Altüberlieferte Sprachen als Gegenstand der Texttechnologie

Ancient Languages as the Object of Text Technology

Guest editors: Armin Hoenen and Thomas Jügel

Text technology predominantly focuses on modern languages. Most of these languages supply text technology with a significant amount of digitised texts. Constantly growing internet resources such as various Wikis produce new analysable data, which serve as input and testing grounds for statistical models, rule-based analyses and various other research tools. While the number of available tokens might easily reach billions, historians usually deal with a fixed set of data and can deem themselves lucky if their texts are complete. If, then, the corpus size reaches to several thousands of words, it is even better.

The aim of this volume is to represent several projects on historical corpora: Corpus Avesticum (Old Iranian), Mercurius Corpus of Early New High German, Referenzkorpus Altdeutsch (Old German Reference Corpus), Old Lithuanian Reference Corpus (SLIEKKAS). As such, the focus is on the humanities' perspective, so to say, on the user's side of text technological tools.

Historical corpora can have several drawbacks, and their evaluation depends heavily on the expertise of specialists. Such corpora are inherently limited, texts might be incomplete, it is likely that we only have an incomplete knowledge of the grammatical system, and there remain uncertainties in lexical meaning. Language change cannot always be clearly located in space and time, so that ambigue phrases in ancient texts could allow for two different interpretations: an old and a new one, or as aptly put by PAULY et al. (this volume): "die Ausgangs- oder die Zielstruktur". Gaps in transmission (be it loss of material, or be it loss of text when transmitted orally), divergent variants due to the copying process, and the uncertainty of place, time, and author of texts all make the identification of originals a difficult task. Hence, the digitisation and computational evaluation of historical linguistic corpora faces problems that can issue challenges to text technology. It is obvious that techniques like machine learning suffer from lack of training data—but there is more to text technology.

With a good annotation for a corpus at hand (the building of which is a labour intensive task), more sophisticated analyses can be conducted by means of technological applications, such as a search tailored for the historian or linguist. Combined queries for word forms allow the investigation of grammatical rules: Which mood follows specific conjunctions? Which case is governed by prepositions? Can prepositions be used as postpositions as well? Which kind of word order is possible? Furthermore, an analysis of co-occurrences of a term under consideration in texts that belong to different eras can reveal semantic changes. Repetitions can reveal the original structure of a text (e.g., of a ceremony), if, e.g., these have been obscured by a modern division into chapters. Detecting intertextual dependencies enables us to trace the path a text took through time and languages, i.e., to trace its reception.

Annotation is the very starting point of historical text technology and, thus, it is a main focus of the present volume. The articles of MITTMANN and LINDE discuss automatic pre-annotation of glossaries of Old German and the problems that occur when defining tags for information from printed media. PAULY et al. examine the representation of ambiguous and discontinuous phrases in Early New High German under the scope of language development. Similar in method is the project on annotating Old Lithuanian texts by GELUMBECKAITĖ et al.

Another complex contains articles representing issues of the Old Iranian language Avestan. GIPPERT gives an overview of the encoding strategies of the complex Avestan writing system. JÜGEL considers the problems concerning the automatic generation of stemmata for Avestan manuscripts. By means of this volume, we hope to bring the disciplines of humanities and text technology closer to one another and to support the exchange of information between these fields of study.

At long last, we would like to thank Dr. Timothy Price, for he not only checked the English articles for spelling, flow, and style, but also gave many helpful comments that helped to improve the articles enormously. We would also like to express our gratitude to Dr. Lothar Lemnitzer and the *Gesellschaft für Sprachtechnologie und Computerlinguistik* for accepting this volume to be published in the JLCL series. Last but not least, our appreciation goes to the many reviewers who so generously spent their time in helping us to improve our articles.

The Editors
Frankfurt am Main, 2012

Content