Jolanta Gelumbeckaitė, Mindaugas Šinkūnas, Vytautas Zinkevičius

Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation

1 AIMS

The *Old Lithuanian Reference Corpus* (Lith. *Senosios lietuvių kalbos korpusas*; acronym SLIEKKAS, cf. Lith. *sliekas* "earthworm") is a comprehensive, deeply annotated diachronic reference corpus of Old Lithuanian, being developed in cooperation between the Goethe-University of Frankfurt/Main (Germany), the Institute of Lithuanian Language (Vilnius, Lithuania), and the University of Pisa (Italy)[1]. The aim of the project is to create a multimodal (alignment of the annotated texts with facsimile reproductions of the original), annotated (header-information, hierarchic, structural, palaeographic, textological, lexical, and grammatical annotations) reference corpus (meta-linguistic information about Old Lithuanian, its diatopic variations, characteristic vocabulary). The ultimate goal is to develop a qualitative multilevel electronic retrieval engine for multilateral linguistic research of Old Lithuanian which will lead to reliable results for diachronic Lithuanian language studies. It will enable the implementation of the two biggest desiderata of Baltic linguistics: the Old Lithuanian grammar, and the historic dictionary of Lithuanian.

The most suitable technological and scientific basis for the multi-layer stand-off annotations is to be established on the basis of 10 selected texts (cf. Section 2): lemmatising (main word form and attested word form, the latter both in a transliterated form and as a normalised form in Modern Lithuanian), glossing (standard form of the lemma and of the attested word as well as their meanings), hierarchic grammatical description, predominantly restricted to morphology (part-of-speech tagging, flexional morphology of the lemmata and single attested word forms), and alignment of the annotated Lithuanian texts with each other and with their Polish, Latin, German etc. translation source texts.

The main endeavours of SLIEKKAS are the following: 1) securing a high philological standard as well as a textological and a palaeographic annotation of the selected Old Lithuanian texts, 2) setting up a basic-XML-structure, which is relevant for a further annotation, and 3) digitisation of the Lithuanian lexica and word indices, which are relevant for a further lemmatising and glossing of the texts.

The *Old Lithuanian Reference Corpus* is designed to provide an innovative scientific resource for historical and comparative linguistics as well as literary, religious, and cultural studies of the Baltic countries. This also includes materials related to the controversy between pre-Christian and Christian cultures and the confessional spin-off processes of the area as well as their backgrounds. In this way, the essential knowledge of the cultural development of Lithuania and the Baltic countries in the given period will be gained. With regard to historical linguistics, the *Old Lithuanian Reference Corpus* is expected to provide a basis for an efficient development and implementation of further research programmes concerning the diachronic grammar and the lexicon of Lithuanian.

This paper focuses on the main steps towards a semi-automated human-controlled grammatical annotation of Old Lithuanian, the available resources, and the most suitable software for this purpose.

2 MATERIAL

Old Lithuanian covers a period of ca. 300 years, from the 16th to the 19th centuries. The earliest known coherent Lithuanian text consists of three so called "Dzūkian prayers" in the copy of *Tractatus sacerdotalis* by NICOLAUS DE BLONY, preserved at the Vilnius University library (Straßburg: Martin Flach; Sign.: VUB RS II–3006). The year 1800, with the grammar by CHRISTIAN GOTTLIEB MIELCKE (1732–1807) *Anfangs-Gründe einer Littauischen Sprach-Lehre* (Königsberg: Hartung), marks the beginning of the standardisation and codification of Lithuanian based on a more or less single dialect, i.e. the southern group of the West High Lithuanian (=West Aukštaitian) dialect.

In total, the corpus will consist of over 10 million text words. Due to such a huge amount and to the complex, multilayered structures, which are needed for such a diachronic corpus, it seems reasonable to start with a smaller test corpus. Ten Old Lithuanian texts comprising ca. 350 000 tokens were chosen for this test corpus:

1. *DzP* ca. 1520—"Dzūkian prayers" (consisting of *Pater noster, Ave Maria, Credo*), the oldest known Lithuanian text; manuscript; translation from Latin, Polish, and/or German.
2. *MžK* 1547—MARTYNAS MAŽVYDAS, *Catechismusa prasty ßadei*; the oldest printed Lithuanian book; partly translated from Latin and Polish, and partly original written text.
3. *MžGA* 1549—MARTYNAS MAŽVYDAS, *Giesme S. Ambraßeijaus*; print; partly translated from Latin and Polish, and partly original written text.
4. *MžFK* 1559—MARTYNAS MAŽVYDAS, *Forma Chrikstima*; print; translation from German.
5. *WP* 1573— *Wolfenbüttel Postil*; manuscript; partly translated from Latin and partly original written text.
6. *VE* 1579—BALTRAMIEJUS VILENTAS, *Enchiridion*; print; partly translated from Latin and German, and partly original written text.
7. *DK* 1595—MIKALOJUS DAUKŠA, *Kathechismas*; print; translation from Polish and Latin.
8. *LyK* 1719—HEINRICH JOHANN LYSIUS, *Mažas Katgismas*; manuscript; translation from German.

