The BLARK concept and BLARK for Arabic

Bente Maegaard (1), Steven Krauwer (2), Khalid Choukri (3), Lise Damsgaard Jørgensen (1)

(1) CST, University of Copenhagen, Njalsgade 80, DK-2300 Copenhagen S, {bente,lise}@cst.dk
(2) UiL OTS, University of Utrecht, steven.krauwer@let.uu.nl
(3) ELDA, 55-57 rue Brillat-Savarin, F-75013 Paris, choukri@elda.org

Abstract

The EU project NEMLAR (Network for Euro-Mediterranean LAnguage Resources) on Arabic language resources carried out two surveys on the availability of Arabic LRs in the region, and on industrial requirements. The project also worked out a BLARK (Basic Language Resource Kit) for Arabic. In this paper we describe the further development of the BLARK concept made during the work on a BLARK for Arabic, as well as the results for Arabic.

1. The BLARK concept

The BLARK defines, ideally in a language independent way, the minimal set of language resources to do any precompetitive language and speech technology research at all for a language. After the first BLARK article (Krauwer, 1998) in the ELRA Newsletter, the idea was taken up by the Dutch Language Union (DLU). Daelemans & Strik (2002) give an overview of the steps taken by DLU to define the contents of the BLARK for Dutch and to assign priorities. Unfortunately this document is only available in Dutch. A summary of the work of the DLU and the results of the Dutch BLARK exercise can be found in Binnenpoorte et al (2002). Later the ENABLER project also contributed to the definition of the BLARK concept.

The starting point of the definition process in Binnenpoorte et al were 8 classes of applications, seen as being 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 large 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.

NEMLAR took this as the point of departure. We distinguish the BLARK *definition* which is the general concepts governing the BLARK, and the BLARK *specification* which is its instantiation for a given language, here Arabic. For the general discussion of the BLARK concept we first discuss a few important issues: availability, quality, quantity and standards.

1.1. Availability

In Binnenpoorte et al (2002) the availability of the existing resources was expressed on a 9-point scale. However, neither this paper nor the underlying report by Daelemans & Strik (2002) explain how they were assigned, or what they exactly mean. NEMLAR therefore proposes a different approach to availability. Three factors play an important role here: accessibility, affordability and customizability. We will distinguish 3 classes of *accessibility* (numbering based on penalties): (3) existent but only company-internal, (2) existent and freely usable for precompetitive research, (1) existent and freely usable for both precompetitive research and product development.

The second factor is *affordability*. Resources that are actually existing, but only at a very high cost (e.g. a morphological analyser for 40,000 €) 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 10,000 €, (3) between 1000 and 10,000 €, (2) between 100 € and 1000 €, (1) less than 100 € or free.

Third, the inherent exploratory nature of precompetitive research 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).

1.2. Quality

An LR may exist but be of bad quality. Binnenpoorte et al. do not provide an account of the way quality was measured or expressed, but of course it has been taken into account even if implicitly. NEMLAR suggests taking the following attributes with corresponding criteria into account: 1) Standard-compliance (values: no standard, standard but not fully compliant, standard and fully compliant), 2) Soundness (well defined specs; values: no specs, specs but not fully compliant, specs and fully compliant), 3) Task-relevance (in terms of information, size and domain coverage), 4) Inter-operability with other LRs (same as 3).

1.3. Quantity

In Binnenpoorte et al (2002) no quantitative figures were provided for the various resources needed: how many words in a corpus, how many hours of speech, 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.

We believe that a specification has to give figures for the size of the various components, if necessary based on estimations of minimal requirements and on best practice. Although most of the BLARK definition will be language independent, so that figures may be taken over, it may be possible that some figures may vary according to language.

1.4. Standards

There are relatively few existing official standards for language and speech resources. As the adoption of standards is crucial for the longevity of LRs, de facto standards have to be recommended.

1.5. BLARK Definition, BLARK Specification, BLARK Content

We will use the term *BLARK Definition* to refer to the proposals for the items to be incorporated, and the term *BLARK Specification* to refer to more detailed specification (in terms of quality, quantity, standards, etc) of these items.

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.

