Report on

Survey on Arabic Language Resources and Tools in the Mediterranean Countries

Main authors

Mahtab Nikkhou, ELDA
Khalid Choukri, ELDA

First version 7 June 2004,
Revised 7 March 2005

The NEMLAR project is supported by the INCO-MED programme

CONTENT

1. Executive Summary

This is the pre-final report of the state of art of the situation of Human Language Technologies (HLT) for Arabic in December 2003. This document aims to describe the work done with respect to surveying existing institutions and experts involved in the development of Arabic Language Resources mainly in Europe and the Southern Mediterranean countries. The present report constitutes an overview; the appendices will give more detailed descriptions of the relevant resources.

2. Introduction

Language Resources (LRs) are recognised 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 and/or devoted to the Arabic language is now getting prominent; the lack, on the one hand, of resources, and, on the other hand, of real-world 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. This also applies to other local/regional languages not part of this report.
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 partly overcome such a handicap, the NEMLAR partners would like to ensure that the 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. ELRA and ELSNET have been promoting the concept of a Basic LAnguage Resources Kit (BLARK) which constitutes a must for each and all languages to allow for automatic processing of the language.

This is the reason that one of the goals of the NEMLAR project 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 aimed to collect some basic and general information about NEMLAR partners' involvement in HLT and to extend our contact database through the identification of new players in each sector and geographical area.

The second phase was 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 aims at drafting a comprehensive report that may serve as the basis for the work of NEMLAR about the industrial needs for resources and the commissioning initiatives that will be carried out as well as listing the recommendations for the future.

3. Methodology

3.1. Questionnaire

This study is targeted towards the players of Arabic Language Technologies in academia and industry in the region. It aims to identify existing Arabic language resources and to define current and future needs regarding Arabic language data. A questionnaire was used for that purpose (see Annex 7.1 Questionnaire).

3.2. Survey

The questionnaire was first sent to the NEMLAR partners who filled it in and contacted Arabic HLT players in their respective countries for the same purpose.

The following institution categories have been distinguished:
- Those who filled in the survey as NEMLAR partners
- Those contacted by the NEMLAR partners to fill in the survey
- Those who have been mentioned by the contacted institutions as willing to co-operate within the NEMLAR project

The questionnaire was sent by e-mail to identified institutions and the project received 55 replies - from **36 institutions** (including ELDA/MLTC) and **19 individual experts :**

3.2.1. Institutions

- Al-Ahlya Amman University –Faculty of Information Technology, Jordan
- AMRA Information Technology, Palestine
- Arabic Textware, Jordan
- Bank of Jordan, Jordan
- Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, Palestine
- Catholic University Leuven (KUL), Belgium
- CEA -LIST/DTSI/SRSI/Laboratoire d'ingénierie de l'information multimédia multilingue, France
- Cimos, France
- CNRS – Centre Nationale de la Recherche Scientifique - Délégation Rhône-Alpes, Site Vallée du Rhône, France
- Coltec, Egypt
- DEEC-FECU – Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Egypt
- ELDA/MLTC, France
- 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
- FCIS – Faculty of Computer & Information Sciences, Egypt
- Hariri Canadian Academy of Sciences and Technologies, Canada
- ILSP –Institute for Language and Speech Processing, Greece
- IMAGINET, Egypt
- IBM Egypt's branch, United States
- Isra Software and Computer Co., Palestine
- Istituto di Linguistica Computazionale – CNR – Italy
- Jinny Paging company, Lebanon
- King Abdulaziz City for Science and Technology, Saudi Arabia
- LibanCell company, Lebanon
- Lyon2 – Université Lumière Lyon 2 Faculté des Langues, France
- Millenium Software S.A.L., Lebanon
- RDI – The Engineering company for computer systems development, Egypt
- Sakhr, Kuwait
- SOTETEL – Information Technology – Société Tunisienne d´Entreprises de Télécommunications, Tunisia

- Systran, France
- 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
- Xerox Research Centre Europe, USA

3.2.2. Individuals

- Abdelhadi Soudi, Morocco
- Abdelkader Fassi Fehri, Morocco
- Aderrahim Benabbou, Morocco
- Antoine Ghaoui, Ceo of Millenium Software S.A.L., Lebanon
- Chenfour Noureddine, Morocco
- Fawaz Al-Anzi, Kuwait
- Fiyad Odeh, Palestine
- Ghassan Qadan, Palestine
- Isan Hamayel, Palestine
- Martine Petrod, Denmark
- Mohamed Maamouri, USA
- Mohammed Benkhalifa, Morocco
- Muhammad Afeefi, Egypt
- Nagy Fatehy, Egypt
- Nashat Al-Aqtash, Palestine
- Rached Zantout, Assistant Professor, Hariri Canadian Academy of Sciences and Technologies, Canada
- Saleh Arar, Palestine
- Salwa Hamada, Egypt
- Tajj-eddine Rachidi, Morocco

4. Results

The figures of the responses to the survey are displayed in the annex (7.2 Statistics).
The key figures which came out of the survey are explained below.

4.1. Information on the institutions

4.1.1. Country of origin

The institutions and individuals who filled in the questionnaires are mostly located in North Africa (Morocco, Tunisia), and in the Near and Middle East (Egypt, Lebanon, Jordan, Palestine, Kuwait, Saudi Arabia).
We also noted institutions and individuals from Europe (France, UK, Italy, Greece, Denmark) and North America (Canada, USA).

Institutions
The majority of the institutions (28%) are in Egypt; in 2nd place is France with 14% and in 3rd Lebanon with 11%. In addition, 8% of the institutions have USA, Palestine and Jordan as country of origin. Finally, we identified institutions from Kuwait, Saudi Arabia, Morocco, Tunisia, Greece, Italy, Denmark and Belgium (3%).

Individuals
32% of the identified experts come from Morocco, which is the highest percentage. At the second and third position, we noted experts from Palestine (26%) and Egypt (16%). With a less significant percentage, we identified experts from Lebanon (11%) and Kuwait, Denmark, Canada and USA (5%).

4.1.2. Type of institution

The majority of the institutions are companies (29%). We also listed a large number of universities (20%) but few public organisations (7%).

4.1.3. No of employees

27% of the institutions employ over 100 people. These are multinationals and universities. 24% of the institutions employ between 10-49 employees. The number of institutions employing between 50-99 employees or less than 10 people is quite low (respectively 7% and 5%).
One should bear in mind that the respondents indicated the total number of employees and not those involved in Human Language Technologies or related areas.

4.1.4. Main activities

Most of the institutions (24%) deal with software development activities. A large number of the identified institutions (about 20%) are involved in education and research, which is not surprising as they are universities. Then, we counted 15% institutions involved in technology transfer, 13% in telecommunications, 11% in Content and 9% in HLT products distribution.
The figures related to activities such as culture, interpretation, translation & localisation, e-commerce and banking are not significant (about 5%).

4.1.5. Involvement in HLT

The large majority (62%) of the institutions are involved in HLT with 38% of them involved in written technologies, 35% in language resources and machine translation and 33% in speech technologies. We also noted institutions involved in text and knowledge mining (27%) and language learning (24%).
And finally, a few institutions (9%) mentioned their main activities include social studies, media and the organisation of conferences.

4.1.6. Main products including Arabic

The following tools and LRs for Arabic HLT have been identified. The language resources and tools which will be described in depth and used for the **BLARK** (Basic LAnguage Resources Kit) are marked with an asterisk (*). The BLARK initiative was launched in The Netherlands with the aim of setting up a central and organised infrastructure for Dutch-based HLT. ELRA and ELSNET are extending this initiative to other languages including Arabic.

Arabic NLP technologies and tools

Arabic NLP technologies and tools	Description	Provider
AL DAAL	Arabic Search Engine	Arabic Textware
Arabic Pen	Arabic Handwriting Recognition	Arabic Textware
Arabic Braille Translator	Printing Braille for Arab Blind	Arabic

		Textware
Term finder for financial domain		Amman University
Spell checker (Under development)		Amman University
ArabMorpho©.	Morphological analyser	RDI
ArabDiac©	Automatic diacritiser	RDI
ArabPOSTag©	POS tagger	RDI
ArabDiction©.	Automatic on-line dictionary binding	RDI
Swift©	Derivative search engine	RDI
Type written OCR for Arabic		RDI
Document indexing and retrieval engine		RDI
ArabMorpho©	Arabic morphological analyser	RDI
English-to-Arabic Interlingua-based MT system		IERA
Arabic morphological generator		ENIM
Arabic grammar checker		Sakhr
ArabDox	Document management system with special support to the Arabic language specifics	Sakhr
Al-Idrisi	Search Engine on the Internet with special support to the Arabic language specifics	Sakhr
AlQari' Alali	Arabic offline type written OCR	Sakhr
A computer system for morphological synthesis and analysis		Sakhr
Language identifier		Sakhr
Morphological parsers		Sakhr
Machine learning systems		
Entity extractors		Xerox
Fact extraction systems		Xerox
Cross-lingual information retrieval		Xerox
Categorisation, federated search		Xerox
Conversion of unstructured documents into XML		Xerox
Text mining		Xerox
Cross-lingual search engine		CEA
Cross-lingual filtering		CEA
Cross-lingual clustering		CEA
Automatic Arabic diacritiser		Cimos
Arabic morphological analyser		Cimos
Arabic syntactical analyser		Cimos
Topic analyser		Cimos
Automatic summarisation		Cimos
***Araterm**		IERA
***Aragen**		IERA
Pertinence mining	Automatic extraction	Pertinence
Automatic reader	Transforms scanned images into a grid of millions of dots, optically recognizes the characters found in them and ultimately coverts them into text. Support Arabic, Persian, English, French and 15 other languages	Sakhr
TTS	Text-to-speech	Sakhr
Corrector	Linguistic tool that is used to analyse content	Sakhr

Speech processing technologies

Speech processing technologies	Description	Provider
ArabTalk©	Arabic text-to-speech	RDI
Speech verification for the self learning of Holy Qur'an's Tadjweed		RDI
Very low bit rate speech compression		RDI
Dictation		RDI
Automatic speech and speaker recognition		Sakhr
Voice communication		LibanCell
Pitch detection		? Morocco
Grapheme to phoneme transcription		? Morocco
Multimodal recognition system		Arab Academy for Maritime Sciences and Transportation
Ibsar	Screen Reader for the blind	Sakhr
Speech synthesis		KACST
Speech recognition		KACST, Sakhr

