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

Arabic Language Resources and Tools Conference, NEMLAR
Cairo, Egypt September 22-23, 2004
PRESENTATION OUTLINE

- ARABIC LINGUISTIC BACKGROUND
- ARABIC DIALECTAL SPEECH: METHODOLOGICAL TRANSCRIPTION PRINCIPLES AND TECHNOLOGICAL GOALS OF THE PROJECT
- AMADAT: LDC’S ARABIC MULTI-DIALECTAL TRANSCRIPTION TOOL
- METALANGUAGE: RT-04 ARABIC TELEPHONE SPEECH TRANSCRIPTION CONVENTIONS
- BRIEF OVERVIEW OF LEVANTINE ARABIC TRANSCRIPTION GUIDELINES

ARABIC LINGUISTIC BACKGROUND

- “ARABIC LANGUAGE CONTINUUM” WITH ARABIC DIGLOSSIA
  FUSHA = Modern Standard Arabic (=MSA) + ARABIC DIALECTS
  + INTRALINGUAL CODESWITCHING & CODE-MIXING

- SIGNIFICANT LINGUISTIC DISTANCE BETWEEN MSA & DIALECTS

- SIGNIFICANT INTER- LINGUISTIC VARIATION AMONG DIALECTS

- SIGNIFICANT INTRA- LINGUISTIC VARIATION WITHIN DIALECTS

- IMPORTANT COMMON CORE OF MUTUAL INTELLIGIBILITY

  ➔ HIGH LEVEL OF FORM AND STRUCTURE SIMILARITY
  ➔ COMMON LEXICAL CORE WITH SIGNIFICANT SEMANTIC DIFFERENTIATION
ARABIC LANGUAGE BACKGROUND

- EXISTENCE OF LIVING MSA WRITING AND READING COMMUNITY
- INTERNALIZED KNOWLEDGE OF MSA BY EDUCATED AND SEMI-LITERATE NATIVE ARABIC SPEAKERS
- EXISTENCE OF UNDERLYING MSA COGNATE STRUCTURES
- USE OF MSA-BASED “ACCOMMODATION FILTERS”
- DOMINANCE OF MSA-BASED GRAPHEMIC TRADITIONS AND EVIDENCE OF MSA-BASED GRAPHEMIC INTERFERENCE
- EXISTENCE OF STANDARD MSA-BASED GRAPHEMIC KNOWLEDGE

→ PRODUCTIVE BASE FOR CONVERSATIONAL DIALECTAL ARABIC SPEECH-TO-TEXT TRANSCRIPTION SKILLS
DIALECTAL ARABIC SOUND CHANGE

DIALECTAL SOUND CHANGE PATTERNS

/θ/ ➞ /t/  /δ/ ➞ /d/  /δ/ ➞ /d/
/s/ ➞ /z/  /z/ ➞ /z/

/ʔ/ ➞ /g/  /q/ ➞ /q/  /k/ ➞ /y/
ARABIC DIALECTAL VARIATION

In Egyptian Arabic, MSA /θ/ becomes both /t/ and /s/ while /g/ is used to replace / j / and /ʔ/ to replace /q/. In Sudanese Arabic, MSA /q/ is pronounced /g/ and [γ] while the same phoneme/letter is pronounced /q/, /g/, /ʔ/, and /k/ in Levantine Arabic.

Example: Iraqi.q.h.C.wav

EXISTENCE AND USE OF ARABIC SCRIPT “ARCHIGRAPHEMES”
Q: 什و القصة؟
$w AlqSp?
"What's the story?"

A/T1: كلتلك مو كوف مش معتقل وما في قصة
ya zlmy kltlk mwkwf m$ mEtkl wmA fy kS~p

A/T2: كلتلك موقوف مش معتقل وما فيه قصة
ya zlmy qltlk mwqwf m$ mEtql wmA fy qS~p
"Hey ‘dude’ I told you arrested not indicted and there is no story"
Need to distinguish the transcription approach from the alphabet used.

