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

OUR FOCUS WILL BE ON THE ARABIC DIALECTAL TRANSCRIPTION RATIONALE, THE TECHNOLOGICAL GOALS OF THE PROJECT, THE ANNOTATION TOOL STRUCTURE AND THE LEVANTINE CONVERSATIONAL 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

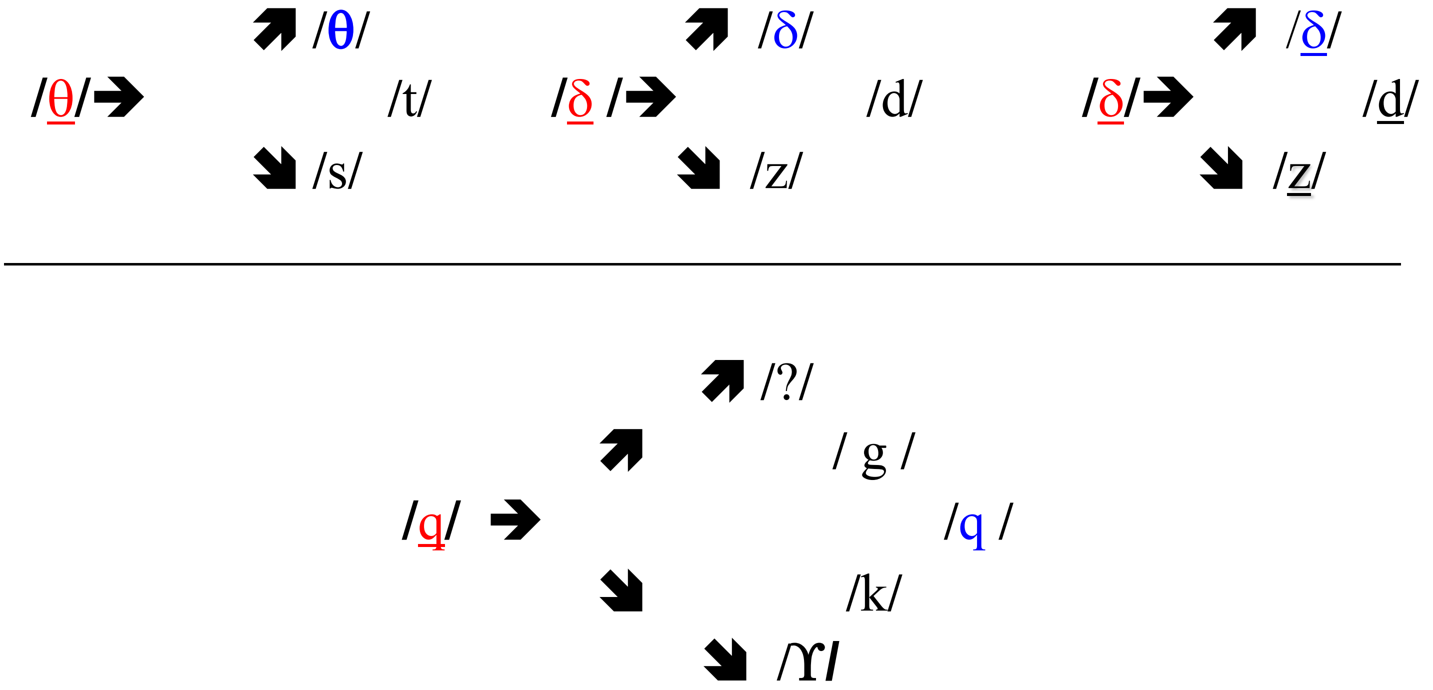
ARABIC LANGUAGE BACKGROUND

- 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/ ↗ /δ/ ↗ /δ/

/θ/→ /t/ /δ /→ /d/ /δ/→ /d/

↘ /s/ ↘ /z/ ↘ /z/

↗ /?/

↗ / g /

/q/ → /q /

↘ /k/

↘ /Ɣ/

ARABIC DIALECTAL VARIATION

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

Example: Iraqi.q.h.C.wav

EXISTENCE AND USE OF ARABIC SCRIPT

“ARCHIGRAPHEMES”

LEVANTINE ARABIC EXAMPLE

Q: شو القصة?

