Computer-Aided Error Analysis for Researching Baltic Interlanguage

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

Abstract: In teaching and learning languages, error analysis as a methodology has been criticized by many scholars. However, errors have also been noted as an important part of interlanguage which should not be ignored if the aim is to improve learners' language skills. A part of the issues can be solved by using an error-annotated learner corpus for computer-aided error analysis, and one such corpus is being built for Baltic languages. It contains texts written by beginner learners of the second Baltic language (Latvian for Lithuanians and Lithuanian for Latvians), therefore, the variety of language represented in it could be called the Baltic interlanguage. The texts are annotated for errors using an error taxonomy created for Baltic languages. It is an attempt to address one of the main problems in error analysis – inconsistency in error taxonomies. The article shows how the corpus can be used in order to achieve a deeper understanding of factors that affect learning the second Baltic language for adults. Its aim is to describe main methodological steps by exploring the corpus as well as to acknowledge some of its limitations.

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

Introduction

It is a truism that learners of a new language unavoidably make errors. There is a widespread methodology for researching them, called *error analysis* (Ellis, Barkhuizen, 2005). It has been criticized for being too unclear and unreliable to be considered a scientific method (James, 2013), but various scholars still point out that errors are an important part of interlanguage (Granger, 2003a; Ramonienė, Brazauskienė, 2012) and that analysing them could give a great insight into language learning processes (Ellis, 1994; Vanhaegendoren, 2002).

Some of the problem areas for error analysis are unclear division or error categories as well as the data not being homogeneous enough to draw any conclusions, and those issues can be solved by creating and using learner corpora (Dagneaux, Denness, 1998). A learner corpus is "an electronic collection of authentic texts produced by foreign or second language learners" (Granger, 2003b, 538). Learner corpora are a relatively new invention, originating in the late 1980s (Granger, 2002). They have evolved from error collections which were comparatively small – the whole volume of one such collection rarely exceeded 2000 words and usually had no more than ten informants (Мальцева, 2011). A good part of learner corpora made nowadays consist of 50 000 – 150 000 words (Мальцева, 2011), and a learner corpus can be considered big if it has approximately 250 000 words or more, such as the learner corpus of English *NOSE* which contains over 300 000 words (Díaz-Negrillo, 2012).

The greater volume is not the only advantage learner corpora have over error collections. Being digital, they also allow researching various topics – such as lexis, grammar, and, of course, errors – using specialized software which can be useful not only for scholars, but also for teachers of the respective language (Камшилова, 2009).

Just like any other kind of corpora, learner corpora are designed so that texts are collected according to specific criteria (Granger, 2002), therefore, the homogeneity of the data is more easily achieved for the researcher. As for the error classification problems, a solution can be found by error-annotating the learner corpus, therefore making the error types more clear and consistent in order to fit one system (Granger, 2003a). A new publicly accessible error-tagged learner corpus has been made for Latvian and Lithuanian – it includes texts in Latvian that are written by learners of Lithuanian background, and texts in Lithuanian that are written by learners of Latvian background. This makes new kinds of analysis available for investigating learning of Baltic languages.

Usage of a learner corpus is the difference between traditional error analysis and computer-aided error analysis. It has been used by a number of researchers of other languages, especially English (Granger, Tyson, 1996; Gilquin, 2007). The aim of this paper is to explain how computer-aided error analysis can be done in order to investigate the errors made by learners of the second Baltic language. In Baltic studies, the term *second Baltic language* is used when talking about a person of Baltic (Latvian or

Lithuanian) background who is learning another Baltic language (Lithuanian for Latvians, Latvian – for Lithuanians; Butkus, 2008).

Material

The material studied used in this study is *Esam* – a learner corpus of the second Baltic language (more information and access to the corpus available online: www.esamkorpuss.lv). New material keeps being added to the corpus, and the annotation work is still ongoing, so the results of certain searches may differ at a later phase. Since the material added so far is not yet fully annotated during the time of writing the article, this paper's objective is not to reveal reliable results about the target language output of learners but rather to show how it can be retrieved and studied.

