Report on

Survey on Industrial needs for Language Resources

Main authors

Mahtab Nikkhou, ELDA
Khalid Choukri, ELDA

September 30, 2004

1.   Executive Summary

This document aims to describe the work done with respect to surveying industrial needs in Arabic Language resources within some HLT key players in the Arabic world and beyond. The document may be further updated, as we get more replies from the companies contacted by our Arab partners.

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 companies 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 in HLT and to extend our contact database through the identification of new players in each sector and geographical area.

The second phase is to go beyond this first list and the basic information, contacting new companies 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 key players of Arabic Language Technologies in industry.
It aims to identify the industrial needs in Arabic Language Resources. We used a questionnaire for that purpose. (see Annex 6.1 Questionnaire)

3.2. Survey

Our first strategy was to interview the targeted companies via telephone. Once contacted, the companies said they would prefer to receive the questionnaire in order to have enough time to fill it in.

Nemlar's partners were in charge of contacting local companies, while ELDA contacted other organisations which are not located in the partner's countries.

The questionnaire was sent by ELDA to the companies listed below:

1. CEA -LIST/DTSI/SRSI/Laboratoire d'ingénierie de l'information multimédia multilingue, France
2. Cimos, machine translation systems, France
3. ELAN, speech technology systems (TTS), France
4. France Telecom, France
5. Microsoft, USA
6. Scansoft, OCR, speech and language technologies, USA
7. Systran, machine translation systems, France
8. Apptek, machine translation, USA

ELDA and other Nemlar partners got replies from the following companies:

1. CEA -LIST/DTSI/SRSI/Laboratoire d'ingénierie de l'information multimédia multilingue, France
2. Cimos, machine translation systems, France
3. ELAN, speech technology systems (TTS), France
4. Scansoft, OCR, speech and language technologies, USA
5. Systran, machine translation systems, France
6. Birzeit University, Palestinian Territories
7. Apptek, machine translation, USA
8. Libancell, Speech Technologies, Lebanon
9. Millenium, Speech Technologies, Lebanon
10. ImagiNet-Egypt, Dictionaries, OCR, Egypt
11. Sakhr, Speech Technologies, OCR, Egypt
12. IBM, Egypt
13. RDI

We also got a summarised version of the results of 7 companies interviewed by Birzeit
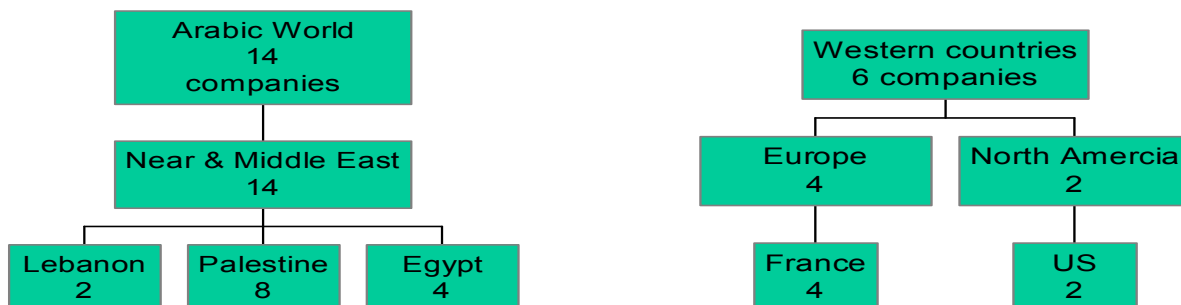
4. Results

We received individual replies from 13 companies and a summarised reply from our contact in the Palestinian Territories. The key results which came out of the survey are explained below. Given the small number of companies interviewed, the survey is intended to reflect an accurate qualitative view of the scene and not necessarily statistical significant data.

