A Large Arabic Broadcast News Speech Data Collection

Niklas PAULSSON[1], Khalid CHOUKRI[1], Djamel MOSTEFA[1], Denise DIPERSIO[2], Meghan GLENN[2], Stephanie STRASSEL[2]

[1]ELDA – Evaluation and Language Resources Distribution Agency, Paris, France
[2]LDC - Linguistic Data Consortium, Philadelphia, USA
E-mail: {paulsson,choukri,mostefa}@elda.org, {dipersio,mlglenn,strassel}@ldc.upenn.edu

Abstract

This paper describes the collection and transcription of large amounts of Arabic broadcast news speech data. More than 4000 hours of satellite data have been collected from various Arabic sources. The data was recorded from selected Arabic TV and radio stations in both Modern Standard Arabic and dialectal Arabic. Also, close to 2400 hours of data from a large selection of Arabic sources were transcribed. The orthographic transcriptions used the Quick Rich Transcription (QRTR) method and included segmentation, speaker turns, topics, sentence unit types and minimal noise mark-up. A validation procedure was also established to apply a quality control measure to the transcripts.

1. Introduction

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. More than 4000 hours of data were collected. The collected data was then audited and a part was selected for transcription. A total of more than 2400 hours of data was transcribed. The transcripts are verbatim, orthographic transcripts with time-aligned section boundaries, speaker turns, sentence types, and speaker identification as well as minimal speaker noise mark-up. The transcriptions followed the Quick Rich Transcription (QRTR) method. 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. Satellite recordings

Part of the collection was carried out in the US by LDC and another part was collected in Morocco by ELDA. The collected data was recorded from satellite channels from the Arabic speaking world.

2.1 Data sources

LDC collects roughly 80 hours per week of Arabic language broadcast programming from the following sources: Abu Dhabi TV, Al Alam News Channel (Iran), Al Arabiyah, Al Iraqiyah, Aljazeera, Al Ordiniyah, Dubai TV, Kuwait TV, Lebanese Broadcasting Corp., Oman TV, Saudi TV, and Syria TV.

The data collected by ELDA was initially based on AlHurra and Radio Sawa programs identified and collected by LDC. This initial collection was done in the US via satellite. The collection was later on moved to Morocco as the program sources became unavailable in the US. To extend the programming, ELDA searched for new sources to identify broadcasts of potential interest and added thus AlBagdadya and Yemen TV to the existing programming. Also more AlHurra programs were added as some ceased to broadcast and others became available. All four sources were recorded from satellite. About 80 programs were recorded each week from October 2007 to March 2009 with about 60 hours of recordings per week.

Both the Moroccan and US collection include Modern Standard Arabic and dialectal Arabic from both 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 while BC programming is more interactive and includes talk shows, interviews, call-in programs and roundtable discussions.

2.2 Tools and format

At the US collection site LDC maintains six satellite dishes that provide access to C-Band, KU-Band, DirecTV and Dish Network programming. A 3.7 meter solid dish and a 3.1 meter sectional dish are installed on movable, horizon-to-horizon mounts and are connected to computer-controlled dish movers. Four small fixed direct broadcast satellite dishes (DBS) provide access to DirecTV and Dish Network. LDC currently operates one Wegener MPEG-1 receiver for SCOLA multilingual broadcasts, two dedicated DVB-S receivers for programming from Chinese broadcasters CCTV and Phoenix TV, one Chapperal M100+ receiver, and six Pansat DVB/MPEG-2 receivers for free to air broadcasts and wild feeds on both C and KU bands. There are eight

[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/

Dish Network receivers which provide the capability to receive all of Dish Network's domestic and international programming and one DirecTV receiver for domestic U.S. programming. A Dresseler active shortwave antenna together with an AOR wideband antenna cover the entire range of the electromagnetic spectrum used for speech communication. The University of Pennylvania's Penn Video Network is an additional broadcast source and provides local, national and international television 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) and sixteen distribution amplifiers to eight Linux-based recording nodes. Each recording node is capable of simultaneously capturing two streams of DV25 digital video+audio 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). A gigabit Ethernet switch connects all broadcast recording hardware to LDC's static storage and backup facilities.

As a program is recorded, the analog audio and video pass through the A/V matrix switch to a Canopus ACEDVio analog to DV converter. The digitized DV25 stream contains DV video (intraframe compression, 4:1:1 color space, 720x480 frame resolution, 30fps) and linear PCM audio (stereo, 48kHz, 16 bit/sample). The raw DV25 stream is captured to disk using dvgrab (http://www.kinodv.org). Once a program has been recorded, the linear PCM audio tracks are extracted from the raw stream. The extracted audio (48kHz, 16bit, stereo) is downsampled to 16kHz and split into separate tracks. Then, the raw stream is converted to an MPEG-4 AVI (30fps, 720x480) with a target bitrate of 1Mbps (896Kbps video, 128Kbps audio). The extracted audio and the MPEG-4 AVI are uploaded to LDC’s fileserver. The audio files are saved as .wav files.

