



MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable D3.1.2

**Update of the Survey of actors, projects, products
Revised and Augmented version**

Author: Khalid Choukri, ELDA

Contributors: Olivier Hamon, H       Mazo, Djamel Mostefa,
September 2009

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-Egypt**, The Engineering Company for the Development of Computer Systems, Egypt
- **Birzeit University:** Center for Continuing Education, West Bank and Gaza Strip
- **University Mohammed V Soussi:** Ecole Nationale Supérieure d'Informatique 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, Center for Sprogteknologi, University of Copenhagen, April 2009
http://www.medar.info, email: nemlar@hum.ku.dk

CONTENT

1. EXECUTIVE SUMMARY	4
2. INTRODUCTION TO THE MEDAR SURVEY #2	4
3. THE MEDAR ONLINE SURVEY AND THE QUESTIONNAIRE STRUCTURE	5
4. THE MEDAR SURVEY 2: SUMMARY OF THE FIGURES AND FACTS.....	10
4.1. IDENTIFICATION OF THE RESPONDENT.....	10
4.2. PROFILE OF THE RESPONDENT.....	11
4.3. INVOLVEMENT OF THE PLAYERS IN HLT & LRS:	13
4.4. MULTILINGUALITY ISSUES	15
4.5. INFORMATION ABOUT THE RESPONDENT'S LRS	16
5. THE MOST IMPORTANT QUESTION WAS ABOUT THE NEEDS FOR LRS.....	17
6. FOLLOW-UP OF THE ONLINE SURVEY(S)	18
7. APPENDIX A: THE MEDAR KNOWLEDGE-BASE	18
8. APPENDIX A: THE NEMLAR REPORT	18
9. APPENDIX B: THE MEDAR DELIVERABLE D3.1, A SURVEY OF ACTORS, PRODUCTS, PROJECTS	18

1. Executive Summary

This is the update report of the state of art of the situation of Human Language Technologies (HLT) for Arabic as drafted in December 2008. This document aims at describing the work done with respect to surveying existing institutions and experts involved in the development of Arabic Language Resources carried out since 2008 (since the first report of NEMLAR) and in 2009. It surveys the activities and projects, existing language resources and tools, as well as the experts. The Summary of the new findings (facts & figures) is given herein.

The update of the original survey was launched on August 2009 and those who had missed the previous surveys were encouraged to fill in the questionnaire for their institution or for themselves. As of September 10, 2009, 33 questionnaires were filled in, 21 have been entirely completed and for 9, the questionnaires were missing some of the answers but were still considered for their usefulness. 3 respondents were returning visitors and only left their details to be re-contacted. This time, respondents were distributed over 16 countries and a number of countries were well represented (e.g. 8 responses from the USA, 4 from Tunisia). A number of key players have taken part to this survey. We feel that the Internet-based questionnaire was more easy to use than by the past (email of word files).

While exploiting the data from September 10, 2009, the survey was left open¹. All the additions to the survey will be part of the next release of this report due by June 2010.

In the meantime, the outcome of all the surveys is compiled and consolidated in a “Knowledge-base” that is being made available to all and that comprises the identified experts, institutions, Language Resources & tools, etc. Such knowledge-base is available from the MEDAR web site (www.medar.info) (and mirrored at: http://www.elda.org/medar_knowledge_base/) and will be maintained and updated regularly.

The previous versions (individual reports) are also still available:

The NEMLAR REPORT is at:

http://www.medar.info/The_Nemlar_Project/Publications/NEMLAR-REPORT-SURVEY-FINAL_web.pdf

The MEDAR one (1st release of this deliverable) at: http://www.medar.info/MEDAR_Survey_I.pdf

2. Introduction to the MEDAR Survey #2

This survey is an update of the original survey and aims at collecting more information on the players, products and projects with respect to language technology for Arabic in the region. As the previous surveys, all run within Nemlar and MEDAR, this update is carried out within the MEDAR project and aimed at providing an overview and an analysis of the situation. 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.

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.

¹ By the 16th of September, the total number of entries in the survey is about 41 (25 Completed and 16 incomplete but exploitable).

