Language Technology for Arabic

Bente Maegaard
Khalid Choukri, Chafik Mokbel, Mustafa Yaseen

July 2005

Table of contents

ISBN 87-90708-15-6

http://www.nemlar.org, email:nemlar@cst.dk

The NEMLAR project was supported by the European Commission. However, views expressed in this document are solely the responsibility of the project, and do not necessarily reflect the opinion of the European Commission

Language Technology for Arabic

A cooperative effort

1. Introduction

This booklet attempts to give perspectives for Arabic language technology. Language technology is of paramount importance for efficient communication, as described below. The booklet addresses decision makers, funding agencies and researchers with the aim of creating awareness and of obtaining a mutual understanding of the possibility for a co-operative effort.

2. The need for language technology

In the modern Information Society, computers are used in relation to a very high proportion of communication. For business as well as for government administration and for the citizens, it is important to be able to produce information efficiently, to translate, to retrieve information, both in written and spoken form. Human Language Technologies (HLT) enable humans to communicate with computers and to use computers and the internet in a more natural way and in their own language, i.e. to participate in the information society in a totally natural way. Native speakers of languages that are not well served by language technology suffer from less access to information, and from less efficient tools, and higher productions costs for documents and translation.

For the Arabic language, some tools already exist which will enable the user to use Arabic in the communication with computers, but there is a long way to go before the Arabic tools reach the level which exists for e.g. English, French or Japanese, thereby enabling an efficient communication and good support for the Arabic business sector, and for administration and culture. Such tools will also be very important for the possibility for the Arab world to have its products, as well as its ideas and its culture exported to the rest of the world.

One of the main building blocks for HLT is language resources (LRs): collections of text and spoken language, dictionaries, grammars, speech recognition modules, speech synthesis components, etc.

Full-fledged HLT for Arabic can only be developed when a reasonable amount of LRs is available. LRs are needed for the language industry, for the translation industry, and in general for any content industry. Large companies may produce their own resources for the languages for which a business can be made, but small companies cannot afford this. Additionally, resources that are built by companies are normally not shareable as they are seen as a competitive advantage and consequently not traded, or traded at a high price.

This is the reason it is so important to create non-expensive, shareable LRs for Arabic.

Extract from Arab Human Development Report 2004, UNDP:
Horizontally, e-government, e-commerce and decision support systems are as an average either non-existent or at their infancy. Government processes do not yet take full advantage of modern information-processing technologies. Information delivery to citizens is severely limited, especially in rural areas. Even high-income sectors like tourism are only starting to take advantage of the Internet as a public relation and advertising mechanism.

3. State-of-the-art for Arabic Language Resources, tools and applications

Fortunately, the interest in Arabic language technology has been emerging for some time, and therefore there are active teams in the universities in almost all countries in the region. Also, there are companies in the region, who develop language technology tools. However, the scarcity of language resources such as large speech databases and large text corpora for training of tools for Arabic, means that the development of tools is not as fast as is needed by the society in the Arab countries. Basic language resources and tools should be available for all teams in the universities, so that they can do research in order to develop next generation of language tools, and they should be available for industry, so that a wide range of basic components can be developed, for the benefit of all using the Arabic language. A need for more enterprises who are involved in development of LRs and tools, and software companies who might integrate language technology within their products, act both locally, regionally, and globally has also been identified.

Examples of integration of language technology are automatic answering services in the telecom sector, automatic call centres, language enhanced information systems where you can ask in your own language, and where language technology is used to filter the information so that only relevant information is presented. Summarisation tools may provide abstracts or short versions of longer texts, which are easier to skim than the full text. News may be read aloud to blind people, or others who cannot read. Doctors may be told the blood pressure etc. automatically during surgery. Banks may use language technology in their automatic customer service on the internet, e-learning with the use of Arabic for teaching language and other themes may be developed. The availability of language technology for Arabic will also have cultural effects as can be seen from the quote from the UNDP ICT-report below: More information will be available in Arabic, and this will lead to a boost for the Arabic language and for the people who speak it.

Extract from Arab Human Development Report 2004, UNDP:

Language

Language usage also plays an important role. Although some regional information mechanisms have been created and are being maintained today, the languages used are primarily English and French, with Arabic developed to a much lesser extent. This exacerbates the hub-and-spoke effect. A number of initiatives are taking place to facilitate the use of ICTs in the Arabic language. They have not yet, however, acquired a substantial critical mass in order to mainstream Arabic language and culture through the Internet in a more pervasive fashion. Lack of standardisation, limits the use of the Arabic language.