9. *EnK* 1722—GABRIEL ENGEL, *Mažas Katgismas*; print; elaborated version of Lysius' catechism.

10. *DM* ca. 1765/1775—KRISTIJONAS DONELAITIS, *Metai*; manuscript, the first Lithuanian poem, autochthonic text. Editions of the text: first edition by LUDWIG J. RHESA (*DMRh*1818); second edition by AUGUST SCHLEICHER (*DMSch*1865); third edition by GEORG H. F. NESSELMANN (*DMN* 1869).

The selected texts represent a characteristic variety of Old Lithuanian text genres, sorts and types—a) religious as well as secular texts, the religious texts being those of the prayers, catechisms, hymnals, and sermons, all of them including Bible quotations; b) prose and poetry, c) translated, original written, and compiled texts, d) translations from Latin, German and Polish, and e) handwritten as well as printed texts. The chosen texts stand for the three language variations of Old Lithuanian, determined according to their dialectal, sociolectal, and confessional features—the Western or so called Prussian (*VE*, *LyK*, *EnK*, *DM*), the Middle (*DK*), and the Eastern type (*DzP*) of Old Lithuanian as well as a compound of several dialects (*MžK*, *MžGA*, *MžFK*, *WP*).

The selected texts also differ in their spelling as well as their accentography, which documents different strategies in indicating a free word stress through the grave, acute, or circumflex accent-mark and in marking two types of syllable accents (for falling resp. rising tonemes) on the one hand, and which also belongs to the system of the diacritical marks (similar to the Neo-Latin practice of accentuation) on the other. Some texts are accentuated (*DK*, *DM*; partly *LyK*, *EnK*), others not. Being heterogeneous as such, the texts determine a rich representativeness of the test corpus by simultaneously causing additional problems for computer processing.

3 ARCHITECTURE

The intended annotation scheme of the Corpus embraces the following structural features:

1. *A thorough linguistic and textological annotation, including header information, lemmatisation, grammatical information (part-of-speech tagging, morphological and basic syntactical information), glossing (in Standard Lithuanian, English, and possibly other languages), information about the text structure (text subdivision into words, sentences, lines, verses, paragraphs, etc.), palaeographic (resp. typographic) and textological information*—

The main purpose is to develop a semi-automated technique that allows establishing the core word form in a historical lexicon (lemmatisation), its glossing in Standard Lithuanian and the determination of its actual meanings in a given Old Lithuanian text. More than 50% of the Old Lithuanian word forms are ambiguous as regards their morphological status. A morphological annotation consists of the unalterable morphological categories of the lemmata as well as of the actual word forms in a given text, and of the flexional morphological characteristics of the latter. For instance, the morphological categories of the lemma and of the attested word form in a

given text are to be annotated differently in such cases, as the masculine adjective *gražiausias* "the most beautiful", which belongs to the *ja*-paradigm (superlative form), while its lemma *gražus* (Masc), *graži* (Fem) belongs to the *u,jo*-paradigm—thus two separate levels for the morphological categories have to be created, one for the token, the other for its lemma. A distinction of the morphological categories of the lemma and of the actual word form helps to trace the alteration of grammatical classes in Old Lithuanian, e.g., substantivisation of adjectives, adjectivisation of participles, adverbalisation, etc. For example, the form *laukan* "to the outside, into the field" is a paradigmatic illative case of the substantive *laukas* "field" in Old Lithuanian, whereas the form *laukan* is considered merely as the adverb "out" in Standard Lithuanian.

These annotation levels (lemmatisation, glossing, part-of-speech tagging, and morphological annotation) are carried out on the basis of the *Toolbox* program (SIL: http://www.sil.org/computing/toolbox/). Afterwards, they will be revised and corrected in the annotation software *ELAN* (Max Planck Institutefor Psycholinguistics in Nijmegen, http://tla.mpi.nl/tools/tla-tools/elan/). Furthermore, the texts will be provided with the basic information on the syntactic structure of the sentences (simple and complex sentences will be marked) in *ELAN* directly.

Single Latin, German, or Polish words and sentences within the Lithuanian texts will be annotated according to the morphology of a corresponding language. Additional annotation levels are required for the taxonomy of both explicit and implicit quotations in the Old Lithuanian texts. It enables a clear distinction between the translated resp. re-narrated text parts and the original written text.

2. *A multi-level architecture of the annotations*—

The aim is to generate an XML-structure that comprises all the intended annotation levels. The experience of the DFG project *Referenzkorpus Altdeutsch*[2] has shown that the software *ELAN* fully serves this purpose. The *ELAN* data structures can either be produced directly from the text data on the basis of the *Toolbox* program, or they can be generated from autonomously programmed components which incorporate the text data with their lexical, grammatical, and other information.

