MEDAR

Mediterranean Arabic Language and Speech Technology

Summary of the General discussion of the Workshop on *LR and HLT for Semitic Languages* at LREC 2010, Malta

Author: Bente Maegaard, University of Copenhagen

June 2010

Summary of the General discussion of the Workshop on *LR and HLT for Semitic Languages* at LREC 2010, Malta

Bente Maegaard, June 29, 2010

The MEDAR project participated in the organisation of the workshop on Language Resources and Human Language Technologies for Semitic languages, - MEDAR being represented by ELDA, France, and University of Copenhagen, Denmark, and the other institutions being Columbia University, USA, and King Abdulaziz City for Science and Technology, Saudi Arabia.

The workshop programme is attached at the end of this report. As can be seen, the programme features many different Semitic languages, and the participation is equally well distributed over the world. The proceedings provide further details.

This short report focuses on the last programme item, the *General Discussion: Cooperation Roadmap for building sustainable Human Language Technologies for the Arabic language within and outside the Arabic world.* The purpose of the discussion was to determine some of the common interests among the participants which could serve as a basis for future collaboration. MEDAR has suggested a cooperation roadmap for Arabic LR and HLT – a short presentation was made - and other roadmaps exist as well: Inspiration for collaboration can be taken in any of these, - and new ideas can be added.

The discussion is presented below.

One of the suggestions of the MEDAR cooperation roadmap is to use language technology to fight illiteracy. This point gave rise to a lively discussion, as one of the participants has been doing that most of his life, and sees this as one of the most important things we can do. He has a long experience working with UNESCO, World Bank etc, on ICT-enabled teaching and learning of Arabic. There is still a gap between what we as researchers see as possibilities, and the understanding of a teacher or a funder – they do not readily see the use of ICT in teaching.
So, instead of trying to get funding for IT tools for teaching illiterates, it may be better to work on tools and methods which can facilitate the reading of Arabic. One idea which had been followed was that the lack of diacritics makes the reading of Arabic more difficult, so an automatic diacritizer was an obvious tool to work on, etc. – I.e. there is strong support for the idea of fighting illiteracy, but the suggestion is to take small steps and suggest something that both teachers and funders can easily understand.
Keywords: tools for reading support, fighting illiteracy.

Another suggestion in the same field of fighting illiteracy was to use the dialects instead of teaching MSA which is far away from the pupils' everyday language. This would lead to a codification of the dialects and there was a general agreement that it would be easier for pupils to learn to read. However, this is a highly political question as many see such a development as a threat to MSA and one unique Arabic (written) language.
Keyword: teach dialects

There was a discussion about the fact that a BLARK for Arabic and any other Semitic language is needed in order to make real progress and avoid doing the same thing over and over again, and that

we need to share resources and tools, and this also requires that they follow standards so that the output from a morphological analyzer can feed into a syntactic analyzer. Standards for input and output should be agreed and we have gone part of the way, but not all, and now the field ought to be sufficiently mature to decide on standards - somebody asked for dictatorship: Maybe the network (see below) can have an important role here. Another point brought up was the fact that sharing resources and tools is not so easy, as LDC and ELRA sell resources at prices which are sometimes prohibitive. There was a discussion about this and it was mentioned that not all LR and tools distributed by agencies are equally expensive, some are even free. However, there was a proposal that it would be good to set up a repository for free language resources and tools, this could be done by NEMLAR (MEDAR) or the ACL SIG for Semitic languages.
Keywords: standards, repositories for free resources, sharing and collaboration.

The discussion then focussed on the fact that there are many morphological analyzers, and it seems that everyone wants to make yet another morphological analyzer, while there are many unsolved problems in the field. One of the answers to this is sharing and standardisation, see above. An approach towards this could be evaluation of morphological analyzers to determine which methods are best etc.
The SIG for Semitic languages has considered organizing such an evaluation but given up as the task is too heavy. It turned out that ALECSO[1] had already organized an evaluation of morphological analyzers, so it was suggested that IE (Information Extraction) could rather be the next field to evaluate. It was suggested that maybe ELRA/LDC/SIG/NEMLAR could prepare something in this field for next LREC.
Keywords: evaluation, IE

It was also suggested that the SIG and NEMLAR join forces so that everyone working on LR and HLT for Semitic languages would be working under the same umbrella. This could give easier and broader cooperation possibilities.
Keywords: networks

Finally Bente Maegaard thanked all participants for the discussion and the suggestions which will be seriously taken into account. Some participants signed up for contributing to further discussion and possible future cooperation.

