Dialectal Arabic Orthography-based Transcription
&
CTS Levantine Arabic Collection

Mohamed Maamouri, David Graff, Hubert Jin, Christopher Cieri, Tim Buckwalter
Linguistic Data Consortium
University of Pennsylvania
maamouri@ldc.upenn.edu, graff@ldc.upenn.edu , hubertj@ldc.upenn.edu,
christopher.cieri@ldc.upenn.edu, timbuck2@ldc.upenn.edu

Abstract

The present paper describes the experience gained at LDC in the collection and transcription of conversational dialectal Arabic. The paper will cover the following: (a) Arabic language background; (b) objectives. principles, and methodological choices of dialectal Arabic transcription, (c) design features of LDC's 'Arabic Multi-Dialectal Transcription Tool' (AMADAT) and metalanguage transcription issues, and finally (d) a summary description of the technical specifications, process, current results and issues of the EARS Levantine Arabic Conversational Telephone Speech Collection.

1.0 Introduction: Arabic Language Background

The Arabic language is a 'linguistic continuum' (Hymes, 1973) with two major poles representing an Arabic Standard, the language of most written and formal spoken discourse, and a collection of related Arabic dialects, which are mainly spoken and which present significant phonological, morphological, syntactic, and lexical differences among themselves and when compared to the standard written forms. This situation, usually referred to as 'diglossia' (Ferguson, 1959), presents some challenging problems for Arabic spoken language technologies, including corpus creation to support Speech-to-Text (STT) systems, since the spoken Arabic dialects are not officially written and have no standardized writing in spite of growing but still relatively small and not wholly conventionalized web activities. A significant amount of linguistic variation occurs and produces many variant forms which are difficult to identify and regroup.

1.1 Arabic Dialectal Variation

The diglossic situation described above represents a significant linguistic distance between all Arabic dialects and the 'fusha,' commonly identified as 'Modern Standard Arabic' (MSA), though the latter term does not cover all features of the former. This linguistic distance is characterized by substantial phonological, morphological, and lexical variation. Arabic dialectal variation is significant not only between major dialects, (e.g. Egyptian, Levantine, Gulf, or Maghrebi) but also between the regional variants of any major dialect (e.g. Northern and Southern Levantine) and even between the subdialects of any regional variant. Since important sound change has occurred in all Arabic dialects, the complexity of the above situation resides in the existence of significant differences between the phonologies of the various Arabic dialects. In Levantine Arabic (LA), for instance, the sound /q/ is pronounced /q/ but also /'/, /g/ and /k/. In Egyptian Arabic, /?/ replaces /q/ with few lexical exceptions and not in all subdialects. In Sudanese Arabic, MSA /q/ is replaced by /g/ and sometimes the uvular [ʏ]. All of the above creates confusion which needs to be addressed and taken into account in any dialectal transcription task.

1.2 Nature of the Dialectal Arabic Transcription Challenge

The description of Arabic dialect differences above, even without considering individual linguistic variation conditioned by age, gender, urbanity, rurality, register or style, shows the complexity of any speech-to-text (STT) transcription task. It also predicts the challenges facing any linguistic transcription methodology which seeks to closely represent sound features to produce a faithful rendering of dialectal Arabic pronunciation regardless of the alphabet used. Such a phonetic transcription would be usful only for and within the framework of a single dialect system. A narrow phonetic transcription, which puts too much emphasis on allophonic alternation necessarily leads to greater difficulty via: (a) a longer disambiguation process, (b) the need for a more comprehensive description of both lexicon and grammar, and (c) significantly longer annotators' training periods resulting from the need for better familiarization with transcription symbols.

1.3 Arabic and the Choice of an Orthography-based Methodology

The search for examples of speech to text transcription practices which have been successfully used to support speech technologies, led us to consider orthographies designed to write different spoken dialects (or different variants of one of them) more similarly than they sound, roughly as English orthography does world Englishes. This idea is not too far-fetched because the Arabic language continuum is similar in many ways to the

English one and presents similar potentially useful features such as (a) an important core of mutual intelligibility between MSA and the dialects, (b) similar morphology and syntactic structures, and (c) a significant common lexical core.

