How You Post Is Who You Are: Characterizing Google+ Status Updates across Social Groups

Evandro Cunha[1,2], Gabriel Magno[1], Marcos André Gonçalves[1],
César Cambraia[2] and Virgilio Almeida[1]
[1]Dept. of Computer Science, [2]College of Letters - Universidade Federal de Minas Gerais, Brazil
{evandrocunha, magno, mgoncalv, virgilio}@dcc.ufmg.br, nardelli@ufmg.br

ABSTRACT

The analysis of user-generated content on the Web provides tools to better understand users' behavior and to the development of improved Web services. Here, we consider a large dataset of Google+ status updates to evaluate linguistic features among members of distinct social groups. Our study reveals that groups hold linguistic particularities – such as a tendency to use professional vocabulary, suggesting that Google+ might be employed, by certain users, for professional activities, or that members do not dissociate from their jobs when interacting in this environment. To illustrate a possible application of our outcomes, we present a classification experiment aiming to infer users' social information through the analysis of their posts, with satisfactory preliminary results. Our findings help to understand not only collective peculiarities of online social media users, but also important characteristics of the textual genre *post*, being one of the first and most comprehensive studies on this topic.

Categories and Subject Descriptors: J.5 [Computer Applications]: Arts and Humanities—*Linguistics*

Keywords: OSNs; Google+; Internet linguistics; Microtext analysis

1. INTRODUCTION

Increasingly, researchers have taken advantage of the vast amount of language data that online applications can provide, which gave rise to a new subfield of knowledge called *Internet linguistics* [5]. According to Crystal [4], the Internet plays an unprecedented role in the study of language, as it allows linguists to use rich documented datasets to investigate language use in various levels and the nature of the language employed by Web users. From this perspective, authors are concerned with understanding and describing computer-mediated communication, as well as developing tools to provide better online services. Opportunities arising in this area include the employment of collections from user-generated content websites as corpora of large-scale natural language data.

Google+ is an online social network (OSN) launched in June 2011. To better understand its typical features, the investigation of formal and functional aspects of the content shared by its members is of utmost importance. Here, 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 texts from male and female members from 10 countries and 15 groups of occupations, since gender, location and job are known as factors that influence language usage in a myriad of domains [12]. Our main hypothesis is that the membership in certain social groups may influence aspects of the language employed by users when posting, reflecting patterns observed in other online and offline situations.

To our knowledge, this work is novel in that it is the first to focus on linguistic aspects of Google+ posts. Moreover, it contributes to the general literature on Internet linguistics by investigating the role of social factors in relevant aspects of language use in social media. Possible applications of this study include the development of improved Web services, as discussed in the following sections.

2. RELATED WORK

On Google+. An analysis of Google+ social graph is presented by Magno et al. [15], who studied structural properties of this network in comparison to other services and found different patterns of its usage across distinct countries. They discovered, among other findings, that Google+ is popular in countries with relatively low Internet penetration rate and that its top users are not celebrities or public figures, but mainly individuals in the IT industry.

A study on how members organize and select audiences for shared content in Google+ was conducted by Kairam et al. [11]. An interesting result is that users weigh limiting factors, like privacy, against the desire to reach a large audience. Gonzalez et al. [9] showed that, despite the recent growth of this OSN, the relative size of its largest connected component has decreased with time and that only a few users exhibit any type of activity.

On language and social factors. Literature on the relations between language and society is now really vast. Labov's [12], Trudgill's [26] and Romaine's [22] works present the main findings of decades of research, considering also the correlations between language variation and the social factors that we contemplate here.

Bell et al. [1] used computational tools to investigate differences in language styles among men and women. Their finding that women use more social words than men could be verified by our analysis. It is also worth mentioning Lakoff's [13] seminal work on language and gender, where the author indicates that a number of linguistic features can distinguish men's speech from women's.

