



MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable 3.1

Survey of actors, resources and products for Arabic HLT

Author: Khalid Choukri, ELDA
November 2010

MEDAR partners

- **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 the Development of Computer Systems, Egypt
- **Birzeit University:** Center for Continuing Education, West Bank and Gaza Strip
- **University Mohammed V Souissi:** Ecole Nationale Supérieure d'Informatique Analyse des Systèmes, Morocco
- **CEA,** Commissariat à l'Energie Atomique: CEA-LIST/LIC2M, Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue, 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 Software Company,** Egypt



European Commission

The MEDAR project is supported by the ICT programme

© The authors and MEDAR, c/o Center for Sprogteknologi, University of Copenhagen, November 2010, <http://www.medar.info>, email: nemlar@hum.ku.dk

CONTENT

1. EXECUTIVE SUMMARY	4
2. INTRODUCTION TO THE SECOND ROUND OF THE MEDAR SURVEYS	5
2.1. THE SURVEYS	5
3. MEDAR SECOND PHASE SURVEY(S) AND INFORMATION GATHERING	6
4. THE MEDAR ONLINE SURVEYS AND THE QUESTIONNAIRE STRUCTURE	7
4.1. THE MEDAR SECOND SURVEY QUICK SUMMARY	8
4.2. THE MEDAR SECOND SURVEY(S) SUMMARY OF RESULTS	8
5. OTHER MEANS FOR INFORMATION GATHERING	13
6. ANALYSIS OF THE ONLINE SURVEY	15
7. ADDITIONAL FINDINGS BY THE MEDAR PARTNERS	16
8. A MEDAR SUSTAINABLE KNOWLEDGE-BASE OF PLAYERS, TOOLS, RESOURCES	16
8.1. THE MEDAR KNOWLEDGE BASE OF RESOURCES	17
8.2. THE MEDAR KNOWLEDGE BASE OF TECHNOLOGIES	17
8.3. THE MEDAR KNOWLEDGE BASE OF PLAYERS	18
9. APPENDIX A: MEDAR 2009 ONLINE SURVEY AND THE QUESTIONNAIRE STRUCTURE	19

1. Executive Summary

The MEDAR survey aims at describing existing institutions and experts involved in the development of Arabic LRs, activities and projects being carried out, existing language resources and tools.

The first MEDAR survey was conducted as an updated version of a report drafted in the NEMLAR project that compiled information collected during 2004 and 2005. The original work carried out in NEMLAR but also in a first phase of MEDAR, used very extensively the consortium partners' networks to collect the raw data and then went through a compilation and correction stage. Part of that work relied on the expertise of members who used a common word-based questionnaire and direct interviews to collect the information.

The second survey partially used a web-based tool that helped participants to fill in a user-friendly questionnaire, leading to more results. This report is based on the second round of surveys conducted by MEDAR in 2009 and 2010. The surveys consisted of a survey conducted in 2009 and then the web site collecting the information was left open and promoted in various events which attracted some new respondents.

In order to consolidate all these findings, MEDAR decided to structure the collected data into a knowledge base that would be accessible through Internet. The knowledge base is available on www.elda.org/medar_knowledge_base/.

The first MEDAR survey allowed us to collect information about 54 players; mostly those who were closely related to the members of the consortium; large majority of them located within Arabic countries and addressing language technology developments as a peripheral activity.

The first round of the second survey allowed collecting more accurate data about 33 key players from all over the world, strongly involved in HLT. The use of various channels to disseminate information about the survey and the project helped on this. The second round, a continuation of the second survey, allowed collecting more information, with about 12 new references.

The list of players, tools, technologies, Language Resources is now part of our knowledge base and will be provided as a web-site (and not as an appendix to this report).

2. Introduction to the second round of the MEDAR Surveys

During the second phase of the project (2009-2010), MEDAR conducted a second survey to enrich its inventories. The surveys consisted of a first round conducted in 2009 and then the web site collecting the information was left open and promoted in various events which attracted some new respondents. The compilation of these new responses constitutes our second round.

A new compilation and consolidation of these two surveys was conducted by July 2010.

2.1. The Surveys

The initial MEDAR survey itself was an updated version of a report in the NEMLAR project that compiled information collected during 2004 and 2005 (First version of the NEMLAR Survey was made available on June 7th, 2004 with an updated version on March 2005). These documents aimed at describing the work done with respect to surveying existing institutions and experts involved in the development of Arabic LR, activities and projects being carried out, existing language resources and tools.

It was planned to review the current status of the domain and update the reports with the new findings. The work carried out for the purpose of updating such findings adopted multiple approaches to overcome some of the drawbacks of the original surveys (2004-2005, 2008).

The original work carried out in NEMLAR and in the first phase of MEDAR, used very extensively the consortium partners' networks to collect the raw data and then went through a compilation and correction stage. Part of that work relied on the expertise of members who used a common word-based questionnaire and direct interviews to collect the information.

The second survey used a web-based tool that helped participants to fill in, in a user-friendly manner, the MEDAR questionnaire, leading to more results.

This final update is based on that survey but also on the exploitation of new instruments that ELRA established (some in conjunction with the EC project FlareNet) to monitor the HLT area with regular snapshots taken at major milestones like LREC Conferences. For instance, the LRE-Map was used to identify resources that were described at LREC2010 and COLING2010 and that could be new with respect to the inventories listed in previous reports. ELRA also exploited its "LR identification task force" that collects data on new resources to identify those related to Arabic. ELDA also set up a specific web site to allow the storage of information of resources usable within WP5 (Language Resources for Machine Translation and its evaluation) for the MEDAR consortium.

The results obtained through these approaches are described in this report.

In order to consolidate all these findings, MEDAR decided to structure the collected data into a knowledge base that would be accessible through the Internet. This is available on www.elda.org/medar_knowledge_base/. This knowledge base will be maintained and regularly updated regarding the Institutions and experts, language technologies and LR. It is also envisaged that such knowledge base would have an interface that allow the interested parties to fill in, update, revise their data. Such work will be moderated to avoid any miss-use.