- Transcription approaches: phonic, orthographic, hybrid
- Alphabets: Arabic, Roman, International Phonetic Alphabet
- One may perform either phonic or orthographic transcription using either Roman or Arabic alphabets

Problems with standard approaches

- Alphabets
  - IPA is hard to learn
  - Roman script looks and feels unnatural to Arabic speakers
  - Few computer systems fully implement Arabic script and bi-directional input.

- Transcription Approaches
  - MSA lacks conventions for many Levantine forms, does fully not address needs of acoustic modeling
  - purely phonic approach hinders language modeling
Speech Recognition

- Original Speech
- Analysis of audio
- Analysis suggests multiple phonetic interpretations.
- Which need to be mapped onto a surface representation
- Sequences of which are compared against existing text to determine probable accuracy. Off-domain written text often substitutes for rare on-domain transcripts of spoken language.
How can we harness the native speaker’s knowledge of Arabic orthography conventions and of the MSA linguistic common core to complete a quick, easy, and low-cost Speech-to-Text transcription of Conversational Dialectal Arabic?

OBJECTIVES OF SPEECH-TO-TEXT TRANSCRIPTION

- FRIENDLY TO WRITERS AND READERS: EASY TO LEARN TO WRITE AND READ
- LEXICALLY CONSISTENT: A GIVEN UTTERANCE WILL ALWAYS BE SPELLED THE SAME
- LEXICALLY DISTINCTIVE: DIFFERENT UTTERANCES WILL ALWAYS BE SPELLED DIFFERENTLY
- ACOUSTICALLY CONSISTENT: TRANSCRIPTION/SPELLING PREDICTS PRONUNCIATION
CONVERSATIONAL DIALECTAL ARABIC TRANSCRIPTION CHALLENGES

MSA-BASED/ARABIC ORTHOGRAPHIC SCRIPT-BASED TRANSCRIPTION

3 MAJOR CHALLENGES

- RARE EVIDENCE OF CONVERSATIONAL DIALECTAL ARABIC TEXT CORPUS WITH STABLE MSA-BASED WRITING CONVENTIONS (POETRY, DRAMA, EPISTOLARY, POLITICAL SPEECHES, WEB & INTERNET CHATROOMS)

- DANGER OF INCONSISTENT CONVERSATIONAL DIALECTAL ARABIC MSA-BASED TRANSCRIPTION PRACTICES

- NATIVE LANGUAGE REPRESENTATION: DANGER OF OVER INTERFERENCE OF MSA WRITING CONVENTIONS IN EXISTING CONVERSATIONAL DIALECTAL ARABIC TRANSCRIPTION PRACTICES
CONVERSATIONAL DIALECTAL ARABIC STT TRANSCRIPTION OBJECTIVE

OBJECTIVE: APPROPRIATE BALANCE BETWEEN THE TWO TENDENCIES BELOW IN ORDER TO AVOID NEGATIVE CONSEQUENCES TO THE SPECIFIC NEEDS OF THE STT SCIENTIFIC RESEARCH COMMUNITY

- Neither too strict an adherence to the use of MSA-based spelling conventions to reconvert dialectal forms to an unnecessary MSA-representation \(\rightarrow\) WITH HIGHER RECONSTRUCTION RATE OF ‘UNDERLYING’ FORMS

- Nor too close an adherence to finer sound /(allo)phonic/ acoustical utterance representation \(\rightarrow\) LEADING TO AN OUTPUT WITH FINER ACOUSTICAL REPRESENTATION BUT WITH LOWER RATE OF SEMANTIC WORD RECOGNITION
“AMADAT” DESIGN SPECIFICATIONS

- ARABIC MULTI-DIALECTAL TRANSCRIPTION AND ANNOTATION TOOL

- TWO TIERS OF TRANSCRIPTION / ANNOTATION

- MODERN STANDARD ARABIC-BASED TRANSCRIPTION (MSAT: ‘ORTHOGRAPHIC LEVEL’)

- ARABIC ORTHOGRAPHIC SYSTEM-BASED TRANSLITERATION (AOST: ‘SURFACE PHONEMIC LEVEL’)

