Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Conventions

Mohamed Maamouri, Tim Buckwalter, Christopher Cieri
Linguistic Data Consortium
University of Pennsylvania
maamouri@ldc.upenn.edu, timbuck2@ldc.upenn.edu, ccieri@ldc.upenn.edu

Abstract

The present paper presents the experience gained at LDC in the collection and transcription of a corpus of conversational telephone speech in dialectal Arabic. The paper will cover the following: (a) Arabic language background; (b) objectives, principles, and methodological choices of dialectal Arabic transcription, (c) conceptualization and design features of LDC's 'Arabic Multi-Dialectal Transcription Tool' (AMADAT), and (d) a brief description of the conversational Levantine Arabic transcription guidelines and annotation conventions.

1.0 Introduction: Arabic Linguistic 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 issues 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 mainly represents a significant linguistic distance between all Arabic dialects and the 'fusha,' commonly identified as 'Modern Standard Arabic' or MSA, though the latter term does not cover all features of the former. This linguistic distance is characterized by substantial linguistic variation, mostly phonological, morphological, and lexical. The Arabic dialectal variation is significant not only between major dialects, for example, Egyptian, Levantine, Gulf, Maghreb, but also between the regional variants of a major dialect, for example, Northern and Southern Levantine.

Sound change has occurred in all Arabic dialects. In Levantine Arabic (LA), for instance, the sound /q/ is pronounced /q/ but also /'/, /g/ and /k/. The glottal stop is mostly deleted in medial and word final position with compensatory lengthening of the word internally (ra?s 'head' becomes ra:s and bi?r 'a well' becomes bi:r). Moreover, interesting cases of chain shifts with counter-feeding rule interactions also occur as in MSA fa?r 'mouse' goes to dialectal fa:r while MSA faqr 'poverty' goes to faqr but also to fa?r -- now meaning 'poverty' while it was 'mouse' earlier on. An important consequence of chain shifts is the multiplication of lexical ambiguity in the language.

The complexity of the above situation is compounded by the existence of significant differences between the sound changes of the various Arabic dialects. In Egyptian Arabic, for instance, MSA /θ/ becomes both /t/ and /s/ while /g/ is used to replace /j/ and /?/ to replace /q/. In Sudanese Arabic, MSA /q/ is replaced by /g/ and the uvular [ʏ]. All of the above creates an important amount of confusion which needs to be addressed and taken into account in any dialectal transcription task.

1.2 Pertinent Linguistic Features and the Dialectal Arabic Transcription Challenge

The description of Arabic dialect differences above, which does not even consider linguistic variation conditioned by age, gender, urbanity, rurality 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 without capturing the distinctions that matter to native speakers. In the case of a conversational Levantine Arabic corpus building, a Romanized orthography-based transcription can bypass the issues of phonemic sound shifts and the resulting variation by, for example, giving a faithful rendering of Levantine pronunciation characteristics. However, such a Romanized transcription would be machine readable and usable only for, and within the framework of, a single dialect system: LA. A Romanized transcription output will necessarily lead to the following tasks: (a) a long LA-related disambiguation process, (b) a comprehensive LA-specific lexicon and grammar, and (c) significantly longer annotators' training periods for better familiarization with transcription symbols.

Looking around us for examples of speech to text transcription practices which have been successfully used to support speech technologies (not just among linguists), one may ponder the wisdom of an orthography 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.

The above idea may seem too far-fetched but the Arabic language continuum is similar in many ways to the English one and presents the following potentially useful features: (a) there exists an important core of mutual intelligibility between MSA and the dialects, (b) there is a high level of similarity in morphological form and

syntactic structure similarity, and (c) there is also a significant common lexical core in spite of important semantic differentiation features. To help the above claim, one can assert the existence of an 'underlying' MSA cognate base with close structural similarities. The internalized knowledge of the above base by educated and even semi-literate Arabs and Arabic speakers is a potential that is available in the MSA writing and reading community in the Arab region and all over the world. Finally, there also exists a standard MSA graphemic knowledge base which could be put to good use to help with dialectal Arabic speech-to-text transcription. So, the question may very well be: How can we harness the native speaker's knowledge of Arabic orthographic conventions and of the linguistic MSA common core to complete a quick, easy, and low-cost Speech-to-Text transcription of Dialectal Arabic?

2.0 Principles, Objectives, and Methodology of Dialectal Arabic Transcription

2.1 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 systems 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.

