DOI: 10.22616/REEP.2018.020

Error Types in the Learner Corpus of the Second Baltic Language

Inga Znotina Mg. philol.
Riga Stradins University, Liepaja University, Ventspils University College, Latvia
inga.s.znotina@gmail.com

Abstract: Errors in language learning are seen as normal and even necessary. However, researching them is often undermined by the need for a clear definition what is or should be considered an error and by the lack of an error taxonomy. This paper shortly discusses the notion of error in various contexts, especially in learner corpora research. Then it offers an error taxonomy that was created for error-tagging a learner corpus of Baltic languages. The aim of the study is to create a taxonomy that is suitable for annotating beginner texts of Latvian and Lithuanian, and efficient in use. The taxonomy is based on the previous work of S. Granger who identified error types for a learner corpus of French. These error types are reviewed, modified and/or replaced where necessary in order to match the structure of Latvian and Lithuanian languages. 5 error types (form; morphology and word-formation; syntax; vocabulary; punctuation) with 29 subtypes are distinguished. Those are described in the article along with examples from the corpus. The taxonomy is now being used for annotation the learner corpus of the second Baltic language which provides researchers with valuable material on language learning outcomes.

Keywords: university education, Baltic languages, errors, learner corpus.

Introduction

One of the most prevalent characteristics of language learners' production of the target language is it having errors. Errors are also often considered when analysing learner-produced language, but not all researchers agree on the notion of error. Some divide all problematic uses of language into *slips, mistakes, errors, solecisms*, especially noting the difference between *errors* and *mistakes*. This difference not entirely clear, though. Errors are sometimes described as persistent mistakes, thus making *mistake* a more general term (Field, 2011). Others place both terms on the same level describing errors as performance issues while mistakes are considered to be based on the learner's knowledge (or lack thereof) of the target language (Cherrington, 2004). Some other languages use one general term such as Latvian *kluda*. The term *error* is used in this paper the same way because further division of errors is still to be discussed.

The need to apply the definition in annotation of corpora requires a simple understanding of errors. In learner corpora, an error usually is a deviation from a reconstruction of a correct target language structure. Such a reconstruction is sometimes called *target hypothesis* (Ellis, 1994, 54; Ludeling et al., 2005; Reznicek, Ludeling, Hirschmann, 2013). This way, the text is corrected by a native speaker of the target language (target hypothesis is created). Each mismatch between the text and the target hypothesis is considered an error. Such an approach is also adopted in the case of the learner corpus which is used in this study.

Various error taxonomies exist for various research problems and finding a suitable one is often seen as a challenge. Lack of a suitable error taxonomy is frequently one of the main problems in error analysis, and corpus linguistics is trying to solve the problem by offering learner corpora which are annotated by certain taxonomies (Dagneaux, Denness, Granger, 1998; Boyd, 2010; Hana et al., 2010). Finding the right taxonomy for a corpus is, however, another issue which this study attempts to deal with. A learner corpus of the second Baltic language (Lithuanian for Latvians, Latvian for Lithuanians) has been created, and it is decided to add error annotation to the data of the corpus. In order to do that, a suitable error taxonomy must be created. Errors in learner data of Baltic languages have been studied before, but categorization has most often been rather blurry because, mostly, only certain types have been analysed, such as errors in use of the Locative case (Laizane, 2014). Some other relevant studies are not so detailed and just give an overview of some error types without discussing classification thoroughly (Zigure, 1999; Dabasinskiene, Cubajevaite, 2009; Zujevaite, Zilinskaite, 2012).

There is a publication on error types for Latvian (Deksne, Skadina, 2014), but the offered classification has been created for an automatic error-correction tool which is supposed to help people who already have a good command of Latvian (Deksne, Skadins, 2011). Because of that, very specific error types have been distinguished, such as the use of Nominative case in debitive mood constructions. It makes the taxonomy very precise but also complicated to use and not very suitable for beginner learner texts where very basic structures are used and the whole variety of errors must be accounted for.