The study of linguistic styles associated with particular professions was performed by Jones [10]. However, our approach that identified the use of professional vocabulary in posts published in an online social network seems to be an original contribution.

On language use in social media. The study of topics from Facebook posts was performed by Wang et al. [28]. They demonstrated that women are more likely to write posts about personal themes,

HT'14, September 1–4, 2014, Santiago, Chile.
Copyright 2014 ACM 978-1-4503-2954-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2631775.2631822.

contrasting with men, who tend to share more public subjects, like politics and sports. Even though we study another OSN, this finding relates to the prevalence of usage of words from categories like *family*, *social* and *affection* by female users in our dataset.

An investigation on how men and women differ when designating hashtags on Twitter was carried out by Cunha et al. [6], who found that, in the context of political debate, Brazilian women are more prone to use approaches based on solidarity, while men tend to employ assertive strategies. Ottoni et al. [16] examined users' descriptions on Pinterest and showed differences in the linguistic style between genders, being women more likely to use words of fondness and affection. Schwartz et al. [24] investigated the relation between language and different variables on Facebook, and found associations between personality and language use of given groups.

Studies that performed gender and location prediction of users based on the written content posted by them will be referenced ahead, where we present the results of our classification experiment.

Although we found these and other investigations on language use in OSNs, we did not find studies that considered, simultaneously, all social factors and linguistic attributes examined here.

3. METHODOLOGY

Data collection. From March 23rd to June 1st, 2012, we collected profile information and posts of Google+ users. For ethical and legal reasons, we gathered only public information revealed in users' pages and did not attempt to obtain access to information set as private. We inspected the *robots.txt* file provided by Google+, followed the corresponding sitemap to retrieve the lists of URLs of profiles to be collected and then made HTTP requests to the pages. Since we collected the complete list of profiles provided, we believe we retrieved information from all users with public pages at the time of the collection, compiling information from 160,304,954 profiles.

Posts. Among the profiles collected, only 8,564,462 set their posts as publicly available. We were able to retrieve up to the last ten status updates from each user's page, totaling 29,366,310 posts.

To select only messages generated in English, we used *langid.py* [14], a language identification tool that identified 20,928,557 posts probably written in this language. In order to increase the confidence that our posts are actually in near-standard English – thus avoiding the analysis of posts only partially produced in this language or written in dialects, mixed varieties or fused lects –, we additionally filtered texts with probability of at least .99 of being in English. After this restriction, we narrowed our dataset down to 7,414,679 posts. A manual evaluation of several filtered posts indicated that they were indeed written in near-standard English.

Since we aimed to analyze language characteristics of individuals, we alleviated the impact of copied posts, like chain letters and other highly replicated texts, by removing duplicated messages. We identified 265,100 types of texts that presented duplication, totaling 1,220,341 repeated posts, and removed them all from the dataset. Therefore, at this point we have 6,194,338 distinct Google+ posts.

Figure 1 displays a general characterization of distinct Google+ posts written in English. The first two graphics show, respectively, cumulative distribution functions of numbers of characters and words per post. On average, posts have 111.2 characters and 25.6 words. The third graphic indicates that the majority of posts have only a few sentences: 53% of them have one sentence, while 26% have two and 10% have three sentences. This shows 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. Besides the posts, we collected information on users' location, gender and professional activity.

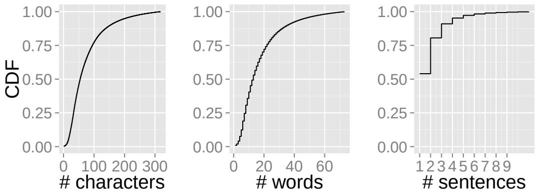

Figure 1: Cumulative distribution functions of numbers of characters, words and sentences per post

Location. We inferred users' location using information available in the field *Places lived*, in which members can create a list of places where they have lived. This is an open field, meaning that users can type any text they want to. Therefore, the same place can be written in different ways (e.g. *New York*, *NYC*, *New York City*) or using distinct geographic levels (e.g. *Los Angeles*, *California*, *USA*).

