Building Annotated Written and Spoken Arabic LRs in the NEMLAR project

Talk by Khalid Choukri (choukri@elda.org)

M. Yaseen1, M. Atiyya2, C. Bendahman3, B. Maegaard4, K. Choukri3, N. Paulsson3, S. Haamid2, H. Fersøe4, S. Krauwer5, M. Rashwan2, B. Haddad6, C. Mukbel7, A. Mouradi8, A. Al-Kufaishi4, M. Shahin2, A. Ragheb2

1Amman University; AU, Jordan.

mYaseen@ammanu.edu.jo

2The Engineering Company for the Development of Computer Systems; RDI, Egypt

{m_Atteya, Salah, Mohsen_Rashwan, Mostafa_Shahin, Ragheb}@RDI-eg.com

3Evaluation and Language resources Distribution Agency; ELDA, France

{Choukri, Paulsson, chomicha}@ELDA.org

4Center for Sprogteknologi; CST, University of Copenhagen, Denmark

{Bente, Hanne,kuadil}@cst.dk

5ELSNET, University of Utrecht, The Netherlands

Steven.Krauwer@ELSNET.org

6University of Petra, Jordan

Haddad@uop.edu.jo

7University of Balamand; UoB, Lebanon

Chafic.Mokbel@balamand.edu.lb

8University of Mohammed V Souissi – Ecole Nationale Supérieur d'informatique et d'analyse des Systèmes; ENSIAS, Morocco

Mouradi@ensias.ma

Nemlar project

Network for Euro-Mediterranean Language Resource

- Network of 14 specialized partners
- Promote and support Arabic & local LRs
- Survey(s) of the HLT sector in the area
- Define an Arabic BLARK
- Identify priority requirements
- Produce a first set of LRs

Nemlar project

Arabic BLARK

- Minimal set of LRs for pre-competitive research
- High priority resources identified through a survey
- Define a clear roadmap to provide and build missing resources

BLARK Steps

- Identify the technology modules and basic components for the language
  - e.g. need for a Vowelizer for Arabic
  - e.g. need for an orthographic to phoneme converter
  -
- Review dependency of technology/applications on such modules

- Identify the required language resources for each component

- Review existing resources that would match the above mentioned requirements

- Design a clear roadmap (recommendations, funding , development, …) to fulfil such requirements.

- Assess first priority requirements.

What are the steps … Blark View?

- Application 1 (Technologies, systems, etc. ??)
  - Module 1.1
    - » LRs 1.1.a
    - » LRs 1.1.b
    - » LRs 1.1.c
  - Module 1.2
    - » LRs 1.2.a
    - » LRs 1.2.b
    - » LRs 1.2.c
  - Module 1.3
    - » LRs 1.3.a
    - » LRs 1.3.b
    - » LRs 1.3.c

- Application N
  - Module N.1
    - » LRs N.1.a
    - » LRs N.1.b
    - » LRs N.1.c
  - Module N.2
    - » LRs 1.2.a
    - » LRs 6.2.b
    - » LRs 9.2.c
  - Module 1.3
    - » LRs 1.3.a
    - » LRs 1.3.b
    - » LRs 1.3.c

The present BLARK Definition for Arabic

Paper by B. Maegaard, S. Krauwer , K. Choukri, Lise Damsgaard

Session O-27 (this afternoon 16:00)

Step 3 …..

1. Detailed Specifications

2. Roadmaps

3. Identification vs Production vs Customization

4. Validation

5. Distribution

Two Surveys

1. Identification of players and Language resources involved in Arabic HLT

2. Industrial needs and requirements in Arabic Language Resources

Nemlar project LRs Selection

Three high priority resources selected:

- Broadcast News Speech Corpus
- TTS Speech Corpus
- Written Corpus

Nemlar project

Broadcast News Speech Corpus

- 40 hours of Standard Arabic
- Broadcast news from Middle East / North Africa
- Annotations in Transcriber format
- Phonetic lexicon in SAMPA

Nemlar project

TTS Speech Corpus - Specifications

- Capitalise/Serve as basis for other projects recommended by the new consortium ECESS (www.ecess.org) that aims at establishing standards for TTS.
- Suitable to build advanced TTS systems (at least for concatenative speech synthesis).
- Two voices: female, male

Nemlar project

TTS Speech Corpus – Text Corpora

Sub-corpus	**Tokens/Speaker**	**Hours**
C1-T Transcribed speech	6 600	1.0
C2-T Written text	16 500	2.5
C3-T Constructed phrases:	10 100	1.5
C3.1-T frequent phrases	*3 500*	
C3.2-T rare diphones	*6 600*	

Speakers Selection

The selection of the base line speakers is done very carefully.

