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

doi:10.3233/FAIA200631

Error Tagging in the Lithuanian Learner Corpus

Jūratė RUZAITĖ [1], Sigita DEREŠKEVIČIŪTĖ, Viktorija KAVALIAUSKAITĖ-VILKINIENĖ and Eglė KRIVICKAITĖ-LEIŠIENĖ
Vytautas Magnus University, Lithuania

Abstract. This paper is a work-in-progress report on error annotation in the Lithuanian Learner Corpus (LLC), which has been developed using the TEITOK environment. The LLC is the first electronic corpus of learner Lithuanian that represents learners of very diverse native language backgrounds and different proficiency levels. In this paper, we have a double aim: firstly, we present the structure of the corpus in its current state; and secondly, we describe the main principles, procedures, and challenges of error annotation in the LLC. The main types of errors that are tagged in this corpus and analysed in this paper are orthographic, lexical, and syntactic.

Keywords. Error annotation, leaner corpus, Lithuanian, TEITOK

1. Introduction

The present study is a work-in-progress report on error annotation in the Lithuanian Learner Corpus (LLC), which is currently still under development but is approaching its final stages. In this paper, we shortly overview the structure of the corpus in its current state and lay our primary focus on the main principles, procedures, and challenges of the process of error annotation mainly focusing on written texts.

Learner corpora have become a conventional empirical resource in studies of Second Language Acquisition (SLA) and language teaching/learning (e.g. [7]). The earliest and most numerous learner corpora have been compiled for English, e.g. the International Corpus of Learner English ([10]), TOEFL11 ([1]), Longman Learners' Corpus ([6]), or the Cambridge Learner Corpus ([15]). In recent years, however, learner corpora have been developed for a large variety of other languages, such as Arabic, Jinan Chinese, Korean, Persian, Czech, Dutch, Portuguese, Spanish, Italian, German, Estonian, Gaelic, Hungarian, Norwegian, Latvian and Lithuanian, Russian, and Slovene ([2]).

The landscape of learner corpora is currently quite diverse not only in terms of the target languages that such corpora represent but also regarding their overall size (ranging from around 50,000 to over 1 million words) and internal constitution. Concerning the latter, learner corpora can represent a different variety of text types (ranging from homogeneous corpora of, for example, solely academic writing to corpora comprising all types of written assignments, exams, and oral communication), L1 backgrounds

[1] Corresponding Author: Jūratė Ruzaitė, Centre of Intercultural Communication and Multilingualism, Vytautas Magnus University, K. Donelaičio g. 58, LT-44248 Kaunas, Lithuania; E-mail: jurate.ruzaite@vdu.lt.

(ranging from a single L1 to more than 60 languages), medium of communication (ranging from exclusively spoken or written texts to both spoken and written texts), educational institutions (covering a single institution or involving multiple institutions), or proficiency level (ranging from a single level to the full scope of A1-C2).

Lithuanian as a foreign language (henceforth LFL), being a lesser used and lesser taught language, in general has been studied to a rather limited extent (e.g. [3], [16], [17], [18]), and learner corpora were not available for a rather long time. This new corpus is the only digital text repository that represents a broad spectrum of LFL in terms of text types, native language backgrounds, and institutions where LFL is taught. It is also the only corpus of this size to be annotated for errors. The corpus ESAM (https://esamtekstynas.wordpress.com/) also represents learner Lithuanian, but it is limited to the beginner level and only Latvian as L1; it is also considerably more limited in size (52,000 tokens) ([22]).

It has become well established that error tagging is important in learner corpus annotation, since it allows for identifying standard and deviant forms, which in turn can help to pinpoint problematic areas in the language learning/teaching process ([9]). Error annotation has been done in a variety of languages, and error taxonomies have been developed for French ([8]), Czech ([11], [19]), Portuguese ([5]), Norwegian ([20]), Hungarian ([14]), Latvian ([4], [22]), and to some extent Lithuanian ([22]). The TEITOK interface, applied in this project, has been used for error annotation in the Croatian Learner Text Corpus (CroLTeC), the Baltic language corpus ESAM, and the Learner Corpus of Portuguese L2 (COPLE2; [5]).

