Cooperation Roadmap

Short version

Bente Maegaard
Mustafa Yaseen, Steven Krauwer, Khalid Choukri

April 2010

Table of contents

ISBN 978-87-90708-17-7

Website: www.medar.info
e-mail: nemlar@hum.ku.dk

MEDAR is 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.

1. Executive summary

This document describes a proposal for a Cooperation Roadmap with the purpose of building sustainable Human Language Technologies for the Arabic language within and outside the Arabic world. The present booklet is a very short version of the full Cooperation Roadmap which can be downloaded from the project website.

The roadmap aspires to address a new perspective on collaboration between the Arabic region and the European Union. In order to do so, the MEDAR consortium adopted a multi-dimensional roadmap that combines various impacting factors helping to derive a coherent view. Such factors are related to the state of players, human resources and education curricula, technology development and R&D, evolution of the e-infrastructures (in particular mobile and Internet penetration, attractiveness of ICT environments, growth of e-content in Arabic). Another dimension that is considered is the market: both the domestic and international ones are reviewed and the market profile is analysed (in particular products versus services).

Last but not least, a set of instruments are elaborated upon with the aim to boost cooperation between universities and industries (both within the region but also with the EU and the West in general), to improve the technology transfer (from local R&D players to local/international industries).

The following summarises those components and directions that will lead into success of the strategy:

- Universities and research centres should provide the basic and applied research in cooperation with industry to produce solid products,
- Universities and other educational institutions should create the proper training and re-training (rehabilitation program for personnel from other disciplines who could be re-trained to fit the new requirements)
- Governments and funding agencies should facilitate, support and help companies and universities to initiate and sustain their products,
- Specialized companies should play a significant role in this area and should build and enhance tools, utilities and applications for Arabic HLT,
- Governments should launch services/applications for citizens (e.g. e-government sub-projects, initiatives) that will be accessed and navigated in Arabic language
- International companies specialized and interested in the HLT (and Arabic HLT) in particular should be encouraged:
    - to maintain the interest in Arabic,
    - to providing services to the region and should be given the facilities to make this attractive for them,
    - to maintain relationships with local companies and task forces, and
    - to utilize what is available locally
- Local mobile companies, internet service providers and telephone companies should provide the support and encourage the local companies and universities to direct their efforts towards producing tools and utilities that could be integrated and added to the provided services.

Please comment!

We invite comments and proposals from all prospective stakeholders for all aspects of this roadmap document (to the email address at page 2 of this document, or to any of the consortium partners).

2. Purpose of MEDAR and of this roadmap

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 the "digital" content available in Arabic, the use of Arabic will grow. 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.

The goal of the **MEDAR** project[1], supported by the European Commission ICT programme, is to establish a network of partner centres of best practice in Arabic dedicated to promoting Arabic HLT (Human Language Technologies). The tasks of the project include surveying present language resource needs, organizing a conference, disseminating information on Arabic language technology, establishing development priorities and creating a proposal for a Cooperation Roadmap for the region. Although the project has a special focus on machine translation and other multilingual tools, the roadmap is directed towards the Arabic HLT in general, where it addresses several areas of interest in the domain; it takes into account relevance, importance, impact and potential for applications and developments in the various areas of ICT in general, and the upcoming of Global Information Infrastructure (GII) and the Information Superhighways (digital communication systems) in particular. The basic building blocks of the GII include: communication facilities, computing technologies, software interfaces/applications and standards tying together facilities, terminals and applications; services (i.e. information, electronic commerce, applications and content) available on these networks; where one of them that provides a base function in the environment is the human-machine interface where the HLT has a central role. If we consider the 90's term Information Superhighway, it was based on providing and dealing with digital electronic content, and language technology is very critical and a "must have" in such an environment.

MEDAR partners have collected knowledge about existing language resources, players, products, most recently initiated projects related to Arabic e-content by governmental, local, business institutions and/or international organizations, etc., and based on this knowledge and additional research, the roadmap has been compiled.

It is the purpose of the roadmap to outline areas and priorities for collaboration.

The full version of the proposed roadmap (prepared in spring 2009) is available from www.medar.info.

