R. Mittmann

OLD GERMAN[1] AND OLD LITHUANIAN[2]: THE CREATION OF TWO DEEPLY-ANNOTATED HISTORICAL TEXT CORPORA

Abstract. The Old German and the Old Lithuanian Reference Corpus are two deeply-annotated corpora of Old German and Old Lithuanian that are created by enriching the digitized texts with additional data. To reduce conceptual effort and to establish harmonized structures, a coordinated approach was chosen. However, large differences in the availability of resources for annotation, but also in the suitability of modern-language resources resulted in considerable discrepancies in the courses of action.

1. Introduction

The creation of deeply-annotated corpora of past language stages constitutes an important task for present-day historical linguistics. The approach to choose for the achievement of this aim depends notably on certain factors concerning the particular corpus: the most crucial one is the question of existing resources that have already been set up in the course of previous analyses of the corpus. Using the example of the Old German and the Old Lithuanian Reference Corpus (below: OG/OGRC and OL/OLRC), it will be shown how differing starting points regarding preliminary work and differing qualities of OG and OL themselves result in diverging approaches to develop the corpora.

2. Description of the corpora

The Old German Reference Corpus (Referenzkorpus Altdeutsch) covers all preserved texts from the oldest stages of German – Old High German and Old Saxon –, dating from ca. 750 to 1050 CE, and comprises a total of 650,000 word tokens[3]. The project at the German

[1] Grants № GI 217/7-1 and GI 217/7-2 (German Research Foundation).

[2] Grant № VAT-42/2012 (Research Council of Lithuania).

[3] cf. **http://www.deutschdiachrondigital.de**

Universities of Berlin (Humboldt University), Frankfurt/Main and Jena has started in 2008, and several subcorpora are already online[4].

The Old Lithuanian Reference Corpus (Senosios lietuvių kalbos korpusas) covers the preserved texts of the oldest stage of Lithuanian, which date from ca. 1520 to 1800 CE, and comprises about 10 million word tokens[5]. A pilot project covering 540,000 word tokens has started in 2012 at the Lithuanian Language Institute (LKI), Vilnius, in cooperation with the Universities of Frankfurt/Main and Pisa (Italy). Due to this time lag and their cooperation in Frankfurt, the concept of the OLRC could draw upon the experiences made with the OGRC.

Both corpora cover religious as well as secular texts, prose as well as poetry and translated or adapted texts as well as independently composed texts. The language of the texts varies due to diachronic, diatopic and diastratic differences. The foreign-language source texts (mostly Latin, for OL also Polish and German) and foreign-language words within the texts are annotated as similarly as possible to the OG or OL word tokens to ensure optimal comparability. They are already comprised in the word token numbers given above. In the case of OL, a balanced choice of texts has been made for the pilot project, considering the aforementioned attributes.

3. The unequal starting points

The texts of OL are considerably closer to Modern Lithuanian than the OG ones are to Modern (High or Low) German. This is not only manifest from their actual age: as the invention of the printing press in the 15th century and the subsequent spread of written texts have significantly slowed down the transformation pace of European literary languages, the moderate development of OL to Modern Lithuanian[6] (and of the High and Low German languages in the same time

[4] cf. **http://korpling.german.hu-berlin.de/ddd**

[5] cf. *Gelumbeckaitė J., Šinkūnas M., Zinkevičius V.* Senosios lietuvių kalbos tekstynas (SLIEKKAS) – nauja diachroninio tekstyno samprata // Darbai ir dienos. Kaunas, 2012. № 58. P. 257–281.

[6] In spite of the smaller changes in phonology, the spelling of OL is

as well) cannot at all be compared to the extensive mutations in the vowel system between OG and Early Modern Times.

These divergent conditions may explain why there is no historic dictionary of Lithuanian[7] and no OL grammar that could be made use of, but also, why dictionaries and grammars of Modern Lithuanian can indeed be helpful when analysing the OL corpus. For OG, though, specific dictionaries and grammars exist. Moreover, there are glossaries for every subcorpus that give all attested inflected word forms and relate them to the corresponding lemmata. So while the OLRC will be the essential basis for the compilation of an OL grammar and glossary, the OGRC will have to question and to amend the existing works.