To identify an user's country, we extracted the geographic coordinates of the last location cited and translated them into a valid country identifier. In this fashion, we were able to identify the country of 22,578,898 members (14.08% of the full dataset). Remaining users set this information as private or simply did not fill this field.

Here, we consider only members located in the ten countries with most posts in English: United States (US), Great Britain (GB), India (IN), Canada (CA), Australia (AU), Indonesia (ID), Germany (DE), Philippines (PH), Malaysia (MY) and France (FR). Figure 2 summarizes the process of posts collection until this point.

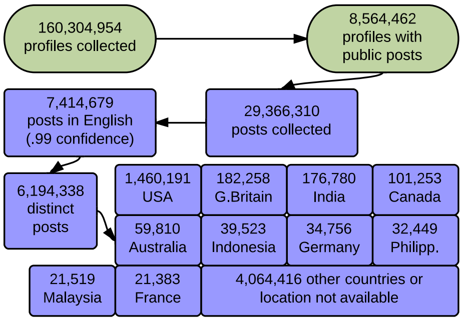

Figure 2: Description of the posts collection

Gender. Gender information is shared by 126,531,842 users (78.93% of the complete dataset) and by 770,997 users with posts collected in the ten countries studied. Considering members who set this information publicly available, 63.77% chose *male*, 34.38% chose *female* and 1.85% chose *other*. Here, we do not consider users who set their own gender as *other*.

Professional activity. The field *Occupation* is an open field, so users can type any text they want to in order to describe their activity. As a result, we gathered a very large number of different occupations and had to summarize the information introduced by users: first, we manually aggregated the most common strings present in the dataset, since the same occupation can be written in different ways (e.g. *student*, *study*, *graduate student*, *go to school*); second, we selected the top 30 occupations; third, we used the Standard Occupational Classification (SOC) by the U.S. Bureau of Labor Statistics [27] to divide these occupations into the major groups of professional activities used here. The occupations *student* and *retired*, although not shown in the SOC, are also considered in our analyses.

Table 1 shows the number of posts and users per social group in our dataset.

4. ANALYSES

In this section, we present the linguistic analyses performed on Google+ posts. They are all independent investigations, not neces-

Social group	# posts	# users	Social group	# posts	# users
Country			Occupation		
United States (US)	1,460k	494k	Student	85k	36k
Great Britain (GB)	182k	62k	Computer and math.	61k	19k
India (IN)	177k	96k	Arts and design	25k	7,9k
Canada (CA)	101k	34k	Archit. and engin.	15k	6,0k
Australia (AU)	60k	21k	Business and financ.	11k	3,9k
Indonesia (ID)	40k	24k	Media	8,3k	2,1k
Germany (DE)	35k	15k	Educ. and library	6,7k	2,2k
Philippines (PH)	32k	14k	Management	5,9k	1,9k
Malaysia (MY)	22k	10k	Sales	4,6k	1,6k
France (FR)	21k	10k	Legal	2,6k	0.8k
			Retired	2,2k	0.9k
Gender			Healthcare	1,9k	0.8k
Male	1,549k	557k	Religious	1,5k	0.4k
Female	526k	203k	Science	1,2k	0.4k
Other/NA	55k	18k	Food preparation	0.7k	0.3k
			Other/NA	1,897k	695k

Table 1: Number of posts and users per social group (round)

sarily examining the same text attributes, which makes it possible to test distinct aspects of language behavior. It is important to note that the results presented here apply only to language behavior in the specific context of Google+ and may not be valid for offline environments or even for other online social networking systems.

4.1 Misspellings

The occurrence of misspelled words in texts may signify unawareness of standard orthographic rules or carelessness during typing, due to negligence or lack of revision. Thus, calculating the extent to which misspellings emerge in our dataset might indicate how high literacy levels in English of the communities are or how concerned individuals are about the quality of their posts – since, for most users, it may not matter whether they make misspellings in OSN posts. In other cases, non-standard spellings may be on purpose, in order to create specific effects on readers.