3. *Multi-modality of the corpus through the alignment of the texts with facsimile reproductions of the original*—

The Old Lithuanian texts will be aligned automatically with facsimile reproductions of the originals (manuscripts resp. prints) on a line level and additionally aligned manually on a word level.

Since most of the Old Lithuanian texts are translations from Latin, German, or Polish sources, the source texts (ca. 190 000 text words in the case of the test corpus) will be annotated in the same way as the Lithuanian ones. This will enable the alignment of the Old Lithuanian texts with their sources with respect to all annotation levels. Furthermore, the Old Lithuanian texts of the same genre will be aligned with each other in order to allow for an assessment of possible mutual influences within a single genre as well as across genres.

4 RESOURCES FOR ANNOTATION

Old Lithuanian can be roughly classified into three main periods of the evolution of orthography. The early period is the most variable and unstable one. Orthography gets more uniform in the middle of the 17th century in Lithuania Minor (Duchy of Prussia), but it has a different variant in Lithuania Major (Grand Duchy of Lithuania). The specific orthography of the texts has to be converted (during which the regular dialectal phonetic features are discarded) to match the one that exists in Modern Lithuanian, in order to be processed by an automatic morphology analyser. Results of these processes will be included in the annotation levels of the "Standardised word form (transliteration)" and "Normalised word form (in Modern Lithuanian)", according to which retrieval tasks can be modified. The conversion of the old orthography to the modern one is done by the transliteration rules that are implemented using the *Consistent Changes Program* (SIL: http://www.sil.org/computing/catalog/show_software.asp?id=4). For the orthography of the early period, special rules have to be created for every individual author (sometimes even every text of the same author). The (ortho)graphy of the texts from the 16th century differs from the one used in the 18th century (cf. Ill. 1 and 2). The transliteration rules are more stable for the later period, though they are also slightly modified for each author (or text) to attain the maximum possible accuracy.

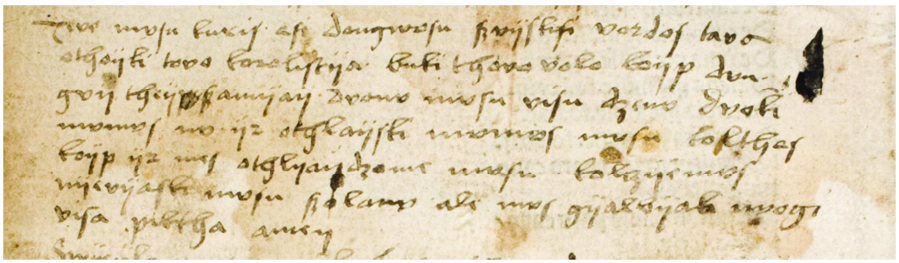


Illustration 1: A fragment of *DzP*, ca. 1520 (*Pater noster*)

