NEMLAR – Network for Euro-Mediterranean Language Resources – Annual Report 2003

http://www.nemlar.org

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 Language Resources (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 protocol for developing a basic LR kit for the major forms of the region's predominant language – Arabic, and other local wide-spoken languages where appropriate.

Language Resources are collections of text and spoken language, dictionaries, grammars etc. Language Resources are basic building blocks for the development of language technology for a language.

The project team comprises 14 partners, to be found at the end of this report.

Summary of 2003 Activities

The project started on February 1st 2003, and the main part of the project activities has been initiated during this first year.

The project has concentrated on 5 main areas:

- Surveying European and Arabic institutions, companies, products, tools, resources, projects and funding agencies within the area of Arabic Language Resources (LR) and Human Language Technologies (HLT), and describing a selection of the surveyed language resources in detail
- Identifying the industrial needs of LR for Arabic
- Producing a **B**asic **La**nguage **R**esource **K**it (BLARK) definition and specification for Arabic
- Preparing an internal workshop to discuss and further elaborate the results of the surveys and the BLARK
- Starting preparations for the international conference on Arabic Language Resources which the project will organise.

Important work areas

- In order to carry out the survey of the European and Arabic institutions, companies, products, tools, resources, projects and funding agencies within the area of Language Resources (LR) and Human Language Technologies (HLT) a set of survey questionnaires were created – one questionnaire for individual persons and one for institutions and companies, for more information see http://www.nemlar.org/Survey-questionnaires .
  The surveys were carried out by all partners of the project (Tunisia, Morocco, Egypt, Lebanon, Jordan, The West Bank and the Gaza Strip, France, Denmark, the Netherlands, the United Kingdom, and Greece).
  We received surveys from 35 institutions and 19 individuals and more than 90 LRs including 26 speech databases, 31 lexicons and dictionaries, 26 text corpora and one multimodal database have been identified.
  As a first outcome of the surveys two lists have been produced showing a) the identified Arabic language resources b) institutions and companies involved in the development of Arabic language resources. For more information see http://www.nemlar.org/Publications . As a second outcome a comprehensive report will be published giving a map of the situation of the Arabic stakeholders, projects, existing resources and tools for Arabic.
  The questionnaires are still available, and new input will be taken into account continuously and result in regular updates of the report on the web site.

- Along with the elaboration of the 'map' of the European and Mediterranean situation mentioned above, a survey on the Industrial needs for language resources for Arabic is also being carried out. A number of companies have been asked about their needs: Scansoft (Language and speech technologies, USA), CEA (Written, France), Systran (MT, France), Cimos (MT, France), ELAN (TTS, France), Aratext (Written, Egypt), IBM (Speech, Egypt), Sakhr (Speech, Written, NLP, Egypt), RDI (Language and Speech technologies). Information from a few more companies is expected in spring 2004 before the report will be published.

- Based on the available results from the surveys on LRs and industrial needs, the work on a **B**asic **LA**nguage **R**esource **K**it (BLARK) for Arabic is being conducted.
  The BLARK is defined as the minimal set of language resources that is necessary for doing any pre-competitive research and education in language technology. The definition is in principle intended to be language independent, but as specific languages come with slightly different requirements, instantiations of the BLARK will vary in some respects from language to language, and the Arabic BLARK is not exactly the same as the Dutch BLARK (which was the first to be made). A BLARK comprises many different items, such as:
  Basic language resources:
    - written and 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 not exhaustive, but serves to illustrate the scope of the BLARK.

- One of the most important ways for the project to convey its messages is via the web. The project has therefore acquired the domain www.nemlar.org and established a project web site. Furthermore, a logo has been designed by the Jordanian partner. The logo is used on the web site, in PowerPoint presentations and on any other material with relation to the project. A PowerPoint template has been created, to be used by all project participants when making presentations at conferences etc.
  Another way of raising awareness about the project and its work is by publishing a quarterly newsletter informing about the mission of the project, its progress, events related to Arabic language technology, LRs and tools, books, software, papers etc. all in the scope of the project.
  In 2003 the newsletter appeared three times. The newsletters may be found at http://www.nemlar.org/Newsletter .
  By the end of 2003 the project had 52 subscribers to the newsletter, but the goal is to reach 100 subscribers before the end of the project. Subscriptions may be made by sending an email to the co-ordinator nemlar@cst.dk .
  Finally, a small folder presenting the project is being prepared.

Promotion and Awareness

Promotion and awareness are key activities for the project. In particular the knowledge acquired through the surveys is of interest to all those involved in Arabic language technology or computational linguistics, and the BLARK definition and specification for Arabic is also expected to attract a good deal of interest. Apart from conveying these findings through the web and the newsletter, this information may be disseminated in various other ways – e.g. through conference papers and articles in magazines.

In 2003 it was still too early to present the project at conferences, but in 2004 the project will be presented at least at 5 conferences, of which 3 in Europe (Malta, Portugal, Switzerland) and 2 in Arabic speaking countries (Morocco and Egypt).

The magazine Multilingual will bring an article on NEMLAR.

The most important promotional activity will be the international conference to be held in Egypt in September 2004, see below.

Future Work

The first part of 2004 will be devoted to finalising the survey report, the report of the Industrial needs and the description of the BLARK for Arabic.
The project will organise an internal workshop where the results of the survey and the BLARK are to be discussed and further elaborated.

Based on these results, decisions will be made on priority needs for Arabic Language Resources update and development. This work will include choosing which LRs among the surveyed ones to be updated, deciding in which way the updates should be carried out (e.g. change of format, change of standards, validation and updating existing LRs etc.)

In the second half of 2004 the project will arrange an international conference on Language Resources for Arabic. The preparations have already started. The venue has been chosen for Cairo, Egypt and the date is settled for September 22-23, 2004. The Egyptian partner, The Engineering Company for computer systems development – RDI, is the local organiser and a programme committee and a scientific committee are being set up.

Further Information

The Nemlar project web site:
http://www.nemlar.org

The Survey questionnaires:
http://www.nemlar.org/Survey-questionnaires

The project publications:
http://www.nemlar.org/Publications

The conference web site:
http://www.nemlar.org

Newsletter:
http://www.nemlar.org/Newsletter

Subscription to newsletter:
nemlar@cst.dk

The NEMLAR Partners

- Center for Sprogteknologi, University of Copenhagen, Denmark
- ELDA Evaluation and Language resources Distribution Agency, France
- RDI, The Engineering company for computer systems development, Egypt
- Université Lumière Lyon 2 - Faculté des Langues, France
- CEA - LIST/DTSI/SRSI/Laboratoire d´ingénierie de l´information multimédia multi-lingue, France
- CNRS - Délégation Rhône-Alpes, Site Vallée du Rhône, France
- Institute for Language and Speech Processing, Greece
- Amman University, Faculty of Information Technology, Jordan
- University of Balamand, Lebanon
- ENSIAS, University of Mohammed V Soussi, Ecole Nationale Supérieur d´informatique et d´analyse des Systèmes, Morocco

- Universiteit Utrecht, The Netherlands
- SOTETEL-IT - Société Tunisienne d´Entreprises de Télécommunications – Information Technology, Tunisia
- The Open University, Computing Department, Maths & Computing Faculty, United Kingdom
- Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank and Gaza Strip