It is worth noting that China faced similar challenges in terms of language content and infrastructure in 1998. A large effort was then made to boost internal connectivity and encourage the creation of content in Chinese. This somewhat decreased the demand for international access and increased domestic exchanges. When the rest of the region joined in, led by Chinese-speaking Hong Kong and Singapore, and several fibre-optic backbones became operative, Asia woke up to a different, more region-centred reality. The effort is today paying its dividends in a substantial increase of regional trade and information sharing.

Compared to Asia, horizontal exchanges of knowledge are still not frequent among Arab State countries, where best practices and lessons learned are shared less regularly. UNDP is keen on establishing region-wide communities of practice (ICTs being one of them) to address regional concerns.

In order to be able to develop Arabic language technology to a level which can be compared with other important languages and upcoming economies, more research should be done at the universities and research centres, preferably in collaboration with industry.

This would lead to a more widespread knowledge about language technology and how it can be used, e.g. in companies producing documentation to go with products, for translation, for summarisation, and for information retrieval and knowledge sharing.

4. The NEMLAR project and its results

The NEMLAR **(Network for Euro-Mediterranean LAnguage Resources)** project was started in order to help pave the way for a collaborative effort for Arabic language resources in the Mediterranean area. The project is supported by the European Union under the INCO-MED programme which supports collaboration between the EU and the countries in the Mediterranean region. The project runs 2003-2005. The project has 14 partners, listed at the end of this booklet.

Now, as NEMLAR is coming to an end, we can summarise the results as follows.

People and institutions:

- **Network**: First of all, the NEMLAR core network, consisting of the 14 partners, has proven to be very well suited for the task. Together the partners cover the important areas of HLT and language resources in a very comprehensive way. And also geographically, the Mediterranean region is well covered. However, in order to promote the NEMLAR ideas and to give more people access to information about Arabic language technology and shareable Arabic language resources, the NEMLAR project has extended its network. One of the important extensions is the regional one: the Arabic language is very important outside the Mediterranean region as well, so members are welcomed from other regions of the world.

Information and documents:

- **Surveys**: Two surveys have been produced by the project partners. *Report on Survey on Arabic Language Resources and Tools in Mediterranean Countries* gives an overview of existing Arabic LRs and tools in the region. As a derivative of this survey, a list of institutions and companies involved in the production and distribution of LRs and tools has been made. The second survey is *Survey on the Industrial Needs for Language Resources and Tools in Mediterranean Countries***.** The needs of industry are important for giving priorities to the development of LRs. The two surveys and the list of institutions and companies may be extended in scope and coverage, through new members of the network and through promotion at conferences, newsletters etc.
- **BLARK for Arabic**: A BLARK (Basic Language Resource Kit) describes the minimal set of language resources that are necessary for developing pre-competitive HLT for a language. The NEMLAR project has elaborated the first BLARK for Arabic. This taken together with the survey on existing LRs, is a very good starting point for deciding on priorities for development of LRs and tools.

Language resources:

NEMLAR has been able to do development of a few language resources, and has chosen the most important ones, based on the BLARK and the needs expressed by industry and research.

- **Written corpus**: An annotated written corpus of Modern Standard Arabic, fully vowelized, POS-tagged (approx. 500K words)
- **Speech database:** A speech database for text to speech synthesis with a male and a female voice with a well designed textual corpus of Modern Standard Arabic

- **Speech database:** An Arabic speech database of broadcast news; fully annotated at various levels (orthographically, named entities, etc.)

Dissemination and meeting place:

- **Conference**: NEMLAR held the first *Arabic Language Resources and Tools Conference* in Cairo 2004. It brought together academics and industry from all over the world to discuss issues in Arabic HLT.

5. Further development of LRs, tools and applications for Arabic

As mentioned above, further development of language technology is absolutely necessary in order for the countries having Arabic as their language to participate fully in the Information Society and take the advantages of efficient information production and processing.

When the NEMLAR project ends mid 2005, a very good development has begun: Experts have been identified, working methodologies have been exchanged and refined etc. A few language resources have been produced, but in order to have a major impact on the availability of Arabic language resources and tools, the collaboration needs to be continued and extended. The NEMLAR network will continue to exist beyond the project, and will probably change into an association, see below.

LRs and tools

The most important activity will be to develop LRs and basic tools, with the goal of filling in the missing elements of the BLARK. Such an activity cannot be done by the NEMLAR association as such, but has to be performed by various consortia. The association could act as a guarantee for quality and relevance, e.g. by maintaining the BLARK and the survey of existing LRs and tools, so that future proposals can always be seen in relation to a publicly available inventory.