The overview of the current situation shows that there currently is no suitable error taxonomy for annotating errors in the learner corpus of the second Baltic language (*Esam*). The aim of the study is to create such an error taxonomy that is suitable for annotating beginner texts of Latvian and Lithuanian, and efficient in use when annotating beginner learner texts. The study aims to do that based on the error taxonomy created by S. Granger for annotating a learner corpus of French (Granger, 2003).

Methodology

The error taxonomies for learner corpora have so far been created based on two kinds of criteria:

- grouping based on the linguistic category the error belongs to (morphology, syntax, vocabulary…);
- grouping based on changes in comparison with the target hypothesis (omission, addition, misformation…).

However, some researchers argue that it is best to merge the two approaches as it lets one adapt the types of error according to the potential usefulness in research (James, 1998, 114). This way, an error taxonomy for a learner corpus of French has been created (Granger, 2003). The resulting system is general enough to cover various aspects of language and detailed enough to give some insight into the character of errors, so it was used as a basis for this study. Since S. Granger's taxonomy was created for French, it deals partly with different language structures from those found in Baltic languages, so it needs to be adapted annotating texts in Baltic languages.

Adaptation is carried out using various descriptions of Latvian and Lithuanian language systems, such as (Kalnaca, 2014; Praulins, 2012) and (Ambrazas, 2006) as well as various publications in the field of teaching and learning languages. The study attempts to answer the research question: what error groups and/or subgroups could form an effective error taxonomy for beginner learner texts of the second Baltic language?

The process is divided into two stages. First, the error types proposed by S. Granger are reviewed and evaluated for matching the systems of Baltic languages. If necessary, certain error types or subtypes are added, deleted, changed or merged. The second stage is annotating texts which allows for further evaluation of each subtype and adding or changing them as needed.

Results and Discussion

The error taxonomy created for the learner corpus *Esam* has already been shortly introduced in (Znotina, 2017). The error types and their respective attributes in annotation have been presented there. In this paper, examples of errors are given instead of attributes. The examples of errors are authentic data from the learner corpus *Esam*.

In each one of the tables, the error matching the respective subtype has been underlined. Where necessary, correction has been provided in brackets, and an approximate translation into English is given. Note that not all error subtypes have examples from the data of both languages.

The first error type is errors of form. Those are the errors that have to do with spelling. This error type consists of four subtypes (Figure 1). Subtype *agglutination* consists of errors where words that are supposed to be written as separate are written together (as one word), or vice versa – a word is erroneously divided into two. Subtype *upper / lower case* is for the errors where capital letters are used unnecessarily or are not used when needed. Subtype *diacritics* is for the issues with any missing or redundant diacritics. The fourth subtype is for any other spelling errors, including typos.

Error subtype	Example in Latvian	Example in Lithuanian
Agglutination	*biju zveru darzā* (zvērudārzā) 'I was in a zoo'	*širdyje kaž kas* (kažkas) *suvirpėja* (suvirpa) 'something trembles in the heart'
Upper / lower case	*...un Viņai* (viņai) *patīk...* 'and she likes'	*Olimpinėje* (Olympinėse) *Žaidynėse* (žaidynėse) 'in the Olympic Games'
Diacritics	*Viņas* (Viņš) *ir uzņemējs* (uzņēmējs) 'he is an entrepreneur'	*dažnai nera* (nėra) *pakankamai laiko* 'often there is not enough time'
Other spelling errors (including typos)	*man patīk tiktis* (tikties) *ar draugiem* 'I like to see my friends'	*kikvieną* (kiekvieną) *dieną* 'every day'

Figure 1. Error classification in learner corpus *Esam*: errors of form.