As to the digital availability of the texts, the two corpora hardly differed in the beginning: for all OG texts, one of the printed editions had already been digitized by the TITUS project[8] in Frankfurt/Main. In the case of the OL pilot project, editions of six out of ten texts are online on TITUS[9]; for one of the remaining, an edition is being prepared, the others are adopted from the OL database of the LKI[10]. The TITUS texts are already provided with a structural annotation, giving, e.g., chapters or lines both for the original document and the edition. This information can directly be adopted, together with the texts.

In the OGRC, a digitized edition of every text serves as the main reference layer; the manual addition of the original text forms and their graphical peculiarities is saved for later and is only performed by way of example. In the OLRC, in contrast, the digitized text editions are directly expanded by the versions of the original manuscripts or prints, together with a detailed representation of amendments, so a digitization of the originals – manuscripts or prints – is required.

very different from the current one, unstable and abounding in variants.

[7] OL dictionaries, however, exist and are made use of, cf. Section 4.2.

[8] cf. **http://titus.uni-frankfurt.de/texte/texte2.htm#ahd** and **http://titus.uni-frankfurt.de/texte/texte2.htm#asachs**

[9] cf. **http://titus.uni-frankfurt.de/texte/texte2.htm#lit**

[10] cf. **http://www.lki.lt/seniejirastai**

4. The courses of action

4.1. Old German Reference Corpus

For the OGRC, a comprehensive automated pre-annotation of the texts was feasible by digitizing the glossaries for the subcorpora into an XML format[11] and linking the part-of-speech and morphological data specified for the word forms to the word tokens in the texts. For the latter, the data required were extracted from the glossary files, and enriched with additional part-of-speech and morphological information manually extracted from the respective grammars[12]. As the glossaries generally give the attestations with their locations in the text, even with ambiguous word forms a one-to-one attribution was mostly possible. A consistent spelling and Modern German translation of the lemmata being aspired, the glossary lemmata had to be adapted to standard dictionaries of Old High German and Old Saxon.

The texts are converted into the format of ELAN[13], software developed by the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, where they are represented with the part-of-speech, morphological, lemmatical and structural pre-annotation in a database structure. This information is amended manually, ambiguities are dissolved, and a simple syntactical annotation is added[14]. Once

[11] cf. *Mittmann R.* Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen // Journal for Language Technology and Computational Linguistics. Berlin, 2013. № 27 (2). P. 39–52. **http://www.jlcl.org/2012_Heft2/3Mittmann.pdf.**

[12] cf. *Linde S., Mittmann R.* Old German Reference Corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries // Corpus Linguistics and Interdisciplinary Perspectives on Language. Tübingen, 2013 (forthcoming). № 3.

[13] cf. **http://tla.mpi.nl/tools/tla-tools/elan**

[14] cf. *Linde S.* Manuelle Abgleichung bei automatisierter Vorannotation: Das Tagging grammatischer Kategorien im Referenzkorpus Altdeutsch // Journal for Language Technology and Computational Linguistics. Berlin, 2013. № 27 (2). P. 53–64. **http://www.jlcl.org/2012_Heft2/4Linde.pdf**.

this is finished, a standardized version of all word tokens is created from the lemmata plus the part-of-speech and morphological data. For this purpose, the morphological knowledge on the corresponding language stages has been conveyed into a Perl program. The standard word forms are also used to detect annotation mistakes by comparing them automatically with the word forms given by the text edition.

4.2. Old Lithuanian Reference Corpus

Lacking any glossaries for the automated pre-annotation of the text, the OLRC is reliant on an annotation tool that is able to learn from the manual annotation and to transfer these specifications to similar cases. To this end, Toolbox[15], software developed by SIL International in Dallas, Texas, is applied, which makes use of expansible dictionaries. For the OLRC, one of the two utilizes the data of Lemuoklis[16], a morphological analyser, lemmatizer and tagger for Modern Lithuanian developed at the LKI, that are enriched by semi-manually classified data from dictionaries on OL, Slavic loanwords in OL and Bible names[17]. The other one uses the Lithuanian language dictionary: the data on all lemmata in the corpus are retrieved from its digital version[18]. The word forms of the OL texts are lemmatized automatically or, if this fails, manually. Standardized word forms are generated by Lemuoklis from lemmata, part-of-speech and morphological annotation. The addition of a Lithuanian-English dictionary enables a lemma

