Specifications the Arabic
Broadcast News Speech Corpus

Khalid Choukri (ELDA)

Salah Hamid(RDI)

N. Paulsson(ELDA)

17th March 2005

European Commission

The NEMLAR project is supported by the INCO-MED programme

Table of Content

I Language Resources to be produced within NEMLAR

I.1 Type of Resources

Following the work carried out within the other work packages of NEMLAR, the consortium agreed to focus on three main resources:

(A) Written corpus
500 K words of annotated, including vowelised corpus fully packaged both the annotated and unannotated parts

Speech corpus for TTS applications
Recordings/ segmentation/ etc. of 10 hours (2x5 hours) for speech synthesis , 5 hours of a Male speakers and 5 hours from a female speaker.

(C) Broadcast news
The data will consist of about 40 hours to be provided by ELDA of Arabic data (mainly Standard Arabic from a number of broadcast companies); Transcriptions will follow the Transcriber conventions as used by ELDA and focus on the orthographic, named entities, speaker/turn segmentation levels. No phonetic transcription/segmentation is planned.

I.2 Language Resources Production within NEMLAR

Within WP5 plans the partners that are in charge of Language Resources productions are:

(A) Written corpus: this part will be carried out by RDI and Amman University (50% each). Tools will be provided by RDI. Raw data will be provided by RDI (IPR cleared and Copyright free) and ELDA.

(B) Speech corpus for TTS applications: Recordings/ segmentation/ etc. of 10 hours will be done at RDI facilities, the processing will be done by RDI for the male speaker and by ENSIAS for the female speaker.

(C) Broadcast news: Transcriptions will be provided by RDI; ELDA will support the RDI team with its expertise in Transcriber exploitation.

I.3 Language Resources Validation within NEMLAR

Data validation will be conducted by ELDA with the support of CST and UU. Some involvement of UOB and Sotelet-IT is foreseen.

I.4 Packaging of Language Resources to be produced within NEMLAR

Once the data is validated and a pre-release version is accepted by the validation center, each producing partner will be in charge of supplying its part to RDI who will be in charge of packaging the three sets.

II The Arabic BROADCAST NEWS SPEECH Corpus (A-BNSC) -- Corpus Design

II.1 Sampling Parameters

Do we have any sampling parameters taken into considerations like:
any special selection of the broadcasting company ,

selection of the Channel (the mean that allows the transmission of the signal and any choice of broadcasting characteristics: studio, microphone, etc.)
TimeSpan,
Communication context (news, interviews, …),
Language variety (colloquial versus Standard, formal versus informal, country/region dependent (Egyptian vs Levantine, maghreb, etc.),
any selection with respect to Speaker characteristics, etc.

Do we sample the corpus with respect of the domain of use and what are the other parameters regarding the genre (political , economic, sports, etc.).

II.2 Corpus design

The corpus design of the resource is a function of two criteria:
a sampling strategy,
a definition of the size of the resource and of each "session/genre"

The criteria must ensure:
To produce corpora that in principle offer a representation for the variety of syntactic, semantic, pragmatic features of modern Arabic
To produce corpora that is a comparable "characteristic" to other BLARK resources for other languages

Given the restriction on the size of the resource due to the cost of the resources, the general size of 40 hours of broadcast News.

Rationals behind the choice of the corpora:
things that exist
things that would lead to a good corpus comparable to others
what we can afford with our budget

We may also want to be very specific wrt genre

Media Broadcast
News
Weather reports
Interviews
Reportage
Scientific press
Sport
Talk shows political debate
Talk shows thematic discussions
Talk shows culture
Talk shows science
Total: in hours

We can also be more specific (e.g. Corpus from CLUL – Portuguese):

Textual typology.	Number of words	Percentage
Administrative and political		

Scientific		
Conversation or familiar		
Educational		
Humanistic		
Instructions (megaphone)		
Juridical		
Ludic (plays with a prize, etc.)		
Journalistic: Debates Sport Documentaries Interviews		
Information services		
Advertising		
Religious		
Technical		
APPROXIMATE TOTAL		

II.3 BNSC Processing Procedure (Transcription, Named entities Labeling, etc.)

Describe the processes behind:

- Use of Arabic scripts within Transcriber
- The conventions related to annotation of speech, music, speakers, turn segmentation, etc.
- See the features of interest in the NET-DC specifications append to this document.

