Network of Data Centres (NetDC)
BNSC – An Arabic Broadcast News Speech Corpus

Khalid CHOUKRI[1], **Mahtab NIKKHOU**[1], **Niklas PAULSSON**[1]

[1]ELDA – Evaluation and Language Resources Distribution Agency, Paris, France
55-57, rue Brillat-Savarin, 75013 Paris, FRA NCE
{choukri, nikkhou, paulsson}@elda.fr
http://www.elda.fr

Abstract

Broadcast news is a very rich source of Language Resources that has been exploited to develop and assess a large set of Human Language Technologies. Some examples include systems to: automatically produce text transcriptions of spoken data; identify the language of a text; translate a text from one language to another; identify topics in the news and retrieve all stories discussing a target topic; retrieve stories directly from the broadcast audio and extract summaries of the content of news stories. BNSC is a broadcast news speech corpus developed in the framework of the European-funded project Network of Data Centres (NetDC). The corpus contains more than 20 hours of Arabic news recordings in modern standard Arabic. The news was recorded over a period of 3 months and were transcribed in Arabic script. The project was done in corporation with the LDC (Linguistic Data Consortium), which has produced a similar corpus of its Voice of America Arabic in the United States. This paper presents the BNSC corpus production from data collection to final product.

1. INTRODUCTION

The Arabic Broadcast News Speech Corpus (BNSC) has been produced in the framework of the Network of Data Centres (NetDC) project, inspired by the debates and conclusions of the LREC'98 MIALM workshop "MultiLingual Information Access and Management: current abilities and future directions", co-sponsored by the EC and the NSF. The project aimed at establishing collaboration between ELDA (Evaluation and Language Resources Distribution Agency) and the LDC (Linguistic Data Consortium). The resource are available for researchers world-wide, from ELDA and LDC's catalogues.

NetDC initiated a large-scale collaborative data collection, involving production, acquisition, normalisation, certification and distribution of spoken and written language resources for research and technology development. The first step towards such collaboration is represented by the production of the Network-DC Arabic Broadcast News Speech Corpus.

NetDC's objective was to set up a network of data centres, thus facilitating the access to electronic language resources currently managed by many different regional data centres. In so doing, the project set up new principles and practices for co-operation between the European Language Resources Association (EU) and the Linguistic Data Consortium (US), covering several areas for the language resources management. SPEX took part in the project as the validation centre for the European side.

The project was co-ordinated by ELDA, started in December 2000 and ended in December 2002.

This paper describes the whole production process of the Arabic BNSC: from the data specifications, collection, recordings, transcription, validation and distribution to strategy, choice of language and radio station. It focuses on the transcriptions tools, methods, standards and also the development of a phonetic lexicon.

2. BNSC Specifications

ELDA and LDC have worked out all the problems that may arise through a practical case of data collection. It is important to implement a specific example of the principles of co-operation that will be agreed upon, and to evaluate if they are realistic, and workable.

In order to do so ELRA/ELDA, LDC and SPEX have designed the Broadcast News Speech Corpus specifications considering industrial requirements and endorsements. Comments from several experts in this field have been incorporated in the specifications.

The BNSC specification document specifies technical aspects of a data collection such as speech file formats, directory structure, label files, annotation levels, orthographic transcription conventions, annotations, transcription tool, recording platforms, signal processing, and also tries to define strategy for data collection and speaker recruitment.

The document has been put into the form of a template (uniform document), which can be used for the specification of several databases. The project participants aimed at reusing some principles derived from experiences learnt in the SpeechDat family projects.

Some of the key points of these specifications are stated herein:

[Introduction

[# Mention database collection in which framework]

[# collectors and owners of database (& collaborations)]

[# overview of data; size of the db]

[=

Broadcast News Speech Corpus (BNSC) contains a minimum of 10-15 hours of speech.

A substantial portion of 'hard news' should be included in the corpus.

Spread of recordings: about two months; no multiple news broadcasts from the same day *and* the same source should be included.

A wide range of speaker variation is recommended.

Recommendation: select periods with various broadcasts (news and chronicles from the same journalists everyday, interview of people changing everyday, and phone calls from the listeners).

