University of Oxford

Global connectivity, information diffusion, and the role of multilingual users in user-generated content platforms

Scott A. Hale,[‡] St. Hilda's College

Thesis submitted in partial fulfillment of the requirements for the degree of DPhil in Information, Communication, and the Social Sciences in the Oxford Internet Institute at the University of Oxford

Michaelmas Term 2014/15 67,092 words

[‡]Please visit `http://www.scotthale.net/` for contact information.

To my wife, Yuko, for her love and support

Acknowledgements

I would like to thank my supervisors, Sandra González-Bailón and Eric T. Meyer, for sharing their insights, encouragement, and support during my DPhil candidacy. This thesis has been substantially improved through their patient engagement with my work and the generous giving of their time.

I owe a debt of gratitude to my examiners, Jonathan Bright and Mike Thelwall. Their work has been an inspiration to me and their thoughtful critiques have enhanced this thesis. I thank Taha Yasseri as well for his ideas and suggestions to refine and improve my work.

Working as a research assistant with Helen Margetts while completing my DPhil has been a wonderful experience, and I have tried to apply the many of the lessons I learned from Helen in this research.

I am thankful to my parents, Brian and Laura, for always believing in me and supporting me. I am also very thankful for their proofreading of my thesis.

Finally, I am grateful to the students and staff of the Oxford Internet Institute, the members of the Networks Journal Club seminar group in Mathematics, and my colleagues and anonymous reviewers at Human Factors in Computing (CHI) and Web Science, where many of the outputs of this thesis were published. I am thankful to you too, dear reader, for engaging with my work.

Global connectivity, information diffusion, and the role of multilingual users in user-generated content platforms

Scott A. Hale, St Hilda's College Michaelmas Term 2014/15

Abstract

Internet content and Internet users are becoming more linguistically diverse as more people speaking different languages come online and produce content on user-generated content platforms. Several platforms have emerged as truly global platforms with users speaking many different languages and coming from around the world. It is now possible to study human behavior on these platforms using the digital trace data the platforms make available about the content people are authoring.

Network literature suggests that people cluster together by language, but also that there is a small average path length between any two people on most Internet platforms (including two speakers of different languages). If so, multilingual users may play critical roles as bridges or brokers on these platforms by connecting clusters of monolingual users together across languages. The large differences in the content available in different languages online underscores the importance of such roles.

This thesis studies the roles of multilingual users and platform design on two large, user-generated content platforms: Wikipedia and Twitter. It finds that language has a strong role structuring each platform, that multilingual users do act as linguistic bridges subject to certain limitations, that the size of a language correlates with the roles its speakers play in cross-language connections, and that there is a correlation between activity and multilingualism. In contrast to the general understanding in linguistics of high levels of multilingualism offline, this thesis finds relatively low levels of multilingualism on Twitter (11%) and Wikipedia (15%).

The findings have implications for both platform design and social network theory. The findings suggest design strategies to increase multilingualism online through the identification and promotion of multilingual starter tasks, the discovery of related other-language information, and the promotion of user choice in linguistic filtering. While weak-ties have received much attention in the social networks literature, cross-language ties are often not distinguished from same-language weak ties. This thesis finds that cross-language ties are similar to same-language weak ties in that both connect distant parts of the network, have limited bandwidth, and yet transfer a non-trivial amount of information when considered in aggregate. At the same time, cross-language ties are distinct from same-language weak ties for the purposes of information diffusion. In general cross-language ties are smaller in number than same-language ties, but each cross-language tie may convey more diverse information given the large differences in the content available in different languages and the relative ease with which a multilingual speaker may access content in multiple languages compared to a monolingual speaker.

Contents

Chapter 1

Introduction

> *It is hardly possible to overrate the value...of placing human beings in contact with persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar.... Such communication has always been, and is peculiarly in the present age, one of the primary sources of progress* —John Stuart Mill

1.1 Information is diverse across languages

Language is a powerful, albeit often under-acknowledged, force influencing the information available to people. Consider, for example, a tourist wanting to locate a restaurant in London. If the tourist performs a quick search on Google, the language she uses will have a large effect on the options seen for dinner. A quick search for "London restaurant" in different languages reveals very different results in French, English, Chinese, Afrikaans, Korean, Russian, Japanese, Italian, and Spanish. The two languages with the most overlap in results, Russian and English, share about half of the top 100 results in common. In contrast, the queries for Chinese, Korean, and Japanese have no results in common with each other or any of the other languages searched (Table 1.1). A restaurant recommendation from an acquaintance on social media is also likely shaped by language as Takhteyev, Gruzd, and Wellman (2012) show language to be a large factor structuring follower–following relationships on Twitter.

