Appendix E: Arabic STT Scoring Protocol

Overview

The following document describes how the STT evaluation rules described in the evaluation plan will be applied to scoring the RT-03 Arabic STT tests. The Arabic STT evaluation will use the same evaluation infrastructure as the English STT tests, (e.g., input UEMs, CTM output format, Word Error Rate computation etc.) For simplicity, the same rule applications will be in effect for both the Broadcast News and Conversational Telephone Speech.

This appendix does not describe the full evaluation infrastructure, but rather covers only those issues not described in sufficient detail for implementing the Arabic STT evaluation.

Text Encoding

The STT systems are required to output UTF-8 encoded Arabic script. While it is recognized that a bi-directional transliteration is available to sites, presenting properly rendered Arabic script to people outside of EARS and the Rich Transcription community, e.g. TIDES, will be a more useful product.

Lexical Pre-Processing

The following language specific textual pre-processing steps will be applied prior to the STT scoring process. The handling of non-speech tokens, fragments, and uncertain lexemes will be as specified in section 3.3 of the evaluation plan. However, for Arabic the following phenomena require additional details: pause fillers, foreign words, compound words, cross-channel speech, and semantically equivalent words.

Filled Pauses

Filled Pauses in both the reference and system transcripts will be converted to a single form for scoring %HESITATION, The Arabic word 'trd~d' تردّد % should be used to mark all forms of hesitation. In the reference, all filled pauses will be marked as optionally deletable [See footnote 24] while they will be normal lexemes in the system output. The following is the set of recognized pause fillers:

Transliteration	Arabic Script
%>h	%هأ
%<yh	%هيإ
%>m	%مأ
%>ww	%ووأ
%hm	%مه
%mhm	%ممم

Backchannel Responses

In the LDC's Arabic transcripts, there is only one notable backchannel response, '%>hh' 'هأ%'. Since there are not several acoustically similar words for backchannel responses, this word will be scored as a regular word token without the percent sign '%'.

Foreign Words

As with any spoken language, Arabic talkers "code switch" into different languages as they speak and the Arabic language has borrowed foreign words. For example, when bi-lingual talkers code-switch, they temporarily change languages to express a thought if they have trouble expressing it in Arabic or for special effect. Alternatively, foreign words like 'cassette' (kasAt in Romanized form) are used on a regular basis and are fully integrated into the language.

The Arabic STT scoring rules for code switched foreign words and borrowed foreign words are consistent with the English STT evaluation rules. Specifically, code switched foreign words will be transcribed in their native orthography and scored as optionally deletable. Borrowed foreign words, which are transcribed in Arabic script, are scored like typical lexemes.

The human-generated reference transcripts do not include transcriptions for code switched data. Bi-directional text rendering algorithms used by editing tools often mislead the user thus making transcription difficult. Instead, a single tag '%>jnby' will be used to indicate each occurrence and a separate file will be provided containing the non-Arabic transcript. NIST will then merge the files to make an STM file for scoring.

Compound Words

Compound words are not common in Arabic, therefore no scoring accommodations will be made.

Cross Channel Speech

Cross talk in the CTS domain is a challenge that STT systems must handle. Systems are required to not emit speech tokens for cross channel speech. The LDC has transcribed significant cross-channel speech between the tags "%tdAxl" and "%tdAxl\". The text pre-processing steps will remove both the tags and all the transcribed speech in between.

Initial Hamza normalization

In colloquial Arabic, hamza attachment to the alif is often a dialect-dependent variation. Whether native speakers do or do not use the lexeme-initial hamza, the meaning of the word does not change. Our practice with hamza is to transcribe it when it is used and not show it when it is not. When the hamza is not pronounced initially, its support, | or A in Tim Buckwalter's code, which is silent, should be transcribed. Since there is no semantic difference between the uttered variant word forms with an initial hamza, and since it is believed that transcribers cannot consistently transcribe them, the scoring code will map all lexeme-initial hamzas, 'A', '<', '>', and '|' into 'A', the hamza without an alif prior to scoring.

Semantically Equivalent Words

In colloquial Arabic, words may be transcribed differently even though the underlying word is the same. The scoring tools will map all colloquial forms of the words to their normalized orthographic form equivalents via a GLM file. The LDC will hand generate the list of variant forms with equivalent normalized MSA-orthographic word forms during the transcription process and show in a separate list the normalized orthographic forms which could be used in post processing text normalization. The following is a partial table of equivalent forms from the RT-04f Arabic development set. This normalization process will be performed prior to lexeme-initial hamza normalization.

variant form	normalized orthographic form
|yh	<yh
>llh	Allh
<strAlyA	>strAlyA
<$yA'	>$yA'
<lly	Ally
<wf	>wf
<ywh	>ywh
Avnyn	<vnyn