The NEMLAR project on Arabic language resources

Bente Maegaard

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S, Denmark
bente@cst.dk

Abstract.
The NEMLAR project is a European Commission supported project with partners from the EU and from Arabic speaking countries in the Mediterranean region. The project aims at surveying the stat-of-the art- of language resources and tools for Arabic in the region, at developing a BLARK definition for Arabic, and at starting development of language resources or updating of existing language resources. The project also aims to create visibility for Arabic language technology, through a newsletter and through an international conference.

1. Motivation

There is abundant evidence that language technologies can only be developed using large bodies of language resources (LRs) for language modelling, as test beds, for evaluation, example bases, and terminology source. The need for LRs applies both for research and for commercial applications.

Not only raw data, but also ‘derived’ LRs, e.g. annotated corpora, lexica and grammars, as well as tools for manipulating data form part of the material of interest. The production of such LRs also enables the linguistic cultural heritage of a community or nation to be preserved in an age of digital access and storage.

This is the reason the NEMLAR project was started. There is a strong interest in supporting the Arabic language, in the region, in Europe and elsewhere. The project runs 2003-2005.

The NEMLAR project covers recognised European centres and recognised partners in 6 non-EU Mediterranean countries, namely Jordan, Morocco, Egypt, Lebanon, Tunisia, West Bank and Gaza Strip.

2. NEMLAR goals

The goal of the NEMLAR (Network for Euro-Mediterranean LAnguage Resources) project is to create a network of qualified Euro-Mediterranean partners to specify and support the development of high priority LRs for Arabic and other local languages in a systematic, standards-driven, collaborative learning context. The project will focus on identifying the state of the art of LRs in the region, assessing priority requirements through consultations with language industry and communication players, and establishing a basic LR kit for the major forms of the region's predominant language - Arabic, and other local wide-spoken languages where appropriate. This knowledge base has appeared in its first version (Nikkhou et al. 2004).

3. Survey: key players, language resources and industrial needs

It is a key part of this project to provide knowledge about the language technology players, projects (ongoing activities), products etc. So a 'mapping' is made covering all Mediterranean countries participating in the project, resulting in a knowledge base with details of all universities, research institutions and companies, as well as ongoing projects, and existing products, - with relation to Language Resources (LRs). This knowledge base is ready in its first instance.

It covers 35 institutional players in the region, and some 20 individual players. It will further develop during the lifetime of the project, but we believe to have identified the most important players already.

Of the 35 institutional players, 22 are based in Arabic speaking countries, obviously in particular in our partner countries. E.g. there are 8 entities in Egypt, 4 in Lebanon, 3 in Jordan and 3 in Palestine. Kuwait houses the Sakhr company with subsidiaries in many countries, incl. Egypt.
In Europe, companies such as Systran and France Télécom also take an interest in Arabic language processing.

3.1 Tools and LRs

Existing Arabic tools and LRs in the region, in Europe or elsewhere have been identified, and the first version of the survey report describes state-of-the-art of LRs for the languages of the region. It should be stressed that the survey has focused on resources in the region, not in the whole world.

Table 1: Number of tools

Arabic NLP technologies and tools	31
Speech processing technologies	11
Text processing technologies	11

Here NLP tools are modules that normally are parts of systems, e.g. morphological analyzer, POS tagger, language identifier, term finder etc. Also classified as NLP tools are research results that have not yet been commercialised, e.g. grammar checker, grapheme recognition for OCR.

It has been encouraging to see that e.g. POS taggers do exist, and not only at the universities, but also as products. Morphological analyzers exist at the universities and also as a component of commercial products, e.g. machine translation. It is important to make morphological analyzers available in a source format, so that researchers can further elaborate on the morphological analysis and can combine this analysis with other components in their efforts to gain new insights and develop ideas for new and better language modules.

Syntactic analyzers exist in some universities, and as an important part of e.g. MT systems. Overall, it seems at present that there is no large scale grammar and parser freely available for researchers. It is foreseeable that commercially developed syntactic analyzers cannot be made available, so we believe that some interest should go into investigating the existence of syntactic analyzers and the possibilities of developing them further.

Speech processing technologies cover Arabic text-to-speech, speech recognition, speaker recognition etc.

