Transcription Conventions for the SRI Speech-based Collaborative Learning Corpus (SBCLC)

(1) Overview

In support of the project "Speech-Based Learning Analytics for Collaboration", small groups of middle school students were audio recorded as they worked together to solve math problems. Each group consisted of three students, each recorded via a head-mounted microphone. The interaction was also recorded by a nearby tabletop stereo array microphone. All audio recordings are sample synchronous. In phase 1 of the project, only two groups interacted in a single room at any one time.

The SRI Speech-Based Collaborative Learning Corpus: Phase 1 (SBCLC) contains audio recordings of 80 Phase 1 interactions, henceforth referred to as "sessions." Thirteen sessions have been annotated to indicate which regions of the audio contain speech. These sessions have also been transcribed orthographically.

The transcriptions can serve as the raw material for computational analysis and processing. Every effort has been made to use consistent transcription conventions and to correct errors. This document presents with examples the transcription conventions.

(2) Transcription file format

The transcription for each session is contained in a separate text file. The names of the files follow the general SBCLC naming conventions and include the file extension "trans":

`SBCLC_`*`observation_problemset_group`*`.trans`

SBCLC	corpus name
observation	code for school visit
problemset	set of problems the students worked on (1 or 2)
group	group name (A or B)

Each line of a transcription file corresponds to a single region of continuous speech. Each line consists of a unique code followed by the orthographic transcription. The code for each speech region is the base name of the corresponding text file plus the speaker's microphone number, the start time, and the end time. The time stamps are in centiseconds and formatted to contain 7 digits. For example:

```
SBCLC_23_2_B_1_021323_021382 would that work?
```

The above example line indicates that during school visit 23 and while working on problem set 2, the speaker from group B wearing microphone 1 said "would that work?" during the time from 213.23 seconds to 213.82 seconds.

(3) Segmentation of audio

The audio from each head-mounted microphone is annotated to mark which regions contain speech from the student wearing the corresponding microphone. The head-mounted microphones are noise cancelling and typically the voice of the student wearing the microphone is the loudest. However, speech from other group members is frequently audible. Occasionally noise from the other group in the room or general school noise is audible.

The audio is not broken into utterances based on meaning or completeness, but rather into continuous regions of speech. In other words, a single meaningful complete sentence may be broken up into several speech regions (and thus occupy several lines in the transcription file) if there were pauses during the sentence. This was done for 2 reasons: (1) to capture the natural discourse style of these interaction - the students frequently speak in short spurts of speech that are often incomplete or interrupted and (2) to approximate the output of an automatic speech activity detection system that breaks up speech based on pauses.

Instances of speech and non-speech vocalization are annotated. This includes vocalizations as short as 100 milliseconds. Only non-speech vocalizations that conveyed some sort of discourse-related meaning are marked. For example, regular breathing and mouth noises are not annotated, but loud sighs or “tsk” sounds that express frustration are annotated.

(4) Orthographic transcription of speech

The goal of this transcription is to produce accurate word-for-word transcripts of the dialogs using consistent conventions. Grammatical errors are not corrected. Standard American spelling is used and non-standard spelling based on pronunciation is avoided.

All disfluent speech, including hesitations, repeated words, and partial words, is transcribed.

The following sub-sections lay out the spelling and punctuation conventions.

Capitalization

The first letter of a name is capitalized. Acronyms and letter names are capitalized. The pronoun “I” is written in lower case. The first word of a sentence is written in lower case, if it is not a name, acronym, or letter name.

Punctuation

The use of punctuation is limited to periods, commas, apostrophes, question marks, and exclamation marks. Quotation marks, colons, and semicolons are not used.

