Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants

# Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants

*associated with*

**The 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)**

13 September, 2013
Hissar, Bulgaria

Adaptation of Language Resources and Tools
for Closely Related Languages and Language Variants
*associated with* THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING 2013

**PROCEEDINGS**

Hissar, Bulgaria
13 September 2013

ISBN 978-954-452-026-7

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

# Preface

Recent initiatives in language technology have led to the development of at least minimal language processing kits for all official European languages. This is a big step towards automatic processing and/or extraction of information especially from official documents produced within the European Union. Apart from those official languages, a large number of dialects or closely related variants are in use, more and more not only as spoken colloquial languages but also in written media. Building language resources and tools is a cost-expensive operation and one can benefit form similarities among languages to reduce the effort in constructing LRs. One should be, however, aware also of the discrepancies which are often visible not only at the lexical level. Two examples could be different variants of Spanish in Latin America, German spoken in Austria and Switzerland, French – in France and Belgium, Dutch – in the Netherlands and Flemish in Belgium, etc. Less attention has been paid up to now to the development of LRs for such languages. This has a major impact on promoting language technology at the educational level, using information processing methods in all-day communication, social media, etc. This workshop intends to draw attention on issues mentioned above by bringing together scientists working with less resourced language variants and producing a roadmap of existing technologies and still existing gaps.

The current workshop aims to discuss topics like:
- Adaptation of monolingual tools for close languages and language variants;
- Case studies of using LRs and tools for standard languages on documents in language variants;
- Machine translation among closely related languages;
- Evaluation of LRs and tools for language variants and close languages;
- Linguistic issues in adaptation of LRs and tools (e.g. semantic discrepancies, lexical gaps, false friends);

We are very happy to include papers addressing topics not only from different language families (Germanic, Romance, Greek, Slavonic) but also going beyond the European borders (e.g. Rio de la Plata Spanish).

We hope that the current workshop will be an impulse for further activities related to the exploitation of language similarities for text technology. Finally, we would like to thank the organizers of the RANLP Conference for making the organization of this workshop possible and the programme committee for a fast and efficient reviewing process.

Cristina Vertan, Milena Slavcheva and Petya Osenova
Organisers of the Workshop on the Adaptation of Language Resources and Tools
for Closely Related Languages and Language Variants,
held in conjunction with the International Conference RANLP-13

**Organizers:**

Cristina Vertan (University of Hamburg)
Milena Slavcheva (IICT, Bulgarian Academy of Sciences)
Petya Osenova (Sofia University “St. Kl. Ohridski” and IICT, Bulgarian Academy of Sciences)

**Program Committee:**

Laura Alonso y Alemany (Univeristy of Cordoba, Argentina)
César Antonio Aguilar (Pontificia Universidad Católica de Chile, Sntiago de Chile, Chile)
Antonio Branco (University of Lisabon)
Gerhard Budin (University of Vienna, Austria)
Jose Castaño (University of Buenos Aires, Argentina)
Walter Daelemans (University of Antwerp, Belgium)
Tomaz Erjavec (Jozef Stefan Institute, Slovenia)
Maria Gavrilidou (ILSP, Greece)
Walther v. Hahn (University of Hamburg,Germany)
Susane Jekat (ZHAW, Winterthur, Switzerland)
Cvetana Krstev (University of Belgrade, Serbia)
Vladislav Kuboň (Charles University Prague, Czech Republic)
John Nerbone (University of Gröningen, the Netherlands)
Maciej Ogrodniczuk (IPAN, Polish Academy of Sciences)
Petya Osenova (University of Sofia, Bulgaria)
Stelios Piperidis (ILSP, Greece)
Laurent Romary (INRIA, France)
Kiril Simov (Bulgarian Academy of Sciences)
Milena Slavcheva (Bulgarian Academy of Sciences)
Daniel Stein (University of Hamburg, Germany)
Marco Tadić (University of Zagreb, Croatia)
Cristina Vertan (University of Hamburg)
Duško Vitas (University of Belgrade, Serbia)
Kalliopi Zervanou (University of Tilburg, the Netherlands)

# Table of Contents

# Conference Program

**13.09.2013**

**(9:15 - 10:15) Invited Talk**

*Combining, Adapting and Reusing Bi-texts between Related Languages: Application to Statistical Machine Translation (invited talk)*
Preslav Nakov

**Session 1: Machine Translation**

**(10:15 - 10:45)**

*Language diversity and implications for Language technology in the Multilingual Europe*
Cristina Vertan and Walther von Hahn

**(11:15 - 11:45)**

*Corpus development for machine translation between standard and dialectal varieties*
Barry Haddow, Adolfo Hernandez, Friedrich Neubarth and Harald Trost

**(11:45 - 12:15)**

*Adaptation of a Rule-Based Translator to Río de la Plata Spanish*
Ernesto López, Luis Chiruzzo and Dina Wonsever

**Session 2: Language processing**

**(13:30 - 14:00)**

*Text segmentation for Language Identification in Greek Forums*
Pavlina Fragkou

**(14:00 - 14:30)**

*Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources*
Yves Scherrer and Benoît Sagot

**(14:30 - 15:00)**

*The Mysterious Letter J*
Andjelka Zecevic and Stasa Vujicic-Stankovic