One can show that there exists an MSA cognate base with close structural similarities 'underlying' all Arabic dialects within the Arabic language continuum. This base, which is both historically motivated and reiterated in formal education, is part of the internalized knowledge of the Arabic language of educated, semi-literate Arabic speakers and is readily available in the MSA writing and reading communities in the Arab region and all over the world. Along with that linguistic base there also exists a knowledge of standard MSA graphemic practice which can be put to use to help with dialectal Arabic transcription supporting speech-to-text. So reflecting our work at LDC, the focus of this paper will be to demonstrate how to harness the native speaker's knowledge of Arabic orthographic conventions and of the MSA common core to reduce the cost of transcribing Dialectal Arabic to support STT research.

2.0 Objectives, Principles, and Methodology of Dialectal Arabic Transcription

2.1 General Objectives of Dialectal Arabic Transcription

Our transcription specifications were developed in the context of a common task technology evaluation program in which the primary goal is the improvement of speech-to-text technologies and in which system building makes use of statistical machine learning techniques. In such an environment, large volumes of data with high quality human annotation are desirable both as training material for learning algorithms and as evaluation material for final systems. Our most immediate goal has been to produce transcripts which, first and foremost, support the development of STT systems within the EARS community at large. We have adopted the following principles for a transcription system in order of priority: (a) friendly to writers and readers: easy to learn to write and read; (b) lexically consistent: a given spoken form is always written the same way; (c) lexically distinctive: different spoken forms will always be written differently; and (d) acoustically consistent: transcription should represent pronunciation.

2.3 Principles of an MSA Orthography-based Dialectal Arabic Transcription

Because the EARS scientific community deemed it extremely important to develop a rapid conversational telephone speech-to-text transcription in the shortest time possible leading to an adequate rendering of a dialectal Arabic text, we adopted the general principle which favored the use, whenever possible, of MSA orthography-based underlying forms to yield a transcription which approximates MSA text as much as is appropriate, especially in orthographic conventions and basic morphology. The closeness to MSA graphemic representation has been deemed important to the transcription task because annotators are believed to be able to easily transfer their MSA-based literacy skills to the transcription task, which makes it relatively easier, and because the alternative, 'phonetic-phonemic' transcription, which would have been very costly in training time, yielded fewer transcribed words per hour of effort and made word form recognition more difficult than necessary.

2.2 Methodological Objectives of a Dialectal Arabic Transcription

Some recent research in recognition technology is drawing attention to methodologies and research techniques that can quickly learn to process new languages and language varieties with relatively small amounts of training material and time. Dialectal Arabic speech poses important problems for ASR and technologies that rely on its output, however, and some researchers have already started to investigate ways to exploit MSA resources in the processing and analysis of dialectal Arabic. Rambow (2003) uses MSA text to model dialectal Arabic. He addresses the portability problem by "converting Modern Standard Arabic (MSA) corpora to (an approximation of) dialectal Arabic text." Our approach is somewhat different. We believe that annotated dialectal Arabic data is better-suited – in fact necessary – to building successful dialectal language models. However, we also believe that dialect transcription should be based on the pragmatic use of the Arabic orthographic symbols and rules. In other words, transcriptions should represent real and authentic dialectal forms written in Arabic script, without short vowels, and otherwise following the general orthographic conventions of MSA. MSA orthography-based 'underlying' forms, should only be used when regular sound change has created predictable correspondence between MSA and dialectal sounds. Instances of this involve MSA interdentals /δ, θ/ and their corresponding dialectal dentals /d, t/ or fricatives /z, s/. Another important example is that of the MSA velar /q/ and its numerous acoustic variants /ʔ, g, k, γ/. Using the 'underlying' MSA representation of the many dialectal variants of a word/form is an efficient way of rendering the complex acoustic reality of that form while respecting the principle of lexical consistency. Our transcription tool and guidelines seek an optimal solution to the requirements of ease of use, lexical consistency and distinctiveness and acoustic consistency. For a more elaborate explanation, see
http://www.ldc.upenn.edu/Projects/EARS/Arabic.