The speech for this project comes from the Linguistic Data Consortium's Fisher Levantine Arabic project, in which more than 9400 speakers of the Northern, Southern and Bedwi dialects of Levantine Arabic (involving Jordan, Lebanon, and to a lesser extent Syria and the Palestinian territories) were recruited to participate in one to three telephone calls. Calls are up to ten minutes in duration and subjects speak to each other about assigned topics. A robot operator initiates most calls though subjects local to the robot operator may dial a toll-free number to initiate calls. Calls are recorded digitally from the native telephone network, and subjects are compensated for each successful call in which they participate. To date, LDC has collected 1670 calls from 1802 speakers.

Because our goal is to produce transcripts which, first and foremost, support the development of STT systems, we adopt the following principles for a transcription system in order of priority:
Friendly to writers and readers: easy to learn to write and read; lexically consistent: a given spoken form is always written the same way; lexically distinctive: different spoken forms will always be written differently; and acoustically consistent: transcription should represent pronunciation.

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

The advantages of an MSA-based strategy for dialectal Arabic transcription come from the fact that while writing Arabs benefit from their knowledge of stable MSA sounds and forms which keep to standard orthographic writing conventions. Arabs can read the same MSA words with the same or a closely similar level of recognition and comprehension. When faced with the task of writing down a dialectal Arabic speech form, native Arabs use their knowledge of the 'underlying' sounds of the Arabic word in order to transcribe its MSA-reconstructed form with Arabic script letters.

Native Arab transcribers' knowledge of the Arabic language and their familiarity with the rules of Arabic script constitute the basis of a strategy for the transcription of Arabic dialects. This strategy uses practical MSA-based orthographic conventions and a reasonable reliance on MSA to produce an acceptable output and guarantee a high rate of consistency and an easy retrieval of meaning structures. A significant advantage of this strategy is that native transcribers do not need to go through long training periods to learn difficult and often complex symbols.

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

An MSA-based Arabic orthographic script transcription faces three major challenges. The first is that there is little or no evidence of a dialectal Arabic text corpus with stable MSA-based writing conventions. In a concordance generated from a corpus of newswire written primarily in MSA, Levantine dialectal forms were found which attest written Arabic colloquial communication. However, the resulting concordance clearly shows that the result of this practice is a mixture of MSA and LA, with a significant use of foreign loan-words. 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 and 'filters' which seek to elevate the level of the dialectal forms toward MSA written standards. This mixture has led and usually leads to inconsistent transcriptions, characterized by two opposing tendencies, namely: (a) a register remaining at the level of LA and (b) another register rising from LA to MSA.

Low-literate Arabs use the Arabic script if and when they have to write anything down. Their practices constitute an idiosyncratic 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 the way they say it without worrying about or being aware of the relationship of the written forms to an underlying MSA structure. On the other hand, educated Arabs tend to obey MSA (over)correctness filters when they write/transcribe LA /kama:n/ ("also") as MSA /?ayDan/ and LA /?ana šuft/ ("I saw") as MSA /?ana ra?ayt/ -- in spite of the fact that they are totally different words.

So we find ourselves confronting two pitfalls: (a) the real danger of the interference of MSA writing conventions and MSA dominance in the budding dialectal Arabic transcription practices, and (b) the danger of inconsistencies, thus the lack of stability, in the emerging

dialect transcription practices and conventions using Arabic orthography.

2.2 Principles of an MSA-based Dialectal Arabic Transcription

The developments described above argue for a transcription of Arabic dialects which pragmatically uses the Arabic orthography, both symbols and rules. That is, Arabic colloquial transcriptions should be written in Arabic script, without short vowels and with no other diacritical marks except for *nunation*, and otherwise following the general orthographic conventions of MSA.

In order to increase the rate of word stability, this system must reach a consistent and steady balance between the two poles of currently observed transcription practices and tendencies. That is to say:
(1) neither too strict an adherence to MSA-based spelling conventions (via the use of linguistic filters) that would shoe-horn LA utterances unnecessarily into MSA form (2) nor too close an adherence to the phonetic reality of the dialect that would lead to a finer acoustical representation but with a lower rate of semantic word recognition.

The transcription of Arabic dialects is a difficult act to balance, and the speech technology community seems divided along the same lines with two equally important goals and tradeoffs: (1) to produce a finer phonetic representation in order to accommodate acoustic modeling or (2) to produce transcripts with maximal similarity to MSA in order to accommodate language modeling. In the final run, transcribers and their research managers have to decide on the amount of MSA-underlying forms they would like to see in the overall corpus/transcripts.