Arabic text-to-speech exists in several versions as products. Arabic text-to-speech is of good quality which can be easily compared to similar tools for other languages. It would be important to identify open source modules which can be used by researchers for further improvement and research.

At present we have identified the following amount of language resources:

Table 2: Number of language resources

Speech databases	22
Lexical databases	29
Text corpora	24
Multimodal resources	1

E.g. LDC has 15 Arabic language resources, and ELRA has 3. On the basis of this knowledge base, a survey report is written, describing state-of-the-art of LRs for the languages of the region. Industry in the Mediterranean countries and other industry working with Arabic is consulted with respect to needs for LRs for the Arabic language and/or multilingual LRs for communication for global networks. This is detailed in a second survey report which provides a record of the LR needs of industry and an analysis of missing LRs in the current situation ('LR gaps'). At the time of writing this report is work in progress.

3.2 Availability of language resources and tools

Some of the resources surveyed belong to universities or other academic institutions, others to commercial companies.

As several companies produce products in the field of Arabic language technology, e.g MT, speech technology etc., LRs do exist within these companies. Such resources are e.g. large corpora, lexica, morphological components, speech corpora etc. However, such basic resources are normally not available to others. It is a well-known fact, and not specific for these companies that the LRs developed have been expensive and constitute a competitive advantage that companies can usually not share with others.

On the other hand, resources developed by universities should be more easily available. Here however, another problem presents itself: Often such resources have been developed for a specific purpose and therefore the resource does not have a general value. Or it was developed within a specific project, and when the project stopped, no time was left to provide it the additional effort that would raise its value form a project specific resource to a general resource.

It is one of the aims of this project to identify such resources and to provide the extra small amount of effort that will make these resources valuable to a larger audience.

According to the present version of the survey, the situation wrt. lexica seems to be positive, in particular lexica with morpho-syntactic information.

3.3 Distribution of LRs

Over half of the interviewed institutions and experts wish to make their resources available to others according to a negotiated standardised distribution agreement.

Only 19% said they do not want to distribute their resources and this due to legal (9%), commercial (6%) and strategic (4%) reasons. This is encouraging, and we believe that the efforts spent in the project to identify the resources are very well spent. Some of the resources may be fully ready for distribution, others after a slight updating. The list is made available at the NEMLAR web site, www.nemlar.org.

4. BLARK

In parallel with the survey, work is ongoing to specify the Basic Language Resource Kit for Arabic. The BLARK constitutes what is seen as the minimum requirements with respect to language resources in order to be able to develop language technology, incl. translation tools. The BLARK concept has not been developed for Arabic before, and it is interesting to note the differences in comparison with other languages. E.g. for Dutch for which the BLARK concept was first developed, a diacritizer tool (for vowelisation) is not relevant or necessary, but for Arabic it is.

The surveys mentioned above provide input about what is already available, and where there are gaps, or resources that have to be updated and improved in order to fit the specifications. Consequently, we have the necessary basis for detailed work on updating or creating languages resources for the Arabic language.

4.1 BLARK concept

The BLARK definiition is in principle intended to be language independent, but as specific languages may come with different requirements, instantiations of the BLARK may vary in some respects from language to language. A BLARK comprises many different items such as:

Basic language resources:

- written language corpora
- spoken language corpora
- bilingual (written) corpora (comparable, parallel, aligned, ...)
- mono- and bilingual dictionaries
- terminology collections
- grammars (i.e. formal standard rule sets such as; a Syntactic Grammar, a Phonetic Grammar, a Lexical Grammar, …)
- Benchmarks for evaluation

Basic tools:

- modules (e.g. taggers, morphological analyzer, parsers, speech front-ends, grapheme-to-phoneme converters, statistical disambiguators, …)
- annotation standards (or best/common practice usage) and tools
- corpus exploration and exploitation tools
- etc

This list is far from exhaustive but serves to illustrate the scope of the BLARK. In addition it should first consider partnering with existing infrastructures for the management, maintenance and distribution of the resources. A BLARK should not be seen as a static object: over time. It may gradually evolve as new technologies and application areas emerge, with new requirements in terms of resources. See Binnenpoorte et al 2002.