2.1.1 Advantages of an MSA-based Strategy for Dialectal Arabic Transcription

Arabs can read words written in MSA orthography with the similar levels of recognition and comprehension. When faced with the task of writing a spoken dialectal, native speakers of Arabic use their knowledge of the 'underlying' MSA sounds in order to transcribe the targeted form with corresponding Arabic script letters. Native Arab transcribers' knowledge of the Arabic language and their familiarity with the rules of Arabic script constitute the basis of our strategy for the transcription of Arabic dialects. This strategy uses: (a) the annotator's native knowledge of their dialectal sounds and structures, (b) a practical knowledge of MSA orthography-based conventions, and (c) a reasonable reliance on MSA in order to both produce an acceptable output and guarantee a high rate of consistency and an easy retrieval of meaning. A significant advantage of this strategy is that native transcribers do not need to undergo a long training period to acquire a new set of transcription symbols. Annotators can easily transfer their MSA-based literacy skills to the transcription task. Our belief is that by avoiding exaggerating the differences between MSA and Arabic dialects we will also help developers adapt their parsing and tagging tools developed for MSA to the peculiarities of the Arabic dialects.

2.1.2 Pitfalls of an MSA Orthography-based Strategy for Dialectal Arabic Transcription

An MSA orthography-based transcription faces three major challenges. The first is that there is little or no evidence of a dialectal Arabic text corpus with stable MSA orthography-based writing conventions. Because dialects are considered to be a 'degraded' form of Arabic, occurrences of written dialectal Arabic have been scarce and largely dominated by MSA writing conventions, or 'filters', which seek to elevate the level of the dialectal forms toward the MSA written standards. This mixture has led and usually leads to inconsistent transcriptions, characterized by two opposing tendencies, namely: (a) to produce a register remaining at the level of the dialectal forms and (b) to correct toward MSA. Low-literate Arabs use the Arabic script if and when they have to write anything down. Their practices usually constitute an idiosyncratic and inconsistent corpus of forms which often manifest a closer adherence to dialectal speech forms than the practices of educated Arabs. Low-literate Arabs write what they say in the way that they say it without being aware of or worrying about the relationship of the written forms to an underlying MSA structure. On the other hand, educated Arabs tend to filter, to over-correct toward MSA forms when they write or transcribe dialect. For example, there is a tendency to write the dialectal relative pronoun /?illiy/ as MSA /?allaðiy/ or to try and complete the contracted forms /ᶜaša:n/ or /taru:H/ as /ᶜan+ša:n/ and /Hattay+?aru:H/.

So we find ourselves confronting two pitfalls: (a) the real danger of the interference of MSA writing conventions and MSA dominance in our budding dialectal Arabic transcription practices, and (b) the danger of inconsistencies, thus the lack of stability, in the resulting corpus. In order to ensure consistency, our transcription practice must achieve and document a balance between the two poles described above. We need to avoid: (1) too strict an adherence to MSA-based spelling conventions that would shoe-horn dialect utterances unnecessarily into MSA form (2) too close an adherence to the phonetic reality of the dialect that would lead to a better acoustic representation but at the cost of word recognition. The transcription of conversational dialectal Arabic is a difficult balancing act, and the speech technology community seems divided between two equally important goals: to produce a finer phonetic representation in order to accommodate acoustic modeling or to produce transcripts with maximal similarity to MSA in order to accommodate language modeling.