2.3 Methodological Choices of 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 speech recognition and technologies that rely on ASR output, however, and some researchers have already started to look at how to use MSA in the processing and analysis of Arabic dialects.

Owen Rambow (2003) uses MSA text to model dialectal Arabic. Rambow 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 and in fact, necessary to building successful dialectal language models. The transcription guidelines we developed for conversational Levantine Arabic use MSA underlying forms, whenever possible and only when linguistically appropriate, to render an approximation of MSA text, especially in orthography and basic morphology. Our transcription guidelines aim at retaining much of the specificities of the dialectal forms even when doing so means some loss of graphemic closeness to MSA (see http://www.ldc.upenn.edu/Projects/EARS/Arabic for a more elaborate explanation).

We chose the above methodological approach because we believe that annotators can easily transfer their MSA-based literacy skills to the transcription task, instead of having to learn a demanding Latin orthography-based phonetic-phonemic transcription. Our assumption is that using dialectal text transcripts that are close enough to MSA will help introduce the adjustments necessary in order to address the specificities of the dialect to parsing and tagging tools that already exist and have been successfully used with MSA texts. This approach will also allow the resulting text to be morpho-analyzed and syntactically annotated using existing tools and fed to downstream processes based upon more common MSA texts.

Because it was deemed extremely important to develop a rapid transcription in the shortest time possible leading to a 'good enough' rendering of a dialectal Arabic text, there was a preference to use MSA underlying forms to render a transcription with a satisfactory approximation of MSA text whenever possible, especially in orthography conventions and basic morphology. The graphemic closeness to MSA is very important to the transcription task because: (a) we believe that annotators can easily transfer their MSA-based literacy skills to the transcription task, which makes it a relatively easier task, and (b) without the closeness to MSA and without the Arabic script conventional practices annotators bring to bear in their transcription task, we would have had to use a Roman-based 'phonetic-phonemic' transcription, which would have been very costly in training time and quite unsatisfactory in level of linguistic form recognition.

The general methodological principles that we followed and the transcription guidelines that we developed for our conversational Levantine Arabic transcription for instance, focus on keeping all or as many of the specificities of the dialectal forms as possible in one or the other of two layers of transcription, which are as follows: (1) a layer which focuses on anchoring transcribed dialectal forms to MSA graphemically similar utterances whenever possible thus establishing a kind of 'underlying' MSA semantic structure base which helps with word recognition and identification and (2) a second layer which uses the output of the first layer and focuses on enriching that transcription within principles of closer adherence to the dialectal specificities of the speech forms under consideration .

The first layer serves as a 'Quick Transcription' (QT) and establishes a 'good enough' dialectal text capable of robust shallow semantic analysis. The second layer adds most functionally necessary vowels, marks important sociolinguistic variants, morphophonemic features (such as assimilation and 'sandhi' phenomena) and other major sound change phenomena. In fact, this second layer is in reality a 'Careful Transcription' (CT) and the annotation it presents also serves other specific linguistic research purposes such as speech recognition for instance, or sociolinguistic analysis.

3.0 Conceptualization and Design Features of LDC's 'Arabic Multi-Dialectal Transcription Tool' (AMADAT)

3.1 The Arabic Multi-Dialectal Transcription Tool Features

The textual representation of LA conversations was done at LDC with the use of the Arabic Multi-Dialectal Transcription tool AMADAT version 1.2 , which is a tool designed and developed at LDC in 2003. AMADAT spans the diglossic gap and linguistic distance which exist between any and all Arabic dialects and Modern Standard Arabic (MSA). It is designed to provide a multi-layered transcription by extending links between the two sets of linguistic structures and connecting transcribed dialectal forms to their underlying MSA-based forms, whenever possible. AMADAT is designed to keep a close annotation of the complicated and idiosyncratic linguistic variation features occurring between individual Arabic speakers in multi-dialectal communication.

