Twitter Users #CodeSwitch Hashtags! #MoltoImportante #wow #헐

David Jurgens, Stefan Dimitrov, Derek Ruths
School of Computer Science
McGill University
Montreal, Canada
jurgens@cs.mcgill.ca, stefan.dimitrov@mail.mcgill.ca,
druths@networkdynamics.org

Abstract

When code switching, individuals incorporate elements of multiple languages into the same utterance. While code switching has been studied extensively in formal and spoken contexts, its behavior and prevalence remains unexamined in many newer forms of electronic communication. The present study examines code switching in Twitter, focusing on instances where an author writes a post in one language and then includes a hashtag in a second language. In the first experiment, we perform a large scale analysis on the languages used in millions of posts to show that authors readily incorporate hashtags from other languages, and in a manual analysis of a subset the hashtags, reveal prolific code switching, with code switching occurring for some hashtags in over twenty languages. In the second experiment, French and English posts from three bilingual cities are analyzed for their code switching frequency and its content.

1 Introduction

Online platforms enable individuals from a wide variety of linguistic backgrounds to communicate. When individuals share multiple languages in common, their communication will occasionally include linguistic elements from multiple languages (Nilep, 2006), a practice commonly referred to as code switching. Typically, during code switching, the text or speech in a language retains its syntactic and morphological constraints for that language, rather than having text from both languages conform to one of the language's grammatical rules. This requirement enables code switching to be separated from borrowing, where foreign words are integrated into a native language's lexicon and morphology (Gumperz, 1982; Poplack et al., 1988; Sankoff et al., 1990).

While work on code switching began with conversational analyses, recent work has examined the phenomena in electronic communication, finding similar evidence of code switching (Climent et al., 2003; Lee, 2007; Paolillo, 2011). However, these investigations into code switching have largely examined interpersonal communication or settings where the number of participants is limited. In contrast, social media platforms such as Twitter offer individuals the ability to write a text that is decoupled from direct conversation but may be read widely.

Twitter enables users to post messages with special markers known as hashtags, which can serve as a side channel to comment on the post itself (Davidov et al., 2010). As a result, multilingual authors have embraced using hashtags from languages other than the language of their post. Consider the following real examples:

- Eating an apple for lunch while everyone around me eats cheeseburgers and fries. #yoquiero
- Jetzt gibt's was vernünftiges zum essen! #salad #turkey #lunch #healthy #healthylifestyle #loveit
- Hasta mañana a todo mundo. Que tengan linda noche. #MarketerosNocturnos #MarketingDigital #BlackVirs #SocialMedia
- 1% มันสำคัญมากนะ เพราะมันอาจเปลี่ยนจากD+ เป็น C และ B+เป็นA เกรดเฉลี่ยคงดีกว่านี้อ่ะ #พลาด #เสียดาย #fail

Here, the first author posted in English with a Spanish hashtag reflecting the author's envious disposition. In the second, the author comments in German on sensible food, using multiple English hashtags to describe the meal and their attitude. In the third and fourth, the authors comment

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 51–61,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

on sleep and school, respectively, and then each use hashtags with similar meanings in both their native language and English.

Hashtags provide authors with a communication medium that also has broader social utility by embedding their post within global discussion of other posts using the same hashtag (Letierce et al., 2010) or by becoming a part of a virtual community (Gupta et al., 2010). These social motivations resemble those seen for why individuals may code switch, such as to assimilate into a group or make discussions easier (Urciuoli, 1995). Twitter and other hashtag-supporting platforms such as Instagram and Facebook offer a unique setting for code switching hashtags for two reasons: (1) potential readers are disconnected from the author, who may not know of their language fluency, and (2) text translation is built into the platform, which enables readers to translate a post into their native language. As such, authors may be motivated to include a hashtag of another language to increase their potential audience size or to appear as a member of a multilingual virtual community.

Despite the prevalence of non-English tweets, which are approaching 50% of the total volume (Liu et al., 2014), no study has examined the prevalence of hashtag code switching. We propose an initial study of hashtag code switching in Twitter focusing on three central questions: (1) for which language pairs do authors write in the first language and then incorporate a hashtag of the second language, (2) when tweets include a hashtag of a different language, which instances signal code switching behavior, and (3) the degree to which bilingual populations code switch hashtags. Here, we adopt a general definition of code switching as instances where an individual establishes a linguistic context in one language and then includes elements (such as words) from one or more other languages different from the first. Two experiments are performed to answer these questions. In the first, we test general methods to identify which languages adopt the same hashtags and whether those shared hashtags are examples of code switching. In the second, we focus on three bilingual cities to examine hashtag code switching behavior in French and English speakers.

