Quick Rich Transcriptions of Arabic Broadcast News Speech Data

Chomicha BENDAHMAN[1], Meghan GLENN[2], Djamel MOSTEFA[1], Niklas PAULSSON[1], Stephanie STRASSEL[2]
[1]ELDA – Evaluation and Language Resources Distribution Agency, Paris, France
[2]LDC - Linguistic Data Consortium, Philadelphia, USA
E-mail: {chomicha, mostefa, paulsson}@elda.org, {mlammie,strassel}@ldc.upenn.edu

Abstract

This paper describes the collect and transcription of a large set of Arabic broadcast news speech data. A total of more than 2000 hours of data was transcribed. The transcription factor for transcribing the broadcast news data has been reduced using a method such as Quick Rich Transcription (QRTR) as well as reducing the number of quality controls performed on the data. The data was collected from several Arabic TV and radio sources and from both Modern Standard Arabic and dialectal Arabic. The orthographic transcriptions included segmentation, speaker turns, topics, sentence unit types and a minimal noise mark-up. The transcripts were produced as a part of the GALE project.

1. Introduction

Quick Rich Transcription (QRTR) provides a means to transcribe large amounts of data in a limited time frame and with minimal but useful mark-up. QRTR differs from Quick Transcription (QTR) in that each sentence unit is time stamped and labeled for its type. QRTR differs from careful transcription (CTR) in the amount of detail contained in the transcript markup, the number of features identified, the degree of accuracy and completeness of the transcript, the amount of time taken to complete the file, and the number of quality checks that are performed on the finished product.

The Arabic Broadcast News Speech data was collected from several satellite sources, Arabic channels from both TV and radio, and with different speaker styles. Both Broadcast News and Broadcast Conversation data were collected. The target languages were both Modern Standard Arabic and dialectal Arabic. A total of more than 2000 hours of data was transcribed. The transcripts are verbatim, orthographic transcripts with time-aligned section boundaries, speaker turns, sentences types, and speaker identification as well as a minimal speaker noise mark-up. In addition, a quick verification procedure was established to apply a degree of quality control to the transcripts. The transcripts were produced as a part of the GALE program[1]. The goal of the DARPA GALE program is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. Automatic processing "engines" will convert and distill the data, delivering pertinent, consolidated information in easy-to-understand forms to military personnel and monolingual English-speaking analysts in response to direct or implicit requests[2].

2. Data Sources

Data is recorded from Broadcast News (BN) and Broadcast Conversation (BC) programs. BN programming consists of "talking head" style broadcasts, i.e., generally one person reading a news script. BC programming is more interactive and includes talk shows, interviews, call-in programs and roundtable discussions. A program's classification as BN or BC is intended to reflect that program's dominant genre, though both genres may occur within a single program. Recordings are typically between 30 and 60 minutes in duration, and are collected from radio and television sources via satellite and web broadcasts.

During Phases I and II of the GALE program, LDC collected and released over 1600 hours of broadcast news and over 1500 hours of broadcast conversation audio recordings. Programs were selected for collection based on their content (news or news-related discussion) and on their availability in terms of recording accessibility and licensing issues. Sources that LDC has collected locally and sent to ELDA for transcription include Abu Dhabi TV, Al Alam News Channel, Al Arabiyah, Al Iraqiyah, Aljazeera, Al Ordiniyah, Dubai TV, Kuwait TV, Lebanese Broadcasting Corp., Oman TV, Saudi TV, SCOLA Foreign Language Network (SCOLA), and Syria TV (Arabic).

While the target language is Modern Standard Arabic, dialectal Arabic from the countries in North Africa and Middle East has been included in the recordings. During the auditing process, which is described in more detail in Section 3.2, auditors identify the percentage of Modern Standard Arabic or dialectal Arabic that is spoken in the recording. Typically, LDC selects files for transcription that contain a minimum of non-MSA speech. However, due to the spontaneous and casual nature of broadcast conversation speech data, many BC files will contain at least some amount of non-MSA.

[1] This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

[2] See http://projects.ldc.upenn.edu/gale/

3. Collection

3.1. Recordings

The Linguistic Data Consortium (LDC) receives the selected programs via satellite and records them in its in-house recording lab. LDC maintains 6 satellite dishes, providing access to C-Band, Ku-Band, DirecTV, and Dish Network Programming. A control computer coordinates the activities of all satellite dishes and receivers and CATV tuners/demodulators routing signals via two Knox AV matrix switches (64 inputs / 32 outputs), twelve distribution amplifiers and a 12-channel digitization system to six Linux-based recording nodes. Each recording node is capable of simultaneously capturing two streams of DV25 via firewire direct to local disk. The broadcast collection system also includes substantial, flexible monitoring capabilities via an integrated LCD monitoring matrix (nine separate video monitors, 4 channels of audio). Three NT servers running BBN speech recognition software for English, Chinese and Arabic provide automatic audio indexing.