AMADAT uses a two-tiered transcription, which provides a Modern Standard Arabic-based transcription (MSAT) of speech data in a first pass (orthographic level) followed by a second annotation pass (surface phonemic level), which uses an Arabic Orthographic System-based Transliteration (AOST) and provides pertinent phonemic and missing pronunciation features of the target dialect (such as distinctive dialect short vowels, consonantal sociolinguistic variation, *shaddah*, etc.). AMADAT has three mutually-exclusive operation modes: (a) GREEN PASS for the MSAT, which uses an Arabic keyboard; (b) YELLOW PASS for the AOST, which uses a Latin keyboard, and (c) RED PASS, which uses a Latin keyboard and serves for editing and correction of MSAT or/and AOST errors and typos. It is probable that the acronyms MSAT and AOST are somewhat confusing.

3.2 AMADAT Tool-based Tags and Metalanguage Annotation

The AMADAT transcription tool includes a set of buttons which are used to annotate the various metalinguistic features of the targeted speech. This metalinguistic annotation includes the following: (1) metalinguistic tags. which include non-speech sounds in the recording; (2) interjections, which are speech sounds (non-lexemes) communicating hesitation, surprise, agreement, etc.; (3) linguistic and sociolinguistic phenomena describing language change, variation, deletion, and other linguistic processes; (4) the dialectal identity of the speaker. Finally, 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)
(Velarized Cons)
(Voc Variant)
(Hamzah Drop)
(Diphthong)
(-h Deletion)
(Cons Deletion)

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 Transrciption: Partial words and restart guidelines

When a speaker breaks off in the middle of the word, annotators should transcribe as much of the word as can be made out. 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.

4.0 Conversational Levantine Arabic Transcription Guidelines and Annotation Conventions

4.1 Transcription of Levantine Arabic

In order to develop orthographic convention guidelines for transcribing Levantine Arabic (LA) conversational speech, we first engaged in a survey of authentic samples written in Arabic script and published on the Web. In order to locate these samples we used the Google search tool and we looked for relatively unambiguous high-frequency LA words, such as شو /šu:/ "what," ليش /le:š/ "why," بدك /biddak/ "you want,", and هون /ho:n/ "here," as well as high-frequency phrases such as يا زلمه /ya: zalameh/ "hey dude", and شو هالحكي / šu: hal-Haki:/ "say

what?" (See Table1). The samples typically came from data posted recently on the Web, where the participants linguistic identity was not doubtful. This allowed us to observe current trends in the orthography of Levantine and other major dialects. We observed fairly stable spelling trends and some variation as well. In cases where orthographic variation occurred, we checked the frequencies provided by Google for the two or more variant spellings, and we arbitrarily chose the most frequent. A good example of this approach was our choice to write the 1st person singular conjugation of the imperfect verb *without* the *alif* prefix (e.g., انا بحكي) instead of with it (e.g., انا باحكي), simply because the Google frequency of the former outnumbered that of the later by about 8 to 1.

إحنا /?iHna/ "we"; cf. نحنا /niHna/
اللي /?il:i:/ "who; which" (with prepositions: باللي /bil:i:/, للي /lil:i:/)
إمبيرح /?imbe:riH/ "yesterday"
إمراة /?imra:/ "woman; wife" (إمراتك /?imra:tak/ "your wife") see also مرة /mara/
إنتي /?inti:/ "you" (fem.sg.)
إنتوا /?intu:/ "you" (masc.pl.)
أنو /?anu:/ "which" من أنو جامعة؟ "from which university?"
إيد /?i:d/ "hand"
إيش /?e:š/ "what"; cf. ليش /le:š/ "why"
إيمتى /?e:mta/ "when"
أيوه /?ay:uwah/, also /?ay:uwa/ "yes"
بد /bid:-/ "want" (بدي /bid:i/ "I want", بده /bid:o/ "he wants", بدها /bid:a:, bidha/ "she wants")
بس /bas:/ "only; just"
بعدين /ba`de:n/ "afterwards"
بكرة /bukra/ "tomorrow"
بلكي /balki/ "maybe"
بـ /bi:-/ "with" (with following clitic: بيك /bi:k/ "to/with you (masc.sg.)", بيكي /bi:ki:/ "to/with you (fem.sg.)"): بيكوا أهلاً /?ahlan bi:ku/ "welcome!"
بيناتـ /be:na:t-/ "among; between" (with following clitic: بيناتهم /be:na:thum/ "among them"

Table 1: Levantine Arabic High-Frequency Words (Partial Listing)