By using a list of 4,238 common misspellings in English (available at `http://bit.ly/1ieaEOa`), that encompasses 31.3% of the whole vocabulary employed in the dataset, we investigated the occurrence of these non-standard linguistic elements in Google+ posts produced by different social groups. This list, that considers spelling differences in distinct varieties of the language, comprises misspelled items and their corresponding standard spellings, which are, therefore, the only words susceptible to misspelling in our analysis. Only cases of sequences of letters representing no standardly spelled words in English are included in the list, and homographs are not considered. We applied this approach instead of running a spelling checker on the posts in order to avoid the classification of intentional non-standard spellings (such as *gonna*, *doin'*, *ur* and many other common spellings on Web environments, which are not included in the list) as misspellings.

We calculated the fraction of misspellings per post by dividing the number of misspelled words by the number of words susceptible to misspelling. To avoid biases due to the small number of words susceptible to misspelling in some posts (e.g. if a post has only one word susceptible to misspelling, its fraction of misspellings is either 0 or 1), we did not consider posts with less than five words that appear in our list, thus evaluating 758,233 posts.

Figure 3 exhibits the average fractions of misspellings per post. It shows that, as expected, non native English speakers, with exception of French users, are more prone to make misspellings in English written posts. We also found that, in general, women's fraction of misspellings is higher than men's: we believe that the difference between topics discussed by men and women – as will be seen in Section 4.4 – does not force women to be so demanding on the formal linguistic attributes of the content published.

Figure 3 also states that workers who deal more with written texts make fewer misspellings in Google+ posts: while food and health professionals have the highest fractions of misspellings, media, legal and education professionals have the smallest ones. It is worth remembering that, by the nature of these occupations, review of written material is sometimes an activity performed daily.

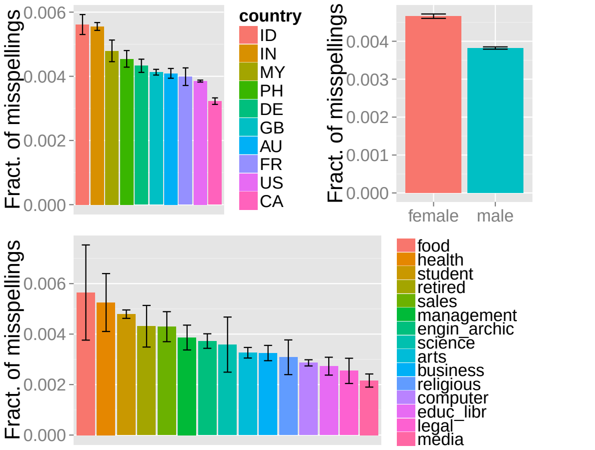

Figure 3: Average fractions of misspellings per post for different countries, genders and occupations ± standard errors

4.2 Readability and complexity

The readability of a text is the ease in which readers can properly comprehend it. A series of formulas that return numerical scores estimating the level of difficulty of texts have already been proposed [8] and should not be seen as metrics of quality of documents, since *easier* or *more difficult* texts are not necessarily *worse* or *better* texts. Here, we employ a readability index to diagnose differences in the organization of speech by distinct groups in Google+.

The Unix command *style* returns results for the Automated Readability Index (ARI), which calculates the readability of a text using the formula $ARI = 4.71 \cdot \frac{\#of characters}{\#of words} + 0.5 \cdot \frac{\#of words}{\#of sentences} - 21.43$. The ARI relies mostly on a factor of characters per word and, on a lesser extent, on a factor of words per sentence. Thus, its assumption is that the adoption of big words and the construction of large sentences are features that enhance the complexity of a text: other aspects being equal, smaller words and shorter sentences should result in increases of comprehension [25].