Web search and information from social media sites are among the predominant

		fr	en	zh	af	ko	ro	ja	it	es
French	(fr)	100								
English	(en)	34	100							
Chinese	(zh)	0	0	100						
Afrikaans	(af)	35	32	0	100					
Korean	(ko)	0	0	0	0	100				
Russian	(ro)	44	53	0	38	0	100			
Japanese	(ja)	0	0	0	0	0	0	100		
Italian	(it)	7	5	0	4	0	7	0	100	
Spanish	(es)	27	15	0	18	0	18	0	6	100

Table 1.1. Each cell shows the number of URLs in common in the results when searching for "London restaurant" in different languages. Searches performed on 1 April 2012, and the first 100 results were recorded and analyzed. All queries were performed on google.co.uk with a UK IP address minutes apart from one another. Search query translations were derived from interlanguage links on Wikipedia.

ways users seek information online. While some attention has focused on the effects of personalization algorithms on the information available (e.g., Pariser, 2011), there has been little research looking at how the language a user employs affects the information available to that user online. Even though the raw number of webpages in English is increasing, growth in other languages has resulted in a steady year-on-year decrease in the percentage of webpages in English since at least 1998 (Pimienta, Prado, & Blanco, 2009). What research does exist shows a huge diversity in the information available in different languages (see Section 2.1).

The internationalization of content has been matched by an increase in the diversity of users. The percentage (but not the raw number) of web users speaking English is in steady decline as Internet use increases worldwide (Pimienta et al., 2009), and the number of Internet users in China now exceeds the total number of Internet users in North America or the European Union (Graham, Hale, & Stephens, 2012). This trend is likely to continue with the introduction of high-speed Internet in Africa (Juma & Moyer, 2008), the increasing use of mobile phones for Internet access throughout the world (Stork, Calandro, & Gillwald, 2012; Pew Research

Internet Project, 2014), and the continued rising standards of living in many less developed countries (e.g., Stork et al., 2012).

1.2 Design matters

Internet-based companies, particularly those relying on user-generated content, face many questions about the best way to adapt their platforms to accommodate users of different languages. These companies must decide whether separate, localized copies of their platforms (i.e., an approach more similar to Wikipedia) or integrated platforms accommodating multiple languages (i.e., an approach more similar to Twitter or Facebook) are appropriate. In addition, companies must decide what language tools (machine translation, allowance of human/crowd translations, dictionary tools, etc.) to incorporate into their platforms.

Design is important and can affect user behavior and by consequence the diffusion of information in user-generated content platforms (see Section 2.2). Despite the fact that "multilingualism...[is] the norm for most of the world's societies" (Birner, 2005), with over half of Europe and over a fifth of the US multilingual (Erard, 2012), many platforms are designed only with monolingual users in mind. The Google Play store is an illustrative example. Google Play allows users to find and install applications (apps) and media (books, music, video, magazines) on Android tablets and phones. The linguistic design of Google Play follows standard interface internationalization and localization practices with separate user interfaces for speakers of different languages (translated text, right-to-left design where appropriate, etc.). The particular point of interest here is how user-generated content (specifically user reviews of apps) is handled. More specialized apps are often reviewed in only one language in Google Play. However, reviews are grouped by the user-interface language settings of users' devices. As implemented, a user cannot see reviews of an app in another language without temporarily changing the user-interface language

Figure 1.1. Google Play store listing for an Okinawan language dictionary app. The Android OS user-interface language is set to US English on the left and to Japanese on the right. The red box added at the top indicates the summary of user reviews available in both languages. The second red box highlights the レビュー (“Reviews”) section, which gives the details of the 27 user reviews and is only available when the OS is set to Japanese.

of the entire Android operating system—a process that takes several minutes and affects all apps on the device. Confusingly, apps are rated with a number of stars (1-4) averaged across reviews from all languages. This can lead to the rather odd case shown in Figure 1.1 where the English interface shows an average rating of 3.6 stars from 27 user reviews, but then gives no indication of how the user could see these reviews. This can be particularly frustrating for multilingual users who could read the reviews in another language if the option were provided. It is further true that some users write a review in a language different from their user-interface language (e.g., a user with a Japanese UI language writing a review of an app in English). Such a user may think he is helping users in another language, but in fact other-language users are very unlikely to see the review as it remains accessible only to users with the same user-interface language selection as the author of the review.

As in the case of Google Play, issues in internationalization often arise in the handling of user-generated content, where user actions cannot fully be anticipated in advance. Google Play is built on the assumption that a user only ever writes and reads reviews in the same language as his or her user-interface setting—an assumption upon which this thesis casts great doubt.

1.3 Users have diverse linguistic abilities

One final illustrative story. An Australian user at the Stack Exchange travel question and answer forum posted a question asking if "there [is] any car museum or collection of cars that is open to the public in Okinawa" while he was visiting the island.[1] Another user answered the following:

> The Okinawa Classic Car Association (沖縄クラシックカー協会) has a small showroom atop a garage specializing in, you guessed it, classic cars. All visitors welcome. ガレージルマン (Garage Le Mans) 沖縄県宜野湾市大謝名2-3-3

[1] `http://travel.stackexchange.com/a/24874/1765`

The answer went on to provide further information about the exact location and linked to a blog by a third person describing that person's visit to the garage. The blog post had some English, but was mainly written in Japanese. The first all-English question has now been answered in English, but with further details provided in Japanese, including the address that the question-asker could now copy-and-paste into Google Maps easily.