Text processing technologies

Text processing technologies	Description	Provider
An English-to-Arabic MT System tested on 500 interlingual Finance domain		A.Soudi
IMAGiNET Pocket Arabiser Suite o Pocket Qur'an o Pocket Qur'an (MIPS) o Pocket Qur'an Audio Manager o Pocket Arabic On-line Handwritten OCR. (AlArabi AlKatib) o Pocket English-Arabic Dictionary o Pocket Finger Clix (English) o Pocket Finger Clix (Arabic/English)		IMAGiNET
ArabDox	Document management system with special support to the Arabic language specifics	Sakhr
NasherNet	Electronic Publisher on WWW	
Al-Idrisi	Search Engine on the Internet with special support to the Arabic language specifics	Sakhr
Johaina		Sakhr
AlQari' Alali	Arabic offline type written OCR	
E-Portal		Sakhr
Books Publisher		Sakhr
Ajeeb	English<->Arabic MT system	Sakhr
Software Translator	Bi-directional English<->Arabic	Cimos
Software Translator	Bi-directional French <->Arabic	Cimos
Systran MT system	English – Arabic planned for 2004	Systran

Language Resources

Speech Language Resources	Description	Provider
***A SpeechDat like database**	With more than 100 speakers French/Arabic	UOB
Arabic-French speech database		UOB
***Speech database in 4 languages**	about 10K announcements with 10 Words/Announcements	LibanCell
***Labelled Database for TTS**		Millenium
***Network-DC Arabic BNSC (broadcast news speech corpus)**	20 hours of recordings in Modern Standard Arabic with transcriptions	ELDA-LDC
Corpus of di-syllables		Chenfour
Prosodic corpus	200 words	Chenfour
CALLHOME Egyptian Arabic Speech	120 Egyptian Colloquial Arabic telephone conversations	LDC
CALLFRIEND Egyptian Arabic	60 telephone conversations between native speaker of Egyptian dialect of Arabic	LDC
CALLHOME Egyptian Arabic Speech Supplement + CALLHOME Egyptian Arabic Transcripts	20 telephone conversations transcripts for 120 Egyptian Colloquial Arabic telephone conversations	LDC
1997 HUB5 Arabic Evaluation + 1997 HUB5 Arabic Transcripts	20 transcribed conversations 20 transcribed conversations	LDC
EARS Levantine Arabic Fisher Telephone Collection	about 300 hours of speech to be transcribed in 2004	LDC
Arabic acoustic corpus mono-speaker	corpus acoustique monolocuteur de logatomes en arabe contenant toutes les formes di-syllabiques de l'arabe	Benabou
Arabic acoustic corpus multi-speaker	corpus acoustique multilocuteurs d'une centaine de phrases arabes de différentes modalités	Benabou
Arabic Phonetic Database		Kacst
***Saudi Accented Arabic Voice Bank**		Kacst
Isolated Arabic digits speech	50 speakers of different nationalities and both genders, Arabic names speech (10 names; 50 speakers)	AAST
Large Arabic continuous and isolated speech recordings	from more than 2,500 Arabic native speakers distributed over different ages and from both genders, along with their phonetic transcription (with very high accuracy) covering the formal Arabic as well as different dialects of the various geographic region of Egypt	IBM
Long (> 4 hours) speech recordings	along with their glottal signals of one Arabic male speaker and one Arabic female speaker who both have superior performance of formal Arabic pronunciation	IBM
*Multi-speaker colloquial/formal Arabic speech DB for speaker independent small vocabulary ASR	office environment Speech + revised phonetic transcription; 5,500 sentences over 80 speakers (extensible)	RDI
Holy Qur'an multi-speaker speech DB	total 60 hours for Tajweed verification (speech + revised Tajweed phonetic transcription) over 30 male/female/kids speakers of variable degrees of Tajweed skills. (extensible)	RDI
***Single male speaker concatenative Arabic TTS data bases**	1,300 sentences over 1 hour clear speech + revised phonetic transcription + revised phonetic segmentation, (repeatable process)	RDI

***Single female speaker concatenative Arabic TTS data bases**	3,000 sentences over 4 hours clear speech + Electric Glottogram (EGG) signal + revised phonetic transcription + revised phonetic segmentation, (repeatable process)	RDI
Male and female speakers concatenative Arabic TTS data bases		Sakhr
Speech recordings	hundreds of Arabic speakers that cover multi-dialect slang, formal, and English utterances with the corresponding transcription	Sakhr

Lexical Databases	**Description**	**Provider**
Comprehensive Arabic lexicon		Sakhr
Comprehensive Arabic word net		Sakhr
Lexical semantic analyses of Arabic multi-dom text corpus	> 1 Mega words along with a standard formalism, (Arabic Lexical Semantics set and hierarchy)	Sakhr
Common transliterated foreign names, Acronyn Science fields, Cities		Sakhr
***Greek-Arabic dictionary**		ILSP
***Arabic-Danish Dictionary**	30,000 entries in XML format	Petrod
Torjomane	An electronic bilingual Arabic/English dictionary with two versions, one labelled and coupled with the translation engine and one more simplified stand-alone which can be used on line	SOTETEL-IT
***DIINAR.1 monolingual Arabic language da**	129 000 entries (with specifiers) between nouns, verbs, deverbals, and function words, 6.2 million existing generated lexical units can be generated	(Lyon2-ENSSIB-SOTETEL-IT) (available on ELDA's catalogue)
OPTAR	Arabic-French-English Optics terminological db	Lyon2 (soon available on ELDA's catalogue)
***KALIMAT French Arabic lexical db**	47.000 : 8500 nouns, 7.300 verbs ; 1200 adjectives ; information on type, gender, number ; Verbs : aspect ; adjectives : type ; pronouns, conjunctions, numerals, relations between entries	Lyon2 (not available yet)
Terminology databases	3000 terms in the financial domain (Arabic/English)	Amman University
***Dictionnaires de formes simples arabes**		CNRS (available on ELDA's catalogue)
***Dictionnaires de formes fléchies sim agglutinées arabes**		CNRS (available on ELDA's catalogue)
***DixAF French-Arabic bilingual dictionary**	125,000 binary links between ca. 43,800 French entries and ca. 35,000 Arabic entries. This dictionary is available as in Access format. The majority of Arabic words are voweled. A number of grammatical categories are indicated (names, adjectives, verbs, adverbs, pronouns, prepositions, etc.). This dictionary may be used for applications such as bilingual French-Arabic, Arabic-French indexing, translation, information retrieval, etc.	CNRS/ENS (available on ELDA's catalogue)
DicNom-Ism	French-Arabic bilingual dictionary for proper nouns- 3,122 proper nouns	CNRS
Buckwalter Arabic Morphological Analyser	78K stems, 45K lemmas	LDC
Arabic Newswire, POS tags, morphological ana	1. ATB Part 1 (AFP Corpus): 140 K words.	LDC

	Completed on September 20, 2002 and released (Catalogue number: LDC2002E55 – ftp distribution) 2. ATB Part 2 (UMAAH Corpus): 84437 words. Completed July 2002 3. ATB Part 3 (AL-NAHAR Corpus) 350,000 words . POS annotation will be released in the Spring of 2004	
Egyptian Colloquial Arabic Lexicon	electronic pronunciation dictionary of Egyptian Colloquial Arabic	LDC
Database of Arabic roots, verbs, nouns, adjecti with statistical studies on them		Mrayati
Arabic/multilingual dictionaries and thesauri		Coltec
Bilingual Arabic/English lexicon		Imaginet
Arabic lexicon	Brief version: 2,800 roots & 30,000 stems. Expanded version: 4,500 roots, 60,000 stems	RDI
Comprehensive Arabic dictionary entries for Arabic morphological entities		RDI
Lexique d'urbanisme		IERA
Lexique de Terminologie Linguistique		IERA
Computer Science Lexicon	French, English, Arabic	IERA
Botanic Lexicon	French, English, Arabic	IERA
Arabic-English dictionary		IERA
Military Lexicon		IERA
Building and construction lexicon		IERA
Food industry lexicon		IERA
***Al Ghani Arabic Lexicon**		Abdelghani Abou Al Azm
Multilingual dictionary	English, French, Arabic; 75 000 basic entries	Cimos
Bilingual Arabic-English general dictionary	English ->Arabic: 80 000 basic entries Arabic -> English: 170 000 basic entries	Cimos
Bilingual Arabic-French general dictionary	French ->Arabic: 75 000 basic entries Arabic -> French: 110 000 basic entries	Cimos
Bilingual Arabic-English specialised dictionaries	Arabic<->English - Accounting: 12 000 basic words - Agriculture: 2000 basic words - Business: 3500 basic words - Computer: 2500 basic words - Economy: 14 000 basic words - Environment: 4600 basic words - Financial: 2000 basic words - Medical: 24 000 basic words - Military: 1500 basic words - Science and Techniques: 65 000 basic words	Cimos

Text Corpora	**Description**	**Provider**
Al-hayat Arabic data set	18,639,264 distinct tokens in 42,591 articles, organised in 7 domains. Mark-up, numbers, special characters and punctuation have been removed. The size of the total file is 268 MB.	Open University (available on ELDA's catalogue)
An-nahar text corpus	6 years archives, 45 000 articles and 24 million words, articles in Arabic (Lebanon) from 1995 to	ELDA

	2000 (6 years) stored as HTML files on CDRom media. Each year contains 45 000 articles and 24 million words. Each article includes information such as title, newspaper's name, date, country, type, page, etc.	
***SOTETEL Arabic text corpus**	8 million words from different genres and periods including literature, journalistic writing, and academic materials. The texts are not organized in a data base form but are currently used for lexicographic research.	SOTETEL-IT
***Tagged corpora Arabic-Italian**		ILC
***Bilingual aligned corpora Arabic-Italian**		ILC
***Monolingual reference corpora**		ILC
UN Arabic English Parallel Text		LDC
Umaah Arabic English Parallel News Text	3,039 stories	LDC
Arabic-English Parallel Translation	13,027 sentence pairs	LDC
10K word AFP Arabic Newswire corpus translated into English		LDC
Multiple Translation Arabic	141 stories, 10 human, 2 COTS translations	LDC
Arabic Treebank: Part 1	10k-word English Translation	LDC
TDT 3 Arabic Text		LDC
TDT4 Multilanguage Corpus		LDC
TREC Cross-Language Topics		LDC
Arabic text corpora collected as of fall 2002 in total: 480 million words		LDC
Arabic Newswire Part 1 Agence France Press Corpus	165K words	LDC
Arabic Newswire Part 2: Umaah Corpus	140K words	LDC
***Arabic journalistic text corpus**	Des corpus de textes contenant quelques dizaines de textes arabes journalistiques, dialogués et littéraires	Benabou
Annotated corpus of handwritten Arabic text patterns	from hundreds of writers, is lexically and graphically labelled	Imaginet
***POS/Semantic tagged annotated Arabic corpora**		Salwa Asayyid Hamada)
***Morphologically analysed and manually revised (according to RDI's formalism) text corpus**	size : around 300,000 words (and persistently growing) covering News domain, Dictionary explanations, Literature domain, Business domain, and the Holy Qur'an	RDI
***The Arabic POS tagging of the same corpus just mentioned above**		RDI
DIINAR-MBC (INCO-DC 961791-EC)	Arabic, 10 M words	Nijmegen University, SOTETEL-IT, co-ordination of Lyon2
Morphologically analysed Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Morphological model)	Sakhr
POS tagged Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Arabic POS tags set and tags vector model)	Sakhr
Phonetically transcripted Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Arabic Phonetic Grammar)	Sakhr
Large corpus of labeled scanned pages of multi-domain Arabic documents	for training type written OCR's	Sakhr
Annotated domain specific parallel (esp. Arabic-English) text corpora	prepared for narrow domain machine translation tasks	Sakhr