$w AlqSp?

"What's the story?”

A/T1: يا زلمي كلتلك موكوف مش معتكل وما في كصّة

yA zlmy kltlk mwkwf m$ mEtkl wmA fy kS~p

A/T2: يا زلمي قلتلك موقوف مش معتقل وما فيه قصّة

yA zlmy qltlk mwqwf m$ mEtql wmA fy qS~p

"Hey ‘dude’ I told you arrested not indicted and there is no story”

Specification Issues

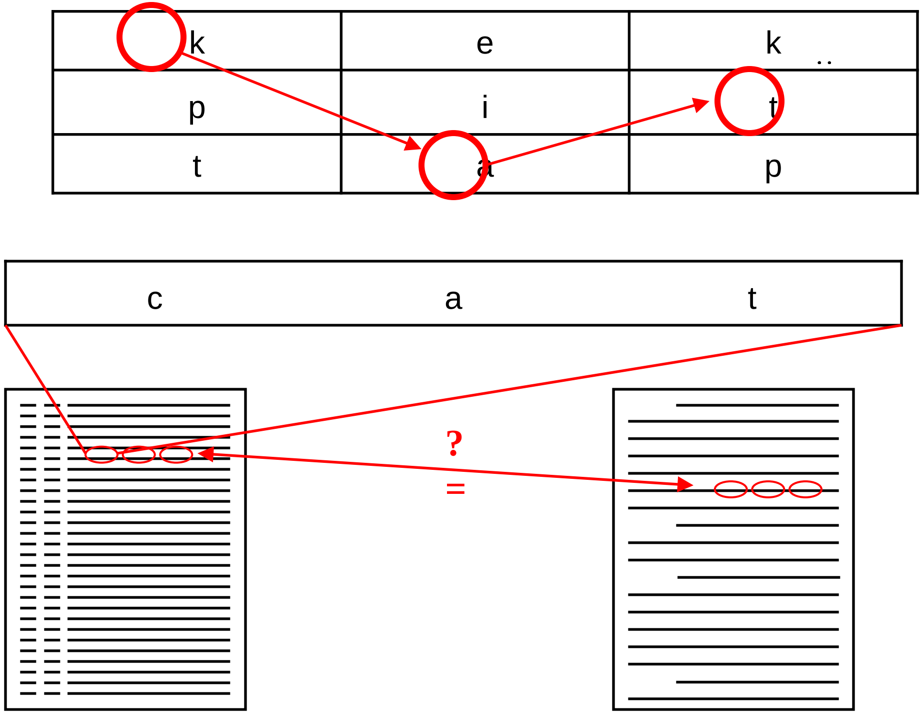
- **Need to distinguish the transcription approach from the alphabet used.**
  - **Transcription approaches: phonic, orthographic, hybrid**
  - **Alphabets: Arabic, Roman, International Phonetic Alphabet**
  - **One may perform either phonic or orthographic transcription using either Roman or Arabic 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

α1	α2	α3	α4	α5	α6	α7	α8	A9
β1	β2	β3	β4	β5	β6	β7	β8	B9
γ1	γ2	γ3	γ4	γ5	γ6	γ7	γ8	γ9
δ1	δ2	δ3	δ4	δ5	δ6	δ7	δ8	δ9

- Analysis suggests multiple phonetic interpretations.



k	e	k
p	i	t
t	a	p

- 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.

LDC CONVERSATIONAL DIALECTAL ARABIC STT RATIONALE

"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 → WITH HIGHER RECONSTRUCTION RATE OF 'UNDERLYING' FORMS
- Nor too cloose an adherence to finer sound /(allo)phonic/ acoustical utterance representation → 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’