Our study provides three main contributions. First, we demonstrate that hashtag code switching is widespread in Twitter. Second, we show that Twitter as a platform includes multiple phenomena that can be falsely interpreted as code switching and therefore must be accounted for in future analyses. Third, in a study of French and English tweets from three cities, we find that an increased rate of bilinguality decreases the frequency of including hashtags from another language but increases the overall rate of code switching when such hashtags are present. Furthermore, all data for the experiments is made publicly available.

2 Related Work

Research on code switching is long standing, with many theories proposed for the motivations behind code switching and how the two languages interact linguistically (Poplack and Sankoff, 1984; Myers-Scotton, 1997; Auer, 1998). Most related to the present work are those studies examining code switching in online communications.

Climent et al. (2003) examined the use of Spanish and Catalan in newsgroups, finding it occurs 2.2% and 4.4% of the Catalan and Spanish contexts, respectively. Lee (2007) analyzed a corpus of Cantonese and English emails and ICQ instant messages and surveyed Hong Kong users of each form of communication. She found that the users preferred mixed-language communication, with no user indicating that they communicated in only Cantonese. Furthermore, the shorter, more informal ICQ messages were more likely to be code switched (99.4%) than emails (41.3%).

Paolillo (2011) measured code switching amongs English, Hindi, and Punjabi in both IRC and Usenet forum posts, finding similar to Climent et al. (2003) that the shorter, more conversational IRC posts had higher rates of code switching. Paolillo (2011) also note that code switching rates differed between Hindi and Punjabi speakers.

The present work differs significantly from these three studies in two aspects. First, we assess code switching across all language communities on Twitter, rather than examining individual groups of bilingual speakers. Second, we focus our analysis only on the code switching of a post's hashtag due to its unique role in microtext (Gupta et al., 2010), which has yet to be examined in this context.

3 Hashtag Use in Twitter

Hashtags provide general functionality on Twitter and prior works have proposed that they serve

Name	Description	Examples
ANNOTATION	Serves as an annotation about the author’s feelings or comments on the content of a tweet.	#happy #fail #cute #joking #YoloSwaggins
COMMUNITY	A topical entity that links the tweet with an external community, which is commonly topical but also includes ”team-like” groups	#music #friends #BecauseItIsTheCup #TeamEdward
NAMED ENTITY	Refers to a specific entity that has a universally recognized name.	#Glee #TeenChoiceAwards #WorldCup2014
PLATFORM	Refer to some feature or behavior specific to the Twitter platform.	#followback #lasttweet #oomf
APPLICATION	Generated by a third-party application, which automatically includes its hashtag in the message.	#AndroidGames #NowPlaying #iPhone #Android
VOTING	Created as a result of certain real-world phenomena asking individuals to tweet with specific hashtags as a way of voting.	#MtvHottest #iHeartAwards
ADVERTISING	Promoting an item, good, or service, which can be sought out by interested parties.	#forsale #porn
SPAM	Used by adversarial parties to appear on trending lists and to make spam accounts appear real.	#NanaLoveLingga #681team #LORDJASONJEROME

Table 1: A taxonomy of hashtag according to their intended use.

a dual role as (1) bookmarking content with the tag’s particular expression and (2) functioning as a method for ad hoc community formation and discussion around a tag’s topic (Gupta et al., 2010; Davidov et al., 2010; Yang et al., 2012). However, the diverse user base of the Twitter platform has given rise to additional roles for hashtags beyond these two. For example, many popular hashtags focus on promoting users to follow each other,[1] such as #followback and #openfollow. Similarly, contests are run on Twitter, which have individuals vote by posting using a specific hashtag, e.g., #MtvHottest.

Given hashtags’ flexible roles, some may be used in multiple languages without being examples of code switching, such as the contest-based or follower-promotion hashtags noted above. Therefore, we first propose a taxonomy for classifying all types of hashtags according to their primary observed use in order to disentangle potential code switching behavior from Twitter-specific behavior. To construct the taxonomy, two annotators independently reviewed several thousand hashtags of different frequency to assess the differences in how the tag was used in practice. Each annotator then proposed their own taxonomy. The final taxonomy was produced from a discussion of differences, with both annotators initially proposing highly similar taxonomies.[2]

Table 1 shows the proposed taxonomy, containing eight broad types of hashtags. The first two types of hashtags correspond to the main hashtag roles proposed in Yang et al. (2012). The NAMED ENTITY tags also serve as method for individuals to link their content with a specific audience like the COMMUNITY type; however, NAMED ENTITY tags were treated as a separate group for the purposes of this study because the entities typically have a common name which is used in all languages and therefore would not be translated; in contrast, COMMUNITY hashtags refer to more general topics such as #soccer, which may be translated, e.g., #futbol. Hashtags of the five remaining types would likely not be observed in instances of code switching, with such hashtags often being used for purposes other than interpersonal communication.

