COMM 220 Final

Connie Bainbridge

**Introduction**

The purpose of this research is primarily exploratory, setting up multiple natural language processing techniques adaptable to future texts of interest. In order to explore a rudimentary comparison of different writing voices, this research 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, and a literature course assignment, where students elected to discuss a theme, character, or narrator and wrote based on a close-reading analysis of a set passage. The styles of the prompts are described as “Personal, involved style” and “Formal style”, respectively.

My planned analyses involved comparisons of sentiment, specifically looking at the frequency of “positive” and “negative” words. Additionally, various exploratory looks into frequencies are included and speculated on.

**Methods**

The USE dataset was used to extract essays from Swedish individuals learning English (a link to download the dataset is available in the r code). In order to set up a system for various text-based analyses to later evaluate self- versus other-communication, I elected to compare two essay prompts: the first of which explored a literary analysis of novels, either elaborating on characters, themes, or the narrators; and second of which explored “My English”, reflecting on the experience of learning English as a secondary language. In order to compare the two different prompts, I elected to explore a few difference natural language processing methods: (1) sentiments 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” 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 flanked with HTML tags, I employed the “rvest” package. All analyses and treatments of the data occur in the R code (available here).

**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 in frequency across groups (means = 0.047 and 0.048, respectively; 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 distributions of 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 in on frequency distributions rather than statistical comparisons. The figures below depict the word clouds and word frequency distributions, where maximum 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. Similarly, 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 presence of many small, green words in the cloud for the “My English” prompt may reflect greater diversity of word usage, as this reveals many words of lower frequency compared to the high frequency usage of “English”. The word cloud for the literature prompt clearly shows more variation in size and color of terms. This difference is also visible in the frequency plots (Figs. 3 & 4): a very clear high outlier, “English”, appears 3.382 times more than the second most frequent word, “language”, with frequencies at 4410 and 1304, respectively. 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 (with “harriet” appearing 1.009 times more frequently).

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 examining each feature at a time, 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 prompted 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. A I pivot in the future to different style 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 similar proxies to the essay prompts investigated here. The next necessary step would then be to map emotion and mental health states onto the use of language to see what interactions arise.

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 evaluating 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). There may be some interesting analyses to consider here in follow-up analyses – 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 additional language-related topic to consider is misspellings – a quick look through the words (after removal of stop words) shows several that may reveal information about their English vocabularies and where mistakes may be made. Some interesting examples that may be informative of misunderstand rather than mere typing errors include the presence of both “interessted” and “interessting”, “ame-rican”, “speking”, “paragraphes”, and “disctionari”. 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 that could be used to filter out correct spellings, then proceed by investigating frequency of misspelled words overall across essay prompts. Another interesting potential analysis could be looking at which words or specific spellings appear with any degree of frequency – one would expect a typing error to appear with a limited frequency, whilst a systematic error would appear at greater frequencies (either within a single essay or across several).

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 an HTML paragraph tag that has failed to be parsed 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”. 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 partially 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".

**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 from a purely visual standpoint the distribution appears Zipfian-like.