3.0 Design Features of LDC's 'Arabic Multi-Dialectal Transcription Tool'

The general methodological principles described above and the transcription guidelines that we developed and follow in our transcription of Levantine Arabic seek to accommodate the goals of acoustic and language modeling by producing a two-tiered transcript in which one layer focuses on anchoring transcribed dialectal forms to similar MSA orthography-based utterances whenever possible thus establishing a kind of 'underlying' semantic structure based upon MSA to assist word recognition and identification, and a second layer uses the output of the first layer enriching it to produce a closer representation of the dialect pronunciation. The first (green) layer is believed to be adequate as STT training material and, since it contains all of the important markup in the Rich Text transcription specification, serves as 'careful' transcription for purposes of technology evaluation. The second (yellow) layer adds most functionally necessary vowels, marks important sociolinguistic variants, morphophonemic features (such as assimilation, 'sandhi' phenomena, etc.) and other major sound change phenomena. This second layer, which serves a similar function to a traditional pronunciation lexicon has not yet been in demand by the EARS community which has opted for larger, cheaper corpora over more carefully transcribed and more costly ones.

3.1 AMADAT: LDC's Arabic Multi-Dialectal Transcription Tool Features

The Arabic Multi-Dialectal Transcription tool, AMADAT version 1.2, was designed and developed at LDC in 2003. using Python and QT (a multiplatform GUI application framework). AMADAT is capable of handling bi-directional UTF8 text and displaying correctly. The LDC team uses AMADAT to audit, segment, play, transcribe

and display transcriptions of the Levantine Arabic conversational telephone speech. The Buckwater transliteration is used as the internal and external representation. However annotators are not required to know the Buckwater transliteration, as long as they know how to use the Arabic keyboard. The Arabic input is accomplished by intercepting the keyboard event and remapping it to the corresponding Arabic character (UTF8) in real time. The Arabic text is then displayed in the text input/edit widget. AMADAT has been ported to all major platforms such as UNIX/Solaris, Linux, Windows and MacOs. With some modest effort to come up with an appropriate transliteration representation and the corresponding keyboard remapping, AMADAT can easily be revised to do transcription in another language.

AMADAT is designed to provide a multi-layered transcription by extending links between the two tiers of linguistic structure to reflect the links between forms as pronounced and as written in our orthography which emphasizes the connection to MSA forms. AMADAT also permits annotation of linguistic variation occurring between individual Arabic speakers in multi-dialectal communication. AMADAT supports two-tiered transcription (GREEN and YELLOW), which provides an MSA orthography-based transcription in a first pass (orthographic level) and a more elaborate second pass (surface phonemic level), which adds phonetic detail (such as distinctive dialect short vowels, consonantal sociolinguistic variation, and *shaddah and nunation* if missing). AMADAT has three mutually-exclusive operation modes: (a) the green pass area for transcription, in which an Arabic keyboard is used; (b) the yellow pass area for the refinement of the 1st pass transcription using the Buckwalter transliteration and a Latin keyboard, and (c) the red pass area used if necessary for correction of errors.

3.2 Tool embedded Metalanguage Annotation features

LDC's AMADAT transcription tool includes a set of buttons which are used to add various metalinguistic annotations of the targeted speech including: (1) non-speech sounds in the recording; (2) interjections, which are speech sounds (non-lexemes) communicating hesitation, surprise, agreement, etc.; (3) linguistic and sociolinguistic phenomena reflecting language change and; (4) the dialect of the speaker (LA for Levantine dialectal Arabic versus Egyptian, Iraqi or Gulf Arabic). We also tag 'foreign' words without transcribing them and only transcribe those words that are considered true borrowings. Foreign words and place names in dialectal Arabic are spelled according to the MSA orthography-based local/regional conventions in the GREEN area (e.g., Levantine 'kara:j' versus Egyptian 'jara:j' for 'garage'). However, the use of "extended" Arabic characters, such as the Persian letters /p/ پ, /č/ چ, /ž/ ژ, and /g/ گ, are available on the keyboard in the YELLOW area.

The set of keyboard symbols used for annotation of speech is summarized below and is more information is found in http://ldc.upenn.edu/Projects/Transcription/rt-03/RT_Transcription_V2.2.pdf and http://www.ldc.upenn.edu/Projects/EARS

3.2.1 Metalinguistic tags

Smt	'silence'
tnf~s	'breath'
DHk	'laugh'
mwsyqY	'music'
sEAl	'cough'
Dj~p	'noise'
Dj~p\	'noise/'
ETs	'sneeze'
>SwAt	'peopletalk'
<nqTAE	'pause'
tdAxl	'overlap'
tdAxl\	'overlap/'

3.2.2 Interjections

%>ah ;' %<yh ; %>m ; %>ww; %hm; %mhm; %>ahh.