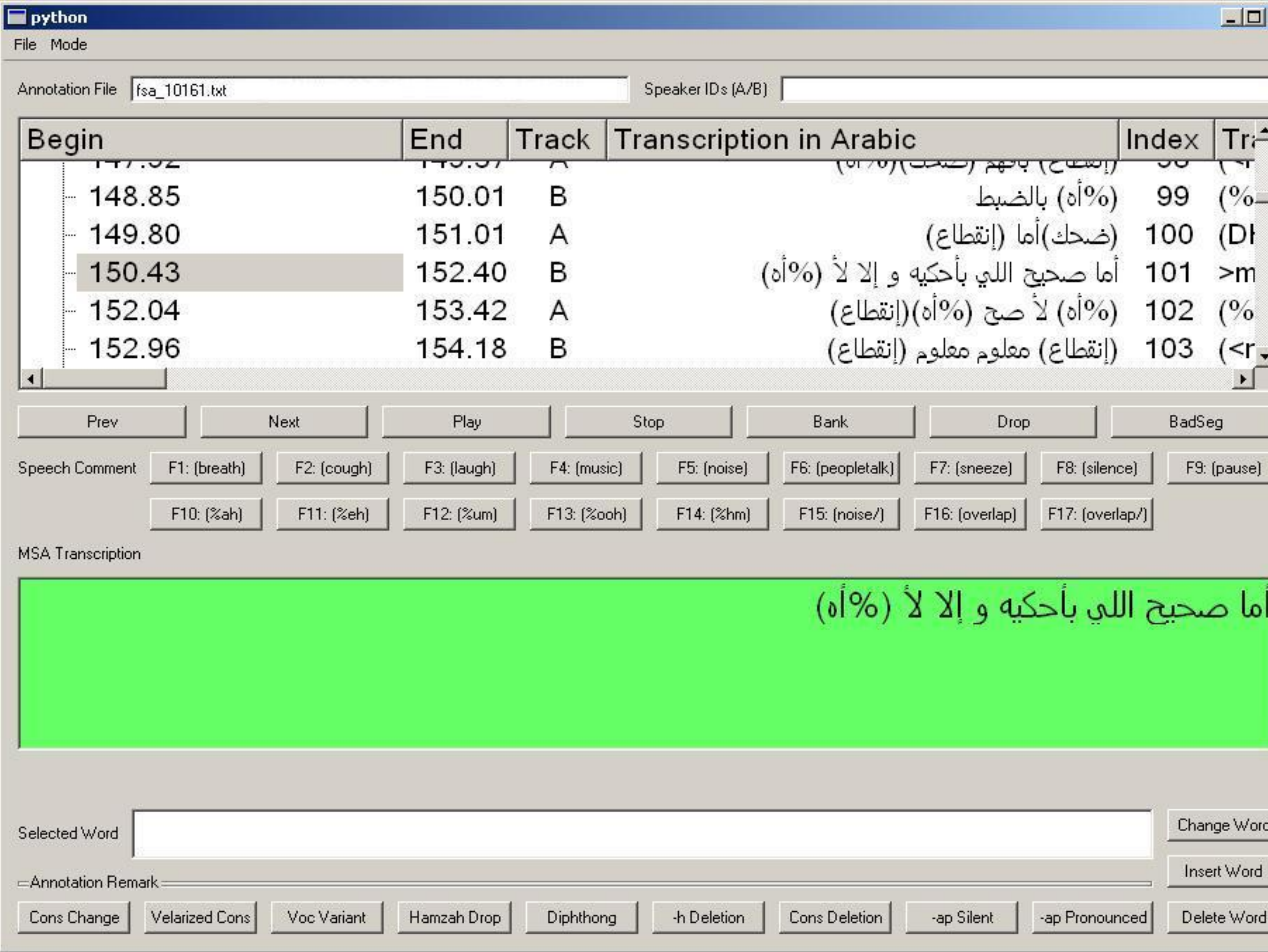
- USE OF NORMAL ARABIC KEYBOARD FOR TRANSCRIPTION
- FIRST PASS WITH MSA-BASED APPLICABLE CONVENTIONS
- METALANGUAGE ANNOTATION (CTS RT-04 ANNOTATION)
  OBJECTIVE: OPTIMIZED OUTPUT FOR LANGUAGE MODELING

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



Annotation File fsa_10161.txt

Speaker IDs (A/B)

Begin	End	Track	Transcription in Arabic	Index	Tra
147.92	149.37	A	(إنقطاع) بافهم (ضحك)(%أه)	98	(<r
148.85	150.01	B	(%أه) بالضبط	99	(%
149.80	151.01	A	(ضحك)أما (إنقطاع)	100	(DI
150.43	152.40	B	أما صحيح اللي بأحكيه و إلا لأ (%أه)	101	>m
152.04	153.42	A	(%أه) لأ صح (%أه)(إنقطاع)	102	(%
152.96	154.18	B	(إنقطاع) معلوم معلوم (إنقطاع)	103	(<r