II.4 Files & Filenames conventions

II.4.1 Audio Files

For the broadcast news: will be provided by ELDA (Digital format: mono or stereo .wav files (Windows PCM) Sampling *frequency***:** 22050Hz; 16 bit :

Text files (transcriptions) are generated automatically by TRANSCRIBER (with its own internal structure of xml files called .trs)

II.4.2 Filenames conventions

Filenames of the resources bear three ordered types of information:

a) The language represented (varieties of Arabic .eg, .ma, .tu etc.) if applicable
b) The broadcast company name/acronym if relevant
c) The text type; that is the field and sub field to which each text belongs in the corpus structure
d) The serial number identifying each text in its sub-field if applicable

II.4.3 Directory structure

To be defined for each sub-corpus and it depends on how the used tools output the intermediate data:

- For the broadcast news: <ELDA>: raw data, segmented data using Transcriber with all layers (see the NET-DC conventions as an appendix to this document)

II.5 Languages and Character Sets Encodings

What do we need to agree on:

- At least the use of Arabic scripts versus Romanization
- The vowelization aspects

We prefer to use Arabic scripts (at least during process of transcription) as they are more natural for our transcribers.
In the demo file
n1_011101_ori_ar__.trs
There are three types of writing

1. Topics are translated to English.
2. Transcriptions in Romanization.
3. Speaker names are transliterated in French characters.

Transcription Format: Re-use of the conventions adopted by ELDA in the NET-DC and other projects (ESTER, TC-STAR, etc.)

II.6 Tools/software used for Broadcast News Transcription

-- Reference : http://................/Transcriber , Arabic patch:
-- specifications if needed
-- Performance as measured on other data sets (details of the data descriptions, metrics, etc.) if applicable

III VALIDATION CRITERIA

The validation should focus on the formal aspect of the database (e.g. documentation, quality of the package, etc.) but also on the content and then assess the accuracy of the main features of the database such as:

- Selection of news
- Character codings
- Text Codings
- vowelisation
- Orthographic Transcription errors,
- Segmentation quality
- Named entities labeling

Such features should be agreed upon with the producers.

IV Here is a template of what validation report can address (from NET-DC)

SPEX VALIDATION REPORT (Quick Quality Check Format)

TITLE DATABASE:	NET-DC BROADCAST NEWS DB (BCAST1AR)
LANGUAGE:	Arabic
DATABASE SIZE:	37 recordings (approx. 21 hrs)
DATABASE OWNER / PRODUCER:	ELDA

DISTRIBUTOR:	ELRA
CATALOGUE NUMBER:	
AUTHORS OF REPORT:	Henk van den Heuvel
DATE:	23 December 2003
VERSION:	1.1. Transcriptions not checked

SUMMARY SHEET:

GENERAL REMARKS:

Database part	Quality value		
	*	**	***
1. Documentation		**	
2. Db structure		**	
3. Design & contents			***
4. Speech signals			
5. Annotation files		**	
6. Speakers		**	
7. Environments	***N/A***		
8. Transcriptions			
9. Lexicon			

For each applicable part a star assessment is given.

1. *** This value is given if this part of the database is not acceptable as it did not pass the relevant checks as presented below.**
2. **** This value is given if this part of the database passed the relevant checks reasonably well, but not fully.**
3. ***** This value is given if this part of the database passed the relevant checks, as presented below, perfectly well.**

QQC Report

The document with specifications for broadcast news in the NET-DC project (version 1.2) was used as reference for the prevalidation.

1. Documentation

The most important topics should be covered and clearly described in the documentation. Here follow comments per section in the DESIGN.DOC file.

Section 1:

- ***CD's contain a different number of broadcasts than documented:***
    - ***NETDC1AR_00: 8 broadcasts***

- ***NETDC1AR_01: 10 broadcasts***
- ***NETDC1AR_02: 10 broadcasts***
- ***NETDC1AR_03: 9 broadcasts***

Section 1.3:

- ***Example file, table 1.6: the values for SCn and SXn are empty in the actual database***

Section 2.2:

- ***There is only one recording from 22h55. It is questionable if this recording should be contained in the database. This section needs to be updated accordingly.***

Section 3:

- ***Table 3.1: The duration of files recorded at 22h55 is a little longer than 45 min.: 49:50***
- ***There is a file starting with N2. This is due to the fact that there are two recordings on the same day from the same source, which is not allowed according to the specifications.***

README.TXT contains wrong information as to the file names of annotation files.

