A Linguistic Characterization of Google+ Posts across Different Social Groups

Evandro Cunha[1, 2], Gabriel Magno[1], Marcos André Gonçalves[1],
César Cambraia[2], Virgilio Almeida[1]

[1] Computer Science Department, Federal University of Minas Gerais, Brazil
[2] College of Letters, Federal University of Minas Gerais, Brazil
{evandrocunha,magno,mgoncalv,virgilio}@dcc.ufmg.br, nardelli@ufmg.br

Introduction

A major research topic in linguistics is the effect that social factors – that is, aspects from outside the linguistic system – exert on the use of language by speakers. Just as language usage varies due to geographical factors, giving birth to different *dialects*, it also varies according to social factors, generating different *sociolets* [4]. At the same time, research in social computing has been performed with the aim of describing, analyzing and understanding users' behavior in online social networks (OSNs). The analysis of human attitude in OSNs can not only characterize interaction in online environments, but also help to find answers to questions in the fields of social sciences and humanities.

Increasingly, scientists have taken advantage of the vast amount of language data that online applications can provide, giving rise to a new subfield of knowledge that has been called "Internet linguistics". According to [2], the Internet plays an unprecedented role in the study of language, as it allows linguists to use rich documented datasets to investigate the rate and the reach of language change in various levels. In this perspective, linguistic studies are concerned with understanding and describing computer-mediated communication, as well as developing tools to provide better and safer online services, often employing collections from user-generated content websites as corpora of large-scale natural language data.

As a social networking service, Google+ allows users to share content with their friends and followers. This content can be of different natures: text messages – such as status updates –, images, videos and URLs [6]. In this work, we study one kind of content published in Google+: status updates, usually called *posts*. Our focus is to characterize Google+ posts and to identify differences and similarities among linguistic aspects of texts produced by users considering their distinct social characteristics. We analyze male and female network members from four countries and 15 groups of occupations, since gender, location and professional activity are notably known as factors able to influence language usage in a myriad of domains [4]. Our main findings include:

- Although Google+ does not limit posts to a small number of characters like other OSNs do, the standard status updates are still very short;
- Certain social groups organize differently their posts, so that the structural complexity of the messages may be quite distinct among users from particular countries, genders and occupations;
- The fraction of misspellings in Google+ posts varies significantly among different social groups. We found a relationship between this fraction and the nature of individuals' professional activities;
- Social groups are not homogeneous with regard to the use of categories of words. Particularly, we discovered that vocabulary employed in Google+ posts is highly dependent on the users' occupations, which may indicate that this OSN is often seen as a professional network or that members maintain a professional writing style even in this environment.

Data collection

From March 23rd to June 1st, 2012, we collected profile information and posts from the 8,564,462 users who set their shared content in Google+ as publicly available. For ethical reasons, no attempts were made to obtain access to information set as private, and we gathered the public data revealed in the users' profiles by making HTTP requests to the pages. We were able to retrieve the last ten status updates from each user's profile page, totaling 29,366,310 posts.

Posts

Our work is focused on the analysis of posts written in English. To select only messages generated in this language, we used langid.py [5], a tool that provides the probability of a text being in a particular language. We filtered texts with probability of at least .99 of being in English. Then, we alleviated the impact of copied posts, like chain letters and other highly replicated texts, by removing duplicated messages. After these restrictions, we narrowed our dataset down to 6,194,338 distinct posts.

On average, posts have 111.2 characters and 25.6 words. The majority of posts have only a few sentences: 53% of them have one sentence, while 26% have two sentences and 10% have three sentences. This confirms the hypothesis that, even though Google+ posts are not compulsorily limited to a small number of characters like Twitter updates and Foursquare tips, they can still be considered microtexts.

Social information

Since we study language use of different groups of people, we also collected information on users' gender, location and occupation. Non native speakers have higher probability of transferring linguistic patterns of their mother tongues to the language spoken as second language [1], so we performed an extra filtering, considering only members located in Canada, India, Great Britain and United States, ultimately limiting the number of posts analyzed to 1,920,482. We used the Standard Occupational Classification (SOC) published by the US Department of Labor [9] to divide the professional activities most cited in the profile field "occupation" into the major groups of professions used in this study.