3. Introduction to the MEDAR Roadmap

We can define a roadmap as *"a document that indicates directions for a planned journey, that shows how and in what order goals can be reached and that indicates distances"*.

As mentioned above, it is the purpose of the roadmap to outline areas and priorities for collaboration, in terms of collaboration between EU countries and Arabic speaking countries, as well as cooperation in general: between countries, between universities, and last but not least between universities and industry. This cooperation should lead to a stronger Arabic HLT community, more technology for Arabic and more

[1] MEDAR is building on the results of the earlier NEMLAR project and has decided to keep the NEMLAR logo.

products on the market. Our primary focus has been on multilingual tools, in particular on machine translation and multilingual information retrieval, but other areas are mentioned as well.

Usually one focuses either on a roadmap as reflecting expected “technology developments and trends” (technology roadmap) or as “time to market“ for a new product (market dimension). In our case we have added a new and essential dimension, which is the cooperation between Arabic and European Union countries (cooperation roadmap). So, the Cooperation Roadmap in a sense consists of three interconnected roadmaps although we do not develop each of them independently, but rather take aspects from all of them into consideration, of course with a strong focus on cooperation.

Until now, we have seen that much cooperation between European and Arabic partners is based on third party incentives (e.g. the European ICT programmes). Some other initiatives are conducted in cooperation with the US. In order to initiate cooperation and in order to support a first relation-building, it is excellent that such incentives exist, and we would need more of this type of support in the future, but at the same time we also face a challenge: to turn these partnerships into *strategic partnerships*, i.e. long term partnerships based on mutual benefit.

Taking into account the three dimensions listed above, we provide in the roadmap report an analysis and report on the present situation in the participants’ countries, we describe the conditions that need to be fulfilled in order to arrive at particular key achievements and at some strategic partnerships and we describe the steps that need to be taken to get there from where we stand. It is very likely that some of these actions (or similar ones) are being implemented by national ICT efforts and through national roadmaps that are emerging in the region, but we had no evidence at the time of writing. We do hope to consolidate this report with valuable input from other official sources within the Arabic countries.

The proposed roadmap resulting from the analysis, is given below, preceded by a few examples from the analysis, whereas we refer to the full version for the full analysis part.

4. A few examples from the analysis phase of the roadmap work

Education

In the universe of HLT many different players fulfil their tasks in the long chain from research idea to end user product or service. In this section we will highlight a few of these players, which can be both organisations and humans.

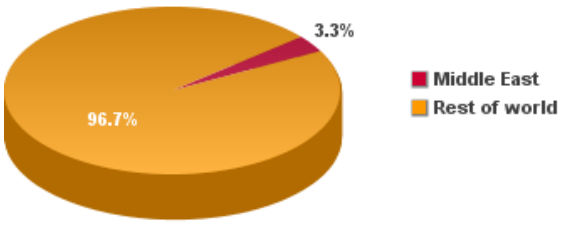
The MEDAR Survey (http://www.medar.info/MEDAR_Survey_I.pdf) shows that the number of Arabic HLT professionals is very low, and by no means sufficient to maintain (or even start building) a strong HLT industry in the Arab states. The Arab states however, do educate general ICT professionals, and here we want to focus on the special, additional skills required to build HLT for Arabic, which include both knowledge about Arabic language and linguistics, knowledge about language and speech processing, machine learning techniques, signal processing, statistics, cognitive sciences, for linguists the capability to communicate and collaborate with software engineers, and for software engineers the capability to communicate with linguists. If we want to increase the number of people with the required skills we have to look at the opportunities for the education of a new generation of researchers and developers with adequate skills in HLT.

The main players in the education system are universities and other institutes for higher education. The number of institutions offering HLT education is minimal, and our recent survey confirmed this. In our view the education system should aim at providing HLT training both to students who want to graduate from university and to professionals who are already working in the ICT field but who lack specific knowledge about HLT and language in general. In addition to that there is also a need to train people to become HLT educators, as this is necessary for a sustainable supply of *HLT-enabled* professionals.