3. MEDAR second phase survey(s) and information gathering

This survey is an update of the original surveys of NEMLAR and MEDAR (going back to 2004) and aims at collecting more information on the players, products and projects with respect to language technology for Arabic in the region. Although MEDAR focuses on tools related to machine translation and information retrieval, the ultimate goal is to draw an accurate knowledge base of the language technology players, projects (ongoing activities), products etc. at large.

Already, a web-based knowledge base has been built from the information collected in the previous surveys and proposes an interesting directory of available information on players, both individual entrepreneur or from larger institutions (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.

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.

This work has provided 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 survey conducted in 2009 was kept active and regular reminders were sent to different mailing lists in addition to individual contacts. It was also publicised at various events e.g. Conferences, meetings, etc. This approach attracted a few new players who did not respond to our initial requests.

During the first word-based and initial web-based surveys (refereed to herein as MEDAR SURVEY 2009), about 54 questionnaires were filled in, some have been entirely completed (37), and some (17) questionnaires are missing some of the answers but were still considered for their usefulness. Part of the responses have been analysed within the first MEDAR Report.

During the second phase (referred to as MEDAR SURVEY Knowledge-base) about 33 questionnaires were filled, all of them exploitable but 9 (this is the main part of the previous report).

During the second round of this second survey (as will be reported on herein) we managed to collect an additional 12 new responses, leading to some 45 new contacts.

During these two last surveys we have managed to identify major players that emerged since the set up of MEDAR or some players that have initiated some strong activities on Arabic; For instance we have identified and set some partnership with Meedan.net (An Arabic-English forum using Machine Translation with expert corrections that also derive aligned data from its forum). We have noticed the participation of a number of new teams from several universities e.g. from Algeria, a country that was missing in our previous inventory. We have also enjoyed the participation of some major players in the Speech area which have a well

established record on Arabic audio transcription (LIMSI, LIGrenoble France, INESC Portugal). We have also seen the participation of players that do work on other Semitic languages. All these institutions are now part of our knowledge base.

Some facts and figures about the two surveys are given herein one after the other to stress some contrast (see structures in appendix A, B). The idea now is to clean all the data and offer it as a single knowledge base of players, resources, tools and technologies.

4. The MEDAR Online Surveys and the questionnaire structure

In order to ensure the largest number of replies to the MEDAR surveys, we opted for an online questionnaire using a web-based tool for online question-and-answer surveys called Limesurvey (<http://docs.limesurvey.org/>), an open source tool that allows to set up user-friendly surveys and also to collect the information in various format that render them very easy to analyze and exploit. The tool also allows to define and set up conditions to display a question or a group of questions if certain conditions are met (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 wished and the tool saved partially finished surveys.

The main challenge was to ensure that filling the questionnaire would not take more than 5 minutes for the new respondents. The questionnaire was set up on the basis of 3 groups of questions.

MEDAR Knowledge Base 2

The goal of this new survey is to update the MEDAR Knowledge base. This base consists of information about the existing experts, organizations, projects, products and language resources. It also aims to give a new opportunity to those who did not contribute to previous surveys to be listed in this knowledge base.

There are 20 questions in total and only 3 for those who have already answered previous surveys.

The structure of the questionnaire is described briefly below and the details are given in the appendix A.

The questionnaire was structured into 2 groups plus an initial one named as "Welcome information", meant to determine whether the respondent had already participated into a MEDAR survey or if this was his/her first visit. In case of returning participants, we did not require any identification information as we had that already stored in our databases.

The new respondent had to give their details, including the profile of their institution (e.g. academic versus industry, public versus private, etc.), their main activities in the language technology area, indicate their projects or products, etc.

The final Group (Group 3) is the Information about Language Resources. This group of questions was meant to collect information about the language resources that the respondent or his/her institution has been using and/or developing, and also list the needs in terms of LRs.

The questionnaire is attached as Appendices A and B to this report.

4.1. The MEDAR second survey quick summary

For this update to the initial MEDAR survey, we have managed to obtain 33 responses (3 are returning visitors, 21 full responses and 9 responses not completely filled out but still provide good information). The results given below comprise the detailed number of responses to each question and the percentages are computed on the basis of the 30 responses.

For this second round of the second survey, we got 12 new participants leading to 45 responses all in all during this second round (26 full responses, 19 responses not completely filled out but still exploitable).

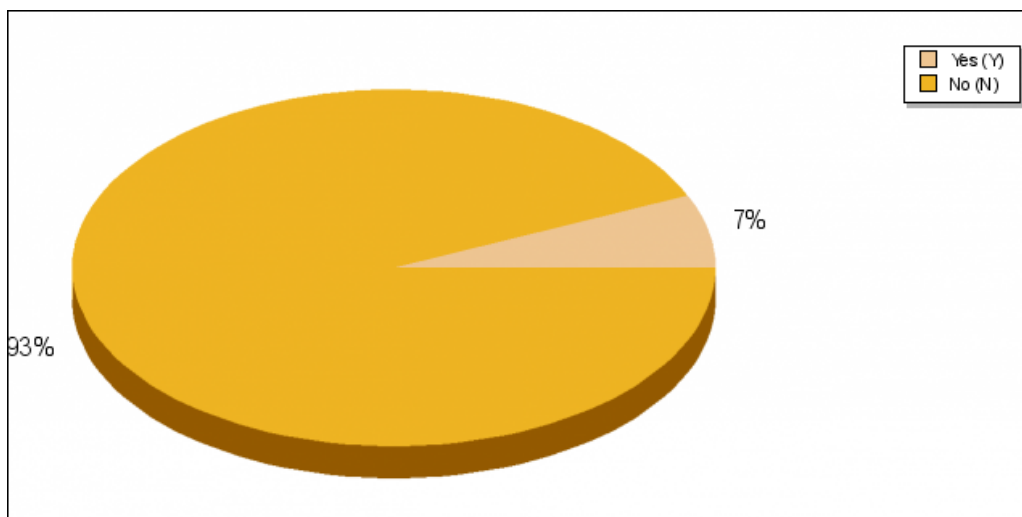
The survey re-used the web based questionnaire indicated above. A summary of both update 1 and update 2 is given herein.

