Report on

BLARK for Arabic

Main authors

Steven Krauwer, UU
Bente Maegaard, CST
Khalid Choukri, ELDA
Lise Damsgaard Jørgensen, CST

4 October 2004
Revised 19 September 2006

Table of Contents

European Commission

The NEMLAR project is supported by the INCO-MED programme

Introduction

This document aims at presenting the concept of the Basic Language Resource Kit (BLARK) and at defining a first instantiation for a BLARK for Arabic.

In section 1 we explain the concept and its purpose. In section 2 we address availability, quality, quantity and standards. Section 3 presents the current BLARK Definition and a first overview of the existence of the components.

1. The BLARK Concept

1.1. Description of the concept

We define the Basic Language Resource Kit (abbreviated BLARK) as the minimal set of language resources that is necessary to do any precompetitive research and education at all. The definition is in principle intended to be language independent, but as specific languages do come with different requirements, instantiations of the BLARK may vary in some respects from language to language. A BLARK comprises many different things, such as:

- Basic language resources:
    - written language corpora
    - spoken language corpora
    - bilingual (written) corpora (comparable, parallel, aligned, ...)
    - mono- and bilingual dictionaries
    - terminology collections
    - grammars (i.e. formal standard rule sets such as; a Syntactic Grammar, a Phonetic Grammar, a Lexical Grammar, …)
- Benchmarks for evaluation
- Basic tools:
    - modules (e.g. taggers, morphological analyzer, parsers, speech front-ends, grapheme-to-phoneme converters, statistical disambiguators, …)
    - annotation standards (or best/common practice usage) and tools
    - corpus exploration and exploitation tools
    - etc

The list is far from exhaustive but serves to illustrate the scope of the BLARK. A BLARK should not be seen as a static object: over time it may gradually evolve as new technologies and application areas emerge, with new requirements in terms of resources. The idea was first launched in the ELRA Newsletter in 1998 (Krauwer 1998). It should be noted that in order for the BLARK to serve its purpose it should be accompanied by a (not necessarily very heavy) infrastructure to support its maintenance (keeping it up to date) and the distribution of the resources included in it.

The underlying idea is to make a common generic BLARK definition, applicable in principle to all languages, based on the collective experience and expertise gained with many different languages by the members of the language and speech technology community at large. This common definition will save time and effort (no reinvention of wheels), it will allow for porting of knowledge between languages, it will ensure interoperability and interconnectivity (especially for multilingual or cross-lingual application areas), and it will help making realistic estimates of costs and efforts required to produce them. In addition a broadly supported common definition may be used as an external reference point in discussions with funding agencies about the best way to create a good starting point for language and speech technology, both in academic & industrial (precompetitive) research and academic & professional training.

In order to make a BLARK for a language maximally impactful the language resources of which it consists should be easily and reliably accessible, inexpensive, and usable.

1.2. How to use it

The target audience of the BLARK is researchers (both in academia and in industry), and educators. It is used to train students, to serve as material for research experiments and application pilots (and benchmarking of various algorithms and techniques). Commercial companies should in theory be able to use the BLARK for the development of commercial products, but in general it is unlikely that BLARK components will be usable for commercial applications as they are, because a BLARK will always be limited and will not focus on specific domains needed by industry; also for industry however, a BLARK may constitute a good starting point which will help avoid duplication of work. Because a BLARK is only a starting point, it is of crucial importance that -in principle- the BLARK should come with tools for the production and annotation of new corpora, and that all modules and resources are available in source format, so that industrial developers can freely adapt them to the specific requirements of their applications (e.g. domain, footprint, application environment).

1.3. How to arrive at it

At this moment ELSNET and its sister project ENABLER, that ended last year, are in the process of producing an initial general BLARK definition. ELSNET will continue this activity in close collaboration with the participants in the ENABLER project, COCOSDA and its newly created sister committee ICCWLRE (working title), and with others who want to contribute or who are interested in adopting the BLARK for their own language. ELRA will also want to contribute to this activity.

Even if in the long run we hope that bodies like COCOSDA and ICCWLRE will be able to come up with general guidelines and recommendations for the BLARK definition we are now still in a pioneering phase, where we try on the one hand to contribute to the further elaboration and refinement of the BLARK concept as such, and on the other to arrive at a concrete proposal for the BLARK for the Arabic language.

1.4. First steps towards the creation of a BLARK

After the publication of the first BLARK article in the ELRA Newsletter the idea has been taken up by the Dutch Language Union (DLU), the intergovernmental body created by the Dutch national and Flemish regional government to take care of their common language. A number of publications have followed from these activities, describing both the result (a fairly concrete enumeration of components that should be included in the BLARK for the Dutch language) and the process that led to this result. An excellent summary of the process and the results of the Dutch BLARK exercise can be found in an article by Binnenpoorte et al (2002) in the proceedings of the LREC 2002 workshop "*Towards a Roadmap for Multimodal Language resources and Evaluation*" organized by ELSNET.

