**Text Toolkit Notes:**

**Adrienne:** Great work! I hope you enjoyed learning about dictionaries and natural language processing. The main thing that needs to be done is modifying the perceptual distinctiveness code to use the Lexicon found in Balota et. al, 2017; the rest can be modified after Holly gives her input. But again, very well done!

**Measure Issues:**

1. **Speech Acts (Assertive, Commissive, Directives):**

* The database wants to use REGEX to find speech acts, but the actual REGEX wasn’t shared.
* I have a ML model I trained last year that we could use instead. Another option is to train a new model using human coding, though that might be more costly and may not match the original method.

1. **Subject User vs Project:**

* This is a bit complicated since the cited paper’s context is a bit different.
* The current code records most common subject (i.e. if the brand is the subject 5 times whereas the consumer is the subject 4 times, then the brand will be recorded since 5 > 4). Perhaps a ratio would be better?

1. **Message Balance:**

* There is a specific package linked in the database – but it is not coded that way in the folder.
* The linked function also doesn’t return contrast and concession scores; it just returns relationships. Maybe how it already is coded is better since it is closer to the definition’s intent?

**Coding Issues:**

1. **Lexical Diversity (Type Token Ratio):**

* The ratio is the number of unique words which appear in a text by the number of total words. Since the ratio can be misleading for shorter texts, the paper recommends normalising the score based on predefined text lengths (e.g. the number of unique words every 10 words). This normalisation wasn’t coded in, but the function itself is fine.
* Just wanted to confirm whether we wanted to normalise or not.

1. **Perceptual Distinctiveness (Levenshtein Distance):**

* “Orthographic LD is the mean number of operations necessary to transform a given word into its 20 most orthographically similar words in the lexicon”. The Lexicon is from [Balota et al. 2017](https://pubmed.ncbi.nlm.nih.gov/17958156/).
* This needs to be recoded to follow the definition using the lexicon.

Balota, David A., Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman (2007), “The English Lexicon Project,” Behavior Research Methods, 39 (3), 445–59.

**Tom’s To-Do:**

1. **Incorporate LIWC in TTT**
2. **Harvard Dictionary:** I think I might just be dumb. Are the comments on the notions the links to the github?
3. **Rhetoric and Alliteration:** I coded that last year, will implement into TTT.