4.1. Identification of the institution

4.1.1. Country of origin

We received answers from companies located in Arabic countries (Lebanon, Egypt, Palestinian territories) in Europe (France) and finally in America (US).

```
            Arabic World                          Western countries
            14                                     6 companies
            companies
                                              ┌──────────┴──────────┐
         Near & Middle East                 Europe              North Amercia
         14                                  4                    2
    ┌────────┬────────┐                       │                    │
 Lebanon   Palestine  Egypt                France                 US
 2         8          4                      4                     2
```

The schema here above represents companies expressing their need in Arabic Language Resources in the Arabic world and western countries

4.1.2. Turnover last year - Number of employees

Most of these companies' turnover was below 6 Mio Euros last year. In majority, the interviewed companies employ less than 50 people.

4.1.3. Involvement in Language Technologies?

We noted about half of the companies were involved in speech (TTS, Embedded ASR, Telephony, ASR, Dictation, Audio-mining); another half in written technologies such as OCR, machine translation and terminology.

4.1.4. Use of Arabic language resources

The great majority of the interviewed companies reported they use Arabic LRs for their technologies. In fact, half of them use LRs as part of an HLT-product for instance in machine translation systems and the ¾ of them integrate LRs as sub-functionality in a broader application such as search engines. A minority of the companies use the LRs in connection with various translation services or generation of dictionaries (use of corpora).

4.2. <u>Arabic language resources</u>

Most of the interviewed companies mentioned they need written LRs and half of them expressed the need in speech LRs. The reason that more companies express the need for written LRs is that written LRs and tools are needed also as the basis for speech products.

Also over half of the interviewed companies said they need lexical data, in most cases monolingual ones; half of them mentioned the need in bilingual and slightly less in multilingual ones.

Half of the interviewed companies mentioned text corpora as needed LRs, and this for monolingual, bilingual, multilingual, parallel, equivalent and enriched ones, plus translation memories. We noticed their preference in parallel, equivalent or enriched bilingual corpora rather than in multilingual ones.
The need in monolingual, bilingual and multilingual terminological data was also expressed by 1/3 of the companies.
Very few of them mentioned the need in semantic networks and wordnets.

Regarding the speech resources type, the needs are the following according to the importance of the speech data: read, prepared, prompted, elicited, and spontaneous. The quoted band was in most cases telephone speech and broadcast news. Conversational telephony and in car were also mentioned by half of them. As preferred annotation standard we noted SGML and XML. SAM and NIST/LDC were also mentioned by a few.

And finally, half of the companies reported a need in multimedia resources.

For more information on needed language resources, please see the related section in annex 6.3 Needed Language Resources.
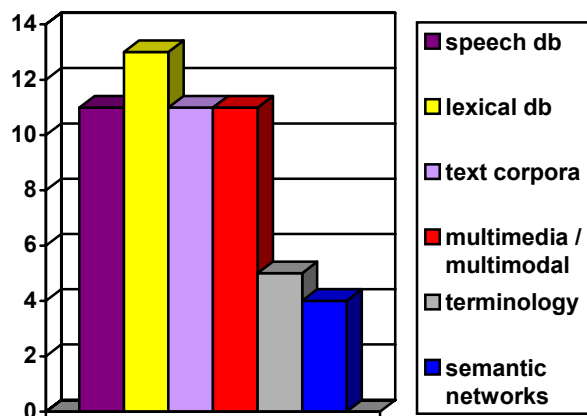
Here above the chart represents the industrial needs – according to our survey - in different types of Arabic Language Resources.

4.3. Acquisition of Arabic language resources

We did not get many answers to this question.
Among those who said they have acquired data, most of the companies reported to acquire data from external vendors. Some others reported that in connection with incorporating the acquired data, they did not benefit from existing standard interchange formats. A few companies said the acquired data fulfilled their requirement.

Among those who said they have not purchased any data, the following reasons were given:
- the LRs were not available .
- the available data were too expensive
- the available data did not live-up to standard quality requirements .
- Lack of portability, in particular lack of adaptability, lack of conformance and lack of reusability.
- Problem with functionality, in particular lack of data quality, lack of coverage, lack of adequate information types
- And others

Regarding the portability issue the opinions are different according to companies. One company reported that "it is very important to comply to standard as it is the only way to use practically the resources without wasting too much effort". However, another company said: "for the moment portability is not really an issue: the type of resources needed for our technology is more demanding than any of the interchange formats - so there is no hope to just plug external resources, whatever is their format".

