# COLING 2004

## WORKSHOP ON

# Computational Approaches to Arabic Script-based Languages

**University of Geneva**
**Geneva, Switzerland**
**August 28, 2004**

# PREFACE

# WORKSHOP THEME

Recently, there has been a surge of interest in the study of the languages of the Middle East, especially Arabic, Persian (Farsi), Pashto, Kurdish and Urdu. This sudden and urgent interest is manifested by the availability of funding for rapid development of practical systems for processing large volumes of data in these languages. Computational applications for proper name identification, entity recognition, categorization, information retrieval, summarization, machine translation and other implementations are currently in high demand. This comes at a time when advances in formal and computational linguistics over the last fifty years are being consolidated, while work on machine learning and statistical methods has been showing great promise.

There exists a considerable body of work in computational linguistics specifically targeted to these middle eastern languages. Much of the research and development has been the result of initiatives by individual research establishments or industry firms. Furthermore, the usage of the Arabic script gives rise to certain issues that are common to all these languages despite their being of distinct language families. Hence, these languages share properties such as the absence of capitalization, right to left direction, lack of clear word boundaries, complex word structure, a high degree of ambiguity due to non-representation of short vowels in the writing system, and related encoding issues.

## ORGANIZING COMMITTEE

**Ali Farghaly,** SYSTRAN Software, Inc.
**Karine Megerdoomian,** Inxight Software, Inc. and University of California, San Diego

## INVITED SPEAKER

**Martin Kay,** Stanford University

## PROGRAM COMMITTEE

**Jan W. Amtrup,** Bowne Global Solutions
**Tim Buckwalter,** Linguistic Data Consortium
**Miriam Butt,** Konstanz University, Germany
**Violetta Cavalli-Sforza,** Carnegie Mellon University
**Joseph Dichy,** Lyon University
**Abdelkadir Fassi Fehri,** Mohammed V University-Souissi Rabat, Morocco
**Andrew Freeman,** University of Washington
**Nizar Habash,** University of Maryland, College Park
**Masayo Iida,** Inxight Software, Inc
**Simin Karimi,** University of Arizona
**Martin Kay,** Stanford University
**Kevin Knight,** USC/Information Sciences Institute
**Farhad Oroumchian,** University of Wollongong in Dubai
**Ahmed Rafea,** The American University in Cairo
**Jean Senellart,** SYSTRAN Software
**Bonnie Glover Stalls,** University of Southern California
**Rémi Zajac,** SYSTRAN Software

## FURTHER INFORMATIONS

**Emails:**

**Ali Farghaly,** AliFarghaly@aol.com
**Karine Megerdoomian,** karinem@inxight.com

**www:** http://members.cox.net/karinem/COLING2004

# WORKSHOP PROGRAM

OPENING AND OVERVIEW

8:30 – 9:00 *Computer Processing of Arabic Script-based Languages: Current State and Future Directions*
Ali Farghaly

SESSION 1: LEXICON AND CORPORA

9:00 – 9:30 *Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools*
Mohamed Maamouri and Ann Bies

9:30 – 10:00 *Preliminary Lexical Framework for English-Arabic Semantic Resource Construction*
Anne R. Diekema

10:00 – 10:30 *The Architecture of a Standard Arabic Lexical Database: Some Figures, Ratios and Categories from the DIINAR.1 Source Program*
Ramzi Abbès, Joseph Dichy and Mohamed Hassoun

10:30 – 10:45 BREAK

SESSION 2: MORPHOLOGY

10:45 – 11:15 *Systematic Verb Stem Generation for Arabic*
Jim Yaghi and Sane Yagi

11:15 – 11:45 *Issues in Arabic Orthography and Morphology Analysis*
Tim Buckwalter

11:45 – 12:15 *Finite-State Morphological Analysis of Persian*
Karine Megerdoomian

12:15 – 2:00 LUNCH & DEMO SESSIONS

DEMONSTRATIONS

*Urdu Localization Project*
Sarmad Hussain

*FarsiSum – A Persian Text Summarizer*
Martin Hassel and Nima Mazdak

*Stemming the Qur'an*
Naglaa Thabet

*Language Weaver Arabic->English MT*
Daniel Marcu, Alex Fraser, William Wong and Kevin Knight

**INVITED SPEAKER**

2:00 – 2:45 *Arabic Script-Based Languages Deserve to be Studied Linguistically*
Martin Kay

SESSION 3: STATISTICAL APPROACHES

2:45 – 3:15 *An Unsupervised Approach for Bootstrapping Arabic Sense Tagging*
Mona T. Diab

3:15 – 3:45 *Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm*
Mohamed El Kourdi, Amine Bensaid and Tajje-eddine Rachidi

3:45 – 4:00 BREAK

SESSION 4: SPEECH PROCESSING

4:00 – 4:30 *A Transcription Scheme for Languages Employing the Arabic Script Motivated by Speech Processing Applications*
Shadi Ganjavi, Panayiotis G. Georgiou and Shrikanth Narayanan

4:30 – 5:00 *Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition*
Dimitra Vergyri and Katrin Kirchhoff

5:00 – 5:30 *Letter-to-Sound Conversion for Urdu Text-to-Speech System*
Sarmad Hussain

5:30 – 6:00 Discussion and Closing
Ali Farghaly and Karine Megerdoomian

# Contents

## Papers

## Demonstrations

# Index