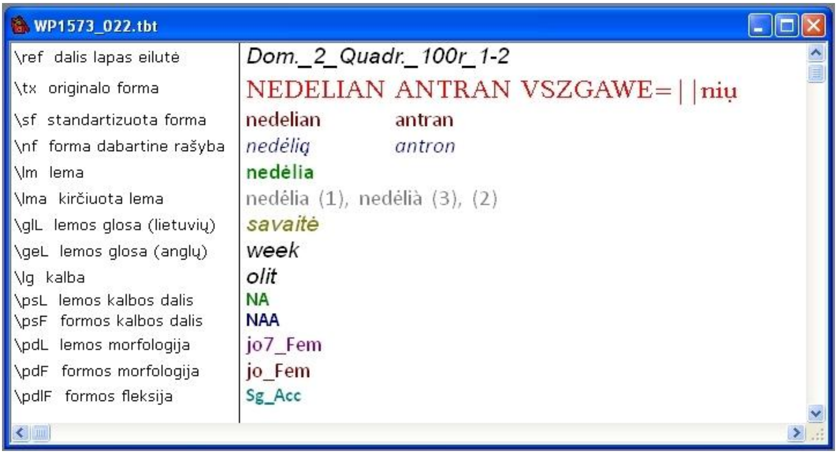
[15] cf. **http://www.sil.org/computing/toolbox**

[16] cf. **http://donelaitis.vdu.lt/~vytas/tool/tool.ppt** and *Zinkevičius V.* Lemuoklis – morfologinei analizei // Darbai ir dienos. Kaunas, 2000. № 24. P. 245–273. **http://donelaitis.vdu.lt/publikacijos/zinkevicius.pdf**.

[17] cf. *Gelumbeckaitė J., Šinkūnas M., Zinkevičius V.* Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation // Journal for Language Technology and Computational Linguistics. Berlin, 2013. № 27 (2). P. 83–96. **http://www.jlcl.org/2012_Heft2/6GelumbeckaiteEtAl.pdf**.

[18] cf. *Zinkevičius V.* The Digitization of the Dictionary of the Lithuanian Language // The Third Baltic Conference on Human Language Technologies. Vilnius, 2008. P. 349–355.

translation. To convey the attested word tokens into a standardized spelling, SIL’s Consistent Changes Program[19] is used. Primarily for the older texts, specific rules need to be created for every single author.



\ref dalis lapas eilutė	Dom._2_Quadr._100r_1-2	
\tx originalo forma	NEDELIAN ANTRAN VSZGAWE=||nių	
\sf standartizuota forma	nedelian	antran
\nf forma dabartine rašyba	nedėlią	antron
\lm lema	nedėlia	
\lma kirčiuota lema	nedė́lia (1), nedėlià (3), (2)	
\glL lemos glosa (lietuvių)	savaitė	
\geL lemos glosa (anglų)	week	
\lg kalba	olit	
\psL lemos kalbos dalis	NA	
\psF formos kalbos dalis	NAA	
\pdL lemos morfologija	jo7_Fem	
\pdF formos morfologija	jo_Fem	
\pdlF formos fleksija	Sg_Acc	

Fig. 1. Annotation in Toolbox (OLRC)

In Toolbox (cf. Fig. 1), the texts are joined with Lemuoklis’ data, and disambiguation is performed by hand. As Toolbox lacks a chart structure and cannot handle the amount of annotation layers required, the data are then transferred into ELAN where, e.g., information on multiword expressions, quotations and glossing of words is added manually. As well as for the OGRC, during the conveyance into ELAN, all word forms are split up into graphemes, enabling an annotation of their specific features. To annotate the facsimiles of the original documents, the data are again converted into the format of the image annotation tool ImAnTo, developed at Frankfurt University. Here, details of the images can be selected and linked to an annotation.

[19] cf. **http://www.sil.org/computing/catalog/show_software.asp?id=4**

4.3. Parallel processing

For the part-of-speech and morphological annotation of the OGRC, the TIGER Morphology Annotation Scheme[20], based on the Stuttgart-Tübingen Tagset (STTS) and developed for Modern German, was adapted for historical stages of German. This tagset, developed together with the Middle High German Reference Corpus[21], was in turn used for the creation of the tagset for the OLRC. Both for the part-of-speech and for the morphological annotation, it distinguishes between the lemma-specific and the record-specific qualities of the word tokens, as there are various cases in which this proves necessary. The language of the word tokens is given according to ISO 639-3[22].