MSA Transcription

أما صحيح اللي بأحكيه و إلا لأ (%أه)

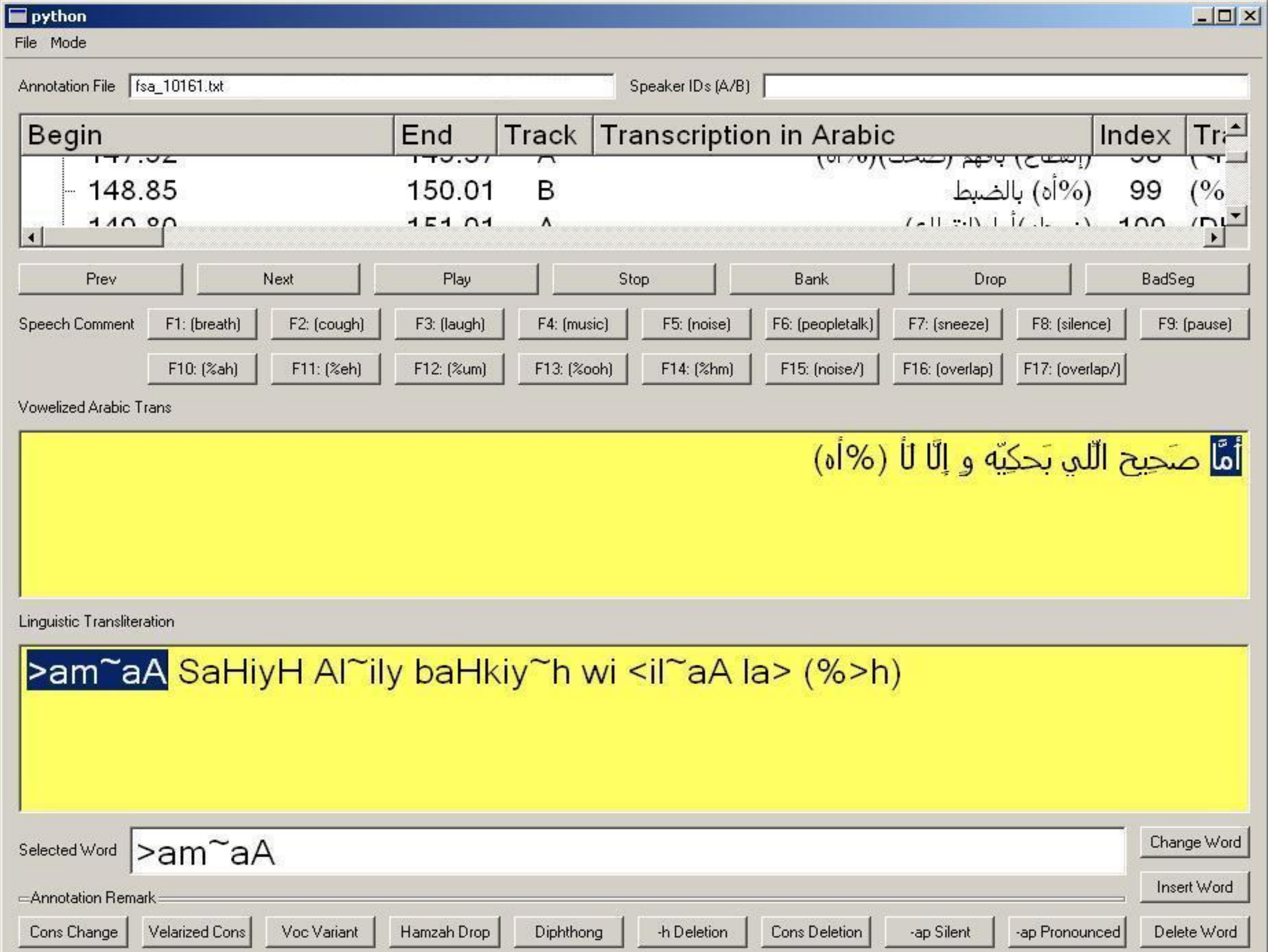
Selected Word

‘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”