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.

3. The MEDAR Online Survey 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 survey in more than one visit if they wish and the tool saves partially finished surveys.

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

This is the introduction to the questionnaire and the 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 below:

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.

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

[Only answer this question if you answered 'Yes' to question 'Q1 '] * Q1b: In this case, just leave your name and email. Thank You! Please write your answer(s) here:
--

First Name:	<input type="text"/>
Last Name:	<input type="text"/>
E-mail address:	<input type="text"/>

[Only answer this question if you answered 'Yes' to question 'Q1 ']

Q2a: Thank you for your participation!

If needed, please contact us at medar@elda.org to update your data.

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:	<input type="text"/>
Last Name:	<input type="text"/>
E-mail address:	<input type="text"/>
Web site:	<input type="text"/>

[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:	<input type="text"/>
Address:	<input type="text"/>
Zipcode:	<input type="text"/>
City:	<input type="text"/>
Web site:	<input type="text"/>

[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:

<input type="checkbox"/>	Company & for profit organization
<input type="checkbox"/>	University
<input type="checkbox"/>	Public Research Center
<input type="checkbox"/>	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:

<input type="checkbox"/>	Software Development
<input type="checkbox"/>	Teaching/training Organization (e.g. university)
<input type="checkbox"/>	HLT Product Vendor
<input type="checkbox"/>	Culture/Museum
<input type="checkbox"/>	Technology Transfer Institution
<input type="checkbox"/>	Minority Language Organization
<input type="checkbox"/>	Content Provider

<input type="checkbox"/> Interpretation/Translation/Localization <input type="checkbox"/> Telecommunication <input type="checkbox"/> E-commerce <input type="checkbox"/> 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:

☐ Monolingual
☐ 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:

☐ Yes

☐ 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:

☐ Speech Resources

☐ Written Resources

☐ Multimedia/Multimodal Resources

Other:

[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?

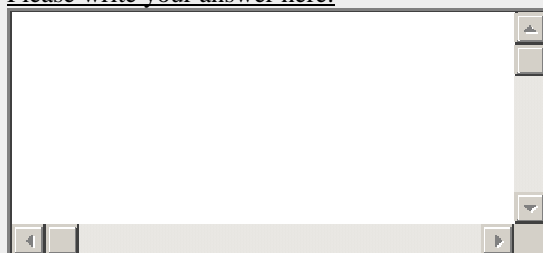
☐ that are distributed by data centres?

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.

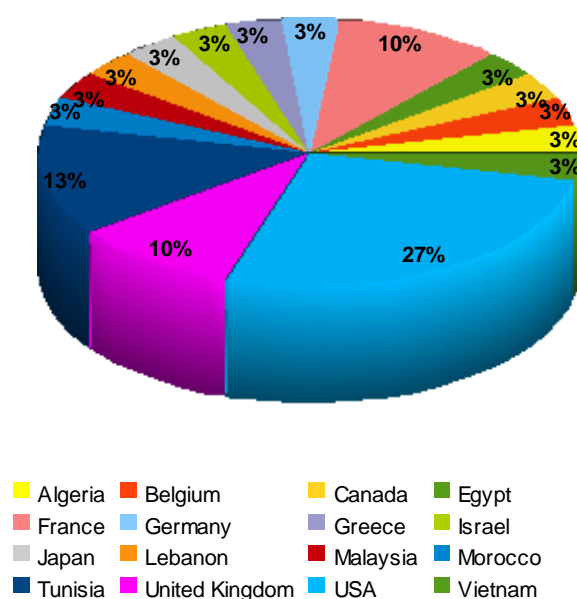
4. The MEDAR survey 2: summary of the figures and facts

For this update to the previous survey, we have managed to obtain 33 responses for this survey (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.

4.1. Identification of the respondent

The 33 respondents are identifiable by name, first name, etc.

