MATERIAL Program Transcription Conventions

This document contains MATERIAL transcription conventions for speech files. Transcription in MATERIAL was performed in two ways, “Quick Transcription” described Section 1, and “Final Transcription” described in Section 2.

1 “Quick” Transcription Conventions

1.1 Introduction

The following conventions apply to all “quick” Transcriptions produced for the Material program. “Quick” Transcriptions were provided for all collected *broadcast* data. Please note: a subset of the collected broadcast data (along with the conversational speech data) will go through a full transcription pass. This subset was nominated based on its responsiveness to queries. As a result, the nominated subset of broadcast speech data will go through a “quick” transcription pass and will ultimately be reworked to comply with the *Final Transcription Conventions* (Section 2 below) .

This data is intended to provide the audio files’ speech content in text format to be used in all subsequent processing of the audio data such as domain annotation and query relevance. It replaces the previously discussed “gisting” effort. The purpose of this transcription is not intended for ASR development and as such, effort was not invested in transcription aspects that support ASR development such as acoustic tagging and spelling normalization. Consequently, transcriptions were done in a single-pass and were non-standardized. Existing autocorrect and spell checking tools (where available) were applied to this transcription but further spelling standardization or other post-processing activities were not performed.

1.2 Speech Events

1.2.1 Hesitancies

Hesitancies (fillers) are sounds made by the speaker to indicate that they are still continuing but have made a mistake or are thinking of what to say next. In English, these would include fillers such as “ah”, “um”, “er”, “hm”, etc.
These will all be tagged using the <hes> tag.

1.2.2 Fragments

If a speaker stumbles mid-word it was transcribed up to the cut-off point and hyphenated.
For example: "to- tomorrow" -> to- tomorrow
For example, if the word is not repeated correctly: "to- the day after tomorrow"-> to- the day after tomorrow
In less frequent cases the hyphen will occur at the beginning of the word. For example: “-morrow”

1.2.3 Unintelligible Speech

If a word is unintelligible the (()) tag was utilized.
The unintelligible tag was used regardless of whether it is a single word or a string of speech.

1.2.4 Truncations

Truncations refer to cases where a fragment of a word appears because the recording cut off the full word or the segment boundary crosses a word. Truncations were treated like fragments (see Section 5.2.2) in the “quick” transcription.

1.2.5 Foreign Words and Code-Switching

The <foreign> tag was used when the transcriber does not understand a word or string of words from another language. This utterance was not transcribed and the <foreign> tag was inserted instead.

1.2.6 No Speech

This tag is to be used for any period greater than one second in which there is no speech from the main speaker. Even if there are some foreground sounds, only the <no-speech> tag was used if there is no actual speech for more than one second.
For example: "silence <breath> silence" should be transcribed as "<no-speech> "

1.2.7 Overlap

Overlap occurs when two foreground speakers talk at the same time.
The overlapping words were not transcribed and the <overlap> tag was inserted instead.

1.2.7.1 Backchanneling

Backchanneling occurs when one foreground speaker is talking, and another foreground speaker is simultaneously producing short words that show active listening (for instance, saying 'yes' or 'uh-huh' or 'mmm'). In these cases, transcribers will treat the backchanneling speaker as background noise and transcribe the speech of the main speaker.

1.2.8 Multiple Foreground Speakers

This is expected to be common in the data. All foreground speakers were transcribed.

Speaker turns were not tagged.

1.3 Restarts

A restart occurs when a speaker interrupts or repeats themselves, causing a sentence or phrase to restart. In such cases, the double dash (--) symbol was placed at the point where the speaker stops and starts again.

For example: I will go -- I will go there tomorrow.

1.4 Punctuation

Transcribers will make use of punctuation that is natural to speakers/readers of the relevant language in order to facilitate easy and fluent reading of the text by annotators and other data users.

2 Final Transcription Conventions

2.1 Introduction

The following conventions apply to all Final Transcriptions (as opposed to "quick transcription") produced for IARPA's MATERIAL program by Appen. Final Transcriptions were provided for all transcriptions completed in accordance with the BABEL program specifications: IARPA_Babel_Specification-08262013 (nist.gov).

Note: Speech files that belong to the MATERIAL *analysis set* underwent "quick transcription" initially, followed by a second pass to produce final transcriptions". When these files underwent the second pass to produce the final transcription, the following methodology was applied:

- The transcriptions were modified to adapt to the conventions described in this section.
- Changes to spelling of lexical items were limited to:
  - Correction of spelling errors, and
  - Standardization of spelling to conventions described in section 2.4below as well as to the approach described in the language specific LSDD. The spelling standardization decisions will be consistent with those applied in the transcription of the conversational telephony data transcribed as part of the BABEL or MATERIAL program.