Figure 4 depicts average values of ARI for distinct groups. Higher scores indicate higher complexity, as they correspond to bigger words and sentences. According to our results, texts of German, French and Indian users on Google+ are the most complex ones; on the other side, posts of Malaysians, Filipinos and Indonesians are the least complex. Interestingly, native speakers of English – from Australia, Great Britain, Canada and USA – present the central values, which seems to indicate that non native English speakers must have transferred linguistic patterns of their mother tongues to the foreign language [3]. This hypothesis is strengthened when we observe that users from countries with prevalence of speakers of Indo-European languages have the highest values of ARI and those from countries with prevalence of speakers of Austronesian languages have the lowest indices. We also observed that the average number of characters per word is very similar across countries, showing that, in this case, the discriminant factor of the readability index is the number of words per sentence, which may be highly influenced by the linguistic structures of mother tongues.

ARI scores for female and male users show that posts written by men are, on average, more complex than those written by women. This fact is observed for most countries and professions. The exami-

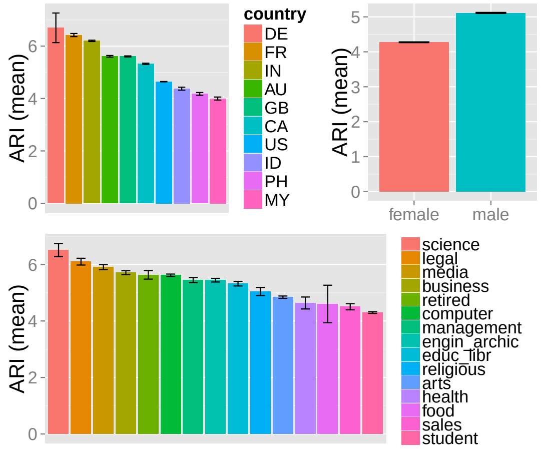

Figure 4: Average values of ARI for posts of users from different countries, genders and occupations ± standard errors

nation of the complexity of posts of users with different occupations can be related to the previous analysis on misspellings: in the same way that workers from fields more associated with written communication and traditionally elaborated texts, like legal and media professionals, publish texts with fewer misspellings, they also produce more complex posts than those from fields that do not necessarily deal with written texts, like food preparation and sales professionals. Ahead, in Section 4.4, 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, about topics that ask for either more or less elaborated linguistic constructions, according to their respective occupations.

4.3 Vocabulary variability

We also considered vocabulary variability – through an entropy-based approach – across different groups, since this could add relevant insights into statistical regularities of the language employed by users. Differences of entropy values are related to the specific style of each community: lower values mean more predictable word usage, while higher ones mean more vocabulary variability.

After removing stopwords and applying stemming based on Porter's algorithm [19], we calculated Shannon's entropy of the concatenation of all posts from each group. Since the number of users in each group differs and the number of unique words is directly affected by the total number of words, we applied an undersampling methodology across our three categories of social groups. We repeated this process 25 times and calculated the mean.

No significant differences among entropy values of different groups were found, indicating that they are not discriminant on the variability of vocabulary in the context of Google+ posts and denying our hypothesis that vocabulary variability in this OSN varies among posts written in English by users from different countries, genders and occupations.

4.4 Semantic categories of words

An interesting way of investigating language differences across groups is through the analysis of the vocabulary used by their members. Since vocabulary is a system of mapping the world, this kind of investigation reveals how groups perceive reality, indicating what the main concerns and interests of certain communities might be.

We aim to identify if some given semantic categories of words are more common in texts produced by members of particular countries, genders and occupations. To accomplish this task, we used the Language Inquiry and Word Count (LIWC) [18], a tool that examines texts and verifies the occurrence of words previously classified as members of functional/grammatical (e.g. pronouns, articles, prepositions etc.) or semantic (e.g. social, money, religion etc.) categories. A comprehensive list of all LIWC categories, including examples of words that are part of each category, is available at **`http://www.liwc.net/descriptiontable1.php`**.