Starting point of the definition process in Binnenpoorte et al (2002) were 8 classes of applications, which were claimed to be the most relevant application categories at that moment: computer assisted language learning, access control, speech input, speech output, dialogue systems, document production, information access and translation. For each of them it was established which modules would be needed to make them (e.g. morphological analysis, text to phoneme converter), and for each of these modules it was analyzed which language data (e.g. data sets, descriptions) they would require, as well as their relative importance. The results were put together in a huge matrix, on the basis of which one can determine which components serve most applications, and which data are most needed for most applications, i.e. which elements should be part of the BLARK. We briefly summarize them here to illustrate the outcome of this process:

For language technology the following elements were identified:
Modules:

- robust text pre-processing (tokenization, named entity etc.)
- morphological analysis
- syntactic analysis
- semantic analysis

Data:

- monolingual lexicon
- annotated corpus (tree-bank)
- benchmarks for evaluation

For speech technology:
Modules:

- automatic speech recognition (incl. prosody, non-natives etc.)
- speech synthesis (incl. tools for unit selection)
- Tools for speaker, language and dialect identification
- Speaker identification/verification tools
- tools for (semi-)automatic annotation of speech corpora

Data:

- speech corpora for specific applications
- multi-modal speech corpora
- multi-media corpora
- multi-lingual speech corpora
- benchmarks for evaluation

When the list of modules and data was completed, an inventory was made in order to determine their availability. As availability is not really a binary distinction (materials may exist, but may not be freely usable, or they may not have the desired quality or coverage) a ten point scale was used to describe availability status.

On the basis of a comparison of the definition of what was most needed (the BLARK) and the availability analysis, a priority list was made and used as the starting point for a plan to complete the BLARK for the Dutch language.

1.5.Towards an Arabic BLARK

In the spirit of the underlying philosophy of the BLARK (porting of knowledge and expertise between languages) we have taken the DLU BLARK exercise as our starting point and tried to transpose the results to the Arabic language. This has led to an initial Arabic BLARK definition, which is based on the general concept but adapted to the needs of the Arabic language.

On the basis of the language specific BLARK definition for Arabic it has then been determined which components are already available, and which ones are missing. The amount of missing components may vary dramatically from language to language, as some of the major languages such as English may already be fully covered, whereas others may have to start from scratch. Once the gaps have been identified, priorities will be assigned to the components to be produced, in order to make a realistic plan for the gradual completion of the BLARK.

2. Some remarks on availability, quality, quantity and standards.

Before we can start we have to address a few important issues: availability, quality, quantity and standards.

2.1.The notion of availability

Let us start out repeating that the BLARK and its components are not intended to serve as a direct basis for commercial applications: its goal is to support pre-competitive activities by researchers, developers, integrators, educators, etc. We will use the abbreviation *PreR&D* for all *precompetitive R&D activities* and we will use the standard abbreviation *R&D* to include activities that may be directly aimed at the *creation of commercial products or services*.

The PreR&D orientation of the BLARK means that we cannot expect e.g. a large corpus of annotated patent applications to be a natural part of a BLARK *definition*, although a BLARK *instantiation* might very well contain such a corpus as sample corpus for a limited domain with specific properties. The production of language resources produced with the explicit goal to serve a specific commercial application developed by some company would normally be the responsibility of the company, as part of its investments in the development of the product. The BLARK and its components should in principle be easily accessible for precompetitive purposes. If a company owns specific resources that are not (or can not) be made available to others they can hardly be considered as *available* BLARK components.

In Binnenpoorte et al (2002) we see that the availability of the existing resources was expressed on a 9-point scale. Even if these figures give some impressionistic idea of the urgency of the creation of some of the components the empirical consequences of the various scores are not immediately clear. We will therefore propose a different approach to availability.

We will distinguish 3 classes of availability (numbering based on penalties): (3) existent but only company-internal, (2) existent and freely usable for PreR&D, (1) existent and freely usable for both PreR&D and R&D.

The second (related) observation is that resources that are actually existing, but only at a very high cost (e.g. a morphological analyser for 40 keuro) should not be listed as fully available, as most SMEs or research labs could most probably not justify the expense if it is not part of an operation aimed at recuperating the investment. We will distinguish four cost classes: (4) over 10 keuro, (3) between 1 and 10 keuro, (2) between 100 euro and 1 keuro, (1) less than 100 euro or free.

Third, the inherent exploratory nature of PreR&D will often require a high degree of customizability and adaptability of the resources, both qualitatively and quantitatively. For this reason it is important to distinguish three types of resources: (3) black box resources (you get them as they are, but you cannot change them, e.g. object code), (2) glass box resources (you can inspect the inside but you are not allowed to touch it), and (1) open resources (freely manipulable, e.g. source code).

We will try to associate with each BLARK Content item a three digit code expressing its availability. Resources scoring (1) in all three categories are the ideal components of a BLARK. If a resource item doesn't exist it doesn't get a score at all.

This system can of course be made more fine-grained than this, but we hope that the idea is clear enough to make an initial categorization.

2.2. The notion of quality

Quality is a difficult concept, as it comes in types. It can be absolute (e.g. in the sense of sloppiness in the definition of the annotation rules, or in the way the annotators have done their job on the basis of an otherwise well-defined annotation scheme). It could also be relative (e.g. a high quality lexicon and an equally high quality grammar constitute a useless pair if their annotation schemes do not match). Quality can be a matter of size (too few entries in a lexicon), or of selection (lots of entries, perfectly coded, but not the ones needed for the task at hand).

Binnenpoorte et al (2002) do not provide any account of the way quality was measured (if at all) or expressed, so we have to provide our own quality marker system.