Multimodal Resources	Description	Provider
***Training corpus of Arabic typewritten OCR**	composed of scanned typewritten Arabic documents parallel to the correct text files of their content. The size of this corpus is over 600 A4 documents covering the 20 most famous Arabic fonts	RDI

Others	Description	Provider
Letters and diacritics for speech synthesis		AlAnzi
Transliteration database	Geographic information and local names transliteration	AlAnzi
Penn Arabic Treebank		LDC
Arabic Treebank ATB Part 1 (AFP Corpus)	fully morphological and syntactic annotation of 734 files representing 160,275 words and 4113 trees -- completed in December 2002 - Electronic release of Arabic Treebank ATB Part 2. v 1.0 - UMAAH Corpus	LDC
Grammatical analyses of Arabic multi-domain large text corpus	> 1 Mega words along with the standard formalism. (Complete formal Arabic grammar)	Sakhr
AlArabi AlKatib	Pocket Arabic On-line Handwritten OCR	Imaginet
Multilingual ontology	Arabic, English, French 400 000 words, phrases and verbs	Cimos

Language Resources or tools planned to be produced within the 2-5 coming years as indicated by the respondents (some are plans others are based on on-going projects):

Planned LRs (2-5 coming years)	Description	Provider
Smaller version of REUTERS data corpus	using a new split	Benkhalifa
Speech Units database for synthesis		Chenfour
Resources for speech recognition		Chenfour
Morphology database		Chenfour
Computer system for morphological synthesis and analysis		Mrayati
Database of Arabic roots, verbs, nouns, adjectives,	with statistical studies on them	Mrayati
Multi-dialect Arabic speech corpora		AAST
Off-the-shelf annotated corpus of online hand-written Arabic text patterns	from hundreds of writers, lexically and graphically labelled, and also offline typewritten documents that cover the different documents layout as well as fonts and styles for N. Fatey's own usage in his PhD thesis	Nagy Fatey
Specialised Arabic dictionaries		Cimos

4.2. Information on language resources

4.2.1. Type of LRs

Most of the institutions and experts use and/or develop LRs. Whereas 62% of them deal with **written LRs**, only 13% use or develop **multimedia and multimodal** data.
We noted 47% institutions and experts involved in the use and development of **text corpora** and **lexical databases**, 40% monolingual and 49% multilingual.
The figures for **speech** (25%) and **terminology** (24%) data are however low.

Regarding types of applications, we noted the creation of an automatic lexicon for morphological engine (lexicon application) and the following applications for text corpora: monolingual reference corpora, automatic lexicons, bilingual aligned corpora and tagged corpora.

As for "other" LRs, the institutions mentioned linguistic tools and resources for bilingual Italian/Arabic corpora and the following related applications: morphological engine for the Arabic language, aligning system for Italian and Arabic parallel texts, automatic tagging system for Italian and Arabic texts and access tools (and relevant query systems) for the texts of the bilingual corpora at each text-processing step.

4.2.2. Use of language resources

67% of the interviewed institutions and experts reported using internally produced LRs, 25% also use LRs produced by specific contracted vendors and 33% use those distributed by data centres such as ELRA and LDC.

4.2.3. Tools used to develop LRs

About half of the institutions mentioned they use MS Office tools to develop their LRs. Another half said they use internal tools. More precisely, the following tools were mentioned:

For speech:

- HTK (HiddenMarkovModels Tools Kit).
- CoolEdit package.
- RDI's Arabic diacritizer and phonetic transcriptor; ArabDiac©.
- Internal tools built using C++ and HTK
- Transcriber
- Shure Micro
- DSP signal processing card
- UNICE : visualisation of temporal signals , energy and F0 ; tagging and detection of F0
- WinF0 (produced internally) : visualisation of temporal signals, energy and F0 ; tagging, F0, calculation of topline & baseline, stylisation,…

For written resources:

- RDI's Arabic morphological (lexical) analyser and disambiguator; ArabMorpho©.
- RDI's Arabic diacritiser and phonetic transcriptor; ArabDiac©.
- RDI's Arabic POS tagger; ArabTagger©.
- Textual analysis procedures
- Morphological engines
- Taggers
- Aligner

- STRAND (locally developed) to identify parallel text on the Web, and locally developed document processing tools to extract translation lexicons from bilingual corpora
- Handler
- Morpho-syntactic tagger & lemmatiser
- Named Entity Recogniser
- Syntactic Parser
- Term Recogniser
- Term Normaliser
- MS-Access forms for lexica
- Arabic morphological analyser
- Arabic POS tagger,
- Arabic syntax analyser,
- Arabic text normaliser
- Arabic morphological (lexical) analyser,
- Arabic grammar checker,
- advanced Arabic text indexer,
- text normaliser
- Statistical tools for disambiguation at all levels of Arabic NLP (from the previous L&H in Belgium).
- Multilingual Word Nets for MT (esp. for English & French).
- Multilingual lexicons and thesauri for MT (esp. for English and French).
- Internally produced technologies and tools built using C++ such as Arabic morphological (lexical) analyzer, Arabic diacritiser and phonetic transcriptor, semantic similarity measuring system, …, etc.

Other tools

- HNC pre-processing environment
- LEXIS environment for encoding the LEXIS lexicon
- MARKER tool for multi-layer text annotation
- CMU (SLM)
- XMLSPY Editor and style sheet designer
- IVR by dialling special codes
- Other common public tools such as MS-Office's Word, Excel, Access, and C++ compilers.
- Home-made tools
- SQL (Structured Query Language)
- (partial) morphological analyzer
- Automatic Reader of Sakhr;
- Finite-State tools
- Rule-based tools (Compiler for Reversible Morphology Rules)
- Knowledge-based tools
- Xerox Finite-State programming languages and libraries
- software to train HMM (Hidden Markov Models) taggers
- machine learning techniques
- HTK (of CMU)
- MatLab
- VB
- ESPS Wave
- Morpho-Conceptual classification system
- Language independent formalism for word and phrase processing
- Language independent formalism for syntactic parsing
- Word, Phrase, Sentence, Structure, and Term disambiguation systems
- Intelligent phrase processor
- Word, phrase, and sentence correction systems
- Internal DSP and DB tools
- HTS

Example of tools used by LDC (Linguistic Data Consortium):

Most of the innovations in LDC's new offices involve data collection infrastructure where multiple collection systems partially overlapping in function provide broad coverage and defence in depth (redundancy).

Three NT servers running BBN speech recognition software for English, Chinese and **Arabic** provide automatic audio indexing. A gigabit Ethernet switch connects all broadcast recording hardware to 12TB of disk and a 24 TB SDLT robotic tape library.

1. LDC created all the annotation tools that relate to **Arabic: POS Annotation tool** (Hubert Jin, 2001)
2. Treebank Tree Editor (Hubert Jin, 2002)
3. AMADAT: an **Arabic Multi-Dialectal Transcription tool** designed and developed by M. Maamouri and H. Jin (LDC, 2003). AMADAT provides a multi-layered transcription that spans the diglossic gap and linguistic distance which exist between any and **all Arabic dialects and Modern Standard Arabic** (MSA) by extending links between the two sets of linguistic structures and connecting transcribed dialectal forms to their underlying MSA-based forms whenever applicable.

4.2.4. Validation of language resources

The majority (67%) of the interviewed experts and institutions validate their data internally. 20% of the interviewees reported that they ask external organisations to validate their data, and 13% reported they do not validate them at all. As given validation references, they mention MULTEXT (handling), PAROLE/SIMPLE (tagging and LEXIS lexicon), TEI and NERC (corpus annotation and tagging), MUC-7 (named entity recognition), EAGLES (corpus structural and linguistic annotation), MATE (functional relations and co-reference annotation), Standards of the specifications of digital speech corpora and accuracy thresholds set by IBM's WRC (Watson Research Center), IEEE, ELDA, IBM, LucentTechnologies and finally EuroSpeech.

Other methods of validation being used are mentioned in the following references:

- Hamza, W.M., "A Large Database Concatenative Approach for Arabic Speech Synthesis", PhD. thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 2000.
- Attia, M., "A Large-Scale Computational Processor of The Arabic Morphology, and Applications.", MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.
- "Magma' Allugha Al-Arabiyya" (The Institute of Standard Arabic) also called "Magma' Al-Khalideen" in Egypt, and the counter part one in Syria-Damascus.

Only 35% of the interviewees follow specific validation standards; and only in 18% of the case this leads to the production of a validation report.

4.2.5. Distribution of LRs

Over half of the interviewed institutions and experts wish to make their resources available to others according to a negotiated standardised distribution agreement. Only 15% said they do not want to distribute their resources, and this due to legal (9%), commercial (5%) strategic (4%) and technical (2%) reasons.

Less than half of those who expressed their wish to distribute their LRs are ready to license their data to researchers and developers. And over a third are willing to license their data to end-users as well. Finally 20% of them would like to license them also to students in the fields of computer science, computer linguistics and IT & communication.

4.2.6. Participation in LRs and Language Technology projects

Most of the institutions and experts are involved in projects related to LRs and language technologies. Among the quoted projects, we listed DINAR-MBC, AREF, ALIF, ALMA, Systran Fr., NAPLUS, CEDRE (Lebanese-French project), LOGOS, DIALOGOS, LE-PAROLE, SIMPLE, INTERA, SpeeCon, NetDC and Orientel. Many of the mentioned projects are EU-funded.

