Language Matters in Twitter: A Large Scale Study

Lichan Hong[1], Gregorio Convertino[1], and Ed H. Chi[2]

[1]Palo Alto Research Center, Palo Alto, CA, USA
{hong, convertino}@parc.com

[2]Google, Mountain View, CA, USA
chi@acm.org

Abstract

Despite the widespread adoption of Twitter internationally, little research has investigated the differences among users of different languages. In prior research, the natural tendency has been to assume that the behaviors of English users generalize to other language users. We studied 62 million tweets collected over a four-week period and found that more than 100 languages were used. Only half of the tweets were in English (51%). Other popular languages including Japanese, Portuguese, Indonesian, and Spanish together accounted for 39% of the tweets. Examining users of the top 10 languages, we discovered cross-language differences in adoption of features such as URLs, hashtags, mentions, replies, and retweets. We discuss our work's implications for research on large-scale social systems and design of cross-cultural communication tools.

Introduction

Despite the widespread adoption of Twitter internationally, little is known about cross-language differences in Twitter. While recent research has indicated that multiple languages are used in Twitter, there is no systematic analysis about the differences across these languages. A small qualitative analysis by Honeycutt and Herring (2009) found that English, Japanese, and Spanish were the most used languages. Examining 2.8 million tweets collected over 48 hours, Semiocast (2010) found that the top five languages were English (50%), Japanese (14%), Portuguese (9%), Malay (6%), and Spanish (4%).

To the best of our knowledge, our work is the first to systematically study how users of different languages behave in Twitter. We address two main questions:
(1) What is the frequency distribution of the *top languages* used in Twitter? Can we validate and extend prior results?
(2) Are there noticeable *behavior differences* exhibited by users of different languages? How do users of various languages differ in their inclusion of URLs, hashtags, and mentions? Do they reply or retweet at similar rates?

To address these questions, we analyzed 62 million tweets, grouped users into language communities, and then compared key behaviors across communities. In this paper we report two contributions. First, we identify the top 10 languages used in Twitter, extending prior research (Honeycutt & Herring 2009, Semiocast 2010). Second, we show significant differences across the top 10 languages in the use of Twitter-specific conventions such as URLs, hashtags, mentions, replies, and retweets.

Related Work

Several studies have characterized how people use Twitter. Honeycutt and Herring (2009) studied the functions of the @ symbol in English tweets. boyd et al. (2010) examined the practice of retweeting. Kwak et al. (2010) found that 67.6% of users were not followed by any of the people whom they followed (i.e., their followees), and conjectured that these users adopted Twitter as an information source rather than a site for social networking.

Unsurprisingly, most of these findings and methods were developed using English tweets. We hypothesize that they may not apply as well to non-English tweets. Only a few studies explicitly accounted for the international nature of Twitter's proliferation (Java et al. 2007, Krishnamurthy et al. 2008). In contrast, since the Web is multilingual, researchers studying cultural differences have been fascinated by the role that languages play online, and how they influence the adoption of online tools (Kayan et al. 2006, Herring et al. 2007, Hecht & Gergle 2010). These findings suggest that language and associated cultural differences matter indeed and need to be considered when designing cross-cultural communication tools or when studying their usage.

Characterizing the Top Languages in Twitter

We collected 62,556,331 tweets from the Spritzer sample feed of Twitter over a four-week period (April 18 - May 16, 2010), using the streaming API. The Spritzer sample represents a random selection of all public messages. On average, we gathered 2.2 million tweets per day, representing 3-4% of all public messages. Then, we identified the language of each tweet using a combination of LingPipe's text classifier[1] and Google Language API[2].

We identified 104 languages from the 62 million tweets. Table 1 shows the top 10 languages, ordered by decreasing number of tweets. Note that if a user posted tweets in multiple languages, we counted her multiple times. The top 10 languages accounted for 95.6% of all the tweets. We conducted a human-coding study of a random sample of 2,000 tweets, and found a substantial agreement between human judges and the language detection algorithm (Cohen's kappa above 0.90 for the top 7 languages).

