Recorded Debating Dataset: Transcription Guidelines

March 11, 2018

This document details the guidelines and notation used by the transcribers when creating the dataset described in Mirkin et al., 2018. We detail them in two parts: those guidelines that refer to the format of the transcript and those that are meant to ensure its accuracy with respect to the audio.

Format

- **Case**: Do not capitalize letters – not at the beginning of sentences, nor for names and abbreviations.
- **Initialisms**: separate the letters with an underscore, as in: "u_s_a", "i_b_m".
- **Numbers**: Numbers should we written as words, not digits, e.g. "nineteen seventy two" and not "1972".
- **Punctuation**: Aim to add commas (","), periods (".") and question marks ("?") when natural, but don't agonize about it. Please use double quotes (") to delimit direct speech (do not use single quotes). You may also use exclamation marks ("!"), and colons (":") when necessary.
  - For the above-mentioned punctuation marks, each one should be separated from the word it follows by a space, e.g. "lies , damned lies , and statistics ."
  - You may use hyphens in compound words, like "sugar-free"
  - Apostrophes may be used in contractions (e.g. don't, can't) and to show possession (e.g. John's dog).
  - Do not add any other punctuation marks.

Accuracy

Transcribe what the speaker says as accurately as possible, and do not make any corrections:

- **Contractions and other abbreviations**: Do not use short forms such as "we'll" instead of "we will" or "etc." instead of "etcetera" if the speaker used the longer form.

- **Handling disfluencies**:
  - Cut-off words, stutters, repetitions: e.g. "sec- second one". Write them as they sound, adding a hyphen followed by a space after the word that is incomplete.
  - Wrong word uttered by the speaker, as in "menial label labor": include the wrong word. In this case the speaker said "menial label" and corrected to "labor", thus he said "menial label labor" and that's what you should put in the transcription.
  - Repeated words, as in "this is this is" – include the repeats.
  - Filler words: The text may include the tag "%hes". For uniformity sake, please use "uh", "um" and '"ah" to represent hesitation sounds, and if a filler word or repetition is unclear, leave or use the generic "%hes" tag. Try your best to represent each filler you hear separately.
  - Incoherent/unintelligible speech: Sometimes you will not be able to understand what was said, even with careful listening. Represent the unintelligible utterance as "
  - Mispronounced words: Please use the notation {pronounced word/real word}. E.g., if the speaker says "infair" instead of "unfair", transcribe this as {infair/unfair}. This notation should **not** be used to correct grammar or improve upon the debater's speech.
  - "Broken" words: E.g., if the speaker says, "im (pause/break) portant", please transcribe this as {im--portant/important}: use two hyphens.
  - Words whose beginnings are inaudible: Please use the notation {pronounced word/real word}. E.g., if the speaker says "scuse me" instead of "excuse me", transcribe this as "{scuse/excuse} me".
- **Spell-checking**: Please make sure to run spell-checking before submitting your transcript.

References

[1] Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, Noam Slonim. A Recorded Debating Dataset. LREC 2018