MEDAR

Mediterranean Arabic Language and Speech Technology

An intermediate report on the MEDAR Survey of actors, projects, products

Khalid Choukri

Evaluation and Language resources Distribution Agency; ELDA, France
E-mail: choukri@elda.org

Abstract

After the successful completion of the NEMLAR project (2003-2005) and its work on identifying the activities, players, projects, technologies and resources, within the Arabic geographic area and within institutions involved/interested in Arabic HLT, a follow up was made possible by the European Commission that decided to fund a new project called MEDAR. MEDAR addresses International Cooperation with the Arabic region on Speech and Language Technologies. One of the three core activities planned within MEDAR is to produce a knowledge base on Human Language Technology (HLT) players, existing language resources (LRs) and processing tools, activities and products for Arabic. Such activity would help achieve some of the MEDAR objectives in particular consolidating the network of players in all areas of HLT, established under the NEMLAR project, develop a Cooperation Roadmap based on a clear picture of the foreseeable technological trends, market potentials, and cooperation possibilities, and exploit such knowledge to update the Basic Language Resource Kit (BLARK, the minimum set of resources and tools necessary for carrying out research on Arabic LRs and HLT). One of the instruments that has been used to gather such knowledge is a web-based survey to identify the activities and projects, existing language resources and tools, as well as the experts. This paper summarizes the MEDAR project's findings related to the situation of Human Language Technologies (HLT) for Arabic. Such state of the art reflects the current situation in the Arabic world as well as those players identified outside this geographical region but who are involved in technology and/or Language Resources development for Arabic. This work builds on the task carried out within NEMLAR (www.nemlar.org). The Summary of the findings (facts & figures) is given herein.

.

1. Introduction to the MEDAR Survey

The goal of **MEDAR i**s to consolidate and further develop a network of qualified Euro-Mediterranean partners to specify and support the development of high priority tools and language resources for Arabic and other local languages in a systematic, standards-driven, collaborative context. Such network will be developed with a view to creating a longer term Cooperation Roadmap for the region and the Euro-Mediterranean stakeholders.

To this end, MEDAR has surveyed the state-of-the-art in the region, analyzed the language resources needs, and has established a first development priority as required by the deployment of a Machine Translation system for Arabic. Such development will comprise a minimum set of language resources that will support the linguistic diversity in the Southern and Eastern Mediterranean region for digital information and communication. The project is also working on a survey and analysis of the key strengths, weaknesses, opportunities and threats to the development of Arabic and other language resources in the region, leading to a set of key priorities for developing machine translation tools.

The exploitation of the survey results will raise awareness of the state of play of Arabic and local language resource and tools development among all stakeholders, ensuring information feeds to existing networks and information sources as well as decision and policy makers.

The survey was launched on April 2008 and all MEDAR partners were encouraged to fill in the questionnaire for their institution and having it filled by their partners.

By end October 2008, 57 questionnaires were filled in. Most of them have been entirely completed (37), some 17 questionnaires are missing some of the answers but this 54 (37 and 17) are all considered for their usefulness.

These figures are the ones we adopted to derive our statistics in the detailed tables of the following sections. A number of countries are over represented (e.g. 11 responses from Egypt, 10 from Morocco), some were not well represented and required more attention from the project partners. Many players are listed for the first time (e.g. a Syrian and a Turkish lab), compared to the NEMLAR report. We feel that the Internet-based questionnaire was more easy to use than by the past (email of word files).

An important part of the survey was related to the technologies our respondents felt important for Arabic. They listed a large number and many of them consolidate our own finding. Among the technologies listed, Machine Translation (MT), CrossLingual Information Retrieval, MultiLingual Information Retrieval (CLIR/MLIR), and Automatic Speech Recognition (ASR) were on the top. They also listed a number of crucial resources that should be better specified and defined by MEDAR in the framework of its updating of the BLARK (Basic Language Resource Kit www.blark.org).

In addition to the survey, the ELDA team also collected information about MT and CLIR/MLIR tools and products that address Arabic as one of the languages. This

is part of the survey report and also reported on in this paper.

2. The MEDAR Survey instruments

The survey was conducted by MEDAR partners using an online web questionnaire 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 tools and Language Resources (LRs), in particular for MT, information retrieval and indexing. The partners, as far as possible, attempted to contact the players to collect information about existing Arabic LRs and tools for Arabic.

In addition to the objective of updating the directory of players, resources, and tools, the survey aims at identifying for the technologies mentioned above (MT, CLIR/MLIR) what is already available, and where there are gaps, or tools or resources that have to be updated and improved in order to fit the specifications.