2. Db structure

- The file names and directory structure should correspond to the documentation

- ***The CD with volume NETDC1AR_03 contained the same files as NET1DCAR_02***
- ***Label file N2_011102.sam should be renamed (to N2_011102_ORI_AR.sam) or deleted with the corresponding recording.***

3. Design and contents

- **All mandatory files according to the documentation should be included and contain the documented information**

OK

4. Speech signals

- **Acoustic measurements on the speech files will be made, and the results reported. The acoustical measurements involved are:**
  - **Clipping rate**
  - **SNR**
  - **Mean amplitude**

- *The provided SAMPSTAT.TXT file does not contain the correct information. An update is needed. SPEX offers to make it.*

5. Annotation files

- **Do all speech files have a corresponding annotation file?**

Yes

- **A random selection of the annotation/label files will be checked. They should be**
  - **Readable**
  - **Contain the information described in the documentation**

- ***The DIR label should not contain the name of the speech file. We propose to add the SRC label for this information.***
- ***SC1, SX1 contain a list. The specs say that each speaker has his/her own number (SC1 and SX1 for the first speaker, SC2 and SX2 for the second, etc.). SCn and SXn are empty now.***
- ***All SAM files have the same value for DUR: , which is wrong.***

- **All annotation files in SGML (or XML) should not violate the corresponding DTD file.**

- The DTD is not valid by itself. If there are no attributes to be specified for the elements "Speakers" and "Topics", there should also not be an ATTLIST statement:

<!ELEMENT Speakers (Speaker*)>

<!ATTLIST Speakers>

...

<!ELEMENT Topics (Topic*)>

<!ATTLIST Topics>

- In the XML files the path to the DTD is not correct:

<!DOCTYPE Trans SYSTEM "trans-13.dtd"> should be <!DOCTYPE Trans SYSTEM "../DOC/trans-13.dtd">

6. Speakers

- **Speaker distribution should comply with documentation**

There is no speaker information in the documentation.
The SPEAKER.TBL file is OK. The SESSION.TBL file is OK.

7. Environments

- **Environment distribution should comply with documentation**

N/A

8. Transcription

Not done for QQC report

- **All markers should be described in documentation**

- **A sample of a total of 2 hrs speech is selected (consisting of contiguous parts of 10-**

15 minutes). Transcriptions are validated (*e.g. – under discussion*):

- **Max. 2% of wrong word transcriptions in orthographical transcription is permitted**
- **Max. of 5% of errors in non-speech markers is permitted**
- **Max. of 2% of errors in background tags is permitted**
- **Max. of 2% of errors in speaker turns is permitted**
- **Max. of 2% of errors in type classifications is permitted**
- **Max. of 2% misalignments of a segment is permitted**

9 Lexicon

- **The correct set of phone symbols should be used (according to documentation)**

  Phoneme symbols are not separated by spaces.

- **All words in the (orth.) transcriptions should be present in the lexicon**

  This was not checked

10. Other remarks

V Appendix 1: The NET-DC table of content and key features (see the document for details)

- Executive Summary

- BROADCAST NEWS SPEECH DB FOR *<$language>*

Introduction

Speech file formats
- Directory structure
- File nomenclature
- Label files
- SAM label files
- TRS label files

Annotation
- Annotation levels
- Orthographic transcription conventions
- List of foreign words in the database
- Speakers and speaker turns

Example of annotation
- Speaker information
- Time annotation

Transcription tool

The lexicons

Database collection
- Strategy for data collection
- Recording platforms and signal processing
- Speaker recrutment

Error! Not a valid link.

VI Appendix 2: The TC-STAR Augmented transcription conventions (to be adapted from English)

TC-STAR: European Parliament Plenary Session Transcription Guidelines
(Amended by ELDA on 2005-01-13)

A. Process of transcription

```
  1. Transcriber 1st pass, Initial Markup
       * Check audio file is OK
       * Create skeleton transcription file with date, time,
         and program name
       * Make initial segmentation (see detailed rules below)
          * Mark speaker changes
          * Mark changes in background conditions
```

Note: Apart from the distinction translator - original speaker (in debating hall), background conditions are not annotated.

2. Transcriber 2nd pass (recommended resolution: 30 seconds)
    * Transcribe all speech segments (see detailed rules below)
    * Transcribe (frequent) noises
    * Verify uncertain orthography (Esp. names)

3. Transcriber 3rd pass, Validation (recommended resolution: 10 seconds or smaller)
    * Verify transcription
    * Fine tune boundary times
[NEW] * Resegment long segments
[NEW] * Check punctuation
    * Spell check

Validation must be done by someone else than the 1st transcriber.

