EACL 2009

Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education

LaTeCH – SHELT&R 2009

30 March 2009
Megaron Athens International Conference Centre
Athens, Greece

Production and Manufacturing by
TEHNOGRAFIA DIGITAL PRESS
7 Ektoros Street
152 35 Vrilissia
Athens, Greece

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Preface

The LaTeCH–SHELT&R workshop aims to present in one forum two strands of research in language technology, which we believe have many common concerns, but also complementary viewpoints.

Museums, archives, and libraries around the world maintain large collections of cultural heritage objects, such as archaeological artefacts, sound recordings, historical manuscripts, or preserved animal specimens. Large scale digitization projects are currently underway to make these collections more accessible to the public and to research. The natural next step after digitization is the development of powerful tools to search, link, enrich, and mine the digitized data. Language technology has an important role to play in this endeavor, even for collections which are primarily non-textual, since text is the pervasive medium used for metadata. Two previous LaTeCH (*Language Technology for Cultural Heritage*) workshops (at ACL 2007 in Prague and LREC 2008 in Marrakech) have shown that there is an interest among language technology researchers in providing intelligent infrastructure and tools for working with cultural heritage data.

Similarly, in research in the Social Sciences, Humanities and Education (SHE), text – and speech, i.e., *language* – are central as both primary and secondary research data sources. In today's world, the normal mode of access to text, speech, images and video is in digital form. Modern material is born digital, and, as already mentioned, older material is being digitized on a vast scale in cultural heritage and digital library projects. Language technology and language resources has an equally important role to play here, as in the cultural heritage area, and for more or less the same purposes. A clear sign of this is the newly launched European research infrastructure initiative CLARIN, which addresses exactly the use of language technology and language resources as research infrastructure in the humanities and social sciences. Against this background, it now seems natural to add a component to the workshop reflecting this development: SHELT&R (*Language Technology and Resources infrastructure for text-based research in the Social Sciences, Humanities and Education*).

The CH and SHE domains are not mere passive consumers of ready-made language technology solutions. Rather, they make up interesting and challenging testbeds, where the robustness and the generality of existing language technology are subjected to the acid test of messy and multilingual reality, more so than in many other application areas, since they have to deal with, *inter alia*, historical, non-standardized language varieties in addition to a number of modern standard languages. Our workshop thus aims to foster interaction between researchers working on all aspects of language technology applied to the CH and SHE domains, and experts from institutions who are testing deployed technologies and formulating improved use cases.

The papers accepted for the LaTeCH–SHELT&R workshop after a thorough peer-review process give a good sense of the current breadth of this exciting and expanding area. We are happy that our keynote speakers, Dr. Martin Doerr and Dr. Tamás Váradi, agreed to join us and deliver excellent topics for a complete workshop program. We would like to thank all authors who submitted papers for the hard work that went into their submissions. We are also extremely grateful to the members of the programme committee for providing thorough reviews and multi-faceted input.

Some of the workshop costs were covered from a project grant awarded to Lars Borin by the Swedish Research Council (VR Dnr 2007-7430: *Safeguarding the future of Språkbanken*), which is hereby gratefully acknowledged.

Lars Borin • Piroska Lendvai

LaTeCH – SHELT&R 2009

Workshop Chairs:

Lars Borin, University of Gothenburg (Sweden)
Piroska Lendvai, Tilburg University (The Netherlands)

Organizing Committe:

Piroska Lendvai (Co-chair), Tilburg University (The Netherlands)
Lars Borin (Co-chair), University of Gothenburg (Sweden)
Antal van den Bosch, Tilburg University (The Netherlands)
Martin Reynaert, Tilburg University (The Netherlands)
Caroline Sporleder, Saarland University (Germany)

Program Committee Members:

Ion Androutsopoulos, Athens University of Economics and Business (Greece)
Timothy Baldwin, University of Melbourne (Australia)
David Bamman, Perseus (USA)
Lars Borin, University of Gothenburg (Sweden)
Antal van den Bosch, Tilburg University (The Netherlands)
Andrea Bozzi, ILC-CNR, Pisa (Italy)
Paul Buitelaar, DERI Galway (Ireland)
Kate Byrne, University of Edinburgh (UK)
Claire Cardie, Cornell University (USA)
Paul Clough, Sheffield University (UK)
Milena P. Dobreva, CDLR, University of Strathclyde (UK)
Mick O'Donnell, Universidad Autonoma de Madrid (Spain)
Claire Grover, University of Edinburgh (UK)
Ben Hachey, University of Edinburgh (UK)
Erhard Hinrichs, Tübingen University (Germany)
Graeme Hirst, University of Toronto (Canada)
Christer Johansson, University of Bergen (Norway)
Jaap Kamps, University of Amsterdam (The Netherlands)
Dimitrios Kokkinakis, University of Gothenburg (Sweden)
Stasinos Konstantopoulos, NCSR Demokritos (Greece)
Piroska Lendvai, Tilburg University (The Netherlands)
Christina Lioma, University of Leuven (Belgium)
Anke Lüdeling, Humboldt University (Germany)
Veronique Malaisé, Free University of Amsterdam (The Netherlands)
Steven van der Mije, Trezorix (The Netherlands)
John Nerbonne, Rijksuniversiteit Groningen (The Netherlands)
Marco Pennacchiotti, Saarland University/Yahoo! Research (Germany)
Georg Rehm, vionto GmbH, Berlin (Germany)
Martin Reynaert, Tilburg University (The Netherlands)
Michael Rosner, University of Malta (Malta)
Caroline Sporleder, Saarland University (Germany)
Tamás Váradi, Hungarian Academy of Sciences (Hungary)
Andreas Witt, Tübingen University (Germany)
Svitlana Zinger, Eindhoven University of Technology (The Netherlands)

Table of Contents

Conference Program

Monday, March 30, 2009

9:00–9.15 Opening

9:15–10:15 Invited Talk by Martin Doerr

10:15–10:30 Moderated discussion

10:30–11:00 Coffee break

11:00–11:30 *Content Analysis of Museum Documentation in a Transdisciplinary Perspective*
Guenther Goerz and Martin Scholz

11:30–12:00 *An Intelligent Authoring Environment for Abstract Semantic Representations of Cultural Object Descriptions*
Stasinos Konstantopoulos, Vangelis Karkaletsis and Dimitris Bilidas

12:00–12:20 *Multiple Sequence Alignments in Linguistics*
Jelena Prokić, Martijn Wieling and John Nerbonne

12:20–12:45 *Evaluating the Pairwise String Alignment of Pronunciations*
Martijn Wieling, Jelena Prokić and John Nerbonne

12:45–14:00 Lunch

14:00–15:00 Invited talk by Tamás Váradi

15:00–15:20 *A Web-Enabled and Speech-Enhanced Parallel Corpus of Greek-Bulgarian Cultural Texts*
Voula Giouli, Nikos Glaros, Kiril Simov and Petya Osenova

15:20–15:40 *The Development of the "Index Thomisticus" Treebank Valency Lexicon*
Barbara McGillivray and Marco Passarotti

15:40–16:00 *Applying NLP Technologies to the Collection and Enrichment of Language Data on the Web to Aid Linguistic Research*
Fei Xia and William Lewis

16:00–16:30 Coffee break

Monday, March 30, 2009 (continued)

16:30–16:55 *Instance-Driven Discovery of Ontological Relation Labels*
Marieke van Erp, Antal van den Bosch, Sander Wubben and Steve Hunt

16:55–17:20 *The Role of Metadata in the Longevity of Cultural Heritage Resources*
Milena Dobreva and Nikola Ikonomov

17:25–18:00 Moderated discussion; closing