Analyses and results

The analyses we performed are all independent investigations, in that they are not examining the same text attributes, which makes it possible to test distinct aspects of language behavior. All these experiments have already been conducted in different situations and are relevant for linguistic studies.

Readability and structural complexity

The readability of a text can be described as the ease in which readers can properly comprehend it. A series of formulas, that aim to quantify how simple a text is to be understood, have already been proposed [3]. These formulas return numerical scores that estimate the level of difficulty of the analyzed texts and should not be seen as metrics of the quality of documents, simply because *easier* or *more difficult* texts are not necessarily *worse* or *better* texts. One of these formulas is the Automated Readability Index (ARI), defined by

$$ARI = 4.71 \cdot \frac{\text{number of characters}}{\text{number of words}} + 0.5 \cdot \frac{\text{number of words}}{\text{number of sentences}} - 21.43$$

As one can see, the ARI is associated with structural aspects of the texts, since it relies mostly on a factor of characters per word, and on a lesser extent on a factor of words per sentence. Thus, the ARI's assumption is that the adoption of big words and the

construction of large sentences are features that enhance the complexity of a text. Naturally, it considers only the structural complexity of the passages, not their conceptual complexity.

According to our results, texts of Indian users on Google+ are more complex than those of users from other countries: although the number of characters per word is very similar, the number of words per sentence is higher among Indian members compared to the Westerners. We must take into account that, in India, there is the question of multilingualism, and substrates from local languages may influence the shaping of texts in English.

The ARI scores for male and female users show that posts written by men are structurally more complex than those written by women. This fact is observed for all countries – except for India, where scores for male and female members are equivalent – and for most professions. Concerning occupations, our results state that workers from fields more associated with written communication and traditionally elaborated texts, like scientists and legal professionals, publish more complex posts than those from fields that do not necessarily deal with written texts, like food preparation and sales professionals. Ahead, we will advocate that: (a) men and women make distinct use of this OSN, which could explain the differences in the complexity of the posts between genders; and (b) Google+ users are often talking about their own professional activities and, therefore, talking about topics that ask for either more or less elaborated linguistic constructions.

Misspellings

The occurrence of misspelled words in texts may indicate unawareness of standard orthographic rules, revealing low literacy levels, or carelessness during writing, due to negligence or lack of revision. Thus, calculating the extent to which misspellings emerge in certain texts might indicate how high are literacy levels of the studied communities or how concerned are individuals about the quality of their posts.

By using a list of 4,238 common misspellings in English[1], we investigated the occurrence of these non-standard linguistic elements in posts produced by different social groups. Our experiment calculates the fraction of misspellings per post, which is obtained by dividing the number of misspelled words by the number of words susceptible to misspelling present in our list. To avoid biases in the results due to the small number of words susceptible to misspelling in some posts, we excluded from this analysis texts with less than five words that appear in our list.

The outcomes indicate that Indian members are more prone to make misspellings. On the contrary, the values for Canadian users are lower than those for all other groups. The Human Development Index [8] may suggest an interpretation of these results: according to the report, Canada is one of the few countries with 100% of the population aged 25 or older with at least secondary education; United States and United Kingdom have, respectively, 94.5% and 99.7%; and India, on the other side of the table, has only 38.7%. Moreover, the rankings of these countries in a reading score of 15-year-old students index follow the same positions as our calculation of the fraction of misspellings.

Also, we found that, in general, women's fraction of misspellings is higher than men's. To explain this, we can suppose that the difference between the topics of the posts written by men and women – a fact that will be explained in the next analysis – does not force women to be so demanding on the formal linguistic attributes of the content published.

The examination of the fraction of misspellings in posts of users with different occupations can be related to the previous analysis on structural complexity. In the same way that professionals that deal more with written texts produce more structurally complex posts, they also make fewer misspellings: while media and education professionals have the smallest fractions, health and food preparation professionals have the highest ones. It's worth remembering that, by the nature itself of these occupations, the review of written material is sometimes part of the activities performed daily by media and education professionals.

Semantic categories of words

An interesting way of studying language differences across social groups is through the analysis of the vocabulary used by their members. This kind of investigation reveals how particular groups perceive reality, showing, for example, what are the main concerns of certain communities.