2. Design and Main Features of the Lithuanian Learner Corpus

The LLC contains written and spoken data collected from LFL learners not only in Lithuania but also other countries, such as Germany, Sweden, Georgia, and China. It includes texts written by beginning (level A1; 102,952 tokens), pre-intermediate (level A2; 99,303 tokens), intermediate (level B1; 62,940 tokens), and upper-intermediate learners of Lithuanian (level B2; 37,639 tokens). In total, the corpus consists of 302,834 tokens. The disbalance between the lower and upper levels results from the fact that there are relatively few learners of Lithuanian who reach levels B1 and B2.

As the distribution of spoken and written texts presented in Table 1 shows, written texts form the majority of texts in each level (from 80 % to 62 %) and are more numerous in A1-A2 mainly because the oral output at this level is still rather restricted in length.

Table 1. Distribution of spoken and written texts

Mode	A1	A2	B1	B2
Written	75,561 (73 %)	79,842 (80 %)	39,514 (63 %)	23,165 (62 %)
Spoken	27,193 (27 %)	19,461 (20 %)	23,426 (37 %)	14,474 (38 %)
Total	**102,952**	**99,303**	**62,940**	**37,639**

Since written texts dominate in the LLC, we focus here on error tagging in this mode; besides, the scope of the paper does not allow discussing in greater detail the amendments that the spoken part requires.

The age span in the LLC ranges from 16 to 70 years of age, but most of the speakers are 18-26 (totalling 220,025 tokens, or 72.7 % of the entire corpus); see Figure 1.

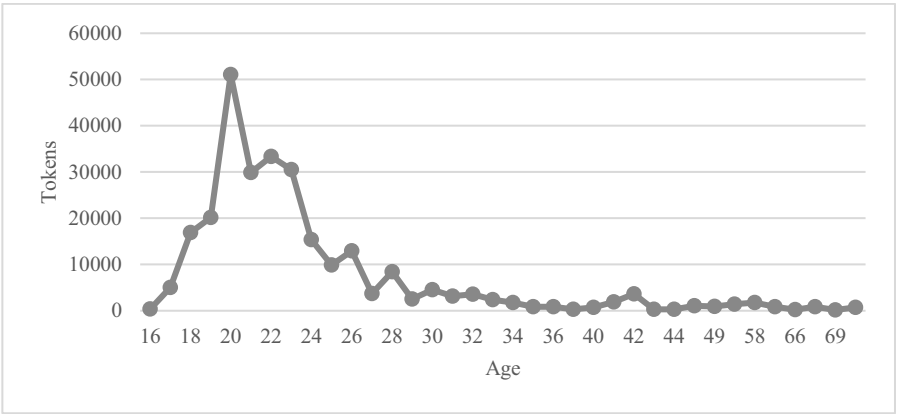

Figure 1. Corpus distribution by age

The dominant age span reflects the fact that the majority of learners in the LLC are undergraduate and graduate students. Approximately two thirds of the learners are female speakers (208,755 tokens as opposed to 94,079 tokens produced by male speakers). The learners come from over 50 different L1 backgrounds, and 55 learners indicated that they are bilingual or multilingual.

In terms of genres, the written subcorpus contains mainly descriptive essays (184,022 tokens), epistolary texts including letters, postcards, and emails (15,963 tokens), argumentative essays (6,409 tokens), and narrative texts (6,118 tokens). None of the other genres (literary essays, chats in a social network, or written dialogues) exceeds 1,000 words, and thus they form only a small minority of texts. In the spoken part, most of the recordings are semi-structured interviews of a teacher with a student, and only a small portion includes presentations (5,609 tokens).