The recording software is a custom-built extension of the open-source software dvgrab, which runs on the lab's Linux machines. Once a program is recorded, the audio is stripped from the recording, and the audio files are saved as .wav files on LDC servers.

3.2. Selection for transcription

3.2.1 Info audit process

LDC performed a manual audit of all programs that are recorded, in order to check the program quality, language, and content. Native speakers of Arabic listened to several 30-second samples from the beginning, middle, and end of each recording. The auditors could choose to listen to additional 30-second samples, chosen randomly within the recording, to resolve any uncertainty in answering questions about the recording. Using a web-based broadcast audio auditing interface, LDC auditors determined if the recording came from the intended program, was in the correct language, and was of good audio quality. The results were recorded in a mySql database. Recordings with poor or problematic audio quality, that do not fit the target program description, or that are in the wrong language, were rejected.

3.2.2 Selection criteria

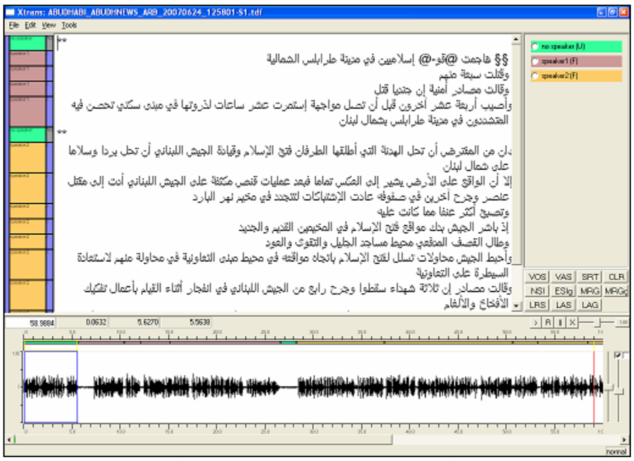
Audio recordings that passed the audit process were selected for transcription based on genre (BN vs. BC), data amount, source, program, and date epoch requirements of the GALE program. In general the batches of files LDC sent to ELDA for transcription were from recent epochs and represented a variety of sources and programs. LDC excluded GALE evaluation epochs from the transcription pool.

4. Transcription

4.1. Data and tools

Once recordings were done and the auditing process complete, the data was sent to ELDA in sets of 20-300 hours at a time to be transcribed. In total 24 sets were sent with more than 2000 hours of recorded broadcast news data. A team of about 40 trained transcribers worked on transcribing the data. About half of the team worked in the office while the other half worked at home.

The transcribers used XTrans, a tool customized by LDC for transcribing broadcast news and conversation data and that allows for orthographic transcriptions in UTF-8 as depicted in Figure 1. XTrans is a multi-lingual, multi-platform transcription toolkit. Powered by Qt's international language support, it can be used for transcription tasks in many different languages. Based on Qt, it can be easily ported to most UNIX derivatives, Microsoft Windows and Mac OS X. XTrans consists of several re-usable components such as text pane and waveform display. Most of the components are written in Python with some components written in C++. The current version of XTrans runs on FreeBSD, Linux and Windows platforms. (LDC, 2007).



§§ هاجمت @قو-@ إسلاميين في مدينة طرابلس الشمالية
وقتلت سبعة منهم
وقالت مصادر أمنية إن جنديا قتل
وأصيب أربعة عشر آخرون قبل أن تصل مواجهة إستمرت عشر ساعات لذروتها في مبنى سكني تحصن فيه
المتشددون في مدينة طرابلس بشمال لبنان

كان من المفترض أن تحل الهدنة التي أطلقها الطرفان فتح الإسلام وقيادة الجيش اللبناني أن تحل بردا وسلاما
على شمال لبنان
إلا أن الواقع على الأرض يشير إلى العكس تماما فبعد عمليات قنص مكثفة على الجيش اللبناني أدت إلى مقتل
عنصر وجرح آخرين في صفوفه عادت الإشتباكات لتتجدد في مخيم نهر البارد
وتصبح أكثر عنفا مما كانت عليه
إذ باشر الجيش بدك مواقع فتح الإسلام في المخيمين القديم والجديد
وطال القصف المدفعي محيط مساجد الجليل والتقوى والعودة
وأحبط الجيش محاولات تسلل لفتح الإسلام باتجاه مواقعه في محيط مبنى التعاونية في محاولة منهم لاستعادة
السيطرة على التعاونية
وقالت مصادر إن ثلاثة شهداء سقطوا وجرح رابع من الجيش اللبناني في انفجار أثناء القيام بأعمال تفكيك
الأفخاخ والألغام

Figure 1: XTrans, two speakers

The transcription output of XTrans is a TDF (Tab Delimited Format) file which is easy to process and which is compatible with other transcription formats, such as the Transcriber format and AG format. Each line of a TDF file corresponds to a speech segment and contains 13 tab delimited fields as described in Table 1.