4 Experiment 1: Popular Hashtags

Persistently popular hashtags reflect established norms of communication on Twitter. We hypothesize that these hashtags may be adopted by the speakers of multiple languages for joining a global discussion. Therefore, the first experiment examines the most-used hashtags over a five month period to measure two aspects: (1) which languages adopt the hashtags of other languages and (2) which hashtags used in multiple languages are evidence of code switching.

[1] In Twitter, following denotes creating a directional social relation from one account to another.

[2] We note that a small number of hashtags did not fit this taxonomy due to their idiosyncratic use. These hashtags were typically single-letter hashtags used when spelling out words, e.g., “tonight is going to be #f #u #n,” or when the author has mistakenly used punctuation, which is not included in Twitter’s definition of a hashtag, e.g., “#I’mAwesome,” which has the hashtag #I rather than the full expression.

4.1 Experimental Setup

Data Hashtag frequencies were calculated from 981M tweets spanning March 2014 to July 2014. Frequencies were calculated over this five month period in order to focus on widely-used hashtags, rather than bursty hashtags that are popular only for a short time, such as those studied in Huang et al. (2010) and Lin et al. (2013). For each hashtag, up to 10K non-retweet posts containing that hashtag were retained, randomly sampling from the time period studied when more than 10K were observed. To enable a more reliable estimate of the language distribution, we restrict our analysis to only those hashtags with more than 1000 posts, for a total number of 19.4M posts for 4624 hashtags, with an average of 4204 posts per hashtag.

Language Identification The languages of tweets were identified using a two-step procedure. First, message content was filtered to remove content such as usernames, URLs, emoji, and hashtags. Tweets with fewer than three remaining tokens were excluded (e.g., a message with only hashtags). Second, the remaining content was processed using `langid.py` (Lui and Baldwin, 2012), a state of the art language identification program that supports the diversity of languages found on Twitter.

Determining the language of a hashtag in a general setting for all languages is difficult due to the presence of acronyms, abbreviations, and slang. Therefore, we adopt a heuristic where a hashtag's language is set as the language used by the majority of its tweets. To quantify the accuracy of this heuristic, two annotators inspected the tweets of 200 hashtags to identify the language of the hashtag and for the majority of the tweets. This analysis showed that the heuristic correctly identifies the hashtag's language in 96.5% of the instances.

4.2 Hashtag Sharing by Languages

The adoption of a hashtag by a second language was measured by calculating the frequency with which tweets using a hashtag with language l_1 were labeled with language l_2. The noisy nature of microtext is known to make language identification difficult (Bergsma et al., 2012; Goldszmidt et al., 2013) and can create spurious instances of second-language hashtag adoption. Therefore, we impose a minimum frequency of hashtag use where l_2 is only said to use a hashtag of l_1 if at least 20 tweets using that hashtag were labeled with l_2. To quantify the accuracy of our hashtag adoption measure, two annotators inspected the second-language tweets of 200 hashtags, sampled from the data and representing 40 language pair combinations; this analysis showed that with the filtering the assertion that at least one author from language l_1 used a hashtag of language l_2 was correct in 67% of the instances.

Hashtag	# Langs.	Primary Lang.	Type
#lastfm	39	en	APPLICATION
#WaliSupitKEPO	32	id	SPAM
#RenggiTampan-DanKece	32	id	SPAM
#NP	32	en	APPLICATION
#Np	32	en	APPLICATION
#MTVHottest	31	en	VOTING
#SidikLoveTini	30	id	SPAM
#np	30	en	APPLICATION
#GER	29	en	NAMED ENTITY
#User_Indonesia	29	id	APPLICATION
#Soccer	29	en	COMMUNITY
#RobotKepo	29	id	APPLICATION
#KeePO	27	id	APPLICATION
#NowPlaying	28	en	APPLICATION
#Hot	28	en	ADVERTISEMENT

Table 3: The hashtags associated with the most number of languages having at least 20 tweets using that hashtag

Table 2 shows the frequency with which authors using the 15 most-commonly observed languages (shown as columns using their ISO 639-1 language codes) adopt a hashtag from another of the most-common languages (shown as rows), revealing widespread sharing of hashtags between languages. English hashtags are the most frequently used in other languages, likely due to it being the most common language in Twitter. However, other languages' hashtags are also adopted, with Spanish, Japanese, and Indonesian being the most common after English.