[1] http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines

We aim to identify if particular semantic categories of words are more common in texts produced by users of certain groups. To accomplish this task, we used the Language Inquiry and Word Count (LIWC) [7], a tool that examines texts and verifies the occurrence of words previously classified as members of grammatical (e.g. pronouns, articles, prepositions etc.) or semantic (e.g. social, money, religion etc.) categories.

We analyzed 64 categories of words. Figure 1 shows the ones with the most significant differences across the distinct social groups considered in this study. However, significant differences were found in most of the categories.

As in the previously described investigations, the Indians hold a distinct pattern from the Westerners analyzed. Indian users have the highest scores in the use of words from categories such as “friend”, “humans”, “social”, “positive emotions”, “health” and “religion”. On the other hand, they have the lowest scores in categories like “negative emotions”, “anger”, “time”, “space” and “money”. These categories might be revealing the topics more covered in the posts and are a sign of cultural differences between Indian and Western Google+ users – that, regardless of the country, have homogeneous scores in almost all categories.

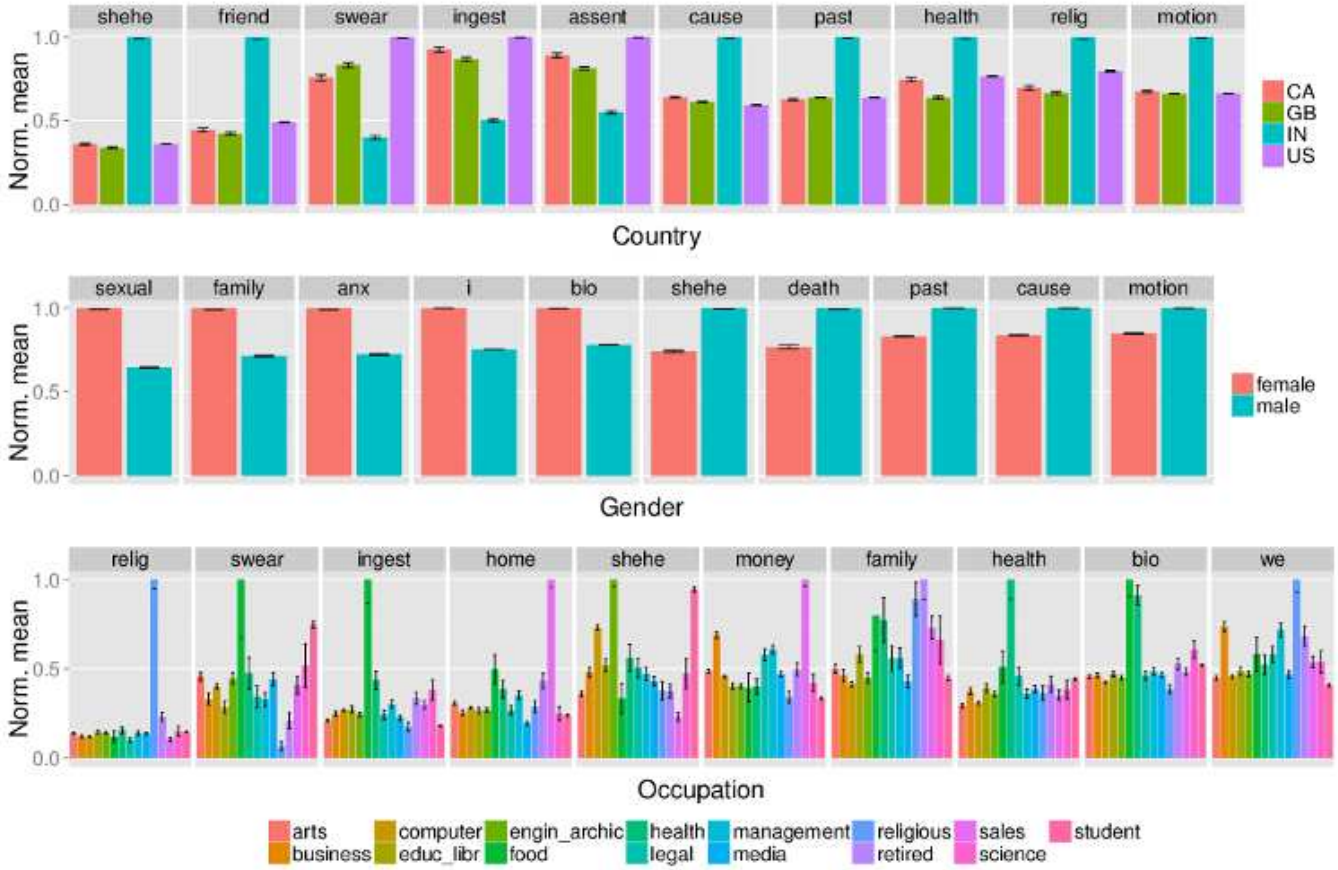

Figure 1: Categories of words with most significant differences across distinct groups of users (normalized mean ± std error)

Considering gender, we found that women are more prone to use words from categories such as “family”, “social”, “friend”, “humans”, “affection”, “emotions” and “home”, while men are the main adopters of words from categories like “numbers”, “money”, “causal”, “motion” and “space”. We interpret these results suggesting that men have a tendency to use Google+ to talk about technical topics, their achievements and professional activities, while women are more likely to use this OSN to talk about their social and familial relations. This may be the reason why men’s posts are more structurally complex and more formally accurate, having fewer misspellings, as described above.

We also found a clear correlation between word usage and the users’ occupations. For instance, words related to religion are extremely more frequent in posts from religious workers; the same for money vocabulary in posts from salespeople and body-related words in posts from health workers, among many others (interestingly, the category “family” is adopted mainly by retired

users). This fact suggests that the vocabulary employed in Google+ posts is highly dependent on the users' working activities, indicating that members' professional writing style is maintained even in this environment.

As far as we are concerned, these significant differences among the vocabulary of users with different occupations have been found for the first time in the context of online social media.