[1] The Arab League Educational, Cultural and Scientific Organization (ALECSO) and King Abdul-Aziz City of Technology (KACT) initiative on morphological analyzers of Arabic text aims to encourage research on developing an open source morphological analyzer for Arabic text of high accuracy, easy to develop, which can be integrated into higher levels of applications for processing Arabic text. ALECSO and KACT with the cooperation with the Arabic Language Academy (Damascus), organized the workshop of the experts of morphological analyzers for Arabic text. The workshop held at the Arabic Language Academy in Damascus, Syria, 26th-28th April 2009. Participants from Algeria, Czech Republic, Egypt, France, Jordan, Morocco, Saudi Arabia, Syria, Tunis, United Kingdom and United States of America, presented their morphological analyzers and evaluation methods. (personal communication from Salwa Hamada)

Programme

9:15-9:30 Welcome and Introduction
Khalid Choukri, Owen Rambow, Bente Maegaard, and Ibrahim A. Al-Kharashi

Oral Session 1: Syntax, Semantics, and Parsing

9:30-9:50 Structures and Procedures in Arabic Language
André Jaccarini (1), Christian Gaubert (2), Claude Audebert (1),
(1) Maison méditerranéenne des sciences de l'homme (MMSH), France
(2) Institut français d'archéologie orientale du Caire (IFAO), Cairo, Egypt

9:50-10:10 Developing and Evaluating an Arabic Statistical Parser
Ibrahim Zaghloul (1) and Ahmed Rafea (2)
(1) Central Lab for Agricultural Expert Systems, Agricultural Research Center, Ministry of Agriculture and Land Reclamation.
(2) Computer Science and Engineering Dept., American University in Cairo

10:10-10:30 A Dependency Grammar for Amharic
Michael Gasser, Indiana University, USA

10:30-11:00 Coffee break

11:00-12:20 Poster Session 1: Morphology & NLP Applications I

A syllable-based approach to Semitic verbal morphology
Lynne Cahill, University of Brighton, United Kingdom

Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet
Lahsen Abouenour (1), Karim Bouzoubaa (1) and Paolo Rosso (2)
(1) Mohammadia School of Engineers, Med V University Rabat, Morocco
(2) Natural Language Engineering Lab. - ELiRF, Universidad Politécnica Valencia, Spain

Light Morphology Processing for Amazighe Language
Fadoua Ataa Allah and Siham Boulaknadel, CEISIC, Royal Institut for Amazigh Culture, Madinat Al Irfane, Rabat, Morocco

Using Mechanical Turk to Create a Corpus of Arabic Summaries
Mahmoud EL-Haj, Udo Kruschwitz and Chris Fox, School of Computer Science and Electronic Engineering, University of Essex, United Kingdom

DefArabicQA: Arabic Definition Question Answering System
Omar Trigui (1), Lamia Hadrich Belguith (1) and Paolo Rosso (2)
(1) ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia
(2) Natural Language Engineering Lab. – EliRF, Universidad Politécnica Valencia, Spain

12:20-13:50 Lunch break

13:50-15:10 Poster Session 2: Morphology & NLP Applications and NLP Tools

Techniques for Arabic Morphological Detokenization and Orthographic Denormalization
Ahmed El Kholy and Nizar Habash, Center for Computational Learning Systems, Columbia University, USA

Tagging Amazigh with AncoraPipe
Mohamed Outahajala (1), Lahbib Zenkouar (2), Paolo Rosso (3) and Antònia Martí (4)
(1) IRCAM,
(2) Mohammadia School of Engineers, Med V University Rabat, Morocco,
(3) Natural Language Engineering Lab. - ELiRF, Universidad Politécnica Valencia, Spain,
(4) CLiC - Centre de Llenguatge i Computació, Universitat de Barcelona, Barcelona, Spain

Verb Morphology of Hebrew and Maltese - Towards an Open Source Type Theoretical Resource Grammar in GF
Dana Dannélls (1) and John J. Camilleri (2)
(1) Department of Swedish Language, University of Gothenburg, Sweden;
(2) Department of Intelligent Computer Systems, University of Malta, Malta

Syllable Based Transcription of English Words into Perso-Arabic Writing System
Jalal Maleki, Dept. of Computer and Information Science, Linkping University, Sweden

COLABA: Arabic Dialect Annotation and Processing
Mona Diab, Nizar Habash, Owen Rambow, Mohamed Al Tantawy and Yassine Benajiba, Center for Computational Learning Systems, Columbia University, USA
A Linguistic Search Tool for Semitic Languages
Alon Itai, Knowledge Center for Processing Hebrew, Computer Science Department, Technion, Israel

13:50-15:10 Poster Session 3: Speech & Related resources

Algerian Arabic Speech database Project (ALGASD): Description and Research Applications
Ghania Droua-Hamdani (1), Sid Ahmed Selouani (2) and Malika Boudraa (3)
(1) Speech Processing Laboratory (TAP), CRSTDLA, Algiers, Algeria;
(2) LARIHS Laboratory, University of Moncton, Canada;
(3) Speech Communication Laboratory, USTHB, Algiers, Algeria.