4.2. The MEDAR second survey(s) summary of results

The first question aimed to identify participation to the survey of new participants.
Did you participate to our previous surveys? (If so you are already part of the MEDAR Knowledge Base)

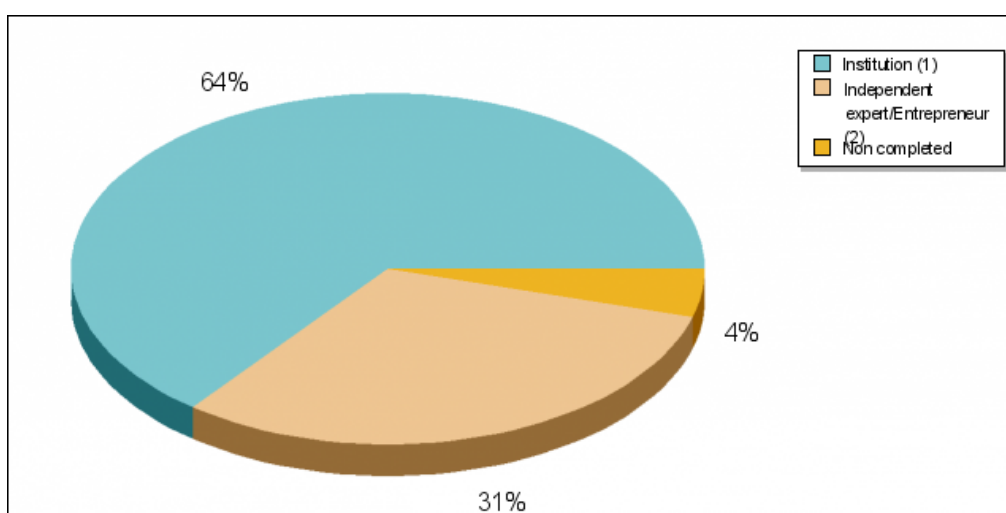
Did you participate in our previous surveys?		
Answer	Count	Percentage
No answer	0	0
Yes (Y)	3	6.67%
No (N)	42	93.33%
Non completed	0	0

This means that most of our participants did not participate directly to our initial surveys. Some were already listed in our inventories thanks to MEDAR partners but many were not.



The profiles of the participants with respect to our list were:

Profiles of participants		
Answer	Count	Percentage
No answer	0	0
Institution (1)	29	64.44%
Independent expert/Entrepreneur (2)	14	31.11%
Non completed	2	4.44%

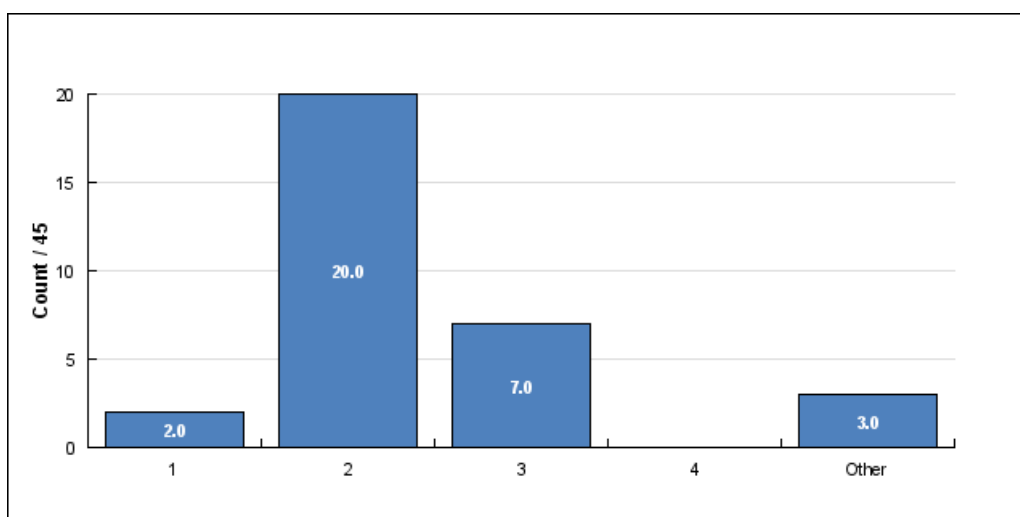


Then about the countries, the ones that have been listed are (please keep in mind that many big players were identified already in previous surveys hence the list of countries).

Countries			
Country	Mentioned	Country	Mentioned
Algeria	1	Lebanon	1
Belgium	1	Malaysia	1
Brazil	1	Morocco	1
Canada	1	Portugal	1
Egypt	2	Tunisia	5
France	5	Turkey	1
Germany	1	United Kingdom	3
Greece	1	USA	11
Israel	1	Vietnam	1
Japan	1		
Jordan	1		

Regarding the institutions' profile we got:

Institution profile		
Answer	Count	Percentage
Company & for profit organization (1)	2	4.44%
University (2)	20	44.44%
Public Research Center (3)	7	15.56%
Other Public Organization (4)	0	0
Other	3	6.67



An important question was related to the core activity (-ies) of the respondent. We have kept only the activities that were selected by the respondent in our summary, except for the HLT Product vendor which we had explicitly mentioned in our categories. We see now that very few state that their main activity is HLT vendor. The “Other” is from R&D groups within Companies.

