ISSN 2335-2019 (Print), ISSN 2335-2027 (Online)
Darnioji daugiakalbystė | Sustainable Multilingualism | 7/2015
http://dx.doi.org/10.7220/2335-2027.7.8

Inga Znotiņa
Liepāja University, Ventspils University College, Rīga Stradiņš University, Latvia

LEARNER CORPUS ANNOTATION IN LATVIA AND LITHUANIA

Summary. Learner corpora are gaining popularity in the Baltic States as well as elsewhere in the world. The aim of the article is to discuss what kinds of annotation have been used in learner corpus research in Latvia and Lithuania so far and to describe which ones of them would be most suitable for the newly created learner corpus of the second Baltic language – *Esam*. A lot of learner corpus research in Latvia and Lithuania is undertaken without any annotation. The most common types of annotation are the ones based on the theory of levels of language – morphological and syntactic annotation. There is little collaboration between researchers of neighbour countries, but linguists of each country collaborate closely with each other using similar annotation schemes and creating corpora that are comparable in some aspects. The learner corpus of the second Baltic language should try to fit in the picture to some extent. Part of speech annotation and simple syntactic annotation could help in that. However, things that have not yet become so popular in learner corpus research in this region could also be useful. Therefore, error annotation and lemmatization have been chosen to be included in the annotation plan of the corpus *Esam* as well.

Keywords: learner corpora, learner language, annotation, Latvia, Lithuania, Baltic States.

Introduction

Learner corpus research is a quickly growing field, and a lot of new learner corpora are being built all over the world. Sometimes, researchers base the design of their corpora on the way other corpora have been built before, in order to make the data comparable to some extent. That is not always the case – some corpora are built entirely independently, but comparability is often seen as an asset, as it allows re-evaluating earlier made assumptions about learner language.

However, it is not always clear to which previously built corpora the new ones should be compared. In the cases when there is no a clearly dominating model, it is sometimes chosen to combine the most fitting and desirable features of various already existing corpora rather than choosing to follow one not fully. That seems to be a good choice especially if the nature of the data to be included in the new corpus is not really suitable for direct comparison with the previous corpora. This is the case with the learner corpus *Esam* – a publicly accessible learner corpus of the second Baltic language[45] – which is currently being built. The new corpus consists of texts written by beginner students of the second Baltic language – Latvian learners of Lithuanian and Lithuanian learners of Latvian – and it is likely to attract interest mainly from researchers in Latvia and Lithuania. Although the texts collected for *Esam* are not very similar to the data of other, already existing corpora, the annotation schemes used in it do not necessarily have to differ too much. Therefore, the aim of this article is to collect information and to describe how learner corpora have been annotated in Latvia and Lithuania so far. This background is then used to make the decision about annotation schemes that should be used in the annotation of *Esam*.

The material used in this overview is the publications of the field of learner corpus research in Latvia and Lithuania, as these countries form the area which is most relevant to the new corpus. Only publicly accessible information is used. It should be noted that quite a large part of learner corpus research in Latvia and Lithuania has been done without any annotation so far (see, for example, Bikelienė, 2009; Burneikaitė, 2009; Juknevičienė, 2009, just to name a few). In this paper, only annotated learner corpora are mentioned. The information about annotation of these corpora is extracted and compared in order to see possible tendencies. Based on these tendencies, a model is chosen for the corpus *Esam*. Four main parts make up the article, and each one of those is dedicated to a specific class of annotation types. The first of those parts discusses annotation that is based on the theory of levels of language (e. g., morphology, syntax, etc.). The second part talks about error annotation. The third part describes problem-oriented annotation, while the fourth one pays some attention to another kind of annotation that could still be considered for using in a learner corpus. In the end of each part, the more suitable annotation types for the *Esam* corpus are emphasized, and the conclusions section in the end of the article brings them together to describe the annotation model chosen for the new corpus.

The examples provided in this paper are not taken from the respective authors of the reviewed publications because not every author provides examples in their work. Since the existing corpora are made for different languages, the examples could even be too different to compare. It was decided that the same

[45] The corpus is currently being built, and a raw sample corpus of about 15’000 tokens has been made publicy accessible on June 14, 2015. For more information as well as access to the sample corpus, see: http://esamcorpus.wordpress.com

text schematically annotated in various ways could be the best illustration for annotation types, so the author of this article used a short sentence in Lithuanian and Latvian and wrote full-word descriptions instead of short tags.