In order to understand what kind of conclusions can be drawn from the corpus's data, it is important to describe the texts included in it as well as the profile of the authors. As mentioned before, the corpus *Esam* includes texts written in Latvian by Latvian learners of Lithuanian background, and texts written in Lithuanian by Lithuanian learners of Latvian background. All of the texts have been a part of their author's university studies where they have been learning the second Baltic language as one of the study subjects. Currently, there are texts written in 2007–2014 by students of four universities: University of Latvia and Liepaja University in Latvia, and Vilnius University and Vytautas Magnus University in Lithuania. Most of the authors were philology students at the time when the texts were written. Each author signed a permission before their texts were included into the corpus.

The topic of each text was given by the teacher, but learners could also write on different topics if they had previously agreed with their respective teacher. Some more common topics are *Me, my family and my friends*, *My home*, *My holidays*, and *My studies*. The texts are sometimes graded, but they are written just as a part of study process – the teachers are not given any limitations on what the texts should be about, how long they should be. Such decisions are left to the teachers themselves and not influenced by the prospect of making a corpus. The text collection process did not influence the pedagogical process in any way, and the texts were given to the creator of the corpus only after the end of the semester when they were written.

The length of one text varies from 40 to 500 words – some teachers give requirements on how long the text should be. The title of each text is not counted because it was often given by the teacher, but titles are shown with each text in order to make it more easily understood as well as searchable for users of the corpus. The current size of the corpus is 157 texts and approximately 28 000 tokens, but it is going to increase by the time this article is published, as more texts are already being prepared for adding.

All authors are beginners – the texts included in the corpus were written during the first or second semester of learning the language in question. The first semester is usually intended to give the learner the skills of A1 level according to the Common European Framework of Reference for Languages, while the second semester is supposed to bring it up to A2 level. The actual knowledge level of the students may vary. The language of instruction matches the students' background language – in Latvia, Lithuanian is taught using Latvian as the language of instruction, while in Lithuania, Latvian is taught using Lithuanian as the language of instruction.

The structure of the texts is rather free. Although a teacher might sometimes have given some more specific requirements on what the text should contain, most of the time the only prompt is the title of the text. The students are also allowed to use any materials – textbooks and dictionaries – when writing the text, and there is no time limit, as this is usually given as a task to be done at home. If a teacher suspects a student of having written the text in another language and used any automatic translation tools or any professional or non-professional human translation services to have it translated, such a text is rejected and not included in the corpus.

Methodology

The current study aims to investigate some features of the Baltic interlanguage – the interlanguage that forms when a speaker of one Baltic language learns the other Baltic language. It especially emphasizes the first steps of computer-aided error analysis – namely, the data collection and annotation criteria – as well as the information retrieval process, to show how the corpus helps in error analysis. Error analysis and computer-aided error analysis as a method has already been described (Ellis, Barkhuizen, 2005;

Dagneaux, Denness, 1998; James, 2013), and error analysis has already been used to describe the Latvian (Žīgure, 1999, Laizāne, 2014, Zujevaitė, Žilinskaitė, 2012) and Lithuanian (Dabašinskienė, Čubajevaitė, 2009) interlanguage. Therefore, the error analysis methodology is not described as a whole, but the added value of the learner corpus and the things that are specific to the corpus *Esam* are discussed in more detail.

Results and discussion

The corpus *Esam* runs on TEITOK, a program created by M. Janssen especially for highly specialized annotated corpora (Janssen, 2016). The XML files of the corpus can be downloaded and used with any software that is compatible with the TEI standard. However, the corpus is intended to mainly be used via an online interface which does not require installation of any specialized software. The website's URL is *www.esamkorpuss.lv.* The user of the corpus does not have to register or provide any data in order to be able to access it. The corpus is annotated in four levels:

- syntactical level – sentences are marked and annotated for sentence types;
- morphological level – the words used in the texts are annotated for parts of speech;
- lexical level – the corpus is lemmatized;
- error annotation – the texts are corrected, and errors are annotated for error types.

The error classification used in this corpus is shown in Table 1. It shows the types of errors distinguished as well as the attributes given to those specific error types. The attributes are included into the annotation tags, and that makes it possible to retrieve errors marked as a specific type.