- THREE MUTUALLY EXCLUSIVE OPERATION MODES
‘AMADAT’ STT TRANSCRIPTION MODES

MSAT MODE: QUICK TRANSCRIPTION → ‘GREEN AREA’
- USE OF NORMAL ARABIC KEYBOARD FOR TRANSCRIPTION
- FIRST PASS WITH MSA-BASED APPLICABLE CONVENTIONS
- METALANGUAGE ANNOTATION (CTS RT-04 ANNOTATION)
  OBJECTIVE: OPTIMIZED OUTPUT FOR LANGUAGE MODELING

AOST MODE: CAREFUL TRANSCRIPTION → ‘YELLOW AREA’
- USE OF LATIN KEYBOARD FOR TRANSLITERATION
- USE OF MODIFIED TIM BUCKWALTER CODE WITH SOUND VALUES
- OBJECTIVE: OPTIMIZED OUTPUT FOR ACOUSTIC MODELING

EDIT MODE: ANNOTATION CORRECTION → ‘RED AREA’
- USE OF LATIN KEYBOARD FOR A TOKEN-BY-TOKEN EDITING
- ACCESS ONLY TO ANNOTATION MANAGEMENT AND QUALITY CONTROL
<table>
<thead>
<tr>
<th>Begin</th>
<th>End</th>
<th>Track</th>
<th>Transcription in Arabic</th>
<th>Index</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>147.32</td>
<td>148.37</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>148.85</td>
<td>150.01</td>
<td>B</td>
<td></td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>149.80</td>
<td>151.01</td>
<td>A</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>150.43</td>
<td>152.40</td>
<td>B</td>
<td>أما صحيح اللي بأحكية و إلا لا (% أه)</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>152.04</td>
<td>153.42</td>
<td>A</td>
<td></td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>152.96</td>
<td>154.18</td>
<td>B</td>
<td></td>
<td>103</td>
<td></td>
</tr>
</tbody>
</table>

MSA Transcription:

أنا صحيح اللي بأحكية و إلا لا (% أه)
‘MSAT’ SPECIFICATIONS AND ISSUES

- MACHINE-READABLE UNVOCALIZED WRITTEN TEXT DATA
- NO DIACRITICS IN GENERAL. HOWEVER, USE OF SHADDAH AND INITIAL HAMZA NEED TO BE RE-DISCUSSED BY THE SCIENTIFIC COMMUNITY’ USERS
- FOCUS ON CONSISTENT TRANSCRIPTION OF SAME FORMS
- FOCUS ON IDENTIFICATION OF SPECIFIC DIALECTAL FORMS (DEFINITIONAL NEEDS TO BE DISCUSSED)
- ANCHORING OF SOME DIALECTAL FORMS TO MSA-SIMILAR UTTERANCES AND AN ‘UNDERLYING’ MSA SEMANTIC STRUCTURE (DEFINITIONAL NEEDS TO BE DISCUSSED)
- CAUTIOUS/CONSERVATIVE USE OF RECONSTRUCTED ‘UNDERLYING’ FORMS: “NO REVERSE MSA ENGINEERING”
<table>
<thead>
<tr>
<th>Begin</th>
<th>End</th>
<th>Track</th>
<th>Transcription in Arabic</th>
<th>Index</th>
<th>Tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>147.62</td>
<td>148.85</td>
<td>A</td>
<td>ﺗَأْنَي (Ala%)</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>148.85</td>
<td>150.01</td>
<td>B</td>
<td>ﺗَأْنَي (Ala%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>149.20</td>
<td>151.01</td>
<td>A</td>
<td>ﺗَأْنَي (Ala%)</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Speech Comment:
- F1: (breath)
- F2: (cough)
- F3: (laugh)
- F4: (music)
- F5: (noise)
- F6: (people talk)
- F7: (sneeze)
- F8: (silence)
- F9: (pause)
- F10: (%ah)
- F11: (%eh)
- F12: (%um)
- F13: (%ooh)
- F14: (%hm)
- F15: (noise/)
- F16: (overlap)
- F17: (overlap/)

Vowelized Arabic Trans:

Linguistic Transliteration:

>am~aA SaHiyH Al~iIy baHkiiy~h wi <il~aA Ia> (%>h)

