MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable D3.1.

Survey of actors, projects, products

Author: Khalid Choukri, ELDA
Contributors: Olivier Hamon, Djamel Mostefa, Mathieu Robin-Vinet, ELDA
December, 2008

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

CONTENT

1. EXECUTIVE SUMMARY 4
2. INTRODUCTION TO THE MEDAR SURVEY 4
3. THE MEDAR ONLINE SURVEY AND THE QUESTIONNAIRE STRUCTURE 5
4. THE MEDAR SURVEY: SUMMARY OF THE FIGURES AND FACTS 8
   4.1. IDENTIFICATION OF THE RESPONDENT: 8
   4.2. PROFILE OF THE RESPONDENT: 11
   4.3. INVOLVEMENT OF THE PLAYERS IN HLT & LRS: 15
   4.4. MULTILINGUALITY ISSUES 18
   4.5. INFORMATION ABOUT THE RESPONDENT'S LRS: 19
   4.6. SOME MARKET FIGURES 25
5. THE MOST IMPORTANT QUESTION WAS ABOUT THE NEEDS FOR LRS: 30
6. ANALYSIS OF THE ONLINE SURVEY 32
   6.1. A SHORT LIST OF KEY APPLICATIONS AS REPORTED WITHIN THE SURVEY: 32
   6.2. A SHORT LIST OF KEY RESOURCES AS INDICATED BY THE RESPONDENTS: 32
7. ADDITIONAL FINDINGS BY ELDA AND THE MEDAR PARTNERS 33
   7.1. MT FINDINGS 33
   7.2. CLIR/MLIR FINDINGS 37
8. LIST OF PLAYERS 38
   8.1. INSTITUTIONS 38
   8.2. INDIVIDUALS 39
9. CONTRIBUTIONS TO FULFILLING THE REMAINING "GAPS" AS DEFINED BY MEDAR 41
10. APPENDIX A: THE NEMLAR REPORT 41

1. Executive Summary

This is the pre-final report of the state of art of the situation of Human Language Technologies (HLT) for Arabic as drafted in December 2008. Following the work carried out within NEMLAR (www.nemlar.org , see the web site for more details at:
(http://www.medar.info/The_Nemlar_Project/Publications/NEMLAR-REPORT-SURVEY-FINAL_web.pdf)
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 in 2008 (since the first report of NEMLAR). It surveys the activities and projects, existing language resources and tools, important language resources and tools, as well as the experts. The Summary of the findings (facts & figures) is given herein.

The survey was launched on April 2008 and all partners were encouraged filling the questionnaire for their institution and having it filled by their partners. As of today 57 questionnaires were filled in. Some of them have been entirely completed (37), some 17 questionnaires are missing some of the answers but are still considered for their usefulness. A number of countries are well represented (e.g. 11 responses from Egypt, 10 from Morocco). A large number of 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 is related to the technologies our respondents feel important for Arabic and they listed a large set. Many of them consolidate our own finding listing MT, CLIR/MLIR, and ASR 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.

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

2. Introduction to the MEDAR Survey

This survey is carried out within the MEDAR project and aimed at providing an overview and an analysis of the situation 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.

So a 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.

3. 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 survey in more than one visit if they wish and the tool saves 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):

MEDAR Survey

MEDAR & NEMLAR

A Network for Euro-Mediterranean LAnguage Resource
and
Human Language Technology development and support

A Follow-up of a FIRST SURVEY ON HLT Experts
AND LANGUAGE RESOURCES
Conducted within the NEMLAR project in 2003

Dear Colleague,

Language Resources (LRs) are recognized as a central component of the linguistic infrastructure, necessary for the development of Human Language Technologies (HLT), and therefore for industrial development. Other purposes may be served by the availability of LRs such as content industry, cultural heritage safeguarding, etc. The availability of adequate LRs for as many languages as possible and, in particular, of multilingual LRs, is a pre-requisite for the development of a truly multilingual Information Society.

The issue of HLT based on Arabic language is getting more and more prominent; the lack, on the one hand, of useful resources, and, on the other hand, of real-world publicly available applications, highlights the need for improving R&D in this area and for promoting the use of HLT among the potential partners, in particular to safeguard some of the cultural heritages of this geographical area.

In many areas and business sectors, large companies produce their own resources for the languages for which some business can be made, and often no resources are built for the less "lucrative" languages.

In order to overcome such handicap, the NEMLAR project (February 2003 - July 2005) and its follow-up MEDAR (February 2008 - August 2010) would like to ensure that Arabic language obtains the necessary funds to produce the required resources and tools, and to make them widely available as for many other major languages.

In order to have a better picture of the Arabic HLT scene, the MEDAR project would like to update the data collected during the Nemlar project (Final Survey) and that helped promote the Arabic HLT in academic and industrial world as well as vis-à-vis the potential funding agencies.

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).

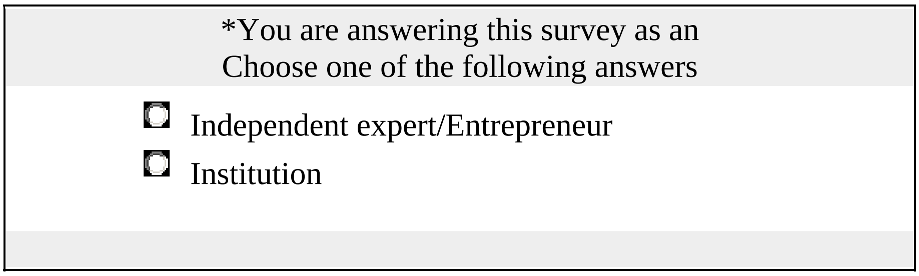
The final phase will aim at drafting a comprehensive report that may serve as the basis for a work plan about the needs for multilingual resources targeting customization of Machine Translation (including speech to speech translation), Cross-lingual information Retrieval, and other speech recognition 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 (Yes / No questions).

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

<u>Group 1 is the contact information (name, email, etc.)</u>

The final question of this group is:



*You are answering this survey as an
Choose one of the following answers

○ Independent expert/Entrepreneur
○ Institution

Group 2: Information about your institution and its language technology

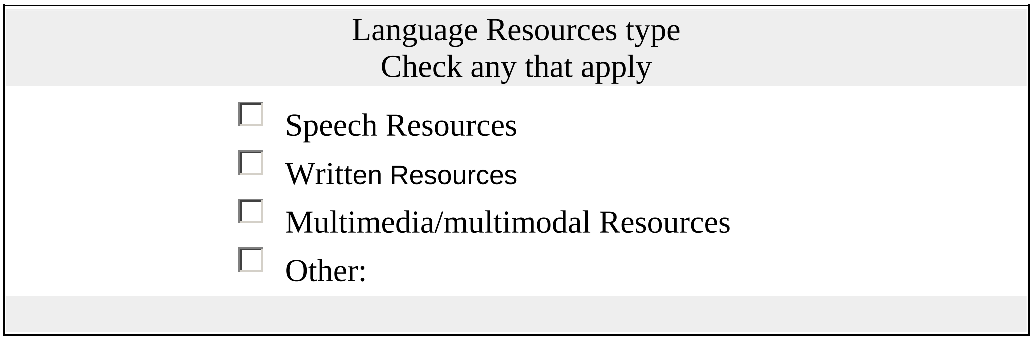
If you are not answering for an institution, please go back to the previous page using the "<<prev" button on this page and select "Independent expert/entrepreneur" at the last question.

This group 2 consists of 16 questions including
"Number of employees (directly or indirectly) involved in language technologies"
"Your institution's main activity (you may choose more than one choice)"
"Is your institution involved in Language Technologies"
"If Speech technologies", please specify"
"If Written technologies", please specify"
"What are your institution's main products and/or services (please list)"
"Do they include the Arabic language",
etc.

Group 3: Information about your language resources

Please select as many boxes as appropriate, please list the Languages whenever appropriate (e.g.; the ones containing Arabic) and Please add details about nature, size, etc. whenever appropriate and possible e.g. for a corpus of business documents, you may state it consists of 2 million words, Arabic-English dictionary, 50,000 entries, etc.)

This group (certainly the most important for our work on MEDAR) is subdivided into almost 20 questions, organized into a hierarchical "tree" like 3.1:



Language Resources type
Check any that apply

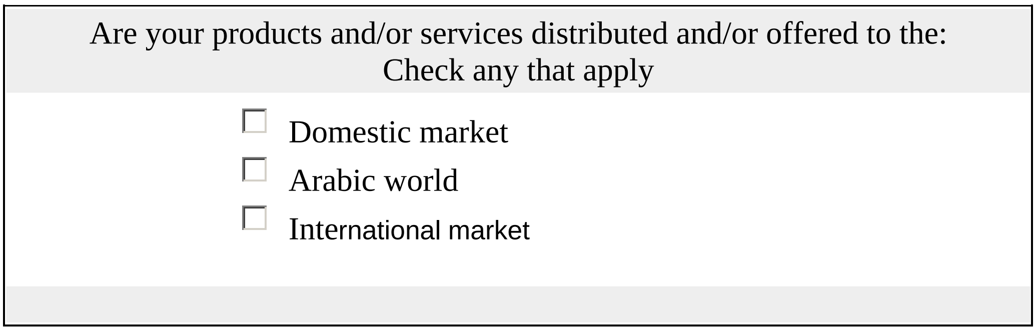
□ Speech Resources
□ Written Resources
□ Multimedia/multimodal Resources
□ Other:

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 (see details in the annexes).

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.

The first question (4.1) is:

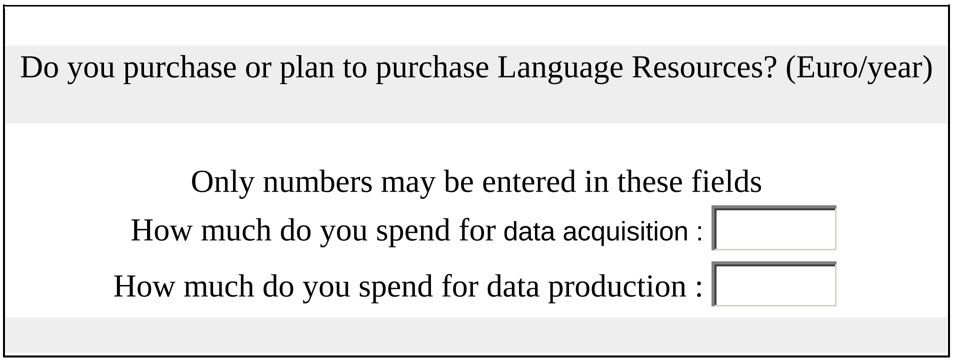


Are your products and/or services distributed and/or offered to the:
Check any that apply

□ Domestic market
□ Arabic world
□ International market

Question 4.2 is about partnership between the respondent institution and other players to identify potential cooperations with the Arabic world organizations. A question was asked about the names of such partners.

The following question (4.3) is also directly impacting our project as it asks about the financial plans of our experts:



Do you purchase or plan to purchase Language Resources? (Euro/year)

Only numbers may be entered in these fields

How much do you spend for data acquisition :

How much do you spend for data production :

Group 5 is about the needs for LRs:

This is the crucial group of the future tasks of the project. It requests information about the potential expectations of the surveyed expert if he/she had to decide (the answers are free texts)
5.1. Which Language Resources should be available?
5.2 For which applications?
5.3. Which design and structure of Language Resources in general would you prefer?

Group 6 is request for more contacts

Group 6 is a request to give us more contacts so we can circulate the questionnaire widely.

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

Thanks to the involvement of all MEDAR partners we managed to obtain 54 responses for this survey (37 full responses, 17 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 44 responses.