In order to achieve the goal of filling the BLARK, and of developing new technology for Arabic, ideally an R&D programme for the development of Arabic language technology should be started. This could be done either by Arab countries or probably in collaboration between Arab countries and the EU. However, this will take time, and it is important that activities are going on. Consequently the aim is to start at least two major projects, probably one funded from Europe, one from the Arab countries, in LRs and core language technology.

To measure the progress in the development and coverage of LRs and tools for Arabic, key performance indicators will be developed, e.g. money spent on research in Arabic HLT, number of persons involved, number of tool kits, etc in the directory, and number of projects. How many real applications/systems for the Arabic language exist after 3 years? A desirable evolution would be something like this:

	Applications	Prototypes/demo	Components	LRs
After 3 years	5	6	4	2
After 5 years	10	12	8	5

More specifically translation technology should be available, not only for English-Arabic and Arabic-English, but also for other important language pairs with Arabic; search engines enriched with language technology for Arabic, speech technology e.g. for reading newspapers aloud and for dialogue; document authoring systems for easy document production, business sector specific applications, e.g. for banks.

Arab League: *Laying Due Concern to Arabic Language No Less Important than Political Issues, Secretary General Says (9-3-2005)*
In his speech in the celebration organized by the Egyptian Ministry of Education at the Arab League's General Secretariat to honor the best Arabic language readers, Secretary General Moussa said the concern laid to perfecting the Arabic language has become minimal, a matter he described as regrettable in view of how important this language is as the first means of inter-communication among all Arabs. He further pointed out how clear it is to Arabic readers that linguistic errors keep increasing and recurring in speeches, articles and books without even trying to remedy them.
In the meantime, Moussa has expressed great enthusiasm over maintaining the classic Arabic language and promoting it in view of the complaints made by many Arab communities over the retreat of the Arabic language in the face of other foreign languages.
He has also promised to lay further concern to the Arabic language and maintaining it through cooperation among the ministries of education in the Arab states and concerned Arab organizations. In this regard, he has stressed the importance of the role of the Arab Organization for Education, Sciences and Culture as regards promoting and supporting the Arabic language and the Arab cultural movement.
Moussa has regarded the celebration as part of the celebration marking the Arab League's sixtieth anniversary which will correspond the holding of the upcoming Arab Summit in Algeria.
On the other hand, Secretary General has honored a number of Arabic readers who won in the subject competition from Egypt, Jordan, Syria, Saudi Arabia, Sultanate of Oman, United Arab Emirates, Bahrain and Kuwait.

Information about language resources

Surveys

The survey on existing LRs and tools should be maintained, otherwise it will be obsolete in a very short time. The NEMLAR association (see below) will be an excellent source and a relevant hosting institution.

BLARK

The BLARK document is more stable, but a BLARK will develop over time as new types of applications and hence new types of resources emerge. Such a document should be maintained by the community, and the NEMLAR association will be in a very good position to be the host of this.

6. The actors

One type of main actors in this desired development is the language technology community: academia and companies, but they cannot alone drive this development. Like for any other language, funding is needed.