Selected Word:
>am~aA
‘AOST’ SPECIFICATIONS AND ISSUES

- **FOCUS ON** CLOSE ADHERENCE TO SOUND SPECIFICITIES
- **FOCUS ON** FULL FUNCTIONAL VOCALIZATION WITH SUKUN LIMITED TO SYLLABIC DIVISION WHEN NEEDED FOR PRONUNCIATION
- **NO REPRESENTATION OF** VOCALIC QUALITY VARIATION BUT LENGTHENING OF UNDERLYING DIPTHONGS
- **INCLUSION OF** RELEVANT SOUND FEATURES EXCEPT MORPHOPHONEMIC ASSIMILATION PHENOMENA (EXAMPLE: AL-), AND EPENTHETIC AND JUNCTURE PHENOMENA
- **USE OF** PERSIAN LETTERS FOR CAREFUL TRANSCRIPTION OF UTTERANCES IN WHICH SOUNDS WHICH DO NOT EXIST IN THE ARABIC ORTHOGRAPHY OCCUR
- **WHILE RECORDING AND ANNOTATING** DIALECTAL SOUND FEATURES IN AOST, THE LINKED MSAT TOKENS AND QUICK TRANSCRIPTION BASELINE REMAIN UNCHANGED/STABLE
<table>
<thead>
<tr>
<th>Begin</th>
<th>End</th>
<th>Track</th>
<th>Transcription in Arabic</th>
<th>Index</th>
<th>Tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>147.32</td>
<td>148.85</td>
<td>A</td>
<td>(لا)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>148.85</td>
<td>150.01</td>
<td>B</td>
<td>(لا) بحكيه و إلا لا (؟أه)</td>
<td>99%</td>
<td></td>
</tr>
<tr>
<td>150.01</td>
<td></td>
<td>A</td>
<td></td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Speech Comment:
- F1: (breath)
- F2: (cough)
- F3: (laugh)
- F4: (music)
- F5: (noise)
- F6: (peopletalk)
- F7: (sneeze)
- F8: (silence)
- F9: (pause)
- F10: (؟ah)
- F11: (؟eh)
- F12: (؟um)
- F13: (؟ooh)
- F14: (؟hm)
- F15: (noise/)
- F16: (overlap)
- F17: (overlap/)

Vowed Arabic Trans:
>امَا صَحِيحُ اللَّهُ بِحْكِيِّهِ وَ إِلَّا لَا (؟أه)

Linguistic Transliteration:
>am~aA SaHiyH Al~iIy baHkiy~h wi <il~aA la> (>%h)
RT-04 CONVERSATIONAL ARABIC TRANSCRIPTION CONVENTIONS

DISFLUENT SPEECH
- FILLED PAUSES AND HESITATION SOUNDS
- PARTIAL WORDS AND RESTARTS
- CONTRACTED WORDS
- MISPRONOUNCED WORDS
- HARD-TO-UNDERSTAND SECTIONS
- BACKGROUND NOISES
- SPEAKER-PRODUCED NOISES

LINGUISTIC MARKUP
- LINGUISTIC CHANGE FEATURES
- SOCIO-LINGUISTIC VARIATION FEATURES
- FOREIGN WORDS
“whenever possible, follow the spelling conventions and word segmentation of MSA.”  Like this:

"قلت لك"  /?ultil:ak/

"مضبوط"  /mazbu:T/

"مثل"  /mitl/

"مثال"  /masalan/
MSA-based orthography

“whenever possible, follow the spelling conventions and word segmentation of MSA.”  

Avoid this:

أَلْتَلْك  /?ultil:ak/

مَزْبوُط  /mazbu:T/

مِتْل  /mitl/

مِسْلَان  /masalan/
Exceptions

“Note, however, the following exceptions…”

1 list of high-frequency colloquial words
2 conjugation paradigms of colloquial verbs
3 nunation (-an -in -un) is transcribed if heard
MSA-based orthography

Exception 1

High-Frequency Colloquial Words (c. 120)

عثمان
علم
عم
فأ
فيه
فيش
فين
زلما
زي
شو
شوي
شوية
بعدين
بكرة
بلكي
بينات
جوا
دغري
إيد
أيش
إيمتي
أيوه
إمبيرح
إنتي
إنتوا
إحنا
اللي
بني
بر
بس
انو
Exception 2

