# Arabic Language Resources and Tools for Speech and Natural Language

**Mansour Alghamdi*, Chafic Mokbel** and Mohamed Mrayati*****

*King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia. mghamdi@kacst.edu.sa; **University of Balamand, Lebanon. chafic.mokbel@balamand.edu.lb; *** UN-DESA. mrayati@un.org.

**Abstract**

Although 5% of the world population speak Arabic as a native language, the research on Arabic is not up to its number of speakers. However, more research has been don on Arabic in the last decade due to many factors including the widespread of communication and information technology applications. Two of the eminent research centres in the Arab World which contributed markedly on research related to Arabic are King Abdulaziz City for Science and Technology (KACST) and University of Balamand (UOB). This paper is to highlight the efforts of these two institutions on enriching the Arabic language resources including that on natural language processing, speech processing and character recognition.

## Introduction

Arabic speech and language processing has gained a lot of interest in the past decade. Several projects have been launched at the international level. DARPA has launched the GALE "Global Autonomous Language Exploitation" project (Cohen 2007). It started in 2005 and aims to transcribe and translate audio and written documents from Arabic or Mandarian Chinese to English. The target is to build by 2010 devices capable to do such transcription-translation in real-time with an accuracy of 90% to 95%. Europe has also funded several projects relative to the Arabic language. We cite the ALMA (AbiRached ), NEMLAR (Yaseen, 2006) and the MEDAR (Maegaard 2008) projects. The NEMLAR project permitted to draw a map especially in the Arab region of the existing resources, tools, technologies and actors in the Arabic speech and language processing. This led to the definition of needs and availability of resources and tools for the Arabic language; the BLARK "Basic Language Resources Kit" (Maegaard 2006). Nemlar has also permitted to define, collect and develop two databases; a broadcast news database and a text database (Yaseen 2006).

Besides the projects, several international competitions are conducted on a yearly basis in order to assess and improve the technology for the processing of the Arabic language. The National Institute of Standards and Technology (NIST) organizes several evaluation competitions . In Arabic handwritten recognition, ICDAR contest assesses and benchmark the different technologies (Margner 2005).

From resources side, two institutions, the ELRA and LDC distribute useful resources for the development of Arabic speech and language processing technologies and applications.

The projects, resources and tools developed for the Arabic language show the high interest in the Arabic Speech and Language processing in the past decade. This paper focuses on the activities, resources and tools developed in two laboratories in the region, i. e. KACST and UOB, in relation with the Arabic speech and language processing. It shows the major achievements in these areas.

## Natural Language Processing

Natural language processing (NLP) or what is sometimes called computational linguistics is the area of knowledge where the text of a natural language such as Arabic is digitized and computed. It includes natural language understanding and generation. In order to do so, some of the human brain linguistic functions need to be simulated. These include: spell checker, morphological analyser and generator, grammar checker, lexicon and so on.

### Morphological Analysis and Tagging

KACST has two projects at the morphological level. The first is the Arabic language Morphological analyzer. This project aims to develop algorithms for morphological analyzing Arabic language vocabulary according to their morphological and grammatical properties. The algorithm depends on the morphological and grammatical characteristics that are extracted from a large linguistic corpus. The system consists of texts tokenizer tools, storage and retrieval of the grammatical and morphological characteristics, the characteristics of the entry system and expert learning system. The system can be used in analyzing and coding Arabic language properties, grammatical and morphological analysis systems, recognizing the grammatical and morphological properties of Arabic language and statistical linguistic systems.

KACST has developed an Arabic stemmer. The system is used to analyze Arabic vocabulary and linguistically recognizes and isolate suffixes, prefixes and infixes and then extract the corresponding stem. The system depends on regular expressions and serves the users and developers of applications that support Arabic language and help in expanding research process about Arabic sources. The system consists of tools for dividing up the texts, tools for analyzing regular expressions, group of regular expressions that exceed 1400 expression.