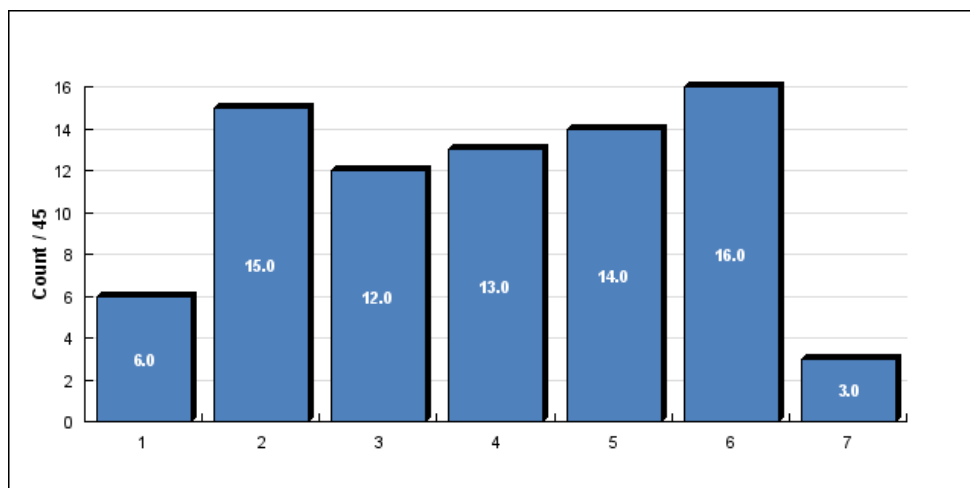
Institution's main activity (choose several options if needed).		
Answer	Count	Percentage
Software Development (1)	11	24.44%
Teaching/training Organization (e.g. university) (2)	21	46.67%
HLT Product Vendor (3)	0	0
Culture/Museum (4)	1	2.22%
Technology Transfer Institution (12)	2	4.44%
Minority Language Organization (5)	0	0
Content Provider (6)	2	4.44%
Interpretation/Translation/Localization (7)	8	17.78%
Telecommunication (8)	2	4.44%
E-commerce (9)	0	0
	50	

The next question is about direct and strong involvement in HLT, unfortunately only half the group answered to this question (24 out of 45), 4 said no and the rest did not reply. We assume that the definition was not clear to everyone.

From the answers of those who responded (they could select more than one technology, hence the sum is over 100%), we obtained the following:

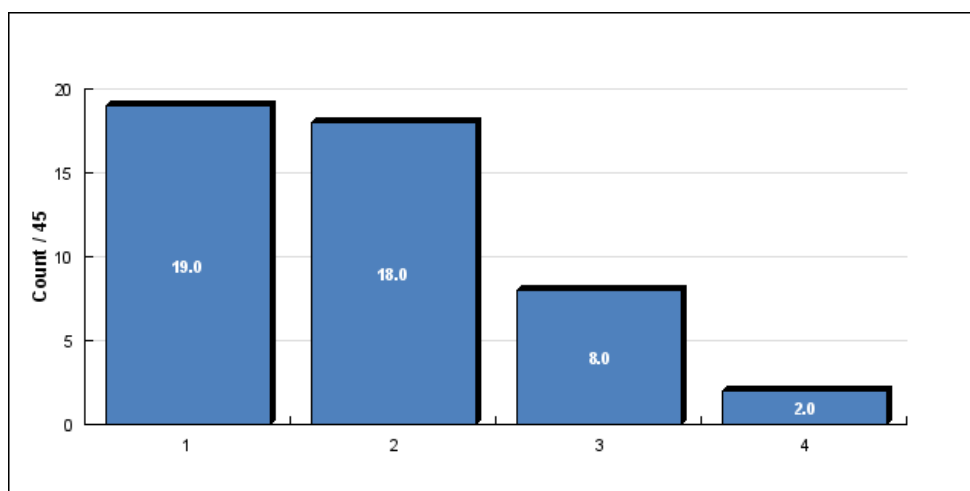
Which Language Technology? Please provide any relevant information.		
Answer	Count	Percentage
Language Learning (1)	6	13.33%
Language Resource Production (2)	15	33.33%
Speech Technologies (3)	12	26.67%
Written Technologies (4)	13	28.89%
Search and Knowledge Mining (5)	14	31.11%
Translation Automation (6)	16	35.56%
Other (7)	3	6.67%

The optimistic result is that about 11 players indicated that they are involved in Search and Translation.



Regarding the products, LR and tools have the largest share (36 mentions and 66.67%) while services are selected only 8 times (14%).

The institution's main products, tools and/or services.		
Answer	Count	Percentage
Language Resources (1)	19	42.22%
Tools (2)	18	40.00%
Services (3)	8	17.78%
Other (4)	2	4.44%



When the participants were asked if those products, tools and/or services were monolingual or bilingual, 10 selected Monolingual, while 20 selected multilingual. When asked if the products, tools and/or services, included Arabic, 18 replied YES and 6 NO while more than 50% did not reply.

When asked about the types of resources, 9 mentioned speech, 13 mentioned written, 7

mentioned multimodal resources. A participant selected the “Other” category and explicitly mentioned “software tools, evaluation tools and metrics”.

Last but not least, the participants expressed their needs and wishes for LRs and the list is given herein as is without any prioritization. One could see that several resources are available already but we could interpret this as “no better data than more data”.

- Arabic dialectology
- Diglossia and code-switching
- Arabisation of foreign terms, Foreign loanwords and absorption
- Acronyms, and expressions,
- Annotated corpora for Arabic.
- Arabic Corpora to evaluate tools of automatic abstracting, parsing, morphological analysis, etc.
- Audio visual materials at specific language proficiency levels
- Corpora of dialects
- Resources for developing multilingual MT software
- Resources from various Arabic speaking countries for localisation (Standard and colloquial Arabic)
- Handwritten resources
- lexical resources (dictionary), formal grammars for syntax,
- Open translation memory for Arabic and English.
- Parallel corpora and translation memory for training our translation engine.
- pathological voice resources
- Resources for research in the fields of speech recognition and synthesis.
- Training, development and evaluation data (speech and text)

5. Other means for information gathering

As stated above, in addition to the two surveys, MEDAR also collected useful information on LRs and tools exploiting three instruments:

Use of the LRE-Map introduced at LREC2010 by ELRA & FlareNet that requires all paper submissions to fill in a form describing the resources used within the paper; For the MEDAR purposes we exploited the data collected at LREC2010 and also from COLING2010 (respectively over 80 references with some duplicates and 16 unique references).

Exploitation of the ELRA universal catalogue: the Universal Catalogue is the outcome of the ELRA “LR identification task force” that collects data on new resources. We tried to identify within the catalogue those resources related to Arabic.