The LDC *Guidelines For Transcribing Levantine Arabic*[1] include a list of all attested high-frequency Levantine Arabic words, as well as tables with verb conjugation paradigms, and provide special notes on difficult areas, such as how to spell the numbers 11-19, pronominal suffixes and verbal objects, the days of the week, etc. Transcribers are instructed to use the writing conventions of unvocalized MSA spelling and word segmentation in all cases. The following example illustrates the application of this principle: the Levantine utterance /?ultil:ak/ ("I told you") is to be transcribed as unvocalized MSA segmented as two words — قلت لك (not قلتلك or قلتللك) and using the accepted MSA orthography قلت for the colloquial pronunciation /?ult/. However, the transcriber is also alerted to the fact that there are three notable exceptions to the above spelling and word segmentation rule:

- If the word is listed among the high-frequency colloquial words then it should be spelled as indicated in the list and no attempt should be made to render it in MSA.
- If the word is a colloquial verb whose morphology deviates substantially from that of its MSA equivalent, then it should be written as indicated in the conjugation paradigms of colloquial verbs.
- *Nunation* (*-an -in -un*) will be transcribed when it is recorded in speech

The *Guidelines* include a list of regular phonological differences between MSA and Levantine Arabic (and many other dialects as well) that do *not* justify departing from MSA orthography when transcribing the colloquial form:

- MSA interdental fricative /θ/ → LA dental stop /t/. Examples: مثل (MSA /miθla/, LA /mitl/), أكثر (MSA /?akθar/, LA /?aktar/)
- MSA interdental fricative /θ/ → LA sibilant /s/. Examples: مثلا (MSA /maθalan/, LA /masalan/)
- MSA velar /q/ → LA glottal stop /?/ or velar stop /k/. Examples: قصة (MSA /qiS:a/, LA /?iS:a/ or /kiS:a/)
- MSA velar stop /k/ → LA palato-alveolar affricate /č/. Examples: كلب (MSA /kalb/, LA (in some village dialects) /čalb/)
- MSA interdental fricative /ð/ → LA dental stop /d/. Examples: تأخذي (MSA /ta?xuði/, LA تاخذي /ta:xudi/), خذ (MSA /xuð/, LA /xud/), هذا (MSA /ha:ða/, LA /ha:da/)
- MSA interdental fricative /ð/ → LA voiced alveolar fricative /z/. Examples: كذاب (MSA /kað:a:b/, LA /kaz:a:b/)
- MSA velarized voiced alveolar stop /D/ → LA velarized voiced alveolar fricative /Z/. Examples: مضبوط (MSA /maDbu:T/, LA /maZbu:T/), بالضبط (MSA /bi-D-DabT/, LA /bi-Z-ZabT/)
- MSA voiced palatal affricate /ǧ/ → LA voiced palatal spirant /ž/. Examples: مجنون (MSA /mağnu:n/, LA /mažnu:n/)

The *Guidelines* give special consideration to the problematic area of *hamza* transcription. After considerable debate it was determined that all glottal stops should be written when and where they occur. This directive requires some elaboration in two different areas of *hamza* orthography:

[1] The Guidelines are updated regularly and are available from the LDC website <http://www.ldc.upenn.edu/Projects/EARS/Arabic/Guidelines_Levantine_MSA.htm >

(1) The writing of stem-initial *hamza* in MSA assists in lexical disambiguation (e.g., أن / إن and باسم / بإسم), and reduces the phonetic "load" already carried by the bare *alif* character. Transcribers who are told to write stem-initial *hamza* only when they hear it may still write it (out of habit) when it is not uttered, or omit it in places where it *is* uttered. Regardless of what transcribers are able to do, the goal is to normalize the orthography of stem-initial *hamza* to full and consistent transcription of *hamza*.

(2) The writing of stem-medial/final *hamza* is lexically determined and not optional in MSA orthography. However, a few MSA words with stem-medial/final *hamza* do show variation, with the glottal stop realized also as vocalic length, but the phenomenon in not widespread as it is among the dialects. When the transcriber attempts to render in MSA orthography colloquial words whose vocalic length could be attributed to underlying glottal stop, there is danger of *reverse-engineering* of ghost MSA words. An example of this would be to render the colloquial utterance /?ana ra:yHa/ ("I am going") as pseudo-MSA أنا رائحة /?ana ra:?iHa/ (rather than the more realistic أنا رايحة). Transcribers are required to write stem-medial/final *hamza* only when they hear it. If both forms occur, one with and one without *hamza*, and this is probable though not frequent, the mapping of colloquial /ğara:yid/ to MSA /ğara:?id/ "newspaper") will provide the needed answer to this variation issue. However, we need to prevent the mapping of dialectal /fa:yiz/ (a proper name "Fayez") to MSA /fa:?iz ("winner").