The recordings in Morocco are captured directly by satellite and feed to a dedicated PC. The satellite dish provides access to the NileSat 101 KU band. All four channels are recorded from this satellite. The DVB-S video stream is then captured by a Hauppauge NOVA-S Plus card that has been installed on a PC running Windows XP. The recording system is currently capturing one satellite. The recording software manages the channel switching according to the schedule. Each program is setup as a scheduled task in Windows that launches WinTV and records the desired channel in MPEG-2 Program Stream format. All files are stored with program name, hour of recording and sample rate. The program stream is next converted into AVI format (360x288 frame resolution; 30 fps) for the video part and linear PCM (mono, 16 kHz, 16 bit) for the audio. A further compression of the audio data was obtained using the lossless compression scheme FLAC. Recordings are then transferred overnight from the recording site in Morocco to the storage server at ELDA in Paris.

Converted Wav and AVI files are stored on external hard drives for backup. Also, the Paris server acts as a backup of all the audio files.

2.3 Post processing

All recordings are post processed during the conversion stage. A post processing script written in Perl takes the MPEG-2 program streams as input and generates AVI, WAV and FLAC files as output. The recorded programs are written to disk with the name and hour but without date which is added during the post processing. The conversion script is launched every day at 23h35 after the last recording. The first action is to check the disk space and report if there is not enough space left to perform the conversion. Next the script checks with the program schedule if every program has been recorded for the selected day. Also, the file size is checked to remove any files less than 15 min in length. The date is added to the filename and all files are grouped into a subset. Every subset contains the recordings of one day and for every two weeks a new set is generated with 13-16 subsets. The audio data is then extracted from the MPEG-2 program stream. The audio is down sampled from 48 kHz to 16 kHz and stored as 16 bit Wav files. A MD5 checksum is created for each file. A further compression is done using the lossless compression scheme FLAC. Wav files are moved to a backup storage while only FLAC files and the MD5 checksum is kept for a transfer to Paris. Following the audio conversion, the script converts the MPEG-2 program stream into AVI files in a resolution of 360x288 and a frame rate of 30 fps. All AVI files are moved to the local backup media.

Once all conversions have been performed a report is sent to ELDA by email with the following information: amount of disk space left, any missing programs in the programming, file size and if the conversion was successful. The FLAC files and MD5 checksum are locally stored in a dedicated directory with the subset file structure. Twice a day a second script attempts to connect with the server in Paris to transfer the whole structure using RSync. The script synchronizes the two directories locally in Morocco and in Paris in order to transfer any missing parts to Paris.

Data deliveries are made once a month to LDC which means that auditing and quality control has to take place shortly after the collection and transfer to Paris.

Once the data has been stored and checked on the server in Paris it is immediately available for LDC via ftp. Only FLAC files are stored on the server in the final version along with audit logs and checksums. This allows for a relatively fast transfer and consumes less storage space.

3. Auditing

Once recorded, all programs are audited manually in order to check quality, language, and content. The auditors performing the task are all native Arabic speakers who listen to 30-second samples from the beginning, middle, and end of each recording. For every sample the auditor answers a set of questions in a web-based tool: if there is a recording, if the audio quality is ok, what the language is, if it is the intended program, what the data type is, and what the topic is. Recordings with poor or problematic audio quality, that do not fit the target program description, or that are in the wrong language, are rejected. All the recordings of the day are audited the following working day. LDC performed a manual audit of all programs that are recorded at the US site, in order to check the program quality, language, and content. Audits made by ELDA were sent to LDC to regroup the audits with the recordings as well as the data recorded in the US.

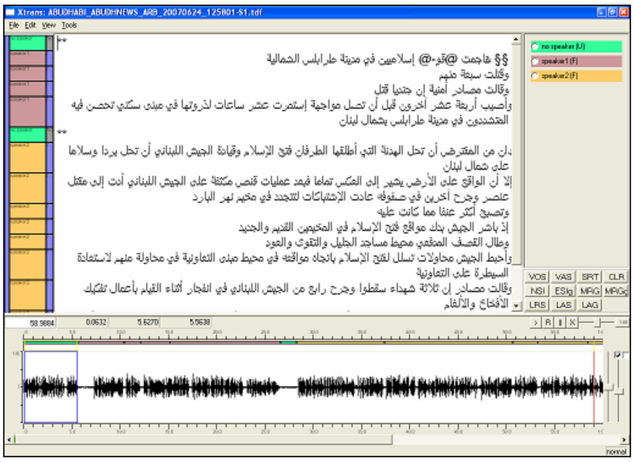
4. Transcription

4.1 Selection for transcription

Audio recordings from all collected sources that passed the audit process were selected for transcription according to the following criteria: proximity to evaluation or test epoch, source and program variety, genre, and data amount. Data that had already been selected for evaluation purposes were excluded from the transcription selection pool. Sources or programs that were previously underrepresented in training data releases to research teams were given priority over those with large volumes of existing available transcripts or transcripts that are available on the web (such as Al-Jazeera).

