

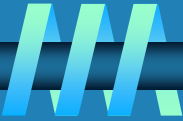
Arabic Speech & Language processing : Tools & Resources

**C. Mokbel, W. Karam,
R. Bayeh, H. Greige**

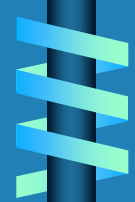
University of Balamand

**Expert Group Meeting on Promoting the Digital Arabic Content
in the ESCWA Region
29-30 April 2008**

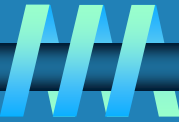
Layout ●



- ∩ **Introduction**
- ∩ **International Context**
- ∩ **Regional Context**
- ∩ **The Balamand Experience**
- ∩ **Recommendations**

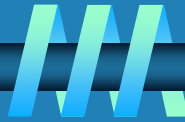


Introduction



- ❧ **The digital Arabic content is present in different forms**
 - Electronic books and articles
 - HTML documents over the web
 - Audio and audio-visual documents (e.g. broadcast news)
 - Videos and films
 - Scanned images
- ❧ **“Human Language Technology” (HLT) is a major mean for ease of access to information**
 - Reduce digital divide
 - Towards knowledge based society
 - Economic and Social development
- ☞ **Develop Speech and Language Processing in the region**

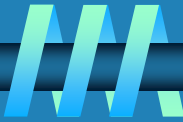




Ω HLT cover:

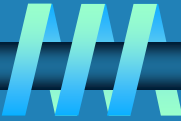
- **Language technologies**
 - Morphological analyzers
 - Taggers
 - Automatic diacritizing
 - MT and SMT
 - Indexing and retrieval
 - Summarization
 - ...
- **Speech technologies**
 - Speech synthesis
 - Automatic Speech recognition
 - Speaker Recognition
 - ...
- **Handwritten recognition technologies**
- **Audio-visual technologies**
 - Video indexing and retrieval
 - Access control (Biometrics)
 - ...

International Context



- Ω **During the last decade, high interest in Arabic Speech and Language Processing:**
- **Originally language technologies:**
 - **France:** CNRS, CEA, Université de Lyon
 - **Germany:** IFN
 - **USA:** University of Pennsylvania, Stanford, Xerox
 - **Czech Republic:** Charles University in Prague
 - **UN:** UNDL
 - ...
 - **Later on, more effort on speech technologies:**
 - **John Hopkins Summer School 2002 for Arabic Speech and Language processing**
 - **USA:** BBN, IBM, CMU, JHU, Microsoft
 - **France:** LIMSI, INRIA/LORIA
 - **UK:** CUED
 - **Belgium:** Babel
 - ...

International Context



∩ Several important projects:

- **Oriental**
 - Industry-oriented project
 - Large speechDat-like databases have been collected covering the different dialects of the region
- **GALE**
 - DARPA project
 - Speech to Speech, Written Text to Text
 - Target language: English

∩ Several competitions:

- **NIST**
- **GALE**

∩ Speech and Language Resources

- **ELRA/ELDA**
- **LDC**

Regional Context



∞ Theoretical Research on Speech and Language

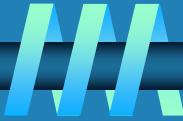
- Speech production: DRM models in Syria
- Phonetic structures: Different parts of the Arab region
- Prosody

∞ Text processing

- Taggers, Morphological analyzers, etc.
 - Amman University, HIAST, Sakhr, RDI, IBM Egypt

∞ Machine Translation

- Rule based
 - Amman University, RDI, Sakhr
- SMT
 - IBM Egypt
- UNDL
 - Jordan



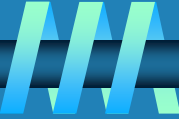
Ω Speech Processing

- Text to Speech
 - Mohammed V Soussi University, Rabat, IBM Egypt, Sakhr
- Speech Recognition
 - University of Balamand, IBM Egypt, Sakhr, FTRD Egypt
- Speaker Recognition
 - University of Balamand