The corpus uses the TEITOK programme developed by Maarten Janssen (2014-, http://www.teitok.org/), which is "a web-based framework for corpus creation, annotation, and distribution, that combines textual and linguistic annotation within a single TEI based XML document" ([13]). The TEITOK interface integrates linguistic annotation and search functions and offers the function of error tagging (for an overview of error tagging options, see [5]).

Thus, the transcriptions in the LLC are stored as TEI compliant XML files consisting of the transcription and a header with metadata. The latter includes the proficiency level, genre of the text, mode of communication, type of the task, use of reference tools, age, sex, the first language(s), foreign languages, mother's and father's first language, home language, education, educational institution, and the length of the text in words. The files are visualised in a user-friendly way in the TEITOK environment, as shown in Figure 2.

Figure 2. Visualisation of a written transcription

The transcribed text appears together with the scanned original, which has multiple advantages: it allows for verifying the accuracy of the transcription, presents the stimulus (the task of the assignment), displays the teacher's corrections (including not only verbal, but also non-verbal mark-up such as underlining, question marks, or explanatory comments), and provides possibility for multimodal research on learner data, which sometimes includes some drawings, schemes, graphs, and other visual elements.

3. Analysis of Error Categories in the LLC

Error tagging in the TEITOK environment is performed on tokenised data by following different types of taxonomies, which include taxonomies marking the source of error (orthography, lexis, and syntax) and taxonomies based on formal types of alternation of the source text (omission, addition, splitting, and merging) (cf. [11], [12], [21]). In addition to the forms suggested by the annotator, the software allows for marking the student form of each token versus the teacher form of the token (see also [5]). TEITOK also provides the possibility to normalise the learner's text by inserting omitted tokens, splitting, and merging them.

Drawing on the model followed in other TEITOK-based corpora, the annotation of deviant language forms in the LLC works at the token level and distinguishes three types of errors: syntactic, lexical, and orthographical errors. Having the same taxonomy for different corpora using the same environment allows for more systematic comparisons across different languages.

The taxonomy used in the LLC is thus based on rather broad categories and offers quite coarse granularity. However, even such a limited level of detail involves challenging tasks and even more so for a language such as Lithuanian, which has rich inflection, derivation, and agreement. Error tagging, which is inevitably guided by the annotator's intuition at least to some extent, does require an analytical framework that would be based on grounded choices made by the annotator to minimise arbitrariness. Thus, further on we overview which more specific categories, or subtypes, fall into the

three broad error types by discussing how different ambiguities in error identification and interpretation were solved and what choices were made by the research team.

3.1. Orthographic Errors

At the orthographical level, errors are limited to the word form. In the LLC, punctuation marks, differently from the Portuguese learner corpus ([5]), are not considered. They are tagged as a distinct error category in Znotina's work ([22]), which we consider to be more relevant than categorising it under orthography. However, in the initial stages of the corpus development we decided not to annotate punctuation, since it often does not receive a sufficiently systematic approach in the language teaching curriculum.

Orthographical errors in the LLC mainly include misspellings resulting in omitted/ substituted letters, misuse of capitalisation, missing or misused diacritics, misuse of long and short vowels, misspelt diphthongs, merging or splitting morphemes (agglutination), and spelling peculiarities arising due to sound assimilation (for examples, see Table 2).

Table 2. Subtypes of orthographic errors

Error subtype	Example in LFL	EN translation
Omission	*negali nuspęsti* (=nuspręsti)	'you can't decide'
Addition	*mokyklios* (=mokyklos)	'school' (sg.gen)
Substitution	*producтus* (=produktus); į sporto clubą (=klubą)	'products (pl.acc); 'to the sports club'
Diacritics	*nera* (=nėra)	'is not'
Capitalisation	*apie Amerikiečių* (=amerikiečių) *kultūrą*	'about American culture'
Long vs. short vowels	*mažas kambaris* (=kambarys)	'small room' (sg.nom)
Sound assimilation	*bendrabutio* (=bendrabučio)	'dormitory' (sg.gen)
Diphthongs	*studijouju* (=studijuoju); *vasių* (=vaisių) *pyragas*	'study' (3sg.pres); 'fruit cake'
Agglutination	*ne susitiksi* (=nesusitiksi); *vistiek* (=vis tiek)	'you will not meet'; 'anyway'