4.2 Data and tools

Once recordings had been done and the auditing process was complete, the data was transcribed in sets of 20-300 hours at a time. In total more than 2400 hours of recorded broadcast news data was transcribed by a team of at most 40 trained transcribers. During more intensive periods, more transcribers were hired to work at home. The transcribers used XTrans, a customized tool for transcribing broadcast news and conversation data and that allows for orthographic transcriptions in UTF-8 as depicted in Figure 1.



§§ 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

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 a set of tab delimited fields.

4.3 Transcription overview

The transcripts were created according to the Quick Rich Transcription (QRTR) guidelines developed by LDC. QRTR provides a mean to transcribe large amounts of data in a limited time frame 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 transcripts include: segmentation of sentence units and overlapping speech, identification of speaker data, and annotation of foreign language regions and speaker noise. The transcripts themselves are verbatim, orthographic and time-aligned in Arabic script, without vowels. No lexicon has been included as this was not part of the project.

4.3.1 Training transcribers

Initial training consisted of learning at first segmentation and the orthographic rules. Next transcribers added speaker names or IDs, sentence types, and finally, noise markers. The transcription process throughout the project follows roughly the same procedure. During the training period a senior transcriber, acting as supervisor, was at hand at all times to help and to check the transcripts. Also, all transcripts were double checked by a supervisor in order to correct any deviations and to give feedback to the transcriber.

Once the training phase had been completed, transcribers entered a second phase where their transcripts were cross checked by another transcriber in the team. Also a random selection of transcripts during this phase was double checked by a supervisor. After the two first phases, which lasted about 2 months, the transcripts from the transcribers only passed a quick quality control of 18 minutes.

The following points are treated during the transcriptions:

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

4.3.2 Audio segmentation

The first stage of transcription involves dividing the audio file into segments of audio, to mark speaker turns and facilitate transcription. The segmentation pass involves a rough division of time stamps into sections to indicate sentence boundaries as well as regions that are not transcribed. Speech segments last on average between 5 and 20 seconds and are classified into one of three categories: news reports, conversations or miscellaneous. News reports are typical "talking head" style news broadcasts with an anchor reading the news. Conversations include interactive broadcasts with more than one speaker like roundtable discussions, call-in segments, interviews, or debates. Miscellaneous sections are not transcribed and typically contain music, commercials, or service announcements.

Transcribers are instructed to use topic changes and audio cues to guide their placement of section boundaries, such as start and end of an utterance, speaker breath, intermittent noises and music. The sections with speech are grouped into speaker turns. Each speaker turn indicates a change of speaker within the same subject. Speaker turns may be either a single-speaker turn or an overlapping turn. Due to its prevalence in broadcast recordings, overlapping speech was segmented and annotated accordingly. The XTrans tool permits easy identification of overlapping speech regions. Each turn also receives a unique speaker ID that identifies the speaker by name if possible. 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. In addition, each sentence is annotated to indicate the Sentence Unit (SU) type: statement, question, or incomplete unit.

Sentence Units

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 three types of sentence units are:

- Statement
- Question
- Incomplete

Statements are declarative sentences or fragments, and in written Arabic would be 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. Periods of silence, music, background noise or other types of non-speech are not annotated as SU's.

Overlapping speech

Much of the programming involves spontaneous discussion or conversation among two or more people, which means there are 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.

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. Foreign languages were not transcribed but marked with one of the following markers: “English”, “French” or “Foreign Language”.

Noise markers

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

Unintelligible speech

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.

4.3.3 Transcription conventions

After segmentation, the files were transcribed according to the QRTR guidelines using standard Arabic orthography. Transcribers also marked and transcribed regions of speech that were not Modern Standard Arabic, and marked regions of foreign speech. Furthermore, transcription conventions were established to treat each of the following phenomena consistently in the transcripts:

- Numbers
- Proper names
- Disfluent speech
- Hesitations
- Partial 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 the established orthography for a speaker name is shared among all transcribers to keep the transcripts homogenous.

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. Mispronunciations and invented words were annotated with respective markers.

Hesitations

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

Partial words

Truncated words were marked with a hyphen at the point where the speaker stops.

5. Quality control

The transcription process included a quick verification procedure to check the quality of the transcripts. The procedure incorporated the verification of three 3-minute segments of the transcripts, selected from the beginning, middle, and end. A Quick Verification should not take more than 18 minutes. Any verification that exceeded this time constraint or transcription that failed the verification was sent back to the transcribers for correction. The Quick Verification focused on the following aspects of the transcript by checking to ensure that the speech matched the transcription, the segmentation was correct, and the orthography of speaker names and transcription orthography were correct and consistent.

6. Conclusion

This paper has described the collection of more than 4000 hours of a large variety of Arabic broadcast news sources as well as the transcriptions of more than 2400 hours of selected audio data. Processing this vast amount of data relies on the Quick Rich Transcription (QRTR) method as well as the Quick Verification procedure to ensure the quality of the transcripts.

7. References

Bendahman C, Glenn M, Mostefa D, Paulsson N, Strassel S. Quick Rich Transcriptions of Arabic Broadcast News Speech Data (2008), LREC 2008 Proccedings.

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

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