It is clear that we will have to include some sort of quality marker in our descriptive system. At this moment we do not see an obvious framework that we could adopt in order to define quality markers, but we would (very tentatively) suggest to start from the following quality attributes, which all have in common that their values can be verified; we list the attributes, the corresponding criteria, and the possible values below:

Attribute	criterion	values
standard-compliance	to what extent is the resource based on a common standard	no standard
		standard, but not fully compliant
		standard, fully compliant
Soundness (internal consistency)	to what extent is the resource based of well-defined specs	no specs
		specs, but not fully compliant
		specs, fully compliant
Task-relevance	to what extent is the resource suited for a specific task X	contains all info needed (yes/no)
		has the proper size (yes/no)
		based on a relevant selection of items (yes/no)
environment-relevance	to what extent is the resource interoperable with its environment (other resources)	information matches (yes/no)
		size matches (yes/no)
		selection matches (yes/no)

Please note that the attributes are not completely independent (e.g. if a resource is fully standard compliant it is necessarily sound, but not vice-versa), and that a fully standard compliant resource might still be useless because it does not match with the task or with the environment. Note also that the first two attributes take just one value out of three, whereas the last two attributes have a yes/no score on all three sub-attributes.

One can easily add a few new attributes, or adopt a more graded scale for each of the attributes, but for the time being we suggest that we try to see how far we get with this simplified scheme.

One of our own immediate conclusions is that in defining the BLARK and in identifying instantiation of the various definition items we should try to maximize the environment-relevance of each single item so that we have maximal chances to interconnect them if we want to use the BLARK for more complex projects.

If we adopt this scheme as our working hypothesis every BLARK Content item will be associated with a quality marker in accordance with the attribute table above, which can be represented as a series of 1+1+3+3=8 values.

In this first version of the BLARK, we have however not been in a position to apply the quality system.

2.3. Quantity

In Binnenpoorte et al (2002) no attempts have been made to provide quantitative figures for the various resources needed: how many words in a corpus, how many hours of speech, how many lexical entries, etc.

It is clear that a BLARK definition should include very clear guidelines for what counts as a sufficiently large corpus, lexicon, etc. In a paper presented at the ELSNET-ENABLER Workshop in Paris (August 2003), Cieri et al suggest that core resources for a language include a written language corpus of at least 100 000 words, and a 10 000 entries (translation) lexicon. These requirements are probably very modest, but given in the context of this paper (mainly concerned with the technologically less well-covered languages) not unrealistic.

In the BLARK for Arabic we have tried to present reasonable figures, based on estimations of the minimal requirements and on best (or current) practice for Arabic and other languages, cf. section 3.3 BLARK Specification for Arabic.

2.4. Standards

There are relatively few existing official standards for language and speech resources; see e.g. Romary et al (2004) and Monachini et al (2003). At the same time it can be observed that a number of de facto standards seem to be evolving in our communities.

Their origin is sometimes based on bottom-up work by committees (TEI), sometimes on top-down actions (often with public funding, and aimed at the creation of standards, such as EAGLES and ISLE), and sometimes on following examples set by specific projects (e.g. MULTEXT, Speechdat, WordNet).

As the adoption of standards is crucial for the longevity of language and speech resources we will, in the definition of the BLARK for Arabic, try to recommend standards for all types of resources, mostly based on best practice considerations.

3. The BLARK for Arabic

3.1. Approach and some terminology to avoid conceptual confusion

As it is hard to believe that what we have now is the final and ideal BLARK definition for Arabic, we will adopt an evolutionary strategy: at each moment in time we will have a current BLARK definition and specification version, but at the same time we keep evaluating and amending it in order to arrive at the best possible one. We will use the term *BLARK Definition* to refer to these proposals, and the term *BLARK Specification* to refer to more detailed specification (in terms of quality, quantity, standards, etc) of the items included in the BLARK definition.

In parallel with the BLARK Definition (but very much depending on it) we will try to maintain an inventory of which parts of the current BLARK are actually available and which ones still have to be developed. We will call this inventory the *BLARK Content*. Each item in the BLARK Definition will correspond to a (possibly empty) set of BLARK Content items instantiating the definition item.

It is important to keep in mind that there is a significant difference: the BLARK Definition and Specification are *prescriptive*, the BLARK Content is *descriptive* in nature.

The present BLARK definition has taken the Dutch BLARK proposals as point of departure, but we have slightly revised it, e.g. the application areas and the types of resources taken into account, which means that our general concept of a BLARK is slightly different from the original Dutch definition. Additionally, an analysis of the specific needs for Arabic made by the members of the project led to certain language specific differences.

A notable difference between the Dutch and Arabic BLARK definitions is the presence of a diacritizer (vowelizer) in the Arabic BLARK. Another difference is the fact that Arabic has two different types of lexica: a lexicon can be based on roots or on stems (where the root lexicon is seen by most as the most correct one).

3.2.The present BLARK Definition for Arabic

In the tables below we first give the ‘traditional’ correspondence which shows a number of general applications and the language modules that are needed in order to build each application. We then go on to show the relationship between language modules and the resources that are necessary to build those modules.

The degree to which the modules are needed is marked by plus signs: ‘+++’ means ‘essential’, ‘++’ means ‘very important’ and ‘+’ means ‘important’. Compared to the Binnenpoorte et al. approach, we have added the ‘+++’ and kept the meaning of the two other markings.

We have split the tables in one for written and one for spoken resources. The reader may note that ASR/dictation and TTS, which are speech applications, occur in the list of written applications. This is because written modules like morphology and POS speech tagging are needed in order to build a good ASR, and even more modules are needed for TTS.

As mentioned above, we have also split the tables in one that shows the correspondence between the applications and the necessary modules for building those applications, and one that shows the language resources that are necessary in order to build the modules. In order to make the correspondence very clear we are using the same list of modules in the left hand side of the tables (e.g. table 1 and table 2).