Syntax error subtypes are given in Figure 2. Syntax type is for errors that have to do with the ties between words, word order, and excess or shortage. There are four error subtypes. Word order subtype consists of cases where mostly correctly chosen words are put in an incorrect sequence. Word missing, and word redundant subtypes include cases where a word is respectively omitted or added unnecessarily. Cohesion error subtype applies to errors of matching words together in a coherent structure if the error is grammatical. If the nature of the error is lexical, then the error belongs to the vocabulary type, compatibility subtype.

Error subtype	Example in Latvian	Example in Lithuanian
Word order	*..radoša tik* (tikai) *dēļ naudas* (naudas dēļ) 'creative only because of money'	*Vieta, kur visada aš* (aš visada) *galiu grįžti...* 'Place where I can always return…'
Word missing	*tā vārds* (vārds ir) *Džekis* 'its name is Džekis'	*Biologijos fakultete yra labai daug* (ko?) 'In the Faculty of Biology there is a lot of (what?)'
Word redundant	*..ceļiauju* (ceļoju) *uz Klaipēdu būt* (-) *brivdienās* (brīvdienās) 'I travel to Klaipeda for holidays'	*ji pasiūlė man kartu su ja reikėjo* (-) *ruošti pjesę* 'she offered me to prepare a play together with her'
Cohesion	*Mans* (Manas) *mātes vārds ir...* 'My mother's name is…'	*aš nešioju kepures, irgi* (ir) *pirštines* 'I wear hats and gloves'

Figure 2. Error classification in learner corpus *Esam*: errors of syntax.

Figure 3 offers morphology and word-formation error subtypes. This is the most diverse error type consisting of fifteen subtypes altogether. Derivation subtype is for those cases when a new word has been derived from elements of source and / or target language. Compounding subtype is similar, only here new words have been created by combining existing words. Inflection subtype consists of errors where the wrong case is used (but if the case is right and there are only issues in the form, then the error belongs to the form type, subtype of other spelling errors). The gender subtype includes the words used in the incorrect gender, while words used in incorrect number belong to the number subtype. The next subtype is for errors where the definite ending is used instead of an indefinite ending or vice versa. Degree of comparison error subtype is for errors where an adjective or adverb is used in superlative instead of comparative or similarly. Person subtype includes verbs used in the wrong person. Tense subtype is for incorrectly used time forms, whether it is using future instead of past (or similar), or using simple tense where complex tenses are needed, or vice versa. In the mood subtype are errors where, for example, conditional mood is used instead of indicative mood or similar. Voice subtype is for misuse of active voice and passive voice. Reflexivity subtype is for the cases when a reflexive form is used unnecessarily or not used when needed; or for incorrect reflexive forms. Participle confusion subtype is for misused or misformed participles. Perfective subtype is for misuse of prefix-verbs that indicate

finished actions in Baltic languages, and interactivity subtype consists of errors using forms that express one-time actions versus forms that express repeated actions.