Despite the strong evidence of using of a single hashtag in multiple languages, the results in Table 2 should not be interpreted as evidence of code switching. Table 3 shows the 15 hashtags used in the most number of languages. The majority of these hashtags are generated by either (1) Twitter-based applications that automatically write a tweet in a user's native language and then append a fixed English-language hashtag or (2) spam-like accounts that use the same hashtag and include random text snippets in various languages, neither of which signal code switching behavior.

Furthermore, given the noise introduced by language misidentification and spam behavior on the

	Language of tweet														
	de	ru	ko	pt	en	it	fr	zh	es	ar	th	ja	id	nl	tr
de		2	1	4	15	6	9	4	9			1	4	6	1
ru	3		3	3	25	7	5	8	7	2	1	7	7	7	1
ko	4	2			13	3	6	5	10	3	10	11	4	2	
pt	14	3			64	45	40	13	63		2	4	3	15	10
en	1705	532	155	1235		1735	2183	1171	2482	362	176	742	1097	1101	342
it	5	2	1	10	29		15	4	22	5		3	6	3	1
fr	38	2	3	36	87	49		28	67	8	1	12	19	29	6
zh	3	4	2	2	12	1	2		4		1	11	1		1
es	67	17	3	321	435	264	206	105		29	5	32	66	66	31
ar	6	2			38	4	9	6	7			8	5	1	2
th	3		7	1	24	5	4	8	8	2		6	4	1	
ja	17	18	11	11	123	17	24	132	45	2	2		14	12	4
id	84	2	6	25	131	88	58	14	92	6	5	11		52	17
nl	13		1	3	17	6	11	2	9			1	1		
tr	17	1		3	28	9	7	7	13	3		1	22	9	

Table 2: The frequency with which a hashtag is used by multiple languages. Columns denote the language in which the tweet is written; rows denote the hashtag's language; and cell values report the number of hashtags where the column's language has used the hashtag in at least 20 tweets. Diagonal same-language values are omitted for clarity.

Twitter platform, we view the initial results in Table 2 an overestimate of hashtag adoption by languages other than the hashtag's source language. A further inspection of language classification errors revealed four common factors: (1) the lack of accents on characters,[3] (2) the use of short words, which appeared ambiguous to `langid.py`, (3) the use of non-Latin characters for emoticons or visual affect, and (4) proper names originating from a language different from the tweet's. Nevertheless, the observed trends do provide some guidance as to which language pairs might share hashtags and also may code switch.

Among the hashtags in Table 3, two are legitimately used by authors in multiple languages: #soccer and #GER, the latter corresponding to the German soccer team. Both hashtags were popular due to the World Cup, which occurred during the time period studied. For both, authors included these hashtags while taking part in a global conversation about the games and event. The hashtag #soccer is a clear case of code switching, where individuals are communicating their interests in multiple languages, even when equivalent hashtags in the tweet's language are actively being used. Indeed, over half of the languages using #football had at least one tweet containing both #football and #futbol. The example of #GER highlights a boundary case of code switching. Here, GER is an abbreviation for the country's name, making it a highly-recognized marker, rather than an example of a language change that results in code switching; however, the country has different names depending on the language used (e.g., Deutschland), which does point to an active choice on an author's part when selecting a particular name and its abbreviation.

4.3 Analysis by Hashtag Type

In a second analysis, we focus specifically on hashtags classified as COMMUNITY and ANNOTATION, which are more associated with intentional communication actions and therefore more likely to be used in instances of code switching. Performing such an analysis at scale would require automated methods for classifying hashtags by their use, which is beyond the scope of this initial investigation. Therefore, we performed a manual analysis of the 100 most-common, 100 least-common, and 100 median-frequency hashtags in our dataset to assess the distribution of hashtag types and cases of code switching among the COMMUNITY and ANNOTATION hashtags. Two annotators labeled each hashtag, achieving 64.6% agreement on the type annotations; disagreements were largely due to mistaken assignments rather that disputed classifications.[4] An adjudication step resolved all disagreements. Additionally, eleven hashtags were excluded from analysis due being made of common words (e.g., #go, #be) which had

[3] In particular, the lack of character accents caused significant difficulties in distinguish between Spanish and Catalan.

[4] In particular, mistakes were more common when analyzing hashtags used in languages outside the annotators' fluency, which required a more careful assessment of why the hashtag was being used.