Annotation File: fsa_10161.txt

Begin	End	Track	Transcription in Arabic	Index	Tr
148.85	150.01	B	(%أه) بالضبط	99	(%

Vowelized Arabic Trans:

أمَّا صَحِيح اللّي بَحكِيّه وِ إلَّا لأ (%أه)

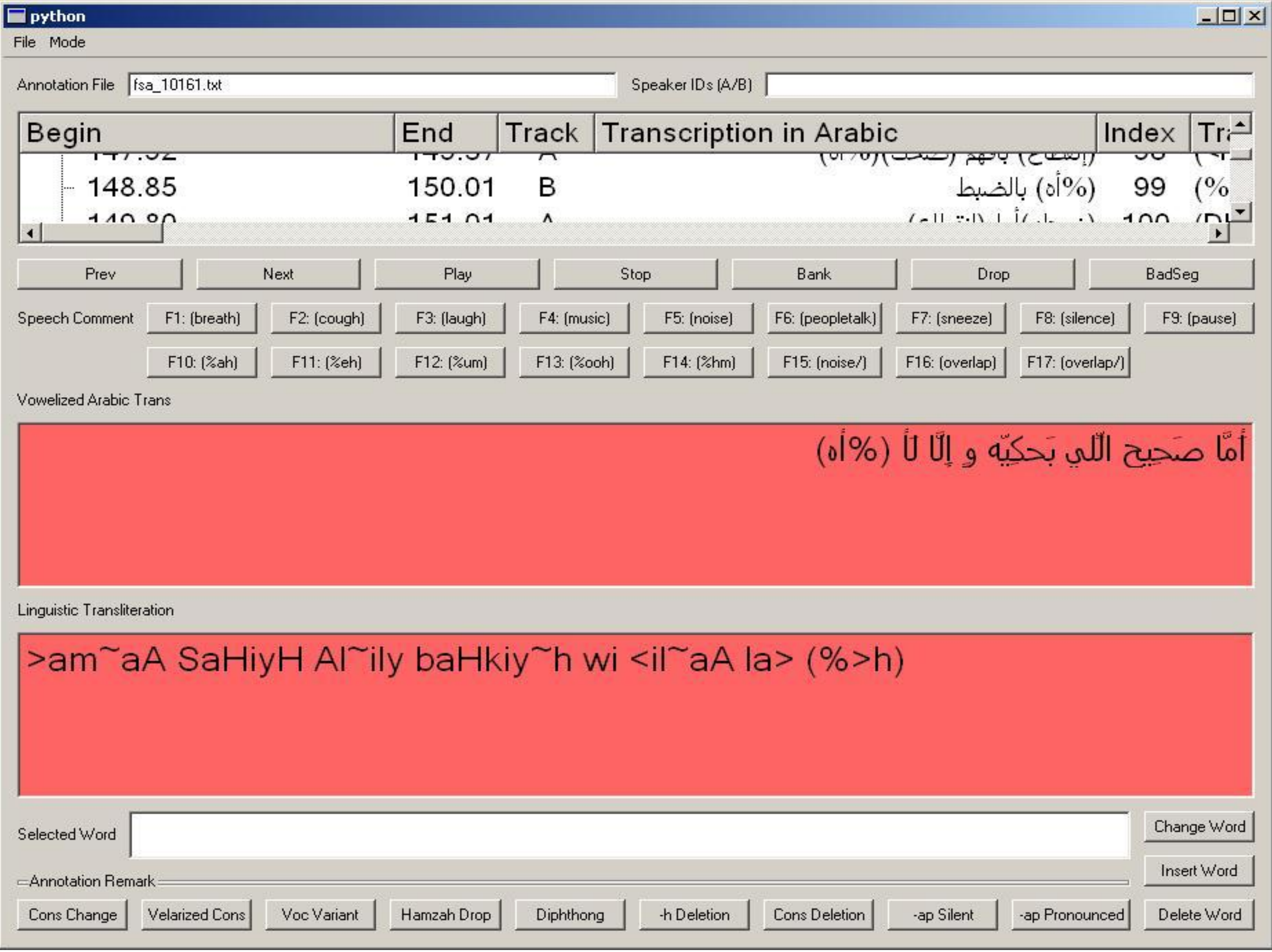
Linguistic Transliteration:

>am~aA SaHiyH Al~ily baHkiy~h wi <il~aA la> (%>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



Annotation File: fsa_10161.txt

Speaker IDs (A/B):

Begin	End	Track	Transcription in Arabic	Index	Tr
148.85	150.01	B	(%أه) بالضبط	99	(%

Vowelized Arabic Trans

أَمَّا صَحِيح اللّي بَحكِيّه وِ إِلَّا لأَ (%أه)

Linguistic Transliteration

>am~aA SaHiyH Al~ily baHkiy~h wi <il~aA la> (%>h)

Selected Word:

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

LEVANTINE ARABIC GUIDELINES

MSA-based orthography

“whenever possible, follow the spelling conventions and word segmentation of MSA.” ***Like this:***

قلت لك	/?ultil:ak/
مضبوط	/mazbu:T/
مثل	/mitl/
مثلا	/masalan/

MSA-based orthography

"whenever possible, follow the spelling conventions and word segmentation of MSA." ***Avoid this:***

ألتلك	`/?ultil:ak/`
مزبوط	`/mazbu:T/`
متل	`/mitl/`
مسلا	`/masalan/`

MSA-based orthography

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)

إحنا
اللي
إمبيرح
إنتي
إنتوا
أنو

إيد
أيش
إيمتى
أيوه
برا
بس

بعدين
بكرة
بلكي
بينات
جوا
دغري

زلمه
زي
شو
شوي
شوية
عشان

علشان
عم
فا
فيه
فيش
فين

MSA-based orthography

Exception 2

Colloquial Verbs Conjugation Paradigm

هو	بيشوف	ما بيشوفش	بيجي	ما بيجيش	بيقرى	ما بيقراش
هي	بتشوف	ما بتشوفش	بتيجي	ما بتيجيش	بتقرى	ما بتقراش
هم	بيشوفوا	ما بيشوفوش	بيجوا	ما بيجوش	بيقروا	ما بيقروش
إنت	بتشوف	ما بتشوفش	بتيجي	ما بتيجيش	بتقرى	ما بتقراش
إنتي	بتشوفي	ما بتشوفيش	بتيجي	ما بتيجيش	بتقري	ما بتقريش
إنتوا	بتشوفوا	ما بتشوفوش	بتيجوا	ما بتيجوش	بتقروا	ما بتقروش
أنا	بشوف	ما بشوفش	بجي	ما بجيش	بقرى	ما بقراش
إحنا	بنشوف	ما بنشوفش	بنيجي	ما بنيجيش	بنقرى	ما بنقراش

MSA-based orthography

Exception 2

Colloquial Verbs Conjugation Paradigm

هو	شاف	ما شافش	إجى	ما جاش	قرى	ما قراش
هي	شافت	ما شافتش	إجت	ما جتش	قرت	ما قرتش
هم	شافوا	ما شافوش	إجوا	ما جوش	قروا	ما قروش
إنت	شفت	ما شفتش	جيت	ما جيتش	قريت	ما قريتش
إنتي	شفتي	ما شفتيش	جيتي	ما جيتيش	قريتي	ما قريتيش
إنتوا	شفتوا	ما شفتوش	جيتوا	ما جيتوش	قريتوا	ما قريتوش
أنا	شفت	ما شفتش	جيت	ما جيتش	قريت	ما قريتش
إحنا	شفنا	ما شفناش	جينا	ما جيناش	قرينا	ما قريناش

MSA-based orthography

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***

45	أنا باحكي	455	أنا بحكي
1,420	أيوا	2,530	أيوه
2,540	برضه	3,180	برضو

Variation in orthography

Issues:

Transcribe hamza when it is pronounced

ممتاز يا ابو محمد	/mumta:z y-abu muHam:ad/
أهلين أبو طارق	/?ahle:n ?abu Ta:riq/
أنا والاولاد	/?ana wa-liwla:d/
الأب والأولاد	/?il-?ab wa-l-?awla:d/

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

LDC [illegible]