A period marks the end of a statement and a question mark marks the end of a question:

```
i think the answer is seven.
why do you think that?
```

When a speaker says a non-question in very excited or agitated manner, an exclamation point is used:

```
oh shoot!
```

Commas are used mainly to mark verbal deletions, when the speaker abandons the course of the utterance and either stops speaking or switches to a new utterance. Additional commas may be used for readability. For example:

```
we know this because,
since the character,
his speed does not, no wait no, his position does not change.
```

If the speaker is interrupted by another and does not finish the utterance, two dashes (`--`) are placed at the end of the partial phrase:

```
A: three X plus eight plus four that's --
B:                                    no that's not right
```

If the speaker is interrupted and pauses before continuing the utterance, the continuation starts with two dashes (`--`). This annotation can be used to join the speech regions into one utterance. For example:

```
A: no it starts --
B:         what?
A:               -- at like thirty one.
```

If the speaker abandons the current utterance and immediately starts a new one because of something another speaker said at the same time, two dashes (`--`) are placed to show this break. This may not be audible in the audio from a single head-mounted microphone. For example:

```
A: to the -- yeah because mine was wrong.
B:   three two,
```

If the speaker's utterance spans more than one speech regions (because it contains pauses), two periods (`..`) are placed before the pause and after the pause. This annotation can be used to join transcriptions into more coherent utterances. For example:

```
know this because his speed ..
                          <pause>
                             .. does not change.
```

Numbers

Numbers are written as complete words without hyphens. The digit "0" when pronounced as a letter is transcribed as "oh." For example:

```
twenty eight divided by two is fourteen.
```

Contractions

Standard contractions are used if the contracted pronunciation is used. The following contractions are allowed:

```
i'm, i've, i'll, i'd
you're, you've, you'll, you'd
he's, he'll, he'd
she's, she'll, she'd
it's, it'll, it'd
that's, that'll, that'd
we're, we've, we'll, we'd
they're, they've, they'll, they'd
who's, who've, who'll, who'd
how'd, how'll
would've, should've, could've
aren't, isn't, wasn't, weren't
haven't, hasn't, hadn't,
don't, doesn't, didn't
won't, wouldn't, can't, couldn't, shouldn't
```

contractions of a noun and "is" or "has" such as "`my brother's nice`" "`my sister's left`"

Hyphenation

Standard American English conventions are used for hyphenated words. Exception: compound numerals are not hyphenated.

Abbreviations

The full spelling of words is used, even for words that are commonly abbreviated like St. (for Saint or Street), Dr. (for Doctor or Drive), etc. (for et cetera).

Acronyms

Acronyms that are pronounced as a single word (NATO, AIDS, NASDAC) are distinguished from acronyms that are pronounced letter by letter (UN, USA, EU, DVD, ID). If it is pronounced as a single word, the acronym is transcribed as a single word using capital letters and no periods or spaces:

```
i work for NATO.
```

If each letter is pronounced, capital letters joined with underscores are used:

```
i bought the D_V_D.
```

Plural acronyms are written without an apostrophe. An apostrophe is used only with possessive acronyms or contractions. For example:

```
i bought five D_V_Ds yesterday.     (plural noun)
the D_V_D's scratched.              (contraction with "is")
```

```
the D_V_D's label is missing.       (possessive)
```

When an acronym contains a digit, the full form of the digit is used and joined with an underscore:

```
MP3  M_P_three
3D   three_D
```

Letters and spelling

When the speaker uses the name of a single letter or spells out a word, letters are transcribed separately using capital letters:

```
so segment B represents the character.
i need more vitamin C.
i got an A on my test.
dog is spelled D O G.
it included everything from A to Z.
his blood is B positive.
```

Short forms

A select set of spellings is used to represent shortened or reduced pronunciations.

When the speaker pronounces a shortened version of a word (often colloquial or slang), a shortened (possibly non-standard) spelling is used.

```
i need more info.
what's your prob?
the food is really delish.
he's cray.
whatev.
```

The lexicalized reduced form of "going to" indicating future action is transcribed as "gonna." If the speaker does not use the fully reduced form (even when indicating a future action), the spelling "going to" is used.

The common shortened pronunciation of "because" is transcribed as 'cause.

Other reduced pronunciations are transcribed with their corresponding full spelling. The following examples list common spellings for common reductions followed by the full spellings that are used.

wanna	`want to`
gotta	`got to`
shoulda	`should have`
get 'em	`get them`
how ya doin'?	`how you doing?`

Interjections

All interjections are transcribed. The following spellings are used:

```
yeah, yep, yup
nope, nah
oh, ooh
uh-huh, mm-hm (agreement or acknowledgment)
uh-uh, nuh-uh (disagreement)
uh-oh, oops, whoops
hm, mm
ach, duh, whew, whoa, oof, jeez
aha, ha
yay, woo-hoo
ew, yuck, euch
huh
yikes
aw
ugh
shh
```

Nonsense words

Nonsense words or syllables are transcribed represent the pronunciation using standard spelling as much as possible. For example:

```
clicker-doodle-doo
boom chaka laka boom
```

Filled pauses or hesitations

All filled pauses (hesitations) are transcribed. Filled pauses that consist of just a vowel are transcribed as `uh`. Filled pauses that consist of a vowel followed by a nasal are transcribed as `um`.

