Selected Papers of Internet Research 16:
The 16th Annual Meeting of the
Association of Internet Researchers
Phoenix, AZ, USA / 21-24 October 2015

LANGUAGE SOCIALIZATION AND GENRE ADHERENCE IN REDDIT.COM

Ashley Titak
Georgia State University

Abstract

Social networking sites (SNSs) are receiving attention from researchers from various theoretical linguistic frameworks. However, there is a paucity research considering language use in digital genres from a language socialization perspective. This study examines Reddit, a popular forum community, with a language socialization perspective, using a comparative corpus analysis to analyze language patterns. A comparative corpus analysis involves comparing the subcorpus in question (e.g., First Year Users) to a larger reference corpus using frequent word, keyword, and n-gram analysis to note the difference in language patterns between comments. Given my interest in socialization, I compare language in posts according to the user's year of use (e.g., First Year users, Second Year users, etc.), the comment's score assigned by other community members (e.g., positive and negative), and the sub-communities (e.g., subreddits such as the *Best of Netflix* subreddit) in which the posts were made, in an effort to explore how language may vary according to these variables and to identify potential trends.

Introduction

With the ubiquity of social networking sites (SNSs), researchers have begun to examine how members of online communities and networks use language, and to what effect. Various theoretical lenses have been used to account for how users learn to use language effectively in these digital settings, including language socialization, genre, digital literacy, authorship analysis, and computer mediated communication (CMC) competence. However, there is still a paucity of research analyzing the language used by novices as they endeavor to become experts in informal online communities, and fewer studies still which examine such socialization within the realm of SNSs. Previous research show that groups develop language norms and shared patterns of communication over time in online communication (Postmes, Spears & Lea, 2000) and that members converge in language style and strategies (Cassell & Tversky, 2005), suggesting that particular internet subcultures develop their own way of speaking. The

Suggested Citation (APA): Titak, A. (2015, October 21-24). *Language Socialization And Genre Adherence In Reddit.Com.* Paper presented at Internet Research 16: The 16th Annual Meeting of the Association of Internet Researchers. Phoenix, AZ, USA: AoIR. Retrieved from http://spir.aoir.org.

prevalence of memes in some online communities also suggests that some language norms are privileged and shared in particular settings.

While it is argued that feedback from other community members and an increased amount of time spent in an environment leads to more generic language use in accordance with community practices, language socialization literature has yet to truly examine the extent to which specific aspects of the community (such as the time spent in the community, larger community practices, and the influence of sub-communities) impact language use.

Though many SNSs sites, such as MySpace and Friendster, move quickly in and out of public use, other SNSs seem to have longevity, as in the case of Reddit and Facebook, each of which have lasted nearly a decade. The longevity of these SNSs make them useful sites to examine long-term language socialization in digital settings. Additionally, Reddit is specifically useful for this kind of research because a user's profile contains the number of years a member has been active, a history of the user's posts, the total points accrued by each post, and the subreddits (i.e., sub-communities) in which a user has posted. These data can be used to examine and compare language patterns between sub-communities, posts with point values assigned by community votes (i.e., positive and negative scoring posts), and between users who have spent different amounts of time using Reddit. Language socialization theory, among other linguistic theories, would predict that the data will reveal linguistic trends over years of use, that more experienced users will collect higher point values per comment, and that individual subreddits will differ slightly in language patterns that emerge because of the influence of sub-community practices. In regards to the language patterns that emerge in sub-communities, sub-communities with a geographic location, such as the "Boston" subreddit, are of particular interest due to what may be revealed in a comparative examination of language use.

Research Questions and Methodology

In the present study, I examine language in comments made in response to original posts (OP) in the forum community, Reddit.com. I use a corpus of comments collected using a Reddit's API and Python programming. The corpus is divided into sub-corpora, according to posts made during different years of use (e.g., first year users, second year users, etc.), sub-community (e.g., the Atlanta subreddit, the News subreddit, etc.) and final score of the post (e.g., positive and negative). This division allows for the comparison of language to identify any patterns that may be available in the data. The number of years a user has been active on Reddit, the sub-community in which a post is made, and the community's feedback in the form of comment scores, are used as variables in this study. While this division is limited in that a user's demographics (such as age and level of education) are not available, it provides an exploratory snapshot of language use by identifying the more common language patterns that occur more frequently in a sub-corpus when compared with others.

In order to study these language patterns, I use a comparative corpus analysis, focusing on frequent word analysis, keyword analysis, and n-gram analysis. A comparative

corpus analysis involves comparing the subcorpus in question (e.g., First Year Users) to a larger reference corpus (e.g., all other users including Second Year, Third Year, etc.). Frequent word analyses show what kinds of words are frequently used within each subcorpora compared to frequent words in reference corpora. Keyword analyses compare common words in the corpus being investigated, which are comparatively rare in a reference corpus. This kind of analysis provides an "aboutness" of the corpus under examination, and also reveals "salient features that are functionally related to that genre" (McEnery, Xiao, and Tono, 2006). N-gram analyses show sets of words that most often collocate together within the corpus, revealing common structures used within each subcorpora, allowing for the examinations of collocations which differ by only one or two words. Through these analyses, differences in language use on Reddit can be examined according to the user's year of use, comment score, and the overarching sub-community in which the comment was made, and reveal a picture of larger language patterns.