]

[# brief summary of contents of each Disk]

[=
Each speech file (extension .wav) has an accompanying ASCII SAM label file with annotational information (extension .sam), and an accompanying file with the transcription in xml format (extension .trs).
The speech data and the text data are stored on separate disks. When the disks are copied onto hard disk, the speech files and the corresponding annotation files of a recording appear in the same directory.
]

Speech file formats

[# which signal codings and file formats are used for the speech signal files]
[=
Encoding is 16 kHz, 16 bits, single channel. Format is raw PCM (.wav) without header information.
]

Directory structure

[# explain the directory structure and the meaning of each directory level/name in it,
preferably in table form]
[=
Speech and annotation data is stored in the directory \BCAST1<LL>\DATA, <LL> being the ISO 639 code for the language.
]

File nomenclature

[=
The following template is used:
`DD_YYMMDD_SSS_LL.<ext>`
where:

`DD`	Database identification code (00-ZZ) For BroadcastNews: N1
`YYMMDD`	Year, month, day of recording
`SSS`	Source of recording (three characters, e.g. 'RAI')
`LL`	Two letter ISO 639 language code, e.g. 'EN'
`<ext>`	File type code, i.e. .wav = speech file .sam = SAM annotation file .trs = transcription in XML format

Table **<$ code>** - Filename convention

A list of separate documentation files, tables and listings follows below:

Directory	File	
\BCAST1<LL>\TABLES	SAMPALEX.TBL	Lexicon with SAMPA transcriptions per word
	DARPALEX.TBL	(Identical) lexicon with DARPA transcriptions per word
	SESSION.TBL	Session table
	SPEAKER.TBL	speaker table
\BCAST1<LL>\DOC	DESIGN.DOC	(this) main documentation file
	ISO8859<n>.PS	ISO 8859 character set
	SAMPALEX.PS	SAMPA phone symbols used in lexicon
	DARPALEX.PS	DARPA phone symbols used in lexicon
	SAMPSTAT.TXT	Acoustic characteristics per speech file
	SPELLALT.DOC	optional; list with alternative spellings
	TRANSCRP.DOC	Manual for orthographic transcriptions
	VALREP.TXT	Validation report by SPEX
\	COPYRIGH.TXT	Copyright statement
\	DISK.ID	ASCII file containing the CD volume name
\	README.TXT	Overview of the db and its directory structure

Table <$ code> **- Documentation files, tables and listings**

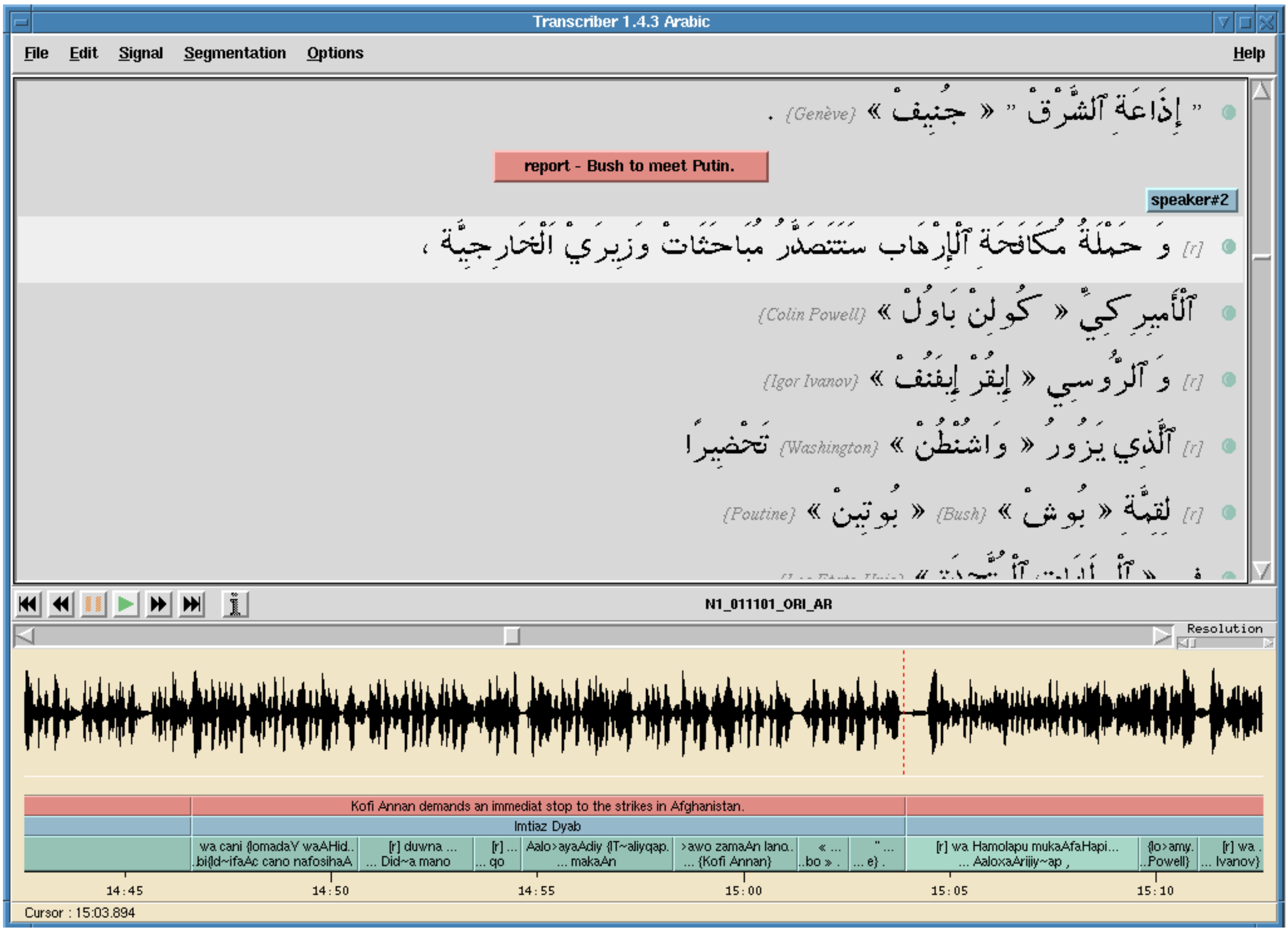
The SESSION.TBL file is an ASCII file that contains the following information of each recorded speech file on one line:
- Directory
Filename
Date of recording
Start time of recording
End time of recording
The fields in each record are separated by TABs.
The SPEAKER.TBL TBL file is an ASCII file that contains the following information of each speaker on one line:
Speaker ID
Speaker name | REPORTER_NN | SPEAKER_NN
Speaker sex
Names of the files in which the speaker appears, separated by commas
The fields in each record are separated by TABs.
The SAMPSTAT.TXT file is an ASCII file that contains the following information of each speech file on one line:
Directory name
File name
Min. sample value
Max. sample value
Clipping rate
Duration (length) in samples**]**