We use the term *HLT-enabled* since we do not believe that the main goal of the education system should be to create a completely new HLT discipline with its own professionals, but rather the *suggestion* is to provide people who have a firm basis in one of the fields relevant for the advancement of HLT with an additional component of knowledge and skills that allow them to use their specific skills to contribute to the development of HLT related products or services. *Typical examples* would be courses in e.g. linguistics, phonetics, language or speech processing offered to software engineers, or (vice versa) courses in e.g. computing, machine learning, language or speech processing offered to linguists or to students in these fields.

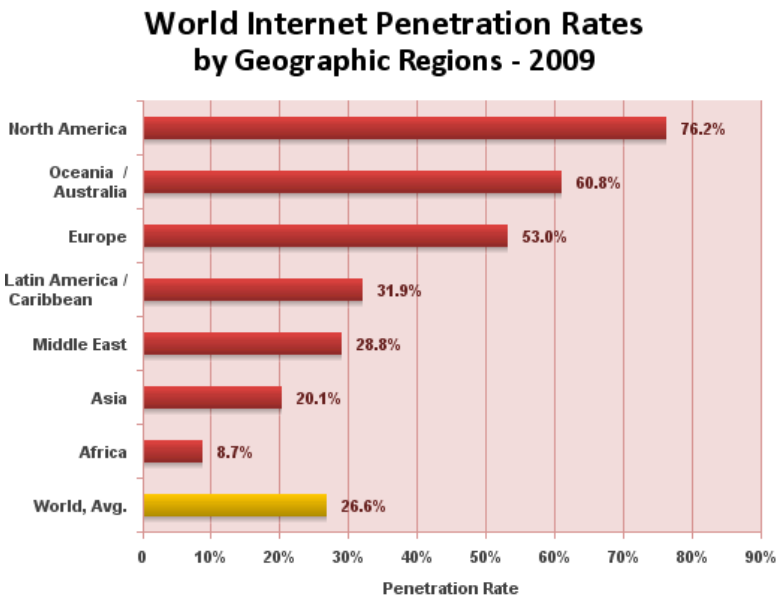
Internet usage in Arab countries 2009

Arabic is one of the top ten languages spoken in the world, yet the Arab content in the web is not encouraging, while penetration is among the fastest growing in the world, and percentage of population penetration is about 29%, compared with e.g. 53% for Europe. As reported in the Internet World Stats Report May 25, 2008[2], "Arabic is seventh place in the Internet. It should be pointed out that the number of Internet users in the Arab countries has been growing at a fast pace lately. Correspondingly, Arabic is now among the top ten Internet languages". It is well known that the Arabic speaking users are divided into two groups: a good percentage are young and they are mainly using the net for entertainment (chat, music, dating, … etc.); for the other more serious user group English or French (depending on the education) is the main language of usage. If the region is to get more people online, and involving other ages' categories and other diverse backgrounds then it must attract them to the Internet through awareness, skills development and Arabic content creation. Researchers, universities, policy-makers, telecom operators and IT companies working in this domain have a critical role to Arabize the Internet.



Internet Users in Middle East
Source: Internet World Stats – www.internetworldstats.com

[2] www.internetworldstats.com



Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 6,767,805,208 and 1,802,330,457 estimated Internet users for December 31, 2010.

5. *The proposed Cooperation Roadmap*

In this section we present the proposed cooperation roadmap. The roadmap does contain actions on aspects on which the consortium partners and similar actors in the countries have no power, but which are a prerequisite for the other actions to succeed. These actions are for governments, funding agencies etc. to decide on. They have been retained in the roadmap because they are part of the overall plan which lies behind the roadmap.

The scenario is applicable to all Arab states represented in the consortium, with partially overlapping phases; we hope it can be extended to include all other countries in the region, and that cooperation, inspired by some of these suggestions will start soon. The roadmap until year 2015 is divided into three phases, with a small overlap between phases 2 and 3. In each phase the actions/activities are classified under the main headings: Political/Policy, Training and Research, and Industry Development.