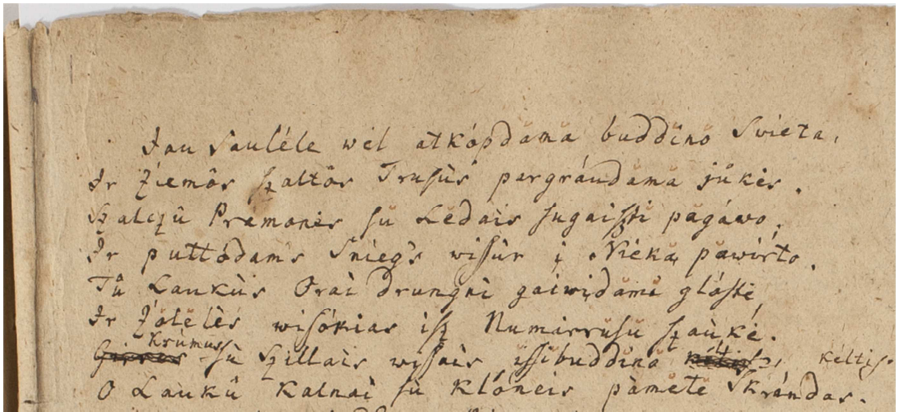
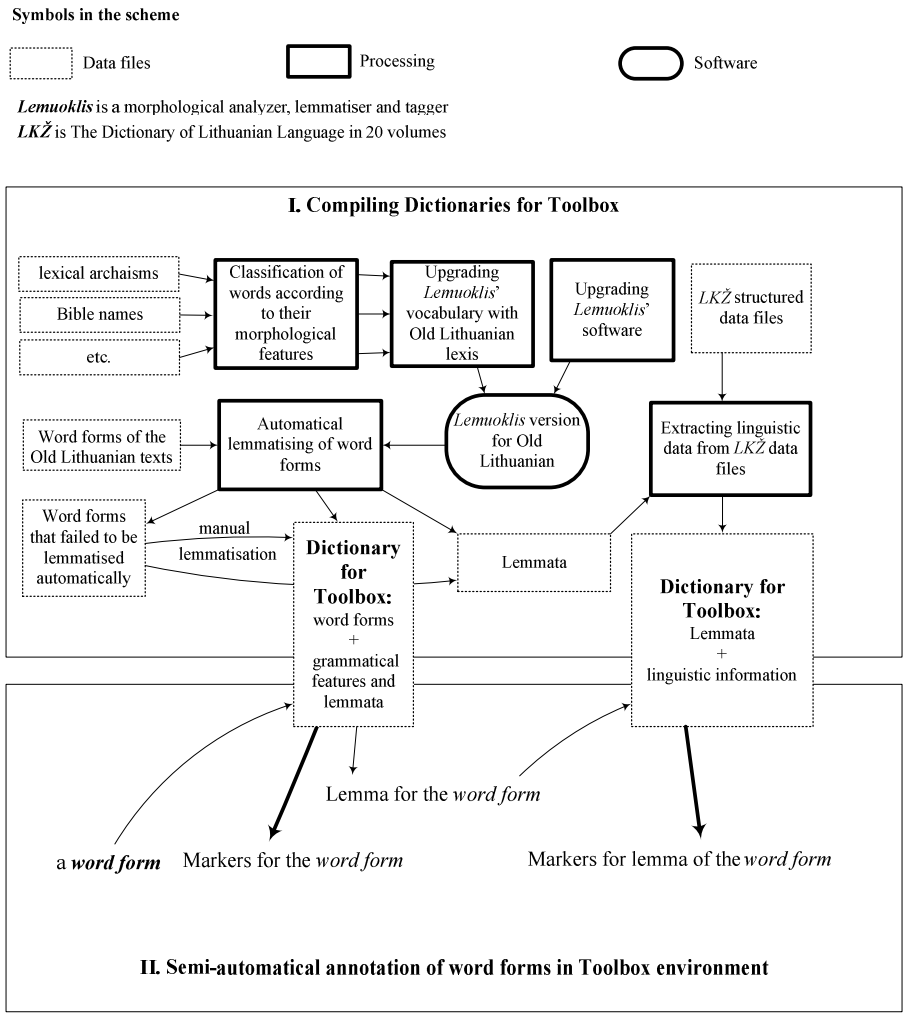


Illustration 2:A fragment of *DM*, ca. 1765/1775

To give an example of the transliteration, in the words *Ape Swetaſti* "about the Sacrament" (chapter name, *MžK* 25) the rules of changing the long <ſ> to the round <s> and <w> to <v> in the form *Swetaſti* are applied, and a form *svetasti* is created. The original word form layer with *Swetaſti* remains unchanged. By implementing additional rules the created form *svetasti* gives possible correspondences, namely *svetastį*, *svėtasti*, and *svėtastį*. Only the latter form will be recognised as a valid entry with standardised orthography in the Old Lithuanian database and can then be analysed further.

Figure 1 shows resources and processes used for developing a semi-automated technique for the grammatical annotation of the Old Lithuanian words. In the annotation process, the *Toolbox* environment utilises two dictionaries, as is shown in the lower part of the scheme. Firstly, a search for a word form in the dictionary of the word forms is performed: lemma, part-of-speech, and other grammatical markers for the word form are extracted. Afterwards, markers for the lemma are searched and extracted from the dictionary of the lemmata. In case the search results are ambiguous, i.e., when more than one record is found in a dictionary, the annotator working with the *Toolbox* program must make a decision and choose the correct variant. In order to enable these processes, two dictionaries—one of the Old Lithuanian word forms and another of the Old Lithuanian lemmata—are currently being compiled, as is shown in the upper part of Figure 1.

Figure 1. The Automated Grammatical Annotation in SLIEKKAS



While producing the Old Lithuanian dictionary, the word forms are lemmatised and POS-tagged using the software *Lemuoklis*, a morphological analyser, lemmatiser, and tagger for Modern Lithuanian (ZINKEVIČIUS 2000, ZINKEVIČIUS, DROŻDŻYŃSKI, HOMOLA, PISKORSKI 2003). *Lemuoklis* is a rule-based system. The lexical and grammatical data of the program

consist of several lexica (organised as letter trees). Three of them store the roots of Lithuanian words, which are associated with certain appropriate morphological rules; morphological rules are presented in the form of digital tables. Both the vocabulary of stems (organized as a tree data structure) and tables of rules are in the original internal digital format. *Lemuoklis* is a library of functions programmed using the C++ language. Other lexica store word forms with no morphological informationor contain lists of abbreviations and acronyms (http://donelaitis.vdu.lt/~vytas/tool/tool.ppt). The software is able both to analyse a word form grammatically and to synthesise a new inflectional form. It performs lemmatisation by means of synthesising new forms (e.g., nominatives for nouns and infinitives for verbs and verbal forms).