Language	Tweets	%	Users	Tweets/user
English	31,952,964	51.1	5,282,657	6
Japanese	11,975,429	19.1	1,335,074	9
Portuguese	5,993,584	9.6	993,083	6
Indonesian	3,483,842	5.6	338,116	10
Spanish	2,931,025	4.7	706,522	4
Dutch	883,942	1.4	247,529	4
Korean	754,189	1.2	116,506	6
French	603,706	1.0	261,481	2
German	588,409	1.0	192,477	3
Malay	559,381	0.9	180,147	3

Table 1. The top 10 languages in Twitter.

Characterizing Differences across Languages

Including URLs and Hashtags

To share information, Twitter users include URLs or links in their tweets. We found that 21% of our 62 million tweets contained URLs. But this percentage changed across languages, as shown in the second column of Table 2. For example, we see that 39% of German tweets and 37% of French tweets included URLs, more than twice the percentages of Japanese and Portuguese tweets.

We identified the most popular linked websites within each language community. Specifically, we expanded shortened URLs into their original URLs and then extracted the domain names. In this way, we computed a list of popular domains cited by each community, ranked by decreasing number of tweets. The top five domains in English tweets were: twitpic.com, www.youtube.com, www.facebook.com, www.twitlonger.com and www.formspring.com. Clearly, some websites were designed for international audiences; for example twitpic.com appeared in the top five domains of eight communities. Others mostly targeted at local users; for example, nicovideo.jp was only popular among Japanese users.

For each pair of language communities, we assume that the number of URL domains cited by both communities indicates how often these two communities visited the same content sources. Based on this assumption, we took the top 100 domains of each community and computed the number of domains shared by each pair of communities, as shown in Table 3 (above diagonal).

Language	URLs	Hashtags	Mentions	Replies	Retweets
All	*21%*	*11%*	*49%*	*31%*	*13%*
English	25%	14%	47%	29%	13%
Japanese	13%	5%	43%	33%	7%
Portuguese	13%	12%	50%	32%	12%
Indonesian	13%	5%	72%	20%	39%
Spanish	15%	11%	58%	39%	14%
Dutch	17%	13%	50%	35%	11%
Korean	17%	11%	73%	59%	11%
French	37%	12%	48%	36%	9%
German	39%	18%	36%	25%	8%
Malay	17%	5%	62%	23%	29%

Table 2. Percentages of tweets using various conventions.

	E	J	P	I	S	D	K	F	G	M	*Av (Sd)*
E		17	23	15	24	22	19	24	21	22	21 (3.2)
J	6		14	13	16	16	18	16	14	16	16 (1.6)
P	9	1		14	22	18	13	19	16	19	18 (3.6)
I	10	2	2		14	16	12	15	14	65	20 (17)
S	10	2	9	7		22	16	22	17	20	19 (3.5)
D	16	3	4	8	9		16	18	18	20	18 (2.4)
K	1	1	1	1	1	1		16	14	15	15 (2.2)
F	22	5	4	8	11	18	1		19	20	19 (2.9)
G	20	6	4	5	9	10	1	26		16	17 (2.5)
M	13	2	3	76	7	11	1	10	6		24 (16)
Av (Sd)	12 (6.7)	3 (2.0)	4 (3.0)	13 (24)	7 (3.5)	9 (5.7)	1 (0.0)	12 (8.6)	10 (8.1)	14 (23)	

Table 3. (Above diagonal) Number of cited URL domains (out of top 100) shared by pairs of language communities. (Below diagonal) Number of hashtags (out of top 100) shared by pairs of language communities.

Hashtags are free-form tags or keywords included in tweets, in the form of #keyword. Including a hashtag parallels using a tag in a social bookmarking system and creates a venue for collecting all tweets about a topic. We found that 11% of our tweets included hashtags, but with clear cross-language differences. For example, the third column of Table 2 shows that hashtags were included in 18% of German tweets but only in 5% of Japanese tweets.

We also tabulated popular hashtags used by each language community, ranked by decreasing number of

[1] http://alias-i.com/ lingpipe/demos/tutorial/langid/read-me.html
[2] http://code.google.com/apis/language/

tweets. The top five hashtags used in Japanese tweets were: #anisonzanmai, #followmejp, #nicovideo, #sougofollow, and #nhkfm. We found that #nowplaying appeared in the top five hashtags of eight communities, while the top three hashtags used by Korean users were all in Korean.