Consequently, this work will provide a substantial part of the necessary basis for detailed work on specifying, updating, or creating languages resources and tools for the MT and CLIR/MLIR with Arabic language as one of the components.
The MEDAR Online Survey and the questionnaire structure
In order to ensure a larger number of replies to the MEDAR survey, we opted for an online questionnaire using a web based tool for interview called Limesurvey (http://docs.limesurvey.org/), an open source survey tool that allows to set up surveys very user friendly and also to collect the information in various format that render them very easy to analyze and exploit. The tool allows also asking a question and continuing the questionnaire according to the answer received (an easy "tree" interface).

The tool was easy to customize so the respondents were presented with questions group by group. Responses were date stamped and IP Addresses have been logged (and Referrer-URL saved) for future exploitation. Participants could reply to the survey in more than one visit if they wish and the tool saved partially finished surveys.

The major challenge was to ensure that filling the questionnaire would not take more than 5mn. The questionnaire was set up on the basis of 6 groups of questions and an introduction.

This is the introduction to the questionnaire and the questions (for more details please refer to MEDAR report D2.1):

The goal of this new survey is to collect information about the existing institutions and Language Resources, and to describe the needs for language resources, etc. This task is being implemented in three phases.

The first phase aims to revise and update the data collected within NEMLAR (general information about Language Resources & Tools for HLT within the members of the NEMLAR network who contributed to the first report) This is the purpose of this first survey.

The second phase is to go beyond this first list and the basic information, contacting new institutions recommended by the partners, and also detailing the descriptions of what has been identified in the first phase, (players, products, Language Resources, needs and requirements).

The final phase will aim at drafting a comprehensive report that may serve as the basis for a work plan about the needs for multilingual resources targeting customization of Machine Translation (including speech to speech translation), Cross-lingual information Retrieval, and other speech recognition tools. The ultimate goal is to commission some work to produce a Basic Kit that would support such customization.

There are 63 questions but most of them are easy to address (simple Yes / No questions).

3. The Survey Structure

The structure of the questionnaire with the 6 groups is briefly described below.

Group 1 is the contact information (name, email, etc.) that aims at identifying the respondent; very few questions were mandatory. The final question of this group is whether the respondent is filling the questionnaire as an Independent expert/Entrepreneur or as a representative of an institution. The idea here is to collect information about the institutional aspects of the HLT R&D and Deployment and avoid a biased statement because of the many independent experts that could respond to the questionnaire.

Group 2: Information about the institution and its language technology: This is asked when the respondent acted on behalf of an organization and not as an independent expert. This group of 16 “easy” questions intends to gather information about the structure of HLT institution both in terms of human-resources devoted/assigned to Language Technologies, the sectors of activities (i.e. speech recognition or parsing written texts domains), to obtain information about the main products or services if any, and details about those addressing Arabic language(s).

Group 3: Information about your language resources:
This group (certainly the most important for our work on MEDAR) is organized in a hierarchical structure and respondent can select as many boxes as appropriate, to allow them to list all the Language resources containing Arabic with the necessary details about nature, size, etc. for a corpus of business documents, you may state it consists of 2 million words, for an Arabic-English dictionary: 50,000 entries, etc.). If the respondent chooses only one type (e.g. written resources), then only questions related to that item are selected in the following sections of the questionnaire.

Group 4: Information and input about the Market:
This group with 5 questions aimed at getting some hints about the market targeted by the respondents. It was clear from previous analysis that the best we could get would be about the players’ offers in terms of products and/or services distributed. We tried within these questions to get accurate qualified input (i.e. targeted market: domestic market, Arabic world,

International market).

Question 4.2 is about partnership between the respondent institution and other players to identify potential cooperations in particular between the Arabic world organizations. Respondent were also were also asked to supply the names of such partners but the outcome is largely lists of potential HLT players, many of them known to the team.

Within this group 4 we also tried to collect quantitative data (not only facts but also some figures) about the market size and the plans of our respondent on their investments for example in terms of LRs acquisitions and/or production. We knew that such figures will not be very accurate (our respondents are very unlikely decision makers in their institutions) and will be hard to interpret.

Group 5 is about the needs for LRs: this is a crucial group of questions for the future tasks of the project that intends to specify the gaps in terms of LRs and boost their development. It requests information about the potential expectations of the surveyed expert if he/she had to decide on which Language Resources should be made available, for which applications, and indication on the design and structure of such Language Resources.

Group 6 is a request for more contacts to ensure that the survey is widely disseminated and that the collected data is as exhaustive as possible.