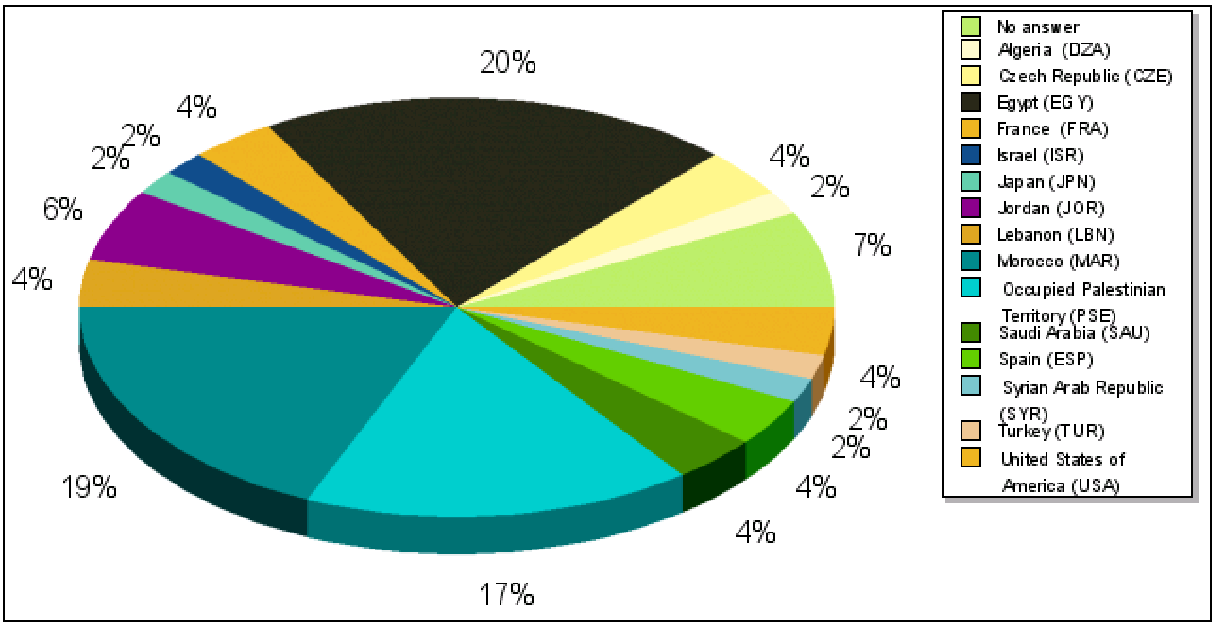
4.1. Identification of the respondent:

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:



The details per country are:

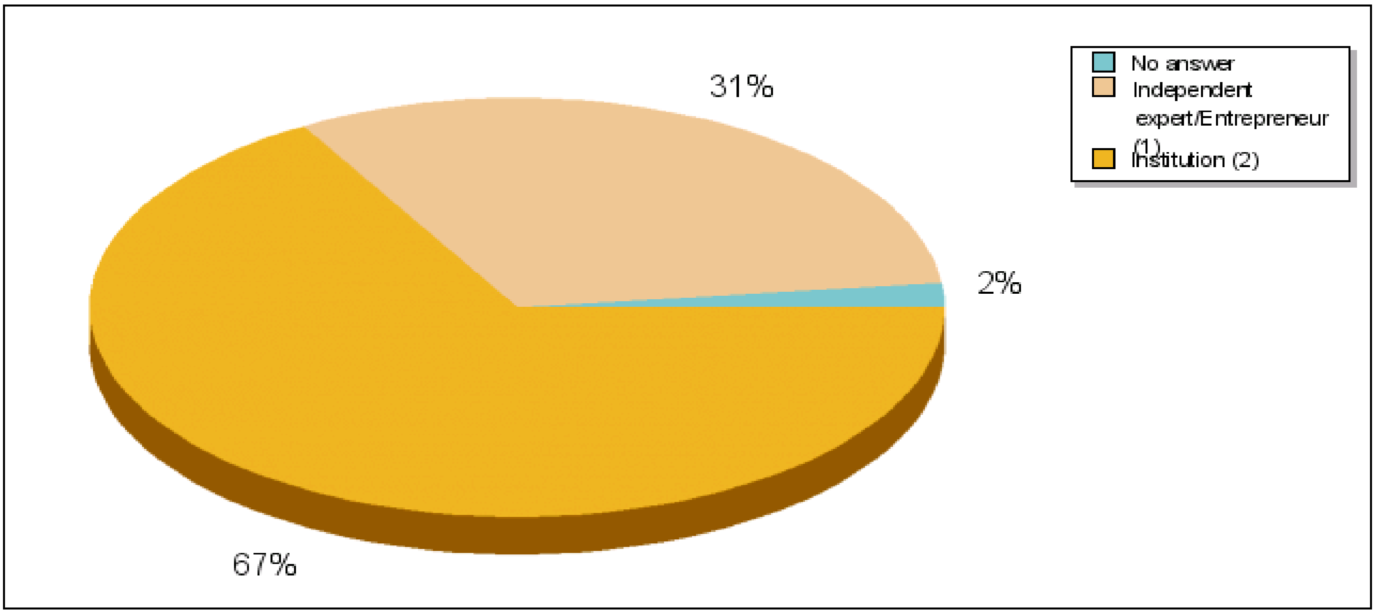
Answer	Count	Percentage
Egypt (EGY)	11	20.37%
Morocco (MAR)	10	18.52%
West Bank & Gaza Strip (PSE)	9	16.67%
Jordan (JOR)	3	5.56%
Czech Republic (CZE)	2	3.70%
France (FRA)	2	3.70%
Lebanon (LBN)	2	3.70%
Saudi Arabia (SAU)	2	3.70%
Spain (ESP)	2	3.70%
United States of America (USA)	2	3.70%
Algeria (DZA)	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.) 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.

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 & LRs.

No answer	1	1.85%
Independent expert/Entrepreneur (1)	17	31.48%
Institution (2)	36	66.67%
Non completed	0	0



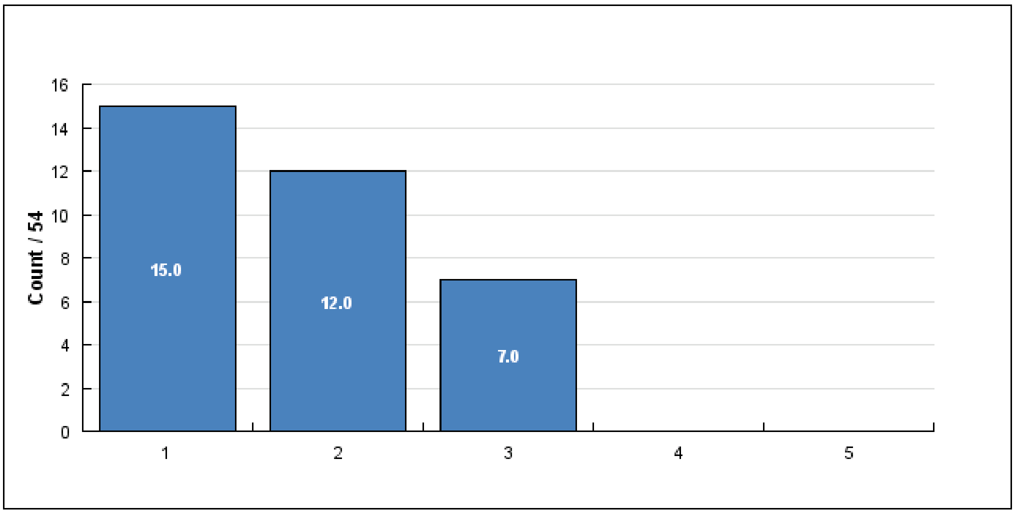
In addition to the individuals that replied without mentioning explicitly their institution, the following ones were listed:

- ACS TechnoCenter
- AlKhawarizmy Language Software
- Arab Academy for Banking and Financial. Sciences,
- ARABIC TEXTWARE
- Arabize;
- Cairo Microsoft Innovation Center in Egypt (CMIC)
- COLTEC
- Columbia University Center for Computational Learning Systems
- CRSTDLA (Scientific & technical Research Center for Arabic Language Development)
- ELDA
- ENSIAS
- European Trading & Technology (eurotec)
- Faculty of Computers and Information, Cairo University
- France Telecom R&D ORANGE Labs
- GIS Int.
- Higher Institute for Applied Sciences and Technology (HIAST)
- IBM (IBM Egypt)
- Indiana University

- Insan Center
- Institute for the Studies and Researches on Arabization (Institut d\'Etudes et Recherches pour l\'Arbisation)
- Isra\' Software & Computer Co. Ltd
- IT College / Birzeit University
- King Abdulaziz City for Science and Technology
- Laboratoire de Recherche en Informatique et Telecommunications Faculte des Sciences
- Lebanese University
- ManarahNet Modern Software Co.
- Millennium Technology
- RDI
- TALP Research Center - Universitat Politecnica de Catalunya
- The CJK Dictionary Institute
- Unit for Learning Innovation- Birzeit University
- University Cadi Ayyad - Faculty of Sciences, Marrakech, Morocco

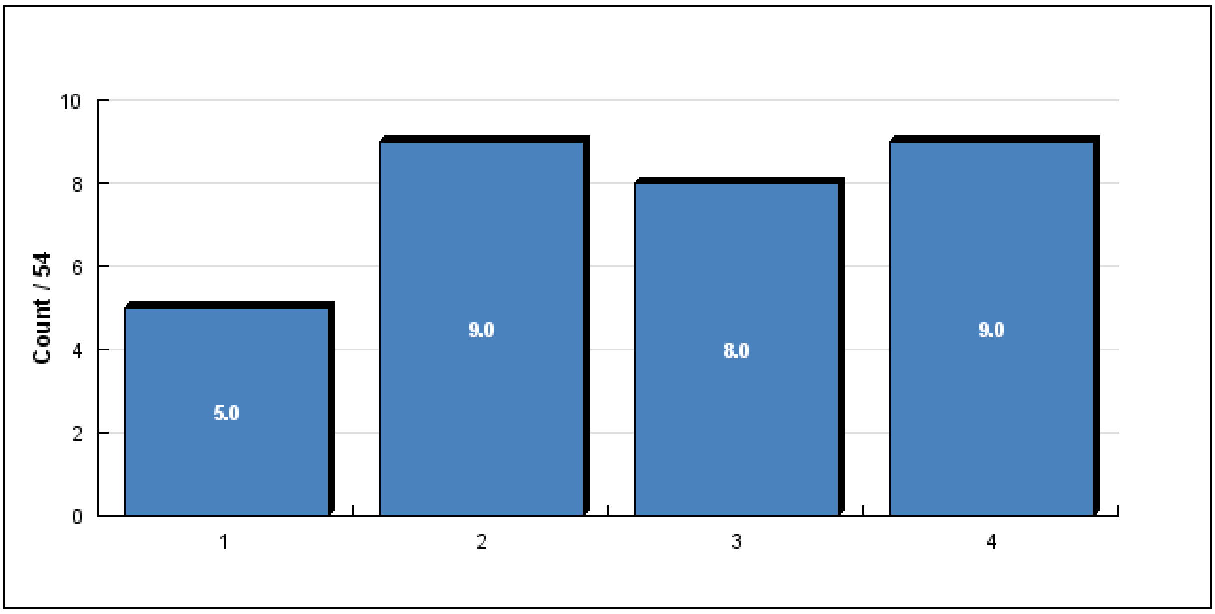
Those who indicated the type of institution they work for listed the following:

Type of institution	Answer Count
Company & for profit organisation (1)	15
University (2)	12
Public Research center (3)	7
Public organisation (4)	0



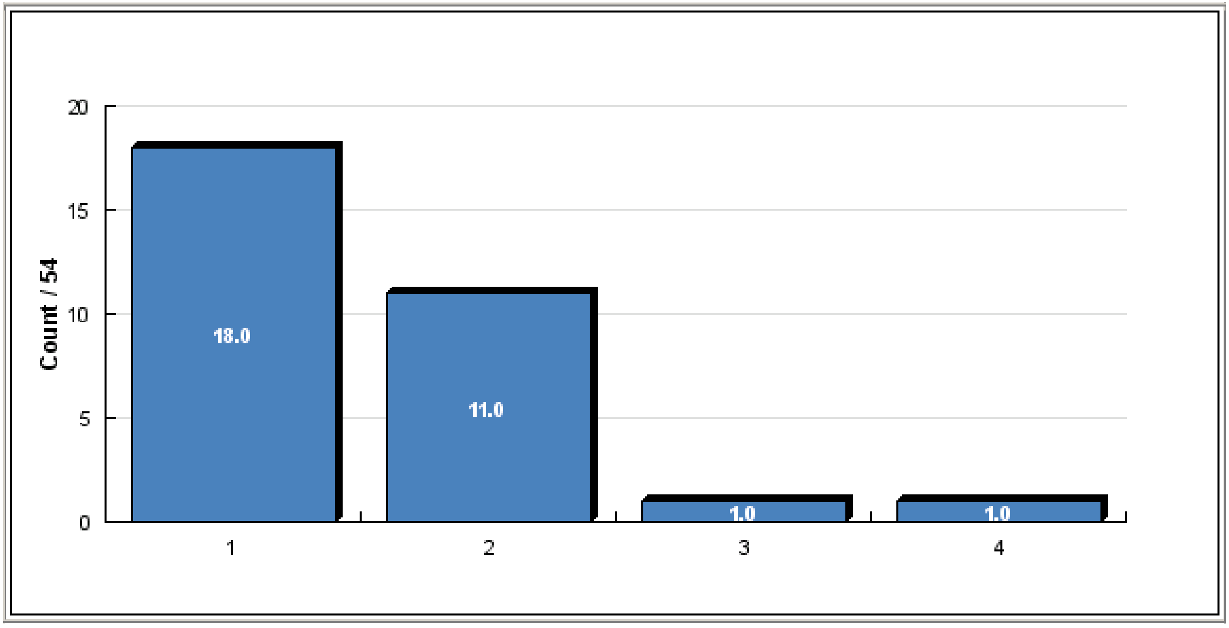
An important question about the number of employees in total versus those who are involved in HLT:

Number of employees	Answer	Count
Less than 10	(1)	5
10-49	(2)	9
50-99	(3)	8
Over 100	(4)	9



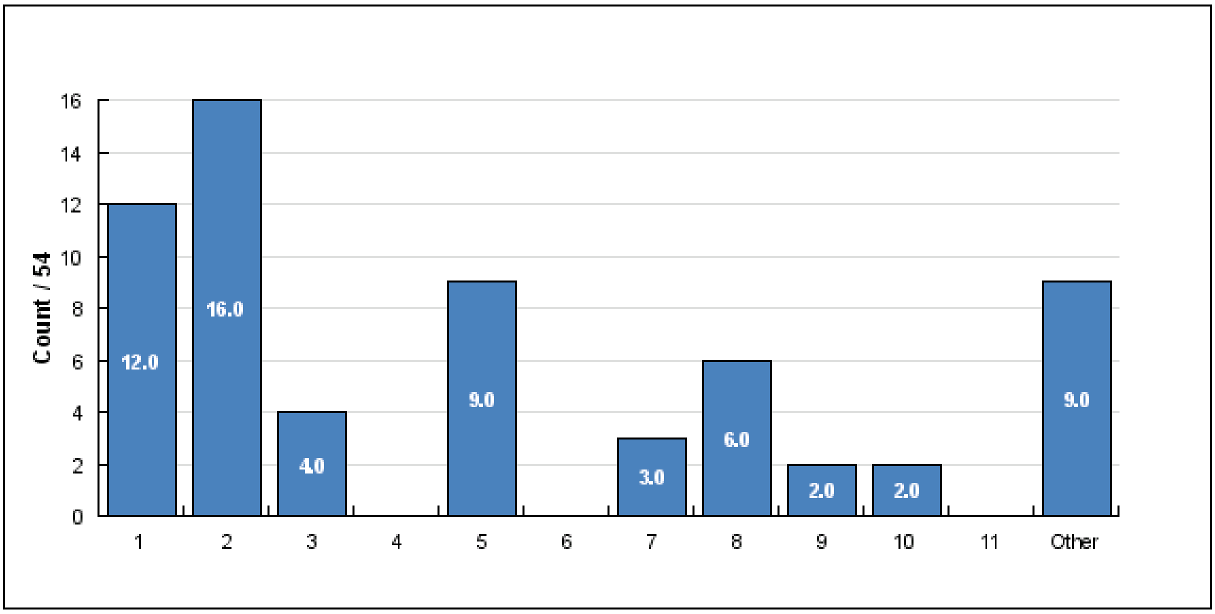
When focusing on the HLT & LRs sectors we obtained the following answers:

Number of employees (directly or indirectly) involved in language technologies	
Answer	Count
Less than 10 (1)	18
10-49 (2)	11
50-99 (3)	1
Over 100 (4)	1



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

Answer	Count	Percentage
Software developer (1)	12	22.22%
Teaching/training organisation (e.g. university) (2)	16	29.63%
HLT Product Vendor (3)	4	7.41%
Culture/Museum (4)	0	0
Technology Transfer institution (5)	9	16.67%
Minority language organisation (6)	0	0
Content provider (7)	3	5.56%
Interpreting/Translating/Localisation (8)	6	11.11%
Telecommunications (9)	2	3.70%
E-commerce (10)	2	3.70%
Banking/Insurance (11)	0	0
Other	9	16.67%



If we browser through the "Other" responses we have:

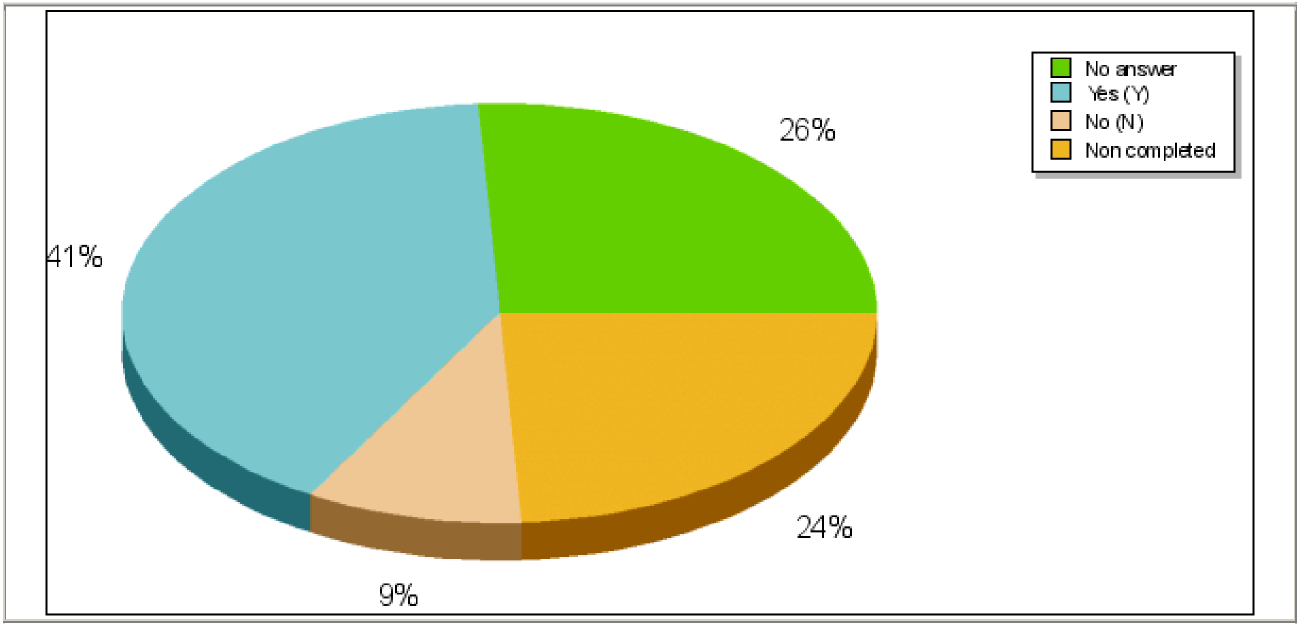
- Center of Excellence for Data Mining and Computer Modeling (DMCM)
- lexical databases development
- Technology Development
- Doing and supporting research within Saudi Arabia
- Research activities
- Database
- legal databases
- computers maintenance

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

4.3. Involvement of the players in HLT & LRs:

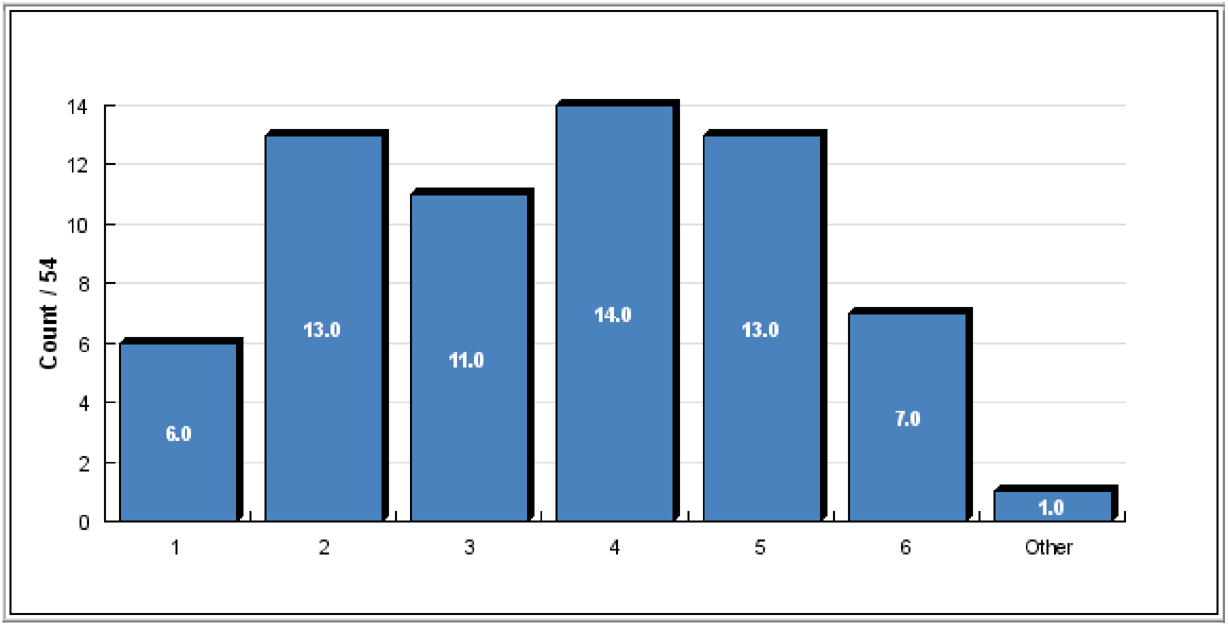
When asked about their involvement in HLT and LRs, we obtained:

Is your institution **involved in Language Technologies**		
Answer	Count	Percentage
No answer	14	25.93%
Yes (Y)	22	40.74%
No (N)	5	9.26%
Non completed	13	24.07%



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)	6	11.11%
Language Resources production	(2)	13	24.07%
Speech technologies	(3)	11	20.37%
Written technologies	(4)	14	25.93%
Search and knowledge mining	(5)	13	24.07%
Translation automation	(6)	7	12.96%
Other (Language Resources)		1	1.85%



The list of technologies as they were mentioned (with the duplicates) is listed herein:

- LRS for speech technology development:
- Text to Speech
- ASR, TTS, Speech Verification for assisting in the self learning of spoken language
- Speech synthesis, speech recognition using open source
- TTS, Speech Recognition
- Speech Recognition and TTS
- Automatic Speech Recognition, Text to Speech, Speaker Verification
- Speech synthesis and speech recognition
- Speech recognition, speech synthesis, machine translation

- LRS for language (text) technology development
- Morphological analysis, PoS tagging, phonetic transcription, lexical semantics, text search, text mining, - Arabic omni font written OCR
- Linguistic Processing
- Morphology, syntax
- OCR
- Arabic Spell/Grammar Checking, Arabic-to-European language transliteration, semantic analysis
- Writing Scorer, Auther Verification, Mining, Translation
- Online handwriting recognition
- Spell Checker, Morphological analyzer
- Natural Langiuage Processing, Machine Translation, Knowledge Representation, Information extraction and question-answer systems
- Semantics
- Research
- Infra structure for Arabic language esp. morphological, PoS taging, and Semantic analysis layers serving for sharpening IR and TM over gigantic Arabic content.

The main products and sector of activities as indicated by the different players are listed herein (without rephrasing or summarizing them):

- Language resources ; Technology Evaluation Services
- Arabic Lexical Semantic Database - News Tracking System (www.Alzoa.com) - Tutor for Arabic Hand Writing - Tutor for Arabic pronunciation (research in progress)
- Arabic ASR (Labeeb) - Arabic TTS (ArabTalk) - Arabic Speech verification (Hafss) - Arabic Morphological Analyzer (Arab Morpho) - Arabic PoS Tagger (Arab Tagger) - Arabic Phonetic Transcriptor (Arab Diac) - Arabic Text Search Engine (Swift) - Arabic Lexical Semantic Analyzer (ALSA) - Arabic omni font written OCR (Clever Page)
- Arabic speech synthesis using diphones and "demi-syllable" Arabic speech recognition using sphinx4
- Araterm CD (multilingual terminological Database) - Aragen CD (general language Database) - terminological lexicons (Banque de données terminologiques Dictionnaires electroniques
- Teaching/training/search/ learning
- Works - (reviews) : linguistic Research
- Arabic Speech Recognition Systems - Arabic TTS - A2E and E2A Machine Translation Systems - Information Retrieval and Extraction
- Arabic Speech recognition Arabic text to speech system Speech Databases , Arabic

name Romanization system

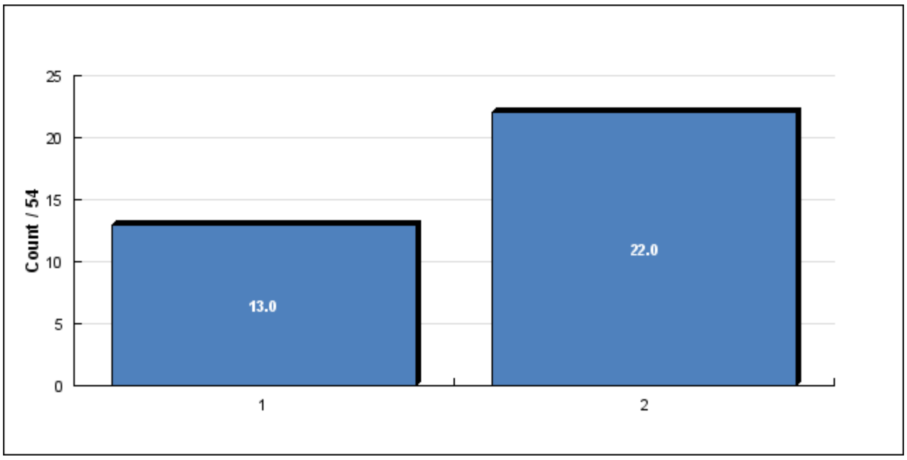
- Accounting Programs Clinics Programs Employee Programs Auto Respond For Telephone Software Web Site Development
- Research activities
- Name variant databases (Arabic, Japanese, Chinese) Phonological database (Japanese) Place name databases (Japanese, Chinese) Orthographic databases (Japanese, Chinese)
- Legal databases legal studies and researches
- ICT private sector activity
- e- enabled curricula Training Evaluation Research Networking
- PhD program in Computational Linguistics
- Information Retrieval (IR), Collaborative Content Services (CCS), Digital Content Services (DCS): DCS is a set of Services built to take advantage of the recent increase in digitized content and books, Research on Information Extraction, Natural Language Processing and Information Retrieval.

