LINGUISTICA COMMUNICATIO: Vol.15, n° 1-2, 2013, p. 5-8

Introduction to the present issue

Joseph Dichy
Professor of Arabic Linguistics,
Université Lumière-Lyon 2 & ICAR lab (UMR 5191-CNRS/Lyon 2),
Lyon, France
joseph.dichy@yahoo.fr

ORIGINAL CALL FOR CONTRIBUTIONS

The *International Conference on Machine Intelligence* (ICMI'05) held in Tozeur, Tunisia, between November 5 and November 7, 2005 included a session on *Linguistic Information Integration in Arabic Character and Text Recognition*. The session, which was coordinated by Dr. Slim Kanoun and myself, aimed at "bringing together specialists in Arabic OCR and Text recognition and specialists in Natural language processing (NLP), with special reference to Arabic, in either monolingual or multilingual perspective". The idea was "to discuss the integration of linguistic information in linguistic-based and statistic-based approaches of Arabic printed and handwritten documents recognition, with special reference to various types of languages resources (LR-s), and to LR based analyzers", in order to contribute to the development of "robust systems for the recognition, indexation and thematic classification of printed or handwritten Arabic texts".

The building of such systems is "of momentous interest for Arabic countries and organisations, considering the significant production of printed and handwritten documents in Arabic witnessed daily in the whole world. Robust OCR systems should also be of great help in the indexing of Arabic historical documents and manuscripts in the context of the conservation and protection of the Arabic culture and civilisation inheritance".

The call for papers also mentioned the following areas of current interest:

1. Arabic character, word and text recognition.
2. Analysis of vowel-free Arabic texts.
3. Arabic language resources (electronic dictionaries, tree-banks, contextual analysis resources, etc.).
4. Robust OCR techniques (printed and handwritten documents).
5. Word and text images digitalization, compression and indexation techniques.

VOLUME CONTENTS

So much for what Dr. Slim Kanoun and myself sought. The session – as could have been expected – only partly answered to the above initial purpose, but was a good go at reflecting the state of the art in the field. Contributions reached us from many parts of the world, and around 50% were retained. The present volume includes the participation of 22 authors, coming, respectively from France (10), Jordan (5), Algeria (2), China (2), Tunisia (1), Belgium (1) and Iraq (1). The international Review Committee, which the coordinators would like to thank warmly for their help in making that session a success, included reviewers from France, Morocco, The Netherlands, Palestine, Tunisia, Great Britain and the U.S.A. Their names are recalled below.

Noticeably enough; the contents of the session currently remain of the same momentum, even a few years after the session took place. The present volume, significantly enough, reflects the actual contents of the Tozeur session: the call for paper focused, at least partly, on character recognition, whereas the title of the volume positions "text recognition" in the first place. Let us, first, summarise the contents of the session and second, try and interpret it.

This special issue of *Linguistica Comunicatio* – التواصل اللساني, divides into four parts. **Part 1** includes the key addresses of the organisers. *Slim Kanoun*'s (Tunisia, Sfax) is concerned with the optical recognition of Arabic printed texts, and focuses on the crucial role of morphological analysis in this respect. *Joseph Dichy* (France, Lyon) considers the assessment of Arabic NLP software that includes a recognition process, and casts light on the relations between (a) language resources (LR-s) dividing into lexical LR-s on the one hand, and other LR-s, such as corpora, tree-banks…, (b) basic tools, i.e., essentially, word-level and sentence-level processors, and (c) the related application software.

Part 2 accordingly deals with "Lexical and textual resources: lexica and corpora". It opens with a set of "Proposals for a normalised representation of Standard Arabic full form lexica", due to *Suzanne Salmon-Alt, Amine Akrout and Roland Romary* (France, Nancy). The standardisation of lexical resources has become necessary, not only for inner coherence reasons, but also for consistency with other languages in today's multilingual context. The *Lexical Mark-up Framework* (LMF), with minor accommodations, is proposed by the authors for Arabic lexica.

Mark Van Mol (Belgium, Leuven) refers to the experience of the Dutch Language Union in compiling corpus-based Dutch-Arabic/Arabic-Dutch bilingual dictionaries, which were entrusted to the Radboud University of Nijmegen. The main endeavour presented in this paper is that of the contribution to this experience of the Catholic University of Leuven. It introduces the use of the lexical database developed by the author for the semi-automatic pre-tagging of a corpus of over 10 million Arabic words.

