COMM 220 Final

Connie Bainbridge

**Introduction**

The purpose of this research is primarily exploratory, setting up multiple natural language processing (NLP) techniques adaptable to future texts of interest, where I intend to delve into self- versus other- communication and different creative methods of altering perceptions, mental health, and quality of experiences (both socially and otherwise). In order to explore a rudimentary comparison of different writing voices, this paper makes use of two datasets of essays from the USE (Uppsala Student English) corpus, where English-learners in Sweden wrote based on various prompts throughout their learning process. The two prompts of interest here were “English, my English”, where students describe their experience of English and evaluate their abilities (“personal, involved style”), and a literature assignment, where students elected to discuss a theme, character, or narrator and wrote based on a close-reading analysis of a set passage (“formal style”). My planned analyses involved comparisons of sentiment, looking at the frequency of “positive” and “negative” words, as is common in sentiment analyses (in future work I hope to look into the degree of positivity and negativity and additional affective terms, see: Aman & Szpakowicz, 2007 – full ref. in [readme](https://github.com/conbainbridge/COMM220_USE_nlp_project)). Various exploratory analyses looking into features and frequencies are included and speculated on.

**Methods**

In order to compare the two different prompts, I elected to explore (1) sentiment analyses of negative versus positive word usage; (2) essay word lengths, and (3) word clouds and word frequency distributions. Sentiment analyses were achieved using “nrc” in the “tidytext” R library, and LSA was conducted using “quanteda”, with the addition of the “quanteda.textmodels” library; in order to parse out the essay texts, which were collected in individual .txt files containing HTML tags, I employed the “rvest” package. Analyses and treatments of the data, as well as some setup for future work are in the R code (available [here](https://github.com/conbainbridge/COMM220_USE_nlp_project/blob/main/USE_text_analysis.R)).

**Results**

To compare sentiment across essay prompts, I ran a two-sided t-test for positive words across groups, as well as for negative words. Positive words were not significantly different on average across groups (mean for literature prompt = 0.047 and English prompt = 0.048; t = -1.207, p = 0.228); however, negative words appeared significantly more frequently in the literature-based essays than in the “My English” essays (means = 0.026 and 0.019, respectively; t = 9.704, p < .001). In addition to the sentiment analysis, I also compared the essay lengths across groups. The literature prompt featured significantly more words on average compared to the “My English” prompt (means = 910.2 and 702.007, respectively; t = 14.622, p < .001).

Given the mostly exploratory nature of this project, remaining results focus primarily on frequency distributions rather than statistical comparisons. The figures below depict the word clouds and word frequency distributions, where “max\_words” is set to 500. As can be seen in the case of the “My English” prompt word cloud (Fig. 1), “English” appears as an extreme high frequency outlier, and a large number of words appear with minimal frequency compared to a few central terms (indicated by small, green-colored text). Several of the most frequent words in the literature prompt word cloud (Fig. 2) are names, as well as the word “theme” and words that likely reflect themes of interest (“family”, “happiness”/”happy”, “life”) . The word cloud for the literature prompt clearly shows more variation in size and color of top feature words, reflecting more diversity and variation in frequency of words. This difference is also visible in the frequency plots (Figs. 3 & 4): the very clear high outlier, “English”, appears 3.382 times more than the second most frequent word, “language”, with frequencies at 4410 and 1304, respectively (frequencies gathered from the topfeatures lists in the [r code](https://github.com/conbainbridge/COMM220_USE_nlp_project/blob/main/USE_text_analysis.R)). The top two words in the frequency distribution for the literature prompt, “harriet” and “family” are far more comparable in frequency, with 786 and 779, respectively.

In investigating the top 20 features in document feature models for each essay prompt, I identified a few words in common between the two: ”like”, “time”, “can”, and “also”. In conducting a series of t-tests for each word, I found a significant difference in frequency only for “like” (mean frequencies: literature prompt = 2.865 and English prompt = 2.017; t =4.124, p < .001). Additional investigations into bigram frequencies are available in the r code - the top 20 bigrams appear to differ substantially across prompts.