	Document prod.	Summa.	Classif.	Indexing	IE	IR/filtering	MAT	MT	ASR	TTS	Dialog Systems
									Dictation		
Morphological comp.(infl, deriv., stemm., diacritic, ...)	+++	+++	+++	+++	+++	++	+++	+++	++	+++	+++
POS disambiguator/tagger	+++	+++	++	+++	+++	++	+++	+++	++	+++	+++
Diacritizer										+++	
Sentence Boundary Detection (punctuation)	+++	+++			+++	++	+++	+++		++	++
Named Entity Recognition	+++	+++	+++	+++	+++	+++	+++	+++		+	+
Word Sense Disambig.		+++	+++	++	+++	++	++	+++		+++	+++
Term extraction	+	+++	+++	+++	+++	+++	+++	+++			++
Shallow parsing	++	++	+		++	+	+++	+++		++	+++
Syntactic analysis comp.	++				++		++	+++		+	+++
Semantic Analysis (incl. Coref.res.)		++	++	+	+++	++	++	++		+	+++
Sentence synthesis and generation	++	+++					++	+++			+++
Transfer tool (Software)								+++			
Alignment							++				

Table 1. Written language applications and corresponding HLT modules, marked with importance

The next table then shows for each module mentioned in table 1 (same left hand side of the table) the resources that are needed to create such a module, e.g. to create a morphological module for Arabic a monolingual lexicon is essential, and annotated corpora are very important.

	Monolingual Lexicon	Multi-/bilingual Lexicon	Proper names	Thesauri, ontologies, wordnets	Unannotated corpora	Annotated corpora	Parallel Multi-ling corpora	Multimodal corpora for (hand) OCR	Multimodal corpora for (typed) OCR
Morphological comp.(infl, deriv., stemm., diacritic,...)	+++					++			
stat.	+					+++			
POS disambiguator/tagger	+++		++						
stat.	+					+++			
Diacritizer	+++		++	++					
stat.						+++			
Sentence Boundary Detection (punctuation)	+++					++			
stat.						+++			
Named Entity Recognition	+++		+++			+			
stat.						+++			
Word Sense Disambig.	+++				++	++			
stat.						+++			
Term extraction	+++				+++				
stat.					+++	+++			
Shallow parsing	+++								
stat.						+++			
Syntactic analysis comp.	+++					+			
stat.						+++			
Semantic Analysis comp.(incl. Coreference res.)	+++			+++					
Sentence synthesis and generation	+++			++	+	++			
Transfer tool (software)		+++							
stat.							+++		
Alignment	+++	+++					+		
stat.							+++		
Grapheme recognition (for typewritten OCR), stat.	++				+++				+++
Grapheme recognition (for handwritten OCR), stat.	++				+++			+++	

Table 2. HLT modules and corresponding written language resources, marked with importance

As rule based and statistics based approaches to language technology have very different demands on resources, we have felt that is was necessary to have two lines in the left hand column, in some (most) cases. E.g. an alignment programme can rely heavily on monolingual and bilingual lexica, or alternatively it can rely heavily on parallel bilingual corpora. (Of course, in a hybrid approach all of these types of resources may be needed).

The following table shows which data are needed for speech application. Some modules are also stand-alone applications (e.g. Dialect/language identification, Speaker recognition/identification, …) they are part of applications (e.g. identification of the language and load of appropriate acoustic models) or independent applications (identification of the language).

	Dictation	Telephony speech	Embedded speech	Transcription of broadcast News	Transcription of conversational speech	Speaker recognition	Dialect / language	“Emotion” Identification	Speaker Adaptation	Lips movement reading :	‘topic’ detection, segmentation, topic boundaries	Speaker 2 speaker mapping	“Emotion/ Prosody” output	– Text to Speech (inc. formatted data e.g. databases)	– Customization to different	– Generation Lips Movement
Acoustic models	+++	+++	+++	+++	+++	++	+++	+++	+++	+++	+++	++	+++	+++	+++	+++
Language models	+++	++	++	+++	+++		++						++	+++		
Pronunciation lexicon	+++	+++	+++	+++	+++							++		+++		
Lexicon Adaptation	+	+	+	+	+							++		+++		
Phoneme Alignment	+	+	+	+	+	+	++					++				
Prosody recognition	+	+	+	+	+	+	+	+++	+			++				
Speech Units Selection													+++	+++		
Prosody prediction													+++	+++		
segmenter Speech / Silence:	++	+	++	++	++	+	++	++	+	+	+		+			
Sentence boundary detection:	+	+	+	+	+	+	+	++	+	+	+		++	+++		
Dialect / language identification	+	+	+	+	+	+	+	+	+	+	+			+		
(word) Boundary identification,	+	+	+	+	+	+	+	+	+	+	+		++			
Speech /Non-speech (music) detection:	+	+	+	+	+	+	+	++	+	+	+					
Speaker recognition/identification	+	+	+	+	+	+	+	+	+	+	+	++				
“Emotion” Identification	+	+	+	+	+	+	+		+	+	+	++	++			
Speaker Adaptation	++	+	++	+	++	+	+	+	+	+	+	++		+		
Lips movement reading										+++						

Table 3. Speech applications and corresponding speech modules, marked with importance