4.4. Delivery media

Most of the companies mentioned CD-ROM and Internet as their preferred delivery media. DVD was also mentioned by a few. The use of tape is minor.

4.5. Evaluation frequency of your Arabic LRs needs

To this question 1/3 of the companies said they re-evaluate their needs in Arabic language once a semester and another third do it once a year. One institution only re-evaluate their needs every month or once a quarter. One affirmed also never doing it.

4.6. Market

Regarding the expenses in Arabic data in Euro, the estimates were quite different from one institution to another.

The estimates for **Arabic data acquisition** goes from zero to 5 Mio. Some estimated it also for 1000-5000, 40 K or 250 K.

For **Arabic data production** the costs go from 20K to 5 Mio and in between we noted 50K, 100K and 220K. One company said they spent between 4 and 8 full time lexicographers!

Regarding the **expected purchasing budget** for Arabic LRs, the amounts were
- from 15 K to 10 Mio and in between: 30K, 100K, 150K and 205K for the current moment
- and from 0 to 20 Mio for the 3 to 5 coming years. In between we also noted 1000, 20K, 35K, 350K, 600K and 1 Mio.

To our question regarding the Arabic LR market today, the interviewed companies estimated it from 10 Mio to 20 Mio.

And finally, to our question regarding comments on the Arabic LRs, one of the companies mentioned their need in proper nouns dictionaries at reasonable price and tools for the transliteration of proper nouns.

5. Conclusion

The investigation showed that there is a need for Arabic Language resources not only in Arabic countries but also in Europe and USA. It allowed to see that a number of key HLT European and American companies are interested in the development of the Arabic language technologies.

Regarding the type of needed Arabic LRs, the interviewed companies showed interest mostly in written LRs, in particular lexica and text corpora (parallel, equivalent, annotated, monolingual, etc). Needs in speech LRs were expressed and this mostly for read and prepared data.

This report will be updated with other surveys mostly from our partners.

6. Annexes

6.1 Questionnaire

The questionnaire used for the survey is presented here after:

Identification of the institution

Turnover last year
[ ] Below 6 Mio. Euros
[ ] More than 6 Mio. Euros

Number of employees
[ ] Less than 50
[ ] More than 50

Is your organisation involved in the areas of speech Technology?
please specify _____

Is your organisation involved in the areas of written Technology?
please specify _____

Is your organisation involved in another Language Technology area?
please specify _____

Do you use Arabic language resources (LRs) for your technology?
[ ] yes

[ ] no
If yes, are they used :
[ ] As part of an HLT-product? *For instance Machine translation systems*
[ ] Integrated as sub-functionality in a broader application such as search engines
[ ] In connection with various translation services or generation of dictionaries *(use of corpora)*

Arabic Language Resources

What types of language resources do you need?

Written language resources

lexical databases
    monolingual
    bilingual
    multilingual

corpora
    monolingual
    bilingual
    multilingual
    parallel
    equivalent (non parallel, but same domain)
    enriched (annotated with e.g. tags/clusters)
    translation memories

terminological databases
    monolingual
    bilingual
    multilingual

[ ] semantic networks (ontologies)

Speech language resources

Speech type
[ ] Read
[ ] Spontaneous
[ ] Elicited speech
[ ] Prepared
[ ] Prompted
Other (please specify) _____

Band
[ ] Telephone Speech (fixed, mobile, etc.)
[ ] Conversational Telephony
[ ] Wide-band microphone
[ ] In-Car (local, telephony)
[ ] Broadcast News
Other (please specify) _____

Annotation standard:
[ ] SAM
[ ] SGML
[ ] XML

[ ] NIST/LDC
Other (please specify) _____

Multimedia/modal language resources

[ ] multimedia data
[ ] multimodal data

Acquisition of Arabic language resources

If you have already acquired Arabic LRs

Was it from
[ ] external vendors?
[ ] others?

