Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200629

Detailed Error Annotation for Morphologically Rich Languages: Latvian Use Case

Roberts DARĢIS [1], Ilze AUZIŅA, Kristīne LEVĀNE-PETROVA and Inga KAIJA
Institute of Mathematics and Computer Science, University of Latvia
Rīga Stradiņš University

Abstract. This paper presents a detailed error annotation for morphologically rich languages. 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*. There is no need for an advanced multi-token error annotation schema, because error annotated texts are written by beginner level (A1 and A2) who use simple syntactic structures. This schema focuses on in-depth categorization of spelling and word formation errors. The annotation schema will work best for languages with relatively free word order and rich morphology.

Keywords. Leaner corpus, error annotation, language acquisition, corpus development

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 with consistently annotated errors. Annotation schema is one of the most important aspects of a learner's corpus. A detailed error annotation schema provides a wide range of statistical analysis, enabling researchers to conduct numerous kinds of quantitative research, and allows the development of fine-grained search that enables research to quickly find the information of interest for qualitative analysis with no need to go through a lot of redundant information.

This paper presents the error annotation schema used in the development of 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. Latvian is a language with rich morphology and a relatively free word order. Latvian can be generally considered a phonetic language, i.e. 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

[1] Corresponding Author: Roberts Darģis, Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Riga, LV-1459, Latvia; E-mail: roberts.dargis@lumii.lv.

and diphthongs, a high degree of inflection and a rather free word order. These properties have to be taken into account in the error-annotation process.

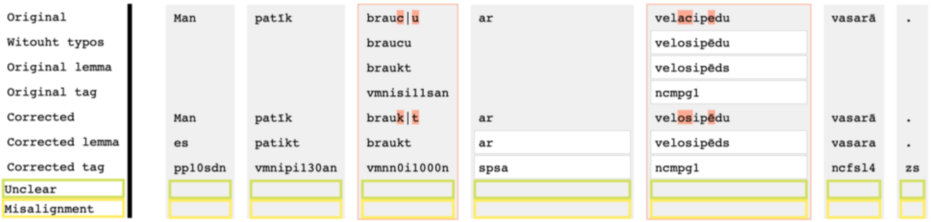
2. Related Work

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 [1] among others. However, more and more learner corpora are being developed for other languages as well, many of which are morphologically rich [2], [3].

Usually, errors are grouped according to language level (phonetics, morphology, syntax, etc.); the linguistic category to which the error belongs and the changes that occur when comparing the original and corrected texts (omission, addition, misformation, etc.) [4], [5], [6]. Although error annotation schemas for morphologically rich languages have more detailed error categories and subcategories [7], [8], [9], manually defined categories will never be comprehensive. In Latvian, there are more than 2,000 morphological tags of which about 200 are used to describe nouns and more than 1,000 to describe verbs, which leads to many possible error combinations. A lot of information would be lost using even as many as 100 error codes. In the LaVA corpus, a different approach is used after text correction. Instead of using a limited set of error codes, only morphological information is annotated and more fine-grained error codes are automatically extrapolated from the morphological information.

3. Error Annotation Schema

The error annotation is done on the alignment between the original text and the corrected text [10]. A commonly used error taxonomy for Latvian includes 5 types (Spelling errors, Punctuation errors, Grammatical errors, Syntactical errors and Lexical errors) and multiple subtypes (for example, subtypes for Spelling errors: Upper/lower case letter, Diacritics, Separately/together spelled words, Missing letter, Redundant letter, Other spelling errors) [6], [10]. These error codes are not directly used in the LaVA corpus; instead, more detailed error codes are extrapolated for five other properties, which are much more easier to annotate. These properties are: original token without typos (the token written by the learner with corrected spelling errors), original lemma, original tag, corrected lemma, and corrected tag (figure 1).



Original	Man	patīk	brauc|u	ar	velacipedu	vasarā	.
Witouht typos			braucu		velosipēdu		
Original lemma			braukt		velosipēds		
Original tag			vmnisillsan		ncmpgl		
Corrected	Man	patīk	brauk|t	ar	velosipēdu	vasarā	.
Corrected lemma	es	patikt	braukt	ar	velosipēds	vasara	.
Corrected tag	pp10sdn	vmnipi130an	vmnn0i1000n	spsa	ncmpgl	ncfsl4	zs
Unclear							
Misalignment							

Figure 1. Error annotation interface

The spelling errors can be determined automatically by comparing the original token with the original token without typos. Character level alignment combined with a rule based system allows to extract exactly which character pairs are used incorrectly. This information can later be used to facilitate qualitative research by providing fine-grained search or quantitatively grouping the extracted character errors. To specify misspellings of together or separately written words, adjacent units are marked/pulled together.

