Reformatting Transcripts for LIWC 2015

Introduction to LIWC:

The files you are transcribing will be analyzed with a computerized text analysis program called the Linguistic Inquiry and Word Count, or LIWC. LIWC is a simple word counting program. It has 80+ internal dictionaries, or word lists. These dictionaries contain words and stems (e.g., *play** captures *plays*, *played*, and *playing*) that represent all frequently used words in various linguistic (e.g., pronouns), psychological (e.g., positive emotion references), and thematic (e.g., words referring to leisure or work) categories. LIWC output tells you the percentage of words in a given text that fall into one or more of its internal categories.

LIWC cannot correctly categorize and count a word if it's not in its internal dictionaries, and it can't tell the difference between a word that's used in two different ways without help. That's why we have to spell-check all files before analyzing them, and why you'll need to learn a few of the tricks below to help LIWC understand what the people in your audio files are saying. When in doubt about whether LIWC will need help recognizing a word, consult the LIWCenome or this manual.

Operator's manual:
https://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_OperatorManual.pdf

Psychometric manual:
https://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

Transcribing and cleaning text so that LIWC can understand it:

1) **Spelling.** Correct all spelling errors using spell check. It is best to use standard United States spelling (although the standard default dictionary also contains most British English spellings as well).
   - Some of the changes requested in the following steps may show up and be changed at this time. Make sure that whatever option that is highlighted as the "Change" is the best option.
   - Fix what might be transcriber errors **without changing the content**
   - **Fix** things like typos or punctuation:
     i. A possessive "its" does NOT have an apostrophe. The contraction "it is" does → "it's". *This is one of the few apostrophes that matters; things like *dont* and *cant* don't need to be changed. See the contractions section.
     ii. Use discretion about changing "?" It could be a statement with a voice tone like a question
     iii. Blatant typos like misspelled words
   - **Don't fix** things like slang or sentence structure:
     i. Long sentences/fragments
     ii. Repeated words "he was just, just not very nice."

2) **Abbreviations.** Meaningful abbreviations should be spelled out. "Jan" should be January. More obscure abbreviations or acronyms, such as "AT&T", can remain as such unless you have reason to want the term to be expanded and counted as four separate words: "American Telephone and Telegraph".

3) **Contractions.** Common verb contractions are in the dictionary and do not need to be changed. These include: don't, won't, isn't, shouldn't, can't, couldn't, I'm, I'll, I'd, we're, we'd, you're, he's, it's, etc.

   Possessive vs. is. Apostrophe + s will be counted as a possessive noun unless you change it: "Sally's shoes will be counted the same as "Sally's going to the store." In the second case, change "Sally's" to "Sally is." Use the "Find feature" to search `s. If unsure about whether a common contraction (e.g., let's) should be spelled out, check the LIWCenome. If the word appear in LIWCenome, leave it as a contraction – the program will count it as such.

   ***It's*.** This word is included in LIWCenome as a present tense contraction for "It is", so anytime the word "it's" is being used in this way, don't change it. However, if it should be "it has" (past tense) you should write it out. If it is being incorrectly used as a possessive word, the apostrophe should be removed to read "its".

   Apostrophe + d. Words ending in apostrophe + d are often ambiguous. For example, "I'd" means "I would" in "I'd rather not" and "I had" in "I'd never tried eel before", etc. If you can tell from the context, change "'d" to would, had, or whatever else it might mean. Use the "Find feature" to search `d.

4) **End of sentence punctuation.** The Words per sentence (WPS) category is based on the number of times that end-of-sentence markers are detected. These include all periods (.), question marks, and exclamation points. One potential problem is that common abbreviations (such as "Dr.", "Ms.", "U.S.A.", "D.O.A.") will be counted as multiple sentences unless the periods are removed.

   Be careful that the removal of the periods doesn't make a new word. For example, the United States, or "U.S.", becomes "US" (1st person plural pronoun) when the periods are removed. In this case, change it to "USA".

   If ellipses (…) are found at the end of a turn, read the context to determine whether to end it with a period, comma, or question mark. If the speaker finishes what they were saying on their next turn, change the … to a comma. If they do not finish what they were saying, use a period. If they are trailing off on a question, use a question mark.

   Sometimes transcribers forget to add punctuation at the end of a turn. Make sure there is appropriate punctuation (period, comma, or question mark) at the end of each turn.

   If a number has a decimal point, change the decimal to a comma. For example, $12.36 → $12,36

5) **Time markers.** Writing out times (e.g., 6 a.m. or 7:30 p.m.) can also be a problem. Because "a.m." without the periods is a verb, "am", change time to 6am or 7:30pm.

6) **Hyphens.** When words start or end with hyphens, they are read by LIWC2015 as part of the word. LIWC2015, for example, lists "chit-chat" as a meaningful word in one of its dictionaries. In cases of hyphenated phrases such as "this-or-that" LIWC2015 will count it as three separate words since "this-or-that" is not in the dictionary.

7) **Common shorthand (and chat transcript) problems.** Below are list common forms of shorthand or alternative spellings. Search for each word in each transcript and replace as needed. Make notes if some are missing that should be changed and added to this list.

Typed entry	***Change to…***
w/	with
b/ or b/w	between
&	and
bc or b/c	because
‘cause* or cuz	because
and/or	and – or
‘an or ‘n	and
mos	months
@	at
All right	Alright
**in' (e.g., eatin')	**ing (e.g., eating)

*Search for "cause" with no apostrophe and read context to determine if they mean "because"

8) **Nonfluencies.** Hm, hmm, uh, uhh, uhm, um, umm, and er are part of the nonfluency dictionary. Other forms will generally not be caught (e.g., ooooh should be changed to um if used as a nonfluency).

   Stuttering can be accommodated by altering the stuttering part of a phrase to a nonfluency marker. For example, "The, the bo-, the boat went into the water" could be changed to "Uh, the boat went into the water." The transcriber will have to decide how many uh's would be appropriate.

   Uh-uh and uh-huh should be changed to "no" and "yes". Huh? should be changed to "what?" Or, if you are very, very proper, to "Excuse me madam, I didn't quite catch what you said."

9) **Transcribers' comments.** LIWC2015 is designed only for spoken language. Transcribers often insert remarks, such as [subject laughs], [shaky voice], [whispers]. We recommend removing these.
   - Look through the transcript to identify how that transcriber made notes for "laughing"/"laughter", "silence", "phone rings" and similar notes. Some transcribers put them in brackets like this [ ] while other may have used < >, ( ), or some other type of notation.
   - After you have identified how the transcriber made these notations, search for those symbols using the "Find" feature in Word.
   - Using careful discretion, one-by-one, evaluate each notation. If it is not speech, delete it (including the examples above).
   - Most of them should be deleted, but be cautious as some transcribers may have used these brackets for important speech or content. For inaudible words, see note below.
   - Fix any punctuation errors these changes might have made.

10) **Inaudible words.** Occasionally, the transcriber cannot understand a word or passage. Rather than writing [can't understand word] or [?], the transcriber should put a nonsense word, such as "xxxx" in its place. LIWC2015 will count the xxxx as a spoken word but not assign it to a dictionary. For up to three inaudible words, use the xxxx method. For entire turns, insert a single xxxx and move to the next line or turn.