Error subtype	Example in Latvian	Example in Lithuanian
Derivation	*patīk futbols, basketbols, vazinātes* (braukāt) *ar ritēni* (riteni) 'I like football, basketball, riding a bike'	*todėl užmiegojome* (užmigome) *anksti* 'so we fell asleep early'
Compounding	-	*aerouostas* (oro uostas) 'airport'
Inflection	*Es gribu pastāstīt par mana* (manu) *ģimeni* 'I want to tell about my family'	*didelė dalis drabužiai* (drabužių) *yra tokios spalvos* 'a great part of clothes are that color'
Gender	*Mans* (Manas) *acis ir brūnas.* 'My eyes are brown.'	*Jos* (Jie) *visi yra šalia* 'They are all nearby'
Number	*es biju ļoti skumīga šoreiz pār* (par) *atvaļinājumiem* (atvaļinājumu) 'I was very sad about the vacation this time'	*įvariuose* (*įvairiomis*) *gyvenimo valandą* (valandomis) 'in various times of life'
Definite / indefinite ending	*Fotoaparātā bija manas skaistas* (skaistās) *fotogrāfijas* 'My beautiful photos were in the camera'	*žmonių kamšatis ir ilgoji* (ilgos) *valandos viešajame transporte* 'crowding people and long hours in public transportation'
Degree of comparison	-	*Aš esu jaunesnioji* (jaunausia). 'I am the youngest'
Person	*Tēvs interesējies* (interesējas) *par automobiliem* (automobiļiem) 'Father is interested in cars'	*aš nebuvo* (nebuvau) *name* (namie) 'I was not at home'
Tense	*viņai patīk ceļot*(,) *un māte apmeklēja* (ir apmeklējusi) *Krieviju, Franciju...* 'she likes to travel, and mother has visited Russia, France…'	*aš pasibundu* (pasibudau), *nes buvau labai alkane* 'I woke up because I was very hungry'
Mood	-	*Aš esu dėkinguma* (dėkinga), *ka* (kad) *sutikčiau* (sutikau) *jai* (ją) 'I am grateful I met her'
Voice	-	*Į ją galėtų* (būtų galima) *įeiti ir iš lauko* 'One could go in it also from the outside'
Reflexivity	*Pie sienām karās* (karāsies) *bērnu zīmējumi.* 'Children's drawings will hang on the walls'	*netrukdėme ir neriejomės* (nesiriejome) 'we did not disturb and did not fight'
Participle confusion	*mēs bijam* (bijām) *ļoti noguras* (nogurušas) 'we were very tired'	*vairuotojas, matytint* (matydamas), *kad bėgtu* (bėgu), (..) *pristabdau* (pristabdo) 'driver stops as he sees me running'
Perfective	-	*Kada ji ėjo* (atėjo) *iš darbo...* 'When she came from work…'
Iterativity	-	*mama man* (mane) *išmokydavo* (išmokė) *nekada nepasiduoti* 'mom taught me to never give up'

Figure 3. Error classification in learner corpus *Esam*: errors of morphology and word-formation.

Figure 4 shows the vocabulary error subtypes of the current taxonomy. These types categorize mistakes of lexical meanings. Of the three subtypes, the first one, meaning subtype, is for cases when a word's meaning does not match that of the sentence in which it is intended to be used. The other subtype, compatibility, is for cases when meanings of each used word match the meaning of the sentence but are not compatible with each other. In the stable phrase subtype are errors where the learner has unsuccessfully

tried to make a construction that exists as a stable phrase in the target language. Even if there is seemingly no meaningful difference, the person who corrects the text chooses a different word.

Error subtype	Example in Latvian	Example in Lithuanian
Meaning	*pelēks, biezs* (resns) *un ļoti labs kaķis Benas* 'grey, fat and very good cat Benas'	*Ne tik katris* (kiekvienas) *latvis* 'not only every Latvian'
Compatibility	*zils paklājs, kurš der* (piestāv) *pie sienu* (sienām) 'blue carpet that matches the walls'	*nes esame tiek* (tokios) *įvairios* 'because we are so different'
Stable phrase	*..braukšu uz ciemus* (ciemos) 'I will go to visit'	*Aš tą* (tai) *labai įvertinu* (vertinu) 'I appreciate it very much'

Figure 4. Error classification in learner corpus *Esam*: errors of vocabulary.

Errors of punctuation are divided into three subtypes as shown in Figure 5. The punctuation confusion subtype consists of errors where the learner has correctly decided that punctuation is needed but chosen the wrong punctuation. Punctuation redundant is for errors where there are unneeded punctuation marks, and punctuation missing is for those where some punctuation is needed but the learner has not used it.

Error subtype	Example in Latvian	Example in Lithuanian
Punctuation confusion	-	*dar kartą užmigau..* (.) 'I fell asleep once more'
Punctuation redundant	*Tāpēc*, (-) *es biju ļoti skumīga* 'Because of that I was very sad'	*Trečią valandą naktį* (nakties), (-) *aš...* 'at three in the night I...'
Punctuation missing	*Viņai patīk ceļot(,) un māte apmeklēja* (ir apmeklējusi)... 'She likes to travel, and mother has visited…'	*Viskas būtu(,) kaip aš norėčiau.* 'Everything would be as I would want.'

