HUMANIORA
KØBENHAVNS UNIVERSITET

The BLARK concept and BLARK for Arabic

Bente Maegaard, University of Copenhagen, Denmark
Steven Krauwer, University of Utrecht, Netherlands
Khalid Choukri, ELDA, France
Lise Damsgaard Jørgensen, University of Copenhagen, Denmark

CST
CENTER FOR SPROGTEKNOLOGI

NEMLAR

HUMANIORA
KØBENHAVNS UNIVERSITET

Overview

The NEMLAR project

The BLARK concept, some extensions

BLARK for Arabic

CST
CENTER FOR SPROGTEKNOLOGI

NEMLAR

HUMANIORA
KØBENHAVNS UNIVERSITET

NEMLAR

Network for Euro-Mediterranean Language Resources

Project on Arabic language resources, 2003-2005, supported by the European Commission

Partners in the EU (Denmark, France, Netherlands, UK, Greece)

Partners in the Mediterranean region (Egypt, Morocco, Jordan, Lebanon, Tunisia, West Bank and Gaza Strip)

NEMLAR made surveys of existing Arabic language resources in the region, as well as industrial needs for language resources, as well as BLARK and LR production

CST
CENTER FOR SPROGTEKNOLOGI

HUMANIORA
KØBENHAVNS UNIVERSITET

The NEMLAR work on BLARK

What is a BLARK?

BLARK: Basic Language Resource Kit

Minimal set of language resources and tools to do pre-competitive research and development of language technology

BLARK concept 1998: Steven Krauwer

Binnenpoorte et al 2002 (LREC)

CST
CENTER FOR SPROGTEKNOLOGI

NEMLAR

The BLARK concept

Binnenpoorte et al

List the most relevant classes of applications

List the modules that are needed to build these applications (e.g. morphological analysis, text to phoneme converter,…)

For each module list the data sets (LRs) are required, as well as their importance

Result: a matrix on the basis of which you can see which components serve most applications and which LRs are needed to produce them.

A BLARK is then used to compare to what exists already, resulting in a list of priorities for LR and tools production.

NEMLAR work on BLARK

BLARK ***definition*** – the general principles
BLARK ***specification***: the instantiation for a given language
BLARK ***content***: the parts that are currently available

General principles, 4 important issues:

- Availability
- Quality
- Quantity
- Standards

These are features that are not explicitly discussed in Binnenpoorte et al

Availability

In Binnenpoorte et al availability was expressed on a 9 point scale. However, it was not explained how these were assigned.

We suggest that three factors play an important role wrt availability:

- Accessibility (existing but company internal, existing and freely usable for research, existing and freely usable for research and product development)
- Affordability (over 10,000 €, 1,000-10,000 €, 100-1000 €, less than 100 €).
- Customizability (black box, glass box (can see it but not touch), open resources (freely manipulable)).

Quality

An LR may exist but be of bad quality

So we need some account of the quality.

We suggest

1) **Standard-compliance** (no standard, standard but not fully compliant, standard and fully compliant)
2) **Soundness** - well-defined specs (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)

Quantity, Standards

Binnenpoorte et al. do not provide quantitative figures for the various resources needed: how many words in a corpus, how many hours of speech etc..

We believe that a specification has to **give figures** for the size of the various components.

Most of the figures can probably be taken over from language to language, i.e. become part of the BLARK definition, but it may be that certain figures vary according to language.

Standards

Few official standards exist, so de facto standards have to be recommended, as adoption of standards is crucial for the longevity of LRs.

Summary of contribution to BLARK concept

- Statement on availability more fine-grained
- Adding Quality, Quantity (size) and Standards,
- Modifying the applications and modules (this will be ongoing in a changing world)
- Separating BLARK definition from BLARK specification

BLARK tables for Arabic

Two pairs of tables, one for written one for spoken language

Language specific

11 written applications (summarisation, MT, IR,...) were related to 13 language technology modules (POS tagger, Named Entity Recognizer, ...)

16 spoken applications (dictation, speaker recognition, ...) were related to 17 language technology modules (acoustic models, language models,...)

Each module is then related to the language resources necessary to create this module, e.g. in order to create a morphological module for Arabic, a monolingual lexicon is essential, and annotated corpora very important.

We use markers for importance.

Some examples

	Dictation	Telephony speech applications	Embedded speech recognition	Transcription of broadcast News	Transcription of conversational speech	Speaker recognition	Dialect / language identification	'topic' detection, segmentation, topic boundaries	"Emotion/ Prosody" output	– Text to Speech (inc. formatted data e.g. databases)
Acoustic models	+++	+++	+++	+++	+++	++	+++	+++	+++	+++
Language models	+++	++	++	+++	+++		++		++	+++
Pronunciation lexicon	+++	+++	+++	+++	+++					+++
Lexicon Adaptation	+	+	+	+	+					+++
Phoneme Alignment	+	+	+	+	+	+	++			
Prosody recognition	+	+	+	+	+	+	+			

Some examples

	Desktop/Micro phone	Telephony	High quality microphone data	annotated Written Corpus	Vowelised corpus	Non-Vowelised Corpus
Acoustic models	+++	+++	+++			
Language models				++	++	++
Pronunciation lexicon				+	++	
Lexicon Adaptation				+	++	+
Phoneme Alignment	++	++	++	++	+	
Prosody recognition	++	++	++	++	+	
Speech Units Selection	+	+	+			
Prosody prediction				++	++	
segmenter Speech / Silence:	++	++	++			

BLARK specification for Arabic

Specify the size and characteristics for each type of resource

Examples.

- Monolingual lexicon, for all components: 40,000 stems with POS and morphology
- Monolingual lexicon, for Named Entity Recognition:50,000 human proper names
- Annotated corpora, for POS tagger: 1-3 mill.
- Audio corpus for speech synthesis: 10-15 hours of male and female speakers (and a minimum of 5 hours)

HUMANIORA
KØBENHAVNS UNIVERSITET

Comparing BLARK spec. and the survey of available LRs

The project decided to develop 3 LRs, based on the specification and the survey of existing LRs in the region

- Written annotated corpus 500,000 words
- Speech annotated corpus for TTS applications 2x5 hours
- Broadcast News speech annotated corpus of 40 hours Modern Standard Arabic

CST
CENTER FOR SPROGTEKNOLOGI

NEMLAR

More information

Download the BLARK for Arabic document

Give comments to the BLARK for Arabic

Give input to the survey

- www.nemlar.org

Provide input for more BLARKs, - share your BLARK with others

- www.elra.info