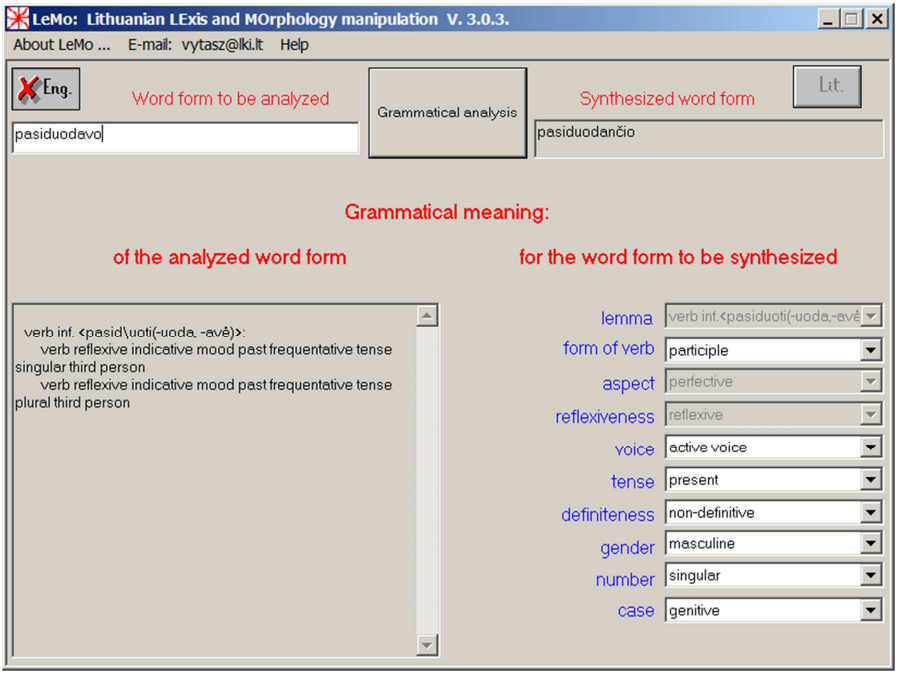


Figure 2: Demo window of *Lemuoklis* – a morphological analyser, lemmatiser, and tagger for Modern Lithuanian

Figure 2 shows a demo and testing window of *Lemuoklis.* The word form *pasiduodavo* (past frequentative, "was used to surrender") is analysed by *Lemuoklis* (in the left part of the window) as having the lemma *pasiduoti* (the infinitive, "to surrender") and is characterised with the tags "verb reflexive", "indicative mood", "past frequentative tense", "plural", and "third person". The infinitive form is supplied with the endings of the two other main forms of a verb, i.e., present and past forms, *pasiduoda* and *pasidavė* respectively. In the right part of the window the annotator can select morphological properties for a new inflectional form of the word *pasiduoti* ("to surrender") which is synthesised automatically in the upper box "Synthesised word form". In case a surface form is homonymous, i.e., it has several grammatical meanings, the program

gives full grammatical characteristics for each possible homograph of the surface form. However, some methods are used to reduce the ambiguity without taking into account the context: one of them is the method of disambiguation between diminutive nouns with the suffix *-yti(s)* and respective verbal infinitive forms. For example, the Lithuanian word form *padaryti* is interpreted as a transitive infinitive form ("to do something") rather than a theoretically possible voc. sg. form of a diminutive *padarytis* from *padaras* ("a creature"); the word form *ginčytis* is interpreted as a reflexive infinitive ("to argue") rather than the nom. sg. of a diminutive *ginčytis* from *ginčas* ("a dispute, argument"). The disambiguation between proper and common nouns is performed through the application of special lexica containing proper noun forms from Modern Lithuanian corpora and other sources (ZINKEVIČIUS, DROŽDŽYŃSKI, HOMOLA, PISKORSKI 2003).

The original *Lemuoklis* is based on the Modern Standard Lithuanian grammar and various modern lexica. In order to enable *Lemuoklis* to recognise words from the Old Lithuanian texts it was enriched through a special vocabulary which comprises the dictionaries of Old Lithuanian (PALIONIS 2004; ca. 8000 words) and of Slavic loanwords in Old Lithuanian (SKARDŽIUS 1998; 4152 words) as well as the dictionary of Bible names (KIMBRYS 2000; 3251 words), and some other lexical material. All added words had to be classified semi-manually (while choosing correct answers to the questions, cf. Fig. 3) according to their morphological features while using a special software, which creates supplemental lexica of the roots associated with morphological rules for *Lemuoklis*.