Table 6

Error classification in learner corpus *Esam*

Error type	**Attribute**	**Error subtype**	**Attribute**
Form	F	Agglutination	FK
		Upper / lower case	FL
		Diacritics	FD
		Other spelling errors (including typos)	FP
Morfoloģija un vārddarināšana	M	Derivation	MA
		Compounding	MS
		Inflection	ML
		Gender	MD
		Number	MN
		Definite / indefinite ending	MG
		Degree of comparison	MQ
		Person	MP
		Tense	MT
		Mood	MI
		Voice	MK
		Reflexivity	MR
		Participle confusion	MV
		Perfective	MB
Error type	**Attribute**	**Error subtype**	**Attribute**
		Iterativity	MX
Sintakse	S	Word order	SV
		Word missing	SI
		Word redundant	SL
		Cohesion	SS
Leksika	L	Meaning	LN
		Compatibility	LV
		Stable phrase	LS
Interpunkcija	I	Punctuation confusion	IN
		Punctuation redundant	IL
		Punctuation missing	IT

Descriptions and examples of the error types given in table one are not given here due to confinements of space; they will be discussed in another article. The corpus offers two kinds of search: simple and advanced. Although it is possible to search for annotations in the simple search form, it is easier to do using the options offered in advanced search view (Figure 1). TEITOK uses a search system that allows using various wildcards (Evert, 2009), but the advanced search, especially the *Word search* view, can offer a more user-friendly interface for some of the features.

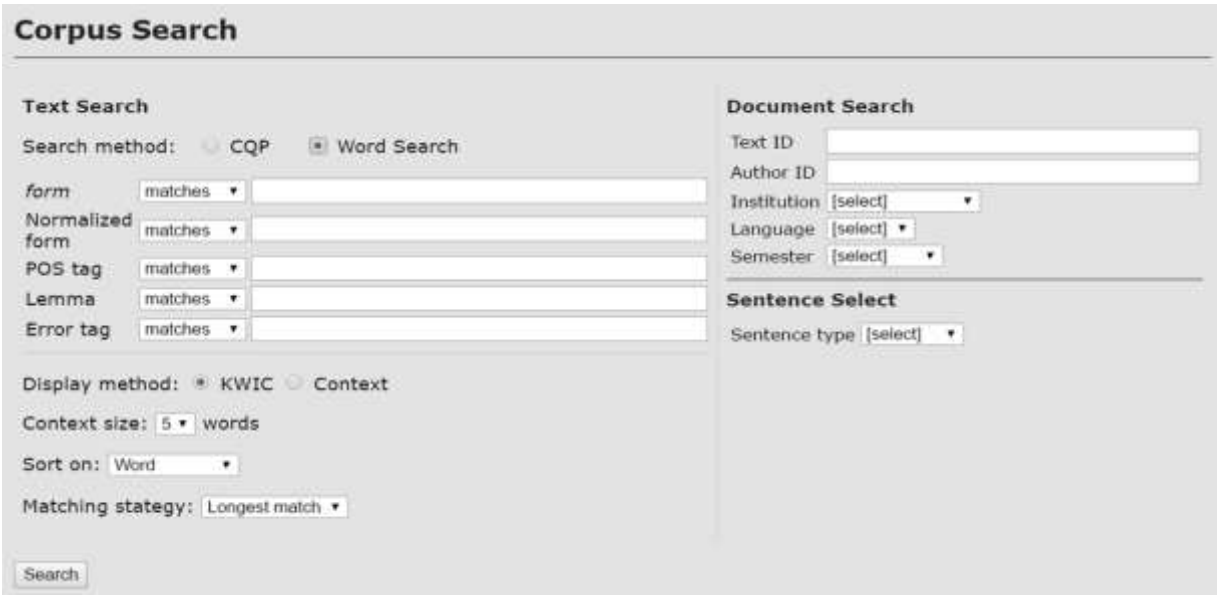

Figure 1. The view of advanced search.