Regarding language technology related projects, the interviewees mentioned TIDES, MALACH, CTU (Denmark's national information centre for technology-supported learning), ELU (Danish government agency), SHF (Danish Research Council for the Humanities), AC/DC Project (supported by the Portuguese Ministry for Science and Technology and SDU),Automatic Geoinformation naming standards and Technolangue (French inter-governmental programme).

Others national and international quoted projects:

US-Morocco: “A Collaborative Human Language Technology Research and Education Agenda”, co-ordinated by a researcher from the Great Plains Network and A.Soudi (ENIM), to be sponsored by NSF and the Great Plains.

Carnegie Mellon University(US)-Ecole Nationale de l’Industrie Minerale (Morocco): A Soudi (ENIM) hosts an NSF grantee from Carnegie Mellon University to work on : Example-based MT and Knowledge-based MT.

Italy-Morocco: “Development of Tools and Linguistic Resources for Translation between Arabic and Italian.”, co-ordinated by a director of research from the Institute of computational linguistics (CNR, Pisa) and A. Soudi (ENIM), to be sponsored by CNRST-Morocco and CNR-Italy

Germany-Netherlands-Morocco: Memory-based Learning of Arabic Morphology + Generating an Arabic Full-Form Lexicon for bi-directional morphology lookup.

AISE (Arabic Intelligent Search Engine)
ARTS (Arabic Thesaurus)
IWRID (Intelligent Word Identifier)
MULDIC (Multilingual Dictionary)
ARWOC (Arabic Word Corrector) <<Crafted for Microsoft and used in its MS-Office>>
AGRAC (Arabic Grammar Checker)
ITRANS (Intelligent Translation System)
The Arabisation of Microsoft’s indexing & searching server. <<Asked by Microsoft>>
ALMA
MORCAS
EURADIC
OUTILEX
NORMALANGUE
EVALDA

Language resources and tools produced within these projects:

- A Generator for Arabic Morphology
- An Arabic Full-form lexicon (not yet complete) for bi-directional morphology lookup
- A Prototype English-to-Arabic MT system
- A Memory-based morphological analysis of Arabic
- An experimental Arabic lexicon (for Nagy Fatehy’s MSc thesis whose title is “An Integrated Morphological and Syntactic Arabic Language Processor Based on a Novel Lexicon Search Technique”).
- Arabic thesaurus, Trilingual dictionary (while Nagy Fatehy was employed in Coltec).
- Arabisation dictionaries of the terminology of modern sciences for the benefit of All the Arabic speaking people.
- Standards of all the components of Arabic language; e.g. setting metrics for Arabic digits, characters, …

4.3. Market:

The majority of the institutions (55%) distribute their products and/or services to the domestic market, and to the Arabic world (53%). A smaller number of them (44%) are present on the international market. 44% of interviewees said they have partnerships with other institutions. The institutions mentioned are the following: ENSSIB, IERA, IRSIT, Cairo University and the Egyptian Society of Language Engineering.

Regarding the needed Arabic LRS, the interviewed experts and institutions stated the following LRs or tools should be made available:

More generally:

- Arabic speech corpora
- Arabic text corpora
- Arabic lexicons
- Domain terminology
- Multimedia resources

And more precisely:

Speech related resources and tools

- A diversified array of Arabic (both MSA and colloquial(s) Arabic) speech corpora
- Arabic speech understanding and synthesis.
- Automatic Arabic large-vocabulary (dictation) Speech Recognition systems for the office environment.
- Automatic Arabic small-vocabulary Speech Recognition systems which are robust versus high noise and channel distortion.
- Basic Resources for Spoken Languages (Moroccan , Algerian, Tunisian Arabic and Berber…)
- Concatenative Arabic Text-To-Speech (TTS) systems.
- Labelled databases for TTS
- Labelled speech corpora
- Large Arabic continuous and isolated speech recordings (from hundreds of speakers distributed over different ages and from both genders) along with their phonetic transcription (with very high accuracy) covering the formal Arabic as well as different dialects of Egypt and other Arab regions.
- Large Arabic continuous and isolated speech recordings along with their phonetic transcription covering the formal as well as different dialects of Egypt and other Arab regions, and also covering the most important channel distortion schemes; direct, wired telephony channels, wireless (mobile) telephony channels, wireless telephony via satellites, etc.
- Long conditioned speech recordings of (at least) one Arabic male speaker and one Arabic female speaker who both have superior performance of formal Arabic pronunciation.
- 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)
- Multi-speaker colloquial/formal Arabic speech databases for speaker independent small vocabulary ASR (office environment Speech + revised phonetic transcription); 25,000 sentences over > 350 or speakers.
- Resources for speech recognition
- Speech units database for synthesis
- Speech verification systems for the (full or computer assisted) self learning of Arabic pronunciation and the Tajweed of the holy Qur'an.
- Very low bit rate Arabic speech compression.

Lexica

- A computational lexicon in the style of Beth Levin's work on Verb Classes in English. This is something that needs to be done in view of its use for many Human Language Technology Applications. There may be some work on an Arabic Computational Lexicon at the University of Maryland (US), sponsored by the defence department, but it seems that access to it is very restricted.

- A standard Arabic lexicon that satisfies the derivative nature of Arabic, along with a standard Arabic dictionary Associated with that lexicon.
- All kinds of electronic bilingual dictionaries whose sources languages are the main world languages (English, French, German, Chinese, …) whose target language is Arabic, and covering the various branches of modern sciences (Physics, Chemistry, Math, Biology, Economy, Politics, …).
- All kinds of traditional and modern electronics monolingual Arabic dictionaries
- Arabic-English-French lexicons
- Bilingual Arabic-English dictionaries
- Computational lexicons including semantic information
- Lexicons of Arabic, English, and French.
- Multilingual dictionaries (where at least Arabic, English, and French are covered)
- Multilingual thesauri (where at least Arabic, English, and French are covered)
- Proper name dictionaries
- Synonym dictionaries
- 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).
- Validated monolingual Arabic lexicon, bilingual Arabic/English lexicon, and statistical data on large Arabic text corpora (for extracting statistical models of graphemes which helps in enhancing the results of pattern recognition in online hand-written OCR).
- Word nets of Arabic, English, and French.

Corpora

- A diversified array of Arabic (both MSA and Arabic colloquials) Text corpora
- Annotated corpus of online hand-written Arabic text patterns (from hundreds of writers) that is lexically and graphically labelled, and also offline typewritten documents that cover the different documents layout as well as fonts and styles for Online hand-written OCR, and Offline type written document analysis and OCR.
- Arabic training text corpus with a size of at least 1 Mega words annotated for all the aforementioned linguistic parameters (morphology, POS tags, grammatical structure, ...).
- Bilingual Italian/Arabic corpora
- Large morphologically, syntactically, phonetically, and semantically tagged Arabic text corpora to use them in developing, testing, and evaluating statistical methods and machine learning techniques.
- Multilingual resources to be used for translator (voice, pager, etc.)
- Parallel corpora
- Semantically annotated corpora
- Text resources to build search engines
- training annotated corpora that are necessary for the statistical selection (disambiguation) of the result, along with the disambiguators
- Validated morphologically analysed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Morphological model).
- Validated phonetically transcribed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic Phonetic Grammar).
- Validated POS tagged Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic POS tags set and tags vector model).

NLP related resources and tools, thesauri

- A canonical standard for the Arabic word structure, and a comprehensive Arabic morphological analyser that follows that standard.
- A comprehensive standard for the Arabic orthography.
- A comprehensive standard for the Arabic phonology including a formal Arabic phonetic grammar.
- A standard Arabic POS tags set, the POS tagging format, and a comprehensive Arabic POS tagger.
- A standard Arabic thesaurus
- A standard for the formal Arabic sentence structure, along with a grammatical analyser that follows that standard.
- A standard operational definition of Arabic semantic analysis and more importantly a standard measure for semantic similarity.

- Arabic derivative/semantic full-text search engines.
- Arabic lexical semantics analyser,
- Arabic Morphological analyser,
- Arabic syntax analyser,
- Arabic text categoriser,
- Arabic text summariser.
- Automatic Arabic morphological analyser and disambiguator
- Automatic Arabic POS tagger.
- Automatic Arabic semantic analyser.
- Automatic Arabic syntactic analyser.
- Automatic Arabic text diacritiser.
- Morphological Analysers
- Off-line type-written Arabic/Latin(mainly English) OCR systems.
- On-line hand-written Arabic/Latin(mainly English) OCR systems.
- Parsers
- Transfer Modules

Machine translation related tools and resources

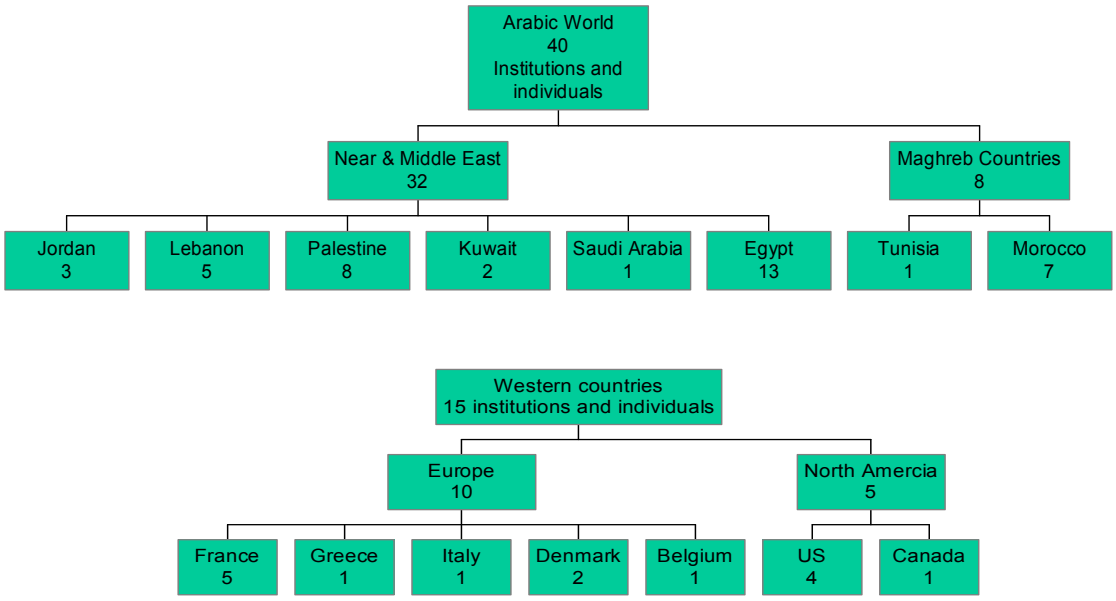
- English-To-Arabic computer assisted MT systems.
- Arabic-To-English assisted MT systems.
- French-To-Arabic Machine Translation systems.
- Arabic-To-French Machine Translation systems.
- Resources that support Machine translation between Arabic and other languages (parallel corpora….)
- Multimedia resources for MT

