Multilinguality in historical documents – challenges and solutions for digital humanities

Maximum Number of Participants: 40

Date: Monday, July 7, 2014 – All Day – 09:00 to 12:00 + 13:00 to 17:00

Facilitator(s):

Stefanie Dipper, Prof. Dr.
Noah Bubenhofer, Dr.
Laurent Romary, Dr. Research Director INRIA
Cristina Vertan, Dr.

Overview:

Recently, the collaboration between the Language Technology community and the specialists in various areas of the Humanities has become more efficient and fruitful due to the common aim of exploring and preserving cultural heritage data. It is worth mentioning the efforts made during the digitisation campaigns in the last years and within a series of initiatives in the Digital Humanities, especially in making old manuscripts and prints available in the form of Digital Libraries.
The availability of old texts on-line produced a revolutionary shift in the way how such objects are analysed. They are no longer restricted to a small number of specialists, knowing the language of the document but to broader groups with various requirements:

1. non-expert users who would like to know what the document is about, understand the main topics, localise places, persons. These users have no or very little knowledge of old languages, and usually are less familiarised with toponyms (especially when these belong to geographical spaces unknown to the user);
2. researchers of neighbour fields, who often have only minimal knowledge of the language but considerable knowledge of the historical context and might be familiarised with historical toponyms and proper names;
3. students and researchers specialising in historical data, who have the required language skills but still can profit from additional information accompanying the texts.

These considerations imply that the storage and visualisation of old texts should be accompanied by a collection of tools empowering the text with suitable information and making it understandable for different user groups. Such tools usually involve automatic language processing methods. In contrast to processing of modern texts, for which language technology made a huge progress in the last years, automatic processing of old texts is still problematic mainly because:

- Historical language data is sparse.
- Texts are often multilingual.

The focus of this workshop is on the second aspect. We think that the challenges posed by multilinguality should be tackled by adapting existing multilingual language resources and tools, and, where necessary, by providing training data in the form of corpora or lexicons for a certain period of time in history.

Audience:

The aim of this workshop is to bring together researchers working in this interdiciplinary domain as well as specialists in machine translation and multilinguality working with languages with sparse resources, to analyse problems and brainstorm solutions in order to implement machine (-aided) translation and processing for (multilingual) historical texts. We envisage also networking with European activites in Digital Humanities like CENDARI, CLARIN, DARIAH.

Topics of interest include but are not limited to:
- character-level MT for normalisation
- historical and modern data as comparable corpora
- historical texts in different languages as parallel or comparable corpora
- MT for translation between language versions
- OCR for multilingual documents
- word- and/or paragraph-level language identification
- crosslingual retrieval in historical documents
- ontologies as language-independent interfaces between collections of historical texts
- particularities of multilingual historical texts and challenges for IT