The simplest way to look for certain errors is to write the specific code – attribute from the Table 1 into the *error tag* field. In addition, it is possible to choose other options, such as the language of the texts to be searched, the semester in which the searched texts should have been written, or the type of the sentences in which the search should be done. Also, there is an option to add more specific criteria to the token itself: one could search for, e. g., inflection errors in pronouns by writing the pronouns' attribute *p* into the *POS tag* field while inserting the *ML* attribute for inflection errors into the *error tag* field.

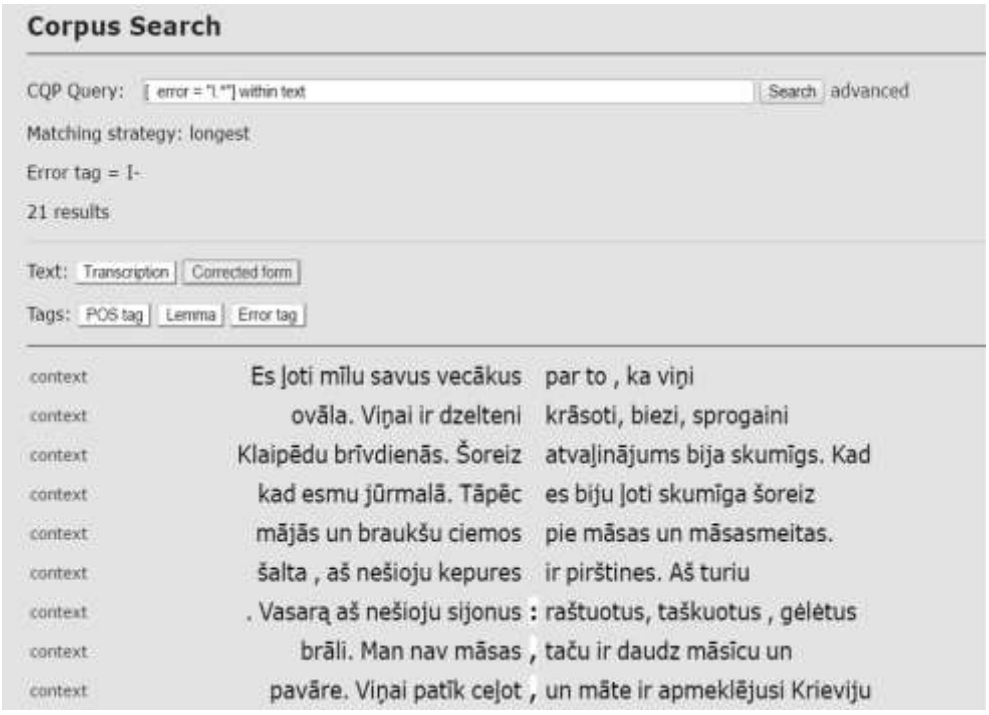

Figure 2. The view of search results.

If there is a need to look for a whole error type rather than subtype, that can be done as well. In order to do that, one has to change the search method of the *error tag* field from *matches* to *starts with*, and insert only the attribute of the whole error type – e. g. for punctuation errors, that would be I. Figure 2 shows some of the results for such a query.

As it can be seen, the program has created the query syntax itself, and the results are given centered around the place that is marked as erroneous. One can easily switch from the original (*transcription*) to the *corrected form* in order to see how the error was corrected. If desired, one can also view the error tags to see how exactly each of the errors has been labelled. This makes it possible to review and work with the material even if the user of the corpus does not fully agree with the corrections done by the annotators.

There is a *context* option on the left side of each result. That opens a whole text view which enables the user of the corpus to analyse the results more thoroughly if the current view does not give enough information. The corpus also offers some options of sorting the results to get the frequencies of certain groups of results (Figure 3).