Annotation based on levels of language

The theory of levels of language classifies language into various levels: phonetics, morphology, syntax, etc. Based on this, linguists work on phonetic analysis, morphological analysis, syntactic analysis, etc., which leads on to respective kinds of corpus annotation, morphological and syntactic annotation being among the most popular ones. They are used in various kinds of corpora, both in Latvia and Lithuania (see, for example, Levāne-Petrova, 2011; Rimkutė, Valskys, & Vaskelienė, 2009, etc.).

Morphological annotation is used in learner corpora built in Latvia, albeit not too widely. While some other corpora have been annotated for word forms or even morphemes, learner corpus annotation seems to be limited to part of speech (POS) annotation. This annotation type assigns each word information about its respective part of speech according to the grammar of that particular language. Schematically, it could look something like this:

<pronoun>Viņa <verb>ir <adjective>skaista. 'She is beautiful.'

<pronoun>Ji <verb>yra <adjective>graži. 'She is beautiful.'

Part of speech annotation schemes cannot vary a lot, since the division of words into parts of speech is relatively stable in each language. There are words which may bring confusion, but most of them can be clearly identified as nouns, verbs, adjectives, etc. There may be some purely technical differences in format, though – for example, if one researcher marks nouns with a tag <N> in their corpus, while another chooses a tag <Lietv> or <Daiktav>. Such differences can be quite easily fixed if needed – various tag or text finding and replacing tools can do it completely automatically (see, for example, the *Find and Replace* function in *Microsoft Word* software).

One of the corpora that makes use of this annotation type is the learner corpus of Latvian, created by a team of linguists under the wing of Latvian Association of Language Teachers (Kalnbērziņa et al., 2011). The tagset used in this annotation is very simple: the ten parts of speech that are traditionally divided in Latvian are identified, and each of them is assigned a tag that consists of one small-caps letter: <n> for nouns, <v> for verbs, etc. The word classes are not divided further into subgroups. The researchers do not elaborate on how exactly (automatically of by hand) the annotation was done. Since no tool is mentioned in the paper, it is assumed to have been done manually.

Various levels of Latvian language skills are represented in the corpus, and word classes were not annotated in the texts of the lowest level (F). This decision was based on the fact that there were too many unrecognizable word forms in the texts of this level. In higher level texts, the word class in ambiguous examples was stated based on the function the word has in the sentence, e.g., morphosyntactic characteristics.

Another part-of-speech annotated learner corpus in Latvia is the learner corpus of English, collected by Zigrīda Vinčela and tagged with the help of automatic tagger CLAWS (Vinčela, 2011; Vinčela, 2014). The tagging tool offers various tagsets as well as formats of output (see Leech et al., 1994), and most of them are much more complicated than the one used in the aforementioned learner corpus of Latvian. This tool works on English language only, so it is not suitable for the use on Baltic languages, and the categorization system used in it is too complicated to follow in manually tagging the learner corpus of the second Baltic language. Therefore, although this research is among the most extensive ones in this field in Latvia, its methodology should probably not be chosen for the corpus *Esam*.

No publications about morphological annotation of learner corpora in Lithuania were found.

Morphosyntactic annotation has been used by one author in the countries viewed. It is in Lithuania and the linguist doing it is Vitalija Kazlauskienė (see Kazlauskienė, 2015). However, her paper does not explain further what classification it was based on and how it was done technically.

Generally, morphosyntactic annotation deals with both morphological and syntactic features and is sometimes understood as simply a combination of morphological and syntactic annotation. However, it is more often seen as tagging words for their functions in the sentences (e.g., see EAGLES, 1996; Rögnvaldsson, 2006, etc.). Therefore, an example of this kind of annotation could look something like this:

<personal pronoun, subject>Viņa <verb, auxiliary>ir <adjective, predicate>skaista.

<personal pronoun, subject>Ji <verb, auxiliary>yra <adjective, predicate>graži.