Phase 1 (2010-2012): Laying the foundations

Political/Policy:

- Organise annual round-table meetings between EU and Arab state players and funders, participants in HLT related actions and in funded projects and grants, and initiate collaboration platform to organise knowledge transfer and take-up.
- Develop e-Government projects; at least few countries should be investigating the possibilities. 50-60% reduction on Internet rates to encourage its use, collaboration scenarios for dealing with illiteracy through linguistic support functions (e.g. reading aloud)

and eLearning on the mobile phone. More enforcements to copyright and IP laws. Tougher legislations to protect software producers from illegal copying.

- Associating and linking the organizations that initiated the projects and initiatives for enrichment of Arabic e-content in Arab countries with major regional and international players identified by the surveys conducted by the consortium. Consultancy services by consortium expertise to HLT related activities in the region could be offered.

Training and Research:

- Develop initial HLT-enabling curricula, training material, faculty members exchange programs, Arab students placements programs in close collaboration between universities in EU and Arab countries.
- Identify high priority essential *BLARK*[3] components needed for training. Strengthening of participation of players from the Arab states in Networks of Excellence in fields related to multilingual HLT or HLT in general.
- Develop methodologies to fight illiteracy through linguistic support functions.

Industry Development:

- Development of MT systems in Egypt, text analysis tools in Jordan, LRs in Egypt and Morocco, and other prototypes or mature products from partners and others identified by the consortium, also other countries have started efforts in Arabic HLT and this will continue to grow. At least one product on the market in this phase.

Phase 2 (2012-2014): Moving forward
Political/Policy:

- Development of schemes of cooperation between the Arab HLT players to be more competitive globally, and continuation of adaptation of copyrights and IP laws.

Training and Research:

- Implement first HLT-enabling curricula in participating Arab states. Teaching staff exchanges, grants for Arab students, Arab student traineeships, and continuation of participation in special grants in EU projects.
- Develop essential BLARK components including tools and LRs.
- Research should support the following areas of application: automatic speech recognition technologies for dictation, language learning, text to speech synthesis in local colloquial, MT for tourism industry, better Arabic search engines, bilingual editing software with grammar checkers, spell checkers, dictation machines (domain specific), translation memories, etc.

Industry Development:

- Focus on: E-Governance and content creation for the general public, and applications via mobile phones for the illiterate, Infotainment services and access to entertainment services. Access to educational services in particular for language learning. At least one application per area and per year should be created in collaboration with universities.
- Development of MT systems, text analysis tools, LRs, search engines. At least one product from each country on the market.
- Bilingual tools: need more enhancements and focus should be on niches, i.e. be domain specific.

[3] BLARK is defined as: The minimal resource kit needed to develop basic R&D activities in language technologies

Phase 3 (2013-2015): First consolidation

Political/Policy:

- Local players can penetrate export markets cooperating with their counterparts internationally.
- The dimension of using and developing Arabic HLT supporting utilities will be a serious growth factor if national agencies consider the use of HLT products and services in the literacy programs. Such huge market will then attract major players. Efforts to obtain this should be made.
- Cooperation between local software industries and international players can be seen here in terms of outsourcing models where local industries could participate with the international players in implementing products in their local markets.
- Cooperation could be seen also between software developers (companies) with initiatives' and projects related to areas where Arabic HLT is core.
- Successful e-Government implementation will be undertaken by some Arab countries.
- Create facilities for HLT entrepreneurship (universities, SMEs).

Training and Research:

- Implementation of improved curricula on the basis of experience gained and new technological developments.
- Regular student and staff exchanges between Arab states and EU, and between academia and industry; and some joint projects and training activities across Arab states.
- Further development of the BLARK and creation of application or domain specific resources and tools for priority areas.
- Joint RTD projects between EU and Arab players to build new applications and services, especially related to multi-linguality.

Industry Development:

- Many small companies will emerge, we expect at least 2 new businesses related to Arabic HLT per university per country per year within this phase.
- By 2015 at least one project should have resulted in an educational product for illiterates on the Arab market.
- By 2015 we expect improvement on the overall books production by utilizing utilities and tools with the aid of MT and other language technology products, raising the average Arabic books production to 5000 books/year.

