A Universal Transcription Format (UTF) Annotation Specification for Evaluation of Spoken Language Technology Corpora

Introduction

There are many transcribed corpora currently in use within the speech research community, each having unique annotation conventions. The creation and maintenance of separate software libraries to support these corpora has multiplied overhead costs borne by the researchers.

These annotation conventions specified in this document are intended to provide an extensible universal format for transcription and annotation across many spoken language technology evaluation domains, beginning in 1998 with Hub-4 and Hub-5. These transcriptions are used to capture the orthography of spoken words in recordings of speech and can also include annotations which associate certain signal, speaker, and content conditions with the speech and its transcription. The convention is formatted using SGML markup and has an associated DTD (document type definition) which describes the structure. SGML tags identify the location of the speech within the recorded waveform using start/end time attributes, and they identify the corresponding transcription by enclosing it within these span of the tags. SGML was chosen to maximize portability and to exploit existing knowledge and software tools. This release represents a re-formatting of existing transcripts into the UTF format which is extensible to many speech domains and to permit the annotations required by the Information Extraction-Named Entity evaluation spoke of the 1998 Hub tests.

As a first step, the UTF format is being used to unify the transcription and annotation formats for both the Hub-4 Broadcast News and Hub-5 Conversational speech corpora.

The UTF format has at its core a notion of a speaker **Turn** which is implemented with the SGML tag *<turn>*. A speaker Turn marks the extent and content of the word(s) spoken by a single speaker, spoken in a particular speaking style. In terms of SGML hierarchy, a Turn is the lowest level of organization within a broadcast or conversation, and represents the lowest common denominator across corpora. The transcribed text within a Turn is formatted identically across corpora.

The differences between corpora are expressed through higher level structures which contain turns as sub-units. For example, the annotation of broadcast news data requires the tagging of topical units (sections) which contain zero or more turns.

Word Tokenization

A **Turn** brackets the orthography of the words spoken during the turn. To make the orthography usable, word division, or word tokenization rules, must be defined. However, different uses of this data may require different word tokenization rules. For example, named entity annotation does not include the possessive 's as part of a named entity while ASR does not make this distinction. Since multiple tokenizations need to be encoded in the orthography, the UTF format uses SGML markup to indicate word boundaries. The ASR tokenization is used as the default and is defined in the UTF DTD via the <separator> tag. The *<separator>* tag indicates that a word boundary has occurred and as a consequence, words are located between *<separator>* tags.

UTF Document Tags and Attributes

The SGML tag structure of the UTF file format is described by the following simplified BNF grammar. There are four categories of tags in a UTF document: 1) structural, 2) state change, 3) pseudo bracketing, and 4) lexical.

1. Structural tags indicate the hierarchical organization of the transcribed recording and mirror the semantic structure of the domain.

2. State change tags indicate attributes of the recording which can occur independently of the spoken content. Currently, the only state changes defined in the UTF format are for that of background acoustics. These tags are not bracketed by turns.

3. Pseudo bracketing tags indicate an annotated attribute of a span of words. Since it is common to annotate overlapping regions with different attributes, these tags do not use the traditional SGML hierarchy of begin and end tags, (i.e. <doc> … </doc>). Instead, explicit begin tags and end tags (i.e. <b_doc> …. <e_doc>) are used for each annotated attribute. These tags are bracketed by turns.

4. Lexical tags indicate attributes of a single spoken word and these tags are bracketed by tuns. Tags in this category include annotations of proper names, word fragments, non-lexemes and etc.

All transcripts, regardless of corpora, begin with the structural tag *<utf>*. The *<utf>* tag spans the appropriate tag hierarchy for the corpus (Hub-4, Hub-5, etc.). As mentioned above, turns for broadcast news episodes in Hub-4 are grouped under section tags, while the conversational speech transcripts in Hub-5 have no such organization. The high level tag structure for all transcripts is as follows:

<utf> :== (<bn_episode_trans> | <conversation_trans>)

<bn_episode_trans> :== (<state_change> | <section>) *

<section> :== (<state-change> | <turn>) *

<conversation_trans> :== (<state_change> | <turn>) *

<turn> :== (<separator> | <state_change> | <pseudo_bracketing> | <lexical_tags> | TEXT) *

The state change tags are used to identify changes in background acoustics and can occur anywhere within the confines of the *<bn_episode_trans>* or *<conversation_trans>* tags. This reflects the property that background acoustics can change independently of topical organization and/or speaker turns. The pseudo bracketing tags and lexical tags pertain to the text portion of the transcript and must be wholly contained within the *<turn>* tags.