4.4. Multilinguality issues

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

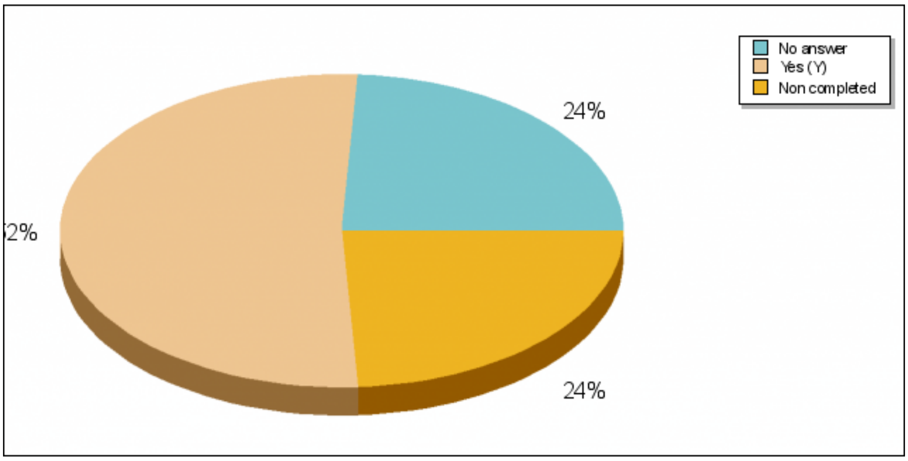
Are your products and or services:

	Answer	**Count**
Monolingual	(1)	13
Multilingual	(2)	22



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

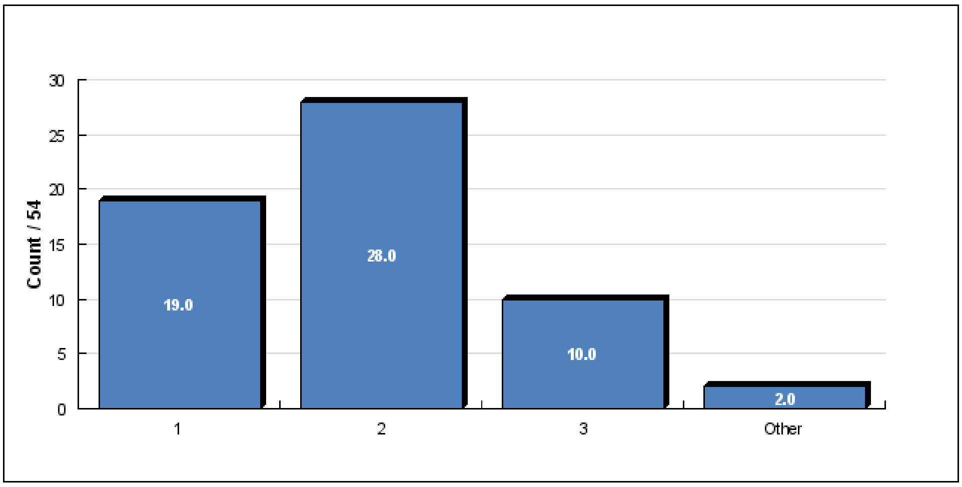
Yes (Y)	28	51.85%
No (N)	0	0
Non completed	13	24.07%
No answer	13	24.07



4.5. *Information about the respondent's LRs:*

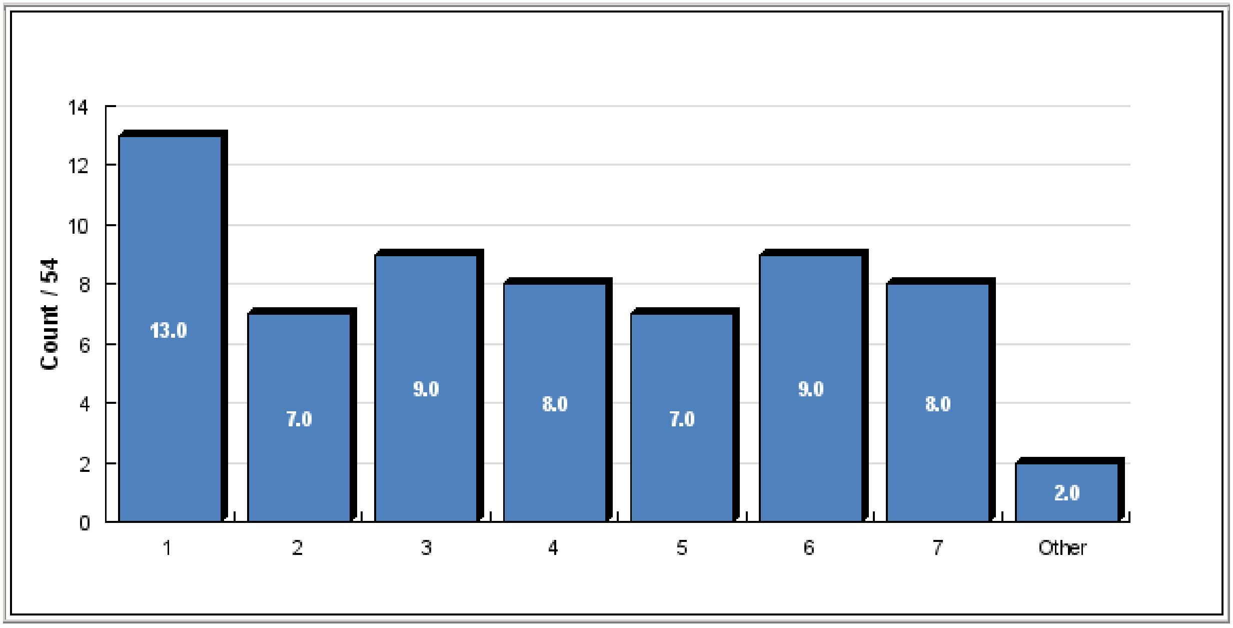
To the question on the Language Resources type we received the following answers:

Answer	Id	Count	Percentage
Speech Resources	(1)	19	35.19%
Written Resources	(2)	28	51.85%
Multimedia/multimodal Resources	(3)	10	18.52%
Other (e.g. biometric data)		2	3.70%

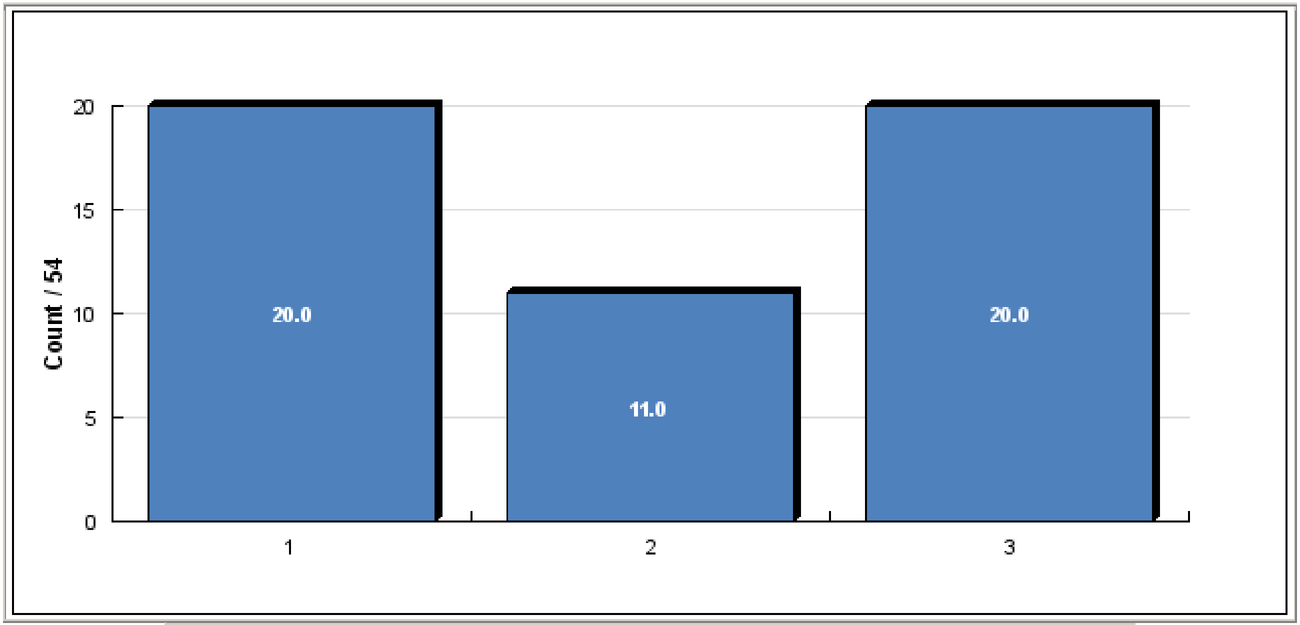


When asked for details about the types of LRs used by the respondent along modalities e.g. speech, text lexica, text corpora, other modalities, we received the following answers (more than one answer):

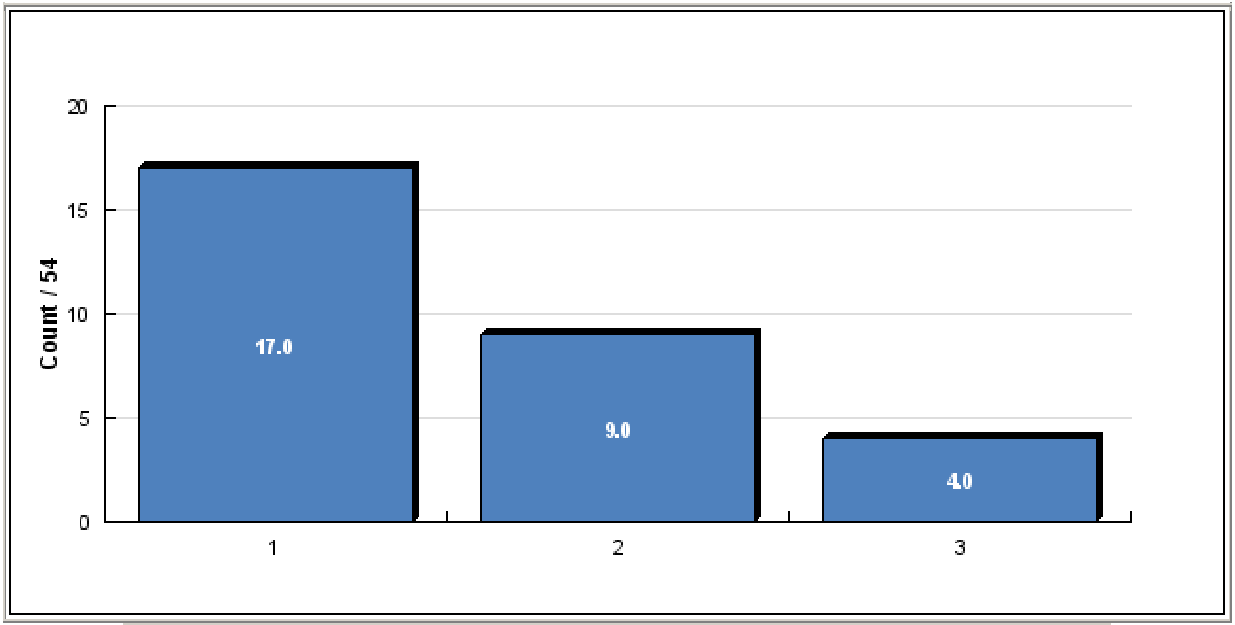
If "Speech Resources" please select		
Answer	Count	Percentage
Broadcast news & conversational speech (1)	13	24.07%
Fixed telephone (2)	7	12.96%
Mobile telephone (3)	9	16.67%
Micro/desktop speech (4)	8	14.81%
In-car recording (5)	7	12.96%
Read newspaper texts (6)	9	16.67%
Pronunciation/phonetic lexica (7)	8	14.81%
Other	2	3.70%