Figure 5. Error classification in learner corpus *Esam*: errors of punctuation.

Each error is given only one, most likely target hypothesis and only one, most likely error type. That is done in order to not make the annotation process too complicated. There are sometimes issues where one error can be considered a mix of several types – in the Latvian example *mans* (manas) *acis ir brunas* 'my eyes are brown' one can see a gender error, a number error or a common spelling error. In such cases, the annotator chooses the type and subtype that he/she finds most fitting. Overly thorough analysis would extremely slow down the annotation process, so the experience and intuition of the annotator is taken into account. Further discussions in each separate case could, however, lead to corrections in the corpus as well.

It is possible that correcting a mistake makes something else in the sentence incorrect. Such an example can be seen in the case of the Latvian sentence *Vinas uzbuve ir smalka, speciga [..]*. 'Her build is fine, strong' the author of the text has correctly matched several adjectives with a noun in feminine, but the corrector changed the noun to a masculine one – *augums* 'figure'. It means that gender of the adjectives must be changed too. It is done in the correction, and the matching error type is selected. Nevertheless, careless evaluation of such examples can lead one to incorrect conclusions about the learners' noun and adjective matching skills (or lack thereof). For that reason, it is necessary to note that the presence of an error in a text does not always point out to actual flaws in the language learner's skills.

Similarly, the number of errors can artificially increase if there is an error type that must be assigned to multiple tokens, such as the word order errors. It means that one error can be counted as double, and such calculations can significantly affect the results if the researched data is not carefully reviewed.

The taxonomy may need some editing later on if texts of higher skill level are being annotated. Such error types as *style* and *register* were offered by S. Granger but rejected in this taxonomy because the first texts someone writes in a target language can hardly be corrected for matching styles and registers.

Further research could also shed some light on some problematic issues where selecting one error type over another can be controversial.

Conclusions

The aim of the study has been reached and the research question of the study has been answered by separating error groups and subgroups and creating an effective error taxonomy for annotating beginner learner texts of the second Baltic language. The taxonomy described in the present paper is being used in annotation of the publicly available learner corpus of the second Baltic language – *Esam* (*http://www.esamkorpuss.lv/*). It can also be used in other corpora as long as the texts are written by beginners and the target language is one of the Baltic languages.

The taxonomy can now be used to discover various peculiarities of learner language, and it also allows for the use of statistical methods, as the number of errors belonging to a certain type can be counted. However, one must bear in mind that numbers can be misleading due to some practical aspects of correcting and annotation, and the quantitative measures should not be expected to always show true correlation with the learners' actual language skill level.