The final transcription is intended to be a broad transcription. Transcribers should not have to agonize over decisions. The transcription is lexical, with tagging to represent audible acoustic events (speech and non-speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance.

The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc.

The character set to be used for the orthographic transcriptions is UTF-8. A Language Specific Design document (LSDD) provides the format of the orthographic transcription for languages that do not use white space as word boundaries and/or do not use a Latin alphabet. In such cases, the LSDD will contain a description of the method used for word boundary detection as well as a complete list of the symbols used in Romanization (e.g., Pin Yins or Romaji syllables).

2.2 Speech Events

2.2.1 Hesitancies

Hesitancies (fillers) are sounds made by the speaker to indicate that they are still continuing but have made a mistake or are thinking of what to say next. In English, these would include fillers such as “ah”, “um”, “er”, “hm”, etc.

These will all be tagged using the <hes> tag.

For each language, a list of possible fillers will be documented, and corresponding pronunciations will be provided in the pronunciation lexicon.

2.2.2 Mispronunciations

If the speaker mispronounces a word, the word will be spelled correctly and annotated using the * * tag. It will enclose the word in stars.

For example: “representive” -> *representative*

2.2.3 Fragments

If a speaker stumbles mid-word, it will be transcribed up to the cut-off point and hyphenated.

For example: "to- tomorrow" -> to- tomorrow

For example, if the word is not repeated correctly: "to- the day after tomorrow"-> to- the day after tomorrow

In less frequent cases the hyphen will occur at the beginning of the word. For example: “-morrow”

2.2.4 Unintelligible Speech

If a word is unintelligible the (()) tag will be utilized.

The unintelligible tag will be used regardless of whether it is a single word or a string of speech.

2.2.5 Truncations

Truncations may occur at the very beginning or end of an utterance if the recording device has cut off a word.

The affected word shall be transcribed in full and marked with the ~ tag.

For example:

~ satisfactory (truncation at beginning of utterance)

unsuitable ~ (truncation at end of utterance)

(()) ~ (truncation with unintelligible word)

While a truncation is similar to a fragment, it will not be marked as a fragment.

2.2.6 **Foreign Words and Code-Switching**

The <foreign> tag will be used when the transcriber does not understand a word or string of words from another language. The foreign words will NOT be transcribed and the <foreign> tag will be inserted instead.

During post-processing of the transcribed data and a review of the unique word list that appears in the transcriptions, words that are deemed as foreign words will be globally replaced with the <foreign> tag to ensure consistency. Individual loan words that are spoken that are commonly used part of the native language will not be replaced with <foreign> and will be transcribed with the accepted loan word spelling. For example, words such as “kimono”, “croissant”, or “falafel” would be considered commonly accepted loan words in the English language. Such words will be written using the same character set as the native language.

For conversational data, please note that speakers will be directed to speak only in the language of the collection. Any calls with excessive amounts of foreign language use or code-switching will be rejected.

2.3 Non-speech ("acoustic") events

These are categorized as either foreground or background sounds.

2.3.1 **Background sound**

Continuous low level background noises do not need to be tagged.

For continuous medium to loud background noise, the <sta> tag will be inserted once at the point where the sound begins.

2.3.2 **Foreground sounds**

Events will only be transcribed if they are clearly distinguishable. Very low-level, i.e., non-intrusive events will be ignored. The event will be transcribed at the place of occurrence, using the defined symbols in angle brackets. For noise events that occur over a span of one or more words, the transcription should indicate the beginning of the noise, just before the first word it affects. If a noise occurs more than once in sequence, the appropriate tag will only be inserted once.

The first four categories of acoustic events originate from the speaker, and the other categories originate from another source. Sounds originating from the speaker usually do not overlap with the target speech, sounds originating from other sources can of course occur simultaneously with the speech.

The categories are:

<lipsmack>	lip smacks, tongue clicks
<breath>	inhalation and exhalation between words, yawning
<cough>	coughing, throat clearing, sneezing
<laugh>	laughing, chuckling
<click>	machine or phone click
<ring>	telephone ring
<dtmf>	noise made by pressing telephone keypad
<int>	any other intermittent foreground noise

If any of the above events overlap with a word and the event is loud enough to render the word useless, the appropriate tag will be inserted, and the word will not be transcribed.

2.3.3 **No Speech**

This tag is to be used for any period greater than one second in which there is no speech from the main speaker. Even if there are some foreground sounds, only the <no-speech> tag will be used if there is no actual speech for more than one second.

For example: "silence <breath> silence" should be transcribed as "<no-speech> "

2.3.4 Overlap

Overlap occurs when two foreground speakers talk at the same time.

The overlapping word(s) will NOT be transcribed and the <overlap> tag will be inserted instead.

