DigHum 2011

Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage

associated with

The 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)

Edited by

Cristina Vertan, Milena Slavcheva, Petya Osenova and Stelios Piperidis

16 September, 2011

Hissar, Bulgaria

INTERNATIONAL WORKSHOP
LANGUAGE TECHNOLOGIES FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE

PROCEEDINGS

Hissar, Bulgaria
16 September 2011

ISBN 978-954-452-019-9

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Foreword

Following several digitization campaigns during the last years, a large number of printed books, manuscripts and archaeological digital objects have become available through web portals and associated infrastructures to a broader public. These infrastructures enable not only virtual research and easier access to materials independent of their physical place, but also play a major role in the long term preservation and exploration.

However, the access to digital materials opens new possibilities of textual research like: synchronous browsing of several materials, extraction of relevant passages for a certain event from different sources, rapid search through thousand pages, categorisation of sources, multilingual retrieval and support, etc.

Methods from Language Technology are therefore highly required in order to ensure extraction of content related semantic metadata, and analysis of textual materials. There are several initiatives in Europe aiming to foster the application of language technology in the humanities (CLARIN, DARIAH). Through initiatives like those, as well as many other research projects, the awareness of such methods for the humanities has risen considerably. However, there is still enough potential on both sides:

• on one hand, there are still research tracks in the humanities which do not sufficiently and effectively exploit language technology solutions;

• on the other hand, there are many languages, especially historical variants of languages, for which the available tools and resources still have to be developed or adapted to serve successfully humanities applications.

The current workshop brings together researchers from the Humanities, as well as from Language and Information Technologies, and thus fosters the above mentioned directions.

As a confirmation of the generated interest in the topic of our workshop, we received a large number of very good submissions. This fact allowed us to provide a programme covering the most important aspects within the area of digital humanities and cultural heritage. Following the workshop programme, the Proceedings of the workshop are thematically structured as follows: Electronic Archives, Language Technology and Resources, Computational Methods for Literary Analysis, Multimodal Aspects in Digital Humanities.

The workshop papers address a multitude of problems and suggest a wealth of developments and solutions related to the digital humanities and the preservation of cultural heritage. The papers represent a whole spectrum of relevant topics: utilizing interlinked semantic technologies for managing and accessing museum data; exploiting topic models in a query classification system for an art image archive; metadata and content-oriented search methods for a multilingual audio-and-video archive; maintaining a digital library of Polish and Poland-related old ephemeral prints; normalization of historical wordforms in German; developing a Bulgarian-Polish on-line dictionary as a technological tool for applications in the digital humanities; semantic annotation models based on ontological representation of knowledge concerning Bulgarian iconography; preparation of an electronic edition of the largest Old Church Slavonic manuscript, the Codex Suprasliensis; literary research support by creating and visualizing profiles of sentimental content in texts; profiling of literary characters in 19th century Swedish prose fiction by interpersonal relation extraction; investigation of diachronic stylistic changes in British and American varieties of 20th century written English language; speeding up the process of creating annotations of audio-visual data for humanities research; automatic transcription of ancient handwritten documents; OCR processing of Gothic-script documents.

We would like to thank the Organisers of the RANLP events, especially Galia Angelova and Kiril Simov, for their unceasing help in the organisation of the workshop.

We are indebted to the Programme Committee members who provided very detailed reviews in extremely short time.

Special thanks are addressed to Gábor Prószéky, who accepted to be our keynote speaker and additionally raised the interest in our workshop.