3. DATA COLLECTION

ELDA has entered in partnership with several broadcasting agencies - e.g., Radio Orient and Radio France Internationale (RFI). The first one broadcasts in modern standard Arabic and the second one in 18 languages including Arabic. This partnership allowed ELDA to collect a substantial number of hours of Arabic broadcasting programmes. The selected recordings implemented news programmes, press reviews and interviews broadcasted during a continuos period of time.

4. RECORDINGS

One news broadcast in Modern Standard Arabic per day was recorded from a radio station. The 18h news, containing 25 minutes of news speech and interviews was selected. Also, some recordings of the 23h news were added since this broadcast contains the chronicles. These recordings contain 35 minutes of news. In total the corpus consist of 37 recordings, collected between 1 November 2001 and 6 January 2002.
The equipment used for this project was Sangean ATS 909 radio receiver connected to a dedicated PC. The



” إِذَاعَةِ ٱلشَّرْق “ « جُنِيفْ » {Genève} .

report - Bush to meet Putin.

speaker#2

[r] وَ حَمْلَةُ مُكَافَحَةِ ٱلْإِرْهَاب سَتَتَصَدَّرُ مُبَاحَثَات وَزِيرَي ٱلْخَارِجِيَّة ،

ٱلْأَمِيرِكِيِّ « كُولِنْ بَاوِلْ » {Colin Powell}

[r] وَ ٱلرُّوسِي « إِيقُرْ إِيفَنُفْ » {Igor Ivanov}

[r] ٱلَّذِي يَزُورُ « وَاشُنْطُنْ » {Washington} تَحْضِيرًا

[r] لِقِمَّة « بُوشْ » {Bush} « بُوتِين » {Poutine}

Kofi Annan demands an immediat stop to the strikes in Afghanistan.

Imtiaz Dyab

wa cani {lomadaY waAHid.. .bi{ld~ifaAc cano nafosihaA | [r] duwna ... ... Did~a mano | [r] ... ... qo | Aalo>ayaAdiy {lT~aliyqap. ... makaAn | >awo zamaAn lano. ... {Kofi Annan} | « ... ..bo » . | " ... ... e} . | [r] wa Hamolapu mukaAfaHapi... ... AaloxaArijiy~ap , | {lo>amy. ..Powell} | [r] wa . ... Ivanov}

Cursor : 15:03.894

Figure 1: Transcription view in Transcriber for Network-DC

software Cybercorder 2000 from Skyhawk Technologies was used for scheduling and managing the recordings.
The selected radio station for the recordings was Radio Orient that broadcast on FM 94.3 MHz in Paris. Radio Orient was created in 1982. Their target audience is the 6 million Arabic people living in France as well as the French population that is interested in the culture and news of the Arabic world.

5. TRANSCRIPTIONS

5.1. Arabic Scripts

ELDA adopted Arabic scripts for the transcriptions. Native Arabic speakers were hired to carry out this work. The transcriptions of the BNSC are orthographic transliterations and have been completed with a phonetic lexicon. One news broadcast of 35 minutes took about 5 days to transcribe.