The remainder of this document is devoted to descriptions of the tags currently supported in the UTF DTD.

Structural Tags

UTF file – <utf> is a spanning tag, terminated by **</utf>**. It spans all of the annotation and transcription information associated with a particular waveform file, and it may contain either a **<bn_episode_trans>** or **<conversation_trans>** tag within its span. The attributes associated with each utf are:

dtd_version: The DTD version number for which the document is formatted. e.g., "utf-1.0".

audio_filename: The name of the file containing the recording's audio signal.

scribe: The name of the transcriber who produced the annotation and the transcription.

language: The language used in the majority of the recording.

version: The version number of the annotation of this recording, starting with "1". Each time the annotation is revised, the version number is incremented by 1.

version_date: The (last) date and time the transcript was edited.

Bn_episode_trans – <bn_episode_trans> is a spanning tag, terminated by **</bn_episode_trans>**. It is bracketed by a *<utf>* tag and it may contain **<section>** and **<Background>** tags within its span. The tag indicates that the transcript is from a broadcast news source. The attributes associated with each bn_episode_trans are:

program: The name of the program that produced the episode. (E.g., "NPR_Marketplace")

air_date: The date and time of the episode broadcast, in "YYMMDD:HHMM" format. (E.g., "960815:1300".)

Section – <section> is a spanning tag, terminated by **</section>**. It spans all of the annotation and transcription information associated with a particular section of a bn_episode_trans, and it may contain **<turn>** and **<background>** tags within its span. The attributes associated with each Section are:

startTime: The start time of the Section, measured from the beginning of the recording in seconds.

endTime: The end time of the Section, measured from the beginning of the recording in seconds.

type: One of the labels "Story", "Filler", "Commercial", "Weather_Report", "Traffic_Report", "Sports_Report", or "Local_News". For the current Hub-4 effort, Commercials and Sports_Reports will not be transcribed and will therefore contain no Segments. Sections of all other Types will be transcribed and will be included in the evaluation.

Topic: An identification of the event or topic discussed in the Section. For example, "TWA flight 800 disaster". Topic is optional and is not currently supplied by the LDC. (Future use and value of Topic will require additional guidance on how to define it.)

Conversation_trans – <conversation_trans> is a spanning tag, terminated by **</conversation_trans>**. It is bracketed by a *<utf>* tag and it may contain **<turn>** and **<background>** tags within its span. The tag indicates that the transcript is from a recorded conversation. The attributes associated with each conversation_trans are:

recording_date: Date on which the recording was made.

Turn – <turn> is a spanning tag, terminated by **</turn>**. It spans the transcription associated with a particular speaker turn, and it may contain any combination of background tags, pseudo empty tags and lexical tags within its span, as well as the transcription text. The **<turn>** tag must be contained within the span of either a **<section>** or **<conversation_trans>** tag. The attributes associated with each Segment are:

startTime: The start time of the Segment, measured from the beginning of the recording in seconds.

endTime: The end time of the Segment, measured from the beginning of the recording in seconds.

speaker: The speaker's name.

mode: One of the labels "Spontaneous" or "Planned".

fidelity: One of the labels "High", "Medium" or "Low.

dialect: One of the labels “native” or “nonnative”.

spkrtype: One of the labels “male”, “female", "child", or “unknown".

channel: The channel of the waveform that contains the speech. If omitted, the default channel is assumed to be 1.

Separator – <separator> [or SGML the space character short reference]. May occur only within turns. This tag indicates that a word boundary has occurred. This construct permits intra-word tagging, which is required by some forms of annotations.

