type	User	System	Both	Correctness
Punctuation	161	22	5696	99.42%
Lexical	1756	0	0	94.47%
Grammatical	1075	745	7000	94.26%
Spelling	299	2090	14657	92.47%

Table 4: The correctness of error prediction system

8. Conclusion and Further Work

The error annotation method suggested in this article proved to be easily understandable and usable for the annotators. The time the annotation process took was similar to the time necessary for classical annotation process. The use of text correction and alignments enables opportunities for a lot more detailed quantitative statistical analysis.

As mentioned earlier, the biggest drawback of this error annotation approach is limitation on word order errors, but there are many flective languages (for example, most of the Slavic languages) for which the word order is not grammatically significant. In the Latvian learners' corpora inter-annotator about word order error was 0 (close to chance).

The development of automated text correction process would give the highest impact to annotator's experience and would reduce the time necessary for the development of the corpus.

Revision of the current automatic error annotation rules and refinement from the lessons learned during the development of the corpus could improve user experience.

9. Acknowledgements

This work has received financial support from the Faculty of Computing, University of Latvia.

The results reported in this paper are part of the Latvian Language agency's research project "Quality of the Latvian language: results of the state language proficiency test".

The tools developed and used in this project has received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219.

10. Bibliographical References

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Granger, S. (2008) Learner corpora. Corpus Linguistics : An International Handbook. Anke Lüdeling, Merja Kytö (eds.). Berlin, New York : Walter de Gruyter, 2008, pp. 259–275.

Granger, S. (2003) Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, Vol. 20, No. 3, 2003, pp. 465–480.

Paikens, P., Rituma, L., Pretkalniņa, L. (2013) Morphological analysis with limited resources: Latvian example, Proceedings of NODALIDA 2013, pp. 267–278.

Znotiņa, I. (2017) Computer-Aided Error Analysis for Researching Baltic Interlanguage. Rural Environment, Education, Personality. Proceedings of the 10th International Scientific Conference, 2017, pp. 238–244. http://llufb.llu.lv/conference/REEP/2017/Latvia-Univ-Agricult-REEP-2017_proceedings-238-244.pdf

Znotiņa, I. (2015) Learner corpus annotation in Latvia and Lithuania. Sustainable Multilingualism, No. 7. 2015, pp. 145–159.