We calculated LIWC scores for a given category of words as the fraction of words of this category in the total amount of categorized words of a particular post. After having calculated LIWC scores for 41 categories of semantic words, we compared them across the social groups. Figure 5 shows the categories of words with most significant differences across the groups considered in this study.

We observed that users from different countries hold distinct patterns in the usage of certain semantic categories of words in their posts. For example, Indians have the highest scores in the use of words from categories such as *friend*, *humans* and *social*, while they have low scores in categories like *negative emotions*, *anger* and *time*. Also, users from most of the Western countries considered here tend to be the main users of words related to *home*, *money* and *work* and the least users of words from the categories *health*, *affection*, *positive emotions* and *family*. These categories might be revealing the topics more covered in the posts and are a sign of cultural differences among users from different countries, which is relevant for the literature on comparative cultural studies, interested in investigating cultures in global and intercultural contexts [20].

Considering gender, we found that women are more prone to use words from categories such as *family*, *home*, *friend*, *social*, *humans*, *affection* and *emotions*, while men are the main adopters of words from categories like *cause*, *motion*, *space*, *numbers*, *money* and *work*. 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. These distinct approaches toward this specific online social networking service may also be the reason why men's posts are more complex and formally accurate, having fewer misspellings, as described in the Sections 4.1 and 4.2 above.

We also found a clear correlation between word usage and users' occupations. For instance, words related to religion are extremely more frequent in posts from religious professionals; the same for money vocabulary in posts from salespeople, 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 vocabulary employed in Google+ posts is highly related to users' working activities, indicating that this OSN may be often used for professional activities or that members do not dissociate from their jobs when interacting in Google+, maintaining their professional vocabulary even in this environment. This result has important implications for the literature on cognitive linguistics, since it reinforces the view that individuals' conceptual maps – represented by their vocabulary – is strongly related to their jobs.

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 online social media.

4.5 Inference of social groups

To illustrate a possible application of these results, we propose the task of inferring social characteristics of users based on linguistic analysis of their posts. This type of application is useful to assist in the development of tools aiming authorship attribution for purposes like personalization of services and identification of fake profiles.

We conducted a preliminary classification experiment using textual metrics contemplated above. For each user, we created a vector

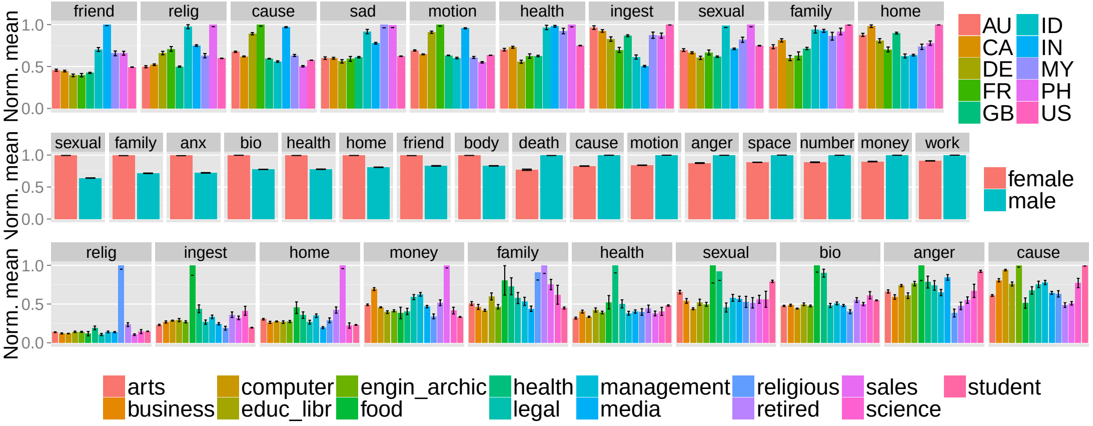

Figure 5: Semantic categories of words with most significant differences across distinct groups of users (countries, genders and occupations, respectively) ± standard errors

containing 76 features: 4 size metrics (numbers of characters, words, sentences and paragraphs per post), 7 readability indices (ARI and other indices provided by the Unix command *style*), 64 LIWC categories (including categories of semantic words and categories of grammatical and function words) and fraction of misspellings.