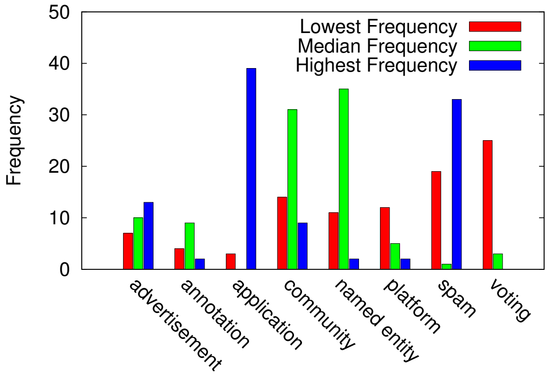

Figure 1: Type distributions of the sets of 100 highest, median, and lowest frequency hashtags used in our dataset

no meaningful interpretation for their use. Following, we describe the results of the analysis and then highlight several types of hashtags.

Figure 1 shows the distribution of hashtag types observed in the three samples. SPAM and APPLICATION hashtags were most common among highest frequency hashtags, whereas the lowest frequency tags in the dataset were also either SPAM or VOTING. Surprisingly, the median frequency hashtags had the majority of the discussion-related hashtag types

Within the ANNOTATION and COMMUNITY types, we selected thirteen hashtags each to manually evaluate if code switching behavior was observed. For each hashtag, two annotators reviewed all associated tweets that were identified as using a different language than that of the hashtag. Annotators were instructed to consider the tweet an instance of code switching only in cases where (1) there was sufficient text to determine the message's actual language and (2) the message was an act of communication (in contrast to spam-like or nonsensical messages).

Code switching behavior was observed for eleven of the ANNOTATION hashtags and twelve of the COMMUNITY hashtags. Table 4 shows those code switched hashtags and the languages in which they were seen, highlighting the varying frequency with which hashtags were used in multiple languages. For example, the primarily Arabic hashtag #Hadith was used in English and Dutch tweets; similarity, all three Spanish hashtags were used in English tweets.

Many hashtags are used primarily with languages that are associated with countries known to have bilingual speakers fluent in English. However, several hashtags were used in a variety of diverse languages. For example, #truth was used with languages such as Arabic, Bosnian, Bulgarian Hindi, and Punjabi. The most widely code switched hashtag was #magic. In English, the hashtag is commonly used with content on magic tricks; however, in other languages, the hashtag often connotes surprise. For example, the Latvian tweet "Es izmeklēju visu plauktu, nekur nav. Mamma piejiet ne sekunde nepagāja, kad viņa atrada. #magic" comments on having an item on the shelf disappear when looking for it, only for it to reappear like magic.

Hashtag	Lang.	Lang. of Code Switched Tweet
#Noticias	es	en
#Facts	en	id th fr es ru
#simple	en	id es fr ms tr tl sw zh ja ko
#bitch	en	ar cs de es fr id it ja ms nl pt ru sv tl tr zh
#delicious	en	ca de es fr id it ja ko ms nl ru th tr zh
#Design	en	ar de es fr ja kr pt th tl zh
#Felicidad	es	ca en
#SWAG	en	de es fr id it pl pt ru
#fresh	en	es fr id it ms nl sv
#BoludecesNO	es	en
#truth	en	ar bs bu es fr hi id ja it ms pa pt ru tl zh
#Hadith	ar	nl en
#Quran	ar	fa ms id sw az it de en
#hadith	ar	fr en
#tech	en	de es nl ar el fr ro id it ja ms no pl pt ru sq sv zh
#RemajaIndonesia	jv	ms
#class	en	ar tr es bg de fr pt he hr id it ja lt lv ms nl ru sw tl uk zh
#animals	en	ar ca de es fr pt it ms ja mk pl pt ro ru tl tr ur vi
#cine	es	ca de en fr ja pt ro ru
#sunday	en	es ar tr fr ca de el gl hu id it ja ms ko pt nl nn no pl ro ru sl sv th tl zh
#Energy	en	ru es de fr it pt tr
#change	en	ar nl es cs de el eu fr pt id it ja ko jv lv ms nb no pl ro uk ru sv ta th tl tr ur zh
#magic	en	nl fr ar ru ca cs de el it es hu id ja jv ko lv ru ms nn pl pt ro sq sv sw sl tl tr zh

Table 4: Code switched hashtags and the languages of the tweets in which they were seen (ANNOTATION types top, COMMUNITY types bottom).

During annotation, we observed that authors were highly productive in their code switching, using these hashtags to generate the types of emotional and sarcastic messages typically seen in same-language messages. For example, in the Swedish tweet "Bussen luktar spya och öl. #fresh" the author is sarcastically commenting on a bus that smells of vomit and beer.