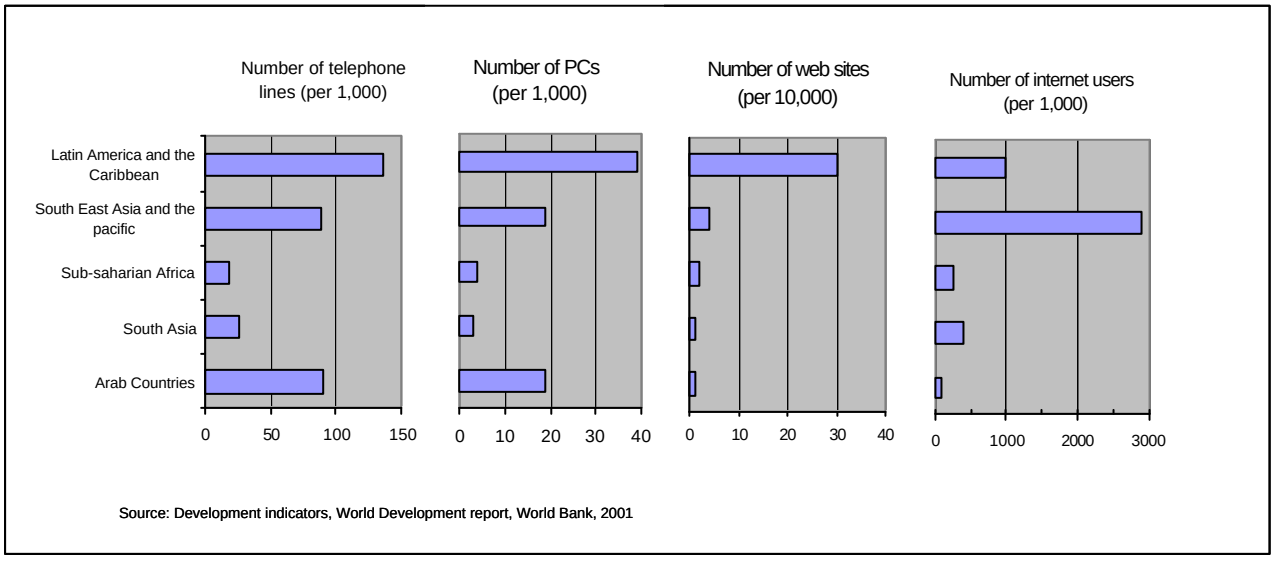
Strategic planners, policy-makers, decision-makers and funding agencies who have an interest in promoting research & development and innovation for the Arabic language in the field of language technology will be one of the main driving forces. Such bodies may want to support the communication and cultural dialogue between e.g. Europe and the Arab countries, or they may want to attack the unemployment and poverty by enhancing the efficiency and effectiveness of the local economies and bridge the digital divide in the Arab countries.

One such institution is the European Commission who has already a programme supporting the collaboration between EU countries and Mediterranean countries, INCO-MED. It is to be hoped that this programme will be continued, or that similar programmes will be decided in the future. The INCO-MED programme has enhanced the development of the cultural dialogue and partnerships across the Mediterranean, as well as the advancement of science to the benefit of all involved parties. It was wise to select language technology as one of the areas to support, and this has to be continued.

United Nations, and its associated bodies, and the World Bank have as one of their objectives to make information accessible for all people in their native languages. It has been proved in several regions of the world that human development may be largely facilitated using Information and Communication Technology (ICT). ICT can help to bridge digital divide and to better integrate the new knowledge-based economy. In 1974, the United Nations adopted the Arabic language as one of its 6 official languages. Being one of the official languages of UN, the Arabic language is a special concern.

In May 2002, the United Nations Development Program (UNDP) has established the Information and Communication Technologies for Development in Arab Region (ICTDAR). This followed the publication by the UNDP Regional Bureau for Arab States (RBAS) of the Arab Human Development Report that stresses the challenge of accessing information. The 2004 edition of the Arab Human Development Report continues to stress the importance of the ICT in human development and the major role that can play the Arabic language to facilitate the access to information. The ICTDAR regional programme enhances and complements UNDP's national and global efforts and issues highlighted in the Arab Human Development Report. The ICTDAR has launched several projects in this direction.

Before going further and in order to illustrate the access to information in the Arab world, we reproduce the following table from the Arab Human Development Report 2002, as it compares to other regional areas.



Source: Development indicators, World Development report, World Bank, 2001

A number of initiatives are taking place to facilitate the use of ICTs in the Arabic language. They have not yet, however, acquired a substantial critical mass in order to mainstream Arabic language and culture

through the Internet in a more pervasive fashion. Lack of standardisation limits the use of the Arabic language. The development of Arabic language technology is a necessary step to ease the access to information in the Arab world. Besides, the economical impact is important. Therefore, we believe that policy makers should initiate the development of Arabic language technologies within targeted programs serving the human and economical development in the Arab region and bridging the digital divide.

The Arab League

The Arab League consists of 22 countries. It is the purpose of the League to create and support collaboration between the member states, for the areas mentioned below.

From the Charter of the Arab League: The League has as its purpose the strengthening of the relations between the member-states, the coordination of their policies in order to achieve co-operation between them and to safeguard their independence and sovereignty; and a general concern with the affairs and interests of the Arab countries. It has also as its purpose the close co-operation of the member-states, with due regard to the Organisation and circumstances of each state, on the following matters:

A. Economic and financial affairs, including commercial relations, customs, currency and questions of agriculture an industry.
B. Communications; this includes railroads, roads, aviation, navigation, telegraphs and posts.
C. Cultural affairs.
D. Nationality, passports, visas, execution of judgments and extradition of criminals.
E. Social affairs.
F. Health affairs.