Arabic lexis morphological rules have been developed by KACST. The aim of this project is to create a database that includes all the morphological rules of Arabic lexis that would lead into the production of the necessary morphological, syntactic and phonological algorisms associated with the generation of lexis through fully computerized systems. The database entries include lists

of the Arabic alphabet, word roots and their features. The entries, also, include morphological measures used in generating Arabic lexis, such as, inert nouns and verbs, derivatives and infinitive forms of verbs and nouns. Morphological features of generated lexis such as the plural and diminution forms, relation and feminine cases, the tense and aspect of verbs and the nominal forms generated from infinitive and derivative nouns are also included in the database. This project is a base for Arabic language processing applications. These applications cover texts generation and analysis, texts compression and compilation, machine search, data coding, machine translation, Arabic language machine teaching and language understanding. The system includes new algorisms that facilitates the process of constant and regular generation of new lexis in Arabic language and to collect and compile rare and infrequent forms in small groups.

Modern Arabic writing includes only the letters that represent the consonants. This means that Arabic vowels and geminates are not represented in the daily writing of Arabic. The absence of the vocalic and geminate symbols does not allow for full usage of other computational systems such as text-to-speech and automatic speech recognition systems and search engines. Therefore, KACST has started doing experiments on Automatic Arabic Diacritization to develop a system that can be integrated in other related computer systems. The result is KACST Arabic Diacritizer (KAD). KAD contains algorithms that calculate the highest probability, algorithms that select the diacritic of the highest probability and 68,378 quad-grams of the Arabic letters and diacritics. The system accuracy is 87% at the letter level including word-final letters. It diacritizes more than 500 words/second. Also, it is small in size, only 3 MB's (Alghamdi and Muzaffar, 2007; Elshafei, et al. 2006).

Transliteration of Arabic names has not been consistent for the reason that Arabic orthography is different from that of the Roman alphabet. The result is that the same name is Romanized in different ways. Such inconsistency has negative effects on individuals and institutions. To standardize the method of Romanizing Arabic names and make it available to others, KACST has developed the KACST Arabic Name Romanizer (KANR). KANR contains algorithms that process Arabic letters, algorithms that transliterate the Arabic letters into the Roman alphabet, algorithms that process Roman letters and more than 50,000 Arabic names written in the diacritized Arabic orthography (Alsalman, et al. 2007).

Although the Arabic alphabet is sufficient to be used in writing and reading in Arabic, as it is the case for other languages and their writing systems, it does not possess symbols for all the Arabic sounds whether they are part of the standard or dialect inventories. Therefore, Arabic speakers find it impossible to transcribe speech sounds using Arabic alphabet. They often either describe the sound in words or use the International Phonetic Alphabet. Due to this fact, KACST has designed a phonetic alphabet based on the Arabic alphabet to represent all speech sounds for all languages. The new phonetic alphabet is called Arabic International Phonetic Alphabet (AIPA). AIPA was further developed as 2 fonts that can be installed on a PC and used in any word processor (Alghamdi, 2006b).

### Language Modelling

The statistical N-gram models are used as the state of the art language models. N-gram define the probability distribution of the vocabulary words in the context of a N-1 words. The Arabic language is characterized by a rich morphology. Therefore, morphologically constrained N-gram (Kirchhoff 2002) have been proposed in order to reduce the complexity of the language models (number of parameters) extending by the same fact their generalization capabilities. In (Ghaoui 2005) a novel approach for the morphologically constrained N-gram has been proposed. It consists in considering the word as a couple (stem, rule), the stem being the root of the word and the rule is the grammatical one used to derive the word from the stem.

$$\Pr(w_i / w_{i-1} \dots w_{i-N+1}) = \Pr[(s_i, r_i) / (s_{i-1}, r_{i-1}) \dots (s_{i-N+1}, r_{i-N+1})]$$

By decomposition and simplification, several models may be derived. For example, one can make independence hypotheses leading to:

$$\Pr(w_i / w_{i-1} \dots w_{i-N+1}) = \Pr[s_i / s_{i-1} \dots s_{i-N+1}] \Pr[r_i / s_i r_{i-1} \dots r_{i-N+1}]$$

This models the probability of a word given its context as the multiplication of the probability of the stem given its stems context by the probability of the rule given the previous rules and the current stem.

Whatever the simplified model applied, the method requires a good morphological analyzer in order to decompose the words into their corresponding stems and rules. A simple automaton based morphological analyzer has been built at the University of Balamand and was used in this morphologically constrained N-gram. The resulting model has shown limited degradation of the perplexity with significant reduction of the number of parameters.