On the one hand, morphosyntactic annotation, just like morphological annotation, should not have a lot of variation between various tagsets for the same language (if only formal differences), because the system of each language and the traditional categories of word functions in sentences define the categories to be tagged. On the other hand, it is much more complicated, especially in learner corpora where sentences can be malformed and unclear. It is done more easily in texts written by advanced students than in beginners' language.

Syntactic annotation deals with sentence types and structures:

<simple sentence>Viņa ir skaista.

<simple sentence>Ji yra graži.

Just like other kinds of annotation that are based on levels of language, the system of each language makes great variation possibility quite unlikely. More complicated annotation schemes are likely to have to be applied manually, at least in learner corpora where the algorithms of automatic analysis carried out by tagging software may be misled by learners' errors.

Syntactic annotation is used by two Latvian linguists Vineta Rūtenberga and Vita Kalnbērziņa. V. Rūtenberga uses it in her dissertation (Rūtenberga, 2014) as well as other connected works in order to research whether syntactic structures can be used in assessing French learners' written performance. The annotation scheme used here is very simple: three tags are used to differentiate between simple sentences, compound sentences, and complex sentences. The researcher herself considers it to be problem-oriented annotation, as it is closely connected with her research question and no other kind of annotation is applied in her corpus (ibid., 108). She also notes that annotation is done manually, because the learners' texts contain errors.

In the research undertaken by V. Rūtenberga and V. Kalnbērziņa in collaboration, texts in French as well as English are investigated, and the classification scheme is more complex. It is still based on the same three sentence types, but the classes are divided further in order to analyse the embedded constructions and clause types for complex sentences (Kalnbērziņa & Rūtenberga, 2012; Rūtenberga & Kalnbērziņa, 2013; Kalnbērziņa, 2015).

Another corpus which uses syntactic annotation is the aforementioned learner corpus of Latvian (Kalnbērziņa et al., 2011). Six kinds of sentences are differentiated in the tagset: in addition to simple, complex and compound sentences, also complex-compound sentences, reduced sentences and unclear sentences with an undefinable structure were tagged in this corpus. No further subcategorization was done.

The author of this article did not manage to find any learner corpora in Lithuania that have been syntactically annotated.

As for the learner corpus *Esam*, it could benefit greatly from both morphological and syntactic annotation. The most suitable kind of morphological annotation seems to be part-of-speech annotation. The relative simplicity of this kind of annotation makes it suitable for the learner corpus of the second Baltic language. Parts of speech are considered to be one of the main features, according to which words are classified, so it could give quite a significant insight into the material for various kinds of analyses. Linguists have already noted that ambiguous forms can be found quite often, though – especially if the language skill level of the learner is lower. Since both Latvian and Lithuanian are rich in grammatical forms, that is a challenge for annotating corpus *Esam*, too. The texts were written by beginners, so they contain their share of errors and therefore also unclear grammar. Firstly, it means that annotation has to be done manually. Latvian and Lithuanian languages have tools for doing this task automatically, but they work best on texts that contain little or no ambiguity (see more in: Rimkutė & Daudaravičius, 2007; Paikens, 2007). When working with learner language, ambiguity increases a lot, and often presents itself in unexpected ways (built up non-existing words; words of not the same word class in the target language may be used as such; etc.). So, even if the analysis was done automatically, reviewing it would still require a lot of work. Secondly, it also means that a system for annotating unclear examples should be made – whether parts of speech should be found by referring to the function the words have in sentences, or another solution could be found – it is still an open question. Thirdly, if even word classes are difficult to annotate, then a deeper morphological annotation (e.g., for morphemes) would be an even greater challenge, so that should probably not be among the first kinds of annotation chosen for *Esam*. The same could be said about morphosyntactic annotation – it is based on the part-of-speech analysis and is more complicated than that, so it is also quite a petty task in a beginner learner corpus.