We assume that the number of hashtags commonly used by a pair of language communities indicates how likely these two communities were to post messages about the same topics. We considered the top 100 hashtags of each community, and then for each pair of communities we counted the number of hashtags that were used in both communities (see Table 3, below diagonal).

As indicators of content sharing, the number of top 100 URL domains and the number of top 100 hashtags shared by a pair of communities were closely related. We found high correlations between the counts of URLs and hashtags shared (i.e., counts below vs. above diagonal in Table 3, Pearson $r=.91$, $p<.01$) and between the average values per language (i.e., rightmost column vs. bottom row in Table 3, Pearson $r=.83$, $p<.01$).

Including Mentions, Replies, and Retweets

Twitter users can refer to a specific user by including a mention anywhere in their tweets, done in the form of @username. We found that mentions were widely adopted. Indeed, 49% of our tweets contained mentions. However, the fourth column of Table 2 shows that 73% of Korean tweets and 72% of Indonesian tweets contained mentions, but only 36% of German tweets did so. A mention is generally used to either attract someone's attention or acknowledge someone's association to the content of the tweet. Both are cases of inherently social acts and resemble public conversations in groups of people. Interestingly, some communities such as Korean and Indonesian exhibited more of this social behavior than others.

A reply, a specific form of mention with @username appearing at the beginning of the tweet, is a tweet responding to a previous message. Inspecting the metadata of our tweets, we found that 31% of them were replies. The percentage of tweets that are replies within a language community might be interpreted as an indicator of the strength of communication, which should reflect social ties. The fifth column of Table 2 shows that only 20% of Indonesian tweets were replies. In contrast, 59% of Korean tweets were replies, suggesting stronger social ties among Korean users and a specific type of use of Twitter for communicating directly with known contacts.

Retweeting is typically used to spread information received from followees to followers (boyd et al. 2010). Similar to forwarding an email, retweeting is both an information-sharing act (i.e., to spread information) and a social act (i.e., to recognize and promote someone's message). A common form of retweeting is "RT @username message", where "message" is a tweet created by "username". Users have also adopted a variety of other syntactical markers such as "RT:@", "retweeting @", "retweet @", "(via @)", "RT (via @)", "thx @", "HT @", and "r @" (boyd et al. 2010). Scanning for these markers, we found that 13% of our tweets were retweets.

The rightmost column of Table 2 shows the differences in how frequently users in different communities retweeted. 39% of Indonesian tweets were retweets, while only 7% of Japanese tweets were retweets. Different from replying, which is a one-to-one communication, retweeting aims to broadcast a message to a broad audience. Our results suggest that Indonesian users were far more likely to spread information via retweets than Japanese users.

Relating Cross-Language Differences in Behaviors

To compare how the above-mentioned behaviors differed across languages, we computed how frequently each behavior was exhibited by users of a language community with respect to English as a fixed language of reference. We used English as our reference, because it is the native and largest language community in Twitter.

To assess the differences we used Chi Square tests and a logistic regression analysis with language as the categorical factor. Across languages, this analysis provided us with odds ratios as a relative measure of differences on each behavior. We fitted a binary logistic model for each behavior (e.g., with vs. without URLs) and considered Language as our categorical predictor with multiple levels (i.e., using a dummy variable per language) (Agresti 2002).

We found that the language communities differed significantly on each behavior. Likelihood Ratio and Wald Chi Square tests confirmed the significant differences by language for each of the five behaviors ($p<.0001$, $df=9$). This was also supported by the computed odds ratios in each of the behaviors. Each ratio indicates that the odds for a given behavior in a language community (e.g., German) are X times the odds for the same behavior in the English community (Agresti 2002). For example, compared to the odds of including a URL in an English tweet, the odds of the same behavior in a German tweet increase 89% (1.89) and decrease 56% (0.44) in Indonesian or Japanese tweets.

Our results indicate that German users were distinctively more likely to include URLs and hashtags than users of other languages, especially when compared to Indonesian, Malay, and Japanese. In contrast, Indonesian and Malay users were more likely to retweet and include mentions, especially when compared to German users. Thus, German users appeared to have a stronger propensity for content-related behaviors. Indonesian and Malay preferred social and message-broadcasting behaviors. This shows two distinctive ways of using Twitter for either information sharing or social networking (Kwak et al. 2010).