Language technology fits very well with the purposes. It concerns in particular A and C, and to a lesser extent B, and is certainly highly relevant for the Arab League.

In addition to the Arab league, the global need for the development of ICT strongly appears in the Economic and Social Commission for Western Asia (UN-ESCWA) countries. In the ESCWA report "Damascus Call Towards Partnership for Building the Arab Information Society" (November 2004) an invitation is issued to the Arab countries for the development of Information technologies and the Arabic language use. The Technology and Capacity-Building Initiatives for the Twenty-First Century in the ESCWA Member Countries (June 2001) shows the clear policy of the concerned countries to develop information technology using the Arabic language. E-government, e-learning and other applications are targeted. As a final example, the Lebanese Research Council has stated the Web and Arabized software technologies as one of the country priorities in Science and Technology in 2005.

Governments in the Arab States, through their policy-making mechanisms, each hold the key to the development of a knowledge-driven society. In the past, already some countries were able to support the ICT sector, and if more governments and research councils take the same type of action, this will give a very important push forward. Governments may also see their own interest in boosting the field as, apart from making administration more efficient and more secure, this will help them introduce e-government with a more human face, and help them boost the economy of their country.

The other actors, apart from these major organisational actors, are of course the community, i.e. universities, research institutions and companies, already mentioned above.

In addition we are suggesting an association which will take the R&D population and anyone else belonging to the community as members, and which will have as its goal to promote the NEMLAR mission.

7. The NEMLAR association

NEMLAR finds it very important that the initiative which was started by the European Commission should be continued. There are various ways in which this should be done. First of all NEMLAR will take the initiative to create an association on the basis of the NEMLAR network. The association will accept as members all those who want to contribute to the NEMLAR goals, be it individuals, institutions or companies. NEMLAR goals have been mentioned several times, in short they are: the production and availability of shareable LRs and tools, the advancement of Arabic language technology, and the collaboration between countries towards these goals.

The association will maintain a newsletter, and will create awareness through conferences etc. The association will probably not have the funds to maintain the BLARK document for Arabic, the directories and surveys, but it will make sure to maintain the framework for elaborating these, and to always display the latest version.

Conference

After the first conference (Cairo, 2004) it was strongly articulated that the effect and impact would only manifest itself if it was repeated in two years' time. It may not be relevant to have international conferences on Arabic in perpetuity, as discussions on Arabic may take place at other international conferences alongside other languages, but for some years to come, this particular focus on Arabic will be necessary. One of the Emirates has already offered to host the next conference.

It is important to keep the expertise together and to build on the strengths that have been developed. The NEMLAR network is much stronger than the sum of its partners. At the same time the network should be extended and encompass more of those who want to contribute to the same goals. This should be partly achieved through the creation of the NEMLAR association.

8. Conclusion

The development of language resources and tools for the Arabic language is important for the economy in the Arab countries; but at the same time it is important for the culture. By focussing on Arabic language technology and making both the technology and content available in Arabic, the use of Arabic will grow and the request for foreign language information will decrease. At the same time language technology can help access information in foreign languages, even without a very good knowledge of these languages. And finally, it can help spread Arabic ideas and culture to non-Arabic languages.

An excellent basis has been laid by the NEMLAR project, and it is now time to follow it up.

It is a major effort to provide the necessary LRs and tools, and this is the reason the term 'cooperative effort' has been used. Cooperation should take place between countries, between structures in different continents, between universities, between universities and industry, and between individuals.

9. The NEMLAR partners

- *Co-ordinator:* CST – University of Copenhagen, Denmark
- *Technical manager*: ELDA – Evaluations and Language resources Distribution Agency, France
- RDI – The Engineering Company for Computer Systems Development, Egypt
- Université Lumière Lyon 2, France
- CEA -Laboratoire d'ingénierie de l'information multimédia multilingue, France
- Centre Nationale de la Recherche Scientifique, France
- Institute for Language and Speech Processing, Greece
- Amman University - Faculty of Information Technology, Jordan
- University of Balamand, Lebanon
- University of Mohammed V Soussi - Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Morocco
- Universiteit Utrecht, The Netherlands
- IT.COM – Information Technology - Société Tunisienne d'Entreprises de Télécommunications – Information Technology, Tunisia
- The Open University – Computing Department, Maths & Computing Faculty, United Kingdom
- Birzeit University, West Bank & Gaza Strip

ISBN 87-90708-15-6

The NEMLAR project is supported by the INCO-MED programme