We sought to make inferences by using support vector machine classifier (SVM) and the *scikit-learn* library [17] to conduct the SVM classification and parametrization. For the experiments, we employed a 5-fold cross-validation technique randomly selecting a fixed number of users per class: 1,000 for countries and genders; 259 – the number of members in the smallest occupation group – for occupations. The results reported in Table 2 are the averages of the 25 runs and their respective confidence intervals at 95%.

Table 2 shows that, when using our vector of linguistic features, the SVM classifier increased in 19.7% (for genders), 83.0% (for countries) and 134.6% (for occupations) the accuracies of the inferences if compared to a random classifier. It also depicts values of F1 per class, indicating that some groups – like Indians ans religious professionals – are much more easily identified by our classifier than others – like Australians and architects/engineers.

	Accuracy random	Accuracy SVM	F1 weighted
Country	0.1000	0.1830±0.0032	0.1788±0.0027
Gender	0.5000	0.5985±0.0093	0.5768±0.0079
Occupation	0.0666	0.1563±0.0054	0.1515±0.0044

Social group	F1	Social group	F1
Country		Occupation	
India (IN)	0.2593	Religious	0.4191
Philippines (PH)	0.2365	Sales	0.2277
Indonesia (ID)	0.2030	Retired	0.1879
United States (US)	0.1910	Media	0.1761
Canada (CA)	0.1851	Business and financial	0.1465
Great Britain (GB)	0.1845	Healthcare	0.1393
France (FR)	0.1605	Legal	0.1364
Germany (DE)	0.1553	Student	0.1354
Malaysia (MY)	0.1148	Computer and mathematical	0.1227
Australia (AU)	0.0990	Arts and design	0.1177
		Education and library	0.1075
Gender		Management	0.0994
Male	0.6179	Science	0.0931
Female	0.5768	Food preparation	0.0672
		Architecture and engineering	0.0463

Table 2: Results of the inference experiments

Other studies already proposed solutions for gender classification in different online social systems. Schler et al. [23], who investigated language use in blogs, achieved up to 80.1% of accuracy in this task; Burger et al. [2], in their Twitter classifier relying only on text attributes, achieved 75.5% of accuracy; and Rao et al. [21], who also studied Twitter, achieved up to 72.33% of accuracy. Although the accuracy of our preliminary gender classifier is not high if compared to these previous ones, we believe that they and other classifiers can benefit from the use of some of the features proposed here.

Eisenstein et al. [7] addressed the issue of inferring users' geographic location from Twitter texts. Differently from us, they only considered users from different states in the United States, which makes comparison between our and their studies quite difficult. The task of predicting the professional activity of OSN users, however, seems to be an unexplored subject, since we did not find studies regarding the inference of occupations in online systems.

We advocate, then, that our vector of linguistic features can be used in conjunction with other metrics, such as profile information, network topology and other linguistic metrics, with the goal of increasing the quality of predictors of social characteristics of members in information networks.

5. CONCLUDING REMARKS

In this study, we considered a large dataset of Google+ posts to evaluate linguistic elements among members of particular social groups. These analyses not only describe the posts, but especially identify how distinct groups differ when posting content on the Web.

To the extent 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 OSN. Also, we contemplated language attributes and social characteristics that have been underinvestigated in other studies on language use in social media.

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 authorship attribution may follow. For this reason, we implemented a preliminary classifier to infer social characteristics of Google+ users, which may be an useful tool to improve the task of automatically detecting fake profiles through the analysis of their linguistic behaviors and to improve language modeling focused on personalization of services.