What kind of Arabic LRs do you get/buy? _____

In connection with incorporating the data acquired, have you ever benefited of existing standard interchange formats?

If yes please specify : _____

Did the acquired data fulfil your requirements?
[ ] yes
[ ] no
If no why? _____

If you have never purchased any Arabic LRs

Was it because:
[ ] The LR data were not available/non-existing? *please specify which ones*

[ ] The available LR data were too expensive? *please specify*

[ ] The available data did not live up to standard quality requirements?

In this case, which more specific quality requirements did the LR data not meet?
[ ] Portability
- Lack of adaptability (smooth integration) [ ]
- Lack of conformance (reg. format issues) [ ]
- Lack of reusability [ ]

[ ] Functionality
- Lack of coverage [ ]
- Lack of adequate information types [ ]
- Lack of data quality (about the content) [ ]
[ ] Other

Regarding the portability issue

In general what is your opinion about the various existing standard interchange formats - are they useful and in which way?

If not what are the problems?
- Too many? (in other words, there is no standard!) *please specify*
- Too difficult to use?
- Too (labour) resource demanding? *please specify*

Delivery media

What delivery support media do you prefer?
[ ] Floppy disk
[ ] CD-ROM
[ ] Via Internet (e.g., FTP)
[ ] Tape / Cartridge
[ ] DVD-ROM
Other (please specify) _____

Evaluation frequency of your Arabic LRs needs

How often do you re-evaluate your Arabic LR needs and seek available databases?
[ ] Monthly
[ ] Once per quarter
[ ] Once per semester
[ ] Once per year
[ ] Once every 1-2 years
[ ] Never
Other (please specify) _____

Market

How much do you spend:
for your Arabic data acquisition? _____ Euros/year
for your Arabic data production? _____ Euros/year

How big is your expected purchasing budget for Arabic LRs?
- now: _____
- in 3-5 years time: _____

Which is your estimate of the Arabic LR market today? _____

Concluding questions

Do you have any other comment regarding the Arabic LRs: _____

Thank you for your participation.

6.2 Statistics

Given the small number of replies and the expectation of some more questionnaires from the contacted companies, we thought it was not necessary to make statistics yet.

Identification of the institution

Country of origin:

	# answers
Europe	**4**
France	4
America	**2**
USA	2
Arabic countries	**14**
Lebanon	2
Egypt	4
Palestinian Territories	8

Turnover last year:

	# answers
Below 6 Mio. Euros	13
More than 6 Mio. Euros	7

Number of employees:

	# answers
Less than 50	12
More than 50	8

Is your organisation involved in the areas of speech Technology?

	# answers
yes	9

Is your organisation involved in the areas of written Technology?

	# answers
yes	8

Is your organisation involved in another Language Technology area?

	# answers
yes	7

Do you use Arabic language resources (LRs) for your technology?

	# answers

yes	17
no	

If yes, are they used :

As part of an HLT-product?

	# answers
yes	9

Integrated as sub-functionality in a broader application such as search engines

	# answers
yes	15

In connection with various translation services or generation of dictionaries

	# answers
yes	6

Arabic Language Resources

What types of language resources do you need?

Written language resources	# answers
lexical databases	13
monolingual	11
bilingual	7
multilingual	6
corpora	11
monolingual	9
bilingual	8
multilingual	4
parallel	8
equivalent	8
enriched	9
translation memories	5
terminological databases	5
monolingual	7
bilingual	6
multilingual	4
semantic networks (ontologies)	4

Speech language resources

Type

	# answers
Read	11
Spontaneous	4
Elicited speech	5
Prepared	10
Prompted	6
Other	1

Band

	# answers
Telephone Speech	13
Conversational Telephony	6
Wide-band microphone	4
In-Car	5
Broadcast News	11
Other	

Annotation standard:

	# answers
SAM	5
SGML	8
XML	7
NIST/LDC	4
Other	1

Multimedia/modal language resources	# answers
multimedia data	11
multimodal data	

Acquisition of Arabic language resources

If you have already acquired Arabic LRs

Was it from

	# answers
external vendors	10
others	4

What kind of Arabic LRs do you get/buy? _____

In connection with incorporating the data acquired, have you ever benefited of existing standard interchange formats?