Colloquial Verbs Conjugation Paradigm

ما بقراش
ما بقري
ما بيجي
ما بيشوفش
هو

ما بترقراش
ما بتقرى
ما بتييجي
ما بتشوفش
هي

ما بفرقروش
ما بقرى
ما بيجوش
ما بتشوفوش
هم

ما بتقرش
ما بتقرى
ما بتييجي
ما بتشوفيش
إنت

ما بتقريش
ما بتتييجي
ما بتشوفيش
إنتي

ما بتقروش
ما بتقرى
ما بتييجوي
ما بتشوفوش
إنتوا

ما بقراش
ما بقري
ما بيجي
ما بشوفش
أنا

ما بنقراش
ما بنقرى
ما بنيجي
ما بنتشوفش
إحنا
Exception 2

Colloquial Verbs Conjugation Paradigm

ما قراش ـ قرى ـ ما جاش ـ إجي ـ ما شافش ـ هو
ما قرتش ـ قرت ـ ما جتش ـ إجت ـ ما شافتتش ـ هي
ما قروش ـ قروا ـ ما جوش ـ إجوا ـ ما شافوش ـ هم
ما قريتش ـ قريت ـ ما جيتش ـ جيت ـ ما شفتتش ـ إنت
ما قريتيش ـ قريتي ـ ما جيتتش ـ جيتي ـ ما شفتتش ـ إنتي
ما قريتوش ـ قريتوا ـ ما جيتوش ـ جيتوا ـ ما شفتتش ـ إنتوا
ما قريتش ـ قريت ـ ما جيتش ـ جيت ـ ما شفتتش ـ أنا
ما قريناش ـ قرينا ـ ما جيناش ـ جينا ـ ما شفناش ـ إحنا
Exception 3

Nunation (tanween) should reflect actual pronunciation

مرحباً /marHaban/
مرحبا /marHaba/
أهلاً وسهلاً /?ahlan wa-sahlan/
أهلا وسهلا /?ahla wa-sahla/
**Variation in orthography**

**Issues:** Choose the variant with the highest frequency of usage

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>أنا باحكي</td>
<td>45</td>
</tr>
<tr>
<td>أيوا</td>
<td>1,420</td>
</tr>
<tr>
<td>برضه</td>
<td>2,540</td>
</tr>
<tr>
<td>أنا بحكي</td>
<td>455</td>
</tr>
<tr>
<td>أيوه</td>
<td>2,530</td>
</tr>
<tr>
<td>برضو</td>
<td>3,180</td>
</tr>
</tbody>
</table>
Variation in orthography

Issues:

*Transcribe hamza when it is pronounced*

محمد أبو ممتاز
/mumtaːz ɣ-abu muHamːad/

أهلين أبو طارق
/?ahleːn ɣ-abu Taːriq/

أنا والابخلاد
/?ana wa-λiwlaːd/

الأب والأولاد
/?il-ɣ-ab wa-l-ʔawlaːd/
CONCLUSION: Collection Update

[September 19, 2004]

- 13604 Recruits (Domestic, International) / 11450 active callers
- 2184 calls completed
- 1662 are available as of today.
- 1400 of them have more than 8 minutes speech.
- Male-Female ratio among the 2184 calls where the genders of both speakers are available: M M 710 / F F 300 / M F 354 / F M 398
  Male to female ratio is: 1086 to 676 = 61.6% to 38.4%
  [Note that when calls involve speakers with no gender information, those calls are excluded from the calculations above].

- 2305 speakers were used for the 2184 calls. 1251 speakers only appeared in 1 call; 381 appeared in 2 calls; 488 appeared in 3 calls. [1 times 1251; 2 times 381; 3 times 488; 4 times 117; 5 times 41]

2 hrs EVALUATION SET/2 hrs DEVELOPMENT SET
68 hours + 32 hours TRAINING SET

For more information, go to:
http://www.ldc.upenn.edu/Projects/EARS/Arabic/Guidelines_Levantine_MSA.htm