Future work should include the analysis of other relevant linguistic and social factors, such as the topic of posts and the educational level of users. Also, it would be interesting to compare the outcomes reported here for Google+ with other popular OSNs, such as Facebook and Twitter. Another related issue to be analyzed in future studies is the question of how these different social groups express their feelings on the Web and which linguistic elements are used to indicate tones of happiness, angriness, hope and hatred, among others: are these elements also distinctive across different social groups in the context of online social networking services?

6. REFERENCES

[1] C. M. Bell, P. M. McCarthy, and D. S. McNamara. Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles. In P. M. McCarthy and C. Boonthum-Denecke, editors, *Applied Natural Language Processing: Identification, Investigation, and Resolution*. Information Science Reference, Hershey, PA, 2012.

[2] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[3] T. Cadierno and L. Ruiz. Motion events in Spanish L2 acquisition. *Annual Review of Cognitive Linguistics*, 4:183–216, 2006.

[4] D. Crystal. *The Language Revolution*. Polity Press, Cambridge, UK, 2004.

[5] D. Crystal. The scope of Internet Linguistics. *American Association for the Advancement of Science*, 2005.

[6] E. Cunha, G. Magno, M. A. Gonçalves, C. Cambraia, and V. Almeida. He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLOSONE*, 9(1):e87041, January 2014.

[7] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[8] E. Fry. Readability. In *Reading Hall of Fame Book*. February 2006.

[9] R. Gonzales, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. Google+ or Google-? Dissecting the evolution of the new OSN in its first year. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW 2013)*, 2013.

[10] N. L. Jones. Talking the talk: the confusing, conflicting and contradictory communicative role of workplace jargon in modern organizations. Master's thesis, University of Rhode Island, 2011.

[11] S. Kairam, M. J. Brzozowski, D. Huffaker, and E. H. Chi. Talking in circles: selective sharing in Google+. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'12)*, 2012.

[12] W. Labov. *Principles of Linguistic Change: Social Factors*. Blackwell, Malden, MA, 2001.

[13] R. Lakoff. *Language and Woman's Place*. Harper and Row, New York, NY, 1975.

[14] M. Lui and T. Baldwin. langid.py: an off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 25–30, 2012.

[15] G. Magno, G. Comarela, D. Saez-Trumper, M. Cha, and V. Almeida. New kid on the block: exploring the Google+ social graph. In *Proceedings of the ACM Internet Measurement Conference (IMC'12)*, 2012.

[16] R. Ottoni, J. P. Pesce, D. Las Casas, G. Franciscani Jr., W. Meira Jr., P. Kumaraguru, and V. Almeida. Ladies first: Analyzing gender roles and behaviors in Pinterest. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, 2013.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. *The Development and Psychometric Properties of LIWC2007*. The University of Texas at Austin and The University of Auckland, New Zealand, 2007.

[19] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14:130–137, 1980.

[20] Purdue University Press. Comparative Cultural Studies. **http://bit.ly/1gDkJUL**, Retrieved in March 2014.

[21] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 37–44. ACM, 2010.

[22] S. Romaine. *Language in Society: An Introduction to Sociolinguistics*. Oxford University Press, 1994.

[23] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

[24] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOSONE*, 8(9):e73791, 2013.

[25] C. Swanson and H. Fox. Validity of readability formulas. *The Journal of Applied Psychology*, 37(2), 1953.

[26] P. Trudgill. *Sociolinguistics: an introduction to language and society*. Penguin, London, UK, 1983.

[27] U.S. Bureau of Labor Statistics. Standard occupational classification and coding structure. **http://1.usa.gov/14INxmQ**, February 2010.

[28] Y. C. Wang, M. Burke, and R. Kraut. Gender, topic, and audience response: An analysis of user-generated content on Facebook. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'13)*, 2013.