State Change tags

Background – <**Background**> is a state change tag that provides information about a particular (single) background signal, specifically regarding the type and level of the signal. This information is synchronized with the transcript by positioning the Background tag at the appropriate point between words in the transcription. (<**Background**> tag locations and times will be positioned at word boundaries so that the word within which the background noise starts or ends will be included in the span of the background noise.) The <**Background**> tag must be contained within the span of either **a <bn_episode_trans>** or **<conversation_trans>**. The attributes associated with each Background tag are:

Time: The time at this point in the transcript, measured from the beginning of the recording in seconds.

Type: One of the labels “Speech”, “Music” or “Other”.

Level: One of the labels “High”, “Low” or “Off”. This attribute indicates the level of the background signal *after* Time. Thus High or Low implies that the signal *starts* at Time, while Off implies that the signal *ends* at Time.

Comment – Comments are annotated using the standard SGML comment tags. The format of an SGML comment tag is "<!-- {misc.text} -->". It spans a free-form text comment by the transcriber, but can not span other SGML tags. Comment can occur anywhere.

Pseudo Bracketing Tags

Pseudo bracketing tags indicate an annotated attribute of a span of words. Since it is common to annotate overlapping regions with different attributes, these tags do not use the SGML hierarchy of begin and end tags, (i.e. <doc> … </doc>). Instead, explicit begin tags and end tags (i.e. <b_doc> …. <e_doc>) are used for each annotated attribute. These tags may occur only within turns.

Foreign Language - <b_foreign> and **<e_foreign>** are pseudo bracketing tags. They are used to mark words spoken in a foreign language. This does not include foreign words that have been incorporated into the target language. The pair of tags span the foreign word(s) and must be contained is bracketed by a **<turn>**. <e_foreign> has no attributes, but <b_foreign> has 1 attribute:

language: Putative language of the spoken words.

Unclear speech - <b_unclear> and **<e_unclear>** are pseudo bracketing tags. They are used to mark unclear speech, where what was said isn’t clear. The tags may be empty or may include a best guess as to what was said. The pair must be contained within the span of a **<turn>**. Neither tag has any attributes.

Overlap – <b_overlap> and **<e_overlap>** are a pseudo bracketing tags. They are used to indicate the presence of simultaneous speech from another foreground speaker. This information is synchronized with the transcript by positioning the pair of overlap tags at the appropriate point in the transcription. The **<b_overlap**> and **<e_overlap>** tags will be positioned between **<separator>** tags so that the all words affected by overlapping speech is indicated. The pair of tags must be contained within the span of a **<turn>**. **<e_overlap>** has no attributes, but **<b_overlap>** has 2:

startTime: The start time of the Overlap, measured from the beginning of the recording in seconds.

endTime: The end time of the Overlap, measured from the beginning of the recording in seconds.

For example:

Speaker A: ... It was a tough game <Overlap S_time=101.222 E_time=102.111> but very exciting </Overlap>

Speaker B: <Overlap S_time=101.230 E_time=102.309> Yes it was </Overlap>

In this example, Speaker B broke into Speaker A's turn. Note that the Overlap times don't coincide exactly because they have been time-aligned to the most inclusive word boundaries for each speaker turn involved in the overlap.

Noscore - <b_noscore> and **<e_noscore>** are pseudo bracketing tags. They are used to explicitly exclude a portion of a transcription from scoring. The pair of tags spans the word(s) to be excluded and must be contained within the span of a **<turn>**. <e_noscore> has no attributes, but <b_noscore> has 3 attributes:

reason: Short free-form text string containing an explanation of why the tagged text has been excluded from scoring. The string must be bounded by double quotes.

startTime: The start time of the excluded portion, measured from the beginning of the recording in seconds.

endTime: The end time of the excluded portion, measured from the beginning of the recording in seconds.

For example:

<b_noscore reason="Mismatch between evaluation index and final transcript" startTime=1710.93 endTtime=1711.71> ... text to be excluded ... <e_noscore>

Speech spoken as an aside - <b_aside> and **<e_aside>** are pseudo bracketing tags. The tags are used to mark speech in conversations that was not directed to the other party in the conversation. The pair must be contained within the span of a **<turn>**. Neither tag has any attributes.