We used the BLARK document with respect to new findings on language resources and tools, and we used the Cooperation Roadmap with respect to list of actors.

The Semitic workshop at LREC2010, co-organised by MEDAR, also gave valuable input.

The set up for the MEDAR consortium of a dedicated web site to allow the storage of information of resources usable within WP5 (Language Resources for Machine Translation

and its evaluation).

Among those resources we have about 6 new annotation tools (annotations of PoS, Arabic Discourse Annotation Tool, etc.), about 40 references to corpora with a majority of new resources, some of them parallel and/or aligned. A few papers mentioned evaluation packages, about a dozen mentioned lexica (about 10 different resources) and then about 10 specialized tools such as Named Entity Recognisers, Spell Corrector, Parsers, etc.

The consortium will continue to compile such resources and enrich the knowledge base.

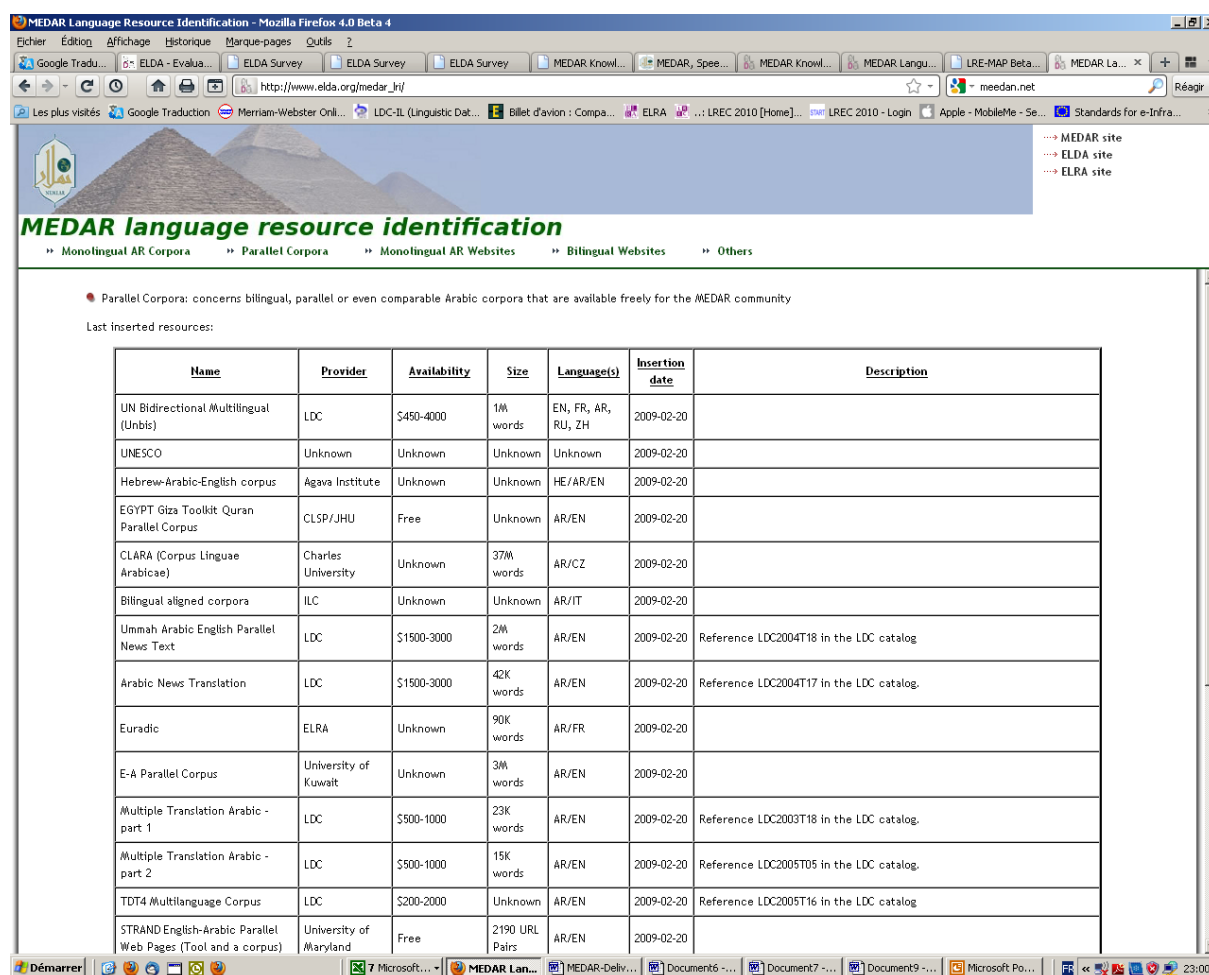
A screen shot is given below to illustrate the outcome of the LRE-Map at LREC and COLING with respect to Arabic as a language.

The screenshot displays the LRE-MAP (Beta Version 1.0) web application. The interface includes a navigation bar with links like 'Account', 'Tool', 'Search', 'Stats', and 'About'. Below this, there's a section titled 'Browse in the LRE MAP' with tabs for 'Type', 'Status', 'Avail', 'Mod', 'Use', and 'Lang'. A search sidebar on the left allows filtering by 'Select Conference' (LREC2010, COLING2010), 'Resource Name', 'Exact Match', 'Search by', and 'Source Language' (Arabic). A callout points to the 'Arabic' selection with the text 'Search on "Arabic"'. Another callout points to the 'Lang' tab with the text 'Search on various parameters (Type, Lang, Modality, etc.)'. The main area shows '33 Resources Found' in a table with columns 'Resource Type', 'Resource Name', and 'Details'. Resources listed include 'SelectPOS', 'TreeEditor', 'ACE data 2007', 'ANERcorp', 'Arabic Propbank', 'Arabic Tree Bank', 'Arabic Treebank', 'Arabic Treebank Part 1 v 4.0', 'Arabic Treebank Part 2 v 3.0', 'Arabic Treebank Part 5 V1.0', 'Arabic Treebank Part3 - Version 3.1', 'Arabic Treebank Part3 - Version 3.1', 'Catib version of the Penn Arabic Treebank part 3 v3.1', 'Corpus of Arabic Speech', 'LDC2005T20', 'List of questions and their answers', 'Penn Arabic Treebank', 'Penn Arabic Treebank 2', 'Pilot Arabic CCGbank', 'The Essex Arabic Summaries Corpus (EASC)', 'The Leeds Arabic Discourse Treebank (LADTB)', and 'the Penn Arabic Treebank (Part 1 v. 2.0)'. A callout points to the 'Sources of information (LREC, COLING)' section. The footer includes logos for ELRA, LREC, and COLING, along with contact information and links for 'Help' and 'Stats 2010'.