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



The details per country are:

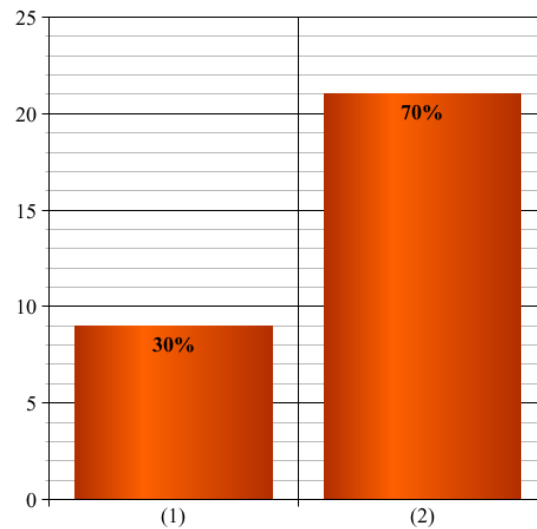
Answer	Count
Algeria	1
Belgium	1
Canada	1
Egypt	1
France	3
Germany	1
Greece	1
Israel	1
Answer	Count

Japan	1
Lebanon	1
Malaysia	1
Morocco	1
Tunisia	4
United Kingdom	3
USA	8
Vietnam	1

4.2. Profile of the Respondent

The profiles of the respondent were collected to ensure that we can distinguish independent experts from institutions and also their involvement in HLT & LR.

Answer	Count	Percentage
Independent expert/Entrepreneur (1)	9	30%
Institution (2)	21	70%

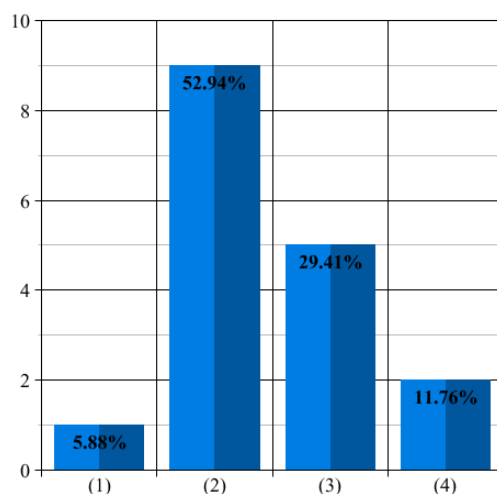


In addition to the individuals that replied without mentioning explicitly their institution, the following ones were listed for the first time (though some respondents refer to a different department from those we had listed by the past) :

- CNRS-LIMSI, France
- Ecole Supérieure des Communications, Tunisia
- EEDIS-UDL SBA, Algeria
- FSEGS, University of Sfax, Tunisia
- FSM, Tunisia
- Hedayet Institute for Arabic Studies, Egypt
- Institute for Speech and Language Processing, Greece
- Laboratory of Informatics of Grenoble, France
- Leuven Language Institute (Katholieke Universiteit Leuven), Belgium
- Linguistic Data Consortium, USA
- Meedan, USA
- Mitre, USA
- School of Computing, University of Leeds, United Kingdom
- University of Constance, Germany

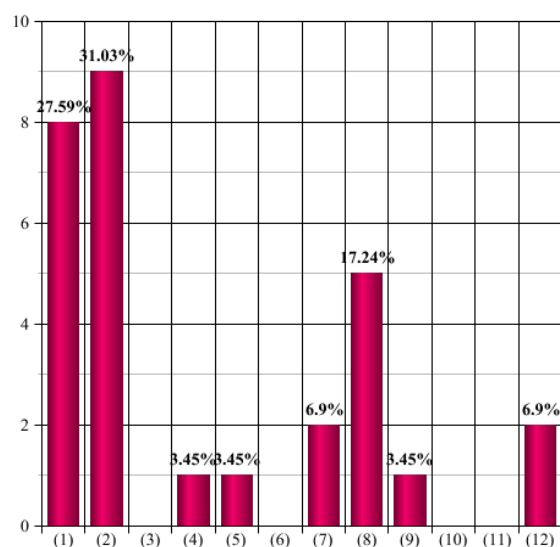
Those who indicated the type of institution they work for listed the following (multiple answers were allowed):

Type of institution	Answer Count
Company & for profit organization (1)	1
University (2)	9
Public Research center (3)	5
Others (4)	2