## Speech Processing

Speech processing covers a set of technologies including speech coding, text to speech, speaker recognition and, speech recognition. While speech coding techniques are universal with the exception of very low rate speech coding, the other technologies highly depend on the language. In this section, several technologies and technology tools developed in UOB and KACST laboratories are described.

### Speech Recognition

Arabic speech recognition has been a major activity of interest as mentioned earlier. This was also the case at the University of Balamand. The activities cover the following directions:

- Arabic broadcast news transcription
- Arabic telephone commands
- Multilingual speech recognition

Systems have been developed in the three areas. In order to build those systems both resources and tools are needed. Speech resources must be annotated. First, for the telephone commands a polyphone like database has been collected. This database contains commands, dates, numbers and, small expressions uttered by 25 male speakers and 30 female speakers, all between 19 and 25 years old. The vocabulary is in two languages; Arabic and

French (Bayeh 2004). In definitive, around 1100 sentences are validated from each language. This database has been experimented in telephone commands and multilingual speech recognition. This database is called BEAF for "Balamand ENST Arabic and French" database.

State of the art speech recognition systems make use of Hidden Markov Models (HMM) to represent the speech units. At the University of Balamand a HMM toolkit has been built. This toolkit is called HCM. It includes tools to compile HMM models, initialize and train those models, and use those models in speech recognition experiments. HCM has a generic structure. Every tool has a configuration file. A BNF language with C-like syntax is used to write the configuration parameters. HCM has been successfully used and shown state of the art results in experiments going from limited vocabulary isolated words to large vocabulary continuous speech recognition.

Experiments have been conducted on BEAF database using the HMM toolkit HCM. The baseline speech recognition results are at the state of the art (i.e. less than 2% of error rate).

For broadcast news transcription, speech resources and tools are also needed. The Nemlar resources (Yaseen 2006) have been used in the experiments conducted at the University of Balamand. This Nemlar database is a broadcast news database formed of about 40 hours. State of the art performance is also achieved for broadcast news transcription (Bayeh 2008), i.e. 86% of accuracy. It is worth noting that a Classification And Regression Tree (CART) algorithm has been developed and used to classify the triphones models in the system.

Experiments have been conducted in multilingual speech recognition. Phonetic models trained on the French language have been used to build Arabic speech recognition models. For this purpose an association between unit models from both languages has to be found. This can be done manually by a human expert or automatically in a data driven approach. The idea of data driven method is to find the optimal association that would permit to better represent few acoustic data from Arabic using French unit models. For both telephone commands and Broadcast News transcription, UOB team has been able to show that better multilingual acoustic models may be obtained if data-driven association is used between the acoustical units of the different languages than when the association is set by human experts.

Although we have been able to build systems using locally developed tools, the resources used have limited size and scope. More important resources exist but they are not very accessible. Accessible resources would permit to better develop the research in this area in the region.

Similar projects on speech has been done in KACST laboratories. KACST team started in the med 1990's to build a phonetic database on Arabic. By the turn of the century, KACST Arabic Phonetic Database (KAPD) was available for researchers and research centres. The most sophisticated equipments were used to collect the database to provide data that can easily be utilized by researchers and those who are interested in speech and phonetics. KAPD contains more than 46,000 files including: 12,000 wave files for all Arabic sounds in different word positions, aerodynamic data, electropalatogrphic data, degree of nasalance, and images of the face, the vocal folds, the epiglottis and the velopharyngeal port,. 7 native speakers of Arabic participated in KAPD experiments (Alghamdi, 2003). KAPD is distributed to researchers and research centres without charge. It has been used in different text-to-speech (TTS) and speech-to-text systems (STT).

KACST has collected another speech database called Saudi Accented Arabic Voice Bank (SAAVB). It represents Arabic native speakers from all the cities of Saudi Arabia. SAAVB has speech that was collected from 1033 speakers. It has 183,518 audio files in addition to their transcriptions. The database has been licensed to IBM which is now integrated in its speech recognition engine.