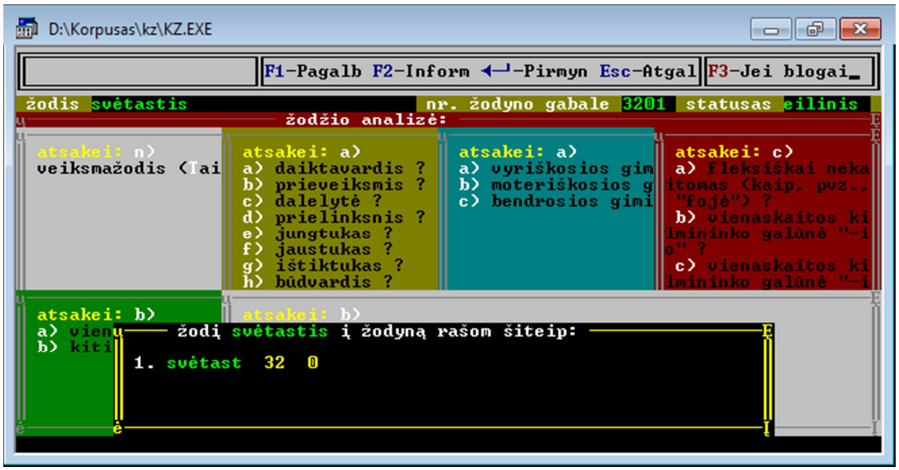


Figure 3: Process of semi-manual classification of words from the Old Lithuanian texts

Figure 3 shows the process of classifying the Old Lithuanian word *svėtastis*. During the first step, the system formulated the question "is it a verb?" and an operator answered by pressing "n" (no). In the second step, the system enumerated names of the parts of speech, and an operator's choice was "a noun" (*daiktavardis* in Lithuanian) by pressing "a"; during the next step, gender was defined (*vyriškoji* masculine, *moteriškoji* feminine, *bendroji* common). Then, an operator was asked whether *svėtastis* is a non-inflective (variant *a*) form, or having the genitive ending *-io*

(variant *b*) resp. *-ies* (variant *c*), and an operator chose the variant "c". Next, the possibility to build a plural form was confirmed, and a type of declension was specified more precisely by choosing the right ending variant for the plural genitive. The two lower windows in the screen were covered by the final black one which indicates the result of the word classification process: "svėtast 32 0", where *svėtast-* is the stem, *32* is an internal number for the inflection type, and *0* indicates the number of letters at stem's end that differs through the inflectional paradigm.

To get back to the above-mentioned example of the transliteration of *Swetaſti*, *Lemuoklis* is provided with the accusative forms *svetasti*, *svetastį*, *svėtasti*, and *svėtastį*. The word *svėtastis* (Nom) does not exist in Modern Lithuanian, but the root was added together with other loanwords from SKARDŽIUS' Old Lithuanian dictionary (SKARDŽIUS 1998), and thus can be processed by the modified version of *Lemuoklis*. In the process of an automatic lemmatisation and of an analysis of the word forms *svetasti, svetastį*, *svėtasti*, and *svėtastį,* only the latter form *svėtastį* is recognised by *Lemuoklis*, its flexional morphological characteristics (Sg_Acc), flexional class (i_Fem), part-of-speech (noun), and lemma (*svėtastis*) are generated. This information is stored in the SLIEKKAS dictionary of the word forms.

While producing the lemmata dictionary, the required grammatical information is obtained from the Lithuanian language dictionary (*LKŽ*; *Lietuvių kalbos žodynas*, 20 volumes, printed in 1941–2002; online: www.lkz.lt/startas.htm; ZINKEVIČIUS 2008). This thesaurus includes ca. ½ million lemmata. The words which are essential regarding the needs of the testcorpus have been extracted according to the token list of the test corpus. For instance, the lemma *svėtastis* was searched in *LKŽ* and marked with the following information: accented lemma (*svẽtastis*), part-of-speech (noun), and accentuation class (1); the flexional class is additionally created by the retrieval engine (i9_Fem). This information is stored in the SLIEKKAS lemmata dictionary.

5 DISAMBIGUATION

Three types of data are created by means of the software and lexical resources mentioned above: 1) a list of transliterated word forms, 2) a dictionary of normalised word forms, which includes information on the part-of-speech, unalterable morphological categories, and flexional morphological characteristics of the actual word form, and 3) a dictionary of the lemmata, which includes the tags for part-of-speech and unalterable morphological categories of the lemma, and also its accentual class. Separate dictionaries for the translation into other languages (English and possibly German) can be added while linking them to the Lithuanian lemmata. The abovementioned three types of data (for transliterated and normalised word forms as well as for lemmata) shall be managed by the *Toolbox* program, in which the annotation levels (lemmatisation, glossing, part-of-speech, and morphological annotation) are created and disambiguation is controlled by a human (Fig. 4).