B. General Transcription Guidelines

What to transcribe:

What not to transcribe, i.e. to exclude by segmentation:

* Music
* Cross talk, e.g. cocktail parties
* Unintelligible speech
* Speech in languages other than English
[NEW] in this case, just put a label identifying the language (if recognised)
e.g. [lang=Danish]
* Applause

C. Rules for Segmentation:

Sections: start a new section at least for:

* Change in date
* Change in location
* Each new speech/debate
    * Name the section, e.g. "EPPS 20. July 2004"
[NEW] if possible, add the topic of the speech/debate
e.g. "EPPS 17. November 2004 - Vote on the Barroso Commission members"

Turns: start a new turn for each change of speaker

* Whenever possible, include the proper name of the speaker.
[NEW]
The format is "LAST NAME(s), First name(s)"

e.g. "COHN-BENDIT, Daniel"

* For unknown original speakers (politicians) use "speaker#1", "speaker#2", ...; for unknown interpreters use "interpreter#1", "interpreter#2", ...
Add the name of the speaker to the name of the interpreter.
E.g. "interpreter#1<-POETTERING, Hans-Gert".

* The id number of anonymous speakers must be unique per recording. I.e. Reuse a number if (and only if) it is the same voice.

* Fill in speaker description:
+ type of speaker: "male" / "female"
+ mark global speaker: true / false
[NEW] + FOUR general classes for the dialect:
"native" for a native English speaker
"non-native" for a non-native speaker speaking in English
"non-native (heavy)" for a non-native speaker speaking in English with a strong accent
"non-native (language)" for a speaker speaking in another language then English
E.g. "non-native (spanish)" for a speaker speaking in Spanish

Placing breakpoints

* A breakpoint must be placed at each (grammatical) sentence boundary. Exceptions of this rule are allowed for short sentences, or where there is no audible separation between the sentences.
[NEW] You can leave two sentences in the same turn provided the total is not longer than two lines of text or 10 seconds in length.

* Insert additional breakpoints wherever you find it convenient.
[NEW] Especially use breaths [b] and pauses [pause] to insert breakpoint.
Resulting segments must not have more than two lines of texts or 10 seconds in duration.
When inserting a breakpoint on [b] or [pause], do it so that the label starts the new segment.
If a segment is too long and there is no [b] or [pause] to segment it, then use grammatical natural boundaries, such as phrases.

* Note: Segments which do not end with a period (.), colon (:), semicolon (;), question mark (?) or exclamation mark (!) will be merged for recognition.
[NEW] Therefore, it is very important to put these punctuation marks!

* Breakpoints should occur at the natural boundaries of speech, such as pauses, breaths, etc. Avoid breakpoint very close to word boundaries.

* Try not to have more than ten seconds of speech or two lines of transcribed text between breakpoints.

* Put non-speech events or other non-usable parts in their own isolated segments, if they are well separated.

[NEW] This means that all noises that are not pauses, breaths, throat noises or articulation noises must be on their separate segment. Create a new turn for these segments and mark it with [no speaker]

* Delimit extended pauses, so that the pause has its own line in the transcription. This also applies to filled pauses, in which case the transcription line will contain hesitations like "uh".

[NEW] Pauses and hesitations are considered extended, and therefore must be segment in an isolated segment, when they last at least 1 second. Pauses, breaths and hesitations that last less than 1 second must be left inside the running segment.

D. Rules for Transcription:

Orthography

How this can affect Arabic corpus?

This section should cover how will deal with situations like (we must agree hoe to handle these situations:-

- The words like "هذا- ذلك - الرحمن- داود" "this" will it be written as this or as it is actually pronunced "هاذا – ذالك – الرحمان - داوود"? . We may leave them as ordinary written and make a list of them.
- When long vowel is eleminated such as in the phrase "فى المكتب" which id pronunced " فِ المكتب"
- When short vowel is extended to long vowels (هاء الكناية)as the case in " هذه أركان" it is pronunced "هذهى أركان"
- How to distinguish between HAMZA WASL "همزة الوصل" when it is "
    1. Correctly pronunced. We suggest to write it "أ" with the pronunced vowlization.
    2. Correctly eleminated. We suggest to write it with out HAMZA "ا " without vowlization. Or we may add the vowel that will be pronunced if it was the start of the phrase. What is suitable??
    3. Mistakenly pronunced

[NEW] When HAMZA WASL "همزة الوصل" in the start of the sentence or in the middle of a sentence but incorrectly pronunced the pronunced vowel will be put on the "ا" like اَ , اِ ، اُ

When it is in the middle of the sentence and corrctly not pronunced will put SELA mark on it like " ٱ"

- We suggest to write long vowels with out vowlisation such as "قَال – فِي - يعلُو" but if they are consonats correct vowleisation will be written "أَهْيَأ - أوْعَى – أيْمَن".
- The letter "ى" when in the last letter if a word is writtenas :
  1. "ى" without vowlization if it is pronunced as the phoneme "a:".
  2. "ي" with the pronunced vowlisation if pronunced "j"
- The letter LAM "" in ".. الـ" definit known prefix will have the pronunced vowelization if it pronunced and written without vowlisation if it is not(despite if that is the correct pronunciation) (as in ORIENTEL).
- Arabized words "أفريقيا - أمريكا" will be written ias pronunced.