A morphological tag contains a lot of information, including part of speech (Pos) tag. There is a tag for punctuation marks, so recognizing punctuation errors is straightforward – if the corrected token is different from the original token and the tokens are punctuation marks, it is a punctuation error.

Lexical errors mean that the lemma of the corrected token is different from the lemma of the original token. Subtype cannot be determined automatically, but it can be added later for unique token pairs only, because the subtype is not context dependent.

The remaining errors are grammatical errors. A very detailed grammatical error analysis can be done based on the morphological tags.

In addition, two more properties can be annotated – *unclear* and *misalignment*. Both of these properties are there just as percussion. *Misalignment* is meant for cases where alignment is not correct, for example, in the alignment, it shows that one token is replaced with another, but actually one is removed and the other one is added independently. *Unclear* is used for cases in which it is not clear what annotations should be added or there is a wider context that impacts the error and that cannot be annotated in the current scheme, for example, a prepositional construction should be used instead of the word form, or an analytic form of verb is used. Such cases are summarised and discussed to decide the correct annotations and update the error annotation scheme if necessary, and to annotate errors at the syntax level.

4. Semi-automatic Annotation Generation

All the values of the properties mentioned in the previous section are generated automatically. There are two types of values: those that can be edited and those that are read-only (figure 1).

The read-only values are obtained from a manually annotated and verified list of tokens which occurred at least 3 times and have only one possible lemma and tag regardless of context.

The suggestions for the rest of the property values are acquired from a morphological annotator [11]. Tag and lemma for punctuation marks and numerals are considered to be correct and are also read-only.

The morphological suggestions for the original tokens are highly inaccurate due to the high amount of typos. If the original form is not in the dictionary and the words are similar, it is considered to be the same token with typos and annotations from the corrected token are suggested.

5. Conclusion

The error annotation method proposed in this paper is tested in the LaVA corpus development. The corpus consists of error annotated texts written by beginner level (A1 and

A2) language learners. There is no need for an advanced multi-token error annotation schema, because beginners use simple syntactic structures. Most of the errors are limited to individual tokens. This schema has more detailed categorization of spelling and word formation errors. These errors are more common and much more diverse for beginner level compared to intermediate and advanced level. Further work includes review of unclear segments and extending annotation schema to support syntax errors and more complex multi-word structure annotation if necessary.

Acknowledgment

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 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.

References

[1] Granger S, Dagneaux E, Meunier F, Paquot M, et al.. International corpus of learner English. UCL, Presses Univ. de Louvain; 2009.

[2] Siemen P, Lüdeling A, Müller FH. FALKO-ein fehlerannotiertes Lernerkorpus des Deutschen. In: Proceedings of Konvens. vol. 2006; 2006. p. 107.

[3] Rakhilina EV, Vyrenkova A, Mustakimova E, Ladygina A, Smirnov I. Building a learner corpus for Russian. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition; 2016. p. 66–75.

[4] Tono Y. Learner corpora: design, development and applications. In: Proceedings of the Corpus Linguistics 2003 conference. University Centre for Computer Corpus Research on Language Lancaster; 2003. p. 800–809.

[5] James C. Errors in language learning and use: Exploring error analysis. Routledge; 2013.

[6] Znotiņa I. Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas. Liepājas Universitāte; 2018.

[7] Štindlová B, Škodová S, Rosen A, Hana J. A learner corpus of Czech: Current state and future directions. Twenty years of learner corpus research: Looking back, moving ahead. 2013:435–446.

[8] Ledbetter S, Dickinson M. Automatic morphological analysis of learner Hungarian. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications; 2015. p. 31–41.

[9] Gayo IDR, Antunes S, Mendes A, Janssen M. Towards error annotation in a learner corpus of Portuguese. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition; 2016. p. 8–17.

[10] Darģis R, Auziņa I, Levāne-Petrova K. 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 11th International Conference on Language Resources and Evaluation (LREC 2018); 2018. p. 4111–4115.

[11] Paikens P, Rituma L, Pretkalniņa L. Morphological analysis with limited resources: Latvian example. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); 2013. p. 267–277.