INTERNATIONAL WORKSHOP

MULTILINGUAL RESOURCES, TECHNOLOGIES AND EVALUATION FOR CENTRAL AND EASTERN EUROPEAN LANGUAGES

held in conjunction with

The International Conference RANLP - 2009

PROCEEDINGS

Edited by

Cristina Vertan, Stelios Piperidis, Elena Paskaleva
and Milena Slavcheva

Borovets, Bulgaria

17 September 2009

International Workshop

MULTILINGUAL RESOURCES, TECHNOLOGIES AND EVALUATION FOR CENTRAL AND EASTERN EUROPEAN LANGUAGES

PROCEEDINGS

Borovets, Bulgaria
17 September 2009

ISBN 978-954-452-008-3

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Programme Committee

Tomaž Erjavec (Jozef Stefan Institute, Slovenia)

Maria Gavrilidou (ILSP, Greece)**;**

Walther von Hahn (University of Hamburg)

Svetla Koeva (Bulgarian Academy of Sciences)

Cvetana Krstev (University of Belgrad)

Steven Krauwer (University of Utrecht, the Netherlands)

Vladislav Kuboň (Charles University Prague)

Petya Osenova (University of Sofia, Bulgaria)

Elena Paskaleva (Bulgarian Academy of Sciences)

Stelios Piperidis (ILSP, Greece)

Adam Przepiórkowski (IPAN, Polish Academy of Sciences)

Milena Slavcheva (Bulgarian Academy of Sciences)

Marco Tadić (University of Zagreb, Croatia)

Dan Tufiş (Romanian Academy of Sciences)

Cristina Vertan (University of Hamburg)

Duško Vitas (University of Belgrade, Serbia)

Organising Committee

Elena Paskaleva (Bulgarian Academy of Sciences)

Stelios Piperidis (ILSP, Greece)

Milena Slavcheva (Bulgarian Academy of Sciences)

Cristina Vertan (University of Hamburg)

Foreword

The workshop on language processing for Central and Eastern European languages is organised this year for the 4th time in conjunction with the RANLP series of conferences. Looking at the titles of previous editions, one can see that they follow the development which NLP for those languages has faced from one edition of RANLP to the other.

Recent activities in the language technology community in Europe are concerned with the combination/ pipelining of already developed systems and use of very large language resources. This approach assumes that large language resources are available, that systems performances have been evaluated on such resources and that input and output are interoperable with other systems. European initiatives like CLARIN and FLAREET offer the frame for the development of a unified approach for languages all over Europe. For the first time methodologies, evaluation campaigns and roadmaps are planed for all European languages.

Language Processing is now seen as the main technology being able to give people access to information (no matter where it has been produced) in their native languages. Unfortunately, despite important developments, language resources for less popular languages, (especially Balkan and Slavic languages) are still far behind the achieved standard for major western European ones.

As most part of the current Language Technology applications rely on corpus-based methods, one major drawback in the development of language resources and tools for those languages is the lack of training and evaluation data, as well as reference systems for comparing results. Although well-known corpora like JRC-ACQUIS or OPUS are a significant step forward, they

- still do not cover all languages in the Balkan area,
- are collections of documents in specialised languages and therefore decrease the performance of systems trained on those data when testing on another domain.

In order to shorten this bottleneck, it is absolutely necessary to develop, promote and make available all data which can be used for training and evaluation. Additionally, it is important to know which systems have been developed for which applications, on which data have been tested, and what qualitative results have come out.

Therefore the workshop's topic focuses this year on *Multilingual resources, technologies and evaluation for Central and Eastern European languages.*

The selected papers for the current workshop proceedings focus on two issues: adaptation of tools for other languages and multilingual systems and language resources. The eight papers cover ten Central and Eastern European languages.

We would like to thank the authors for contributing to the workshop proceedings and the members of the scientific committee for their quality work. We are grateful to the organisers of RANLP 2009 for hosting this workshop as one of its satellite events. Especially we would like to thank Galia Angelova and Kiril Simov for their great support throughout the whole organisation period.

September 2009

Cristina Vertan, Milena Slavcheva, Stelios Piperidis and Elena Paskaleva

Table of Contents

Bulgarian-Polish-Lithuanian Corpus – Current Development
Ludmila Dimitrova, Violetta Koseska, Danuta Roszko and Roman Roszko 1

On the behavior of Romanian syllables related to minimum effort laws
Anca Dinu and Liviu P. Dinu 9

Using JRC-ACQUIS in SMT Experiments for Romanian and German
Monica Gavrila 14

E-Connecting Balkan Languages
Cvetana Krstev, Ranka Stanković, Duško Vitas and Svetla Koeva 19

Converting Russian Treebank SynTagRus into Praguian PDT Style
David Mareček and Natalia Kljueva 26

A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words
Svetlin Nakov and Elena Paskaleva 32

New Issues and Solutions in Computer-aided Design of MCTI and Distractor Selection for Bulgarian
Ivelina Nikolova 40

Workshop Program

Session 1

9:20-9:30 Welcome and opening remarks

9.30-10.00 *SMT Experiments for Romanian and German Using JRC-ACQUIS*
Monica Gavrila

10.00-10.30 *A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words*
Svetlin Nakov, Elena Paskaleva and Preslav Nakov

10.30 11.00 *E-Connecting Balkan Languages*
Cvetana Krstev, Ranka Stanković, Duško Vitas and Svetla Koeva

Session 2

11.30 12.30 Constantin Orasan (invited talk): Multilingual Information Access Using the QALL-ME Framework: An Example for Romanian

Session 3

14.00-14.30 *Bulgarian-Polish-Lithuanian Corpus – Current Development*
Ludmila Dimitrova, Violetta Koseska, Danuta Roszko and Roman Roszko

14.30-15.00 *On the Behavior of Romanian Syllables Related to Minimum Effort Laws*
Anca Dinu and Liviu P. Dinu

Session 4

15.30-16.00 *Converting Russian Treebank SynTagRus into Praguian PDT Style*
David Mareček and Natalia Kljueva

16.00-16.30 *New Issues and Solutions in Computer-aided Design of MCTI and Distractor Selection for Bulgarian*
Ivelina Nikolova

16.30-17.00 Round table / Closing session