	# answers
yes	3
no	2

If yes please specify : _____

Did the acquired data fulfil your requirements?

	# answers
yes	2
no	2

If no why? _____

If you have never purchased any Arabic LRs

Was it because:
The LR data were not available/non-existing?

	# answers
yes	4
no	

The available LR data were too expensive?

	# answers
yes	2
no	

The available data did not live up to standard quality requirements?

	# answers
yes	2
no	

In this case, which more specific quality requirements did the LR data not meet?
Portability

	# answers
yes	4
no	

	# answers
Lack of adaptability	2
Lack of conformance	2
Lack of reusability	3

Functionality

	# answers
yes	4
no	

	# answers
Lack of coverage	2
Lack of adequate information types	2
Lack of data quality	3
other	1

Regarding the portability issue
In general what is your opinion about the various existing standard interchange formats - are they useful and in which way?

If not what are the problems?

	# answers
Too many	2
Too difficult to use	
Too (labour) resource demanding	

Delivery media

What delivery support media do you prefer?

	# answers
Floppy disk	
CD-ROM	18
Via Internet	17
Tape / Cartridge	1
DVD-ROM	8
Other	

Evaluation frequency of your Arabic LRs needs

How often do you re-evaluate your Arabic LR needs and seek available databases?

	# answers
Monthly	1
Once per quarter	1
Once per semester	8
Once per year	7
Once every 1-2 years	
Never	1
Other	

Market

How much do you spend:
for your Arabic data acquisition? _____ Euros/year
0
1000-5000
40 K
250 K (2answers)
5 M

for your Arabic data production? _____ Euros/year
4-8 lexicographer full time
20K
20K-70K
50K (2 times)
5M
100K
220K

How big is your expected purchasing budget for Arabic LRs?
-    now:
15K
10M
30K
100K
150K
205K

- in 3-5 years time: _____

0
1000
35K
350K
600K
1M
20M

Which is your estimate of the Arabic LR market today? _____
Can not say
10M
15M
18M
20M
25M

6.3 Needed Language Resources

According to our survey on Arabic Language Resources and Tools in the Mediterranean countries (see http://www.nemlar.org/Publications/NEMLAR-REPORT-SURVEY-FINAL_web.pdf) - which can also be used as a complement to this report - the following LRs and tools have been mentioned by the 35 interviewed institutions:

More generally:

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

And more precisely:

Speech related Language Resources and Tools	Languages	Description
Speech corpora	Modern Standard arabic (MSA), Colloquial Arabic	
Basic Resources for Spoken Languages	Moroccan, Algerian, Tunisian Arabic, Berber	
Resources for speech recognition	Arabic	
Speech units database for synthesis	Arabic	
Labelled speech corpora	Arabic	
Labelled databases for TTS	Arabic	
Large Arabic continuous and isolated speech recordings	Arabic, Egyptian dialects, other Arabic dialects	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	Arabic	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	Arabic	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	Arabic	3,000 sentences over 4 hours clear speech + Electric Glottogram (EGG) signal + revised phonetic transcription + revised phonetic segmentation
Multi-speaker speech databases	Colloquial/ formal Arabic	For speaker independent small vocabulary ASR (office environment Speech + revised phonetic transcription); 25,000 sentences over > 350 or speakers
Arabic speech understanding and synthesis	Arabic	
Speech Recognition systems	Arabic	For the office environment
Automatic Arabic small-vocabulary ASR	Arabic	Automatic Arabic small-vocabulary ASR which are robust versus high noise and channel distortion
Speech verification systems	Arabic	For the (full or computer assisted) self learning of Arabic pronunciation and the Tajweed of the holy Qur'an.
Concatenative Arabic TTS systems	Arabic	
Very low bit rate Arabic speech compression	Arabic	