Initial Findings

The exploratory pilot study on which the present study is based was limited in representativeness by the limited corpus size and the number of years of use included (i.e., only First, Second, and Third year users). Additionally, the data did not include a each comment's score. However, the pilot study revealed findings which suggest the present study, with a much larger corpus and a wider lens of focus to include sub-community and comment scores, could yield interesting comparative results regarding the language and strategies used by members of the Reddit community, and various subreddits, over time. One of the most interesting findings is that language typically found in memes popular on Reddit were absent; this finding is especially intriguing because memes are, by definition, replicated and reposted.

The pilot study showed a difference in use of hyperlink language, such as *r* (which links to other subreddits on Reddit), and *com* among the three years of use corpora. While hyperlink words do not necessarily represent words by themselves, concordance programs such as ANTconc (Anthony,2010) enables these terms to be considered words in its processing. It is true that these terms most often appear in conjunction with a web address, and for the purpose of this examination, they will be recognized as signal words which indicate that there is a shift in modality from the "reading" mode to the "navigation" mode. These words are especially relevant to digital genres because they introduce a mode of reading that differs from traditional genres. Digital genres feature frequent shifts between "reading" and "navigating" modes, which changes the experience of reading traditional texts (Askehave & Ellerup Nielsen, 2005, p. 125). Users who encounter hyperlinks, for example, may choose to permanently, or temporarily, leave a text at any time (2005: p. 125). Askehave & Ellerup Nielsen (2004) argue that this modality is influential on the discourse used in digital genres.

In the pilot study, three particularly interesting uses of hyperlink language were noted. First, the Third year corpus showed an overall decrease in reliance on links to other subreddits and other sites on the web in comments. This means that Third year users

used hyperlink language less than their First and Second year counterparts. Additionally, Third year users provided more surrounding context when using hyperlinks to other subreddits; for example, they would provide information about why the subreddit was useful to the discussion. A third finding is that First and Second Year users were more interested in managing the community through their use of hyperlink language. For example, First and Second year users may link to another subreddit as a complaint, suggesting that a previous comment or an OP belongs in a different subreddit. This kind of usage was seen less frequently in Third Year user posts. The varied use of hyperlink language suggests that Third Year users may be less interested in managing community posts, and are less likely to provide an invitation to leave the conversation and to go to other subreddits.

The pilot study also revealed other differences among user's posts in the use of Reddit's quote function and in the use of language unique to digital settings. The quote function allows a user to highlight a previous user's words in a different font within their post. While users from all three years of use generally used the quote function in the initial portion of their comment, there were differences in the extent to which the quote function was used. First year users used this function more frequently than Third year users. Third year users seemed to prefer avoiding the use of the quote function in favor of incorporating the context of previous responses into their comments.

Language unique to the digital settings, such as “TL;DR” (“Too long; didn’t read”) and “Edit:” were used differently among the three years of use. Third Year users preferred to use TL;DR as a courtesy to other users. These users would use TL;DR as a means of summarizing their own longer posts, providing other users the opportunity to skip over their longer comments in favor of the summary. First and Second year users used TL;DR much less frequently. All three corpora included instances of "Edit," which was largely used either to provide clarification and additional information, or grammar corrections. While all three corpora showed instances of clarification and elaboration, the Third Year corpus used "Edit" in this way more frequently. Additionally, the Third year corpus showed few uses of "Edit" to show grammatical corrections in their own posts.

While the data used in the pilot study were not sufficient to make representative claims about language socialization, the findings from the pilot study invite further investigation with a larger, more representative corpus. More research into language use over time in various SNSs is needed, as online applications, including SNSs, are becoming commonplace among adults, with users almost seamlessly moving between offline and online, and between platforms (Buck, 2012). Additionally, research is beginning to show that online data may be somewhat representative of the offline population (Gosling, Vazire, Srivastava & John, 2004; Back et al, 2010; Swartz et al., 2013), and digital literacy investigations are beginning to consider how online language may bleed into offline settings. By investigating the language used by members within Reddit, who demonstrate various levels of expertise, we can begin to examine how effectively users recognize and adhere to linguistic norms of the community in digital settings.

References

Anthony, L. (2010). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Askehave, I. and Ellerup Nielsen, A. (2004). Web-Mediated Genres. A Challenge to Traditional Genre Theory, Working Paper no. 6, Centre for Virksomhendskommunikation, Aarhus, 1–50.

Askehave, I. and Ellerup Nielsen, A. (2005). Digital Genres: a challenge to traditional genre theory. *Information Technology and People*, 18(2), pp. 120-141.

Back, M., Stopfer, J., Vazire, S., Gaddis, S., Schmuckle, S., Egloff, B., & Gosling, S. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3), 372-374.

Buck, A. (2012). Examining Digital Literacy Practices on Social Network Sites. *Research in the Teaching of English*, 47(1), 9-39.

Cassel, J. & Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer Mediated Communication*, 10, 16-33.

Gosling, S., Vazire, S. Srivasta, S., John O. (2004). Should we trust web-based studies? A comparative analysis of sic preconceptions about internet questionnaires. *American Psychologist*, 59, 93-104.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based language studies*. Abingdon: Routledge.

Postmes, T., Spears, R., and Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3), 341-371.

Schwartz, H., Eichstaudt, J., Kern, M., Dzuiurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., Ungar, L. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary approach. *PLOS one*. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0073791