3.2.3 Linguistic/Sociolinguistic tags

(Cons Change) such as /finja:l/ and /finja:n/ 'tea cup' or /minšu:f/ for/and / minšu:f/ 'we see'
(Velarized Cons) to be used with emphatic variants of consonants only represented by non-emphatic Arabic Orthography letters such as /r, z etc./
(Voc Variant) such as /wiza:ra/ and /waza:ra/ 'ministry' or /SiHa:fa/ and /SaHAfa/ 'journalism'
(Hamzah Drop) if needed
(Diphthong) as in 'zyt' for /zi:t/ and /zayt/ 'oil' when the latter occurs.
(-h Deletion) as in /ma: fiy ᶜamal l'yu:m/ for /ma: fiyh ᶜamal Al-yu:m/ 'There is no work today.'
(Cons Deletion) such as /nuS:/ for /nuSf/ "half".

3.2.4 Language Identification tags

- Modern Standard Arabic: 'MSA
- Arabic Dialects: 'NA', 'ALG', 'EGP', 'GLF', 'IRQ', 'LEB', 'JOR', 'MOR', 'PAL','SAU', 'SYR', 'TUN', 'YEM',
- Foreign Language(s): 'FOR'

3.2.5 Keyboard symbols used for transcription

((text))	Semi-intelligible speech or Hard-to-understand speech
(( ))	Unintelligible speech
[lg.text]	Foreign Language
+	See Mispronounced words

- See Partial words specs
-- See Restarts specs
? See punctuation specs

3.2.6 Example of Keyboard Transcription: Partial words and restart guidelines

When a speaker breaks off in the middle of the word, annotators transcribe as much of the word as can be recognized. A single dash – is used to indicate point at which word was broken off. For example: wyn$- yEny yn$rwA. Speaker restarts are indicated with a double dash (--) as in the following examples: brAmj -- brnAmj tEArfy yEny and yEny mA -- mA fyh kvyr.

4. Levantine Arabic Conversational Telephone Speech Collection: technical specifications, process, currents results and issues

The Levantine Arabic transcription project receives its data from an ongoing Fisher style telephone speech collection. Fisher Arabic uses the same basic telephone collection strategy as Fisher English: a dedicated Windows-NT workstation ("the platform") controls a T-1 trunk line solely for the use of Arabic participants. Because the focus of the collection is on subjects living in the Levant, virtually all call activity is initiated by the platform: calls are placed from the platform during agreed-upon hours to phone numbers provided by the subjects. A small number of native speakers of Levantine Arabic among LDC staff also participate in the program to help increase the rate of collection and the likelihood of successful calls. During call recording, audio data is captured directly from the digital stream of the T-1 trunk line and stored directly to local disk as two separate single-channel files in raw 8-Khz, mu-law format. Calls are uploaded to network storage at regular intervals (typically once per business day), and multiplexed into two-channel files in NIST SPHERE format, retaining the original 8-Khz mu-law sampling.

Platform call activity and data archiving is managed through an SQL (Oracle) database, which keeps track of enrolled subjects, their phone numbers, and all inbound and outbound calls. During the hours when subjects are available to receive calls, the T-1 lines dedicated to outbound calls are in fairly constant activity: querying the database for an available callee who has not been called within the last couple of hours and has not completed a successful earlier in the current day. On dialing out to an available number, if the platform does not reach a willing participant for whatever reason (e.g. busy signal, ring-no-answer, call refusal or hang-up, etc) the failed attempt is logged to the database and to local text log files, and another query is executed to get another callee. If there is an answer, the platform presents messages in Levantine Arabic to the callee describing the call collection, announcing the title of the day's topic, and asking the person to wait on hold until another callee is located to carry on a conversation. Up to twelve lines are dialing out simultaneously, and there are enough active subjects in the database to keep the dial-outs going on a continuous basis, so in theory there should generally be only a brief waiting period before two callees can be joined for a conversation. Upon being joined, both subjects are presented with a full description of the day's topic, and they are instructed to converse on this topic for 10 minutes. Recording begins immediately after the topic is announced, and continues until the end of the call, which may be the full 10 minutes, or sooner if both speakers hang-up (or are otherwise disconnected from the phone network).

