GSURC Research Proposal by Kelly Carr, John Webb, Jasmin Anderson, Joyce Bullock-Vigdor, and Matthew Wamboldt

Research topic: Understanding Online Language: A Corpus-Based Comparison of Web Registers

Introduction: In this poster presentation, the similarities and differences between common web-based registers are analyzed and interpreted using corpus-based analysis. We learned from the perspective of Sociolinguistics that variation in writing exists and could be investigated across genres (or "registers") by comparing the co-occurrences of many linguistic features. Although there is much research regarding internet-based discourse as a whole, as these internet registers continue to change the way we use language, the research surrounding internet-based registers is still relatively new (Friginal and Hardy, 2014). The sub-registers of e-mail writing, blog writing, and online news and opinion column writing, when analyzed linguistically can provide interesting information about the emerging language online.

Purpose: The primary goal of this study is to identify the characteristic features of online language and compare their distributions across groups of texts. Online language is represented by blogs, micro-blogs, workplace emails, discussion posts and reader feedback, and online newspaper and opinion columns. Interpretation of linguistic patterning, for example, from the texts of emails and Facebook posts may therefore show the influence of production constraints (e.g., Facebook has a default limit of 420 characters per status post; 704 characters for Notes), setting, topic, and target audience.

Methodology: The data used in the present analysis came from an exploratory corpus of online texts (with approximately 16,501,785 words) collected by Dr. Eric Friginal of the Department of Applied Linguistics and ESL at GSU from various online public domains from 2006 to the present. A combined automated and manual collection of texts was conducted and the resulting corpus was analyzed using a combination of computational tools (e.g., concordance, tagger). We processed and identified a range of frequency data of linguistic features and as a group, we then functionally interpreted these patterns.

Results: Interesting similarities and differences in the linguistic properties of these six groups of texts are observed and the functional interpretation of patterns appears to be supported by pertinent text samples from online sources. For example, one might anticipate, from the interactive and informal nature of Facebook and Twitter status updates, that these combined texts might be characterized by familiarity and personal focus. In our study, however, Facebook/Twitter posts were identified more by "nominal" and "informational" style of writing and very limited narrativity (i.e., narration of events). The limited space (based on number of characters) allowed in micro-blogging to communicate details influences writers to focus more on the nominal style of posts or updates. We will be presenting these types of results in our poster.

Conclusion: We examined online registers and explored the linguistic properties of blogs, online newspaper articles, emails, Facebook/Twitter updates, reader comments, and opinion columns. By establishing the characteristic linguistic features in these web registers, we were able to show how prototypical a particular text may be compared to other texts. A functional analysis of the meanings and messages of these features is necessary to completely define the underlying linguistic characteristics of online language.

Recommendation: Results thus far are clearly exploratory and limited, given the current composition of the web registers corpus. Future studies may provide a more detailed and complete description of the co-occurring features within online registers when additional data and more specific texts are included in the analysis. However, we believe that it is a good start and initiative to extract frequency distributions to understand online language.