4.4 Discussion

The process of annotating code switching for hashtags revealed four notable trends in author behavior that occurred with multiple hashtags. First, authors fluent in non-Latin writing systems will often use Latin-transliterated hashtags, which are then adopted by authors of Latin-based systems. For example, the hashtag #aikatsu describes a collectible card game and anime and is heavily used by both Japanese and English authors. Similarity, the transliterated hashtags #Hadith and #Quran are commonly associated with Arabic-language tweets, which rarely include an Arabic-script version of those hashtags even when the tweets include other hashtags in Arabic.

Second, when two or more languages share the same written form of a word (i.e., homographs), the resulting hashtags become conflated and appear as false examples of code switching. For example, #Real was widely used in both English and Spanish, but with two meanings: the English usages denoting something existent (i.e., not fake) and the Spanish usages referring to Real Madrid FC, a soccer club. The hashtag #cine also posed a challenge due to abbreviation. While many Spanish-language tweets include #cine (cinema), tweets in other languages include #cinema and its abbreviated form #cine, which matches the Spanish term, creating false evidence of code switching.

Third, multilingual individuals may adopt a common hashtag for reasons other than code switching, which we highlight with two examples. The hashtag #1DWelcomeToBrazil is used in a large number of English and Portuguese tweets. This hashtag is associated with the travel arrival of the English-speaking band One Direction to Brazil. Similarly, the #100happydays hashtag was spawned from a movement where individuals describe positive aspects of their day. These global phenomena increases the difficulty of automatically identifying code switching instances.

Fourth, spam accounts will occasionally latch onto a hashtag and use it in a variety of languages. For example, the popular hashtag #1000ADAY is used to attract new followers, which resulted in adult content services also using the hashtag to post spam advertisements. Surprisingly, nearly a third of tweets for this hashtag are in Russian and feature fully-grammatical text that appears to be randomly sampled from other sources, such as lists of proverbs. After examining multiple accounts, we speculate that these messages are actually bot accounts who need to generate sufficient number of messages to avoid Twitter's spam filters. Work on detecting fake accounts has largely been done in English (Benevenuto et al., 2010; Grier et al., 2010; Ghosh et al., 2012) and so may benefit from detecting this cross-lingual hashtag use in accounts.

5 Experiment 2: Bilingual Cities

The second experiment measures the prevalence of hashtag code switching in tweets from three cities with different populations of English and French speakers: Montreal, Canada, Quebec City, Canada and Paris, France. All three cities are known to contain bilingual speaker as well, who have been shown to actively code switch (Heller, 1992). To test for differences in the code switching behavior of populations, each city is analyzed according to the degree to which Anglophone and Francophone speakers incorporate hashtags of other languages into their tweets and whether translations of the code switched hashtags are used in the original language.

5.1 Experimental Setup

Data Tweets were gathered for each city by using the method of Jurgens (2013) to identify Twitter users with a home location within each city's greater metropolitan area. Tweets were then extracted for these users over a three year sample of 10% of Twitter. This process yielded 4.4M tweets for Montreal, 203K for Quebec City, and 58.1M for Paris. For efficiency, we restricted the Paris dataset to 5M tweets, randomly sampled across the time period.

Language Identification The language of a tweet was identified using a similar process as in Experiment 1. Because this setting restricts the analysis to only English and French, a different method was used to determine the language of a hashtag. Given a tweet in language l_1, the text of a hashtag is tested to see if it wholly occurs within the dictionary for l_1; if not, a greedy tokenization algorithm is run to attempt to split a hashtag into constituent words that are in the dictionary of l_1. If either the dictionary-lookup and tokenization steps

French hashtags on English tweets			English hashtags on French tweets		
Quebec City	Montreal	Paris	Quebec City	Montreal	Paris
imfc	imfc	comprendraquipourra	lasttweet	gohabsgo	bbl
rilive	charte	sachezle	bbl	fail	teaminsomniaque
relev	seriea	nian	mtvhottest	ind	teamportugal
ceta	bel	hollande	gohabsgo	mtvhottest	ps
preorderproblemonitunes	brasil2014	federer	not	not	findugame
derpatrash	touspourgagner	tropa	fail	soccer	adp
villequebec	2ne1	guillaumeradio	100factsaboutme	wow	lasttweet
tufnations	ma	vousetespaspret	herbyvip	podcast	follow
ta	lavoixtva	bel	foodies	ukraine	teamom
rougeetor	passionforezria	retouraupensionnat	electionsqc2014	int	thebest