1.6. Applications considered

Still belonging to the general BLARK discussion is the number and nature of applications to consider, cf. the 8 classes seen as most relevant by the DLU. Such applications will certainly change over the time, so this can not be a constant part of a BLARK definition. At the same time it is probably rather language independent.

For the NEMLAR work, it was decided to split into written applications (or applications that include written LRs) and spoken applications. For written, 11 applications were listed in the areas of document production and handling, translation, information retrieval and dialogue. For speech 16 applications were listed, in the areas of dictation, transcription, lip movement, emotion etc. The actual use of some of the classes may be seen from Table 2.

1.7. Summary of contribution to the BLARK concept

As can be seen, we have followed Binnenpoorte et al (2002) quite closely. The main contributions lie in making statements on availability more fine-grained, adding quality, quantity (size) and standards, changing the applications and modules, separating BLARK definition from BLARK specification.

2. The BLARK tables for Arabic

Two pairs of tables were made, one pair for written and one for spoken language.

The 11 written applications were related to 13 HLT modules, such as morphological component, POS tagger etc. At this level the BLARK becomes language specific as the HLT modules necessary will to some extent be dependent on language. E.g. one of the important modules for Arabic is the diacritizer (vowelizer). In a separate table the same HLT modules are related to LRs, e.g. a monolingual lexicon is necessary for the morphological component. For each LR its importance for the module is marked.

Similarly for spoken language, the 16 applications were related to 17 HLT modules, which were in turn related to the necessary LRs.

By splitting the BLARK table into four separate tables, we have obtained a clearer and more coherent representation. . Let us illustrate these tables herein, for a more exhaustive version please refer to the NEMLAR report (Maegaard et. al., 2004).

In the tables below we give a few lines/columns from the 'traditional' correspondence which shows a number of general applications and the language modules that are needed in order to build each application. The table shown (table 2) is from speech.

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. However, 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.

The first table (Table 21) shows for each module the resources that are needed to create such a module, e.g. in order to create a morphological module for Arabic a monolingual lexicon is essential, and annotated corpora are very important.

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).

	Monolingual Lexicon	Multi-/bilingual Lexicon	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 1: Written language resources and corresponding HLT modules, marked with importance

	Dictation	Telephony speech applications	Embedded speech recognition	Transcription of broadcast News	Transcription of conversational speech	Speaker recognition	Dialect / language identification	"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)	– Synthesis by Concatenation ;	– Customization to different voices	– 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										+++							
Morphological comp.(infl, deriv., stemm., diacritic,...)	++	+	+	++	++									+++			
POS disambiguator/tagger	++	+	+	++	++		+							+++			
Diacritizer														+++			
Named Entity Recognition	++	+	++	++	++									++			
Word Sense Disambig.														++			
Shallow parsing	++	+	+	++	++									++			
Syntactic analysis comp.	++	+	+	++	++									++			
Sentence synthesis and generation													+	++			
Semantic Analysis	+			+	+						+		+	+			

Table 2: Speech language applications and corresponding HLT modules, marked with importance

In addition to these modules, a large number of the modules described within the tables 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 these tables we avoided duplicating the modules.

In order to carry on this task, we have capitalized on the survey conducted within the project to specify for each LR a number of critical features (its size in terms of number of words/tokens, lexical entries, number of speakers, etc.). This is also a precision compared to previous publications on BLARK.

2.1. 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: We have tried to present reasonable figures, based on estimations of minimal requirements and on best practice for Arabic and other languages; the figures may thus be modified when more information becomes available.

For standards we are recommending (de facto) standards for all types of resources, mostly based on best practice considerations.

2.1.1. Written Resources

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 stems as for all components, but also with subcategorisation, lexical semantic information (concrete-abstract, animate, domain etc.). A wordnet would be good.

Multi-, bilingual lexicon

Same size as monolingual lexicon, depending on application.

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.

Unannotated corpora

For term extraction: 100 mill words

Annotated corpora

A minimum of 0.5 mill. words may be used for a few applications, but for most applications more is needed:

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.

From this can be seen that an annotated corpus of 2 mill. words will meet most requirements.

Parallel multilingual corpora

Alignment: 0.5 mill. tagged corpus is needed to train alignment.