Because our transcription of Levantine Arabic makes no use of short vowels and diacritics (except for *nunation* when it is pronounced, as mentioned earlier), certain orthographic conventions are used for eliminating ambiguity, such as the spelling of the 2nd pers. fem. sg. direct object clitic /-ki/ as كي (e.g., /šufna:ki/ شفناكي "we saw you"). This applies as well to the pronoun suffix /-ki/: (e.g., /ma`a:ki ?inti/ معاكي إنتي "with you." In a similar fashion, the 2nd pers. pl. direct object clitic /-ku/ is written كوا (e.g., /šufna:ku/ شفناكوا "we saw you"), and this applies as well to the pronoun suffix /-ku/ (e.g., /Ha:ritku ?intu/ حارتكوا إنتوا "your neighborhood." These spelling conventions are well attested in data on the Web.

Foreign words and place names in Levantine Arabic are spelled according to the conventions of MSA (e.g., Washington واشنطون, Los Angeles لوس انجلوس). In cases where there is regional variation in MSA spelling, the transcriber is instructed to follow Levantine spelling conventions, as opposed to Egyptian spelling habits, as in the following examples: "garage" (Egyptian: جراج ; Levantine: كراج), "congress" (Egyptian: كونجرس; Levantine: كونغرس). Words that are not attested in MSA are transcribed as expected in MSA, but according to Levantine orthography. Note that although many computers are able to display "extended" Arabic characters, such as the Persian letters /p/ پ, /č/ چ, /ž/ ژ, and /g/ گ, few systems provide the user with an easy way to actually type these characters on the keyboard. So, although these letters are potentially available for representing foreign sounds, the convention in MSA orthography is to substitute the corresponding and easily-available Arabic letters instead. Therefore, according to Levantine practice, for /p/ we use ب (e.g., "Pam" بام), for /č/ we use تش (e.g., "Chet" تشيت), for /ž/ we use ج (e.g., "Mirage" ميراج), and for /g/ we use غ (e.g., "Gilbert" غيلبرت).

4.2 Transcription Challenges

Successful transcription of Levantine Arabic requires knowing when to apply the rules of MSA orthography and when to depart from those rules, specifically by following the list of high frequency colloquial words, or other special tables and lists, such as the verb conjugation paradigms, the numeral 11-19, days of the week, etc. Among the pitfalls we have witnessed is the occasional over-zealous attempt to follow MSA rules, to the extent that the transcriber is actually "translating" Levantine Arabic into its MSA equivalent. An example of this error would be to transcribe LA /nuS:/ ("half") as MSA /niSf/. For the this example, we note that there is no productive phonological rule in LA that calls for deletion of word final /f/, hence the word /nuS:/ is patently colloquial and should not be rendered in its MSA equivalent.

Among the issues of MSA orthography implementation, two interesting examples emerge as they represent a higher level of transcription difficulty for most annotators. The first is the use of the *ta marbuta*, which is rarely pronounced in Arabic dialectal speech since that /-t/ is really morphophonemic and therefore only present in juncture-related phenomena, and the need to always write it with the two dots (ـة) (whether it is pronounced /a/ or /t/) in order to distinguish it from the possessive pronoun (ـه).

The second relates to the acoustic ambiguity relating to the speech occurrence /fiy/, which as a preposition means 'in' but as a prepositional phrase or equational sentence means 'in it' or 'there is in it.' Since all three meanings are represented by the same acoustical reality, transcribers have struggled with this distinction. Since syntactic rules exist and help with the issue, the following 'Negation test' was devised to train transcribers to make the distinction and disambiguate between *fiy* and *fiyh*: (a) if you need *muw$*, *mahuw$*, etc. to negate the targeted form, which is pronounced [fiy], the POS value of [fiy] is then PREP or a PREP PHRASE. In all such cases, [fiy] needs to be transcribed في ; (b) if you can append the negative suffix *-i(y)$* directly on the targeted form [fiy] then transcribe [fiy] فيه in all those cases.

5.0 References

Ferguson, Charles (1959). Diglossia. Word, 15, 325—340.

Hymes, D. (1973). Speech and Language: On the Origins and Foundations of Inequality among Speakers. Deadalus, 59—86.

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