	BNSC	Desktop/Microphone & High quality	Telephony	Audio data with prosodic markers and other	annotated Written Corpus	unannotated written Corpus	Vowelised corpus	Non-Vowelised Corpus	Phonetic lexicon general vocab;	Onomastica (proper names)	Visual data (Faces, lips, etc.)
Acoustic models	+++	+++	+++	+++							
Language models					++	+++	++	++			
Pronunciation lexicon					+		++		+++	+++	
Lexicon Adaptation					+	+	++	+	+++	+++	
Phoneme Alignment	++	++	++	++	++		+		+++	+++	
Prosody recognition		+	+	+++	++		+		++	++	
Speech Units Selection		+	+	+++	++				+	+	
Prosody prediction				++	++		++		++	++	
segmenter Speech / Silence:	++	++	++	++							
Sentence boundary detection:	++	++	++	++					+	+	
Dialect / language identification	++	++	++	+					+	+	
(word) Boundary identification,	+	+	+	+					+	+	
Speech /Non-speech (music) detection:	++	+	+	++							
Speaker recognition/identification	+	+	+	+							
"Emotion" Identification	+	+	+	+					+	+	
Speaker Adaptation	++	++	++	+							
Lips movement reading											+++

Table 4. Speech modules and corresponding spoken language resources, marked with importance

In addition to these speech modules, a large number of the modules described within the tables 1 and 2 above (related to written techniques and applications) are used and usable within speech modules and speech techniques. For instance morphological components are essential for text to speech applications as used in the dictation applications. This is also the case of POS disambiguator/tagger. In order to simplify tables 3 and 4 we omitted to duplicate the modules:

- Morphological comp.(infl, deriv., stemm., diacritic,...) see written
- POS disambiguator/tagger
- Diacritizer
- Named Entity Recognition
- Word Sense Disambig.
- Term extraction
- Shallow parsing
- Syntactic analysis comp.
- Term extraction
- Semantic Analysis (incl. Coref.res.)
- Sentence synthesis and generation
- Semantic Analysis
- Coreference Resolution
- Word Sense Disambig.

- Pragmatic Analysis
- Text generation
- Alignment

3.3. BLARK Specification for Arabic

The BLARK definition above describes the type of resources that are needed, but it does not give an indication of the size or any other characteristic of each type of resource. We have examined the needs for Arabic and give our estimation below. Note: These figures are tentative, building on available experience, and may be changed if further work so suggests.

3.3.1. Written Resources

3.3.1.1 Monolingual lexicon

For all components: 40,000 stems with POS, morphology
For sentence boundary detection: a list of conjunctions and other sentence starters/stoppers
For Named entity: proper names tagged. 50,000 human proper names needed
For semantic analysis: same 40,000 as for all components, but also with subcategorisation, lexical semantic information (concrete-abstract, animate, domain etc.). A wordnet would be good.

3.3.1.2 Multi-, bilingual lexicon

Same size as monolingual lexicon, depending on application

3.3.1.3 Thesauri, ontologies, wordnets

Thesauri: Subject tree with 200-300 nodes for each domain
Ontologies and wordnets should ideally be the same size as the lexicon

3.3.1.4 Unannotated corpora

For term extraction: 100 mill words

3.3.1.5 Annotated corpora

A minimum of 0.5 mill. may be used for a few applications
POS tagger, statistics based: 1-3 mill.
Sentence boundary: 0.5 – 1.5 mill.
Named entity, statistics based: 1.5 mill.
Term extraction: 100 mill
Co-reference resolution: 1 mill.
Word sense disambiguation: 2-3 mill.

Summing up, it seems that an annotated corpus of 2 mill. should meet most requirements.

3.3.1.6 Parallel multilingual corpora

Alignment: 0.5 mill. tagged corpus

3.3.1.7 Multimodal corpora for hand OCR

Grapheme recognition:
Specifications for this will follow in an updated version of the document.

3.3.1.8 Multimodal corpora for typed OCR

Grapheme recognition
Specifications for this will follow in an updated version of the document.

3.3.2. Spoken Resources

3.3.2.1 Acoustic Data

The audio data required for:

- Dictation about 50-100 speakers x 20mn, Transcribed fully vowelized + 10 speakers for testing; (It should be made available with a written corpus of a few mill words and a Phonetic lexicon (size of which depends on the Language Model), derived from a vowelized text (see written corpus below).

- Telephony speech applications requires about 500 speakers uttering around 50 different sentences and other items (SpeechDat family (http://www.speechdat.org/) like (Orientel (http://www.orientel.org/) , UOB project), it should cover both Modern Colloquial Arabic, “middle Arabic” , MSA (Modern Standard Arabic), Fr/Eng, Conditions as for SpeechDat resources including a Phonetic lexicon in SAMPA (emphasise on digits, proper names, cities, companies, named entities, …).

- Embedded speech recognition. One may Use desktop data (dictation), but data similar to Speecon (see details http://www.speechdat.org/speecon/index.html for the acoustic conditions, set of 3-4 microphones, etc.) is preferable.

- Transcription of broadcast News (BNSC: Broadcast News Speech Corpus). Transcribed Audio data. About 50 to 100 hours of well annotated speech (at the orthographic level), about 1000 hours of non transcribed data is useful. Should come with written corpus for Language Models (from newspapers + press-releases + transcriptions) of about 300 mill. of non annotated corpora (partly vowelized), it should come with a lexicon (like the previous ones), lexicon of Proper names with updating mechanisms from newspaper and media.

- Transcription of conversational speech. Data similar to CallHome / CallFriends from LDC (which covers mainly Egyptian Arabic) that may be extended with other varieties of Arabic (Maghrebian, Levantine, etc. ..)

- Speaker recognition: an audio corpus of about 500 speakers for training (labelling with speaker id but also orthographic transcriptions) uttering about

3mn of speech peer speaker, it requires also about 100 speakers for testing (amount of speech 0.5mn , incl. impostors, ….)

- Dialect / language identification: Data similar to LDC/NIST CALLFRIEND or extracted from Broadcast news speech transcripts; we may add a set of varieties of Arabic to extend the Egyptian variety at LDC.