KACST has developed speech related systems including TTS (Elshafei, et al. 2002), STT (Elshafei, et al. 2008; Alotaibi, et al. 2008; Alkanhal, et al. 2008; ) and speaker verification (Alkanhal, et al. 2007; Alghamdi, 2006a).

### Speaker Recognition

State of the art text-independent speaker recognition systems use Gaussian Mixture Models. For this purpose Becars, an open source software, has been developed in a joint effort by the University of Balamand and the ENST. Becars is a GMM toolkit allowing, to compile, initialise, train and test GMM models in different applications. The strength of Becars is in the different adaptation techniques that are included. Actually, MAP, MLLR, and unified adaptation (Mokbel, 2001) procedures are fully implemented. In (Blouet 2004) the Becars software is described.

Using Becars, the University of Balamand has participated to several NIST (speaker recognition) evaluations. Becars has been used also for the segmentation of the audio signal in broadcast news transcription applications. A participation to the ESTER evaluation with such a segmentation system has been done.

Speaker recognition systems may be used in standalone applications or within other applications like the segmentation of audio signals in different speakers. The speech resources needed to build speaker recognition systems and to benchmark those systems are lacking in the Arabic region. Projects need to be launched in order to build such resources.

## Handwritten Document Recognition

Arabic handwritten recognition is also being a subject of large interest. The only freely available resource is the IFN/ENIT database[1], which is built in a joint effort between IFN (Germany) and ENIT (Tunisia). The database lexicon is formed of the Tunisian towns. It includes more than 26 000 city words images written by around 411 writers. A biyearly contest is organized by ICDAR (Margner 2005). A baseline system has been developed based on HCM and showed state of the art results (El Hajj 2005). While the IFN/ENIT database is limited to isolated Tunisian towns names, there is a need for annotated databases resources with full text in order to develop handwritten document recognition.

KACST has developed a system for cursive Arabic text recognition. The system decomposes the document image

[1] http://www.ifnenit.com

into text line images and extracts a set of simple statistical features. The Hidden Markov Model Toolkit (HTK) is used to process the output. The system has applied to a data corpus which includes Arabic text of more than 600 A4-size sheets typewritten in multiple computer-generated fonts (Khorsheed, 2002).

## Conclusions and Perspectives

In this paper we have presented several resources and tools developed in KACST and UOB laboratories. KACST has developed an Arabic stemmer and an Arabic lexis of morphological rules. In addition, an Arabic diacritizer KAD and an Arabic name romanizer KANR have been produced at KACST. At the University of Balamand a French-Arabic telephone commands database has been collected and is available upon request. Two speech databases have been collected by KACST: KAPD and SAAVB. A free open source GMM toolkit, Becars, is being distributed. A full HMM toolkit, HCM, has been developed and used in several Arabic speech recognition experiments and Arabic handwritten experiments.

The available resources developed in the region and the tools built and tested on those resources are at the state of the art. However, there is a need for larger resources and for the development of the research activities on Arabic speech and language processing. This may be done through the organizations that are involved in such activities. Benchmarking, verification and evaluation would certainly help in this direction.


### Acknowledgements

The authors would like to express their gratitude to KACST, UOB and UN-DESA for their support that made this paper possible.


### Bibliographical References


AbiRached M. & Faby J.A. (2004), Communication and information technologies at the international office of water (IOW), In Proceedings International Conference on Information and Communication Technologies: From Theory to Applications, (pp. 157-158), Damascus, Syria.

Alghamdi, Mansour (2006a) "Voice Print": Voice Onset Time as a Model. Arab Journal for Security Studies and Training. 21. 42: 89-118.

Alghamdi, Mansour (2006b) Designing Arabic Fonts to Represent International Phonetic Alphabets. Journal of King Abdulaziz University: Engineering Sciences. 16, 2: 27-64.

Alghamdi, Mansour (2003) KACST Arabic Phonetics Database, The Fifteenth International Congress of Phonetics Science, Barcelona, 3109-3112.

Alkanhal, Mohamed, Mansour Alghamdi, Fahad Alotaibi and Ammar Alinazi (2008) Arabic Spoken Name Recognition. International Workshop on Signal Processing and Its Applications, 18 – 20 March 2008, University of Sharjah, Sharjah, U.A.E.