Selection criteria are:

pleasantness of the voice

suitability for speech synthesis

based on concatenation and pitch synchronous manipulation.

A specific procedure for selection is defined.

Nemlar project

TTS Speech Corpus - Resource

- Studio recordings of professional speakers
- 5 hours female, 5 hours male voices
- Two channels: speech and laryngograph
- 96 kHz sampling rate, 24 bit precision
- Transcribed, POS tags, pitch marks, prosody

Annotation and segmentation is based on :

- All speech recordings are transliterated in normalized text form using Arabic vowelized text scripts.
- All speech transcriptions are tagged (POS) and annotated with specific markers, which are important for selecting speech units (e.g. noise, unintelligible words, etc)
- All speech recordings have to be marked prosodically.
- For baseline voices the speech recordings are completely phonetically transcribed and manually checked listening to the real recordings.
- For baseline voices the speech recordings are completely segmented in phones/syllables on signal level. 2 h of speech are checked manually.
- For baseline voices the speech signal of the speech recordings is completely pitch marked. 2h of speech are checked manually.

Nemlar project

Written Corpus - Design

- Sampling strategy, annotation types, size
- 500K words
- Time span (1990 – 2005)
- Only Standard Arabic
- Miscellaneous domains

Nemlar project

Written Corpus - Domains

Domain	Size	% of total
General news	100 000	20,0%
Dictionary entries	52 000	10,4%
Political news	51 000	10,2%
Scientific press	50 000	10,0%
Sports press	50 000	10,0%
Interviews	49 000	9,8%
Political debate	35 000	7,0%
Arabic literature	31 000	6,2%
Islamic topics	29 000	5,8%
IT business & manageme	20 000	4,0%
Legal domain	20 000	4,0%
Broadcast news texts	8 500	1,7%
Phrases of common words	5 500	1,1%
Total	*500 000*	*100,0%*

Nemlar project

Written Corpus - Contents

Four different files:

- Raw text
- Lexical analysis
- PoS tagging
- Vowelized

Nemlar project

Validation

- All resources validated by external partner
- Two types: Formal and Content
- Formal validation of full corpus to ensure it complies with specifications
- Content validation of a sample to manually check transcriptions and/or recordings

Written corpus validation

- *Meta-data & Administrative Information*, for instance contact person; owner; producer; distributor; IPR/copyright statement; etc.,
- *Technical Information*, for instance read-me file; structure, naming and size of discs, directories and files; format of data and annotation files; associated tools; etc., and
- *Content Information*, which is corpus specific e.g.
    - compliance with the documentation
    - linguistic correctness of the annotations.

- The content checks were made manually on a sample of approximately 5,000 words.
- The criteria were Fragmentation and Integrity of the Material (number of fragmented phrases, number of phrases containing offending material), Lexical Analysis (correctness and consistency of lexical analysis), Vowelization (rate and accuracy of vowelization), and PoS Tagging (correctness and completeness in assignment of PoS tags).

Written corpus validation

- The result of the content validation can be summarized as follows:
  - No fragmented or offending phrases.
  - 0.55% errors in lexical analysis and very few inconsistencies.
  - 53 Arabic words not fully vowelised. Inconsistencies in vowelisation between the four parallel datasets.
  - Some minor errors in PoS-tagging.

Nemlar project LRs Distribution

www.nemlar.org

Three Arabic language resources:

- Broadcast News (40 hours, phonetic lexicon)
- TTS Corpus (female, male, 10+ hours)
- Written Corpus (vowels, lexical analysis, PoS)

Distributed through ELRA