- Speech Synthesis Corpus: (for Text to Speech, TTS) requires a male and female professional speakers; 15 hours (optimal, but realistically 5 hours may be OK) ; generated using a read phonetically balanced text (in some applications one may need 10 speakers x 100 sentences)

- Formant Synthesis/Parametric Corpus: same database as for Speech Synthesis above with hand labelled 'formant' (min. half an hour).

Notes on the applications for which the audio corpus may be used

The audio corpus may be used for

- (word) Boundary identification,
- Speech /Non-speech (music) detection: use audio data from Broadcast News Speech Corpus with the appropriate segmentations,
- Speech / Silence discrimination,
- "Emotion" Identification (if the corpus is adequately annotated),
- Speaker Adaptation
- 'topic' detection, segmentation, topic boundaries (usually use of BNSC with the adequate labelling (e.g. Topic labelling)
- Sentence boundary detection.

3.3.2.2 Multimodal corpora for Lips analysis and generation

- Lips movement reading: the corpus could be similar to M2VTS with some 50 faces (see details http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/)

We anticipate that this would be a good candidate for the BLARK

Written corpus for speech technologies

3.3.2.2.1 Un-annotated corpus

About 300 mill. words, preferably from BNSC or press and media sources.

3.3.2.2.2 Annotated corpus

This may be useful in order to derive phonetic lexicon and language models; may be same as for written technologies (min between 1 and 5 mill., other sizes for specific applications).

3.3.2.2.3 Vowelized corpus and Non-vowelized corpus:

This is important only if there is no way to obtain a vowelization tool and/or a phonetic lexicon.

3.3.2.3 Phonetic Lexion

- Phonetic lexicon (depends on the size of the language model and could be derived from a vowelized text;   may be same size as for written technologies but fully vowelized
- a specific Phonetic lexicon emphasising on digits, proper names, cities, companies, named entities, …)
- Lexicon of Proper names (including foreign names and entities) with updating mechanisms from newspaper and media, about 50K if used in conjunction with named entities.

3.4. The present BLARK Content for Arabic

Below are described resources which have been surveyed in the NEMLAR project (see *Report on Survey on Arabic Language Resources and Tools in Mediterranean Countries*) and for which we have basic information about size, language, provider etc. Many more resources have been surveyed and as soon as basic information about these resources is available, the tables below will be updated.

The rightmost column gives information about availability, price and manipulability as described in section 2.1. 'R' means for research, 'C' means for commercial use. If the availability of an LR is 3 (company internal, i.e. not available for other users), then the other features are irrelevant and not filled in.

Written resources

Monolingual lexicon

Name of lexicon	**Provider**	**Size**	**Other information**	**Availability, price, manip.**
Diinar 1	Lyon2	138,766 entries – 129,000 lemmas		1,3,1 R 1,4,1 C
Arabic Lexicon	RDI	2,800 roots, 30,000 stems	For MT	3
Dictionnaire de formes fléchies simples et agglutinées arabes	CNRS	66 million entries		1 (subject to nego-ciation)
Arabic lexicon	Sakhr	120K MSA & Classic stem		3,4,1
Arabic Idiom lexicon	Sakhr	50K basic idioms	With both lexical and semantic information	3
Selectional restrictions	Sakhr	50K frame	Semantic restrictions associated with senses of verbs, nouns and	3

			adjectives and imposed on the environment in which they occur	

List of conjunctions and other sentence starters/stoppers
No resources have been surveyed for ‘sentence boundary detection’

Name of data	Provider	Size	Other information	Availa bility, price, manip.
Arabic word segment model	Sakhr		MSA & Classic Arabic. Language model for Arabic word segments	3

Named entity:

Name of lexicon	Provider	Size	Other information	Availa bility, price, manip.
DicNom-SM	CNRS	3,122 proper nouns	Lang: Fr-Ar	3
Arabic World knowledge	Sakhr	215K of names	Database of contemporary Arabic Named entity with their English equivalent	3

Multi-, bilingual lexicon

Name of lexicon	Provider	Size	Other information	Availa bility, price, manip.
Greek- Arabic lexicon	IlSP/Amman University	2,386 entries	Lang : Ar, El, En Domain: Financial	1,2,1
OPTAR	Lyon2/ELRA	8,000	Lang : Ar-En-Fr, Domain : science technology	3
Kalimat	Lyon2	47,000 entries	Lang : Ar-Fr	3
Dictionnaire de formes simples arabes	CNRS	1,454,000 entries	Lang: Fr-Ar	3

DixAF	CNRS/ELRA	124,580 bilingual links, between ca. 43,800 French entries and ca. 35,000 Arabic entries	Lang: Fr-Ar	1,4,1
Arab-English and English-Arab dictionary	IT.COM	20,000 entries	Lang: Ar-En, En-Ar	3
Bilingual Arabic-English dictionary	Cimos	80,000 entries (En-Ar)/ 170,000 entries (Ar-En)	Lang: Ar-En	3
Bilingual Arabic-French	Cimos	75,000 entries (Fr-Ar)/110,000 entries (Ar-En)	Lang : Ar-Fr	3
Bilingual Arabic-English specialised dictionary	Cimos	12,000 entries of basic words, and 119,100 of specialised words	Lang: Ar-En. Specialized vocabulary within different domains	3
Arabic-English transfer lexicon	Sakhr	85K stem + idiom sense	Lang: Ar-En	3
English-Arabic transfer lexicon	Sakhr	190K stem + idiom sense 190K stem + idiom sense	Lang: En-Ar	3
Arabic-English	Systran	65.000 lemmas	Lang: Ar-En	3
Arabic-French	Systran	40.000 lemmas	Lang: Ar-Fr	3
English-Arabic	Systran	54.000 lemmas	Lang: En-Ar	3
MULDIC	Coltec		Lang: Ar-En-Fr	3
Lanes' Arabic-English lexicon	Qur'an Institute, Inc.	8 volumes (3162 pages)	Lang:Ar-En	1,1,3
Arabic-English dictionary	Davis Smith, Tufts University		Lang:Ar-En	1,1,
World-translator	Aramedia		Lang: Ar- Ar, Ar-Ar, En, Fr,	3