2.3.5 Prompt

This tag will be used for an electronic voice or automated recording.

This utterance will NOT be transcribed and the <prompt> tag will be inserted instead.

2.3.6 Change of Speaker (For different gender only, in conversational data only)

Demographics of speakers are specified in the session metadata.

The following tags will be used to indicate if the gender of a speaker changes during a call. The appropriate tag will be inserted at the exact point at which the new speaker starts:

<male-to-female>

<female-to-male>

If the speaker changes after a section of overlapping speech, this speaker change tag will be inserted after the overlap tag.

2.3.7 Miscellaneous Segments

The <misc> tag will be used for sections of music, promos, ads, etc. in broadcast data. This tag will be inserted at the point where the miscellaneous segment begins. If such an acoustic event occurs as background to the speech of a main speaker it will be handled as a background noise acoustic event (i.e. the <sta> tag will be used).

2.4 Spelling

2.4.1 Proper Nouns

Proper names will be transcribed in a case-sensitive manner in applicable languages. Initials should be in capital letters with no period following.

For example: George W Bush has confirmed his relationship with the South American government

2.4.2 Titles and abbreviations

All titles and abbreviations will be transcribed as a word.

For example: Dr -> Doctor

Exception: if the abbreviated form was actually pronounced

Speaker says 'Appen Butler Hill Inc' (instead of 'Appen Butler Hill Incorporated'), the word 'Inc' will be transcribed.

2.4.3 **Punctuation**

Word-level punctuation will be used only if it is an essential part of the word. For example:

can't

Sentence-level punctuation will not be used, except in datasets that will be translated (see section 2.5 below).

2.4.4 **Acronyms**

Acronyms will be transcribed as words if spoken as words, and as letters if spoken as letters. When transcribing sequences of letters an underscore will be inserted between each letter

For example: NASA, I_B_M

2.4.5 **Numbers**

Numbers will be transcribed as full words

For example: 16 -> sixteen

112 -> one hundred and twelve

2.4.6 **Phonetic Spelling**

The / / tag is used for letters that are pronounced as the sound, rather than as the word - for example when a person means to convey the letter B but they say the sound 'buh' instead of the word 'bee'. In this case we transcribe it as /B/.

2.5 Additional conventions for transcriptions undergoing translation

In addition to the above set of transcription conventions, the following requirements apply to the subset of broadcast speech and conversational speech documents that will be translated.

The following additional requirements support this and other needs of the translation process.

2.5.1 **Transcription segment boundary**

In order to aid translation, all transcription data which will be translated will be re-segmented in such a way that segments contain meaningful semantic units as much as feasible given the constraints of natural speech. This will create transcription segments that are more coherent for translation and facilitate a bitext format of translation with the source text in each bitext corresponding to a single transcription segment.

As such, each transcription segment should end in either a full stop or a question mark, or be marked %incomplete if the statement was cut short and never finished.

Segment boundary points will be selected as much as possible so that the boundary is the logical end of one sentence-like sequence and the logical beginning of another sentence-like sequence.

For example:

Segment 1: good morning I think you know John Bailey from the other day .
Segment 2: he suggested we meet and discuss how we might solve your problem .
Segment 3: and we will try to do whatever we can to find a resolution .

A segment boundary should not be added just because the speaker hesitates. Short pauses and fillers like “um” or “uh” should not be treated as breakpoints if they occur within a logical sentence-like sequence. For example:

I think you know uh... John Bailey from um... the other day

The above text SHOULD NOT be broken up like this:

NO! Segment 1: I think you know <hes> %incomplete
NO! Segment 2: John Bailey from <hes> %incomplete
NO! Segment 3: the other day .

The text should be transcribed as one segment:

Segment 1: I think you know <hes> John Bailey from <hes> the other day .

If possible, excessively long utterances should be broken into shorter utterances.

2.5.2 **Restarts**

A restart occurs when a speaker interrupts or repeats themselves, causing a sentence or phrase to restart. In such cases, the double dash (--) symbol will be place at the point where the speaker stops and starts again.

For example: I will go -- I will go there tomorrow .

2.5.3 **Punctuation**

Sentence-level punctuation (periods and question marks only) will be used to mark logical break points in speech. For example:

have you been down to the river ? it's lovely on the west bank .

If a sentence was cut short and never finished the %incomplete tag is used. For example:

I'm going to the %incomplete

Periods, question marks and the %incomplete tag will be preceded by a space.

3 Lexicon Specifications

All lexicons will apply the same specifications as the BABEL conversational lexicon. Normalization will be consistent with prior lexicons delivered to BABEL program. See section 6 of the BABEL program specifications: IARPA_Babel_Specification-08262013 (nist.gov) for lexicon information.