Lexical Databases	*Languages*	*Description*
Computational lexicon	Arabic	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 HLT 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.
Computational lexicons including semantic information		
Standard Arabic lexicon	Arabic	that satisfies the derivative nature of Arabic, along with a standard Arabic dictionary Associated with that lexicon.
Automatic Arabic large-vocabulary	Arabic	dictation
All kinds of traditional and modern electronics monolingual Arabic dictionaries	Arabic	
Validated monolingual Arabic lexicon	Arabic	
Bilingual dictionaries	Arabic-English	

All kinds of electronic bilingual dictionaries	Arabic -English, French, German, Chinese, …	whose sources languages are the main world languages whose target language is Arabic, and covering the various branches of modern sciences (Physics, Chemistry, Math, Biology, Economy, Politics, …).
Multilingual dictionaries	Arabic-English-French	
Proper name dictionaries	Arabic	
Synonym dictionaries	Arabic	
Validated comprehensive Arabic lexicon	Arabic	
Validated lexical semantics of Arabic multi-domain large text corpus	Arabic	(>500K words) along with a standard formalism (Arabic Lexical Semantics set and hierarchy)
Standard Arabic thesaurus	Arabic	
Multilingual thesauri	Arabic, English, French	
Word nets	Arabic, English, French	

Text Corpora	*Languages*	*Description*
Parallel corpora	Arabic	
Bilingual corpora	Italian/Arabic	
A diversified array of Arabic Text corpora	MSA and Arabic colloquials	
Annotated corpus	Arabic	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	Arabic	with a size of at least 1 Mega words annotated for all the aforementioned linguistic parameters (morphology, POS tags, grammatical structure, …)
Large morphologically, syntactically, phonetically, and semantically tagged Arabic text corpora	Arabic	to use them in developing, testing, and evaluating statistical methods and machine learning techniques
Semantically annotated corpora	Arabic	
Training annotated corpora	Arabic	that are necessary for the statistical selection (disambiguation) of the result, along with the disambiguators
Validated morphologically analysed Arabic multi-domain large text corpus	Arabic	(> 500K words) along with a standard formalism (Morphological model)
Validated phonetically transcribed Arabic multi-domain large text corpus	Arabic	(> 500K words) along with a standard formalism (Arabic Phonetic Grammar).
Validated POS tagged Arabic multi-domain large text corpus	Arabic	(> 500K words) along with a standard formalism (Arabic POS tags set and tags vector model).
Resources that support Machine	Arabic	

translation between Arabic and other languages (parallel corpora)		
Multilingual resources to be used for translator	Arabic	voice, pager, etc.
Text resources to build search engines	Arabic	

NLP related Resources and Tools	Languages
A canonical standard for the Arabic word structure, and a comprehensive Arabic morphological analyser that follows that standard.	Arabic
A comprehensive standard for the Arabic orthography	Arabic
A comprehensive standard for the Arabic phonology including a formal Arabic phonetic grammar	Arabic
A standard Arabic POS tags set, the POS tagging format, and a comprehensive Arabic POS tagger	Arabic
A standard for the formal Arabic sentence structure, along with a grammatical analyser that follows that standard	Arabic
A standard operational definition of Arabic semantic analysis and more importantly a standard measure for semantic similarity	Arabic
Arabic derivative/semantic full-text search engines	Arabic
Arabic lexical semantics analyser	Arabic
Arabic morphological analyser	Arabic
Arabic syntax analyser	Arabic
Arabic text categoriser	Arabic
Arabic text summariser	Arabic
Automatic Arabic morphological analyser and disambiguator	Arabic
Automatic Arabic POS tagger	Arabic
Automatic Arabic semantic analyser	Arabic
Automatic Arabic syntactic analyser	Arabic
Automatic Arabic text diacritiser	Arabic
Off-line type-written Arabic/Latin(mainly English) OCR systems	Arabic
On-line hand-written Arabic/Latin(mainly English) OCR systems	Arabic
Parsers	Arabic
Transfer modules	Arabic
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	Arabic

Machine Translation related Resources and Tools	Languages
Computer assisted MT systems	English-To-Arabic Arabic-To-English French-To-Arabic Arabic-To-French
Multimedia resources for MT	