Bibliography

1. Ambrazas V. (Ed.). (2006). *Lithuanian Grammar*. Vilnius: Baltos Lankos. Retrieved from http://lukashevichus.info/knigi/ambrazas_lithuanian_grammar.pdf
2. Boyd A. (2010). EAGLE: An Error-Annotated Corpus of Beginning Learner German. In N. Calcoari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the International Conference on *Language Resources and Evaluation*, 7. Valetta, Malta, 1897-1902. Retrieved form http://www.sfs.uni-tuebingen.de/~adriane/papers/boyd-lrec-2010.pdf
3. Cherrington R. (2004). Error analysis. In M. Byram (Ed.), *Routledge Encyclopedia of Language Teaching and Learning*. London, New York: Routledge, 198-200.
4. Dabasinskiene I., Cubajevaite L. (2009). Acquisition of Case in Lithuanian as L2: Error Analysis. In *Eesti Rakenduslingvistika Uhingu aastaraamat*, 5. Tallinn: Eesti Keele Sihtasutus, 47-66.
5. Dagneaux E., Denness S., Granger S. (1998). Computer-aided error analysis. In *System*, 26(2), 163-174.
6. Deksne D., Skadins R. (2011). CFG Based Grammar Checker for Latvian. In B. S. Pedersen, G. Nespore, I. Skadina (Eds.), *NODALIDA 2011 Conference Proceedings*, 11, 275-278. Retrieved from http://dspace.ut.ee/bitstream/handle/10062/17341/0Deksne_Skadins_26.pdf
7. Deksne D., Skadina I. (2014). Error-Annotated Corpus of Latvian. In A. Utka, G. Grigonyte, J. Kapociute-Dzikiene, J. Vaicenoniene (Eds.), *Human Language Technologies – The Baltic Perspective*, 6. Kaunas, Lithuania: IOS Press, 163-166. Retrieved form https://www.researchgate.net/profile/Inguna_Skadina/publication/266402825_Error-Annotated_Corpus_of_Latvian/links/543267ec0cf225bddcc79c92/Error-Annotated-Corpus-of-Latvian.pdf
8. Ellis R. (1994). *The Study of Second Language Acquisition*. Oxford, UK: Oxford University Press.
9. Field F.W. (2011). *Key Concepts in Bilingualism*. New York: Palgrave Macmillan.
10. Granger S. (2003). Error-tagged learner corpora and CALL: A promising synergy. In *CALICO Journal*, 20(3), 465-480. Retrieved from https://www.researchgate.net/publication/228602144_Error-tagged_learner_corpora_and_CALL_A_promising_synergy
11. Hana J., Rosen A., Skodova S., Stindlova B. (2010). Error-tagged learner corpus of Czech. In N. Xue, M. Poesio (Eds.) *Proceedings of the Fourth Linguistic Annotation Workshop*. Uppsala, Sweden: Association for Computational Linguistics, 11-19.
12. James C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York: Longman.
13. Kalnaca A. (2014). *A Typological Perspective on Latvian Grammar*. Warsaw, Berlin: De Gruyter Open.
14. Laizane I. (2014). Difficulties in Acquisition of Latvian as a Foreign Language Learning the Locative. In *Language in Different Contexts*, VI (1). Vilnius: Lithuanian University of Educational Sciences, 218-225.
15. Ludeling A., Walter M., Kroymann E., Adolphs P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*. Birmingham, UK: University of Birmingham. Retrieved from https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf

16. Praulins D. (2012). *Latvian: An Essential Grammar*. Abingdon, UK: Routledge.
17. Reznicek M., Ludeling A., Hirschmann H. (2013). Competing target hypotheses in the Falco corpus. In A. Diaz-Negrillo, N. Ballier, P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data*, 59. Amsterdam, Philadelphia: John Benjamins Company, 101-123.
18. Znotina I. (2017). Computer-Aided Error Analysis for Researching Baltic Interlanguage. In V. Dislere (Ed.), Proceedings of the International Scientific Conference *Rural Environment. Education. Personality (REEP)*, 10, 238-244. Retrieved from http://llufb.llu.lv/conference/REEP/2017/Latvia-Univ-Agricult-REEP-2017_proceedings-238-244.pdf
19. Zigure V. (1999). Biezak sastopamas kludas, apgutot latviesu valodas elementarkursu (Most Common Errors When Learning Elementary Latvian). In A. Veisbergs (Ed.), *Sastatama un lietiska valodnieciba. Kontrastivie petijumi.* Zinatniskie raksti, VIII. Riga: Latvijas Universitate, 107-113. (in Latvian)
20. Zujevaite A., Zilinskaite E. (2012). Latviu kalbos kaip uzsienio kalbos tekstynas (Corpus of Latvian as a Foreign Language). In *Studentu moksliniai tyrimai 2011/2012.* Konferencijos pranesimu santraukos. Vilnius: Lietuvos mokslo taryba, 55-57. (in Lithuanian)