Some of these error subtypes also appear in Znotina's ([22]) taxonomy for Lithuanian and Latvian as second languages; she identifies diacritics, agglutination, upper / lower case (for capitalisation), and 'other spelling errors'.

Perhaps the most challenging are those instances when a deviant form is ambiguous and can be interpreted as an error in orthography or syntax, e.g. *Mano šalis turi jūra*. ('My country has a sea.'). Here the noun *jūra* should appear in the accusative form *jūrą* but '-a' is used without the diacritic and thus has the form of the nominative case. However, it is impossible to know if the learner misused the inflection of the nominative case (which would result in a syntactic error) or intended to use the inflection for accusative but did not add the diacritic to it (which would result in an orthographic error). We followed the principle that if the student form exists in native Lithuanian (e.g. *jūra*), but it does not fit in the grammatical context of the sentence, it is considered to be an error in syntax, not orthography, since a difference in the word form results in a different grammatical form. An orthographic error appears when a deviant form results in a non-existent form in standard native Lithuanian.

3.2. Syntactic Errors

The syntactic level covers grammatically deviant forms, that is, errors that affect syntactic structures. Most of the errors in this category include morphology errors

(illustrated in Table 3). Examples of such errors mainly comprise agreement problems (subject-verb, verb-object, modifier-noun, etc.), inaccuracies in the verb form (mood, voice, conjugation, reflexivity, etc.) and noun form (case, number, declension, gender, etc.), part of speech errors (e.g. adjective vs. adverb), errors in the use of prepositions, and agreement between prepositions and nouns. We also ascribe question words (as a category of function words) to the area of syntax.

Table 3. Subtypes of syntactic errors

Error subtype	Example in LFL	EN translation
Case ending	*Nebeturiu vietą* (=vietos)	'I don't have the place anymore'
Noun declension	*pirkti užsienietiškus prekius* (=užsienietiškas prekes) *savo šalyje*	'buy foreign goods in one's own country'
Number	*Ventspilis turi daug tako ir parko* (=takų ir parkų)	'Ventspils has a lot of paths and parks'
Countable/uncountable	*Aš valgau bandeles, tartus ir duonas* (=tortus ir duoną)	'I eat buns, cakes and bread'
Reflexive verb	Leiskite prisistatyti apie mano šalį (=pristatyti mano šalį)	'Let me introduce my country'
Person	*Kai aš buvo vaikė* (=buvau vaikas)	'When I was a child'
Agreement	*visokie skirtingos renginai* (=skirtingi renginiai)	'all sorts of different events'
Derivation	*Valdauja* (valdo); *radau toksį suoliuką* (radau tokį suoliuką)	'rules'; 'I found such a bench'
Verb conjugation	*Ji užaugė* (=užaugo) *kaime*	'She grew up in the countryside'
Voice	*kada autobusas bus atvažiuotas* (=atvažiuos) *pagal tvarkaraštį*	'when the bus comes according to the schedule'
Mood	*daugelis iš mūsų konservuoja agurkus ..., kad žiemą yra* (=būtų) *atsargos.*	'many of us can cucumbers … so that we stock up for the winter.'
Prepositions	Daug jaunų žmonių išvažiavo užsienyje (=į užsienį)	'Many young people went abroad'
Pronoun form	*Aš* (=Man) *patinka mano miestas*	'I like my city'
Adverb vs adjective	*Jie ieško darbo ir geriau* (=geresnio) *gyvenimo*	'They look for work and a better life.'
Question words	*Is kur tu studiojuje?* (=Kur tu studijuoji?)	'Where are you studying?'