The main activity of the institution (respondents could choose more than one choice):

Answer	Count	Percentage
Software developer (1)	8	27.59%
Teaching/training organisation (e.g. university) (2)	9	31.03%
HLT Product Vendor (3)	0	0%
Culture/Museum (4)	1	3.45%
Technology Transfer institution (5)	1	3.45%
Minority language organisation (6)	0	0%
Content provider (7)	2	6.9%
Interpreting/Translating/Localisation (8)	5	17.24%
Telecommunications (9)	1	3.45%
E-commerce (10)	0	0%
Banking/Insurance (11)	0	0%
Other (12)	2	6.9%

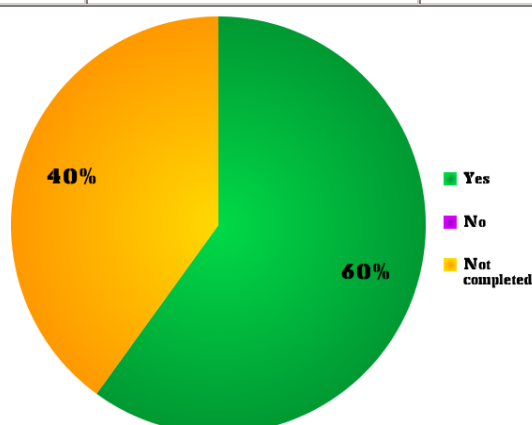


It is important to stress the fact that a large number of key sectors (e-content, Translation and interpretation, software integrator/developer) are represented.

4.3. *Involvement of the players in HLT & LRs:*

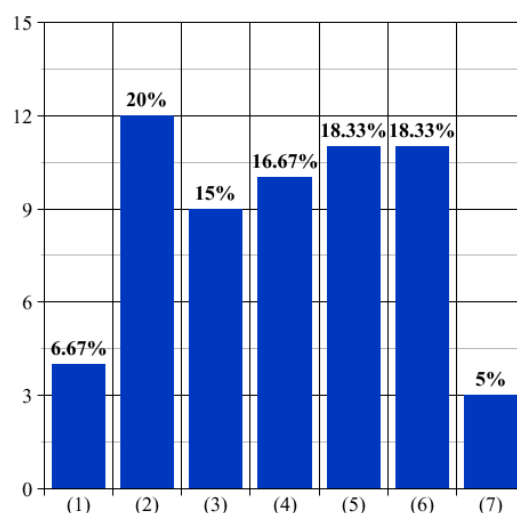
When asked about their involvement in the Human Language Technologies and in the Language Resources, we obtained the following answers:

Is your institution involved in Language Technologies		
Answer	Count	Percentage
Yes (Y)	18	60%
No (N)	0	0%
Non completed	12	40%



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

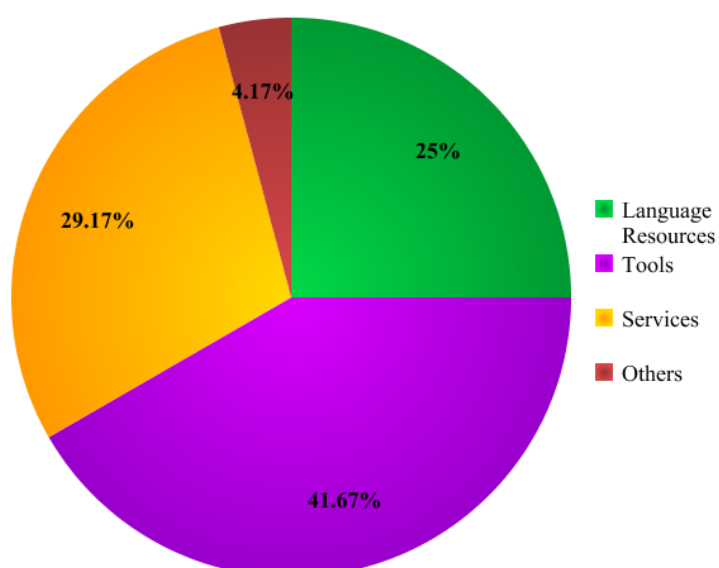
Involvement in HLT and related sectors			
Answer		Count	Percentage
Language learning	(1)	4	6.67%
Language Resources production	(2)	12	20%
Speech technologies	(3)	9	15%
Written technologies	(4)	10	16.67%
Search and knowledge mining	(5)	11	18.33%
Translation automation	(6)	11	18.33%
Other (Language Resources)	(7)	3	5%



The technologies listed as they were mentioned are not very specific and cover areas like resources (e.g., Verbnnet, LFG Grammar) or tools of various levels including TTS, search tools in particular for Quran, platforms for developing grammars, etc.