Syntactic annotation is also desirable for beginner texts of the second Baltic language. It must, however, be decided how complicated the scheme should be because, just like in the case of morphological and morphosyntactic annotation, the tagging of beginner learner texts must be done manually. The syntactic annotation schemes used in Latvia differ in this aspect. Only tagging sentence types (simple, complex, etc.) would probably be the most convenient solution, but the tagset should be borrowed from the learner corpus of Latvian rather than the one used by V. Rūtenberga. A review of the texts included in *Esam* shows that various kinds of sentences are used, and the three-tag set does not fully reflect it while the six-tag set acknowledges also rarer but still present complex-compound sentences as well as reduced sentences.

Error annotation

While levels of language are widely used in annotating various kinds of corpora, error annotation is mostly associated with learner corpora. That being said, it is not the only kind of corpus that can be error-annotated – for example, a corpus of native Latvian has also been annotated for errors in order to help developing grammar checkers (Deksne & Skadiņa, 2014). As for learner corpora, the author of this article also did not manage to find any publications about any learner corpus in Latvia and Lithuania in which errors are annotated.

The way errors are annotated can differ quite a lot in various works, as it depends on the way errors are classified. The aforementioned publication by Latvian researchers lists the categories of errors (and, therefore, tags) which were identified for the needs of this corpus. 22 types of errors were classified into five groups: formatting errors, orthography errors, morphology and syntax errors, punctuation errors, and style errors (ibid., 164). This is a well-fitting classification for tagging the native language but is not so suitable for a learner corpus of beginner language. There are errors that are common in beginner learners' language but are rarely seen in the native language, such as a wrong word / a word that does not have the meaning intended in the text or even a similar one. Besides, the authors of this classification have created some quite specific error types, for example, incorrect noun case if a verb is used in debitive mood. Despite being a common error in native speakers' texts, it is hardly one of the most important ones in the texts written by people who just started to learn the language.

Consequently, the learner corpus of the second Baltic language should not blindly follow the lead and take over the error classification used in the corpus of native Latvian. However, it also ought not to use the error schemes used in the research of other languages without reviewing, as there may be categories that simply do not exist in the Baltic languages, such as articles (see, for example, the scheme offered in Granger, 2003). Rather, a classification that fits the Baltic languages and is able to describe a wide scope of errors produced by beginner learners should be made. It could, however, benefit greatly from the current work as well, because the Latvian tagset was made with a Baltic language being the main target, while other tagsets used abroad provide some insight in error-tagging a learner corpus in general. Since no learner corpora of Latvian or Lithuanian seem to have been error-tagged so far, a classification that fits both languages could perhaps help move towards error annotation in both countries in a comparable manner.

Problem-oriented annotation

Problem-oriented annotation is not a homogeneous class of annotation types. It is rather a way to group all those annotation types that were not created with the aim to make the data of the corpus maximally useful for most purposes – instead, problem-oriented annotation is aimed at tackling narrow, very specific research problems.

Traditionally, problem-oriented annotation is not added to learner corpora that are made for general use, without a very specific aim in mind. When it is not known if that specific annotation will be needed by anyone, it is most often decided not to waste the resources needed for tagging. If researchers need problem-oriented annotation in a general learner corpus, they usually annotate the data themselves.

Lithuanian linguists Jonė Grigaliūnienė and Rita Juknevičienė focused a study of English learner corpora on participial *-ing* clauses (Grigaliūnienė & Juknevičienė, 2012). Therefore, participles with *-ing* were the only thing the researchers annotated in the corpora they used. The tagging procedure was also very simple – automatically replacing all instances of *ing* with *ing PARTICIPLE*, followed by a manual review. This shows how pre-compiled corpora can be annotated for specific research needs, as the corpora used in this research were the Lithuanian components of ICLE[46] and LINDSEI[47] – two international corpora of learner English.

It could be assumed that one of the researchers who have annotated their data for specific problem-oriented elements is Z. Vinčela. In one of her papers, she describes researching linguistic variation in texts that are not only annotated for word classes but in which other features, specifically chosen for this analysis, were also identified (Vinčela, 2011, 2–3). All the features are not listed in the publication, but all of them were chosen to belong to one of three dimensions: *Dimension A – Involved Versus Informational Production*, *Dimension B – Explicit Versus Situation Dependent Reference*, and *Dimension C – Abstract Versus Non-abstract Information.* However, the researcher does not specify if the said features were annotated in the texts or searched without annotation. If annotation was done here, then it is a typical case of problem-oriented annotation.