Table 5: The ten most frequent hashtags occurring in French and English tweets

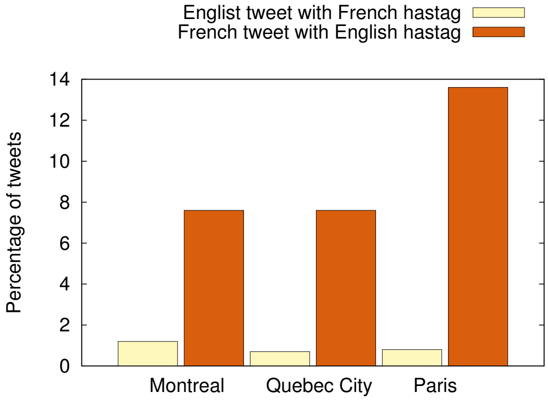

Figure 2: Percentages of tweets with any hashtag that include a hashtag from the other language

succeed, the hashtag is said to be in l_1. Otherwise, the tests are repeated with the second language l_2. If the hashtag cannot be recognized in l_1 or l_2, it is assumed to be in the language of its tweet. The `aspell` dictionaries were used to recognize words. Furthermore, after analyzing the errors made due to missing words, dictionaries were augmented to include common social media terms in each language (e.g., "selfie"). A manual analysis of 100 hashtags each for French and English showed that this language assignment method was correct for 91% of the instances.

5.2 Results

Francophone authors were much more likely to use English hashtags than Anglophone authors were for French hashtags. For tweets in each locale and language, Figure 2 shows the percentage containing a hashtag in the other language relative to the total number in that city using a hashtag in either language. Notably, Paris has a higher rate of using English hashtags than both Canadian cities. We speculate that this difference is due to the high rate of bilingualism in Montreal and Quebec City; because authors are fully fluent in both languages, should Francophone authors need to express themselves with an English hashtag, they may write the entire tweet in English, rather than code switching. In contrast, Parisian authors are less likely to be fully fluent in English (though functional) and therefore express themselves primarily in French with English hashtags as desired. An analogous trend may be seen for French hashtags in the English tweets from Montreal, which has a higher population of primarily Anglophone speakers who might be less willing to communicate entirely in French but will still use French hashtags to connect their content with the dominant language used in the city.

For each language and city, Table 5 shows the ten most popular hashtags incorporated into tweets of the other language. Examining the most popular English tags in French tweets shows a clear distinction in the two populations; French Parisian tweets include more universal English hashtags or those generated by applications, which are not generally instances of code switching. In contrast, the Canadian cities include more ANNOTATION type hashtags, including the sarcasm-marking #not, which are more indicative of code switching behavior.

An established linguistic convention within a population can also motivate authors to prefer one language's expression over another (Myers-Scotton, 1997). To test whether a high-frequency concept was equally expressed in French and English or whether one language's expression was preferred, we created pairs of equivalent English and French hashtags expressing the same concept (e.g., #happy/#heureux) by translating the 50 most-popular English hashtags used in French tweets. Then, the tweets for each city were analyzed to identify which languages were used in expressing each concept as a hashtag. The results in Figure 3 reveal that for nearly half of the hash-

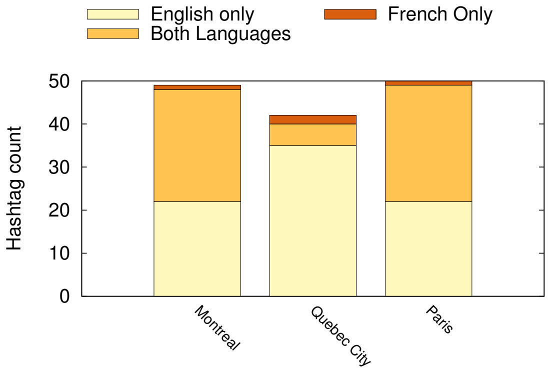

Figure 3: For 50 most-common concepts expressed in equivalent French and English translations, the frequency with which the hashtags for a concept were seen in each language.

tags, equivalent French language versions are in use; however, examining the relative frequencies shows that in all cases, the English version is still preferred, despite the presence of a large Francophone population. For hashtags that were only seen in English, many were of the COMMUNITY type, e.g., #50factsaboutme, which may not have an equivalent French-language version. However, we observed that when both an English hashtag and its French translation were attested, the use of the English hashtag in French was most often an instance of code switching. Hence, testing for the presence hashtag translation pairs may serve as a helpful heuristic for identifying hashtags whose use signals code switching behavior.