The ELRA Universal Catalogue (ELRA-UC) contains about 72 LRs in Arabic. In particular, we should highlight that 17 resources (out of the 72) are available through the ELRA catalogue. Of course there are duplicates between the ELRA-UC and the LRE-Map as LREC proceedings are also reviewed by the team to collect information about resources. Some of these are novelties with respect to the R&D areas being tackled these days. For instance, some of the corpora are about Arabic dialects, Colloquial Arabic spoken databases, a number of bilingual dictionaries (the other languages being other than English or French, e.g. Spanish, Dutch), an Arabic WordNet, a Collection of Arabic Document Images, etc.

The ELRA Catalogue was reported on, in previous reports, it has about 20 resources, many of these to address issues related to speech processing but some appropriate for MT&IR.

The ELDA web site www.elda.org/medar_lri was also used to collect information on Language Resources usable by the participating sites to our MT evaluation. This is part of our final knowledge base and the ultimate storage place for resources on Arabic. It contains already information on Monolingual corpora, Bilingual Corpora (consists of bilingual, parallel or even comparable Arabic corpora), Monolingual Arabic Websites with textual resources that could be used to derive textual corpora, Bilingual Websites with data in more than one language (in addition to Arabic) and that could be used for creating parallel and/or comparable corpora and finally on “others”, miscellaneous items useful for MEDAR activities such as evaluation corpora, corpus tools, etc. The information has been gathered by members of the consortium and verified. For instance the parallel corpus section shows the list of languages that are used in addition to Arabic, it looks like:



The screenshot shows the MEDAR language resource identification website. The page title is "MEDAR language resource identification". The navigation bar includes links for Monolingual AR Corpora, Parallel Corpora, Monolingual AR Websites, Bilingual Websites, and Others. The main content area is titled "Parallel Corpora: concerns bilingual, parallel or even comparable Arabic corpora that are available freely for the MEDAR community". Below this, there is a section "Last inserted resources:" followed by a table.

Name	Provider	Availability	Size	Language(s)	Insertion date	Description
UN Bidirectional Multilingual (Unbis)	LDC	\$450-4000	1M words	EN, FR, AR, RU, ZH	2009-02-20	
UNESCO	Unknown	Unknown	Unknown	Unknown	2009-02-20	
Hebrew-Arabic-English corpus	Agava Institute	Unknown	Unknown	HE/AR/EN	2009-02-20	
EGYPT Giza Toolkit Quran Parallel Corpus	CLSP/JHU	Free	Unknown	AR/EN	2009-02-20	
CLARA (Corpus Linguae Arabicae)	Charles University	Unknown	37M words	AR/CZ	2009-02-20	
Bilingual aligned corpora	ILC	Unknown	Unknown	AR/IT	2009-02-20	
Ummah Arabic English Parallel News Text	LDC	\$1500-3000	2M words	AR/EN	2009-02-20	Reference LDC2004T18 in the LDC catalog
Arabic News Translation	LDC	\$1500-3000	42K words	AR/EN	2009-02-20	Reference LDC2004T17 in the LDC catalog
Euradic	ELRA	Unknown	90K words	AR/FR	2009-02-20	
E-A Parallel Corpus	University of Kuwait	Unknown	3M words	AR/EN	2009-02-20	
Multiple Translation Arabic - part 1	LDC	\$500-1000	23K words	AR/EN	2009-02-20	Reference LDC2003T18 in the LDC catalog
Multiple Translation Arabic - part 2	LDC	\$500-1000	15K words	AR/EN	2009-02-20	Reference LDC2005T05 in the LDC catalog
TDT4 Multilingual Corpus	LDC	\$200-2000	Unknown	AR/EN	2009-02-20	Reference LDC2005T16 in the LDC catalog
STRAND English-Arabic Parallel Web Pages (Tool and a corpus)	University of Maryland	Free	2190 URL Pairs	AR/EN	2009-02-20	

6. Analysis of the online survey

From the survey we extracted information regarding more than 45 key players that were not part of our previous inventory. We also identified a number of hot topics, though there are nothing new compared to our previous surveys, i.e. the results of previous survey are still relevant in this respect.

7. Additional findings by the MEDAR partners

In addition to the Surveys outcomes, we have also collected valuable information through other means by the members of the consortium. The specific interview form used so far will be left active so new players can add their own data to our MEDAR knowledge-base.

8. A MEDAR sustainable knowledge-base of players, tools, resources

The work carried out within NEMLAR and MEDAR¹ that aimed initially to identify key players, projects and tools within the Arabic region has been extended to the identification of technologies, Language Resources etc. that are related to Arabic.

Over the years, such inventories have been drawn up and exploited to set up partnerships and joint projects. Our goal today is to establish a reliable database that would constitute a trustable reference to all players interested on Arabic HLT. These could be funding agencies, policy makers, research labs, HLT specialized corporations and vendors, researchers, as well as educational institutions.

This database will be open to the public with a moderation by the MEDAR partners so to keep the data accurate and focused.

A crucial task, being carried out at present is to consolidate all the findings (MEDAR Surveys, LREC-Map, ELRA-UC, ELRA Catalogue, MEDAR-WP5 activities) before opening it to the community.

The knowledge base itself contains most of the identified players, resources and technologies.