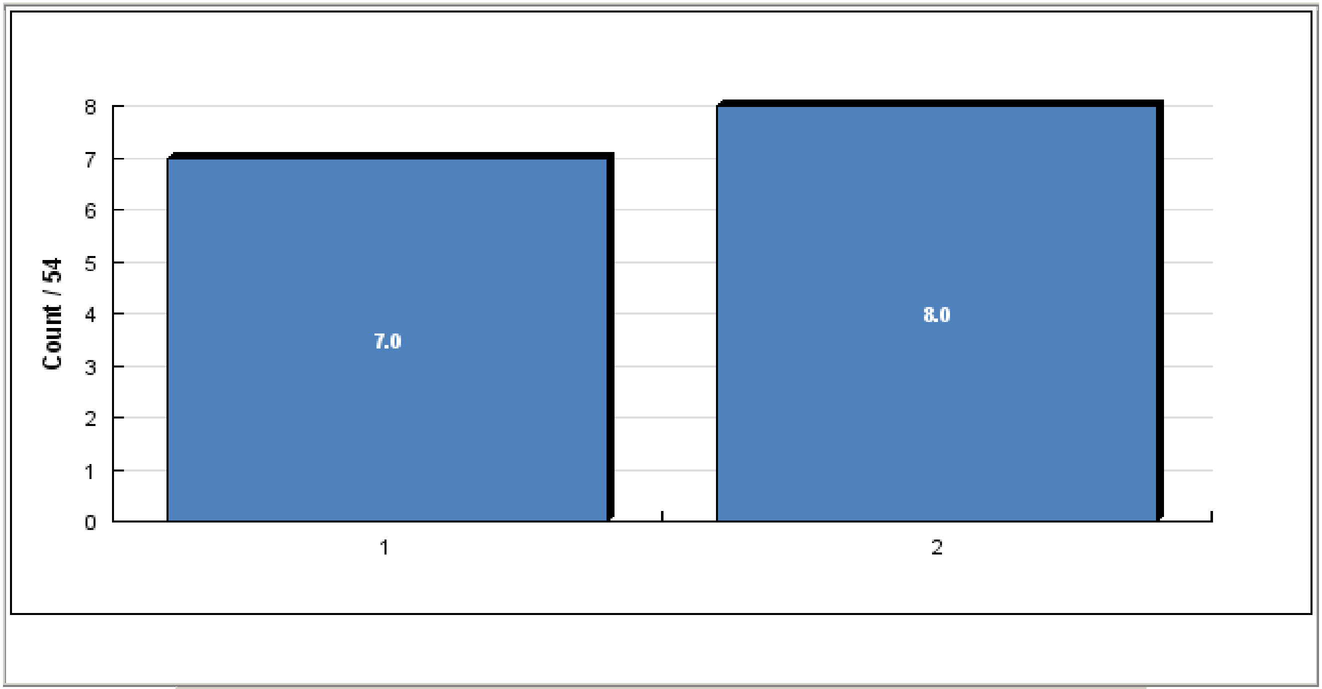
If "Written Resources", please select	
Answer	Count
Lexical databases (1)	20
Terminology and specialized dictionaries (2)	11
Text Corpora (3)	20



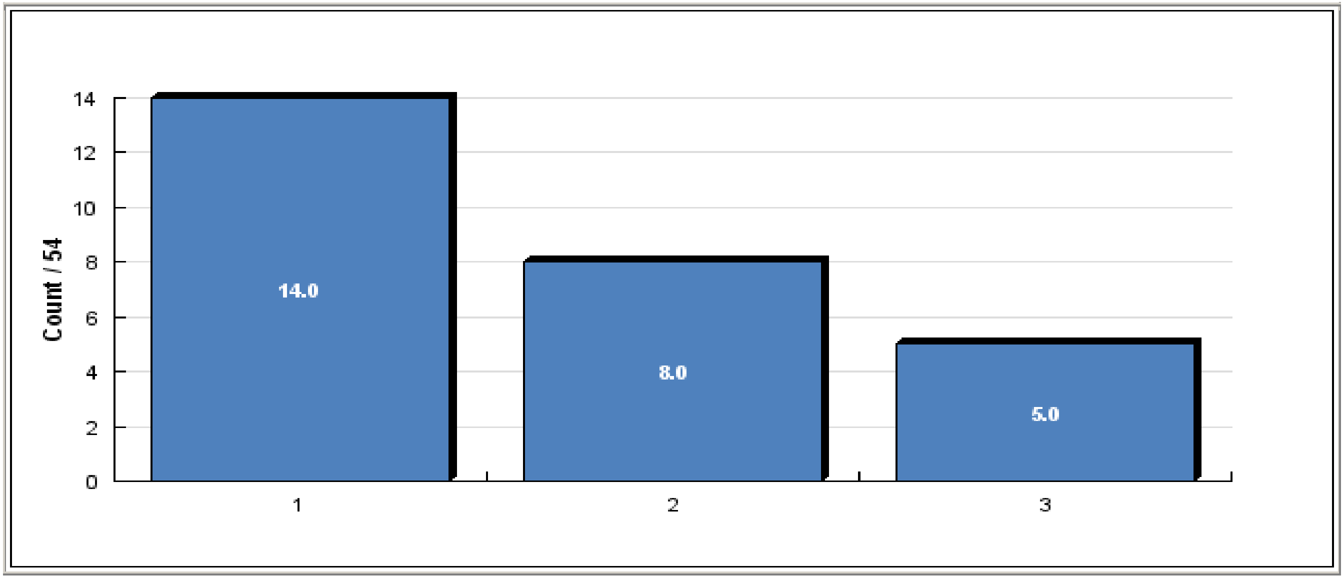
If "Lexical databases", please select	
Answer	Count
Monolingual lexical databases (1)	17
Multilingual lexical databases (2)	9
Onomastica (proper and geographical name lexical) (3)	4



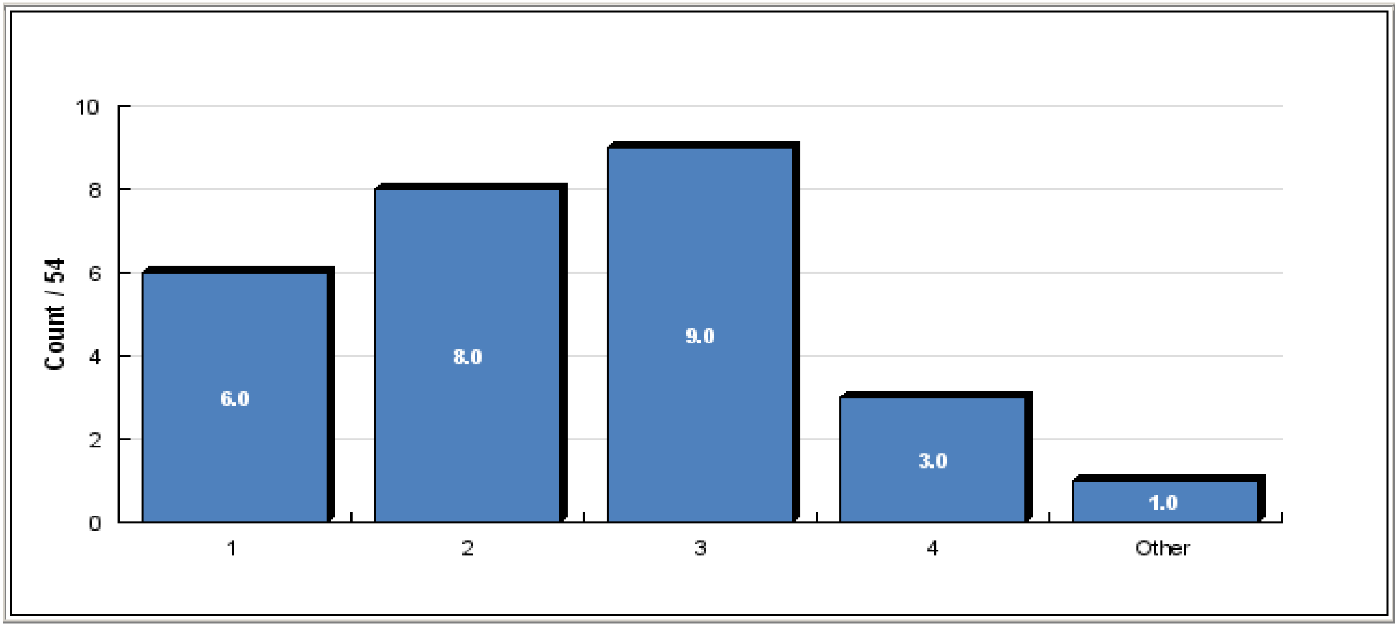
If "Terminology and specialized dictionaries"	
Answer	Count
Monolingual terminology databases (1)	7
Multilingual terminology databases (2)	8



If "Text Corpora", please select	
Answer	Count
Monolingual text corpora (1)	14
Multilingual and parallel text corpora (2)	8
Multilingual and Aligned text corpora (3)	5

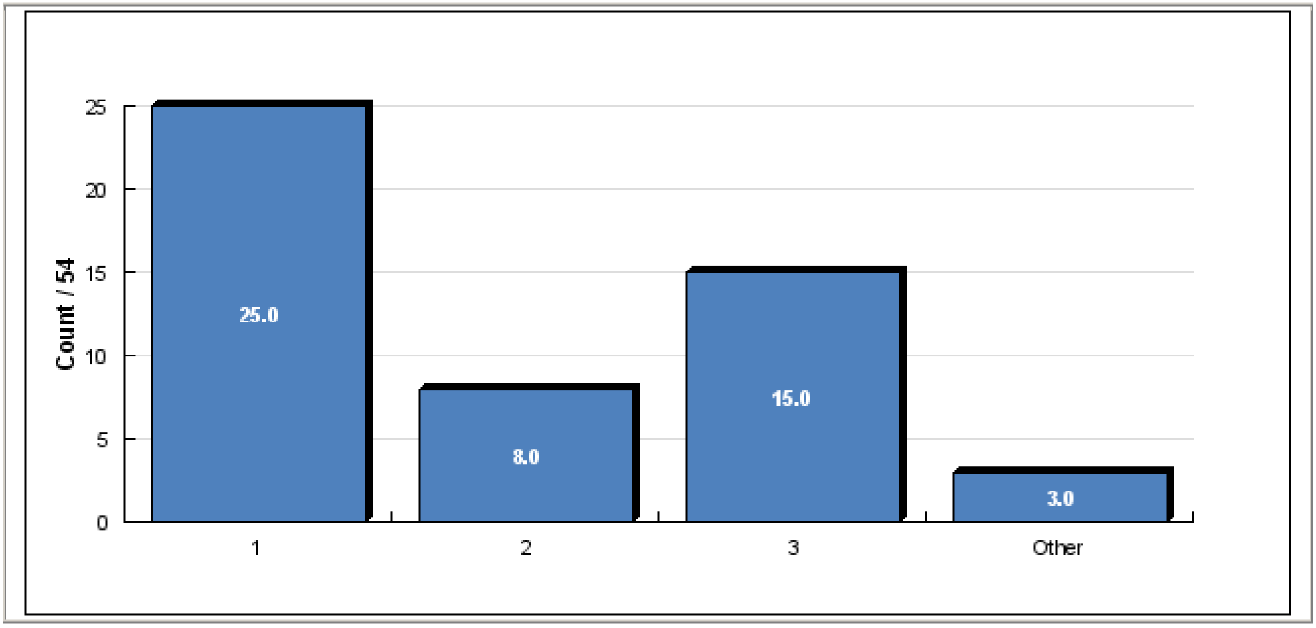


If "Multimedia/multimodal Resources", please select	
Answer	Count
Face (1)	6
Image (2)	8
Video (3)	9
Finger prints (4)	3
Other	1



Regarding the sources of the LRs used by the respondent:

Answer	Count
that are produced internally ? (1)	25
that are produced by specific contracted vendors ? (2)	8
that are distributed by data centres ? (3)	15
Other	3