Problem-oriented corpus annotation cannot be given a general example, because the variations are practically limitless. Any linguistic feature that can be identified can also be annotated in a corpus and therefore investigated by methods of corpus linguistics. If one were to research, say, constructions that consist of the verb *to be* and an adjective, the annotated text could look something like this:

Viņa <be+adjective> ir skaista </be+adjective>.

Ji <be+adjective> yra graži </be+adjective>.

As can be seen in the aforementioned publications as well as in the example given here, any rules and limitations in this kind of annotations are set by the researcher. It makes problem-oriented annotation extremely flexible, but that is also the reason why it cannot be unified. Having often to annotate manually, and the limited use of such an annotation make it necessary to consider whether the gains outweigh the resources spent annotating the data. It is close to impossible to know what kind of problem-oriented annotation could be needed by researchers in the future, so it is usually decided not to choose this kind of annotation before a specific research problem has been identified. Due to this reasoning, the learner corpus of the second Baltic language is also not expected to have any problem-oriented annotation, at least in the initial phase of research.

[46] For more information, see: https://www.uclouvain.be/en-cecl-icle.html.

[47] For more information, see: https://www.uclouvain.be/en-cecl-lindsei.html.

Other annotation types

Although not widely used in learner corpora in Latvia and Lithuania, there are more annotation types that could prove to be useful in a learner corpus. One of the most popular annotation types in corpus linguistics is lemmatization. Despite not being found in any of the publications on learner corpus research in Latvia and Lithuania, it is used in learner corpora elsewhere (e.g., Mönnink 1999; Haan 1998).

Lemmatization means assigning each word form its lemma – the base word for that form (McEnery & Hardie, 2012, 245). The example sentence used in this article could be lemmatized like this:

<viņa>Viņa <būt>ir <skaists>skaista.

<ji>Ji <būti>yra <gražus>graži.

Various forms of the same word can also often be retrieved by using wildcards, for example, if a wildcard * stands for any number of characters, then the forms *gražus*, *graži*, *gražaus,* etc. can be found by searching for *graž**. However, this search would not only return the unwanted word forms which do not belong to the desired lemma (such as *gražuolis*), but also can omit erroneous forms which are especially often found in beginner learner corpora. Unlike the standard forms used in language, erroneous forms cannot be easily predicted and therefore undermine the wildcard searching method. Beginner learner corpora, such as *Esam*, would therefore benefit greatly from lemmatization. Besides, if part-of-speech annotation is done, lemmatization requires no additional analysis because it is necessary to find a word's lemma in order to state which word class it belongs to.

The annotation types mentioned in this article do not, of course, make a complete list of all possibilities. Only the ones currently relevant to learner corpus research in Latvia and Lithuania have been listed here.

Conclusions

The field of learner corpus research is growing, but researchers in Latvia and Lithuania have not collaborated much so far. Since there are not many researchers working with this kind of corpora in the area, the scope of annotation types is also not too wide. This article might not have described every learner corpus created in Latvia or Lithuania, but the overview was made as full as possible. However, researchers sometimes concentrate more on the research questions they are working with than on the explanation of methods, so it is possible that some information about the annotated learner corpora in this area has not reached publications yet.

It should also be noted that the choices made for the corpus *Esam* should not be considered "best" but merely the most fitting in this case, based on the material gathered as well as the expected use of the corpus. Besides, one could also argue that another model would work better, if the importance of specific arguments is seen differently – the criteria are not always perfectly clear and objectively applicable to all cases in the same way. The main opposition being gain vs. resources needed, the evaluation of resources remains a somewhat subjective factor.

There has not been much collaboration between Latvian and Lithuanian researchers of the learner corpus field. However, collaboration between linguists in the same country is common: similar annotation schemes are sometimes chosen for several corpora, and the learner corpora are made comparable when possible. Annotation according to the levels of language seems to be most popular in Latvia and Lithuania at the moment, and problem-oriented tagging has also been done. A great part of work on learner language has been done without any annotation for now.