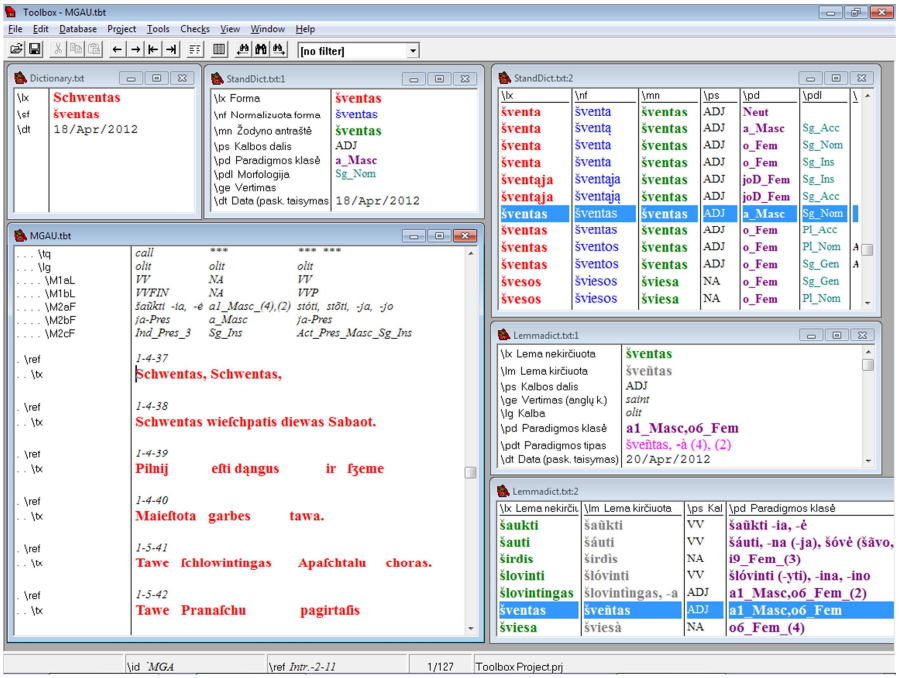


Figure 4: Annotations are created in the *Toolbox* program using the generated lexical and grammatical information

Figure 4 illustrates the *Toolbox* environment, where the word *Schwentas* ("saint") is being processed in the window MGAU.tbt using a list of transliterated forms (window Dictionary.txt), a dictionary of the normalised word forms (StandDict.txt) and a dictionary of the lemmata (Lemmadict.txt). The rich flexion of Lithuanian and the inconsistency of the old orthography result in a very high rate of homographs. The automatic disambiguation is complicated because the analysis is done on the word level only without involving the context or considering punctuation (no tools or rules on the Old Lithuanian syntax are implemented), in the absence of semantic information, without regard to accent marks, and with lack of statistical data. The overall disambiguation has to be controlled manually, as can be seen in Figure 5 (the word form *šventas* can be either Masc_Sg_Nom or Fem_Pl_Acc). After the ambiguous grammatical information is dissolved, the annotation layers are created (Fig. 6).

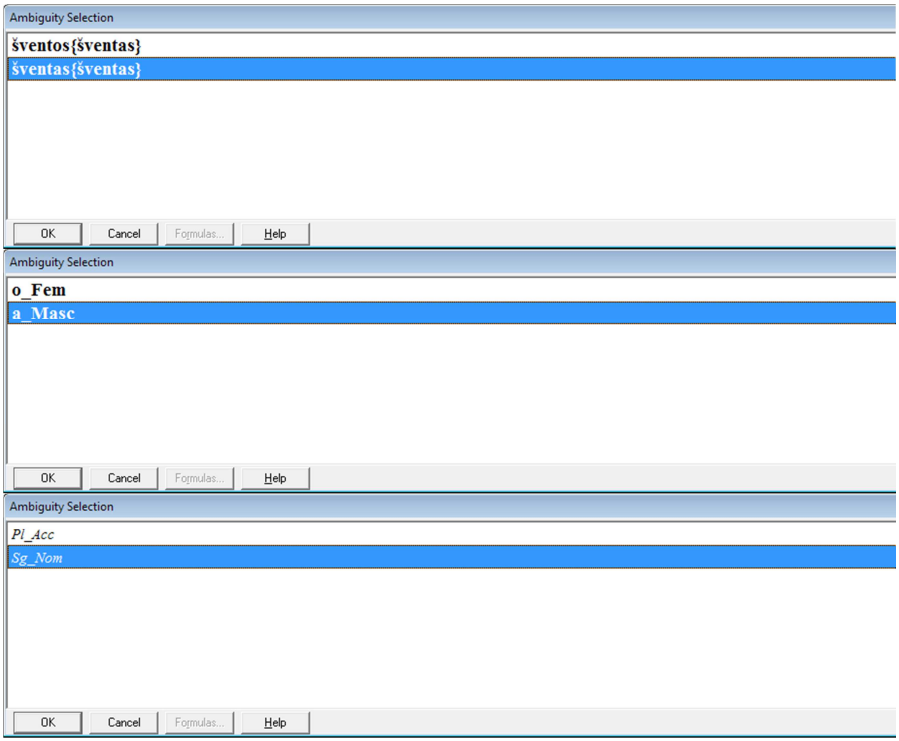


Figure 5: Steps of ambiguity selection for the annotation of the word form *Schwentas* "saint"