Cooperation across sectors:

- Arabic HLT will be utilized more and will be incorporated more efficiently in many areas of applications. The cooperation will be seen in the following:
    - Institutional consumers (e-government, e-health, e-… agencies) could adopt the HLT products and provide web-services for citizens.
    - More R&D activities will take place to enhance the products and tools started in Phase I & II, and to come up with e.g. advanced search engines, speech enabled services, multilingual tools to support cultural and tourism sectors, etc.
    - Start adding languages to enrich the multilingual dimension on both the applications and tools/utilities levels.
    - MT prototypes should develop to more mature products especially considering closed domains for specific applications.

- More enhanced applications and utilities: Text to Speech applications could be very much useful in this area, since public need to access such services, and given the high degree of illiteracy among citizens in some countries.

6. Conclusions and future actions

From the survey of the education systems we can conclude that the academic institutions in the Arab world do not provide enough academic programmes or training courses to ensure the competency needed for the technical tasks of industry. The recommendation for this shortcoming is to develop programmes and curricula for Arabic HLT as mentioned above. This can be done in win-win partnership with international institutions from Europe, USA but also across Arabic countries. It is important to also focus research at the universities towards HLT in collaboration with European and other international partners.

Throughout this roadmap, research is supposed to support all the actions, be it education or industrial development. It is important for the Arab region to have researchers at a very high level so that they can participate in competitive research programmes. It is also important that the universities collaborate to build the BLARK for Arabic, and in general take a lead in promoting Arabic HLT and making it possible.

From the technology and market perspectives, we can conclude that as a result of the spread of ICT in general, more penetration, and consequently more user needs will attract the international major players to this area and emerging, promising market. Major companies such as the mobile and telecommunication companies will have needs to enhance their services by facilitating the utilization of the services in the Arabic language. Major international applications and software providers and manufacturers need to add tools and utilities to their products, such as Microsoft and Google; who are basically dealing with language aspects and always in need to enhance their services with professional, not superficial tools and utilities. Their focus is directed towards utilities for the enhancement of Arabic processing (spell checkers, morphological analyzers, syntactic analyzers, lexicons, search engines based on Arabic main features, etc).

Two scenarios could be seen in this case:

1) The major international players will dominate the market, develop technologies and enforce their vision, methodologies, procedures, and as a result monopolize the whole industry.
2) Local efforts will be undertaken, governments and major funding agents will encourage and incubate such activities, giving the local companies opportunities to grow and develop its own; build the capacity in local forces; and build and maintain national industry that, as a result, could be competitive on the international level.

What we recommend is a hybrid solution between the two scenarios; the Arabic HLT should get the benefit out of the huge interest, huge allocation of investments, and immediate need for the services and applications in the ICT in general, and in related applications to Arabic HLT in particular (search engines, mobile services, e-commerce, e-government, e-learning, etc). The cooperation roadmap is focussing on promoting this view for the benefit of the Arabic speaking societies, the research communities and universities and the industry.

The MEDAR consortium

- **University of Copenhagen:** Centre for Language Technology, Denmark (Coordinator)
- **ELDA:** Evaluations and Language resources Distribution Agency, France
- **University of Balamand:** Research Council - Speech and Image Research Group (SIR), Lebanon
- **Amman University:** Faculty of Information Technology, Jordan
- **University of Utrecht:** Utrecht Institute of Linguistics OTS, the Netherlands
- **Research and Innovation Centre "Athena":** ILSP, Institute for Language and Speech Processing, Greece
- **RDI:** The Engineering Company for Development of Computer Systems, Egypt
- **Birzeit University:** Center for Continuing Education, West Bank and Gaza Strip
- **University Mohammed V Soussi:** Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Morocco
- **CEA, Commissariat à l'Energie Atomique:** LIST, Vision and Content Engineering Laboratory, France
- **CNRS, Centre National de la Recherche Scientifique**, Laboratoire LLACAN - UMR 8135 du CNRS, Langage, langues et cultures d'Afrique Noire, France
- **The Open University:** Computing Department, Maths & Computing Faculty, The United Kingdom
- **Université Lumière Lyon2:** Groupe SILAT, France
- **IBM:** International Business Machines WTC - Egypt Branch, Egypt
- **Sakhr:** Sakhr Software Company, Egypt

ISBN 978-87-90708-17-7

MEDAR is supported by the European Commission's ICT programme