The large number of errors in the texts written by beginners make one do the annotation manually. It means that, in order not to slow the work down tremendously, the annotation schemes used in the *Esam* corpus should not be too complicated.

The decisions made for the learner corpus of the second Baltic language reflect two points of view: fitting in the research context and adopting new things. On the one hand, it is desirable to follow the path started by other researchers because it allows for some contextualization of the findings and might promote collaboration between researchers of neighbour countries. In the context of current research, the newly created learner corpus *Esam* could join the corpora that are morphologically and syntactically annotated, although with not too complex schemes: only part of speech morphological annotation and sentence type syntactic annotation. The annotation schemes in both of those kinds of annotation are not too rich in variation which helps to develop a wider view on specific problems over several learner corpora. However, a deeper subcategorization would require a lot of effort, as the corpus needs to be manually tagged and contains lots of errors. That is also the reason why morphosyntactic annotation is refused in this case. Lastly, problem-oriented annotation is also not going to be implemented in the learner corpus of the second Baltic language, unless specific needs of some research require it in future.

On the other hand, new things can be implemented and may give a greater insight in the material used. Things that seem new to the learner corpus research field in Latvia and Lithuania are not necessarily new for corpus linguistics and learner corpora in general. Rather, it is just not done here yet. One of the things considered here is error annotation with a different tagset from the currently used ones. Although error annotation has been done by Latvian corpus linguists previously, it has concentrated more on native speakers' errors. Errors in learner corpora have been annotated by linguists in other countries, but the classifications differ from language to language and no universal scheme seems to have been made so far. Since the existing error classification schemes do not seem too fitting for the needs of *Esam*, they should not be used in it. A suitable error classification for the features of the Baltic languages should be introduced which would also account for the great variety of errors made by beginner learners.

Another kind of annotation that could be of great use is lemmatization. Although not yet used in learner corpora in Latvia and Lithuania, it is quite easy to do together with part-of-speech annotation and can help one find various erroneous forms of a word which could not be so easily identified otherwise.

References

Bikelienė, L. (2009). Priešpriešos konektorių vartojimas besimokančių anglų kalbos ir anglakalbių studentų rašto darbuose. *Kalbotyra*, 61(3), 21–35.

Burneikaitė, N. (2009). Metadiscoursal connectors in linguistics MA theses in English L1 & L2. *Kalbotyra*, 61(3), 11–16.

De Haan, P. (1998). How 'native-like' are advanced learners of English? In A. Renouf (Ed.), *Explorations in Corpus Linguistics* (pp. 55–66). Amsterdam: Rodopi.

De Mönnink, I. (1999). Parsing a learner corpus? In C. Mair and M. Hundt (Eds.) *Corpus Linguistics and Linguistic Theory* (pp. 81–90). Berlin: Walter de Gruyter.

Deksne, D. & Skadiņa, I. (2014). Error-Annotated Corpus of Latvian. In A. Utka et al. (Eds.), *Human Language Technologies – The Baltic Perspective* (pp. 163–166). IOS Press.

EAGLES (1996). Recommendations for the morphosyntactic annotation of corpora. EAGLES Document EAG-TCWG-MAC/R, Version of Mar, 1996 [PDF document]. Retrieved from http://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3), 465–480.

Grigaliūnienė, J. & Juknevičienė, R. (2012). Corpus-based learner language research: contrasting speech and writing. *Darbai ir Dienos*, 58, 137–150.

Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra*, 61(3), 61–72.

Kalnbērziņa, V. (2015). *Sakārtojuma un pakārtojuma attiecības valodu apguves līmeņos*. Poster presentation in the 73rd conference of University of Latvia, section of Latvian and general linguistics, 06.02.2015. Rīga: Latvijas Universitāte.

Kalnbērziņa, V., Lokmane, I., Kunda, T., Vinčela, Z. & Baiža, K. (2011) Pētījums „Latviešu valodas apguves kvalitāte mazākumtautību skolās". Rīga: LVASA.

Kalnbērziņa, V. & Rūtenberga, V. (2012). Subordinate clauses as critical features in English and French learner examination corpora. *Baltic Journal of English Language, Literature and Culture*, 2, 54–62.