Named Entities – <b_enamex> and **<e_named>** are a pseudo bracketing tags. They are used to annotate named entities for the Hub Evaluation Information extraction spokes. (Consult the Hub-4 Named Entity Task Definition for a complete description of the named entity tags and their use.) These tags are attached to words, (i.e. within **<separator>** tags), but may be placed within word orthographies. The pair of tags must be contained within the span of a **<turn>**. **<e_enamex>** has no attributes, but **<b_enamex>** has 3:

type: The type of named entity can be either an ORGANIZATION, PERSON or LOCATION. The tagging specifications permit any combination of attributes to be joined using the 'OR' connector '|'.

status: When the proper tagging of a named entity is unclear, the status attribute is used to indicate that the markup is optional. The only value of this attribute is "OPT".

alt: The alt attribute is used when the tagged string contains one or more sub-strings that should also be considered correct for the scoring of system responses.

Temporal expressions – <b_timex> and **<e_timex>** are pseudo bracketing tags. They are used to annotate temporal expressions for the Hub Evaluation Information Extraction spokes. (Consult the Hub-4 Named Entity Task Definition for a complete description of the temporal expression tags and their use.) These tags can are attached to words, (i.e. within **<separator>** tags), but may be placed within word orthographies. The pair of tags must be contained within the span of a **<turn>**. **<e_timex>** has no attributes, but **<b_timex>** has 3:

type: The type of temporal expression can be either a DATE or TIME. The tagging specifications permit any combination of attributes to be joined using the 'OR' connector '|'.

status: When the proper tagging of a temporal exporession is unclear, the status attribute is used to indicate that the markup is optional. The only value of this attribute is "OPT".

alt: The alt attribute will be used when the tagged string contains one or more sub-strings that should also be considered correct for the scoring of system responses.

Number expressions – <b_numex> and **<e_numex>** are a pseudo bracketing tags. They are used to annotate number expressions for the Hub Evcaluation Information Extraction spokes. (Consult the Hub-4 Named Entity Task Definition for a complete description of the numeric expression tags and their use.) These tags can are attached to words, (i.e. within **<separator>** tags), but may be placed within word orthographies. The pair of tags must be contained within the span of a **<turn>**. **<e_numex>** has no attributes, but **<b_numex>** has 3:

type: The type of a number expression can be either MONEY or PERCENT. The tagging specifications permit any combination of attributes to be joined using the 'OR' connector '|'.

status: When the proper tagging of a number expression is unclear, the status attribute is used to indicate that the markup is optional. The only value of this attribute is "OPT".

alt: The alt attribute will be used when the tagged string contains one or more strings that should be considered correct for the scoring of system responses.

Lexical Tags

Time – <time> is a non-spanning tag that provides transcription timing information within a Turn. It is positioned within the transcription between **<separator>** tags and gives a time mark at that point. The **<time>** tag must be contained within the span of a **<turn>**. Time tags are a side-effect of the transcription process and are included for synchronization purposes. Time has a single attribute, namely Sec:

Sec: The time at this point in the transcript, measured from the beginning of the recording in seconds.

Word time – <wtime> is a non-spanning tag that provides the time of occurrence of a single word. It is positioned before the word's text, to which it is attributed, but does not have a **<separator>** on both sides of it. (e.g., "<separator><wtime startTime=44.3 endTime=45.1>this<separator><wtime startTime=45.2 endTime=46.3>is<separator>"). The **<wtime>** tag must be contained within the span of a **<turn>**. Wtime has 4 attributes:

startTime: The start time of the Overlap, measured from the beginning of the recording in seconds.

endTime: The end time of the Overlap, measured from the beginning of the recording in seconds.

clust: An optional attribute containing the adaptation cluster used during recognition.

conf: An optional attribute indicating the confidence of the recognized word.