The main products and sector of activities as indicated by the different players are listed hereafter:

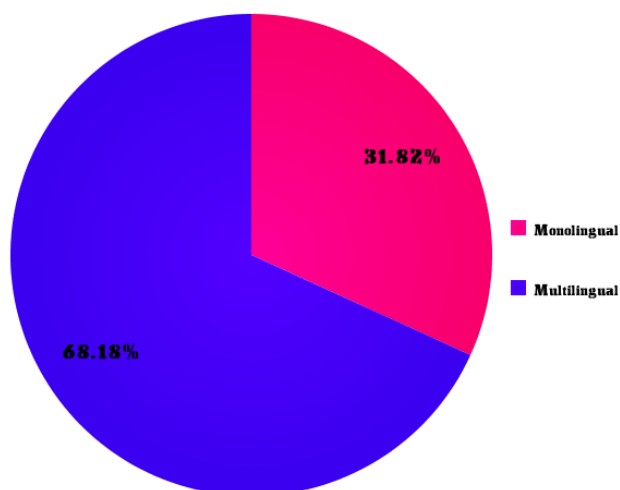
Institution's main products, tools or services		
Answer	Count	Percentage
Language Resources	6	25%
Tools	10	41.67%
Services	7	29.71%
Other	1	4.17%



4.4. Multilinguality issues

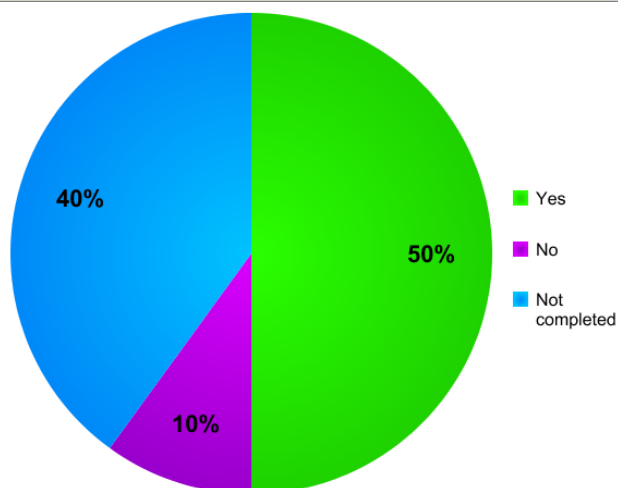
Another important issue is the Monolingual vs. Multilingual aspect of the products offered by the respondent:

Answer	Count	Percentage
Monolingual	7	32%
Multilingual	15	68%



When asked if they do they include the Arabic language, respondent replied

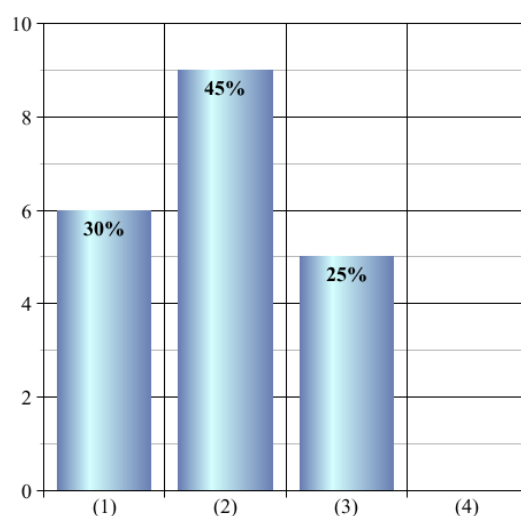
Answer	Count	Percentage
Yes	15	50%
No	3	10%
Not completed	12	40%