In general, syntactic errors also include word order errors, but these were corrected in the LLC only when absolutely necessary. Lithuanian is a highly synthetic language and thus allows for a high degree of flexibility in word order, since usually more than one morpheme indicates the relations between different syntactic units. Alternatives in syntactic patterns in Lithuanian are difficult to assess since they can be used for different stylistic effects but strictly grammatically are still acceptable. Our approach seems to be more flexible than Znotina's ([22]) taxonomy; in her research, a stricter approach to word order is applied and some syntactic patterns presented as examples of inaccurate word order would not be counted as errors in our corpus.

3.3. Lexical Errors

Lexical errors (illustrated in Table 4) are restricted to word choice and meaning. At this level, the word used by the learner is orthographically and grammatically correct but is not the most natural choice for a native speaker in terms of word meaning and/or collocability. In some rarer cases, a lexical unit does not follow the word formation rules

(a derivational affix is misused) or a foreign word is used as a loan with a Lithuanian inflection.

Table 4. Subtypes of lexical errors

Error subtype	Example in LFL	EN translation
Prefixation	*Ilgai supgalvojau* (=galvojau) *apie tai; suspaustas* (=išspaustas) *sultis*	'I was thinking long about it'; 'squeezed juice'
Collocability	*tai yra dalykas, kuris keičiasi laikui skrendant* (=bėgant)	'this is something that changes as times passes'
Word choice	*ne vienas negali keltis nuo stalo kol vienas* (= kas nors) *dar valgo.*	'no one can leave the table while someone is still eating.'
Word formation	*šaltakariu* (=šaltojo karo) *pabaiga*	'the end of the cold war'
Loan	*Stvetaforas* (=šviesoforas) *yra prie teatro.*	'The traffic lights are near the theatre.'

As demonstrated in Table 4, we consider misuse of prefixation as a lexical error. It is an ambiguous subtype since some prefixes can also mark perfectivity (as in *galvojau* vs *sugalvojau*, where the latter refers to a completed action and is perfective) and as such can be assigned to the syntactic error category (cf. [22]). However, we take the stance that prefixation in many cases leads to semantic changes and lexicalization, and its impact on word meaning cannot be explained solely in grammatical terms (as in *suspaustas* vs *išspaustas*, where both forms are perfective, but there is an important semantic difference between the two).

Finally, it needs to be noted that a typical learner of Lithuanian makes errors across all linguistic levels, and a single token may result in more than one correction, e.g. a misspelt word may also be used with a non-standard inflection. Such multi-level errors are also marked using the TEITOK annotation tool.

4. Conclusion

This new error-tagged Lithuanian learner corpus with a rich XML-encoding opens new research areas as well as possibilities for practical applications in language teaching/learning. Error tagging can provide qualitative data about the types of errors in LFL and quantitative information about the distribution of these error types across different learner groups/texts. Such data can help develop an inventory of difficulties typical of the learner population in general and those that are restricted to a certain L1 background. By containing complete metadata, it allows for relating learners' errors to sociolinguistic parameters, e.g. the person's linguistic background, age, or gender.

The error taxonomy discussed here still needs refining as well as further testing by performing an inter-annotator agreement evaluation to assess the accuracy of the system. A more fine-grained annotation could be developed to account for more types/subtypes of errors. Further quantitative analysis of error types could lead to some insights about learners' difficulties; however, such analysis needs to be carried out with caution especially when comparisons between different languages are made since there are some differences in the internal structure of learner corpora and annotation systems even if a common tool for developing them is used. Despite the slippery areas that exist in such research, we believe that this new corpus will provide language instructors and researchers with valuable authentic data about learners' interlanguage so that better-grounded teaching and testing materials can be developed.

References

[1] Blanchard D, Tetreault J, Higgins D, Cahill A, Chodorow M. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service; 2013.