Concluding remarks

Online social networks have opened huge opportunities to understand the role that social factors exert on linguistic aspects of user-generated content on the Web. In this study, we consider about two million Google+ posts in order to evaluate linguistic elements like structural complexity, occurrence of misspellings and use of words of different semantic categories among members of particular social groups. These analyses not only describe the posts, but especially identify how social groups differ when posting content on the Web.

To the best of our knowledge, this work is the first to focus on language aspects of Google+ posts and one of the most extensive investigations of the role that social factors exert on language usage in an online social networking service. Contributions of our study go beyond the mere characterization of posts – which per se is an important supplement to the literature on language use in social media – since implications on misbehavior detection and personalization of services may follow. For instance, our analyses may be useful to improve the task of automatically detecting fake entities in a given social group – since the misalignment with most of the collective language features could mean that a certain element does not belong to that group – or to identify fake profiles by analyzing their linguistic behaviors, with implications for privacy and misbehavior protection. On another direction, our results may help to improve language modeling focused on the personalization of services, such as recommendation, in order to increase the empathy between users and recommendation systems.

References

[1] Teresa Cadierno and Lucas Ruiz. Motion events in Spanish L2 acquisition. Annual Review of Cognitive Linguistics,4:183–216, 2006.

[2] David Crystal. The scope of Internet Linguistics. American Association for the Advancement of Science, 2005.

[3] Edward Fry. Readability. In Reading Hall of Fame Book. February 2006.

[4] William Labov. Principles of Linguistic Change: Social Factors. Blackwell, Malden, MA, 2001.

[5] Marco Lui and Timothy Baldwin. langid.py: an off-the-shelf language identification tool. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 25–30, 2012.

[6] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. New kid on the block: exploring the Google+ social graph. In Proceedings of the 2012 ACM Internet Measurement Conference (IMC'12), pages 159–170, New York, NY, USA, 2012. ACM.

[7] James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The Development and Psychometric Properties of LIWC2007. The University of Texas at Austin and The University of Auckland, New Zealand, 2007.

[8] United Nations. Human Development Report 2013 – The Rise of the South: Human Progress in a Diverse World. United Nations Development Programme, New York, NY, 2013.

[9] U.S. Bureau of Labor Statistics. Standard occupational classification and coding structure, February 2010.