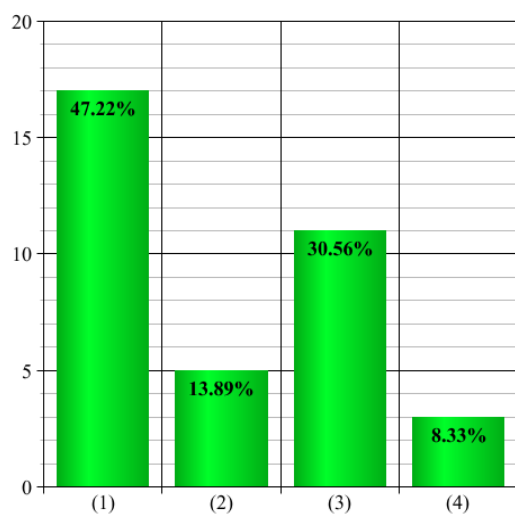
4.5. Information about the respondent's LRs

To the question on the Language Resources type we received the following answers (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).

Answer	Count	Percentage
Speech Resources (1)	6	30%
Written Resources (2)	9	45%
Multimedia/multimodal Resources (3)	5	25%
Others (4)	0	0%



Does the institution you represent use Language Resources		
Answer	Count	Percentage
that are produced internally (1)	17	47.22%
that are produced by specific contracted vendors (2)	5	13.89%
that are distributed by data centers? (3)	11	30.56%
Other (4)	3	8.33%



5. The most important question was about the needs for LRs

We list herein the replies that have been entered by the participants.

- Arabic dialectology
- Diglossia and code-switching
- Arabisation of foreign terms, concepts, jargon, acronyms, and expressions
- Design and population of bilingual dictionaries, glossaries, word-lists, and related references to reinforce language learning
- Foreign loanwords and absorption
- Computer-mediated communication (CMC)
- CALL and CBT , Arabic learners of L2 and acquisition
- Annotated corpora for Arabic
- Arabic Corpora to evaluate our tools of automatic abstracting, parsing, morphological analysis, etc.
- Audio visual materials at specific language proficiency levels are most needed. Graded Readings resources at all levels as well.
- Dictionary
- Multilingual MT software
- Material for dialects of Arabic
- Language-independent processing tools, corpora that are compatible with international standards, annotation interfaces
- Language resources form building language models; handwritten resources
- Lexical resources (dictionary), formal grammars for syntax,
- Lexicons Dictionaries and standards
- Open-source corpora, especially audio and video
- Parallel corpora and translation memory for training our translation engine, open licensed linguistic data.
- Pathological voice resources
- Training, development and evaluation data (speech and text) to build acoustic models, language models and translation models
- Open translation memory for Arabic and English that would be reliable

6. Follow-up of the online survey(s)

From this survey we can see the real emergence of an Arabic HLT community. In addition to the information regarding the players and experts that could be extracted, we also saw a consolidation of the urgent and expressed needs that now covers all areas of language processing: from speech and audio corpora on Arabic and colloquial Arabic(s) to lexica, visual and multimedia resources, tools for MT & IR, etc. This strongly confirms the needs identified previously and we do expect our knowledge-base to reflect such needs in a second release.

7. Appendix A: The MEDAR Knowledge-base

The outcome of all the surveys has been compiled and consolidated in the MEDAR “Knowledge-base” that is being made available to all and that comprises the identified experts, institutions, Language Resources & tools, etc. Such knowledge-base is available from the MEDAR web site (www.medar.info) (and mirrored at: http://www.elda.org/medar_knowledge_base/) and will be maintained and updated regularly.

8. Appendix A: The NEMLAR REPORT

This report is available at:
http://www.medar.info/The_Nemlar_Project/Publications/NEMLAR-REPORT-SURVEY-FINAL_web.pdf

9. Appendix B: The MEDAR Deliverable D3.1, A Survey of Actors, Products, Projects

This report is available at: http://www.medar.info/MEDAR_Survey_II.pdf