Dagmawi Zerihun
CS 401 Program 3 - Text Analysis Report
8 Nov 2023

In this program, we compare part-of-speech tagging discrepancies across three different POS taggers(NLTK, Stanford and CLAWS). We parse sentpos files from the CLAWS, Stanford, and NLTK taggers and identify instances of discrepancy. We perform this analysis on three samples from a sentpos file of Hamlet, The Adventures of Tom Sawyer and the 2017 US Presidential Inaugural Address.

To evaluate the differences in automated POS tagging among these taggers, we map the tags from the CLAWS tagger to the corresponding Penn Treebank tags used by the Stanford and NLTK taggers. This allows us to match up the different tags and compare them directly. We also make sure that words which can be either nouns or verbs are only counted as verbs for our analysis by filtering them based on their part of speech tags. This helps us maintain consistency when looking at the frequencies of certain types of words across texts of different lengths.

As expected , gerunds are tagged differently across the different taggers. For instance, the word "contemplating" is tagged by CLAWS solely as a gerund 'VBG', while NLTK and Stanford list it as both a proper noun 'NNP' and a gerund 'VBG'. This suggests that NLTK and Stanford may be taking into account a broader context which could influence the word's function in a sentence. In addition, these taggers differ as to what they consider “proper nouns” in certain situations. The word "adventures" is tagged by CLAWS as a plural noun 'NNS', suggesting a general use of the word. In contrast, NLTK assigns it both 'NNP' and 'NNS' tags, and Stanford does the same but adds 'NNPS' to the mix, indicating recognition of this word as a proper noun in certain contexts, possibly when it is part of a title. The tagging of hyphenated words also varies. The word "counter-irritation" receives a general noun tag 'NN' from CLAWS, whereas Stanford classifies it as a proper noun 'NNP'. The difference here highlights the ambiguity that hyphenated words present to tagging algorithms, with Stanford potentially interpreting this as a specific term or name These discrepancies highlight the complexity of natural language processing and the challenges in creating a more streamlined approach to tagging parts of speech.

A possible limitation to our exploration might be the relatively simple tagset mapping we use. The direct mapping between different tagsets, while functional, could be reductive. A more nuanced mapping that can account for the subtleties of each tagger's linguistic model would likely reduce the number of discrepancies observed. Beyond mappings, the incorporation of neural networks capable of discerning the subtleties of language through context, rather than

relying solely on the isolated syntactic function of words, could greatly enhance the variances in POS tagging that result purely from context differences.