**Conclusion/Discussion**

In considering the expectation of each prompt, the results I report here may appear rather intuitive; for example, when asked to provide a critical analysis of a literary text (in this case, either investigating the narrator, a specific character, or a theme), it is likely that negative language will permeate given the critical nature of the prompt. In the same vein, such a critical analysis would, as one might expect, provoke longer essays in order to elaborate on said critical text evaluations. As I pivot in the future to different styles of self-communication, I suspect considering these distinctions may provide direction in how to pose such prompts. For example, it is commonly referenced in conversations surrounding mental health that self-talk tends to orient more negatively and critically than how one might talk to others. In exploring prompts that change the nature of self-reflection, there may be optimal interventions to be revealed in the data. For example, critical analysis of one’s own character versus reflection on one’s learning experience of developing themselves and their values may be close proxies to the essay prompts investigated here. The next necessary step would then be to map emotion and mental health measurements onto the use of language to see what interactions arise.

For these specific USE datasets, in the future, it may be interesting to look in depth at the use of non-English words and phrases in the essays. When evaluating various bigrams, it was revealed that several instances where Swedish (or otherwise non-typical English, such as the phrase “Déjà vu”) failed to parse, creating a high frequency of the bigram “�\_�”. If one is to take these symbols to represent non-English bigrams as a whole (ideally, this would instead be evaluated by properly extracting non-English text), it appears there may be a greater instance of such bigrams in the “My English” prompt compared to the literature prompt (frequency = 155 versus 50, respectively). Given that reporting on one’s experience of learning English may prime reflection on one’s native tongue, it may be reasonable to expect such differences in frequency across essay prompts.

Another topic to consider is misspellings – a quick look through the words (sans stop words – this may make more sense to conduct with stop words included) shows several that may reveal information about their English language experience. Some interesting examples include: the presence of both “interessted” and “interessting”, “ame-rican”, “speking”, “paragraphes”, and “dictionari”. While I have not delved into analyses in this domain as of yet, one might be able to approach this by loading in a dictionary to filter out correct spellings, then proceed by investigating frequency of misspelled words overall across essay prompts. Perhaps comparisons with common Swedish spellings and pronunciations may be able to elucidate an interaction of the two languages. It may also be informative to look at which specific spellings appear with any notable degree of frequency – one would expect a typing error to appear with limited frequency, whilst a systematic error would appear at greater frequencies (either within or across essays).

There are additional text cleaning methods I would strive to investigate and employ for future research. For example, the appearance of “p” as a high frequency word in the word cloud of the literature prompt raises a flag, which I suspect may be HTML paragraph tags that have failed to filter out. Additional such details likely remain (some symbols appear to have remained as well), as well as some debate over what constitutes “stop words” or otherwise minimally informative words that have not been filtered out, such as “also” (a high frequency word shared across both groups). Additionally, I suspect that a word such as “like” may pose conflicting results in sentiment analyses, as “like” can mean “similar” as well as an expression of fondness (the former of which might not qualify as “positive”, while the latter might). Given the significant result in comparing frequency of “like” across prompts, it is possible the null finding for positive sentiment when comparing essays may be slightly attributable to this word. Comparing across different sentiment libraries, or delving into methods for words with multiple meanings, may be informative in handling this type of issue moving forward.

Chart

Description automatically generatedChart, histogram

Description automatically generatedText

Description automatically generatedA picture containing text, newspaper, screenshot

Description automatically generated**Figures**

**Figure 4** The word frequency for the "My English" prompt shows a clear outlier, "english". I suspect without the outlier of “english”, this figure might look more similar to Fig. 3.

**Figure 2** Word cloud for "My English" prompt. The dominance of the word “English” is reflected here, and far less variation in the distribution compared to the Literature prompt.

**Figure 1** Word cloud for Literature prompt. Colors and size of words reflect variations in word frequency.

**Figure 3** The word frequency distribution for the Literature prompt. “harriet” and “family” are similarly high in rank, and appears Zipfian-like as one would predict in cases of word usage.

**Github repository:** [**https://github.com/conbainbridge/COMM220\_USE\_nlp\_project**](https://github.com/conbainbridge/COMM220_USE_nlp_project)