2.1.2. Spoken Resources

In this section we focus mainly on audio/acoustic Data required by very well know applications and technologies such as voice dictation, rich audio, transcription, telephony servers, etc. The idea is to offer the R&D community the kits needed to implement such techniques and/or to port existing toolkits and prototypes to the Arabic language.

Voice dictation:

Recordings from about 50-100 speakers, uttering 20mn each and that is transcribed and fully vocalized plus about 10 speakers for testing purposes and a written corpus of a few million words (for language modeling) and a Phonetic lexicon (size of which depend on the Language Model), derived from a vowelized text.

Telephony speech applications

About 500-1000 speakers uttering around 50 different sentences and other items as this was well established within the SpeechDat family (http://www.speechdat.org/) , preferably covering both Modern Colloquial Arabic, “middle Arabic” , and MSA (Modern Standard Arabic), in addition to spoken languages that may be used in Interactive Voice Systems (e.g. French, English, Berber, etc.).

Embedded speech recognition solutions

One may use the desktop data (from voice dictation resources), 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.

Broadcast News Speech Corpus

Applications and techniques related to transcription of Broadcast News require transcribed audio data of about 50 to 100 hours of well annotated speech (at the orthographic level), about 1000 hours of non transcribed data is useful and a written corpus for Language Models (from newspapers + press-releases + transcriptions) of about 300 millions of non annotated corpora (partly vowelized).

Conversational speech

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

Speaker recognition

Speaker recognition requires an audio corpus of about 500 speakers for training uttering about 3 min. of speech peer speaker, it requires also about 100 speakers for testing (amount of speech 0.5 min, including several impostors).

Dialect / language identification

This requires 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 collected at LDC.

Speech Synthesis Corpus

A male and female professional speakers; about 10 to 15 hours (optimal, but realistically 5 hours may be OK), generated using a phonetically balanced text.

Written corpus for speech technologies

Un-annotated corpus

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

Annotated corpus

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

Vowelized corpus and Non-vowelized corpus:

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

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 50,000 if used in conjunction with named entities.

2.2. Comparing the BLARK and the survey, decisions for production

After having compared the LRs stated as necessary with the survey of existing Arabic LRs (Nikkhou et al, 2004), and with the industrial requirements, the project decided to develop three LRs:

- A written corpus of about 500,000 words
- A speech corpus for TTS applications of 2x5 hours.
- A broadcast news speech corpus of 40 hours Modern Standard Arabic.

These three LRs were produced and validated in the last period of the project, cf. Yaseen et al. (2006), and are distributed through ELRA.

3. Conclusion and further work

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).

NEMLAR provided the first BLARK specification for Arabic, and it is our hope that the community will contribute both to the specification and to the overview of available data and tools. It is also our hope that this work may form the basis for similar work on other languages not yet described through the BLARK concept, thereby making it very visible what resources exist and which ones are needed.

NEMLAR took the next step from the BLARK specification and developed three resources that were needed. These resources are made available to the public through ELRA.

In the future, the NEMLAR association will promote work on language resources and tools for Arabic, see www.nemlar.org.

4. Acknowledgement

This work was done within the NEMLAR project supported by the European Commission through the INCO-MED programme. The authors thank all project partners, cf. www.NEMLAR.org, who contributed to the development of the BLARK for Arabic.

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

Daelemans, W & H. Strik eds (2002): Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen, DLU, Den Haag

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

Maegaard, B., L. Damsgaard Jørgensen, S. Krauwer, K. Choukri (2004): NEMLAR: Arabic Language Resources and Tools, In: K. Choukri and B. Maegaard (ed.): *Proceedings of Arabic Language Resources and Tools Conference*, p. 42-54, Cairo.

Maegaard, B. (2004): NEMLAR – an Arabic Language Resources project. In: *Fourth International Conference on Language Resources and Evaluation, Proceedings Vol I*, p. 109-112, Lisboa.

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.

Yaseen, M., M. Atiyya, C. Bendahman, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid, H. Fersøe, S. Krauwer, M. Rashwan, B. Haddad, C. Mukbel, A. Mouradi, A. Al-Kufaishi, M. Shahin, A. Ragheb, Chenfour (2006): Building Annotated Written and Spoken Arabic LRs in NEMLAR Project. In: *LREC 2006 Proceedings*, Genova.