4. Summary of the survey: facts and figures

Thanks to the involvement of all MEDAR partners we managed to obtain 54 exploitable responses for this survey (37 full responses, 17 responses not completely filled out but still provide good information, 3 had some information but not exploitable in our survey).

The results given below comprise the detailed number of responses to each question and the percentages are computed on the basis of the 54 responses.

Identification of the respondent:

The 57 respondent are identifiable by name, first name, etc. The types of positions reported are listed below in alphabetical order to highlight the quality of the respondent and thus the quality of the responses obtained:

Position	Nb of respondents
Assistant professor	4
Associate Professor	2
Associate Research Scientist	1
CEO	6
Consultant of Human Language Technologies	1
Dean	2
Deputy Director	2
Director	3
Founder & Chief Scientist	1
General Manager	3
Head of Department	1
Human Language Technologies Group Manager	1
IT Instructor	1
Lab Technician	1
Laboratory Head	1
Lecturer of English & Arabic	1
Linguist	1
PhD student	2
Professor	9
Project Manager	1
Research Assistant	2
Researcher	3
Student	1
Teacher researcher	1

The countries from which originated the replies are given by this diagram:

Answer	Count	Percentage
Egypt (EGY)	11	20.37%
Morocco (MAR)	10	18.52%
West Bank & Gaza Strip (PSE)	9	16.67%
Jordan (JOR)	3	5.56%
Czech Republic (CZE)	2	3.70%
France (FRA)	2	3.70%
Lebanon (LBN)	2	3.70%
Saudi Arabia (SAU)	2	3.70%
Spain (ESP)	2	3.70%
United States of America (USA)	2	3.70%
Algeria (DZA)	1	1.85%
Israel (ISR)	1	1.85%
Japan (JPN)	1	1.85%
Syrian Arab Republic (SYR)	1	1.85%
Turkey (TUR)	1	1.85%
No answer	4	7.41%

This item has to be interpreted considering the other items (topics of interest, position, etc.) in order to balance the fact that some countries are over represented. Some experts are interested by language technologies and assume they would incorporate some in their own business but they do not claim to be active players.

As indicated above, the profiles of the respondent were collected to ensure that we can distinguish independent experts from institutions and also their involvement in HLT & LRs.

Answer	Count	Percentage
No answer	1	1.85%
Independent expert or Entrepreneur	17	31.48%
Institution	36	66.67%

A large number of organizations were part of the respondents and many others were identified and indicated to us by the consortium members which allowed collecting a reasonable-size directory of HLT players.

Another aspect is the profile of the identified institutions. Those who indicated the type of institution they work for listed the following:

Type of institution	Count
Company & for profit organization	15
University	12
Public Research center	7
Public organization	0

As we can see a large number is labeled as "company" or "for profit" entity.

Regarding the manpower available for these activities and/or the size of these organizations we did ask for the total number of employees within these organizations:

Total Number of employees of the respondent organization	Count
Less than 10	5
10-49	9
50-99	8
Over 100	9

and also asked to "estimate" the number of those who are involved in HLT:

Number of employees (directly or indirectly) involved in language technologies	Count
Less than 10	18
10-49	11
50-99	1
Over 100	1

The respondents' activities and involvement in HLT

The main activity of these institutions (respondents could choose more than one choice) were itemized as follows (without any priority or rank):

Answer	Count
Software developer	12
Teaching/training organization (e.g. university)	16
HLT Product Vendor	4
Technology Transfer institution	9
Content provider	3
Interpreting/Translating/Localization	6
Telecommunications	2
E-commerce	2
Other	9

Those who responded positively to the question on their involvement in HLT indicated the following areas (more than one answer):

Answer	Count
Language learning	6
Language Resources production	13
Speech technologies	11
Written technologies	14
Search and knowledge mining	13
Translation automation	7
Other (i.e. Language Resources)	1

This section allowed us to gather information about a list of technologies as they were mentioned by the respondents. In addition, the main products and sectors of activities were indicated by the different players and are listed in detail in the full report on this survey (Deliverable D31 of the MEDAR project).

