COMM 220 Midterm Project proposal

Connie Bainbridge

My goal for the final project is to set up the methods and technical skills necessary to progress my research into textual analysis, in the hopes of eventually exploring creative writing to the self and others and the ways the two forms of communication may influence each other. As I suspect that emotive language is used very differently depending on individual differences as well as the target of the communicative signal (receiver as self versus other), I will need a toolset for exploring language possibility space, including topical analysis, affective content, and perhaps eventually verb tense. While I will be saving such types of studies for down the road (as they would require development of a study for collecting ideally large amounts of open-response data), this final project will focus on the tools and methods by analyzing available preexisting essay datasets.

Using the USE (Uppsala Student English) Corpus, which comprises of writing from Swedes learning the English language, I will be employing NLP techniques to conduct text analyses, exploring in particular the essay prompt: “a4. Literature course assignment. Students choose between a discussion of theme/character/narrator and a close-reading based analysis of a set passage. Formal style. Written in early April or November.” I have selected this prompt as it poses the opportunity for comparisons – while comparing across different essay assignments would be interesting, I decided to avoid the potential for English abilities being inconsistent from assignment to assignment as students gain more experience writing in English. I will be parsing out the essays where students selected discussion of a character to compare to essays focused on a theme (or if the N proves smaller than ideal, the close-reading based essay option compared to character OR narrator), to evaluate differences in lexicon that may arise when focusing on a person versus a concept. Given that a theme will likely probe more investigative and perhaps opinionated thinking than a character, I suspect that the themes explored will present greater variance in topic and emotive language than when focused on a character. I then intend to venture into exploratory terrain, with an interest in identifying specific differences in use of emotion-based words. Given future interests in exploring negative versus positive emotion language (e.g. more negative emotion language likely used to the self versus to others), I will be exploring these essays on this dimension as well. While I don’t have strong priors on how the essays might differ on emotive content, it may be the case that themes with negative leaning language may be reflected by a character evoking negative affect or opinions as well (either by character flaws or negative experiences).

In order to explore these differences, I intend to study the topic and word usage by generating a topic model using the r package “topicmodels”, where I will be using the Latent Dirichlet allocation (LDA) algorithm. To clean out common, non-topical words such as “a”, “the”, and so on, I will set stopwords to TRUE to ensure removal. To explore emotional content, I will be making use of the r library “tidytext”, using get\_sentiments(“nrc”), which will both assign words to positive or negative categories, as well as the categories of “anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.” I am additionally considering running the get\_sentiments(“AFINN”) function, which assigns words numerical scores between -5 and 5 to reflect negative and positive affect levels, allowing for different statistical analyses than the categorical analyses of the nrc function. Because I have struggled to find specific packages for verb tense, I may save it for future projects, though it may be partially feasible via regex.