6 Discussion

Typically, code switching is distinguished from the related phenomena of borrowing by testing whether the word is being fluently mixed into the utterance instead of simply functioning as a loan word (Poplack, 2001). Hashtags present a unique challenge for distinguishing between the two phenomena due their brief content and unstructured usage: a hashtag may occur anywhere in a tweet and its general content lacks grammatical constraints. Examining the hashtags seen in our study, we find evidence spanning both types of uses. Common hashtags such as #win or #fail are widely recognized outside of English and their uses could easily be interpreted instances of borrowing. However, the complexity of other hashtags gives the appearance that their uses go beyond that of borrowing, e.g., #goingbacktoschool in “Nadie dijo que sería fácil, pero cómo cuesta estudiar después de 4 años de no tener nada académico cerquita #goingbacktoschool” where the author is commenting on the difficulty of returning for a degree. Still other posts include multiple single-token hashtags from a second language, e.g., the earlier example of “Jetzt gibt’s was vernünftiges zum essen! #salad #turkey #lunch #healthy #healthylifestyle #loveit.” Although individually these hashtags may be widely recognized and operate as interlingual markers, their combined presence suggests an intentional language shift on the part of the author that could be interpreted as code switching. Together, the examples point to hashtag use by multiple languages as a complex phenomena where shared hashtag entities exist on a graded scale from simple borrowing to fully signaling code switching. Our study is intended as a starting point for analyzing this practice and all our data is made available to support future discussions on the roles these hashtags play and how they facilitate communication both within and across language communities.

7 Conclusion

The present work has provided an initial study of code switching in Twitter focusing on instances where an author produces a message in one language and then includes a hashtag from a second language. Our work provides three main contributions. First, using state-of-the-art language identification techniques, we show that hashtags are widely shared across languages, though the challenges of correctly classifying the language of tweets limits our ability to quantify the exact scale. Second, in a manual analysis of ANNOTATION and COMMUNITY hashtags, we show that authors readily code switch with these types of hashtags, using them just as they would in single language tweets (e.g., indicating sarcasm). Third, in a case study of French and English tweets from three Francophone cities with bilingual speakers, we find that the cities with more bilingual speakers tended to have fewer occurrences English hashtags in French tweets, which we speculate is due to authors being more likely to write such tweets entirely in English, rather than code switch; however, when English hashtags were observed in French tweets from these more bilingual cities, they were much more likely to be used in instances of code switching. Data for all of the experiments is available at `http://www.networkdynamics.org/datasets/`.

Our work raises several avenues for future work. First, we plan to examine how to improve language identification in microtext in order to gain a more accurate estimation of hashtag sharing and code switching rates for languages. Second, the Twitter platform enables measuring additional factors that may influence an individual's rate of code switching; specifically, we plan to investigate (1) a user's historical tweets to estimate the degree of bilinguality and (2) the impact of a user's social network with respect to homophily and language use.

References

Peter Auer. 1998. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74. Association for Computational Linguistics.

S. Climent, J. Moré, A. Oliver, M. Salvatierra, I. Sànchez, M. Taulé, and L. Vallmanya. 2003. Bilingual newsgroups in catalonia: A challenge for machine translation. *Journal of Computer-Mediated Communication*, 9(1).

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 107–116. Association for Computational Linguistics.

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. 2012. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 61–70. ACM.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. 2013. Boot-strapping language identifiers for short colloquial postings. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*. Springer Verlag, September.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security (CCS)*, pages 27–37. ACM.

John Joseph Gumperz. 1982. *Discourse strategies*. Cambridge University Press.

Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, 12(1):58–72.

Monica Heller. 1992. The politics of codeswitching and language choice. *Journal of Multilingual & Multicultural Development*, 13(1-2):123–142.

Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Carmen K. M. Lee. 2007. Linguistic features of email and icq instant messaging in hong kong. In Brenda Danet and Susan C. Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press.

Julie Letierce, Alexandre Passant, John Breslin, and Stefan Decker. 2010. Understanding how twitter is used to spread scientific messages. In *WebSci10: Extending the Frontiers of Society On-Line*.

Yu-Ru Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. 2013. # bigbirds never die: Understanding social dynamics of emergent hashtags. In *Seventh International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. 2014. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Carol Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Chad Nilep. 2006. Code switching in sociocultural linguistics. *Colorado Research in Linguistics*, 19(1):1–22.

John C. Paolillo. 2011. Conversational codeswitching on usenet and internet relay chat. *Language@Internet*, 8.

Shana Poplack and David Sankoff. 1984. Borrowing: the synchrony of integration. *Linguistics*, 22(1):99–136.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Shana Poplack. 2001. Code-switching (linguistic). In *International Encyclopedia of the Social and Behavioral Sciences*, pages 2062–2065. Elsevier Science Ltd., 2nd edition.

David Sankoff, Shana Poplack, and Swathi Vanniarajan. 1990. The case of the nonce loan in tamil. *Language variation and change*, 2(01):71–101.

Bonnie Urciuoli. 1995. Language and borders. *Annual Review of Anthropology*, 24:pp. 525–546.

Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what@ you# tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 261–270. ACM.