Transcription conventions

Ron Artstein
artstein [at] ict.usc.edu

March 28, 2012

Purpose

1. This manual lays out the conventions used in transcribing audio files for the various Natural Language projects at ICT.

2. The main purpose of the transcription performed is to provide data for the various modules of the project, for the following tasks.

   (a) Training domain-specific speech recognizers (ASR).

   (b) Creating Natural Language Understanding (NLU) resources, including:

   - Utterance-response mappings for classifier training
   - Dialogue-act and semantic frame coding

   Linguistic analysis of the speech and transcriptions is a secondary purpose.

3. The transcription conventions are intended to facilitate the main tasks. They therefore need to be as close as possible to format used for ASR output and NLU input (which should be the same, since ASR feeds into NLU). This format is rather close to the conventions of written English, which has the added benefit that transcriptions are easy for humans to read.

4. The transcription conventions are not intended for linguistic, phonetic or acoustic analysis beyond what is needed by the processing modules. They therefore do not record phonetic detail, and are synchronized with the speech signal only at coarse granularity (at the level of utterances, typically lasting a few seconds).

5. These conventions are driven by the needs of the processing modules. If and when these needs change, the conventions will change as well.

Software

6. Audio recordings come in two forms.
   (a) Short, utterance-length sound files typically come from a user interacting with a virtual character through an implemented system.
   (b) Long, dialogue-length recordings typically come from audio and video recorders, such as used in role plays.
7. Transcription of the short sound files is done by listening to the file in a sound player and transcribing its contents in a text file. Audio signals are associated with the corresponding transcription by noting the audio file identifier next to the transcribed text. There is no temporal alignment of the transcript and the audio signal.
8. Transcription of dialogue recordings is done through a transcription tool; currently we use Transcriber for audio files and Elan for video files. The tools allows dividing the transcript into segments which are anchored to the speech signal at specific time points, and identifying the speaker of each segment. Transcriber also allows grouping multiple utterances into turns. ELAN allows overlapping segments.

Segmentation

9. Recordings transcribed in a tool such as Transcriber or Elan are segmented into Turns and utterances, using the tool.
10. Short audio files transcribed in a text file, such as the ones that come from interaction with a virtual character, are not segmented (even if they contain multiple utterances or distinct speakers).

Turns

11. Turns are used to mark speakers in continuous dialogue transcribed with Transcriber.
12. A set of utterances by one dialogue participant without interruption from other participants constitutes a turn.
13. Pauses do not interrupt a turn: if a speaker takes a long pause and then resumes speech, this is considered to be the same turn.
14. A long pause between speakers is marked as a turn with no speaker.
15. When two speakers speak together, a turn is created for the overlapping period and marked as such. The order of speakers in the turn matches the transcription – if it says, for example, “Alice + Bill”, then line 1 is the transcription of Alice and line 2 is the transcription of Bill.

Utterances

16. Speech is segmented into utterances in Transcriber and ELAN.

17. The definition of an utterance is driven by the needs of the processing modules:

    (a) ASR needs short segments (a few seconds) of connected speech, preferably with clear boundaries at either end.

    (b) NLU needs short segments (a few words) which convey a single idea.

    The overall criterion for segmentation is prosodic, but semantic boundaries should also be identified when possible.

18. Utterances are separated by major intonational breaks, and pauses.

19. Distinct pauses are generally marked as utterance breaks, but short pauses can be included within an utterance if they only separate a single word or hesitation sequence.

20. A pause of more than 1–2 seconds is marked as a separate utterance in Transcriber; shorter pauses between utterances are attached to the preceding or following utterance.

21. A long string of speech without a pause or intonational break is broken into two utterances if a boundary between words can be identified, and the parts before and after the boundary convey distinct meanings.