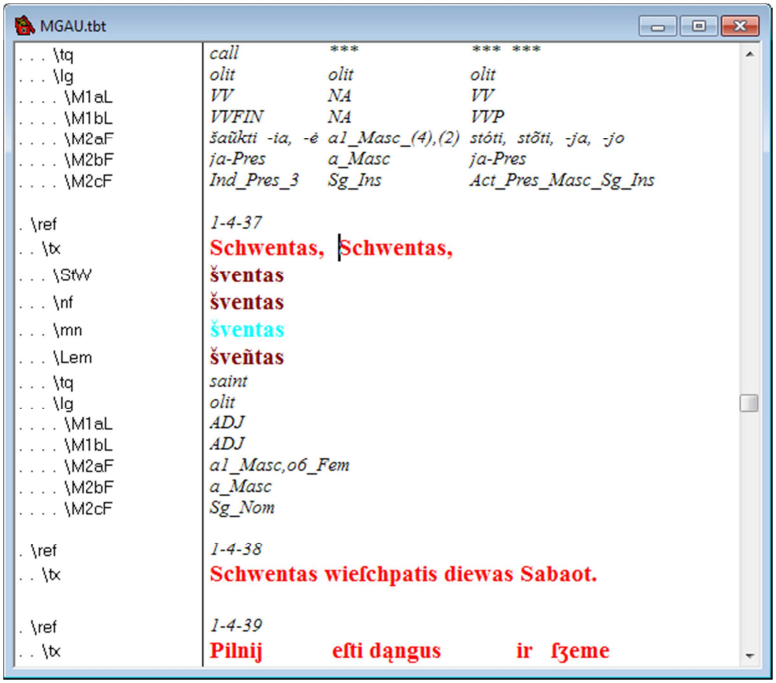


Figure 6: Steps of ambiguity selection for the annotation of the word form *Schwentas* "saint"

Afterwards, the annotations will be revised and corrected by an annotator in the software *ELAN*. Furthermore, the texts will be provided with the basic syntactic information in the software directly. Single Latin, German, or Polish words and sentences that occur within the Lithuanian texts will be annotated by hand correspondingly.

6 CONCLUSION

The multi-layer stand-off architecture (the architecture, in which every layer is a separate document, and nonetheless all layers are synchronised among themselves) of the tags and the amount of the texts in the test phase of the *Old Lithuanian Reference Corpus* require solutions for automated processes that could help to save time. This can be achieved using different databases, compiled from available lexical and grammatical resources. The morphological annotation of the Old Lithuanian word forms can be done using a modified version of the morphology analysis software of Modern Lithuanian. Nevertheless, the rich inflection of Lithuanian results in a very high rate of homographs. Their grammatical disambiguation still has to be solved manually.

[1]In 2010 SLIEKKAS was supported by the Lithuanian Ministry of Education and Science. Since 2012 it is funded by a grant No. VAT-42/2012 from the Research Council of Lithuania and performed in cooperation of the Goethe-University Frankfurt/Main and of the Institute of Lithuanian Language (Vilnius).
[2]The *Referenzcorpus Altdeutsch* is carried through by the Friedrich-Schiller-University of Jena, the Humboldt-University of Berlin, and the Goethe-University of Frankfurt/Main: http://www.deutschdiachrondigital.de/.

Bibliography

Kimbrys, P. (2000). Biblijos vardų žodynas. Aidai, Vilnius.

LKŽ: Lietuvių kalbos žodynas, vol. 1–20. Vilnius, 1941–2002. URL: http://www.lkz.lt/

Palionis, J. (2004). XVI–XVII a. lietuviškų raštų atrankinis žodynas. Mokslo ir enciklopedijų leidybos institutas, Vilnius.

Skardžius, P. (1998[1931]). Die slavischen Lehnwörter im Altlitauischen. In Rinktiniai raštai, vol. 4. Mokslo ir enciklopedijų leidybos institutas, Vilnius. 61–309

Zinkevičius, V. (2000). Lemuoklis–morfologinei analizei. In Darbai ir dienos, vol. 24. Vytauto Didžiojo Universitetas, Kaunas. 245–273. URL: http://donelaitis.vdu.lt/publikacijos/zinkevicius.pdf

Zinkevičius, V., Drożdżyński, W., Homola, P., and Piskorski, J. (2003). Adapting SProUT to processing Baltic and Slavonic languages. In Proceedings of the Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages, held in conjunction with the Conference Recent Advances in Natural Language Processing, 10-12 September 2003, Borovets, Bulgaria. URL: http://www.dfki.de/dfkibib/publications/docs/homola_baltslavir.pdf

Zinkevičius, V. (2008). The Digitization of the Dictionary of the Lithuanian Language. In The Third Baltic Conference on Human Language Technologies (October 4–5, 2007), Kaunas. Vytauto Didžiojo universitetas, Lietuvių kalbos institutas, Vilnius. 349–355