Joseph Dichy, *Mohamed Hassoun* and *Ramzi Abbès,* (France, Lyon) refer to the DIINAR.1 lexical resource (http://diinar.univ-lyon2.fr), the contents of which are analysed with respect to morpho-lexical ambiguities in Arabic. Many ambiguities are not solved by adding secondary diacritics, known as 'vowel-signs', to word-forms, but actually stem form the morphological structures of the language. Statistical evidence is presented.

Part 3 is entitled "Arabic Text Processing and Information Retrieval". *Mabrouka El Hachani* and *Mohamed Hassoun* (France, Lyon) first present a paper on "Knowledge management and interlingua equivalence in the indexing process", which focuses on the cognitive process involved in the indexing of multilingual information and the identification of key concepts in order to represent or summarize multilingual documents, and, ultimately, to build a multilingual thesaurus.

Nasredine Semmar, Faïza Elkateb-Gara and *Christian Fluhr* (France, Paris) present the experience of the French Laboratory of Multilingual and Multimedia Knowledge Engineering (LIC2M) in the devising of a cross-language information retrieval system designed to work on Arabic, Chinese, English, French, German and Spanish. Information retrieval in Arabic involves analysing unvowelled Arabic writing, i.e. an analyzer drawing on a lexical resource.

Amjad M. Daoud (Jordan) presents a text compression algorithm based on the morphological structure of Arabic, and on an affix analysis that takes advantage of statistical studies of Arabic morphological features, such as the top 20 most frequent n-grams. Roots and affixes dictionaries have been built using a corpus derived from diacritical Arabic school texts.

Bassam Haddad, Mustafa Yaseen and *Mamoun Hattab*'s contribution (Jordan) presents a research project on Arabic non-words detection and correction. The method is both linguistic and statistics-based. Morphological (including root-pattern relationships) knowledge and morpho-syntactical relations are complemented by original probabilistic measures: Root-Pattern and Pattern-Root Predictive Values (RPV and PRV). Keyboard and letter-sound effects are also considered.

Part 4 finally comes to character recognition and OCR. *Mohammed Zeki Khedher* and *Ghayda Al-Talib* concentrate on a very difficult task: the recognition of secondary signs in handwritten Arabic. The model they propose uses Fuzzy Logic.

REVIEW COMMITTEE

The Committee included, in addition to the Session organizers, Joseph Dichy and Slim Kanoun:

Pr Mehdi Arar – Birzeit University – Head of the Department of Languages and Translation – marar@birzeit.edu

Dr Sami Boudelaa – Post-doctoral Researcher, Medical Research Council, Cognition and Brain Sciences Unit – University of Cambridge, U.K. – sami.boudelaa@mrc-cbu.cam.ac.umk

Pr Abdelfattah Braham – Département de Lettres arabes de l'Université de La Manouba, Directeur de l'équipe de recherche sur Corpus et traitement automatique de l'arabe – Tunis – abdelfat.braham@flm.rnu.tn

Dr Violetta Cavalli-Sforza – Visiting Researcher, Language Technologies Institute – Carnegie Mellon University, Pittsburgh, U.S.A. – violetta@cs.cmu.edu / violettacavalli@yahoo.com

Dr Everhard Ditters - TCMO, Radboud Universiteit Nijmegen, The Netherlands – e.ditters@let.ru.nl

Pr Salem Ghazali – Institut Supérieur des Langues de Tunis, Université du 7 novembre-Carthage, and IT.COM (Information Technologies and Communication, previously: IRSIT) – Tunis – salem_ghazali@yahoo.com

Dr Malek Ghenima, Directeur de l'École Supérieure de Commerce électronique, La Manouba, Tunis – malek.ghenima@esct.rnu.tn

Dr Nizar Habash – Post-doctoral Researcher, Center for Computational Learning Systems – Columbia University – habash@cs.columbia.edu

Pr Mohamed Hassoun, ENSSIB, Villeurbanne (near Lyon) – France – hassoun@enssib.fr

Dr Lamia Labed, Institut Supérieur de Gestion, Département Informatique, Tunis – lamia.labed@isg.rnu.tn

Dr Abdelhadi Soudi - Center for Computational Linguistics & Center for Languages and Communication, École Nationale de l'Industrie Minérale – Rabat – asoudi@enim.ac.ma