Thesauri, ontologies, wordnets

Name of lexicon	Provider	Size	Other information	Availability, price, manip.
Multilingual ontology	Cimos	400,000 words, phrases and verbs	Lang: Ar-En-Fr	3
Arabic wordnet	Sakhr	Comprehensive	Lang: Ar	3
Arabic thesaurus	Coltec		Named ARTS	3

Unannotated corpora/annotated Corpora

We have the knowledge of many more corpora than the ones mentioned below, but at present we do not have any details about these and will therefore not list them.

Name of Corpus	Provider	Size	Other information	Availa bility, price, manip.
Al-hayat Arabic data set	ELRA	18,639,264 tokens	The tokens cover 42,591 article within 7 domains	1,2,1 R 1,3,1 C
An-nahar newspapers text corpus	ELRA	24 million words	The words are found in 45,000 articles; Arabic from Lebanon	1,2,1 R 1,3,1 C
8 million words of Arabic text	IT.COM	8 million words	Domains: literature, animal life, family, nature, history, geography, economy, civil education, general culture, social science, philosophy	3
Dinar-MBC	Lyon2	10 million	Lit., essays, press	3
Text corpus	RDI	380,000 words	Dictionary explanations, literature, business, Holy Qur'an	1,4,1
Arabic POS tagged corpus	RDI	350,000 words	POS, literature, business, Holy Qur'an	1,4,1
Monolingual unannotated	Sakhr	1.4 billion words	Classified on a coarse grained subject tree	3
Monolingual Arabic POS-tagged corpus	Sakhr	1.2 million words	Manually tagged for Pos and Named entity	3
Fully diacritised monolingual Arabic corpus for Islamic domain	Sakhr	80 million words		3
Manually POS and sense tagged	Sakhr	1.2 million words		3

Arabic collocates				

Parallel multilingual corpora

There exist many bilingual corpora but for some of them we have too little information. More detailed information is being searched for and will be made available in later versions of this document.

Name of Corpus	**Provider**	**Size**	**Other information**	**Availa bility, price, manip.**
Sentence aligned bilingual Arabic English corpus	Sakhr	1.35 million sentences	Lang: Ar-En , En-Ar	3
Arabic/Farsi font library	Sakhr		26 fonts	3
Arabic Omni Data	Sakhr		Arabic script – OMNI data trained for the feature space of Arabic characters covering both Naskh and Kofi font families	3

Multimodal corpora for hand OCR

Name of corpora	**Provider**	**Size**	**Other information**	**Availa bility, price, manip.**
IFN/ENIT	IFN/ENIT		Handwritten scanned pages	2,1,1

Multimodal corpora for typed OCR

Name of corpora	**Provider**	**Size**	**Other information**	**Availa bility, price, manip.**
Training corpus of Arabic typed written OCR	RDI	600 pages of A4	Covering the 20 most famous fonts	1,2,1

Spoken Resources

Acoustic data

Name of data	Provider	Size	Other information	Availability, price, manip.
SpeechDat like database	UOB/ENST		More than 100 speakers French/Arabic, For speech recognition, Lebanese/Syrian/Fr	1,1,1
Arabic digits	UOB		For speech recognition, Lebanese accent	1,1,1
Speech database in 4 languages	LibanCell	10K announcement with 10 words/announcements	Speech database	3
Labelled database for TTS	Millenium			3
Arabic broadcast news speech corpus (BNSC)	ELRA/LDC		Domain: news More than 20 hours of transcribed Arabic news in Modern Standard Arabic.	1,2,1
Arabic acoustic corpus mono-speaker	Benabbou, Morocco			3
Arabic Phonetic database	King Abdulaziz City for Science and Technology		Lang: En-Ar	3
Holy Qur’an multi-speaker	RDI	60 hours		1,4,1
Single male speaker concatenative Arabic TTS database	RDI	1 hour, 1,300 sentences		1,3,1
Single female speaker concatenative Arabic TTS database	RDI	4 hours, 3,000 sentences		1,3,1
Arabic concatenative TTS male	Sakhr	MSA 1.5 hours		3

recording				
Arabic concatenative TTS male recording	Sakhr	MSA 2.5 hours		3
Arabic ASR recording db	Sakhr	56 hours of MSA and Colloquial Arabic		3
Human Names Language Model	Sakhr	500K name	Egyptian and Saudi human names corpus	3
Arabic Acoustic Model	Sakhr			3

Name	**Provider**	**Size**	**Other information**	**Availability, price, manip.**
CALLHOME Egyptian Arabic Speech	LDC	120 Egyptian Colloquial Arabic telephone conversations	calls lasted up to 30 minutes and were originated in N. America	1,2,1
CALLFRIEND Egyptian Arabic	LDC	60 telephone conversations between native speaker of Egyptian dialect of Arabic	Calls lasted between 5 and 30 minutes. Includes documentation. All calls are domestic and were placed inside the continental United States and Canada	1,2,1
CALLHOME Egyptian Arabic Speech Supplement	LDC	20 telephone conversations. Transcripts for 120 Egyptian Colloquial Arabic telephone conversations. 273,681,144 bytes (261 Mbytes) or 8 hours of audio data.	20 data files in sphere format, 8 KHz shorten-compressed 2-channel mulaw.	1,1,1