Transcription

Orthography

22. Words are transcribed using standard American English spelling conventions, except as noted below.

23. Transcriptions use only lowercase letters.

24. All words are spelled out without abbreviation: *doctor perez*, not ~~*dr. perez*~~.

25. Letter names are transcribed as a single letter separated by spaces: *a b*.

26. Numbers:

    (a) All numbers are spelled out: *one*, *fifteen*, *twenty five*.

    (b) Numbers are transcribed as pronounced: *one hundred and first* is distinct from *one o first*.

    (c) We do not use digits for transcriptions.

27. Acronyms

    (a) Acronyms which are pronounced letter by letter are connected by underscores: *a_p_c* (armed personnel carrier).

    (b) Acronyms which are pronounced as a word are transcribed without underscores: *jtac* (joint terminal attack controller).

    (c) In acronyms consisting of a single letter plus a number, the letter is connected by an underscore to the first word in the number: *t_seventy two* (Soviet tank).

    (d) In acronyms consisting of multiple letters plus a number, the letters are connected by underscores to each other but not to the number: *a_h sixty four* (US attack helicopter – Apache).

    (e) Plurals of acronyms are written with an *s* not separated by any punctuation: *a_p_cs*, *jtacs*, *f_sixteens*, *a_h sixty fours* .

    (f) Sequences of letters that do not form a legitimate acronym are not connected. For example, if a speaker starts to say *b_t_r* (Soviet armored personnel carrier) but corrects mid-word to *b_m_p* (Soviet infantry fighting vehicle), the appropriate transcription is *b t b_m_p*. The letters that do not form a full acronym are left unconnected, in order to avoid contaminating the speech models with non-existent acronyms.

28. Military call signs are transcribed as pronounced: *alpha*.

29. Contractions are transcribed in the conventional way: *we'll*, *can't*. This also applies to non-standard contractions which have a conventional spelling: *ain't*, *wanna*, *'cause*.

    – The apostrophe character for contractions in English is the ASCII character (straight quote, U+0027), not any "smart" or curly character. The reason is that some of the software we use is not aware of character encodings or multibyte characters.

30. Minor pronunciation deviations from standard are ignored: *living*, not ~~*livin'*~~. The purpose of this convention is to train the speech models to associate such deviant pronunciations with the intended word.

31. Generally, words are transcribed using unaccented Latin-script letters only.

    (a) Accented characters in English words are converted to their unaccented counterparts: *cafe*, not ~~*café*~~.

    (b) Words in foreign languages that use the Latin script are transcribed using native language conventions, except that accented characters are converted into their ASCII counterparts: *como estas*, not ~~*cómo estás*~~.

(c) Words in foreign languages that do not use the Latin script are transcribed using an accepted transliteration into unaccented Latin-script letters.

The reason for the above conventions is that some of the software we use is not aware of character encodings or multibyte characters, and we have no general way of indicating the character encoding of transcription text files. Specific projects may deviate from the above conventions – for example, Pashto in the CHAOS project is transcribed using native Pashto (Arabic-script) characters. In such cases a character encoding should be agreed upon; the recommended encoding is UTF-8.

32. Generally, we do not use punctuation in our transcriptions.

Disfluencies

33. All repetitions and hesitations are transcribed as pronounced: *i i i think*, *eh*, *um*.

34. If a person misspeaks (e.g. says "happen" instead of "happened"), write the word actually pronounced. The reason for this convention is that such problems are better handled by the NLU component, so we want to train the ASR to give the actual word pronounced and to train the NLU to handle such input.

35. If speech is cut-off, put down the complete intended word, followed by a comment with the part that was actually pronounced in angle brackets: *people <peop>*. The comment is just for human readers; the reason for transcribing a whole word is to avoid confusing the processing modules by training them on non-words.

36. Unrecognizable words are written as *xxx*.

Non-speech

37. When a speaker produces an audible non-speech sound, put it down as a comment between angle brackets: `<cough>`, `<laughter>`, `<sigh>`.

38. Other comments also go between angled brackets (for example speaker identification in short audio files). The precise placement (at the beginning or end of a line) varies by project.