Alkanhal, Mohamed, Mansour Alghamdi and Zeeshan Muzaffar (2007) Speaker Verification Based on Saudi Accented Arabic Database. International Symposium on Signal Processing. Sharjah, United Arab Emirates. 12-15 February 2007.

Alotaibi, Yousef Ajami , Mansour Alghamdi and Fahad Alotaiby (2008) Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus. The 2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition. Las Vegas, USA.

Alsalman, Abdulmalik, Mansour Alghamdi, Khalid Alhuqayl and Salih Alsubay (2007) A Computerized System to Romanize Arabic Names. The First International Symposium on Computers and Arabic Language. 25-28/3/2007.

Bayeh, R., Lin, S.-S. Chollet, G. & Mokbel, C. (2004), Towards multilingual speech recognition using data driven source/target acoustical units association, In Proceedings International Conference on Acoustic, Speech and Signal Processing (ICASSP), (pp. 521-524), Montreal, Canada.

Bayeh, R. Mokbel, C. & Chollet, G. (2008) Broadcast News Transcription baseline system using the Nemlar database, In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco.

Blouet, R., Mokbel, C., Mokbel, H., Sanchez, E., Chollet, G. & Greige, H. (2004) Becars: A free software for speaker verification. In Proceedings of Odyssey (pp 145-148), Toledo, Spain.

Cohen, J. (2007), The GALE project: A description and an update, In Proceedings IEEE workshop on Automatic Speech Recognition and Understanding ASRU (pp 237-240), Kyoto, Japan.

Ghaoui, A. Yvon, F. Mokbel, C. & Chollet, G. (2005) On the Use of Morphological Constraints in N-gram Statistical Language Model, In Proceedings of the Interspeech, 9th European Conference on Speech Communication and technology, Lisbon, Portugal.

El-Hajj, R., Likforman-Sulem, L., Mokbel, C. (2005), Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, In Proceedings of the 8th International Conference on Document Analysis and Recognition, Seoul, Korea.

Elshafei, Mustafa, Husni Al-Muhtaseb and Mansour Alghamdi (2002) Techniques for High Quality Arabic Speech Synthesis, Information Science, 140 (3-4) 255-267.

Elshafei, Mustafa, Husni Al-Muhtaseb and Mansour Alghamdi (2008) Speaker-Independent Natural Arabic Speech Recognition System. The International Conference on Intelligent Systems. Bahrain, 1-3 December 2008.

Elshafei, Mustafa, Husni Al-Muhtaseb and Mansour Alghamdi (2006). Machine Generation of Arabic Diacritical Marks. The 2006 World Congress in Computer Science Computer Engineering, and Applied Computing. Las Vegas, USA. 26-29/6/2006.

Khorsheed, M. (2002) Off-line Arabic character recognition - A review. Pattern Anal. Appl. v5 i1. 31-45.

Kirchhoff, K. et al. (2002), Novel Speech Recognition Models for Arabic, Johns-Hopkins University Summer Research Workshop, Baltimore, USA.

Maegaard, B. et al. (2006), The BLARK Concept and BLARK for Arabic, In Proceedings of the 5th

International Conference on Language Resources and Evaluation (LREC), (pp 773-778), Genova, Italy.

Maegaard, B. et al., (2008), MEDAR: Collaboration between European and Mediterranean Arabic Partners to Support the Development of Language Technology for Arabic, In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco.

Margner, V. Pechwitz, M. & El Abed, H. (2005), ICDAR 2005 Arabic Handwriting Recognition Competition, In Proceedings of the 8th International Conference on Document Analysis and Recognition, Vol. 1 (pp 70-74), Seoul, Korea.

Mokbel, C., Abi Akl, H. & Greige, H. (2002) Automatic speech recognition of Arabic digits over the telephone network, In Proceedings of Research Trends in Science and Technology, Beyrouth, Lebanon.

Mokbel, C. (2001) Online Adaptation of HMMs to real-life conditions: a unified framework, In IEEE Trans. on Speech and Audio Processing, Vol. 9, Issue 4, (pp 342-357).

Yaseen, M. et al. (2006), Building Annotated Written and Spoken Arabic LRs in NEMLAR, In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), (pp 533-538), Genova, Italy.