Like Fisher English, subjects who receive calls from the platform are not required to identify themselves with their assigned Personal Identification Number (PIN). Anyone who answers the phone and understands enough about the collection project to carry on a successful conversation is accepted for participation. PIN validation is only required when subjects dial in to the platform; this is relatively frequent for collections involving subjects who live in the U.S., because they have direct access to a toll-free number for dialing in to the platform. However, only a few Fisher Arabic subjects reside in the U.S., so the number of dial-ins is extremely small.

Also like Fisher English, we have accepted multiple subjects who share the same telephone number. This, combined with the fact that PIN validation is not required when a callee answers a dial-out from the platform, creates a dual uncertainty regarding the identity of speakers in the recordings. On the one hand, when the platform dials out to a specific pin (assigned to a specific individual whose age, sex and other demographics are registered in the database), someone other than this registered individual may answer the phone and carry on with the recorded conversation; each PIN is allowed up to 3 calls, so different voices can be associated with a single PIN. On the other hand, when two or more people have registered using the same telephone, the platform may dial this number on successive occasions based on getting each of the various PIN's from the database, but the same person might answer each time, so the same voice can be associated with different PIN's in such cases.

This indeterminacy of voice identity for PIN's, while troublesome in many respects, is considered acceptable because the primary purpose of the data collection is to support speaker-independent automatic speech recognition research, where voice identity is a relatively minor concern, and because any steps to establish or assure correct speaker identity during the call would necessarily reduce the rate of successful calls (in fact, due the way most Arabic subjects have been recruited, they are typically never aware of the PIN that is assigned to them). Still, it is important to note that while this

indeterminacy is present in both Fisher English and Fisher Arabic, it is relatively more prevalent in Arabic: a higher proportion of registered subjects share phones, and there is a higher proportion of calls where a recorded speaker is obviously different from the person to whom the selected PIN was registered, even in cases where the given phone number was registered to only one PIN.

Because so many phone numbers are being used by multiple subjects, we have chosen to try limiting the scope of indeterminacy as follows: for each phone number that has been registered to multiple PIN's, we have deactivated all but one of the subjects involved, and have increased the number of calls that will be allowed to that one PIN, in proportion to the number of people who are likely to use that phone. This eliminates further occurrences of the same voice being recorded under different PIN's; it also provides a slight improvement to platform efficiency, because there is no longer any chance that we might dial two different PIN's at the same time that happen to use the same phone number.

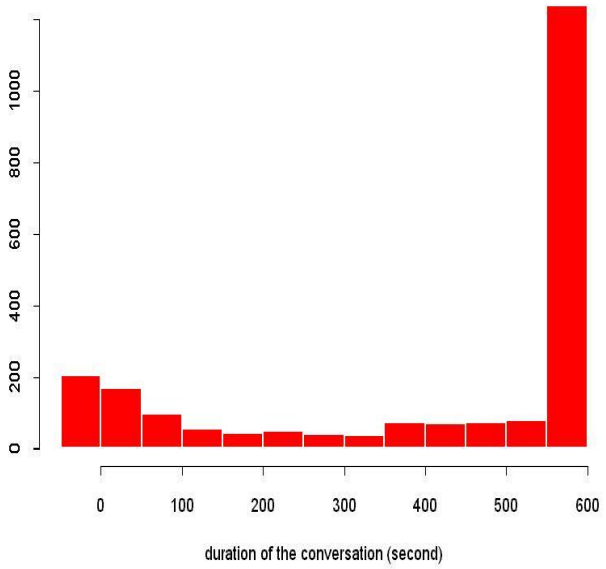
All calls are audited manually to label the gender of participants, and to provide a best guess as to their approximate ages (young, middle-aged or old) based on voice quality and any other clues available in the recording and table of subjects associated with the telephone number. The manual audits also label each speaker's regional dialect, and assign subjective assessments of "good, acceptable or poor" to both conversational involvement and overall signal quality. Given the indeterminacy of voice-ID based on PIN, the audit results are the only reliable basis for measuring the balance of the corpus in terms of gender, age and dialect.