5.2. Segmentation

The transcriptions have been divided into Sections, which denotes untranscribed portions of speech or topic boundaries. Each topic treats a subject and has been documented in the corpus. Turns describe a portion of speech from a single speaker.
The transcriptions were done in two passes: the first dividing the recording into sections and turns. Next each section was transcribed in Arabic.

5.3. Transcription Tool, Transcriber

The transcription tool used for this corpus was Transcriber 1.4.2 developed by DGA, the French General Armament Delegation. Transcriber is a tool for segmenting, labelling and transcribing speech signals and is more specifically designed for the annotation of broadcast news recordings and for creating corpora used in the development of automatic broadcast news transcription systems. It assists the manual annotation of speech signals and provides a graphical user interface for segmenting long duration speech recordings, transcribing them, and labelling speech turns, topic changes and acoustic conditions. The output of this tool is a transliteration in XTML created by DGA for Arabic broadcast news.

```
<Section type="report"
startTime="903.894" endTime="1002.509"
topic="to13">
<Turn speaker="spk2"
startTime="903.894" endTime="1002.509">
<Sync time="903.894"/>

<Event desc="r" type="noise"
extent="instantaneous"/>
 wa Hamolapu mukaAfaHapi {lo&lt;irohaAb
satataSad~aru mubaAHavaAto waziyrayo
```

```
AaloxaArijiy~ap ,
<Sync time="909.571"/>
 {lo&gt;amyirikiy~i « kuwlino baAwulo »
<Comment desc="Colin Powell"/>

<Sync time="910.977"/>

<Event desc="r" type="noise"
extent="instantaneous"/>
 wa {lr~uwsiy « &lt;iyquro &lt;iyfanufo
»
<Comment desc="Igor Ivanov"/>
```

5.4. Quality Assurance

Quality assurance procedures include a transcription manual given to all transcribers before the transcriptions started and a cross verification of the transcriptions by a second transcriber. Also, a meeting was held once a week among the transcribers to discuss changes and modifications of the transcriptions.

5.5. Phonetic Lexicon

Furthermore, a lexicon was developed to give a phonetic representation of each word. The lexicon was generated automatically from the transliterations and samples were taken to verify manually the correctness of the words. Each entry in the lexicon contains a word written in transliteration, a frequency count of the word and a phonetic translation in Arabic SAMPA.

```
&gt;aTofaAlFA     1     ?at'fa:lan
Aal~agiy     276   ?allaDij
Hukuwmapi    22    X/uku:mati
bacoda       343   ba?'da
jawolapi     11    Zawlati
muwsaY       33    mu:sa:
qabola       124   qabla
qanodahaAro  55    qandaha:r
qibali       29    qibali
takuwna      21    taku:na
tisociyn     29    tis?'i:n
wujuwdi      24    wuZu:di
xilaAla      159   xila:la
yaAsiro      95    ja:sir
yawoma       57    jawma
|soyaA       10    ?a:sja:
|xaruwn      9     ?a:xaru:n
```

6. VALIDATION

To ensure good quality and add value to the language resources it distributes, ELRA/ELDA established a validation centre (VC) for Spoken Language Resources which is co-ordinated by its Board. The procedures of validation are supported on the one hand by validation committees linked to the Board and on the other hand by the VCs themselves.

Before distributing the BNSC, ELDA and LDC have had to ensure that the produced language resources correspond to the previously defined specifications and proceed to validation.

ELDA's BNSC validation activity has been carried out by SPEX which has elaborated a Quality Check report. The Quality Check report is available from ELDA's on-line catalogue http://www.elda.fr next to the description of the BNSC (ELRA reference S0157).

7. DISTRIBUTION

The NetDC Arabic Broadcast News Speech Corpus is currently available from ELDA and LDC, with more than 20 hours of transcribed Arabic news in Modern Standard Arabic. Transcriptions include speaker turns, topics, channel information and a phonetic lexicon in Arabic SAMPA. Information related to the distribution of this resource can be found at ELDA's on-line catalogue: http://www.elda.fr under reference S0157.

8. FUTURE WORK

ELRA/ELDA and LDC are willing to extend their partnership to other regional data centres.

The production of the BNSC has been seen as an initial test case for the validation of the ELRA/ELDA and LDC collaboration agreement. This collaboration agreement will be extended to other types of Language Resources (multimedia/multimodal, textual, lexicographic, and terminological data) as well as to other regional data centres.

ELRA/ELDA and LDC are planning on expanding their efforts to assist others in developing multilingual resources of high quality, by developing, documenting and distributing information about the many needed standards and tools. In summary, this project by ELRA/ELDA and LDC clearly advances the efforts of the EC and the NSF to streamline research and educational activities within MLIS, HLT, and the corresponding Human Language Resources.