LT-DHA 2019

Proceedings of the Workshop on Language Technology for Digital Historical Archives - with a Special Focus on Central-, (South-)Eastern Europe, Middle East and North Africa

in conjunction with

The 12th International Conference on Recent Advances in Natural Language Processing (RANLP 2019)

5 September, 2019
Varna, Bulgaria

LANGUAGE TECHNOLOGY FOR DIGITAL HISTORICAL ARCHIVES
IN CONJUNCTION WITH THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING'2019

PROCEEDINGS

Varna, Bulgaria
5 September 2019

ISBN 978-954-452-059-5

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Foreword

During the last decades Digital Humanities evolved dramatically, from simple database applications to complex systems involving most recent state-of-the art in Computer Science. Especially Language Technology plays a major role either for processing the metadata of recorded objects or for analyzing and interpreting content. Applying Language Technology methods to objects from humanities in general and historical archives in particular, is a challenge for NLP-related research: data is heterogeneous (image /text), often incomplete (e.g. OCR errors), multilingual within one document (historic documents with Latin or/and classical Greek paragraphs) and difficult to structure (paragraphs, titles, pages are somewhat different in historical texts).

Corpus-based methods, nowadays standard in NLP research, often cannot be applied as the necessary large training data is missing.

Moreover, requirements for tools in Digital Humanities, especially tools dedicated to cultural heritage objects, are different from the ones applied to modern texts. Thus, performing research in Digital Humanities involves also: adapting existent NLP tools to the historical variants of languages; developing tools for new languages; making tools robust to syntactic deviation; and adapting semantic resources.

Central and Eastern Europe as well as the Middle East and North Africa were always characterized by a high concentration of languages and cultures, interacting with each other. On a relatively small area texts written with at least 10 alphabets (Arabic, Hebrew, Armenian, Georgian, Greek, Cyrillic, Geez, Syriac and Latin, Coptic) can be found. On the other hand, information within these texts is important beyond the borders of a given language or script. (e.g. often documents in Ge'ez are translations of lost Coptic or ancient Greek texts). Places, Persons, Events have language-dependent denominations but refer to the same individual or geographical location.

Unfortunately, especially in this area many historical documents are in bad condition; many languages or dialects became extinct over the time and their written evidence is rare. Digital methods seem the perfect means for preservation and investigation of this rich cultural heritage asset. However, up to now, concentrated activities seem to be absent, probably also due to the lack of adequate NLP resources and tools. Thus, it is very necessary to evaluate existent technology, monitor current activities, network research teams in this area - all aims of this workshop.

This is the second edition of Language technology for Digital Humanities in Central and (South-)Eastern Europe workshop, held in 2017 at RANLP. In the 2019 International Year of Indigenous Languages this edition expands also to Middle East and North Africa.

The Organisers thank the members of the Programme Committee for the valuable help in selecting the papers.

Cristina Vertan, Petya Osenova and Dimitar Iliev

Organizers:

Cristina Vertan, University of Hamburg
Petya Osenova, Bulgarian cdemy of Sciences and St. Kliment Ohridski University of Sofia
Dimitar Iliev, St. Kliment Ohridski University of Sofia

Program Committee:

Martha Yifiru Abate, University of Addis Ababa
Gabriel Bodard, Institute of Classical Studies, SAS, London
Elie Damaoui, University of Balamand
Antske Fokkens Vrije Universiteit, Amsterdam
Walther v. Hahn, University of Hamburg
Vladislav Kubon, Charles University, Prague
Preslav Nakov, Qatar University
Maciej Ogrodniczuk, Polish Academy of Science
Gabor Proszeky , Catholic University, Budapest
Kiril Simov, Bulgarian Academy of Sciences
Stefan Trausan, Politechnics University, Bucharest
Valeria Vitale, Institute of Classical Studies, SAS, London

Invited Speaker:

Alicia Gonzalez Martinez, University of Hamburg

Table of Contents

Conference Program

09:30–09:40 ***Opening***

09:40–10:30 *Graphemic ambiguous queries on Arabic-scripted historical corpora*
Alicia González Martínez

10:30–11:00 ***Coffee Break***

11:00–12:30 **Corpus Annotation**

11:00–11:30 *Word Clustering for Historical Newspapers Analysis*
Lidia Pivovarova, Elaine Zosa and Jani Marjanen

11:30–12:00 *Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition*
Tannon Kew, Anastassia Shaitarova, Isabel Meraner, Janis Goldzycher, Simon Clematide and Martin Volk

12:00–12:30 *Controlled Semi-automatic Annotation of Classical Ethiopic*
Cristina Vertan

12:30–14:00 ***Lunch***

14:00–16:00 **Integration of NLP and Knowledge Representation**

14:00–14:30 *Implementing an archival, multilingual and Semantic Web-compliant taxonomy by means of SKOS (Simple Knowledge Organization System)*
Francesco Gelati

14:30–15:00 *EU 4 U: An educational platform for the cultural heritage of the EU*
Maria Stambolieva

15:00–15:30 *Modelling linguistic vagueness and uncertainty in historical texts*
Cristina Vertan

15:30–16:00 ***Discussions; Concluding Remarks***