Discussion

Here we summarize the findings and draw implications. First, we applied an automatic language detection algorithm over 62 million tweets to identify the top 10 most popular languages in Twitter. We found that only half of the tweets were in English (51%). Other popular languages such as Japanese, Portuguese, Indonesian, and Spanish together accounted for 39% of the tweets. Our results mostly match the findings of prior studies (Honeycutt & Herring 2009, Semiocast 2010).

Second, we found that in their use of common Twitter conventions, users of different languages varied considerably. Some of the variations might be attributed to inherent cultural differences. Others might be due to how long Twitter had been used by a language community, how many people actively used it, whether users were spread out geographically, and how many bilingual brokers there were to spread conventions and practices from one language community to others, etc. In future work, we plan to conduct more research to understand the motivation and practices of these cross-language differences.

Third, we found that users of different languages used Twitter for different purposes. Some language users, like German, tended to include URLs and hashtags more often, while others, like Korean, tended to reply to each other more often. This suggests that some communities used Twitter more for information sharing, while others used it more for conversational purposes (Kwak et al. 2010). These language-specific inclinations should be considered when building cross-cultural tools. For example, a recommendation tool developed for German users may want to promote tweets including URLs. For Korean users, the tool may want to focus more on conversational tweets.

While there has been much work on characterizing general Twitter usage, we know of no large-scale in-depth studies that compare the behaviors of different language users in Twitter. Since Internet usage is a global phenomenon, studies of how users perceive and behave on social websites will become increasingly important. In this paper we have used the phrase "language community" somewhat loosely, but it is clear that languages serve as barriers in information diffusion (Herring et al. 2007).

However, the study has limitations. First, our dataset only represents 3-4% of all public messages in Twitter, and we could not include private messages. Second, the users reported are skewed towards active users. Due to sampling, many users posting few tweets during the collection period were missing from our dataset. Nonetheless, the consistency with prior studies (Honeycutt & Herring 2009, Semiocast 2010) suggests that the language distribution reported should hold for larger datasets. Finally, although our language detection algorithm is imperfect, we believe that automatic techniques like ours, even with their inherent shortcomings, are complementary to manual coding methods. Both types of research methods are needed to study the explosive growth of social media.

Conclusion and Future Work

In prior research, the natural tendency has been to assume that the behaviors of English users generalize to other language users. However, since many communication tools are global, we need to examine critically whether indeed users of non-English languages behave in similar ways.

In this paper, we make two key contributions. First, we identified the most popular languages used in Twitter. Second, we profiled the top 10 language communities and showed that they differed considerably on using specific Twitter conventions. Our findings can help designers of cross-cultural communication tools to take into account the differences between languages. Moreover, we illustrated a method for large-scale study of social media sites (e.g., Twitter, Facebook).

In our future work, we aim at validating the differences observed in the 2010 dataset with a new large sample of tweets. We are also measuring indicators of how information flows across language communities: e.g., number of bilingual brokers, amount of URLs and hashtags that such brokers shared across language communities.

References

Agresti, A. 2002. *Categorical Data Analysis*. J.Wiley & Sons, NJ.

boyd, d., Golder, S., and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *Proc. HICSS'10*, 1-10.

Hecht, B. and Gergle, D. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. *Proc. CHI'10*, 291-300.

Herring, S., Paolillo, J., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L., and Clark, B. 2007. Language Networks on LiveJournal. *Proc. HICSS'07*.

Honeycutt, C. and Herring, S. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. *HICSS'09*, 1-10.

Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. *Proc. WebKDD/SNA-KDD'07*, 56-65.

Kayan, S., Fussell, S., and Setlock, L. 2006. Cultural Differences in the Use of Instant Messaging in Asia and North America. *Proc. CSCW'06*, 525-528.

Krishnamurthy, B., Gill, P., and Arlitt, M. 2008. A Few Chirps About Twitter. *Proc. WOSN'08*.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a Social Network or a News Media? *Proc. WWW'10*, 591-600.

Semiocast. 2010. Half of Messages on Twitter are not in English, Japanese is the Second Most Used Language. http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf.