New Technologies for Old Germanic. Resources and Research on Parallel Gospels in Older Continental Western Germanic

Christian Chiarcos
Jens Chobotský
Gaye Detmolt
Roland Mittmann
Maria Sukhareva

Goethe University Frankfurt am Main, Germany

December 16, 2013

We describe on-going efforts at the Goethe University Frankfurt on the study of older Continental Western Germanic languages, in particular, Old High German (OHG, antecessor of German) and Old Saxon (OS, antecessor of Low German and closely related to the antecessor of Dutch) and their relation to Old English (OE), Gothic, German and other Germanic languages as well as the relation of OHG and OS religious texts to their Latin sources. This line of research is conducted in the context of two larger efforts, the Old German Reference Corpus and the LOEWE cluster "Digital Humanities", in collaboration with the Applied Computational Linguistics group at the Goethe-Universität Frankfurt.

The Old German Reference Corpus is a DFG-funded project that emerged from the Deutsch Diachron Digital (DDD) initiative, conducted in cooperation between HU Berlin, U Frankfurt and U Jena, and aims to provide a morphosyntactically annotatated, exhaustive reference corpus of Old High German and Old Saxon. The LOEWE cluster "Digital Humanities"[1], funded through a programme of the State of Hessen, is a collaboration between U Frankfurt, TU Darmstadt and Freies Deutsches Hochstift Frankfurt aiming to develop methodologies and infrastructures to facilitate information-technological support of research in the humanities.

Here, we concentrate on biblical texts: These are available for a variety of modern and historical European languages and possess high-quality alignment (verses, segments), thus building up a valuable parallel resource for linguistic, philological and historical research questions, as well as for Natural Language Processing, whose methodologies for alignment and annotation projection can be used to support the analysis of these texts:

- The **Old German Reference Corpus** [4] provides a lexicon and an exhaustive corpus of older continental Western Germanic, i.e., Old Saxon (OS) and Old High German (OHG), comprising 650,000 tokens automatically enriched with morphological and morphosyntactic information drawn from existing glossaries which have been digitized by the project, complemented with manual annotations[3] and metadata and published via the ANNIS database [2].

- A **Historical Linguistic Database** was developed in LOEWE from a collection of etymological dictionaries for all Old Germanic languages (incl. OS, OHG, Old English, Gothic, Old Norse) as a relational data base providing user-friendly means of comparing etymologically related forms between historical dialects and their daughter languages, as well as a machine-readable view on these [5].

- Major texts in the corpus are the **gospel harmonies** associated with the names Heliand (OS), Tatian (OHG and Latin) and Otfrid (OHG). Although not direct translations of the

[1] http://www.digital-humanities-hessen.de

Bible and hence not directly alignable with the gospel translations we have for Old English, Gothic, and later stages of English, German, Dutch and North Germanic, a section-level alignment has been manually extrapolated from the literature in a LOEWE project [5].

- This coarse-grained alignment is currently being refined to a **phrase-level alignment** using the linked lexical resources mentioned above as well as statistical models of systematic character correspondences like those applied by [1].
- On the basis of correspondences between historical and modern languages in parallel and quasi-parallel text, **statistical annotation projection** can be applied for the syntactic annotation of Older Germanic. So far, we conducted experiments on the joint projection of dependency syntax to Old English, Middle Icelandic and Early Modern High German corpora following the methodology of [6]. These indicate that projected annotations can serve as training data for mono- and cross-language parsing also for, e.g., OHG.
- These annotations can be applied, for example, to compare linguistic structures in OHG gospel harmonies and their Latin sources, thereby facilitating the research of a LOEWE project that currently uses statistical word alignment and existing *morpho*syntactic annotations only as the basis for a **qualitative, philological comparison** with the TreeAligner [7].

References

[1] Marcel Bollmann. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW/ID-2013)*, pages 11–18, Sofia, Bulgaria, Aug 2013.

[2] Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues (TAL)*, 49(2), 2008.

[3] Sonja Linde and Roland Mittmann. Old German Reference Corpus. Digitizing the knowledge of the 19th century. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics / Korpuslinguistik und interdiziplinäre Perspektiven auf Sprache*, Korpuslinguistik und interdiziplinäre Perspektiven auf Sprache / Corpus linguistics and Interdisciplinary perspectives on language (CLIP): 3, Tübingen, 2013. Narr.

[4] Roland Mittmann. Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(2):39–52, 2013. Special issue *Altüberlieferte Sprachen als Gegenstand der Texttechnologie / Text Technological Mining of Ancient Languages.*

[5] Timothy Blaine Price. Multi-faceted Alignment: Toward Automatic Detection of Textual Similarity in Gospel-derived Texts. In *Proceedings of Historical Corpora 2012*, Frankfurt, Dec 2012.

[6] Kathrin Spreyer and Jonas Kuhn. Data-Driven Dependency Parsing of New Languages Using Incomplete and Noisy Training Data. In *Proceedings of CoNLL*, pages 12–20, Boulder, CO, Jun 2009.

[7] Martin Volk, Joakim Lundborg, and Maël Mettler. A search tool for parallel treebanks. In *Proceedings of the 1st Linguistic Annotation Workshop (LAW-2007)*, pages 85–92, Prague, Czech Republic, Jun 2007.