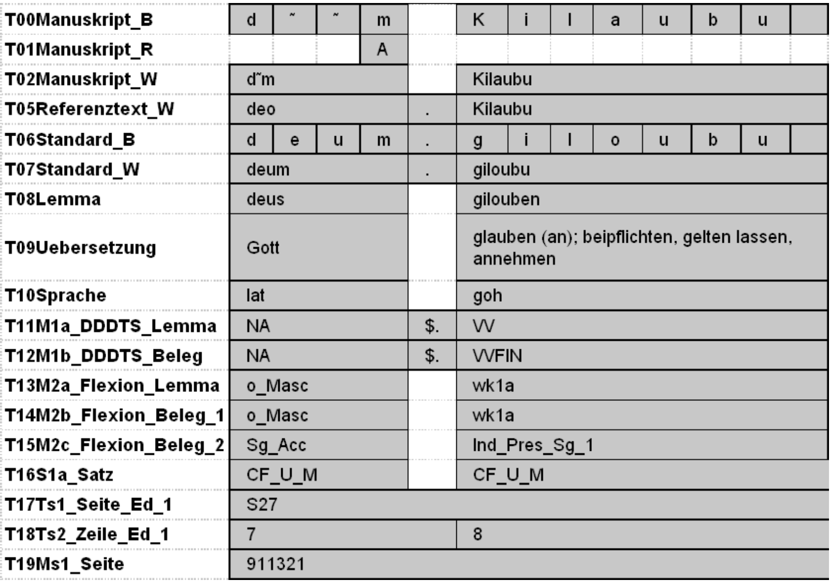
Once completed, the subcorpora of both projects are transferred into the ANNIS database (cf. Fig. 2), hosted at Potsdam University, Germany. All texts are joined with an extensive metadata description, conceived by Middle High German and OGRC and adapted by the OLRC. Complex search patterns comprising both the annotated texts and their metadata can be used to search within the corpora[23].

[20] cf. **https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-morph.pdf**.

[21] cf. *Dipper S.* Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison // Journal for Language Technology and Computational Linguistics. Berlin, 2011. № 26 (2). P. 25–37. **http://www.jlcl.org/ 2011_Heft2/2.pdf**.

[22] cf. **http://www.sil.org/iso639-3/codes.asp**

[23] cf. *Chiarcos C., Dipper S., Götze M., Leser U., Lüdeling A., Ritz J., Stede M.* A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets // Traitement Automatique des Langues. Paris, 2008. № 49 (2). P. 217–246. **http://www.atala.org/A-Flexible-Framework-for**



T00Manuskript_B	d	˜	˜	m		K	i	l	a	u	b	u	
T01Manuskript_R				A									
T02Manuskript_W	d˜m					Kilaubu							
T05Referenztext_W	deo				.	Kilaubu							
T06Standard_B	d	e	u	m	.	g	i	l	o	u	b	u	
T07Standard_W	deum				.	giloubu							
T08Lemma	deus					gilouben							
T09Uebersetzung	Gott					glauben (an); beipflichten, gelten lassen, annehmen							
T10Sprache	lat					goh							
T11M1a_DDDTS_Lemma	NA				$.	VV							
T12M1b_DDDTS_Beleg	NA				$.	VVFIN							
T13M2a_Flexion_Lemma	o_Masc					wk1a							
T14M2b_Flexion_Beleg_1	o_Masc					wk1a							
T15M2c_Flexion_Beleg_2	Sg_Acc					Ind_Pres_Sg_1							
T16S1a_Satz	CF_U_M					CF_U_M							
T17Ts1_Seite_Ed_1	S27												
T18Ts2_Zeile_Ed_1	7					8							
T19Ms1_Seite	911321												

Fig. 2. Representation in ANNIS (OGRC)

5. Conclusion

Although initially various aspects were significantly diverging, the work on the OLRC can benefit from the course of action applied for the OGRC in various ways. All the same, it can also make use of the digitized data and tools already at hand for Modern Lithuanian – an approach inapplicable for OG. The deficiency in glossaries for the OL texts results in a need for an additional adaptive annotation tool. Special approaches are also required for objectives exceeding those of the OGRC, such as a precise annotation of the facsimiles of the original documents. Still, the cooperation of the two research projects has proved and is still proving an advantageous decision, as more time can be devoted to the actual philological work if the wheel, once conceived, does not have to be reinvented.