Proposal 1

Create a context free grammar and lexicon with detailed information collected automatically. Information like selection/subcategorization, parts of speech, location/colocation, how many arguments a verb takes, etc. We would want to pull the sample words from newscast transcripts, movie and tv scripts, and other corpora. This would give us data from authors specifically for both broadcasted news and tv scripts, which would potentially skew our age range to be higher. We would also want to find one more potential source for a younger audience and more varied gender scope. One compiled, ideally our database would allow for enough variation of word usage and structure that our lexicon could feasibly be established through running an automatic program.

Remko Scha and Livia Polanyi discuss doing so in their paper “An Augmented Context Free Grammar for Discourse”, discussing the optimal manner for parsing out this grammar or lexicon as well as touching on the reality of going through it in a left-to-right manner, as we would be in Python. To go one step further with the idea of parsing for a context-free grammar, Eugene Charniak in “Statistical Parsing with a Context-free Grammar and Word Statistics” compares various statistical parsing methods to optimize the context-free grammar program learning as well as find the most efficient and reliable option.

Proposal 1 References

Charniak, E. (1997) Statistical parsing with a context-free grammar and word statistics. *AAAI-97 Brown University*. http://people.csail.mit.edu/mcollins/6891Fall03/aaai97.pdf

Scha, R., & Polanyi, L. (1988) An augmented context free grammar for discourse. *Proceedings of the 12ths conference on Computational Linguistics, 2*(573-577). https://doi.org/10.3115/991719.991756

Proposal 2

Find the relative frequency of words across multiple social media platforms. Word frequency combined with the platform’s demographic information could provide interesting insight into generational linguistic trends. By scraping content (tweets, captions, bios), we could categorize word frequency across multiple age ranges. We could also follow up with further analysis of these words: colocation, capitalization, etc. We would need to build a database of content for multiple (perhaps three) social media platforms. We would prepare this data for manipulation in python and present our findings in a write-up or whatever format is preferable.

Nguyen et al’s "How Old Do You Think I Am?" explores age prediction based on a database of tweets. These tweets are analyzed by both humans and computer programs, with the computer programs making more accurate predictions. This article demonstrates that there are quantifiable linguistic variables that can accurately predict age. These variables can even be observed/detected in such short content as tweets.

Proposal 2 References

Seargeant P., Tagg C. (2014) Introduction: The language of social media. In: Seargeant P., Tagg C. (eds) *The Language of Social Media. Palgrave Macmillan, London.* https://doi.org/10.1057/9781137029317_1

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder. (2021) “How old do you think I am?" A study of language and age in twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 439-48, https://ojs.aaai.org/index.php/ICWSM/article/view/14381