The underlying idea is to make a common generic BLARK definition, applicable in principle to all languages, based on the collective experience and expertise gained with many different languages by the members of the language and speech technology community at large. This common definition will save time and effort, it will allow for porting of knowledge between languages, it will ensure interoperability and interconnectivity (especially for multilingual or cross-lingual application areas), and it will help making realistic estimates of costs and efforts required to produce them. In addition a broadly supported common definition may be used as an external reference point in discussions with funding agencies about the best way to create a good starting point for language and speech technology, both in academic and industrial research.

5. Work on language resources, update and production

Following the surveys of existing LRs and LR needs, as well as the BLARK specifications, the project will decide on priority needs for LR update or development, and develop a work plan for this work. The plan will also take into account the available human resources. The work plan will specify projects/pilot projects for project partners. Such pilot projects may concern the updating of existing resources (e.g. change of format, change of standards, validation and updating of existing LRs etc.). They may also concern collaboration with ongoing projects in order to ensure that the specifications are met. The development of LRs from scratch will be considered only to a very small extent in this project, given the amount of resources this requires.

5.1 Dissemination

The objectives of this project are on the one hand the technical work with the surveys, the BLARK specifications and the language resources and tools. On the other hand, dissemination of the knowledge that has been acquired, and awareness about Human Language Technology in general is also a key objective.

The project web site, www.nemlar.org, and the quarterly newsletter contribute to the general awareness raising.

Additionally, an international conference will be held in September 2004 in order to disseminate the surveys and the specifications, the industrial needs, and research in the field of LRs and tools for Arabic.

Acknowledgements

This project is supported by the European Commission.

The author is the co-ordinator of the project and it is important for me to acknowledge the contribution of all the partners, in particular Khalid Choukri, ELDA, France, Steven Krauwer, Utrecht University, Netherlands, and Muhammad Atiyya, RDI, Egypt.

The NEMLAR partners that contributed through their work are:

- Lise Damsgaard Jørgensen. Center for Sprogteknologi, University of Copenhagen, Denmark
- Mahtab Nikkhou, ELDA European Language Resources Distribution Agency, France
- Joseph Dichy, Université Lumière Lyon 2 - Faculté des Langues, France
- Christian Fluhr, CEA - LIST/DTSI/SRSI/Laboratoire d´ingénierie de l´information multimédia multi-lingue, France
- Fahti Débili, CNRS - Délégation Rhône-Alpes, Site Vallée du Rhône, France
- Stelios Piperidis, Institute for Language and Speech Processing, Greece
- Mustafa Yaseen, Amman University, Faculty of Information Technology, Jordan
- Chafik Mokbel, University of Balamand, Lebanon
- Abdelhak Mouradi, ENSIAS, University of Mohammed V Soussi, Ecole Nationale Supérieur d´informatique et d´analyse des Systèmes, Morocco
- Salem Ghazali, SOTETEL-IT - Société Tunisienne d´Entreprises de Télécommunications – Information Technology, Tunisia
- Anne DeRoeck The Open University, Computing Department, Maths & Computing Faculty, United Kingdom
- Mahdi Arar and Kanan Al-Ali, Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank and Gaza Strip

References

Atiyya, M. (2000), *A Large-Scale Computational Processor of The Arabic Morphology, and Applications*, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University.

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, H. Strik, C. Cucchinari (2002) A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, In: *Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation), Las Palmas de Gran Canaria, Spain*

Fersøe, H. (2004): *Validation Manual for Lexica*, ELRA, Paris

Hamza, W.M. (2000), *A Large Database Concatenative Approach for Arabic Speech Synthesis*, PhD. thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University.

Krauwer, S., M. Atiyya, K. Choukri, B. Maegaard (2004): *BLARK for Arabic*, NEMLAR report, www.nemlar.org

Monachini, M., F. Bertagna, N. Calzolari, N. Underwood, C. Navarretta (2003): *Towards a Standard for the Creation of Lexica*, ELRA, Paris

Nikkhou, M., K. Choukri (2004): *Survey on the existing institutions and Language Resource using or developing Arabic,* NEMLAR report, www.nemlar.org.

Van den Heuvel, H., Louis Boves, Eric Sanders (2000): *Validation of Content and Quality of Existing SLR: Overview and Methodology*, ELRA, Paris.