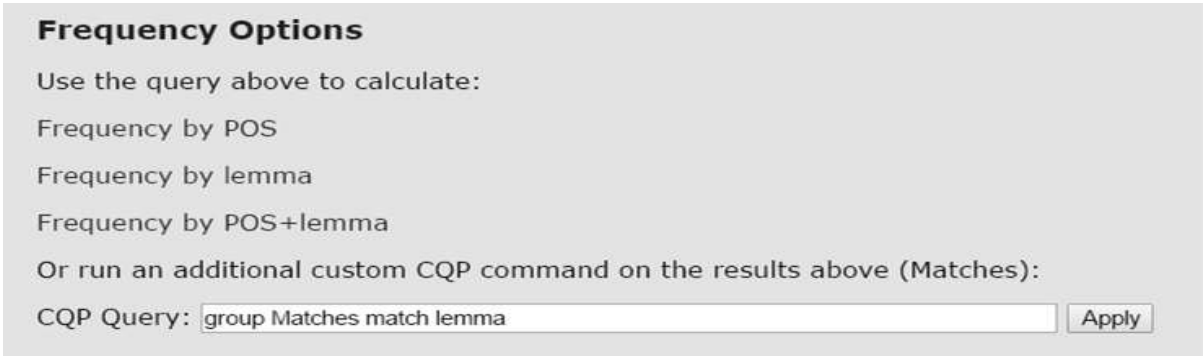

Figure 3. Search results: frequency options.

This kind of sorting does not show each separate sort result but gives some quantitative data on the types of results based on the specific query. Figure 4 shows an example where the initial search is for all errors (search query: *[error=".+"] within text*) and the following sort shows how many examples belong to each kind of errors (sort query: *group Matches match error*).

Frequency Information

Search query: [error=".+"] within text

Group query: **group Matches match error**

Error tag	Frequency	Percentage
FD	33	22.30
ML	24	16.22
FP	21	14.19
IT	14	9.46
LN	10	6.76
LS	6	4.05
IL	6	4.05
MD	6	4.05
LV	5	3.38

Figure 4. Search results: frequency information (top part only).

One thing that the corpus does not offer, is statistical calculations. It gives the number of results and percentage for each specific query, but it will not calculate statistical significance or any other statistical measures. If a researcher wishes to work more with quantitative methods, the calculations have to be done separately either with the help of some other software, or by the researcher him/herself.

Once the results are shown, the rest is left to the researcher – any other qualitative or quantitative descriptions are beyond the automatic tool's capabilities. The conclusions should be drawn carefully, as the corpus does not have enough data to describe the whole Baltic interlanguage. However, it includes a relatively large portion of it, especially at the beginner's level university students, since the corpus has been created in collaboration with a good part of the universities where the second Baltic language is being taught.

The most urgent following research would probably be an overview of the more common errors which would be helpful for the creators of new learning and teaching materials and textbooks. It would also provide information on the necessary updates in dictionaries, since some cases of words used in the wrong meaning seem to originate from not having enough explanation in dictionaries used by the students.

The corpus is also continually being updated with new material. In the future, it is possible to create a comparable corpus of texts written by students from other backgrounds which could help researchers understand which errors are more (or less) likely to be made by students of specific background and which seem to be more universal.

Conclusions

Scholars have been discussing errors that appear in the interlanguage of learners of Latvian and Lithuanian, but the research has been somewhat inconsistent. Challenges posed by error taxonomy and (non-)homogeneity of the data have undermined advances in this field so far.

The learner corpus *Esam* offers new insights into learning Baltic languages in two dimensions: first, it makes actual data available to anyone willing to research the topic even if those people are not involved in the teaching process themselves. Second, it gives more ground to error analysis for Latvian and Lithuanian which can not only help understand the learning processes, but also shed some more light on the similarities and differences between the structures of the languages themselves. Thus, error analysis turns into computer-aided error analysis.

The search and sort examples given in this article do not reflect the Baltic interlanguage because not much of the corpus has been annotated when this article is written. By the time it is published, the same queries will give much more extensive and reliable results.