Another important issue is the Monolingual vs. Multilingual feature of the products offered by the respondent. When asked about their offer, we obtained the following responses only (out of the 57):

	Count
Monolingual Products or Services	13
Multilingual Products or Services	22

And when they stated that their offers are multilingual, the respondents were asked if it includes the Arabic language:

Yes	28
No	0
Non completed	13
No answer	13

To the question on the Language Resources types used by the respondents, we received the following answers:

Answer	**Count**
Speech Resources	19
Written Resources	28
Multimedia/multimodal Resources	10
Other (e.g. biometric data)	2

When the answer was "Speech Resources" then respondents explicitly referred to:

Answer	Count
Broadcast news & conversational speech	13
Fixed telephone	7
Mobile telephone	9
Micro/desktop speech	8
In-car recording	7
Read newspaper texts	9
Pronunciation/phonetic lexica	8
Other	2

And when the answer was “Written Resources” then we got:

Answer	Count
Lexical databases	20
Terminology & specialized dictionaries	11
Text Corpora	20

If the answer was "Lexical databases", then we got:

Answer	Count
Monolingual lexical databases)	17
Multilingual lexical databases	9
Onomastica (proper and geographical name lexical)	4

If the answer was "Text Corpora", then we got:

Answer	Count
Monolingual text corpora	14
Multilingual and parallel text corpora	8
Multilingual and Aligned text corpora	5

These are essential information for the follow-up of the project activities.

If the response was "Multimedia/multimodal Resources", then we got the following figures:

Answer	Count
Face	6
Image	8
Video	9
Finger prints	3
Other	1

Regarding the sources of the LRs used by the respondents, we obtained that:

Answer	Count
produced internally	25
produced by specific contracted vendors	8
Through data centers	15
Other	3

Those who replied to our question regarding production of resources were asked about the tools they use to design and produce LRs. The listed tools cover very basic and general-purpose tools (e.g. Matlab) as well as very specific ones developed on the basis of well known toolkits (e.g. HTK, Cool Edit), many responses referred to internal ones (e.g. graphical editor and viewer for tree-like structures, written Arabic text annotation tool Fassieh, etc.).

Many participants stated that some of the customized tools could be made available.

Standards and best practices are also important to the project and when asked about the standards & best practices they follow to design and produce LRs, the replies were:

Answer	Count
None	8
Internal specifications	25
External specifications	5

This points-out the global problem of information dissemination on the standardization activities being carried out at the international level rather than the awareness problem by the respondents.

Another important dimension we wanted to tackle through the survey is the issue of sharing and distributing LRs. The participants were asked if they are willing to make their resources available to others according to a negotiated distribution agreement and we got:

Answer	Count	Percentage
No answer	31	57.41%
Yes (Y)	19	35.19%
No (N)	1	1.85%
Non completed	3	5.56%

Which clearly indicates that more than half of the respondents did not want to “make” any commitment; some of those who were asked about this in a separate interview mentioned that the status of their data bases is to be investigated both from packaging, quality, and “value” points of view before envisaging any distribution.

Those who answered positively to the question on distribution, were very specific on whom they would agree to supply their data to:

Answer	Count
End-users	10
Tool developers	13
Researchers	12

Another group of questions were devoted to collecting some Market data. We tried to gather information about the market coverage from all the respondents and did ask them about their offers on their domestic market (country-based) as will as Arabic versus International markets. Products and/or services distributed and/or

offered to the:

Answer	Count
Domestic market	17
Arabic world	14
International market	18

Another market question is about the plans to purchase LRs. We knew that this is a hard question and that most of the respondents are not necessarily decision makers with respect to that and therefore the answers should be interpreted carefully as was stated by several participants.

The budgets seem to be steady over the next a few years (Euro/year):

	Now	3-5 Years
Nb of respondents	9	10
Average	10100	12950
Maximum	40000	50000

5. The needs for LRs as derived from the survey

The most important question was about the needs for LRs as this is the crucial input to the other tasks to be carried out by the project.

We list herein the replies categorized into speech, lexica (inc. Wordnet), corpora, multimedia/multimodal, and tools:

✓ Speech:
The needs as expressed by the respondents refer to most of the hot topics in speech processing as covered by the major conferences. Among the LRs listed we have Arabic conversational speech, Multi-speaker colloquial and formal Arabic speech databases for ASR an dictation, high-quality recordings for male and female concatenative Arabic TTS data bases; corpora for language models, recordings in different environments (e.g. Car).

✓ Lexica, wordnets, …
The needs cover a full scale Arabic WordNet, a validated comprehensive Arabic lexicon (general language), Terminological resources, morphosyntactic-lexicons for Arabic words, various ontologies and thesauri, Arabic proper names dictionaries, etc.

✓ Corpora:
The corpora item covered monolingual, bi/multilingual, with various annotations; participants indicated text corpora for special domains, Idiomatic Databases and Corpora, validated morphologically analyzed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Morphological model).

Parallel Corpora and aligned ones were mentioned many times in particular for language pairs e.g. English & French; an aligned corpus between Modern Standard Arabic and its colloquial varieties was also mentioned. A database consisting of a "gigantic corpora" labeled as per proper names was also mentioned. Printed Arabic font-written text images corpus was also listed along with Segmented Arabic Hand written corpus.