Ω Handwritten recognition

- University of Balamand, Sakhr

Regional Context - Associations

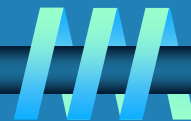


- ❧ Egypt: The Egyptian Society of Language Engineering (ESLE)
- ❧ Syria: The Arabic Language Association مجمع اللغة العربية, Syrian Computer Society
- ❧ Morocco: Arabic Language Institute in Fez (ALIF)

Regional Context – Universities and Institutions

Institution	Country	Domains of Interest
Amman University	Jordan	Translation, Morphological analyzer, Tagger,
Arabic Language Institute in Fez	Morocco	Resources
Hariri-Canadian University	Lebanon	POS tagging
Institut d'Etudes et de Recherche pour l'Arabisation	Morocco	Development of Arabic Language
Royal Scientific Society	Jordan	Translation
Speech Center, King Abdulaziz City for Science and Technology	Saudi Arabia	Resources
University of Balamand	Lebanon	Speech Recognition, Speaker Recognition
University of Mohammed V Soussi – ENSIAS	Morocco	Speech Synthesis
HIAST	Syria	Arabic Speech synthesis, Analysis and Recognition. Emotion Recognition Text tagging
Damascus University	Syria	Natural Language Processing, Speech and Text

Regional Context – NEMLAR



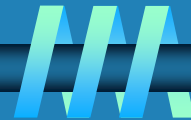
- Ω Network for Euro-Mediterranean Language Resources
 - <http://www.nemlar.org>
- Ω At least one leading LR actor in each country in the network (from research institutes in Morocco, Tunisia, Egypt, Lebanon, Jordan, ...)
- Ω Partnership with recognized European centers of excellence in Arabic and other indigenous speech and text processing
- Ω A ‘map’ of Euro-Mediterranean stakeholders, national and cross-border projects, and existing language resources and processing tools addressing the existing linguistic diversity in the region
- Ω Key strengths, weaknesses, opportunities and threats to the development of Arabic and other language resources in the region and establish a set of key priorities for developing LRs
- Ω **BLARK: Basic Language Resource Kit**
- Ω **Language Resources**
 - Text
 - Speech (Broadcast News)

Regional Context – MEDAR (1/3)



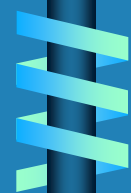
- ❧ **MEDiterranean ARabic Language and Speech Technology**
 - <http://www.medar.info>
- ❧ **An ICT (FP7) project – Objective 9.1 International Cooperation**
- ❧ **Start date: February 2008 for 30 months**
- ❧ **MEDAR is structured around 3 pillars, 4 main objectives, and a number of instruments**
- ❧ **The 3 pillars are:**
 - **Producing a knowledge base on Human Language Technology (HLT) players, existing language resources (LRs) and processing tools, activities and products for Arabic**
 - **Designing a strong cooperation roadmap between the EU and Arabic countries, within the Arabic countries, and between academia and industry**
 - **Focusing on Machine Translation (MT) and Multilingual Information Retrieval (MLIR) for which required technology components, LR, and benchmarking methodologies will be identified.**

Regional Context – MEDAR (2/3)

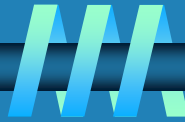


∞ The 4 objectives consist in

- Consolidating the NEMLAR network of players in all areas of HLT
- Developing the Cooperation Roadmap based on a clear picture of the foreseeable technological trends, market potentials, and cooperation possibilities
- Updating the Basic Language Resource Kit: the minimum set of resources and tools necessary for carrying out research and training on LRs and HLT, with a focus on MT and MLIR
- Supporting the development of tools and resources, in particular MT and MLIR on the basis of partners technologies and open source code (e.g. Statistical MT, MLIR, and speech recognition) and the framework for their benchmarking.



Regional Context – MEDAR (3/3)



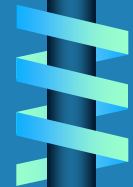
Ω The Consortium:

- University of Copenhagen (Coordinator)
- ELDA S.A
- University of Balamand
- Amman University
- Universiteit Utrecht
- Institute for Language and Speech Processing-Athena Research Center
- The Engineering Company for Digital Systems Development
- Birzeit University
- Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes- ENSIAS
- Commissariat à l'énergie atomique- CEA
- Centre National de la Recherche Scientifique – CNRS
- The Open University
- Université Lumière-Lyon 2
- IBM Egypt
- Sakhr Software Co.