Bibliography

1. Butkus A. (2008). *Baltiškos impresijos (Baltic Impressions).* Kaunas: Aesti. (in Lithuanian)
2. Dabašinskienė I., Čubajevaitė L. (2009). *Acquisition of Case in Lithuanian as L2: Error Analysis*. Eesti Rakenduslingvistika Ühingu aastaraamat 5. Tallinn: Eesti Keele Sihtasutus, pp. 47 – 66.
3. Dagneaux E., Denness S., Granger S. (1998). Computer-Aided Error Analysis. *System*, Vol. 26 (2), pp. 163 – 174.
4. Díaz-Negrillo A. (2012). Learner Corpora: the Case of the NOSE Corpus. *Journal of Systemics, Cybernetics & Informatics*, Vol. 10 (1), 2012, pp. 42 – 47. [online] [06.11.2016.]. Available at http://www.oalib.com/paper/2891896#. Vp0RP1JN-mJ
5. Ellis R., Barkhuizen G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.
6. Ellis R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
7. Evert S. (2009). *The CQP Query Language Tutorial.* [online] [25.09.2016.]. Available at http://cwb.sourceforge.net/temp/CQPTutorial.pdf
8. Gilquin G. (2007). To Err is not All: What Corpus and Elicitation can Reveal about the Use of Collocations by Learners. *Zeitschrift für Anglistik und Amerikanistik*, Vol. 55 (3), pp. 273 – 291.
9. Granger S. (2002). A Bird's-Eye View of Learner Corpus Research. In: S. Granger, J. Hung, S. Petch-Tyson (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, pp. 3 – 33.
10. Granger S. (2003a). Error-Tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, Vol. 20 (3), pp. 465 – 480.
11. Granger S. (2003b). International Corpus of Learner English: a New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, Vol. 37 (3), pp. 538 – 546.

12. Granger S., Tyson S. (1996). Connector Usage in the English Essay Writing of Native and Non-Native EFL Speakers of English. *World Englishes*, Vol. 15 (1), pp. 17 – 27.
13. James C. (2013). *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York: Routledge.
14. Janssen M. (2016). TEITOK: Text-Faithful Annotated Corpora. Proceedings of the Tenth International Conference on *Language Resources and Evaluation.* Paris: ELRA. [online] [04.09.2016.]. Available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf
15. Laizāne I. (2014). Datīva un datīva/instrumentāļa nozīmju apguve latviešu valodā kā svešvalodā (Meaning of Dative and Dative/Instrumental in Acquisition of Latvian as a Foreign Language). Zinātnisko rakstu krājumā *Valodu apguve: problēmas un perspektīva*, Vol. X. Liepāja: LiePA, 154. – 169. lpp. (in Latvian)
16. Ramonienė M., Brazauskienė J., Burneikaitė N., Daugmaudytė J., Kontutytė E., Pribušauskaitė J. (2012). *Lingvodidaktikos terminų žodynas* (Linguo-Diactics Term Dictionary). Vilnius: Vilniaus universiteto leidykla. (in Lithuanian)
17. Vanhaegendoren K. (2002). *Fremdsprachendidaktik in Theorie und Praxis*: *Deutsch als Fremdsprache (Teaching Foreign Languages in Theory and Practice: German as Foreign Language).* Lage: Hans Jacobs. (in German)
18. Zujevaitė A., Žilinskaitė E. (2012). Latvių kalbos kaip užsienio kalbos tekstynas (Corpus of Latvian as a Foreign Language). *Studentų moksliniai tyrimai.* Konferencijos pranešimų santraukos. Vilnius: Lietuvos mokslo taryba, 55.–57. psl. (in Lithuanian)
19. Žīgure V. (1999). Biežāk sastopamās kļūdas, apgūstot latviešu valodas elementārkursu (Most Common Errors while Learning the Basics of Latvian). Zinātniskie raksti *Sastatāmā un lietišķā valodniecība. Kontrastīvie pētījumi*, Vol. VIII. Rīga: Latvijas Universitāte, 107. – 113. lpp. (in Latvian)
20. Камшилова О.Н. (2009). Специальный корпус как составляющая лингвистического обеспечения языкового образования (Specialized Corpus as a Component of the Linguistic Support for Language Teaching). Материалы VIII Международной научно-практической конференции *Иностранные языки в дис-танционном обучении: мат-лы..* Т.2 – Пермь, Россия: ПГТУ. (in Russian)
21. Мальцева М.С. (2011). Учебный корпус как база для лингвистического и лингводидактического анализа в рамках методики преподавания иностранных языков (Learner Corpus as the Basis for Linguistic and Linguo-Didactic Analysis in Foreign Languages Teaching Methodology). *Социально-экономические явления и процессы* Т.9 (031), с. 209 – 212. (in Russian)