Unintelligible speech

Regions of speech that the transcriber recognizes as speech but cannot understand are transcribed with double parentheses.

```
the (()) is getting closer.
```

When the transcriber is uncertain about a region of speech, it is transcribed as best understood but inside double parentheses.

```
i think this is ((what was said)).
```

Fragments or partial words

When the speaker breaks off in the middle of the word or stutters, as much of the word as can be understood is transcribed using standard spelling and followed by a single dash.

```
sh- sh- should we do twenty and four or twenty and six?
```

When the speaker is interrupted and stops speaking in the middle of word, the partial word is transcribed with a single dash, followed by a space and two dashes (`--`) to indicate the interruption.

```
so half of i- --
```

Mispronounced words

When a speaker mispronounces a word due to a speech error, the word is transcribed using standard spelling but preceded by a plus sign (+). This indicates that there is a speech error and it is not the typical pronunciation of this word for this speaker. For example, if a speaker in error says: *I bought flive flowers*, it would be transcribed as: `i bought +five flowers.`

Overlapping speech or crosstalk

There is a lot of overlapping speech in this corpus, but it is not marked directly in the transcriptions. The interruption marker (`--`) indicates that there was speech from another speaker that caused the primary speaker to stop speaking. Overlapping speech can also be deduced from the time stamps.

Non-English

When the speaker speaks in a language other than English, the tag `[non-english]` is used to mark that speech.

(5) Transcription of non-speech

All occurrences of laughter and yawning are annotated. If laughter and yawning do not overlap with speech they are transcribed with the tags `[lau]` and `[yawn]`, resp.

In general, occurrences of non-speech such as normal breathing, coughing, lip smacking, and mouth clicking are not annotated. However, if non-speech is used to express an emotion or has a discourse function, then such non-speech is annotated. For example, a loud exhalation expressing frustration or fatigue is marked with the tag `[sigh]`. An exhalation expressing derision is marked with the tag `[scoff]`. A sharp inhalation to express surprise is marked with the tag `[gasp]`. Loud click noises (commonly spelled “tsk”) that express dissatisfaction or annoyance are marked with the tag `[mouth]`.

Any other non-speech vocalization that is not speech but is salient and plays some sort of role in the social interaction (beat-boxing, humming, whistling) is marked with the tag `[nonspeech]`.

(6) Affected speech

Frequently speakers will speak while laughing, yawning, sighing, whispering, or singing. This is annotated with start and end tags surrounding the words that are affected. The following tags are used:

```
[lau] … [/lau]
[yawn] … [/yawn]
```

```
[sigh] … [/sigh]
[whi] … [/whi]
[sing] … [/sing]
```

For example:

```
[lau] we're racking our brains [/lau] he gets it like that.
[yawn] oh my god this is, [/yawn]
[sigh] they're all the same problem as this. [/sigh]
[whi] wait [/whi] what's this?
[sing] oh, say can you see, by the dawn, [/sing]
```

Unintelligible whispering or singing is marked with the single tag `[whi]` or `[sing]`, resp.

(7) Personally identifying information

The students recorded in this corpus frequently call each other by first name. There are a few instances when students refer to other students in the school by full names. First names are transcribed. Middle names and last names are replaced with the tags `[middlename]` and `[lastname]`, resp.

```
it's not Calvin [lastname].
Neil [middlename] [lastname].
Omar's l- middle name is [middlename].
```

(8) Transcriber comments

Short comments next to the transcriptions are included if the transcriber thinks they could be informative. Comments are contained in curly brackets to separate them from actual speech. Be careful not to merge these with the transcribed speech when doing any automatic processing on the transcripts. Some examples:

```
{fake British accent}
{talking to researcher}
{sounds like gibberish}
{voice over intercom}
```