5. Summary of the outcome of the survey

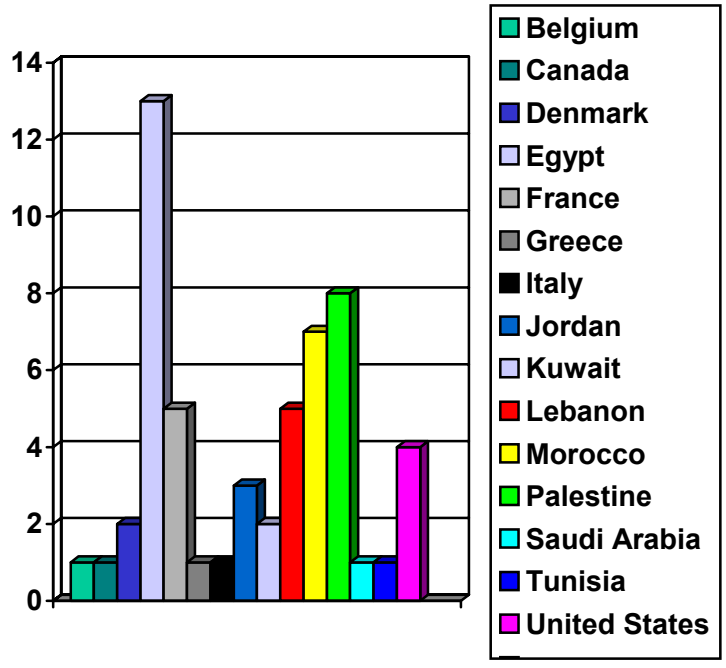
5.1. Geographic distribution of Arabic Language Technologies

We noted Arabic related HLT activities not only in the Arabic speaking world (Lebanon, Morocco, Tunisia, Egypt, Jordan, Palestine, Kuwait, Saudi Arabia), but also in Europe (France, UK, Italy, Greece, Denmark, Belgium) and North America (Canada, USA). This implies a need for Arabic language resources and technologies outside the Arabic countries.

The presence of institutions and experts working on Arabic HLT in France can be linked to the importance of the Arabic speaking population of the country (the most important minority group) who are from North African countries (Algeria, Morocco and Tunisia).



Here above, institutions and experts identified dealing with Arabic HLT in the Arabic world and western countries.



Here above the chart represents the geographic distribution of institutions and experts involved in Arabic HLT.

N.B. As stated in the report one should bear in mind that the survey aimed at collecting information through various networks and primarily through the members of the NEMLAR consortium. We did not intend to be exhaustive but treated all the information received, which explains the unbalanced charts above (we do expect to get more at least from the "big" countries).

5.2. Use and development of Arabic Language Resources

Although some activities around speech technologies have been reported in our survey, only a few Arabic speech resources seem to be used and or developed. However, most of the used Language Resources are developed internally which implies that existing Arabic Language Resources i. are not available on the market or ii. there is a need for very specific resources and the institutions prefer to develop them internally. Meanwhile ELDA and the LDC have been quoted as the providing data centres.

5.3. Validation

The investigation outlines a lack of validation activity on language resources; and even though some institutions and experts do validate their resources (mostly internally), this does hardly lead to the production of a validation report. However, when these institutions and experts expressed their needs on Language Resources, they required validated Arabic data.

5.4. Needed Language Resources

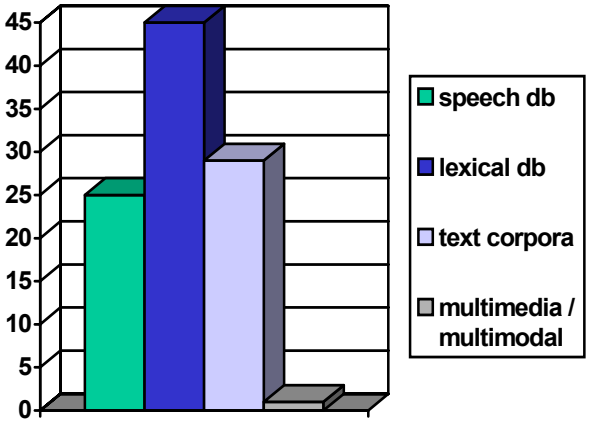
The type of required Arabic Language Resources is linked to the application the institution and or expert is developing. For instance, those which develop or do research on speech, reported they need labelled speech data and phonetic lexicons. As a general fact, the investigation shows there is a need for all type of Language Resources for Arabic: i.e. speech corpora, text corpora (parallel, annotated, parallel, multilingual), monolingual and multilingual lexicons, terminology, multimedia and multimodal data etc.

6. Conclusion

This document aims to reflect the current situation of Human Language Technologies for the Arabic language in the region. The NEMLAR consortium will exploit the given results to draw a working plan according to its objectives of providing the missing resources that would be part of a Basic Language Resource Kit (BLARK) for Arabic and local languages.

The investigation showed that there is a need for Arabic Language Resources in the Arabic world and beyond (western countries) and this is independent of the type and size of the interviewees.
The experts and institutions working on Arabic HLT did not only ask for specific Language Resources but also expressed the need for validated language data.

Regarding the quantity of existing Arabic Language Resources, some institutions reported they do have Language Resources but did not mention their name, number neither describe them. However, **we identified over 100 Language Resources including 25 speech databases, 45 lexicons and dictionaries (monolingual, bilingual and multilingual), 29 text corpora (tagged, annotated, aligned, monolingual, bilingual and multilingual) and 1 multimodal database.**



In the chart here above, is represented the number of Arabic Language Resources identified by college (speech, written, multimedia/multimodal).

The aim of this survey is to get a clear picture of the current situation but also to derive a list of gaps (in term of key basic resources) to be filled to fulfil requirements of R&D and industry. In that context, NEMLAR's partners are investigating on the needs of industrials regarding the languages resources. The results of that investigation will be published in the next step in a report called the Industrial needs in Arabic Language Resources.

7. Annexes

7.1 Questionnaire

The letter and questionnaire used for the survey are presented here after:

Dear NEMLAR partner,

Language Resources (LRs) are recognised 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 now getting prominent; the lack, on the one hand, of resources, and, on the other hand, of real-world 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, NEMLAR partners 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. ELRA and ELSNET have been promoting the concept of a Basic LAnguage Resources Kit (BLARK) which constitutes a must for each and all languages to allow for automatic processing of the language.

This is the reason that one of the goals of the NEMLAR project 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 collect some basic and general information about NEMLAR partners' involvement HLT and to extend our contact database through the identification of new players in each sector and geographical area. 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 the work of NEMLAR about the industrial needs for resources and the commissioning initiatives that will be carried out as well as listing the recommendations for the future.

If you work for an institution, please fill in this questionnaire "Survey on the existing Institutions and Language Resources". If you work as an individual and or consultant, please fill in the "Survey on Individuals and Language Resources".

Best regards

Khalid Choukri

1. INFORMATION ABOUT YOUR INSTITUTION

Name of your institution: _____.

What is your institution's country of origin? _____

Address:
Postcode:
City:
Country:

Phone:
Fax:
Website:

Type
☐ Company
☐ University
☐ Public organisation
☐ Other

Number of employees
☐ Less than 10
☐ 10-49
☐ 50-99
☐ Over 100

Your institution's main activity
☐ Software
☐ HLT Product Vendor
☐ Culture/ Museum
☐ Technology Transfer
☐ Minority language organisation
☐ Content provider

☐ Interpreting/ Translating / Localisation
☐ Telecommunications
☐ E-commerce
☐ Banking/ Insurance
☐ Other, please specify: _____

Is your institution involved in Language Technologies?
☐ Yes
☐ No

If yes, in which one(s) is it involved?
☐ Language learning
☐ Language Resources
☐ Speech technologies
☐ Written technologies
☐ Search and knowledge mining
☐ Machine Translation/Computer-Assisted Translation
☐ Other, please specify: _____

What are your institution's main products and or services? (Please list)

_____
_____
_____
_____
_____

Are your products and or services:
☐ Monolingual
☐ Multilingual

Do they include the Arabic language?
☐ Yes
☐ No

2. CONTACT PERSONAL DATA

First name:
Last name:
Position:
Department:

Address:
Postcode:
City:
Country:

Phone:
Fax:
E-mail:

3. INFORMATION ABOUT YOUR LANGUAGE RESOURCES

Please select as many boxes as appropriate:

Language Resources type

(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.):

☐ Speech Resource
☐ Written Resources
☐ Lexical databases
  ☐ Monolingual
  ☐ Multilingual
☐ Terminology databases
☐ Text Corpora
☐ Multimedia /multimodal Resources
☐ Please specify: _____
☐ Other Language Resources, please specify: _____

Use of your Language Resources

Does your institution use Language Resources,

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

Tools

What kind of tools do you use to produce your Language Resources?

(Please elaborate)_____

Validation of Language Resources

When producing Language Resources, do you follow specific guidelines?

☐ Internal specifications
☐ External specifications - Please name them and give references:
☐ None

Do you follow specific standards?

☐ Yes
☐ No

If yes, please name them and give references:

☐ Validation carried out by independent/external organisation/expert
☐ No

If yes, does this lead to the production of validation reports?

☐ Yes
☐ No

Distribution of Language Resources

Would you be willing to make your resources available to others according to a negotiated standardised distribution agreement?

☐ Yes
☐ No

If no, what are the reasons for not distributing your Language Resources?

☐ Technical
☐ Commercial (pricing policy)
☐ Legal (Copyright, Industrial/intellectual property rights)
☐ Strategic
☐ Other, please specify _____

If yes, whom would you be ready to license your Language Resources to?

☐ End-users
☐ Tool developers
☐ Researchers
☐ Other, please specify _____

Participation in Language Resources Projects

Is/was your institution involved in any Language Resources or Technology project?

☐ Yes
☐ No

4. MARKET

Are your products and or services distributed and or offered to the:

☐ Domestic market
☐ Arabic world
☐ International market

Do you have partnerships with other institutions?