The matrix used herein allows to identify within a glance who is doing what and then with a click on the player access to more information. The access can be done through the players (by simple browsing), the technologies or the resources within a matrix a la BLARK. Here are the corresponding web pages:

¹ Some inventories did even start earlier, within projects such as OrienTel (2001-2003.) (<http://www.speechdat.org/ORIENTEL/index.html>)

8.1. The MEDAR Knowledge base of resources

MEDAR Knowledge Base - Mozilla Firefox 4.0 Beta 4

http://www.elda.org/medar_knowledge_base/

Les plus visités: Google Traduction, Merriam-Webster Onli..., LDC-IL (Linguistic Dat..., Billet d'avion : Compa..., ELRA, LREC 2010 (Home)...

MEDAR site
ELDA site
ELRA site

MEDAR Knowledge Base

» Home » Players » Language Technologies » Language Resources

Institution	Language Learning	Resource Production	Speech Technologies	Written Technologies	Search Knowledge Mining	Translation Automation
ELDA						
BOUZOUBAA Karim						
SOUFI ABDELHADI						
KANAN AIAH						
Insan Center						
ACS TechnoCenter						
IBM						
France Telecom R&D ORANGE Labs						
Higher Institute for Applied Sciences and Technology (HIAST)						
King Abdulaziz City for Science and Technology						
Isra' Software & Computer Co. Ltd						
RDI						
ManarahNet Modern Software Co.						
GIS Int.						
TALP Research Center - Universitat Politècnica de Catalunya						
The CJK Dictionary Institute						
Lebanese University						
SHTAYYAH Mohammed						
Unit for Learning Innovation- Birzeit University						
Indiana University						
Cairo Microsoft Innovation Center in Egypt (CMIC)						
ENSIS						
Millennium Technology						
CRSTDLA (Scientific & technical Research Center for Arabic Language Development)						
Arabize						
Alkharazmy Language Software						
HIYASSAT Hussein						
ARABIC TEXTWARE						
COLTEC						
CMELEK Yakup						
Natural Language Engineering Lab.						
ALBELTAGY Abdallah						
EL-MAHALLAWY Mohamed						

Démarrer Thunderbird 7 Microsoft... 2 Firefox 6 Adobe Re... Favoris réseau MEDAR-Delive... Document6 ... 18:24

8.2. The MEDAR Knowledge base of technologies

MEDAR Knowledge Base - Mozilla Firefox 4.0 Beta 4

http://www.elda.org/medar_knowledge_base/

Les plus visités: Google Traduction, Merriam-Webster Onli..., LDC-IL (Linguistic Dat..., Billet d'avion : Compa..., ELRA, LREC 2010 (Home)...

MEDAR site
ELDA site
ELRA site

MEDAR Knowledge Base

» Home » Players » Language Technologies » Language Resources

Institution	Speech Resources	Written Resources	Multimodal Resources	Other Resources	Produced Internally	Produced by Vendors	Distributed Data Centers	Tools Used
ELDA								
RDI								
ENSIS								
EL-MAHALLAWY Mohamed								
Laboratoire de Recherche en Informatique et Télécommunications Faculté des Sciences								
University Cadi Ayyad - Faculty of Sciences								
RAGHEB Ahmed								
EL HOUSSEINE BOUYAKHF								
EL JHAD Abdelhamid								
BOUZOUBAA Karim								
SOUFI ABDELHADI								
KANAN AIAH								
Insan Center								
ACS TechnoCenter								
IBM								
France Telecom R&D ORANGE Labs								
Higher Institute for Applied Sciences and Technology (HIAST)								
King Abdulaziz City for Science and Technology								
Isra' Software & Computer Co. Ltd								
ManarahNet Modern Software Co.								
GIS Int.								
TALP Research Center - Universitat Politècnica de Catalunya								
The CJK Dictionary Institute								
Lebanese University								
SHTAYYAH Mohammed								
Unit for Learning Innovation- Birzeit University								
Indiana University								
Cairo Microsoft Innovation Center in Egypt (CMIC)								
CRSTDLA (Scientific & technical Research Center for								

Démarrer Thunderbird 7 Microsoft... 2 Firefox 6 Adobe Re... Favoris réseau MEDAR-Delive... Document6 ... 18:24

8.3. The MEDAR Knowledge base of players

If one selects a particular player, this would give more details. This section is with restricted access so far as a number of issues e.g. consent, quality of the data, etc. requires more discussions with each partner. A new version of the knowledge base would allow each part to enter data that would be moderated (and checked) before publication. This will help consolidating the MEDAR community which is now very active. With two major conferences and several workshops, MEDAR is perceived as a major leading initiative within the field.

The screenshot displays the MEDAR Knowledge Base website. The browser window shows the URL http://www.elda.org/medar_knowledge_base/. The page has a header with the MEDAR logo and navigation links: Home, Players, Language Technologies, and Language Resources. On the left, a sidebar titled 'Players:' lists various institutions and individuals, including ELDA, RDI, ENSIAS, EL-MAHALLAWY Mohamed, and others. The main content area on the right provides details for the selected player, ELDA. It includes the institution's name (ELDA, France), website (www.elda.org), and a 'Contact (access restricted)' link. Below this, the 'Language Technologies' section lists 'Resource Production', 'Speech Technologies', and 'Written Technologies', with links to 'More details on speech technologies' and 'More details on written technologies'. The 'Language Resources' section lists 'Type of Speech Resources', 'Type of Written Resources', 'Type of Multimodal Resources', and 'Type of Other Resources'. The 'Products and Services' section lists 'Monolingual', 'Multilingual', and 'Arabic' resources, with a note on 'Main products: Language resources Technology evaluation services'. The footer indicates the page is 'Powered by ELDA © 2009 ELDA/ELRA'.

9. Appendix A: MEDAR 2009 Online Survey and the questionnaire structure

As reported in the previous sections, 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 interviews called Limesurvey (<http://docs.limesurvey.org/>). This is an open source survey tool that allows to set up surveys very user friendly and also to collect the information in various formats which render them very easy to analyze and exploit. The tool also allows asking a question and continuing the questionnaire according to the answer received (an easy "tree" interface). For more information about the questionnaire, please refer to the previous report (http://www.medar.info/MEDAR_Survey_II.pdf)

Group 1 is the Welcome information

This group of questions was meant to determine whether the respondent had already participated into a MEDAR survey or if this was his/her first visit. For returning respondents, the number of questions to answer was limited to 2. The others were brought to the Contact Information group of questions.