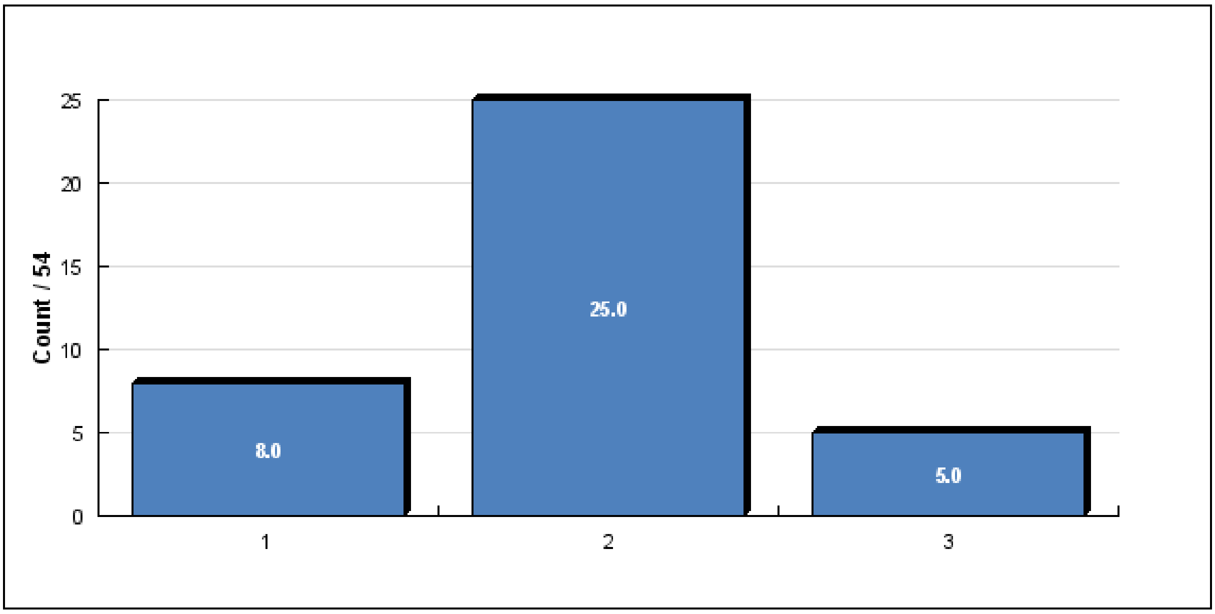
Those who replied to our question regarding production of resources were asked about the tools they use to design and produce LRs, the following ones were mentioned:

- Speech recording platforms
- Lexical Semantic Database: data are acquired from linguistic experts then represented and organized electronically using exiting Data Base Management Systems.
- Other resources are collected and organized by locally developed tools
- Internal tools for written Arabic text annotation (Fassieh); morphogical, phonetic, PoS tagging, semantic, ...
- Internal tools for segmenting speech. (Some of them are built on HTK)
- Cool Edit for the preparatory stages and the monitoring of speech signals.
- MS Office & SQL server are suite is also used at some phases of the production of both kinds of resources.
- Arabic speech recognition using open source tools
- Scanner hp Printer hp 1100
- Matlab
- Audio recording and editing public tools
- HTK HMM
- IBM Internal tools
- TrEd - TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures (http://ufal.mff.cuni.cz/~pajas/tred/index.html)
- Off the shelf tools esp. MS-Office suite esp. MS-Word and MS-Access.
- Plain text editors like MS-Windows\'s NotePad.
- MatLab, Praat, Delphi, Java, PHP/MySql
- lexicon, structure analysis
- Internal text annotation tool.
- Gwave
- MS-Office tools. Translation memories.

When asked about the standards & best practices they follow to design and produce LRs, the replies are:

Do you follow specific standards?

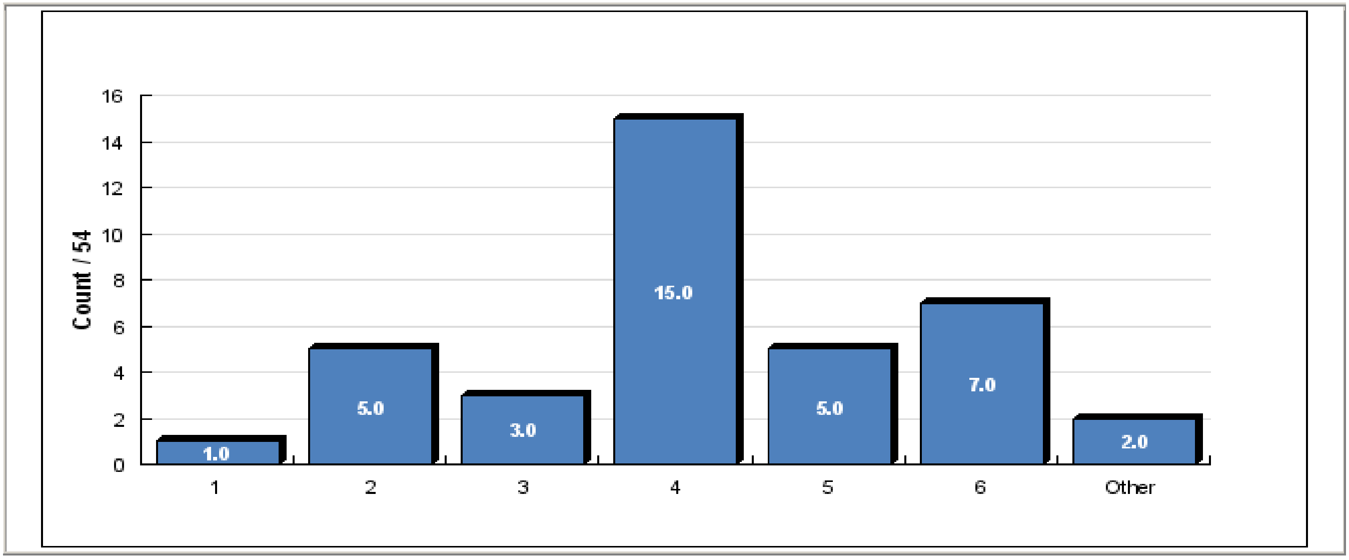
Answer		Count
None	(1)	8
Internal specifications	(2)	25
External specifications	(3)	5



4.6. *Some Market figures*

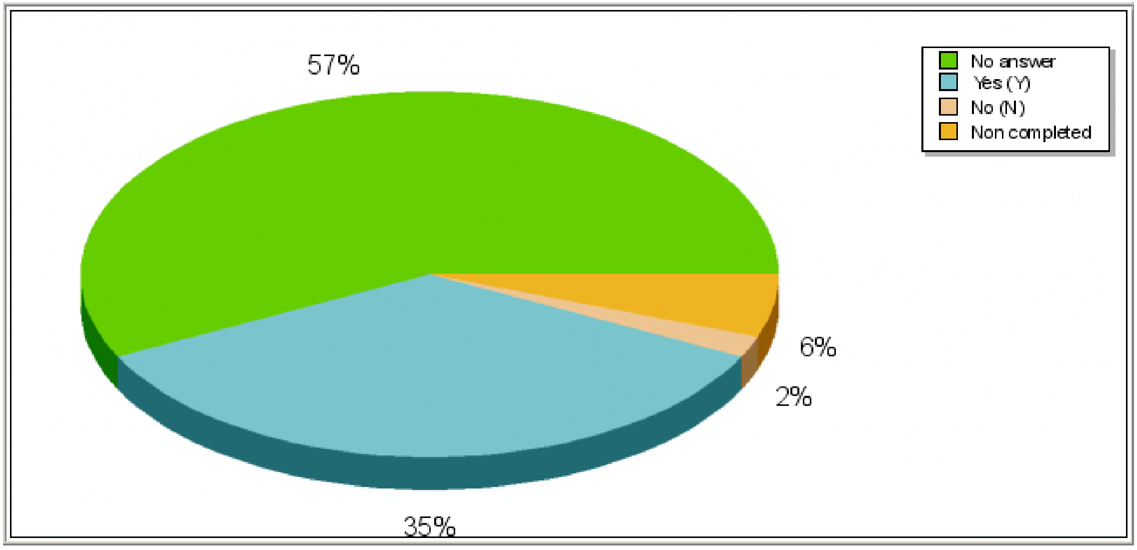
The survey also asked questions about the respondents' visions on their needs and on the market. When asked how often they review their needs for LRs:

How often do you re-evaluate your Language Resources needs and seek available databases?	
Answer	Count
Monthly (1)	1
Once per quarter (2)	5
Once per semester (3)	3
Once per year (4)	15
Once every 1-2 years (5)	5
Never (6)	7
Other	2



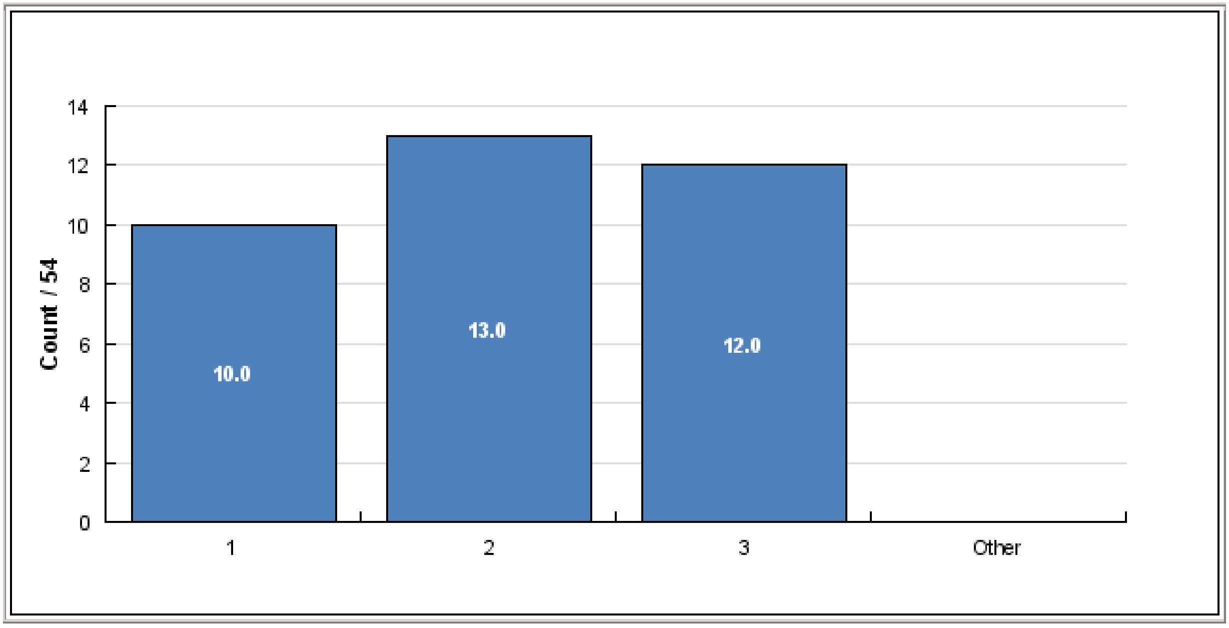
Regarding the important issue of distribution, we got:

Would you be willing to make your resources available to others according to a negotiated distribution agreement?		
Answer	Count	Percentage
No answer	31	57.41%
Yes (Y)	19	35.19%
No (N)	1	1.85%
Non completed	3	5.56%



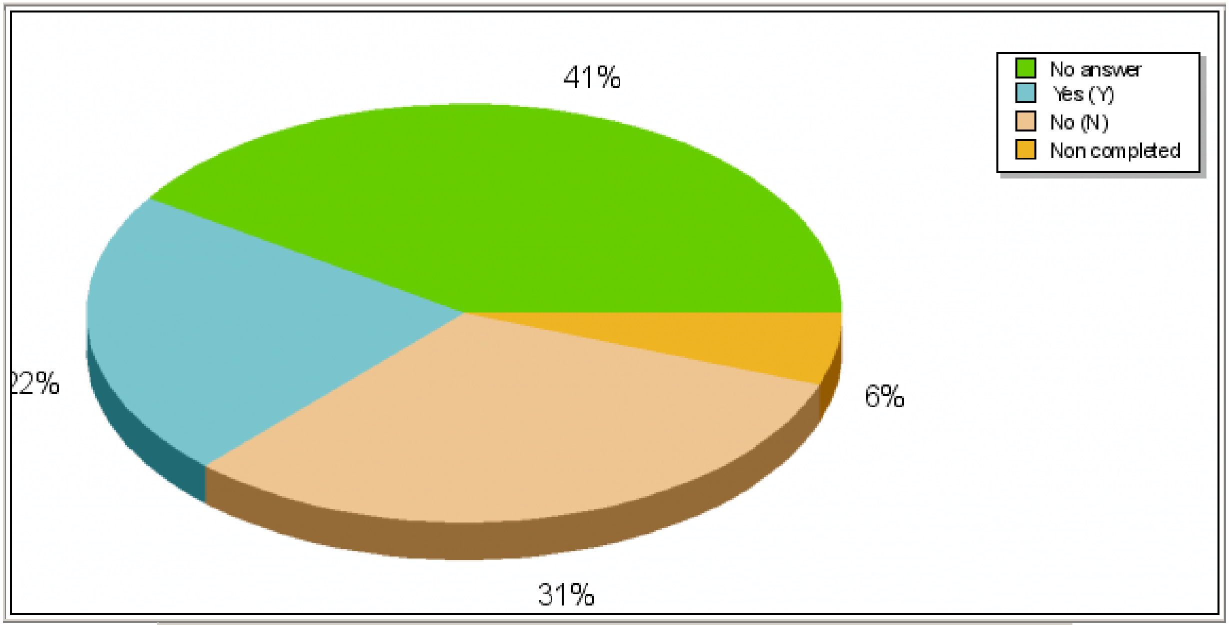
And those who answered positively to the question on distribution, were very specific on whom they would agree to supply their data:

If "Yes", whom would you be ready to license your Language Resources to?	
Answer	Count
End-users (1)	10
Tool developers (2)	13
Researchers (3)	12

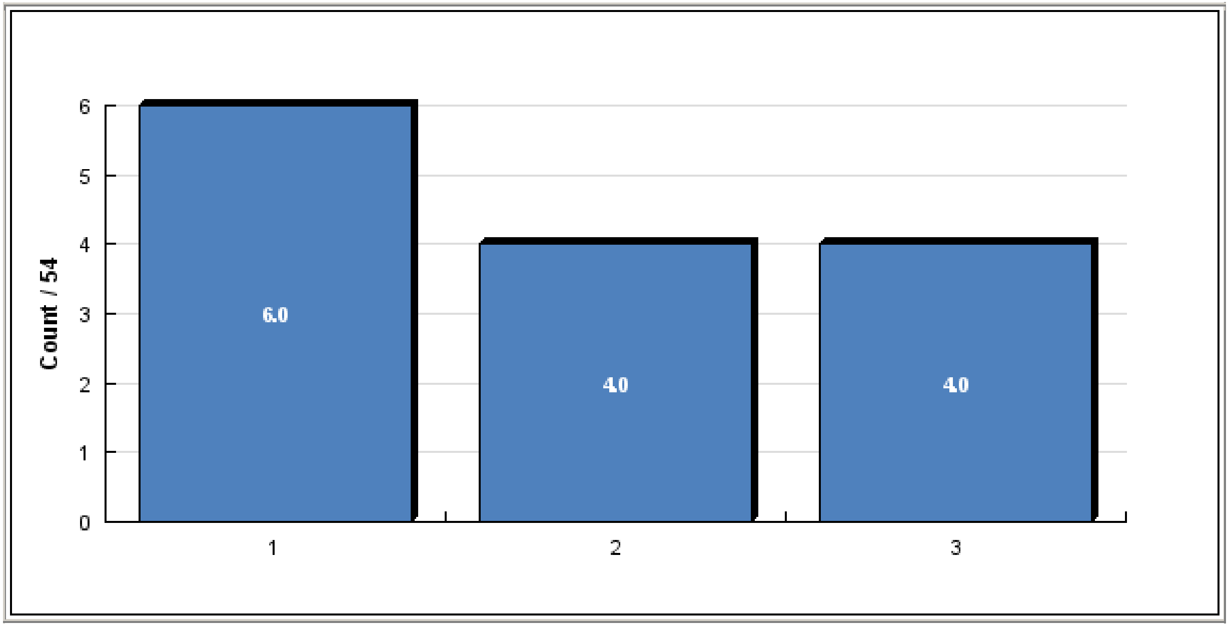


And when they did answer No to the question related to distribution, only one respondent gave the reasons behind that as "strategic".

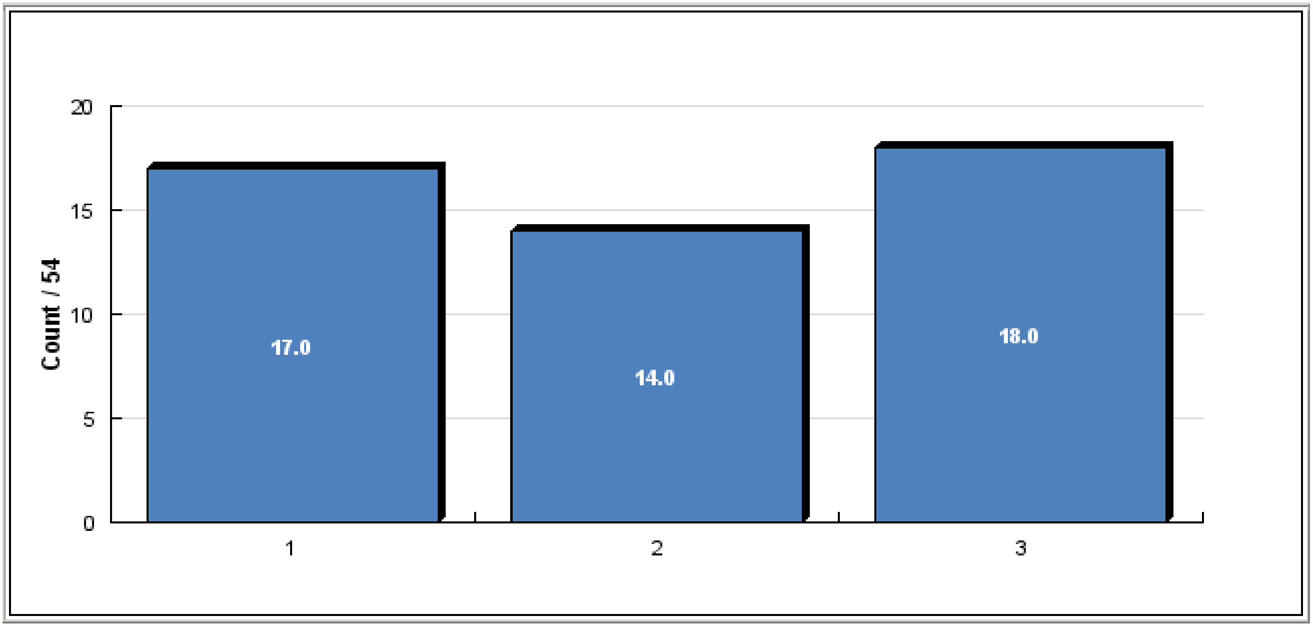
Is/was your institution involved in **any Language Resources** project?		
Answer	Count	Percentage
No answer	22	40.74%
Yes (Y)	12	22.22%
No (N)	17	31.48%
Non completed	3	5.56%



If "Yes", is/was it a project aiming at	
Answer	Count
Language resources production for your own use, please specify the causes, aims etc: (1)	6
Language resources packaging for others, please specify: (2)	4
Other, please specify: (3)	4



Are your products and/or services distributed and/or offered to the	
Answer	Count
Domestic market (1)	17
Arabic world (2)	14
International market (3)	18



Regarding the plans for purchasing LRs, the budgets seem to be steady over the next a few years:

How big is your expected purchasing budget for Language Resources? (Euro/year)		
Calculation	Now	3-5 Years
Nb of respondents	9	10
Average	10100	12950
Maximum	40000	50000

<u>When asked about the LRs market, the replies are hardly exploitable.</u>
The replies were:

- In Egypt (Few tens of Millions of USD; there are many producers with few purchasers)
- In the Arab World (few milliards of USD) (!! Probably billions?)
- Talking about Arabic 1000000
- Growing rapidly and will improve, since the whole world is communicating fast through the internet.
- International LR's market is a tens-of-Billion of USD's market. This market in the middle-east is very tiny compared to such numbers (perhaps just few millions of USD\'s)
- We estimate to need more multilingual aligned parallel corpora
- No reliable estimation is available at my reach
- No accurate idea. But my impressions about any market of Arabic LRs are to be quite small.

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

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

✓ Speech:

- Arabic conversational speech
- Multi-speaker colloquial/formal Arabic speech DB for speaker independent small vocabulary ASR (office environment Speech + revised phonetic transcription); 25,000 sentences over > 350 or speakers.
- Male and female speakers concatenative Arabic TTS data bases; (3,000 sentences over 4 hours clear speech + Electric Glottogram (EGG) signal + revised phonetic transcription + revised phonetic segmentation).
- corpus for acoustic models corpus for language models
- Speech Resources for Arabic dialects and the Amazigh language
- Speech recordings in cars

✓ Lexica, wordnets, …

- wordnets, Full scale Arabic WordNet –
- Arabic Verb Classes à la Beth Levin
- Validated comprehensive Arabic lexicon.
- Validated lexical semantics of Arabic multi-domain large text corpus (> 500K words) along with a standard formalism. (Arabic Lexical Semantics set and hierarchy).
- Arabic (general language) Terminological resources
- morphosyntaxic-lexicons for Arabic words
- Lexicon, ontologies,
- Thesauri.
- Arabic proper names dictionaries.

✓ Corpora: monolingual, bi/multilingual, various annotations

- text corpora special jargon
- Segmented Arabic Hand written corpus
- Parallel Corpora
- Idiomatic Databases and Corpora
- Validated morphologically analyzed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Morphological model).
- Validated POS tagged Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic POS tags set and tags vector model).
- Validated phonetically transcribed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic Phonetic Grammar).
- Parallel corpora for the language pairs (Arabic-other languages)
- Parallel bi- and multilingual corpora

- Arabic Resources needed to learn the basic information of the language
- Arabic- English (in an ideal world, millions of sentences for every language and every dialect, annotated on all levels)
- Written LRs consisting of gigantic corpora labeled as per proper names, terminologies, ...,
- Validated labeled printed Arabic font-written text images corpus.
- Evaluation corpora

✓ Multimedia/multimodal:

- Multimodal LRs consisting of AudioVisuals + content-rich textual material (emails, blogs, Wikis, Forums).
- multimedia

✓ Tools:

- Basic tools for the (text and speech) processing of the least processed languages (Amazigh language and other languages spoken in Africa)
- A baseline of the NLP infrastructure LRs with Phonetically, morphologically, syntactically, semantically, proper nouns/named-entities

✓ Other

- Arabic printed omni font database
- Arabic grammar

And the corresponding applications (not prioritized yet though MT, CLIR/MLIR and ASR are mentioned many times):

- Arabic language learning for non native speakers
- Information Retrieval + Text Mining.
- Applications using the semantic level
- ASR (speech recognition)
- Bilingual News Tracking
- CALL
- Computer software
- Discourse analysis with dictionary use
- Document Management Systems (DMS). With a special focus on Arabic within either monolingual or multilingual applications.
- E-learning
- Handwriting Recognition
- IR and text search engines, web and search engine applications
- Knowledge Mining
- Language labs
- Linguistics developing parsers
- Morphology,
- MT,
- Omni font written OCR.
- Part-of-speech tagging.
- Question/Answering,
- Screen Readers for the blind or visually impaired people.
- SD-retrieval,

- Spell checkers
- Speech Verification.
- Text mining.
- Tools for (CCS) Collaborative Content Services
- TTS.
- Tutorial/e-learning of written/spoken languages.
- Tutors for Teaching Handwriting
- Voice Car navigation systems

A global platform was also mentioned as: For all NLP applications esp. Text Mining & IR, Text-to-Speech, OCR, MT & MAT, and Language Learning.

6. Analysis of the online survey

From the survey we can at least extract information regarding the key players, the "hot" topics and applications, and the needed resources. Such findings can reinforce our assumptions as stated in the technical annex of the project:

6.1. A short list of key applications as reported within the survey:

The key technologies seem to be **MT, ASR and CLIR/MLIR**. Others were also frequently listed (including some applications based on a combination of technologies) among which:

Information Retrieval + Text Mining.
ASR (speech recognition)
Bilingual News Tracking
IR and text search engines, web and search engine applications
Part-of-speech tagging and Parsers
Spoken Document Retrieval

But as the MEDAR project focuses on multilingual tools, we will concentrate on MT and CLIR/MLIR.

6.2. A short list of key resources as indicated by the respondents:

The short list per type of resources:

- ✓ Speech:

  - Arabic conversational speech
  - Speech Resources for Arabic dialects and the Amazigh language
  - Speech recordings in cars

- ✓ Lexica, wordnets, …

  - wordnets, Full scale Arabic WordNet
  - Validated comprehensive Arabic lexicon.
  - Validated lexical semantics of Arabic multi-domain large text corpus (> 500K words) along with a standard formalism. (Arabic Lexical Semantics set and hierarchy).

✓ Corpora: monolingual, bi/multilingual, various annotations

- Text corpora special jargon
- Parallel Corpora
- Validated morphologically analyzed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Morphological model).
- Parallel corpora for the language pairs (Arabic-other languages)
- Arabic- English (in an ideal world, millions of sentences for every language and every dialect, annotated on all levels)
- Written LRs consisting of gigantic corpora labeled as per proper names, terminologies, ...,

7. Additional findings by ELDA and the MEDAR partners

Following the analysis of the questionnaire, ELDA is working on a specific interview form to collect more detailed information from Partners and those who replied to the survey regarding MT and CLIR/MLIR tools and systems that could be made available.

While doing this ELDA has collected information from various sources such as LREC proceedings, European funded projects, Evaluation campaigns workshops (e.g. NIST/MT, Evalda-CESTA, etc.).