☐ Yes
☐ No

If yes, please cite some of your key partners: _____

5. CONCLUDING QUESTIONS

In general - if you were to decide:
Which Language Resources should be available?

Would you like us to send this survey to other institutions within the NEMLAR focus (which are involved in Human Language Technologies and Language Resources for Arabic and/or related languages)?

☐ Yes
☐ No

Please indicate here the details (please duplicate this section as many times as needed)

Name of the institution:
What is the institution's country of origin?
Contact person
Address:
Postcode:
City:
Country:

Phone:

Fax:
Website:
Email

Please return this questionnaire to choukri@elda.fr

7.2 Statistics

Number of institutions: 36
Number of individuals: 19
Total number of interviewees: 55
The statistics figures are rounded up.

1. INFORMATION ON YOUR INSTITUTION

What is your organisation's country of origin?

(on a total of 36 institutions)	# answers	%
Jordan	3	**8**
Lebanon	4	**11**
Palestine	3	**8**
Kuwait	1	**3**
Saudi Arabia	1	**3**
Egypt	10	**28**
Morocco	1	**3**
Tunisia	1	**3**
France	5	**14**
Greece	1	**3**
Italy	1	**3**
United States	3	**8**
Denmark	1	**3**
Belgium	1	**3**

What is your country of origin (survey on HLT experts) ?

(on a total on 19 individuals)	# answers	%
Denmark	1	**5**
Lebanon	2	**11**
Canada	1	**5**
Egypt	3	**16**
USA	1	**5**
Kuwait	1	**5**
Morocco	6	**32**
Palestine	5	**26**

The rest of the statistics will be on a total of 55 institutions and individuals

Type of institution

	# answers	%
Company	16	29

University	11	20
Public organisation	4	7
Other	1	2

Number of employees

	# answers	%
- Less than 10	3	5
- 10-49	13	24
- 50-99	4	7
Over 100	15	27

Your institution's main activity

	# answers	%
☐ Software	13	24
☐ HLT Product Vendor	5	9
☐ Culture/ Museum		0
☐ Technology Transfer	8	15
☐ Minority language organisation		0
☐ Content provider	6	11
☐ Interpreting/ Translating / Localisation	3	5
☐ Telecommunications	7	13
☐ E-commerce	2	4
☐ Banking/ Insurance	1	2
☐ Other, please specify: *Education, Research, Research in computational linguistics, HLT, LRs identification, production, validation and evaluation, HLT promotion, conferencing, science and technology research, dissemination of proceedings, dissemination of knowledge and research results*	13	24

Is your institution involved in Language Technologies?

	# answers	%
☐ Yes	34	62
☐ No	2	4

If yes, in which one(s) is it involved?

	# answers	%
☐ Language learning	13	24
☐ Language Resources	19	35
☐ Speech technologies	18	33
☐ Written technologies	21	38
☐ Search and knowledge mining	15	27
☐ Machine Translation/Computer-Assisted Translation	19	35
☐ Other, please specify: *teaching, social studies, media, LRs identification, production, validation and evaluation, HLT promotion, NLP, Arabisation of science*	5	9

What are your institution's main products and or services? (Please list)

Arabic NLP technologies and tools

Arabic NLP technologies and tools	Description	Provider
AL DAAL	Arabic Search Engine	Arabic Textware
Arabic Pen	Arabic Handwriting Recognition	Arabic Textware
Arabic Braille Translator	Printing Braille for Arab Blind	Arabic Textware
Term finder for financial domain		Amman University
Spell checker (Under development)		Amman University
ArabMorpho©.	Morphological analyser	RDI
ArabDiac©	Automatic diacritiser	RDI
ArabPOSTag©	POS tagger	RDI
ArabDiction©.	Automatic on-line dictionary binding	RDI
Swift©	Derivative search engine	RDI
Type written OCR for Arabic		RDI
Document indexing and retrieval engine		RDI
ArabMorpho©	Arabic morphological analyser	RDI
English-to-Arabic Interlingua-based MT system		IERA
Arabic morphological generator		ENIM
Arabic grammar checker		Sakhr
ArabDox	Document management system with special support to the Arabic language specifics	Sakhr
Al-Idrisi	Search Engine on the Internet with special support to the Arabic language specifics	Sakhr
AlQari' Alali	Arabic offline type written OCR	Sakhr
A computer system for morphological synthesis and analysis		Sakhr
Language identifier		Sakhr
Morphological parsers		Sakhr
Machine learning systems		
Entity extractors		Xerox
Fact extraction systems		Xerox
Cross-lingual information retrieval		Xerox
Categorisation, federated search		Xerox
Conversion of unstructured documents into XML		Xerox
Text mining		Xerox
Cross-lingual search engine		CEA
Cross-lingual filtering		CEA
Cross-lingual clustering		CEA
Automatic Arabic diacritiser		Cimos
Arabic morphological analyser		Cimos
Arabic syntactical analyser		Cimos
Topic analyser		Cimos
Automatic summarisation		Cimos
***Araterm**		**IERA**
***Aragen**		**IERA**
Pertinence mining	Automatic extraction	Pertinence
Automatic reader	Transforms scanned images into a grid of millions of dots, optically recognizes the	Sakhr

	characters found in them and ultimately coverts them into text. Support Arabic, Persian, English, French and 15 other languages	
TTS	Text-to-speech	Sakhr
Corrector	Linguistic tool that is used to analyse content	Sakhr

Speech processing technologies

Speech processing technologies	Description	Provider
ArabTalk©	Arabic text-to-speech	RDI
Speech verification for the self learning of Holy Qur'an's Tadjweed		RDI
Very low bit rate speech compression		RDI
Dictation		RDI
Automatic speech and speaker recognition		Sakhr
Voice communication		LibanCell
Pitch detection		? Morocco
Grapheme to phoneme transcription		? Morocco
Multimodal recognition system		Arab Academy for Maritime Sciences and Transportation
Ibsar	Screen Reader for the blind	Sakhr
Speech synthesis		KACST
Speech recognition		KACST, Sakhr

Text processing technologies

Text processing technologies	Description	Provider
An English-to-Arabic MT System tested on 500 interlingual Finance domain		A.Soudi
IMAGiNET Pocket Arabiser Suite o Pocket Qur'an o Pocket Qur'an (MIPS) o Pocket Qur'an Audio Manager o Pocket Arabic On-line Handwritten OCR. (AlArabi AlKatib) o Pocket English-Arabic Dictionary o Pocket Finger Clix (English) o Pocket Finger Clix (Arabic/English)		IMAGiNET
ArabDox	Document management system with special support to the Arabic language specifics	Sakhr
NasherNet	Electronic Publisher on WWW	
Al-Idrisi	Search Engine on the Internet with special support to the Arabic language specifics	Sakhr
Johaina		Sakhr

AlQari' Alali	Arabic offline type written OCR	
E-Portal		Sakhr
Books Publisher		Sakhr
Ajeeb	English<->Arabic MT system	Sakhr
Software Translator	Bi-directional English<->Arabic	Cimos
Software Translator	Bi-directional French <->Arabic	Cimos
Systran MT system	English – Arabic	Systran

Language Resources

Speech Language Resources	Description	Provider
*A SpeechDat like database	With more than 100 speakers French/Arabic	UOB
Arabic-French speech database		UOB
*Speech database in 4 languages	about 10K announcements with 10 Words/Announcements	LibanCell
*Labelled Database for TTS		Millenium
***Network-DC Arabic BNSC (broadcast news speech corpus)**	**20 hours of recordings in Modern Standard Arabic with transcriptions**	ELDA-LDC
Corpus of di-syllables		Chenfour
Prosodic corpus	200 words	Chenfour
CALLHOME Egyptian Arabic Speech	120 Egyptian Colloquial Arabic telephone conversations	LDC
CALLFRIEND Egyptian Arabic	60 telephone conversations between native speaker of Egyptian dialect of Arabic	LDC
CALLHOME Egyptian Arabic Speech Supplement + CALLHOME Egyptian Arabic Transcripts	20 telephone conversations transcripts for 120 Egyptian Colloquial Arabic telephone conversations	LDC
1997 HUB5 Arabic Evaluation + 1997 HUB5 Arabic Transcripts	20 transcribed conversations 20 transcribed conversations	LDC
EARS Levantine Arabic Fisher Telephone Collection	about 300 hours of speech to be transcribed in 2004	LDC
Arabic acoustic corpus mono-speaker	corpus acoustique monolocuteur de logatomes en arabe contenant toutes les formes di-syllabiques de l'arabe	Benabou
Arabic acoustic corpus multi-speaker	corpus acoustique multilocuteurs d'une centaine de phrases arabes de différentes modalités	Benabou
Arabic Phonetic Database		Kacst
***Saudi Accented Arabic Voice Bank**		Kacst
Isolated Arabic digits speech	50 speakers of different nationalities and both genders, Arabic names speech (10 names; 50 speakers)	AAST
Large Arabic continuous and isolated speech recordings	from more than 2,500 Arabic native speakers distributed over different ages and from both genders, along with their phonetic transcription (with very high accuracy) covering the formal Arabic as well as different dialects of the various geographic region of Egypt	IBM
Long (> 4 hours) speech recordings	along with their glottal signals of one Arabic male speaker and one Arabic female speaker who both have superior performance of formal Arabic pronunciation	IBM
*Multi-speaker colloquial/formal Arabic speech DB for speaker independent small vocabulary ASR	office environment Speech + revised phonetic transcription; 5,500 sentences over 80 speakers	RDI

	(extensible)	
Holy Qur'an multi-speaker speech DB	total 60 hours for Tajweed verification (speech + revised Tajweed phonetic transcription) over 30 male/female/kids speakers of variable degrees of Tajweed skills. (extensible)	RDI
*Single male speaker concatenative Arabic TTS data bases	1,300 sentences over 1 hour clear speech + revised phonetic transcription + revised phonetic segmentation, (repeatable process)	RDI
***Single female speaker concatenative Arabic TTS data bases**	**3,000 sentences over 4 hours clear speech + Electric Glottogram (EGG) signal + revised phonetic transcription + revised phonetic segmentation, (repeatable process)**	RDI
Male and female speakers concatenative Arabic TTS data bases		Sakhr
Speech recordings	hundreds of Arabic speakers that cover multi-dialect slang, formal, and English utterances with the corresponding transcription	Sakhr