Kazlauskienė, V. (2015). Daiktavardinis žodžių junginys kaip gramatinės kompetencijos įsisavinimo elementas prancūzų kalbos baigiamojo egzamino rašto darbuose [Abstract]. The conference „Sustainable Multilingualism: Language, Culture, and Society", May 29–30, 2015. Kaunas: Vytauto Didžiojo universitetas. Retrieved from http://daugiakalbyste.vdu.lt/wp-content/uploads/docs/03/abstracts/Kazlauskiene.pdf.

Leech, G., Garside, R. & Bryant, M. (1994). *CLAWS4: The tagging of the British National Corpus*. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) (pp. 622–628). Retrieved from http://ucrel.lancs.ac.uk/papers/coling1994paper.pdf.

Levāne-Petrova, K. (2011). Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. *Vārds un tā pētīšanas aspekti*, 15(1), 187–193.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Paikens, P. (2007). Lexicon-Based Morphological Analysis of Latvian Language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies* (pp. 235–240). Kaunas.

Rimkutė, E. & Daudaravičius, V. (2007). Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos*, 11, 30–35.

Rimkutė, E., Valskys, V. & Vaskelienė, J. (2009). Lietuvių kalbos leksemų morfologinis anotavimas: ypatumai ir sunkumai. *Kalbų studijos*, 15, 63–70.

Rögnvaldsson, E. (2006). The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation. *Copenhagen studies in language*, 32, 133–145.

Rūtenberga, V. & Kalnbērziņa, V. (2013). Syntactic indicators of language acquisition levels in English and French written language learner corpora. *Lublin Studies in Modern Languages and Literature*, 37, 111–126.

Rūtenberga, V. (2014). *Syntactic Criterial Features in Assessing Written Performance in English and French* [Unpublished doctoral dissertation]. Rīga: University of Latvia.

Vinčela, Z. (2011). Linguistic Variation in EFL Students-Composed Virtual Texts in Different Registers. Corpus Linguistics Conference 2011 [PDF document]. Retrieved from http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-223.pdf.

Vinčela, Z. (2014). Tagging Errors in Non-Native English Language Student-Composed Texts of Different Registers. *Baltic Journal of English Language, Literature and Culture*, 4, 122–129.

Inga Znotiņa
Liepojos universitetas, Ventspilio universiteto koledžas, Rygos Stradinio universitetas; inga.s.znotina@gmail.com

BESIMOKANČIOJO TEKSTYNO ANOTAVIMAS LATVIJOJE IR LIETUVOJE

Santrauka. Besimokančiųjų tekstynai populiarėja tiek Baltijos šalyse, tiek ir visame pasaulyje. Šio straipsnio tikslas – išnagrinėti, kokios anotavimo rūšys, analizuojant besimokančiojo tekstyną, buvo iki šiol naudojamos Latvijoje ir Lietuvoje bei pateikti tas, kurios būtų tinkamiausios antrosios baltų kalbos naujai sukurtam besimokančiojo tekstynui *Esam* nagrinėti. Nemaža besimokančiųjų tekstyno tyrimų dalis atliekama be anotavimo. Dažniausiai naudojami anotavimo būdai grindžiami kalbos lygių teorija, t. y. morfologinis ir sintaksinis anotavimas. Kaimyninių šalių tyrėjai bendradarbiauja nedaug, bet kiekvienos šalies kalbininkai prisideda prie bendros veiklos, naudodami panašias anotavimo schemas ir kurdami tam tikrais aspektais palyginamus tekstynus. Antrosios baltiškos kalbos besimokančiojo tekstynas turėtų iš dalies derėti su bendra struktūra. Tam galėtų pasitarnauti kalbos dalių anotavimas ir paprastas sintaksinis anotavimas. Tačiau ir kiti aspektai, kurie dar nėra tokie populiarūs besimokančiojo tekstyno tyrimuose, šiame regione galėtų būti naudingi. Dėl šios priežasties klaidų anotavimas ir lematizavimas taip pat įtraukti į *Esam* tekstyno anotavimo planą.

Pagrindinės sąvokos: besimokančiųjų tekstynai, besimokančiojo kalba, anotavimas, Latvija, Lietuva, Baltijos šalys.