Contractions – <contraction> is a non-spanning tag that indicates the intended expansion of a contraction. It is positioned before the word's text, to which it is attributed, but does not have a **<separator>** on both sides of it. (e.g., "<contraction e_form="[that=>that]['s=>is]">that's"). The **<contraction>** tag must be contained within the span of a **<turn>**. Contraction tags have 1 attribute:

e_form: The e_form attribute indicates the proper expansion of the transcribed contraction. The notation of the e_form attribute is described by the following BNF grammar:

e_form :== <contraction_part><contraction_part>+

<contraction_part> :== "[" <transcribed_text> "=>" <expanded_text> "]"

<transcribed_text> :== is the orthography as spoken (with contractions).

<expanded_text> :== is the expanded orthography (uncontracted form).

For example, if the speaker says: "Don't stop". The proper expansion of "don't" is "do not", so the SGML markup is '<contraction e_form="[do=>do][n't=>not]">don't do'.

Word fragments – <fragment> is a non-spanning tag that indicates that the speaker spoke only a partial word. It is positioned at the point in the word's text where the speaker stopped production (e.g., "good**<separator>**b**<fragment><separator>**bye"). The **<fragment>** tag must be contained within the span of a **<turn>** and has no attributes.

Period – <period> and the SGML short reference "." indicates a period, the syntactic sentence end. It is positioned at the end of a word's text. The **<period>** tag must be contained within the span of a **<turn>** and has no attributes.

Question mark – <qmark> and the SGML short reference "?" indicates a period, the syntactic interrogative sentence end. The tag is positioned at the end of a word's text. The **<qmark>** tag must be contained within the span of a **<turn>** and has no attributes.

Comma – <comma> and the SGML short reference "," indicates a period, the syntactic intra-sentence pause. The tag is positioned at the end of a word's text. The **<comma>** tag must be contained within the span of a **<turn>** and has no attributes.

Mispronounced word– <mispronounced> and the SGML short reference "+" indicates that a word was mispronounced by the speaker. The tag is positioned at the beginning of a word's text and applies only to that word. The **<mispronounced>** tag must be contained within the span of a **<turn>** and has no attributes.

Misspelled word – <misspelling> and the SGML short reference "@" indicates that the transcriber was unsure of the spelling. The tag is positioned at the beginning of a word's text and applies only to that word. The **<misspelling>** tag must be contained within the span of a **<turn>** and has no attributes.

Acronyms – <acronym> and the SGML short reference "_" indicates that a word is part of a spelled acronym. The tag is positioned at the beginning of each letter of the acronym and applies only to the letter after the tag. The **<mispronounced>** tag must be contained within the span of a **<turn>** and has no attributes. An example transcription of the acronym AT&T would be _A _T and _T.

Proper Names – <pname> and the SGML short reference "^" indicates that a word is a proper name. The tag is positioned at the beginning of a word's text and applies only to that word The **<pname>** tag must be contained within the span of a **<turn>** and has no attributes.

Idiosyncratic word– <idiosyncratic> and the SGML short reference "*" indicates that an idiosyncratic word. The tag is positioned at the beginning of a word's text and applies only to that word. The **<idiosyncratic>** tag must be contained within the span of a **<turn>** and has no attributes.

Non-Lexeme word– <nonlexeme> and the SGML short reference "%" indicates that pause filler that carries no semantic meaning. English examples are: "um", "uh", "eh" etc. The tag is positioned at the beginning of a word's text and applies only to that word. The **<nonlexeme>** tag must be contained within the span of a **<turn>** and has no attributes.

Non-speech event – <nonspeech> and the SGML short reference "{" indicates that a momentary noise produced by a speaker, that is not considered a lexeme. The text of the word indicates the type of non-speech event. Typical non-speech events are "breath", "cough", etc. The tag is positioned at the beginning of a word's text and applies only to that word. The **<nonspeech>** tag must be contained within the span of a **<turn>** and has no attributes.

Brief acoustic event – <idiosyncratic> and the SGML short reference "*" indicates that a breif noise, produced by something other than a speaker has occurred. Typical brief acoustic noises include "doorslam", "ring", "static", etc. The tag is positioned at the beginning of a word's text and applies only to that word. The **<idiosyncratic>** tag must be contained within the span of a **<turn>** and has no attributes.