<p>Q1: Did you participate to our previous surveys? (If so you are already part of the MEDAR Knowledge Base)</p> <p>Please choose *only one* of the following:</p> <p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>
--

<p>[Only answer this question if you answered 'Yes' to question 'Q1 ']</p> <p>* Q1b: In this case, just leave your name and email. Thank You!</p> <p>Please write your answer(s) here:</p> <p>First Name: <input type="text"/></p> <p>Last Name: <input type="text"/></p> <p>E-mail address: <input type="text"/></p>
--

<p>[Only answer this question if you answered 'Yes' to question 'Q1 ']</p> <p>Q2a: Thank you for your participation!</p> <p>If needed, please contact us at medar@elda.org to update your data.</p>

Group 2 is the Contact Information

This group of questions was meant to collect all the contact information of the respondent in addition to details on his/her institution, including the field of activity, the type of service or tool developed, the use of Arabic language.

[Only answer this question if you answered 'No' to question 'Q1 ']

*** Q3: Please enter your contact information in this field.**

Please write your answer(s) here:

First Name:

Last Name:

E-mail address:

Web site:

[Only answer this question if you answered 'No' to question 'Q1 ']

*** Q4: Please specify your status below.**

Please choose *only one* of the following:

☐

Institution

☐

Independent expert/Entrepreneur

[Only answer this question if you answered 'Institution' to question 'Q4 ']

*** Q5: Details of your institution.**

Please write your answer(s) here:

Institution Full Name:

Address:

Zipcode:

City:

Web site:

[Only answer this question if you answered 'No' to question 'Q1 ']

*** Q5a: Country.**

This can be the country of your institution or the country where you reside.

Please write your answer here:

[Only answer this question if you answered 'No' to question 'Q1 ']

Q6: Phone number.

Please enter using the proper international format:

+ccc (aaa) nnn

where ccc stands for country code, (aaa) stands for area code and nnn stands for the number.

Please write your answer here:

[Only answer this question if you answered 'No' to question 'Q1 ']

Q6a: Fax number.

Please enter using the proper international format:

+ccc (aaa) nnn

where ccc stands for country code, (aaa) stands for area code and nnn stands for the number.

Please write your answer here:

[Only answer this question if you answered 'No' to question 'Q1 ' *and* if you answered 'Institution' to question 'Q4 ']

*** Q7: Type of institution**

Please choose *all* that apply:

- ☐ Company & for profit organization
- ☐ University
- ☐ Public Research Center
- ☐ Other Public Organization

Other:

[Only answer this question if you answered 'No' to question 'Q1 ' *and* if you answered 'Institution' to question 'Q4 ']

*** Q7c: Institution's main activity (choose several options if needed).**

Please choose *all* that apply:

- ☐ Software Development
- ☐ Teaching/training Organization (e.g. university)
- ☐ HLT Product Vendor
- ☐ Culture/Museum
- ☐ Technology Transfer Institution
- ☐ Minority Language Organization
- ☐ Content Provider
- ☐ Interpretation/Translation/Localization
- ☐ Telecommunication
- ☐ E-commerce
- ☐ Banking/Insurance

Other:

[Only answer this question if you answered 'No' to question 'Q1 ' *and* if you answered 'Institution' to question 'Q4 ']

*** Q8: Is your institution involved in Language Technologies?**

Please choose *only one* of the following:

- ☐ Yes
- ☐ No

[Only answer this question if you answered 'Institution' to question 'Q4 ' *and* if you answered 'Yes' to question 'Q8 ']

*** Q8a: Which Language Technology? Please provide any relevant information.**

Please choose all that apply and provide a comment:

<input type="checkbox"/> Language Learning	
<input type="checkbox"/> Language Resource Production	
<input type="checkbox"/> Speech Technologies	
<input type="checkbox"/> Written Technologies	
<input type="checkbox"/> Search and Knowledge Mining	
<input type="checkbox"/> Translation Automation	
<input type="checkbox"/> Other	

[Only answer this question if you answered 'Yes' to question 'Q8 ']

Q9: What are the institution's main products, tools and/or services? Please provide any relevant information.

Please choose all that apply and provide a comment:

<input type="checkbox"/> Language Resources	
<input type="checkbox"/> Tools	
<input type="checkbox"/> Services	
<input type="checkbox"/> Other	

[Only answer this question if you answered 'Yes' to question 'Q8 ']

*** Q9a: Are those products, tools or services**

Please choose **all** that apply:

<input type="checkbox"/> Monolingual
<input type="checkbox"/> Multilingual

[Only answer this question if you answered 'Yes' to question 'Q8 ']

*** Q9b: Do they include Arabic language?**

Please choose **only one** of the following:

<input type="checkbox"/> Yes
<input type="checkbox"/> No

Group 3 is the Information about Language Resources

This group of questions was meant to collect information about the language resources that the respondent or his/her institution has been using and/or developing, and also list the needs in terms of LRs.

[Only answer this question if you answered 'Language Resources' to question 'Q9 ']

*** Q10: Language Resource Type**

Please choose **all** that apply:

<input type="checkbox"/> Speech Resources
<input type="checkbox"/> Written Resources
<input type="checkbox"/> Multimedia/Multimodal Resources
Other: <input type="text"/>

[Only answer this question if you answered 'No' to question 'Q1 ']

*** Q11: Does the institution you represent use Language Resources**

Please choose *all* that apply:

- ☐ that are produced internally?
- ☐ that are produced by specific contracted vendors?
- ☐ that are distributed by data centres?

Other:

[Only answer this question if you answered 'No' to question 'Q1 ']

Q12: What are your needs in terms of Language Resources? Please provide specific information.

Please write your answer here:



[Only answer this question if you answered 'No' to question 'Q1 ']

Q14: Thank you for completing the survey.

Submit Your Survey.

Thank you for completing this survey.

Please fax your completed survey to: +33 1 43 13 33 30.