7.1. MT findings

The first result of this information collection is given in the following matrix that lists all identified MT systems that includes Arabic either as a source or a target language:

Source of the "product"	Product name	Arabic->English	Arabic->French	Arabic->Spanish	English->Arabic	French->Arabic	Spanish->Arabic
Commercial	Ajeeb	x			x		
Commercial	Al Misbar				x		
Commercial	Al Mutarjim Al Arabey	x			x		
Commercial	Al-Wafi	x			x		
Commercial	Ambassador	x			x		
Commercial	Angusman’s Translator				x		
Commercial	An-Nakel El-Arabi	x	x		x	x	
Commercial	Applied Language Solutions	x			x		
Commercial	BBN Technologies	x					
Commercial	Golden al-Wafi	x			x		
Commercial	Google	x			x		
Commercial	IBM	x					
Commercial	Interpret	x					
Commercial	Johaina	x					
Commercial	Language Weaver SMTS	x	x	x	x	x	x
Commercial	LEC	x			x		
Commercial	LEC Passport Premium	x	x	x	x	x	x
Commercial	LEC Translate DotNet	x	x	x	x	x	x
Commercial	Maximum Edge	x			x		
Commercial	MITRE Corporation	x					
Commercial	MutarjimNet	x			x		
Commercial	Sakhr Enterprise Translation	x			x		
Commercial	Systran	x			x		
Commercial	Tarjim	x			x		
Commercial	Transclick	x			x		

Commercial	Translate-Net	x	x		x	x	
Commercial	Translution	x			x		
Commercial	TranSphere	x			x		
Commercial	WebTrans	x			x		
Commercial	Windows Live Translator	x			x		
Academic	Fitchburg State College	x					
Academic	Johns Hopkins University	x					
Academic	Queen Mary University of London	x					
Academic	RWTH University of Aachen	x					
Academic	Technical University of Catalonia (UPC)	x					
Academic	U.S. Army Reasearch Laboratory	x					
Academic	Université du Maine	x					
Academic	University of Cambridge	x					
Academic	University of Edinburgh	x					
Academic	University of Maryland	x					
Academic	University of Southern California, Information Science Institute	x					

A number of tools are also considered important by the community; these were also identified and are listed herein:

	Name
Morphological Analyzers:	ArabMorpho
	Xerox Arabic Morphological Analyser
	Raramorph
	Buckwalter Arabic Morphological Analyser
	Sebawai
	Morphological Analyser (CRL, New Mexico State University)
Stemmer	Al-Stem
	Light10
	Larkey
POS Tagger	ArabTagger
	MorphTagger
	Stanford Log-linear Part-Of-Speech Tagger
	Brill's POS tagger for Arabic
Parser	Stanford Arabic Parser
Statistical Machine Translation Toolkit	Egypt
Syntactic Analyzer	Syntactic Analyser (Cimos)

Finally, ELDA identified a number of LRs that could be used within the project activities to train the selected tools or to better tune them to Arabic and the given domains:

Type	Name
Dictionaries:	Al-Misbar
	Al-Wafi Quick Dictionary
	ATA-NTS
	Babylon-Pro
	FreeDict
	LingvoSoft
	Pan Images
	Partner
	PocketTran
	TranslationBooth
	WordPoint
	Xpro7
	ArabDictions
	Sakhr Multilingual Dictionary

Parallel Corpora	UN Bidirectional Multilingual (En, Fr, Ar, Ru, Zh)
	UNESCO
	Hebrew-Arabic-English corpus (Agava Institute)
	EGYPT Gizza Toolkit Quran Parallel Corpus (Ar-En)
	CLARA (Corpus Linguae Arabicae) (Ar-Cz)
	Bilingual aligned corpora (Ar-It, ILC)
	Umaah Arabic English Parallel News Text
	Arabic-English Parallel Translation (LDC)
	10k words AFP Arabic Newswire corpus translated into English (LDC)
	Euradic (Ar-Fr)
	E-A Parallel Corpus (University of Kuwait)
Bilingual Corpora	Multiple Translation Arabic (LDC)
	TDT4 Multilanguage Corpus (LDC)
Evaluation corpora	Arcade II Evaluation Package
	CESTA Evaluation Package

7.2. CLIR/MLIR findings

Regarding the CLIR and/or MLIR, ELDA has identified the following tools and resources. A deep analysis of CLEF & TREC campaigns will be conducted to obtain more data.

Tools:	Product
Text Search Engine	Swift
	Google
	Yahoo
	4Arabs
	Ayna
	Arabo
	Yamli
	MSN
	Exalead
	IDRISI (Sakhr)
	URSA
	MG System
	Araby
Question Answering	AQAS

Resources:

Monolingual Corpora	Agence France Presse (LDC, ELRA)
	Al-Hayat Arabic Corpus (ELRA)
	An-Nahar Arabic corpus 5ELRA)
	Leuven Corpus
	Nijmegen Corpus
	DINAR corpus

	General Scientific Arabic Corpus Classical Arabic Corpus SOTETEL Corpus of Contemporary Corpus
Treebank	Penn Arabic Treebank

Among the work to be done after this information collection is to select a short list of tool kits and resources. This will be very crucial as it impact the whole project. After that the partners will have to conduct all the tasks described above (customization, production of adequate LRs for training and testing, etc.).

8. List of players

This is 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. A coming version will elaborate on profiles and sector of activities.

8.1. Institutions

- Al-Ahlya Amman University –Faculty of Information Technology, Jordan
- ACS TechnoCenter, Morocco
- AlKhawarizmy, Egypt
- AMRA Information Technology, West Bank & Gaza Strip
- Arabic Textware, Jordan
- Arabize, Egypt
- Bank of Jordan, Jordan
- Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank & Gaza Strip
- Cairo Microsoft Innovation Center in Egypt (CMIC)
- Catholic University Leuven (KUL), Belgium
- CEA -LIST/DTSI/SRSI/Laboratoire d’ingénierie de l’information multimédia multilingue, France
- Cimos, France
- CJK Dictionary Institute, Japan
- CNRS – Centre Nationale de la Recherche Scientifique - Délégation Rhône-Alpes, Site Vallée du Rhône, France
- Coltec, Egypt
- Columbia University Center for Computational Learning Systems, USA
- CRSTDLA (Scientific & technical Research Center for Arabic Language Development), Algeria
- DEEC-FECU – Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Egypt
- ELDA, France
- MLTC, Morocco
- ENSIAS – University of Mohammed V Soussi - Ecole Nationale Supérieur d´informatique et d´analyse des Systèmes, Morocco
- ESLE – The Egyptian Society of Language Engineering, Egypt
- European Trading & Technology (eurotec), West Bank & Gaza Strip
- Faculty of Computers and Information, Cairo University, Egypt

- FCIS – Faculty of Computer & Information Sciences, Egypt
- France Telecom R&D ORANGE Labs, France
- GIS Int., Israel
- Hariri Canadian Academy of Sciences and Technologies, Canada
- Higher Institute for Applied Sciences and Technology (HIAST), Syria
- IBM Egypt’s branch, Egypt
- ILSP –Institute for Language and Speech Processing, Greece
- IMAGINET, Egypt
- Indiana University, USA
- Insan Center, Saudi Arabia
- 'Isra Software and Computer Co., West Bank & Gaza Strip
- Institute for the Study and Research on Arabisation, Morocco
- Istituto di Linguistica Computazionale – CNR – Italy
- IT College / Birzeit University, West Bank & Gaza Strip
- Jinny Paging company, Lebanon
- King Abdulaziz City for Science and Technology, Saudi Arabia
- Laboratoire de Recherche en Informatique et Telecommunications Faculte des Sciences, Morocco
- LibanCell company, Lebanon
- Lyon2 – Université Lumière Lyon 2 Faculté des Langues, France
- ManarahNet Modern Software Co., West Bank & Gaza Strip
- Millenium Software S.A.L., Lebanon
- Millennium Technology, Israel
- RDI – The Engineering company for computer systems development, Egypt
- Sakhr, Kuwait (& Egypt)
- SOTETEL – Information Technology – Société Tunisienne d´Entreprises de Télécommunications, Tunisia
- Systran, France
- TALP Research Center - Universitat Politecnica de Catalunya, Spain
- The Arab academy for Sciences and Technology, Egypt
- The Egyptian Society for the Arabisation of Science, Egypt
- University of Maryland, College Park, United States
- UOB – University of Balamand – Department of Computer Engineering, Lebanon
- SDU – University of Southern Denmark, Denmark
- Unit for Learning Innovation- Birzeit University, West Bank & Gaza Strip
- University Cadi Ayyad - Faculty of Sciences, Morocco
- Xerox Research Centre Europe, USA

8.2. Individuals

- Ramzi Abbès, France
- Abdelhamid El Jihad, Morocco
- Abdelmajid Benhamadou, Tunisia
- Muhammad Afeefi, Egypt
- AlAli Kanan, West Bank & Gaza Strip
- Fawaz Al-Anzi, Kuwait
- Nashat Al-Aqtash, West Bank & Gaza Strip
- Rami Al-Hajj Mohamad, Lebanon
- Saleh Arar, West Bank & Gaza Strip
- Ken Beesley, USA
- Zied Ben Tahar, Tunisia
- Aderrahim Benabbou, Morocco
- Yassine Benajiba, Spain
- Mohammed Benkhalifa, Morocco
- Viktor Bielicky, Czech Republic
- Christian Boitet, France
- Malek Boualem, France
- Karim Bouzoubaa, Morocco

- Chris Brew, USA
- Tim Buckwalter, USA
- María J. Castro-Bleda, Spain
- Violetta Cavalli-Sforza, USA (/Morocco)
- Achraf Chalabi, Egypt
- Noureddine Chenfour, Morocco
- Gerard Chollet, France
- Khalid Choukri, France
- Christopher Cieri, USA
- Daoud Maher Daoud, Jordan (formerly France)
- Fathi Debili, France
- Mona Diab, USA
- Joseph Dichy, France
- Everhard Ditters, Netherlands
- Said El Hassani, Morocco
- Bouyakhf El Houssine, Morocco
- Khaled Elghamry, Egypt
- Mohamed El-Mahallawy, Egypt
- Salwa Elramly, Egypt
- Ossama Emam, Egypt
- Mohammed Erradi, Morocco
- Salvador España-Boquera, Spain
- Aly Fahmy, Egypt
- Mohamed Waleed Fakhr, Egypt
- Ali Farghaly, USA
- Abdelkader Fassi-Fehri, Morocco (& UK)
- Nagy Fatehy, Egypt
- José A. R. Fonollosa, Spain
- Jean-Luc Gauvain, France
- Wasel Ghanem, West Bank & Gaza Strip
- Antoine Ghaoui, Lebanon
- Gregory Grefenstette, France
- Ahmed Guessoum, UAE
- Nizar Habash, USA
- Lamia Hadrich Belguith, Tunisia
- Jan Hajic, Czech Republic
- Sonia Halimi, Switzerland
- Salwa Hamada, Egypt
- Isan Hamayel, West Bank & Gaza Strip
- Abdelfattah Hamdani, Morocco
- Mohamed Hassoun, France
- Ihab Jabari , West Bank & Gaza Strip
- Haddar Kais, France
- Reem Kanjawi-Faraj, USA
- Iveta Kourilova, Czech Republic
- Jakub Kracmar, Czech Republic
- Abouenour Lahcen, Morocco
- Azzeddine Lazrek, Morocco
- Mohamed Maamouri, USA
- BenteMaegaard, Denmark
- Abdel. Messaoudi, France
- Outahajala Mohamed, Morocco
- Emad Mohamed, USA
- Chafic Mokbel, Lebanon
- Abdelhak Mouradi, Morocco
- Fiyad Odeh, West Bank & Gaza Strip
- Martine Petrod, Denmark
- Ghassan Qadan, West Bank & Gaza Strip

- Tajj-eddine Rachidi, Morocco
- Ahmed Ragheb, Egypt
- Owen Rambow, USA
- Mohsen Rashwan, Egypt
- Horacio Rodríguez, Spain
- Paul Roochnick, USA
- Mike Rosner, Malta
- Paolo Rosso, Spain
- Salim Roukos, USA
- Jean Senellart, France
- Khaled Shaalan, UAE
- Mohammed Shtayyah, West Bank & Gaza Strip
- Otakar Smrz, Czech Republic
- Abdelhadi Soudi, Morocco
- Emna SOUISSI, Tunisia
- Dekai Wu, Hong Kong
- Mustafa Yaseen, Jordan
- Tawfiq Yazidy, Morocco
- Francisco Zamora-Martínez, Spain
- Rached Zantout, Canada

9. 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 the BLARK for MT & CLIR/MLIR. This survey has identified a large set of requested resources and a few available ones. The following important task 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.

10. Appendix A: The NEMLAR REPORT

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