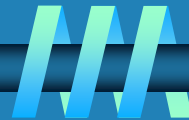
Regional Context – ALMA



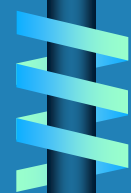
- ∩ **Arabic Language Multilingual Application**
- ∩ **Directed towards Machine Translation**
- ∩ **Supported by EC**



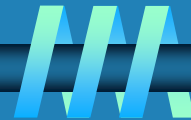
Regional Context – Workshops 2007



- Ω "Arabic Language in the Information Age", 5th annual conference of Arabic Academy in Damascus Nov 20-22 2006.
- Ω "Natural Language Processing in Arabic", IEEE 2nd Information and Communication Technologies International Symposium (ICTIS'07), Fez, Morocco, April 3-5, 2007.
- Ω "The 1st International Workshop on Natural Language Processing Using the Universal Networking Language (UNL)," Bibliotheca Alexandrina, Alexandria, Egypt, May 4-7, 2007.
- Ω "Arabic Electronic Dictionary in Arabic Academy," Damascus, June 11-13 2007.
- Ω "International Colloquium on Arabic Language Processing" (CITALA 2007), Rabat, Morocco, June 18-19, 2007.
- Ω Seventh Conference on Language Engineering (SOLE'07), Cairo, Egypt, 05-06 December 2007.



Regional Context – ISCA-WANA (1/2)



Ω ISCA-WANA Subcommittee

(International Speech Communication Association - West Asia and North Africa)

- <http://www.isca-speech.org/>

Ω ISCA is a non-profit association aiming "to promote Speech Communication Science and Technology, both in the industrial and Academic areas"

Ω WANA regional subcommittee established in Spring 2006 to promote ISCA and more particularly Research and Development in Speech Communication in the region

- Dominancy of the Arabic language
- Significant number of different Arabic dialects
- Speech communication research work is being conducted in the region using the resources available for the various languages and dialects
- The WANA subcommittee aims to reinforce communication within the speech scientific and technical community in the region and their interaction with the international community

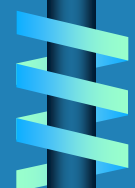


Regional Context – ISCA-WANA (2/2)



Ω WANA Subcommittee's members:

- Dr. Oumayma AL-Dakkak (Syria)
- Dr. Ossama Emam (Egypt)
- Dr. Chafic Mokbel (Lebanon)
- Dr. Mohammad Mrayati (Saudi Arabia)
- Dr. Mustafa Yasseen (Jordan)



The Balamand Experience - HLT



∩ Arabic Speech Recognition

- Voice Commands
- Broadcast News Transcription

∩ Arabic Language Modeling

- Morphological based Language Modeling

∩ Speaker Recognition

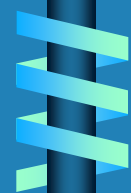
- AudioVisual
- Participation to NIST, BioSecure competitions

∩ Arabic Handwritten Recognition

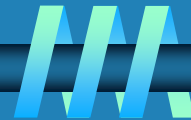
- Participation to ICDAR

∩ Video Indexing and Retrieval

- Participation to NIST Trecvid



The Balamand Experience – HLT Tools



∩ Developed at Balamand

- **Becars (a freeware)**
- **HCM (HMM toolkit)**
- **CART**
- **NN-MLP**

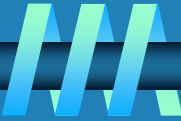
∩ Other freeware experimented

- **SRILM**
- **HTK**
- **SPHINX**

∩ HLT Resources

- **NEMLAR resources**
- **CEDRE database**
- **IFN/ENIT**
- **Annahar**

Recommendations



- ❧ **More Resources**
 - **Statistical models need data to get better performances**

- ❧ **Regional competitions to develop technologies**
 - **Connect with international competitions**

- ❧ **Workshops and conferences to be organized in the region**

- ❧ **ISCA-WANA and NEMLAR consortium to support these efforts**

- ❧ **Connect with private sector**