Written corpus for speech technologies

Name of data	**Provider**	**Size**	**Other information**	**Availability, price, manip.**
Corpus for di-	Abdelhak		Domain: text-to-	1,2,1

syllables	Mouradi, Noureddine Chenfour		speech	

Name of data	**Provider**	**Size**	**Other information**	**Availa bility, price, manip.**
CALLHOME Egyptian Arabic Transcripts	LDC	contiguous 5 or 10 minute segments taken from 120 unscripted telephone conversations	The transcripts are timestamped by speaker turn for alignment with the speech signal and are provided in standard orthography.	1,2,1

Phonetic Lexicon

Name of lexicon	**Provider**	**Size**	**Other information**	**Availa bility, price, manip.**
Special pronunciations dictionary	Sakhr	20K entry	Dict. For handling pronunciation anormalities such as borrowed words and supporting special patterns that requires irregular pronunciation	3
Name master dictionary	Sakhr	100K name		3

3.5. Discrepancies between what is needed and what is available

Based on the results of the conducted survey, see chap. 3.4 of this report, a first priority list of needed language resources has been produced, cf. table 5:

	Type of resources	Size
Written resources	ArabWordNet	
	Un-annotated Corpus	50M-100M
	Annotated corpus (POS, named entities, sentence boundaries)	2 M

	Parallel Corpus AR//FR and AR//Eng, aligned at the sentence level , unannotated	500K
	OCR Typed Corpus (PDF Files as OCRed and corresponding validated texts), non vowelized texts	5000 pages
	Vowelised corpus	1M
Speech resources		
	Dictation corpus (all varieties of MSA?)	2*50 speakers x 20mn, Transcribed … fully vowelised + 10 speakers for testing
	Broadcast News	100 hours annotated (orthographic, named entities, topics, etc.) + 500 Non transcribed but validated data. Select 20 Hours per country (Egypt, Morroco, Jordan, Lebanon, International (Medi1), may be with video?
	Conversational speech	100 speakers * 15 mn, exc. Egypt
	Speech synthesis	1 male and 1 female speakers 5 Hours each well recorded and annotated

Table 5

The produced list contains both language resources for written as well as speech resources. An estimate has been given for the size of the resources that would be most suitably developed (i.e. highest priority). The NEMLAR project will have to decide on priorities within this list, as the project does not have the necessary time, nor funds to take on the full list.

4. Acknowledgements

The authors want to thank all project partners for their input and comments, and especially Muhammad Atiyya, RDI for his rich input to this report.

The NEMLAR partners that contributed through their work are:

Muhammad Atiyya, RDI, Cairo, Egypt
Mahtab Nikkhou, ELDA European Language Resources Distribution Agency, France
Joseph Dichy, Université Lumière Lyon 2 - Faculté des Langues, France
Christian Fluhr, CEA -LIST/DTSI/SRSI/Laboratoire d´ingénierie de l´information multimédia multi-lingue, France

Fahti Débili, CNRS - Délégation Rhône-Alpes, Site Vallée du Rhône, France
Stelios Piperidis, Institute for Language and Speech Processing, Greece
Mustafa Yaseen, Amman University, Faculty of Information Technology, Jordan
Chafik Mokbel, University of Balamand, Lebanon
Abdelhak Mouradi, ENSIAS, University of Mohammed V Soussi, Ecole Nationale Supérieur d´informatique et d´analyse des Systèmes, Morocco
Salem Ghazali, SOTETEL-IT - Société Tunisienne d´Entreprises de Télécommunications – Information Technology, Tunisia
Anne DeRoeck The Open University, Computing Department, Maths & Computing Faculty, United Kingdom
Mahdi Arar and Kanan Al-Ali, Birzeit University – Birzeit Information technology UNIT (BIT) & Arabic Department, West Bank and Gaza Strip

5. References

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, H. Strik, C. Cucchinari (2002) A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, In: *Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation)*, Las Palmas de Gran Canaria, Spain.

Cieri, C., M. Maxwell, S. Strassel (2003): Core Linguistic Resources for the World’s Languages. In: *International Roadmap for Language Resources*, Workshop Paris 2003, http://www.enabler-network.org/documents/workshop/Cieri-Maxwell-Strassel.zip

Fersøe, H. (2004): *Validation Manual for Lexica*, ELRA, Paris

Krauwer, Steven (1998): ELSNET and ELRA: A common past and a common future. In: *The ELRA Newsletter, Vol. 3, n. 2*, Paris

Monachini, M., F. Bertagna, N. Calzolari, N. Underwood, C. Navarretta (2003): *Towards a Standard for the Creation of Lexica*, ELRA, Paris

Nikkhou, M., K. Choukri (2004): *Survey on the existing institutions and Language Resource using or developing Arabic*, NEMLAR report, www.nemlar.org.

Romary, L., N. Ide (2004): Towards a roadmap for standardization in language technology, In: *Building the LR&E Roadmap* Workshop at LREC2004, http://www.elsnet.org/lrec2004-roadmap/Romary-Ide.ppt

Van den Heuvel, H., Louis Boves, Eric Sanders (2000): *Validation of Content and Quality of Existing SLR: Overview and Methodology*, ELRA, Paris.

Arabic Information Retrieval and Computational Linguistics Resources, *http://www.glue.umd.edu/~dlrg/clir/arabic.html*