Lexical Databases	**Description**	**Provider**
Comprehensive Arabic lexicon		Sakhr
Comprehensive Arabic word net		Sakhr
Lexical semantic analyses of Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism, (Arabic Lexical Semantics set and hierarchy)	Sakhr
Common transliterated foreign names, Acronyms, Science fields, Cities		Sakhr
*** Greek-Arabic dictionary**		ILSP
***Arabic-Danish Dictionary**	**30,000 entries in XML format**	Petrod
Torjomane	An electronic bilingual Arabic/English dictionary with two versions, one labelled and coupled with the translation engine and one more simplified stand-alone which can be used on line	SOTETEL-IT
***DIINAR.1 monolingual Arabic language database**	**129 000 entries (with specifiers) between nouns, verbs, deverbals, and function words,** 6.2 million existing generated lexical units can be generated	(Lyon2-ENSSIB-SOTETEL-IT) (available on ELDA's catalogue)
OPTAR	Arabic-French-English Optics terminological db	Lyon2 (soon available on ELDA's catalogue)
***KALIMAT French Arabic lexical db**	47.000 : 8500 nouns, 7.300 verbs ; 1200 adjectives ; information on type, gender, number ; Verbs : aspect ; adjectives : type ; pronouns, conjunctions, numerals, relations between entries	Lyon2 (not available yet)
Terminology databases	3000 terms in the financial domain (Arabic/English)	Amman University
***Dictionnaires de formes simples arabes**		CNRS (available on

		ELDA's catalogue)
***Dictionnaires de formes fléchies simples et agglutinées arabes**		CNRS (available on ELDA's catalogue)
***DixAF French-Arabic bilingual dictionary**	125,000 binary links between ca. 43,800 French entries and ca. 35,000 Arabic entries. This dictionary is available as in Access format. The majority of Arabic words are voweled. A number of grammatical categories are indicated (names, adjectives, verbs, adverbs, pronouns, prepositions, etc.). This dictionary may be used for applications such as bilingual French-Arabic, Arabic-French indexing, translation, information retrieval, etc.	CNRS/ENS (available on ELDA's catalogue)
DicNom-Ism	French-Arabic bilingual dictionary for proper nouns	CNRS
Buckwalter Arabic Morphological Analyser	78K stems, 45K lemmas	LDC
Arabic Newswire, POS tags, morphological analysis	1. ATB Part 1 (AFP Corpus): 140 K words. Completed on September 20, 2002 and released (Catalogue number: LDC2002E55 – ftp distribution) 2. ATB Part 2 (UMAAH Corpus): 84437 words. Completed July 2002 3. ATB Part 3 (AL-NAHAR Corpus) 350,000 words . POS annotation will be released in the Spring of 2004	LDC
Egyptian Colloquial Arabic Lexicon	electronic pronunciation dictionary of Egyptian Colloquial Arabic	LDC
Database of Arabic roots, verbs, nouns, adjectives, etc. with statistical studies on them		Mrayati
Arabic/multilingual dictionaries and thesauri		Coltec
Bilingual Arabic/English lexicon		Imaginet
Arabic lexicon	Brief version: 2,800 roots & 30,000 stems. Expanded version: 4,500 roots, 60,000 stems	RDI
Comprehensive Arabic dictionary entries for all the Arabic morphological entities		RDI
Lexique d'urbanisme		IERA
Lexique de Terminologie Linguistique		IERA
Computer Science Lexicon	French, English, Arabic	IERA
Botanic Lexicon	French, English, Arabic	IERA
Arabic-English dictionary		IERA
Military Lexicon		IERA
Building and construction lexicon		IERA
Food industry lexicon		IERA
***Al Ghani Arabic Lexicon**		Abdelghani Abou Al Azm
Multilingual dictionary	English, French, Arabic; 75 000 basic entries	Cimos
Bilingual Arabic-English general dictionary	English ->Arabic: 80 000 basic entries Arabic -> English: 170 000 basic entries	Cimos
Bilingual Arabic-French general dictionary	French ->Arabic: 75 000 basic entries Arabic -> French: 110 000 basic entries	Cimos
Bilingual Arabic-English specialised dictionaries	Arabic<->English - Accounting: 12 000 basic words - Agriculture: 2000 basic words - Business: 3500 basic words - Computer: 2500 basic words	Cimos

	- Economy: 14 000 basic words - Environment: 4600 basic words - Financial: 2000 basic words - Medical: 24 000 basic words - Military: 1500 basic words - Science and Techniques: 65 000 basic words	

Text Corpora	Description	Provider
Al-hayat Arabic data set	18,639,264 distinct tokens in 42,591 articles, organised in 7 domains. Mark-up, numbers, special characters and punctuation have been removed. The size of the total file is 268 MB.	Open University (available on ELDA's catalogue)
An-nahar text corpus	6 years archives, 45 000 articles and 24 million words, articles in Arabic (Lebanon) from 1995 to 2000 (6 years) stored as HTML files on CDRom media. Each year contains 45 000 articles and 24 million words. Each article includes information such as title, newspaper's name, date, country, type, page, etc.	ELDA
***SOTETEL Arabic text corpus**	8 million words from different genres and periods including literature, journalistic writing, and academic materials. The texts are not organized in a data base form but are currently used for lexicographic research.	SOTETEL-IT
***Tagged corpora Arabic-Italian**		ILC
***Bilingual aligned corpora Arabic-Italian**		ILC
***Monolingual reference corpora**		ILC
UN Arabic English Parallel Text		LDC
Umaah Arabic English Parallel News Text	3,039 stories	LDC
Arabic-English Parallel Translation	13,027 sentence pairs	LDC
10K word AFP Arabic Newswire corpus translated into English		LDC
Multiple Translation Arabic	141 stories, 10 human, 2 COTS translations	LDC
Arabic Treebank: Part 1	10k-word English Translation	LDC
TDT 3 Arabic Text		LDC
TDT4 Multilanguage Corpus		LDC
TREC Cross-Language Topics		LDC
Arabic text corpora collected as of fall 2002 in total: 480 million words		LDC
Arabic Newswire Part 1 Agence France Press Corpus	165K words	LDC
Arabic Newswire Part 2: Umaah Corpus	140K words	LDC
***Arabic journalistic text corpus**	**Des corpus de textes contenant quelques dizaines de textes arabes journalistiques, dialogués et littéraires**	Benabou
Annotated corpus of handwritten Arabic text patterns	from hundreds of writers, is lexically and graphically labelled	Imaginet
***POS/Semantic tagged annotated Arabic corpora**		Salwa Asayyid Hamada)

***Morphologically analysed and manually revised (according to RDI's formalism) text corpus**	**size : around 300,000 words (and persistently growing) covering News domain, Dictionary explanations, Literature domain, Business domain, and the Holy Qur'an**	RDI
***The Arabic POS tagging of the same corpus just mentioned above**		RDI
DIINAR-MBC (INCO-DC 961791-EC)	Arabic, 10 M words	Nijmegen University, SOTETEL-IT, co-ordination of Lyon2
Morphologically analysed Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Morphological model)	Sakhr
POS tagged Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Arabic POS tags set and tags vector model)	Sakhr
Phonetically transcripted Arabic multi-domain large text corpus	> 1 Mega words along with a standard formalism (Arabic Phonetic Grammar)	Sakhr
Large corpus of labeled scanned pages of multi-domain Arabic documents	for training type written OCR's	Sakhr
Annotated domain specific parallel (esp. Arabic-English) text corpora	prepared for narrow domain machine translation tasks	Sakhr

Multimodal Resources	Description	Provider
***Training corpus of Arabic typewritten OCR**	**composed of scanned typewritten Arabic documents parallel to the correct text files of their content. The size of this corpus is over 600 A4 documents covering the 20 most famous Arabic fonts**	RDI

Others	Description	Provider
Letters and diacritics for speech synthesis		AlAnzi
Transliteration database	Geographic information and local names transliteration	AlAnzi
Penn Arabic Treebank		LDC
Arabic Treebank ATB Part 1 (AFP Corpus)	fully morphological and syntactic annotation of 734 files representing 160,275 words and 4113 trees -- completed in December 2002 - Electronic release of Arabic Treebank ATB Part 2. v 1.0 - UMAAH Corpus	LDC
Grammatical analyses of Arabic multi-domain large text corpus	> 1 Mega words along with the standard formalism. (Complete formal Arabic grammar)	Sakhr
AlArabi AlKatib	Pocket Arabic On-line Handwritten OCR	Imaginet
Multilingual ontology	Arabic, English, French 400 000 words, phrases and verbs	Cimos

Language Resources or tools planned to be produced within the 2-5 coming years as indicated by the respondents (some are plans others are based on on-going projects):

Planned LRs (2-5 coming years)	Description	Provider
Smaller version of REUTERS data corpus	using a new split	Benkhalifa
Speech Units database for synthesis		Chenfour
Resources for speech recognition		Chenfour
Morphology database		Chenfour
Computer system for morphological synthesis and analysis		Mrayati
Database of Arabic roots, verbs, nouns, adjectives,	with statistical studies on them	Mrayati
Multi-dialect Arabic speech corpora		AAST
Off-the-shelf annotated corpus of online hand-written Arabic text patterns	from hundreds of writers, lexically and graphically labelled, and also offline typewritten documents that cover the different documents layout as well as fonts and styles for N. Fatey's own usage in his PhD thesis	Nagy Fatey
Specialised Arabic dictionaries		Cimos

Are your products and or services:

	# answers	%
Monolingual	22	40
Multilingual	30	55

Do they include the Arabic language?

	# answers	%
Yes	30	55
No		

3. INFORMATION ABOUT YOUR LANGUAGE RESOURCES

Language Resources type

	# answers	%
Speech Resources	14	25
Written Resources	34	62
Lexical databases	26	47
Monolingual	22	40
Multilingual	27	49
Terminology databases	13	24
Text Corpora	26	47
Multimedia /multimodal Resources	7	13
Other Language Resources	7	13

Use of SLR

Does your organisation use LR:

	# answers	%
that are produced internally?	37	67
that are produced by specific contracted vendors?	14	25
that are distributed by data centres?	18	33

Tools

What kind of tools do you use to produce your Language Resources?