Data coming from the new "text styles" were also suggested (emails, blogs, Wikis, Forums).
A few multimodal and/or multimedia were indicated.

✓ Tools:
Some of the tools that were listed included basic tools for the (text and speech) processing. Also listed a baseline of the NLP infrastructure with phonetically, morphologically, syntactically, semantically, proper nouns/named-entities annotators.

In addition to the LRs required, the contributors listed a large number of technologies and applications that process Arabic language(s). As expected MT, CLIR/MLIR, and Automatic Speech Recognition, are mentioned several times.

Other applications such as Arabic language learning for non native speakers, Bilingual News Tracking, Arabic Discourse analysis with dictionary use, Document Management Systems (DMS), with a special focus on Arabic within either monolingual or multilingual applications, Handwriting Recognition, IR and text search engines, web and search engine applications, Linguistics developing parsers, Morphology, Omni font written OCR (which could help collect the huge aligned corpus as required by MT), Question/Answering, tools (web page reading) for the blind or visually impaired people, Spoken Document-retrieval, Spell checkers, Tools for Collaborative Content Services (CCS), Voice Car navigation systems, etc.

A summary of these findings was given in the project deliverable.

6. Additional findings by ELDA and the MEDAR partners for MT

In order to have a more exhaustive view of the Arabic HLT, the consortium decided to extend this work and identify other players who did not participate and/or more resources and tools, with a strong focus on MT and CLIR. The consortium has collected information from various sources such as LREC proceedings, European funded projects, Evaluation campaigns workshops (e.g. NIST/MT, Evalda-CESTA, etc.).

The first summary is related to MT systems that were identified. The consortium identified over 40 releases (some systems with more than one version) that handle Arabic and another language. The other languages constituted the following pairs: Arabic->English, Arabic->French, Arabic->Spanish, English->Arabic, French->Arabic, Spanish->Arabic. About 27 systems are offered by Commercial entities while a dozen are developed within academic and not for profit

organizations. Regarding the technical approach used we noticed that 15 claim to be SMT, three based on a more specific approach related to Translation Memories, at least one is based on hybrid approach (rule-based and statistics), a large number did not disclose their approach and are labeled as unknown (about 20). The MEDAR project is disseminating such information (that will be regularly updated) and is also working towards the development of an MT baseline that will be made widely available.

In addition to MT systems and components, the consortium also identified a large number of tools that could help carry out NLP research without starting from scratch as usual. These tools cover topics like morphological analyzers and PoS Taggers, stemmer, Parsers, and different Syntactic Analyzers.

Regarding the CLIR and/or MLIR, we also identified a number of tools such as Text Search Engines or Question Answering tools.

For these areas (MT, CLIR) we also identified useful language resources that could help customize them and/or assess their performance.

7. Directory of HLT Players

As planned within the MEDAR project we wanted to revise and update a directory of players involved in Arabic Human Language technology activities and projects. We have collected such information from projects such as Oriental, NEMLAR, CESTA, MEDAR as well as from contributions to workshops/conferences on Arabic.

This first draft is a list of institutions and individual experts with some input about their activities. A coming version will elaborate on profiles and sector of activities for each. This directory could be shared with all interested parties if legal, ethical, or privacy issues do not prevent us from doing so.

8. Contributions to fulfilling the remaining "gaps" as defined by MEDAR

This survey has focused essentially on identifying the players, LRs and Tools. The LRs and the tools are those that could be part of a revised Arabic BLARK for MT & CLIR/MLIR. It has identified a large set of requested resources and a few available ones.

The following important task within MEDAR is to list the LRs & Tools identified during this survey phase, drawing conclusions about which items are usable and which are not. MEDAR will also prioritize these items according to the BLARK as defined by NEMLAR both in terms of importance and availability.

Although, the BLARK concept was introduced to serve as a support for pre-competitive activities by researchers, developers, integrators, educators, etc. and not as a direct basis for commercial applications, it is important to pave the way to several levels of systems with various performances and with different requirements if this can be achieved by available resources and open source systems.

Our primary target is to specify and try to fulfill requirements of the precompetitive R&D activities that may indirectly lead to commercial products or services.

9. Acknowledgements

We would like to thank the European Commission for its support to this project.

This paper builds on work done in NEMLAR, as well as the preparation and the first part of MEDAR. MEDAR has 15 partners, and we want to acknowledge the contribution of all of them (see the list on www.MEDAR.info).

10. References

Choukri, K., *Mediterranean Arabic Language and Speech Technology, "Deliverable D3.1.Survey of actors, projects, products", December*, 23d, 2008, V2.0 (www.medar.info)