September 2011

Cristina Vertan, Milena Slavcheva, Petya Osenova and Stelios Piperidis

Workshop Organizers:

Cristina Vertan (University of Hamburg, Germany)
Milena Slavcheva (Bulgarian Academy of Sciences, Bulgaria)
Petya Osenova (Sofia University and Bulgarian Academy of Sciences, Bulgaria)
Stelios Piperidis (ILSP, Greece)

Programme Committee:

Galia Angelova (Bulgarian Academy of Sciences, Bulgaria)
David Baumann (Perseus, Tufts University, USA)
Núria Bel (University of Barcelona, Spain)
António Branco (University of Lisbon, Portugal)
Nicoletta Calzolari (University of Pisa, Italy)
Günther Görz (University of Erlangen, Germany)
Walther v. Hahn (University of Hamburg, Germany)
Fotis Jannidis (University of Würzburg, Germany)
Steven Krauwer (University of Utrecht, the Netherlands)
Éric Laporte (Université Paris-Est, Marne-la-Vallée, France)
Anke Lüdeling (Humboldt University, Berlin, Germany)
Adam Przepiórkowski (Polish Academy of Sciences, Poland)
Gábor Prószéky (MorphoLogic, Hungary)
Laurent Romary (LORIA-INRIA, Nancy, France)
Manfred Thaler (Cologne University, Germany)
Tamás Váradi (Hungarian Academy of Sciences, Hungary)
Martin Wynne (University of Oxford, UK)

Invited Speaker:

Gábor Prószéky (MorphoLogic & Pázmány University, Budapest, Hungary)
Endangered Uralic Languages and Language Technologies

Table of Contents

Multimodal Aspects in Digital Humanites

Workshop Programme

Friday, 16 September 2011

9:00–9:15 Opening

9:15–10:15 *Endangered Uralic Languages and Language Technologies*
Gábor Prószéky

10:15–10:45 Coffee Break

Electronic Archives

10:45–11:10 *A Framework for Improved Access to Museum Databases in the Semantic Web*
Dana Dannélls, Mariana Damova, Ramona Enache and Milen Chechev

11:10–11:35 *Query classification via Topic Models for an art image archive*
Dieu-Thu Le, Raffaella Bernardi and Ed Vald

11:35–12:00 *Unlocking Language Archives Using Search*
Herman Stehouwer and Eric Auer

12:00–12:25 *Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage*
Maciej Ogrodniczuk and Włodzimierz Gruszczyński

12:25–14:00 Lunch

Language Technology and Resources

14:00–14:25 *Rule-Based Normalization of Historical Texts*
Marcel Bollmann, Florian Petran and Stefanie Dipper

14:25–14:50 *Survey on Current State of Bulgarian-Polish Online Dictionary*
Ludmila Dimitrova, Ralitsa Dutsova and Rumiana Panova

14:50–15:15 *Language Technology Support for Semantic Annotation of Icono-graphic Descriptions*
Kamenka Staykova, Gennady Agre, Kiril Simov and Petya Osenova

Friday, 16 September 2011 (continued)

15:15–15:40 *The Tenth-Century Cyrillic Manuscript Codex Suprasliensis: the creation of an electronic corpus. UNESCO project (2010–2011)*
Hanne Martine Eckhoff, David Birnbaum, Anissava Miltenova and Tsvetana Dimitrova

15:40–16:10 Coffee Break

Digital Methods in Literary Analysis

16:10–16:35 *SentiProfiler: Creating Comparable Visual Profiles of Sentimental Content in Texts*
Tuomo Kakkonen and Gordana Galic Kakkonen

16:35–17:00 *Character Profiling in 19th Century Fiction*
Dimitrios Kokkinakis and Mats Malm

17:00–17:25 *Diachronic Stylistic Changes in British and American Varieties of 20th Century Written English Language*
Sanja Štajner and Ruslan Mitkov

17:25–17:40 Short break

Multimodal Aspects in Digital Humanites

17:40–18:05 *AVATecH: Audio/Video Technology for Humanities Research*
Sebastian Tschöpel, Daniel Schneider, Rolf Bardeli, Oliver Schreer, Stefano Masneri, Peter Wittenburg, Han Sloetjes, Przemek Lenkiewicz and Eric Auer

18:05–18:30 *Handwritten Text Recognition for Historical Documents*
Veronica Romero, Nicolas Serrano, Alejandro H. Toselli, Joan Andreu Sanchez and Enrique Vidal

18:30–18:55 *Reducing OCR Errors in Gothic-Script Documents*
Lenz Furrer and Martin Volk

18:55–19:15 Conclusions and closing