For speech:

- HTK (HiddenMarkovModels Tools Kit).
- CoolEdit package.
- RDI's Arabic diacritizer and phonetic transcriptor; ArabDiac©.
- Internal tools built using C++ and HTK
- Transcriber
- Shure Micro
- DSP signal processing card

Softwares :

- UNICE : visualisation of temporal signals , energy and F0 ; tagging and detection of F0
- WinF0 (produced internally) : visualisation of temporal signals, energy and F0 ; tagging, F0, calculation of topline & baseline, stylisation,…

For written resources:

- RDI's Arabic morphological (lexical) analyser and disambiguator; ArabMorpho©.
- RDI's Arabic diacritizer and phonetic transcriptor; ArabDiac©.
- RDI's Arabic POS tagger; ArabTagger©.
- Textual analysis procedures
- Morphological engines
- Taggers
- Aligner
- STRAND (locally developed) to identify parallel text on the Web, and locally developed document processing tools to extract translation lexicons from bilingual corpora
- Handler
- Morphosyntactic tagger & lemmatizer
- Named Entity Recognizer
- Syntactic Parser
- Term Recognizer
- Term Normalizer
- MS-Access forms for lexica
- Arabic morphological analyzer
- Arabic POS tagger,
- Arabic syntax analyzer,
- Arabic text normalizer
- Arabic morphological (lexical) analyser
- Arabic grammar checker
- Advanced Arabic text indexer
- Text normaliser
- Statistical tools for disambiguation at all levels of Arabic NLP (from the previous L&H in Belgium).
- Multilingual Word Nets for MT (esp. for English & French).
- Multilingual lexicons and thesauri for MT (esp. for English and French).
- Internally produced technologies and tools built using C++ such as Arabic morphological (lexical) analyser, Arabic diacritiser and phonetic transcriptor, semantic similarity measuring system, …, etc.

Other tools

- HNC pre-processing environment
- LEXIS environment for encoding the LEXIS lexicon
- MARKER tool for multi-layer text annotation
- CMU (SLM)

- XMLSPY Editor and style sheet designer
- IVR by dialling special codes
- Other common public tools such as MS-Office's Word, Excel, Access, and C++ compilers.
- SQL (Structured Query Language)
- (partial) morphological analyser
- Automatic Reader of Sakhr;
- Finite-State tools
- Rule-based tools (Compiler for Reversible Morphology Rules)
- Knowledge-based tools
- Xerox Finite-State programming languages and libraries
- software to train HMM (Hidden Markov Models) taggers
- machine learning techniques
- MatLab
- VB
- ESPS Wave
- Morpho-Conceptual classification system
- Language independent formalism for word and phrase processing
- Language independent formalism for syntactic parsing
- Word, Phrase, Sentence, Structure, and Term disambiguation systems
- Intelligent phrase processor
- Word, phrase, and sentence correction systems
- Internal DSP and DB tools
- HTS
- Home-made tools
- Encoded typing in computer

Validation

When producing LRs, do you follow specific guidelines?

	# answers	%
Internal specifications	37	67
External specifications - Please name them and give references: *MULTEXT (handling), PAROLE/SIMPLE (tagging and LEXIS lexicon), TEI and NERC (corpus annotation and tagging), MUC-7 (named entity recognition), EAGLES (corpus structural and linguistic annotation), MATE (functional relations and co-reference annotation)*	11	20
None	7	13

Do you follow specific standards?

	# answers	%
Yes	19	35
No	19	35

If yes, does this lead to the production of validation reports?

	# answers	%
Yes	10	18
No	7	13

DISTRIBUTION

Would you be willing to make your resources available to others according to a negotiated standardised distribution agreement?

	# answers	%
Yes	31	56
No	8	15

If no, what are the reasons for not distributing your resources?

	# answers	%
Technical	1	2
Commercial (pricing policy)	3	5
Legal (Copyright, Industrial/intellectual property rights)	5	9
Strategic	2	4

Other (please specify) _____

If yes, whom would you be ready to license your resources to?

	# answers	%
End-users	22	40
Tool developers	27	49
Researchers	27	49
other	11	20

Participation in Language Technology Projects

	# answers	%
Yes	25	45
No	21	38

TIDES (http://tides.umiacs.umd.edu), MALACH (http://www.clsp.jhu.edu/research/malach)
CTU - Denmark's national information centre for technology-supported learning (now part of Learning Lab Denmark (LLD)
ELU - a Danish government agency which supports project development in connection with further education for university graduates
SHF - The Danish Research Council for the Humanities
The AC/DC Project, supported by the Portuguese Ministry for Science and Technology
The University of Southern Denmark (SDU)
ALMA
MORCAS
EURADIC
OUTILEX
NORMALANGUE
EVALDA

MARKET

Are your products and or services distributed and or offered to the:

	# answers	%
Domestic market	30	55
Arabic world	29	53
International market	24	44

Do you have partnerships with other institutions?

	# answers	%
Yes	24	44
No	15	27

In general - if you were to decide:

Which Language Resources should be available?

The interviewed companies also stated that the following LRs or tools should be made available:

More generally:

- Arabic speech corpora
- Arabic text corpora
- Arabic lexicons
- Domain terminology
- Multimedia resources

And more precisely:

Speech related resources and tools

- A diversified array of Arabic (both MSA and Arabic colloquials) speech corpora
- Arabic speech understanding and synthesis.
- Automatic Arabic large-vocabulary (dictation) Speech Recognition systems for the office environment.
- Automatic Arabic small-vocabulary Speech Recognition systems which are robust versus high noise and channel distortion.
- Basic Resources for Spoken Languages (Moroccan , Algerian, Tunisian Arabic and Berber…)
- Concatenative Arabic Text-To-Speech (TTS) systems.
- Labelled databases for TTS
- Labelled speech corpora
- Large Arabic continuous and isolated speech recordings (from hundreds of speakers distributed over different ages and from both genders) along with their phonetic transcription (with very high accuracy) covering the formal Arabic as well as different dialects of Egypt and other Arab regions.
- Large Arabic continuous and isolated speech recordings along with their phonetic transcription covering the formal as well as different dialects of Egypt and other Arab regions, and also covering the most important channel distortion schemes; direct, wired telephony channels, wireless (mobile) telephony channels, wireless telephony via satellites, etc.
- Long conditioned speech recordings of (at least) one Arabic male speaker and one Arabic female speaker who both have superior performance of formal Arabic pronunciation.
- 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)
- Multi-speaker colloquial/formal Arabic speech databases for speaker independent small vocabulary ASR (office environment Speech + revised phonetic transcription); 25,000 sentences over > 350 or speakers.
- Resources for speech recognition
- Speech units database for synthesis
- Speech verification systems for the (full or computer assisted) self learning of Arabic pronunciation and the Tajweed of the holy Qur'an.
- Very low bit rate Arabic speech compression.

Lexica

- A computational lexicon in the style of Beth Levin's work on Verb Classes in English. This is something that needs to be done in view of its use for many Human Language Technology Applications. There may be some work on an Arabic Computational Lexicon at the University of Maryland (US), sponsored by the defence department, but it seems that access to it is very restricted.

- A standard Arabic lexicon that satisfies the derivative nature of Arabic, along with a standard Arabic dictionary Associated with that lexicon.
- All kinds of electronic bilingual dictionaries whose sources languages are the main world languages (English, French, German, Chinese, …) whose target language is Arabic, and covering the various branches of modern sciences (Physics, Chemistry, Math, Biology, Economy, Politics, …).
- All kinds of traditional and modern electronics monolingual Arabic dictionaries
- Arabic-English-French lexicons
- Bilingual Arabic-English dictionaries
- Computational lexicons including semantic information
- Lexicons of Arabic, English, and French.
- Multilingual dictionaries (where at least Arabic, English, and French are covered)
- Multilingual thesauri (where at least Arabic, English, and French are covered)
- Proper name dictionaries
- Synonym dictionaries
- 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).
- Validated monolingual Arabic lexicon, bilingual Arabic/English lexicon, and statistical data on large Arabic text corpora (for extracting statistical models of graphemes which helps in enhancing the results of pattern recognition in online hand-written OCR).
- Word nets of Arabic, English, and French.

Corpora

- A diversified array of Arabic (both MSA and Arabic colloquials) Text corpora
- Annotated corpus of online hand-written Arabic text patterns (from hundreds of writers) that is lexically and graphically labelled, and also offline typewritten documents that cover the different documents layout as well as fonts and styles for Online hand-written OCR, and Offline type written document analysis and OCR.
- Arabic training text corpus with a size of at least 1 Mega words annotated for all the aforementioned linguistic parameters (morphology, POS tags, grammatical structure, ...).
- Bilingual Italian/Arabic corpora
- Large morphologically, syntactically, phonetically, and semantically tagged Arabic text corpora to use them in developing, testing, and evaluating statistical methods and machine learning techniques.
- Multilingual resources to be used for translator (voice, pager, etc.)
- Parallel corpora
- Semantically annotated corpora
- Text resources to build search engines
- training annotated corpora that are necessary for the statistical selection (disambiguation) of the result, along with the disambiguators
- Validated morphologically analysed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Morphological model).
- Validated phonetically transcribed Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic Phonetic Grammar).
- Validated POS tagged Arabic multi-domain large text corpus (> 500K words) along with a standard formalism (Arabic POS tags set and tags vector model).

NLP related resources and tools, thesauri

- A canonical standard for the Arabic word structure, and a comprehensive Arabic morphological analyser that follows that standard.
- A comprehensive standard for the Arabic orthography.
- A comprehensive standard for the Arabic phonology including a formal Arabic phonetic grammar.
- A standard Arabic POS tags set, the POS tagging format, and a comprehensive Arabic POS tagger.
- A standard Arabic thesaurus
- A standard for the formal Arabic sentence structure, along with a grammatical analyser that follows that standard.
- A standard operational definition of Arabic semantic analysis and more importantly a standard measure for semantic similarity.
- Arabic derivative/semantic full-text search engines.
- Arabic lexical semantics analyser,

- Arabic Morphological analyser,
- Arabic syntax analyser,
- Arabic text categoriser,
- Arabic text summariser.
- Automatic Arabic morphological analyser and disambiguator
- Automatic Arabic POS tagger.
- Automatic Arabic semantic analyser.
- Automatic Arabic syntactic analyser.
- Automatic Arabic text diacritiser.
- Morphological Analysers
- Off-line type-written Arabic/Latin(mainly English) OCR systems.
- On-line hand-written Arabic/Latin(mainly English) OCR systems.
- Parsers
- Transfer Modules

Machine translation related tools and resources

- English-To-Arabic computer assisted MT systems.
- Arabic-To-English assisted MT systems.
- French-To-Arabic Machine Translation systems.
- Arabic-To-French Machine Translation systems.
- Resources that support Machine translation between Arabic and other languages (parallel corpora….)
- Multimedia resources for MT

CONCLUDING QUESTIONS

Would you like us to send this survey to other institutions within the NEMLAR focus (which are involved in Human Language Technologies and Language Resources for Arabic and/or related languages)?

	# answers	%
Yes	37	67
No	6	11