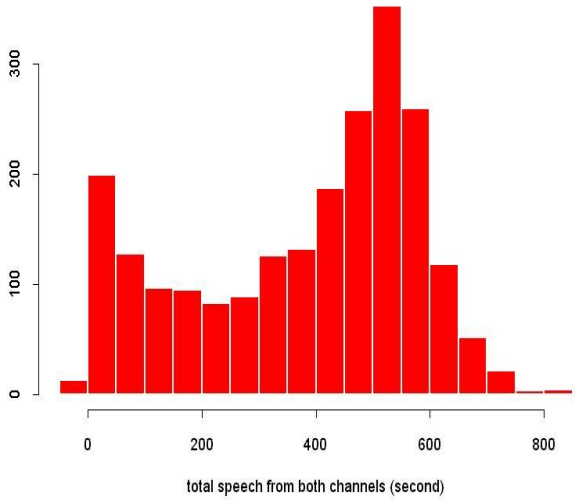
Another difference from Fisher English is the relative difficulty of making connections to participants and completing successful, full-length recordings. The nature of public telephone network (PTN) connectivity between the U.S. and the Middle East is such that, in general, a dial-out to an ME phone number is simply less likely to succeed. When dial-outs do get through, the connection is less stable and more prone to unexpected interruption. When the connection is sustained, there is still the problem of callees being unfamiliar with the goals or procedures for the project, or being unaccustomed to waiting on hold for anything more than a few seconds. All subjects who are currently active in the study have been recruited by third-party coordinators outside the LDC. While the arrangements with outside recruiters have yielded a very large subject pool fairly efficiently, this method precludes any direct communication between LDC staff and recruited subjects. This not only raises the risk that subjects won't know what to expect or what is expected of them, but also impedes feedback to the LDC regarding platform performance as perceived by the subjects. As a result, we have been limited in our ability to diagnose the relatively high rate of dial-out failures to the Middle East.

The following tables and figures provide details about the collection as of this writing (Oct. 26, 2004). About 2000 successful calls have been collected to date; calls are considered successful when the two-channel recording period lasts at least 5 minutes, and the total duration of utterances, as determined by automatic speech segmentation on both channels, is at least 3 minutes. At present, only about one fourth of these calls (fewer than 700) have received manual audits, but we will be increasing the pace of audits in order to eliminate the backlog of unaudited calls. (Also, as indicated by the varying number of unaudited sides in the tables below, the current audit interface needs to be improved to assure that all required audit decisions are entered for each call side.)

Signal Quality	**Call sides**
Poor	50
Acceptable	551
Good	726
Unaudited	**2651**
Gender	
Male	835
Female	538
Unaudited	**2605**
Age (estimated)	
Young	85
Middle	1175
Old	29
Unaudited	**2869**
Dialect	
Lev. (NC)	3
Lev. (LEB)	621
Lev. (PAL)	191
Lev. (JOR)	427
Lev. (SYR)	57
Egyptian	34
Gulf	2
Iraqi	23
Moroccan	2
Saudi	3
Yemeni	8
Other	2
Unaudited	**2604**

Figures below show the distributions of calls in terms of total speech (silence excluded) and overall call durations from automatically segmented utterances within each call. They confirm our decision about the 3 minute cutting point of acceptable calls.





5. References

Charles Ferguson, "Diglossia" in Word 15, pp.325-340, 1959.

Hymes, D. "Speech and Language: On the Origins and Foundations of Inequality among Speakers" in Deadalus, pp. 59-86, 1973.

Owen C. Rambow, "Arabic Dialect Modeling in Speech and Natural Language Processing", NSA Award Abstract #0329163, 2003.