* Non-ASCII characters (if any) are written in the usual way
(I.e. no special conventions like LaTeX). Therefore
the character encoding of transcriber is set to UTF-8.

* Punctuation (full stop, question mark, call sign) need to be set where it seems to be orthographical correct. Other punctuation (comma, semicolon, etc.) is optional. (It is assumed that punctuation is generally not spoken.)
Note: Separate a punctuation marker by a space from the previous word.
[NEW] Try to put as many optional punctuation markers as is grammatically correct, especially the commas.

* Note: full stop (.), colon (:), semicolon (;), question mark (?) call sign (!) are significant for language modelling.

* Do NOT use the full stop (.) for abbreviations!

* Transcribe abbreviations as pronounced. E.g. ج م ع "جِيمْ مِيمْ عَيْنْ" if pronounced as separate letters c.

Transcribe a spelled out word in SEPARATE LETTERS.
E.g. "إسمى سهيل، سِينْ هَاءْ يَاءْ لَامْ ."

* Do transcribe like it is spoken:

8.23 جنيه	ثمانية جنيهات وثلاثة وعشرون قرشا
0.9583 دولار	خمسة وتسعون سنتا أمريكيا وثلاثة وثمانون بالمائة
1998	ألف وتسعمائة وثمانية وتسعون
أ.محمود	الأستاذ محمود

* All numbers must be written in letters.

Spontaneous Speech

* For hesitations use: uh, uhm, oh and eh. These are transcribed like ordinary words (not like noise events).

[NEW] All other hesitations sounds must be mapped to these four options.

* For incorrectly pronounced words use tag [pron=*]. This is not needed for common pronunciation variants.

[pron=*] Incorrectly pronounced

[NEW] This label must always refer to;

- the previous word, e.g. "word+[pron=*]"
- the next word, e.g. "[pron=*]+word"
- a sequence of words, e.g. "[pron=*-]word word word[-pron=*]"

but it can never be a spontaneous event.

* Use the hyphen for interrupted words. E.g. "polit- uh political"

* Mark falsestart of a speaker. E.g. "we saw [pron=fs-] the gor- [-pron=fs] uh the two gorillas at the zoo"

[NEW] False starts apply if and only if the semantic directions changes between the false start and the rest of the sentence. Repetitions are not false starts.

False starts must never be spontaneous event, exactly like mispronunciations (see before).

[pron=fs] Falsestart

* If you still cannot understand a word after listening several times use [???].

[???] Non-intelligible utterance

Note: For transcriber these are treated as noise events.

[NEW] Therefore, they must be on an isolated segment.

Non-Speech Events

* Noises: There are four general classes for non-verbal events: [artic], [noise], [sound] and [voice]. Because of their high frequency [throat] (which is a special case of [artic]) and [rustle] (which is a special case of [noise]) are tagged separately.

[b] Breath (This is optional. You may use this to denote audible breath noise, but you are not required to do so.)

[voice] Untranscribed speech, e.g. background speech in speech pauses, especially interference by non-english speech.

[throat] Clear one's throat, coughing.

[laugh] Laughing

[artic] Other non verbal articulatory noise of the
speaker, e.g. smack, swallowing, etc.

[pause] Silence, e.g. long speaker pause (>1 second)

[applause] Applause

[sound] All non-articulatory harmonic noises, e.g. short
music parts, beeps, other sound effects, etc.

[music] Music, jingle, radio, mobile phones?

[rustle] Rustle, e.g. paper or microphone rustle, etc.

[noise] All other noises, inharmonic noise,
non-articulatory events like, e.g. knocking,
babble of voices, machines, etc.

Lexical Tags

[lex=sp?] Spelling unclear (correct spelling of a word
is unknown or could not be found.

[lex=w?] Poor intelligibility of word or unknown word.

* Proper names and other semantics are not annotated.

* Words from foreign languages are tagged appropriately:
[lang=XYZ] preceding word is from language XYZ and is
pronounced correspondingly.

No tagging is needed for words commonly used in English or for well-known names, such as "Kindergarten", "", ... Words with English pronunciations must not be tagged. e.g. Hanover, Paris, ...

For noun words like "البنتاجون" a commented is added with the English spelling

If this noun has a connected prefix or suffix like "بالبنتاجون" the comment contains only th noun word .