CALL FOR PAPERS
Workshop on
Language Resources (LRs) and Human Language Technologies (HLT)
for Semitic Languages
Status, Updates, and Prospects

To be held in conjunction with the 7th International Language Resources and Evaluation Conference (LREC 2010)
17 May 2010, Mediterranean Conference Centre, Valetta, Malta
Deadline for submission: 26 February 2010

Description

The Semitic family includes languages and dialects spoken by a large number of native speakers (around 300 million). Prominent members of this family are Arabic (and its varieties), Hebrew, Amharic, Tigrinya, Aramaic, Maltese and Syriac. Their shared ancestry is apparent through pervasive cognate sharing, a rich and productive pattern-based morphology, and similar syntactic constructions. In addition, there are several languages which are used in the same geographic area such as Amazigh or Coptic, which, while not Semitic, have common features with Semitic languages, such as borrowed vocabulary.

The recent surge in computational work for processing Semitic languages, particularly Modern Standard Arabic (MSA) and Modern Hebrew (MH), has brought modest improvements in terms of actual empirical results for various language processing components (e.g., morphological analyzers, parsers, named entity recognizers, audio transcriptions, etc.). Apparently, reusing existing approaches developed for English or French for processing Semitic language text/speech, e.g., Arabic parsing is not as straightforward as initially thought. Apart from the limited availability of suitable language resources, there is increasing evidence that Semitic languages demand modeling approaches and annotations that deviate from those found suitable for English/French. Issues such as the pattern-based morphology, the frequently head-initial syntactic structure, the importance of the interface between morphology and syntax, and the difference between spoken and written forms (especially in Colloquial Arabic(s)) exemplify the kind of challenges that may arise when processing Semitic languages. For language technologies, such as information retrieval and machine translation, these challenges are compounded by sparse data and often result in poorer performance than for other languages.

This Workshop intends to follow on topics of paramount importance for Semitic-language NLP that were discussed at previous events (LREC, MEDAR/NEMLAR Conferences, the workshops of the ACL Special Interest Group for Semitic languages, etc.) and which are worth revisiting.

The workshop will bring together people who are actively involved in Semitic language processing in a mono- or cross/multilingual context, and give them an opportunity to update the community through reports on completed or ongoing work as well as on the availability of LRs, evaluation protocols and campaigns, products and core technologies (in particular open source ones). We also invite authors to address other languages spoken in the Semitic language area (languages such as Amazigh, Coptic, etc.). This should enable participants to develop a common view on where we stand and to foster the discussion of the future of this research area. Particular attention will be paid to activities involving technologies such as Machine Translation and Cross-Lingual

Information Retrieval/Extraction, Summarization, etc. Evaluation methodologies and resources for evaluation of HLT will be also a main focus.

We expect to elaborate on the HLT state of the art, identify problems of common interest, and debate on a potential roadmap for the Semitic languages. Issues related to sharing of resources, tools, standards, sharing and dissemination of information and expertise, adoption of current best practices, setting up joint projects and technology transfer mechanisms will be an important part of the workshop.

Topics of Interest

This full-day workshop is not intended to be a mini-conference, but as a real workshop aiming at concrete results that should clarify the situation of Semitic languages with respect to Language Resources and Evaluation. We expect to launch at least two evaluation campaigns: Comparative evaluation of Morphology taggers and Named Entities Recognizers.

Among the many issues to be addressed, below follow a few suggestions:

- Issues in the design, the acquisition, creation, management, access, distribution, use of Language Resources, in particular in a bilingual/multilingual setting (Standard Arabic, Hebrew, Colloquial Arabic, Amazigh, Coptic, Maltese, etc.)

- Impact on LR collections/processing and NLP of the crucial issues related to "code switching" between different dialects and languages

- Specific issues related to the above-mentioned languages such as the role of morphology, named entities, corpus alignment, etc.

- Multilinguality issues including relationship between Colloquial and Standard Arabic

- Exploitation of LR in different types of applications

- Industrial LR requirements and community's response

- Benchmarking of systems and products; resources for benchmarking and evaluation for written and spoken language processing;

- Focus on some key technologies such as MT (all approaches e.g. Statistical, Example-Based, etc.), Information Retrieval, Speech Recognition, Spoken Documents Retrieval, CLIR, Question-Answering, Summarization, etc.

- Local, regional, and international activities and projects and needs, possibilities, forms, initiatives of/for regional and international cooperation.

We invite submissions on computational approaches to processing text/speech in all Semitic and Semitic-area languages. The call is open for all kinds of computational work, e.g., work on computational linguistic processing components (e.g., analyzers, taggers, parsers), on state-of-the-art NLP applications and systems, on leveraging resource and tool creation for the Semitic language family, and on using computational tools to gain new linguistic insight. We especially

welcome submissions on work that crosses individual language boundaries, heightens awareness amongst Semitic-language researchers of shared challenges and breakthroughs, and highlights issues and solutions common to any subset of the Semitic languages family.

Workshop general chair:
Khalid Choukri, ELRA/ELDA, Paris, France

Workshop co-chairs:
Owen Rambow, Columbia University, New York, USA
Bente Maegaard , University of Copenhagen, Denmark
Ibrahim A. Al-Kharashi, Computer and Electronics Research Institute, King Abdulaziz City for Science and Technology, Saudi Arabia

Organizing Committee information

The Organizing, Program, and the Scientific Committees will be listed on the web pages.

Important Dates

Deadline for abstract submissions:	26 February 2010
Notification of acceptance:	15 March 2010
Final version of accepted paper:	11 April 2010
Workshop full-day:	17 May 2010

Submission Details

Submissions should comply with LREC standards (including the LREC Map initiative) and must be in English. Abstracts for workshop contributions should not exceed Four A4 pages (excluding references). An additional title page should state: the title; author(s); affiliation(s); and contact author's e-mail address, as well as postal address, telephone and fax numbers.

Submission will use the LREC START facility. Expected deadline is 26 February 2010.

Submitted papers will be judged based on relevance to the workshop aims, as well as the novelty of the idea, technical quality, clarity of presentation, and expected impact on future research within the area of focus.

Registration to LREC'2010 will be required for participation, so potential participants are invited to refer to the main conference website for all details not covered in the present call (http://www.lrec-conf.org/lrec2010/)

Formatting instructions for the final full version of papers will be sent to authors after notification of acceptance and will be identical to LREC main conference instructions.

When submitting a paper through the START page, authors will be kindly asked to provide relevant information about the resources that have been used for the work described in their paper or that are the outcome of their research. For further information on this new initiative, please refer to http://www.lrec-conf.org/lrec2010/?LREC2010-Map-of-Language-Resources.