[2] Centre for English Corpus Linguistics: Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain; 2020. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[3] Dabašinskienė I, Čubajevaitė L. Acquisition of case in Lithuanian as L2: error analysis. Eesti Rakenduslingvistika Uhingu Aastaraamat. 2005; 5:47-66.

[4] Deksne D, Skadina I. Error-annotated corpus of Latvian. In: Utka A, Grigonytė G, Kapočiūtė-Dzikienė J, Vaičenonienė J, editors. Human Language Technologies – The Baltic Perspective; 6. Amsterdam: IOS Press; 2014. p. 163-166.

[5] Del Río I, 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. Umeå: LiU Electronic Press; c2016. p. 8-17.

[6] Gillard P, Gadsby A. Using a learners' corpus in compiling ELT dictionaries. In: Granger S, editor. Learner English on Computer; London: Longman; 1998. p. 159-171.

[7] Granger S. A bird's-eye view of learner corpus research. In: Granger S, Hung J, Stephanie Petch-Tyson S, editors. Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins; 2002. p. 3-33.

[8] Granger S. Error-tagged learner corpora and CALL: a promising synergy. CALICO Journal. 2003; 20(3): 465-480.

[9] Granger S. Computer learner corpus research: current status and future prospects. In: Connor U, Upton T, editors. Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam & Atlanta: Rodopi; 2004. p. 123-145.

[10] Granger S, Dupont M, Meunier F, Naets H, Paquot M, editors. International Corpus of Learner English. Version 3. UCL: Presses Universitaires de Louvain; 2020. 227 p.

[11] Hana J, Rosen A, Škodová S, Štindlová B. Error-tagged learner corpus of Czech. In: Proceedings of the Fourth Linguistic Annotation Workshop. Uppsala: Association for Computational Linguistics; c2010. p. 11-19.

[12] James C. Errors in language learning and use. Exploring error analysis. London: Longman; 1998. 320 p.

[13] Janssen M. TEITOK: Text-faithful annotated corpora. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož; c2016. p. 4037-4043.

[14] Langman J. Analyzing second-language learners' communication strategies: Chinese speakers of Hungarian. Acta Linguistica Hungarica 1997;44(1/2):277-299.

[15] Nicholls, D. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In: Archer D, Rayson P, Wilson A, McEnery T, editors. Proceedings of the Corpus Linguistics 2003 Conference; Lancaster: Lancaster University; c2003. p. 572-581.

[16] Ramonaitė J. Kaip lietuviškai šneka užsieniečiai? Lietuvių kaip antrosios kalbos veiksmažodžio įsisavinimas. Baltistica 2015; L (2):295-330.

[17] Ramonaitė J. Bendratis lietuvių kaip antrojoje kalboje. Baltistica 2017; LII(1):81-104.

[18] Ramonaitė J. Ką sako tokios užsieniečių sudaromos formos kaip *valgu* ar *žinėjau*? Lietuvių kalba 2017;11:1-25.

[19] Rosen A, Hana J, Štindlová B, Feldman A. Evaluating and automating the annotation of a learner corpus. Language Resources and Evaluation 2014; 48:65-92.

[20] Tenfjord K, Meurer P, Hofland K. The ASK corpus – A language learner corpus of Norwegian as a second language. In: Proceedings from 5th International Conference on Language Resources and Evaluation (LREC), Genova: ELRA; c2006. p. 1821-1824.

[21] Tono Y. Learner corpora: design, development and applications. In: Proceedings of the Corpus Linguistics 2003 Conference. Lancaster: Lancaster University; c2003. p. 800-809.

[22] Znotina I. Computer-aided error analysis for researching Baltic interlanguage. In: Dislere V, editor. Proceedings of the 11th International Scientific Conference. Education. Personality (REEP); Jelgava: Latvia University of Life Sciences and Technologies; c2017. p. 238-244.