Integrating Annotated Spoken Maltese Data into Corpora of Written Maltese
Alexandra Vella (1,2), Flavia Chetcuti (1), Sarah Grech (1) and Michael Spagnol (3)
(1) University of Malta, Malta
(2) University of Cologne, Germany
(3) University of Konstanz , German

A Web Application for Dialectal Arabic Text Annotation
Yassine Benajiba and Mona Diab, Center for Computational Learning Systems, Columbia University, USA

Towards a Psycholinguistic Database for Modern Standard Arabic
Sami Boudelaa and William David Marslen-Wilson, MRC-Cognition & Brain Sciences Unit, Cambridge, United Kingdom

Oral Session 2 : Resources and tools for Machine Translation

15:10-15:30 Creating Arabic-English Parallel Word-Aligned Treebank Corpora
Stephen Grimes, Xuansong Li, Ann Bies, Seth Kulick, Xiaoyi Ma and Stephanie Strassel, Linguistic Data Consortium, USA

15:30-15:50 Using English as a Pivot Language to Enhance Danish-Arabic Statistical Machine Translation
Mossab Al-Hunaity, Bente Maegaard and Dorte Hansen, Center for Language Technology , University of Copenhagen, Denmark

15:50-16:10 Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons
Nasredine Semmar, CEA LIST, France

16:10-16:30 Coffee break

16:30-17:20 General Discussion
Cooperation Roadmap for building a sustainable Human Language Technologies for the Arabic language within and outside the Arabic world.
Chair: Bente Maegaard

17:20-17:30 Concluding remarks and Closing

List of registered participants

First name	Last name	Country	Institution
Mossab	Al-Hunaity	Denmark	University of Copenhagen, CST
Amal	Al-Saif	United Kingdom	University of Leeds
Mohamed	Altantawy	United States	Columbia University
María Jesús	Aranzabe	Spain	University of the Basque Count
Fadoua	Ataa Allah	Morocco	IRCAM
Yassine	Benajiba	United States	CCLS, Columbia University
Ann	Bies	United States	Linguistic Data Consortium
Sami	Boudelaa	United Kingdom	MRC-CBU
Stephen	Boxwell	United States	The Ohio State University
Lynne	Cahill	United Kingdom	University of Brighton
John J.	Camilleri	Malta	University of Malta
Khalid	Choukri	France	ELRA/ELDA
Angelo	Dalli	Malta	Malta U
Ahmed	El Kholy	United States	Columbia University
Mahmoud	El-Haj	United Kingdom	University of Essex
Ray	Fabri	Malta	University of Malta
Paul	Felt	United States	Brigham Young University
Michael	Gasser	United States	Indiana University
Albert	Gatt	Malta	University of Malta
Sarah	Grech	Malta	MaltaU
Stephen	Grimes	Bahrain	LDC
Jan	Hajic	Czech Republic	Charles University in Prague
Dorte	Haltrup Hansen	Denmark	University of Copenhagen, CST
Hamdy	Hussein	Egypt	Sakhr Software Company
alon	itai	Israel	technion
Don	Killian	Finland	University of Helsinki
Mariama	Laib	France	CEA LIST
Veronika	Lux	France	CNRS
Bente	Maegaard	Denmark	University of Copenhagen
Jalal	Maleki	Sweden	Linkoping University
Joseph	MarianiI	France	LIMSI-CNRS & IMMI
Michael	Maxwell	United States	University of Maryland
Abdul-Baquee	Muhammad	United Kingdom	University of Leeds
Mohamed	Maamouri	United States	Linguistic Data Consortium
Mohamed	Outahajala	Morocco	IRCAM
Stephanie	Poisson	United States	U.S. Department of Defense
Owen	Rambow	United States	Columbia U -- CCLS
Abbes	Ramzi	France	ICAR

Eric	Ringger	United States	Brigham Young University
Michael	Rosner	Malta	University of Malta
Paolo	Rosso	Spain	Valencia U
C. Anton	Rytting	United States	University of Maryland
Fatiha	Sadat	Canada	UQAM
Majdi	Sawalha	United Kingdom	University of Leeds
Judith	Schlesinger	United States	IDA/CCS
Nasredine	Semmar	France	CEA LIST
Michael	Spagnol	Germany	University of Konstanz
Richard	Sproat	United States	OHSU
Gregor	Thurmair	Germany	Linguatec
Lamia	Tounsi	Ireland	Dublin City University
Omar	Trigui	Tunisia	MIRACL laboratory
Martin	Volk	Switzerland	University of Zurich
Shuly	Wintner	Israel	University of Haifa
Ait ouguengay	Youssef	Morocco	IRCAM