In total about 2050 hours of data was transcribed. In order to transcribe such a vast amount of data a limit of 20 hours of transcription per hour of audio was fixed as a target, i.e. a transcription factor of 20. Initially the transcription factor was about 30 – 35 but soon reached the 20 hour limit as transcribers got more used to the tool and the recorded material. Initial training consisted in learning at first segmentation and the orthographic rules. Next transcribers added speaker names and type of phrases and finally they added the noise markers. The transcription process throughout the project follows roughly they same procedure. Throughout the training a senior transcriber, acting as supervisor, was at hand at all times to help and to check the transcriptions. Also, all transcriptions were double checked by a supervisor in order to correct any deviations and to give feedback to the transcriber.

	Field	**Data type**
1.	File	unicode
2.	Channel	int
3.	Start	float
4.	End	float
5.	Speaker	unicode
6.	speakerType	unicode
7.	speakerDialect	unicode
8.	Transcript	unicode
9.	Section	int
10.	turn	int
11.	Segment	int
12.	sectionType	unicode
13.	suType	unicode

Table 1: TDF format

Once training had been completed, transcribers entered a second phase were their transcriptions were cross checked by another transcriber in the team. Also a random selection of transcriptions during this phase was double checked by a supervisor. After the two first phases, which lasted about 2 months, the transcriptions from the transcribers only passed a quick quality control of 18 min. The latter transcription procedure combined with the Quick Rich Transcription rules allowed transcribers to reach a transcription factor of less than 20 as depicted in Figure 2.

Noticeable variations in the transcription factor have been caused by the recruitment of new transcribers early in the project as well as by the variability in the audio files. Some audio file deliveries contained more dialectal and overlapping speech than others, thus increasing the transcription factor.

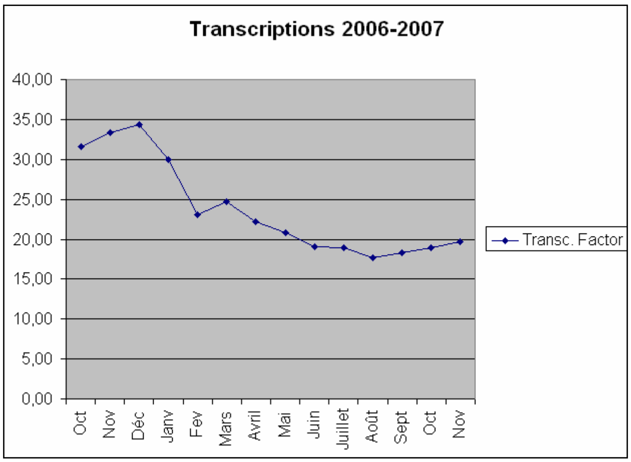


Figure 2: Transcription Factor

4.2 Transcription rules

The transcriptions follow the Quick Rich Transcription guidelines developed by LDC.

The transcripts are verbatim, orthographic and time-aligned in Arabic script and without vowels. No lexicon has been included as this was not part of the project.

The following points are treated during the transcriptions:

- Segmentation
- Sentence Units
- Overlapping speech
- Foreign language
- Noise

4.2.1. Transcription conventions

Besides orthography, a transcription convention was established to treat each of the following details in the transcripts:

- numbers
- proper names
- disfluent speech
- hesitations
- incomplete words

Numbers

All numbers are written out in words, as well as dates, times and amounts.

Proper names

Proper names are not marked specifically, but transcribers annotate the name, gender and native language for each speaker. The established orthography for a speaker name is shared among all transcribers to keep the transcripts homogenous. This applies to all other proper names, as well.

Disfluent speech

Disfluent speech can be quite difficult to transcribe. The transcribers were thus instructed not spend too much time trying to precisely capture difficult sections of disfluent speech, but to make their best effort after listening to the segment once or twice, and then move on.

Hesitations

A list of words used for transcribing hesitations was established and shared among transcribers.

Incomplete words

Truncated words were marked with a hyphen at the point where the speaker stops. Mispronunciations and invented words were annotated with respective markers.

4.2.2. Segmentation

The transcription involves several steps, starting with segmentation. The segmentation pass involves a rough division of time stamps to indicate a silence or pause, or a sentence boundary as well as parts that are not transcribed. These sections last in average between 5 and 20 seconds and are classified into one of three categories: news reports, conversations or miscellaneous. News reports are typical news broadcasts with an anchor reading the news. Conversations include interactive broadcasts with more than one speaker like roundtable discussions, call-in segments, interviews, debates, aso. Miscellaneous sections are not transcribed and typically contain music, commercials, service announcements, etc. Transcribers are instructed to rely on audio cues for creating section boundaries, such as start and end of an utterance, speaker breath, punctuation, intermittent noises and music. Next the sections with speech are grouped into speaker turns. Each speaker turn indicates a change of speaker within the same subject. Speaker turns could be either a single-speaker turn or an overlapping turn. Each turn also receives a speaker ID to identify the speaker by name if possible.

4.2.3. Sentence Units

Each sentence is annotated to indicate the Sentence Unit (SU) type. The purpose of sentence units is to group utterances into semantically- and syntactically-cohesive clusters of words that constitute a reasonable sentence-like unit. The four types of sentence units are:

- Statement
- Question
- Incomplete
- Non-speech

Statements are declarative sentences or fragments, and are usually punctuated by a period or exclamation point.

Questions are complete sentences that functions as an interrogative and ending with a question mark.

Incomplete sentences are utterances that are not grammatically complete. Typically this occurs in two situations: when a speaker interrupts himself to restructure his speech or when a speaker is interrupted by another speaker. Examples of non-speech SUs are periods of silence, music, background noise or other types of non-speech.

4.2.4. Overlapping speech

Much of the programming involves spontaneous discussion or conversation among two or more people, which means there is portions of speech that overlap among speakers. Overlapping speech was segmented and annotated accordingly. However, overlapping speech can be very challenging for transcribers and if proven too difficult in combination with audio quality and dialect variants, these regions were marked as non-speech segments.

4.2.5. Foreign language

In addition, markers for language and dialect were used whenever languages or dialects other than Modern Standard Arabic were encountered. Non-MSA dialectal speech was marked and transcribed as the transcribers heard it, using Arabic orthography writing conventions. Foreign languages was not transcribed but marked with one of the following markers: “English”, “French” or “Foreign Language”.

4.2.6. Noise markers

Minimal noise markers were included in the transcripts. Long periods of non-speech within a speaker turn were marked with a non-speech SU. Speaker noises were marked with one of the following four tags: {laugh}, {cough}, {sneeze}, {lipsmack}.

Sections of speech that were impossible to understand were indicated by a double parentheses (( )) and with a separate time stamp. The double parentheses were also used to indicate parts of speech that were difficult to understand, and where the transcribers made a best guess at what was said.

5. Quality Control

Due to the time constraints for quick transcription, quality assurance measures are necessarily limited. In order to apply a degree of quality control to the transcripts, a Quick Verification procedure was established. The procedure included the verification of three 3-minute segments – selected from the beginning, middle, and end – of the transcripts. A Quick Verification should not take more than 18 minutes. Any verification that exceeded the 18 minutes or transcription that failed the verification was sent back to the transcribers with a report to recheck and correct the transcripts. The Quick Verification focused on the following transcript features: the speech matched the transcription, the segmentation was correct, the orthography of speaker names and transcription orthography were correct and consistent.

6. Conclusion

Quick Rich Transcription provides a very useful method for transcribing large amounts of broadcast audio data to support human language technology development. System developers require large volumes of annotated data for building language models and training systems. Machine translation is a core technology of the DARPA GALE program. Before audio data can be translated, however, it must be converted into text. The transcripts created using the QRTR method described here are used to train Automatic Speech Recognition systems, which comprise a key component of a machine translation system.

As with any human transcription or annotation effort, such work requires a large team of skilled transcribers. In addition to the challenges inherent to managing a large group of people, the data in this project presented many challenges to the transcribers, such as overlapping speech, dialectal variation, and audio signal distortion. Furthermore, as the transcripts are carried out within a limited timeframe, it is useful to apply a certain degree of quality control to correct careless errors. Such a scheme was proposed and implemented within the project.

7. References

Vandecatseye A, Martens J.P. (2004). The COST278 pan-European Broadcast News Database (2004), LREC 2004 Proceedings Vol. III, pp. 873 – 876

Choukri K, Nikkhou M, Paulsson N. Network of Data Centres (NETDC) BNSC – An Arabic Broadcast News Speech Corpus (2004), LREC 2004 Proceedings Vol. III, pp. 889 – 892

LDC (2007). Using XTrans for Broadcast Transcription: User Manual, Version 3.0. http://projects.ldc.upenn.edu/gale/Transcription/XTransManualV3.pdf

LDC (2008). Audit Procedure Specification, Version 2.0. http://projects.ldc.upenn.edu/gale/task_specifications/Audit_Procedure_Specificationv2.0.pdf