The true significance of the interaction, however, is clear in the follow-up comments. The question-asker comments, "How on earth did you find this? It's perfect because I'm heading to Ginowan [(the location of the garage)] tomorrow anyway..." The user providing the answer responds that he simply "Googled 'classic car okinawa' in Japanese, and this was the second hit!"

Thus, the second user was able to leverage his ability to search in Japanese to find the answer to the question that eluded the first user. The story, however, doesn't stop there. As a result of this content being posted on the question and answer platform in English, a search for "classic car okinawa" in English on Google finds this question and answer thread and thereby enables any future monolingual English user to discover the garage.

There are a number of take-aways from this story, but the important one for this research is to recognize that the two users had different linguistic abilities and this diversity of skills allowed the second user to succeed where the first user could not. Diversity has been recognized as a key contributor to innovation (see Section 2.3). While most research has focused on cognitive diversity (differences in how individuals think and approach problems), linguistic ability may be of increasing importance in the modern information-based economy.

1.4 Structure

These three stories illustrate the three main starting points of this thesis: information can be diverse across different languages, site design can affect how users behave and thereby affect how information diffuses, and differing linguistic abilities may be just as important as cognitive diversity for information discovery.

These three themes lead to two major research questions motivating this thesis. The first concerns how connected users are across languages on user-generated content platforms and to what extent multilingual users bridge language divides. The seconds asks what the benefits and challenges are for global, multilingual user-generated content platforms and how these differ across platforms with different design choices.

The rest of the thesis proceeds by first examining broadly related work on these themes using literature from Computer-Mediated Communication (CMC) and specifically the field of Human-Computer Interaction (HCI) as well as literature on diversity and social networks in Chapter 2. That chapter presents and justifies the overarching research questions about the role design and multilingual users play in the diffusion of information between speakers of different languages online. The chapter also details the selection of Twitter and Wikipedia as the two platforms to be studied and compares them from a technological design or HCI perspective.

The thesis examines the role of language in Twitter (Chapter 3) and Wikipedia (Chapter 4) at a global level. The chapter on Twitter analyzes the global connectivity of the Twitter retweet and user-mentions (@messages) network and the role of multilingual users. The network is heavily structured by language with most mentions and retweets directed to users writing in the same language. Users writing in multiple languages are more active, authoring more tweets than monolingual users writing in one language. Despite this, multilingual users are no more likely to be retweeted or mentioned. These multilingual users play a unique bridging role in the global connectivity of the network. The level of introversion from speakers

in each language collectively does not correlate straightforwardly with the size of the user base as predicted by previous research. Finally, the English language does collectively play more of a bridging role than other languages, but the role played collectively by multilingual users across different languages is the largest bridging force in the network.

In a similar manner to the chapter on Twitter, the chapter on Wikipedia examines the role multilingual users play on Wikipedia. Multilingual users may serve an important function in diffusing information across different language editions of the project, and prior work has suggested this could reduce the level of self-focus bias in each language edition of the encyclopedia (e.g., Hecht & Gergle, 2009). This chapter also finds multilingual users are much more active than their single-edition (monolingual) counterparts. Multilingual users are found in all language editions, but smaller-sized editions with fewer users have a higher percentage of multilingual users than larger-sized editions. About a quarter of multilingual users always edit the same articles in multiple languages, while just over 40% of multilingual users edit different articles in different languages. When non-English users do edit a second language edition, that edition is most frequently English. Nonetheless, several regional and linguistic cross-editing patterns are also present.

The third empirical chapter (Chapter 5) narrows the scope from a global view to a specific, linguistically interesting region: Okinawa, Japan. The selection of Okinawa is motivated by geolinguistic factors and the findings of the earlier chapters. Geographically closer to Taipei than Tokyo, the islands were once part of a prosperous independent kingdom built on trade in the region. The islands were separated from Japan after World War II and administered by the United States, which has maintained a strong presence since. Between military members, contractors, and dependents, Okinawa is home to a large population of English as well as Japanese speakers in a relatively small geographic area. The narrowed focus on Okinawa allows for a more in-depth, mixed-methods examination of the content shared across

languages and the users sharing that content.

Examining the content of edits in addition to the edit meta-data used in Chapter 4 on Wikipedia, the chapter on Okinawa finds that multilingual users editing both the English and the Japanese editions of Wikipedia are among the most active and dedicated users in their primary languages, where they make many large, high-quality edits. However, when these users edit in their non-primary languages, they tend to make edits of a different type that are overall smaller and more often restricted to the narrow set of articles that exist in both languages. From these findings it is argued that multilingual users are more likely to make incremental additions to existing content rather than contribute larger amounts of information in a second language. Design changes to motivate wider contributions from users in their non-primary languages and to encourage multilingual users to transfer more information across language divides are presented.

Finally, the thesis concludes in Chapter 6 by comparing the findings from all chapters in light of the literature. This final chapter also links the research to wider themes such as language loss and identifies opportunities for further research.

Chapter 2

Background and Motivation

If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart. —Nelson Mandela

2.1 Language bubbles

The causes of divisions online are many and varied—censorship/copyright restrictions (Goldsmith & Wu, 2006), personalization (Pariser, 2011), geography (Graham & Zook, 2011; Wilkinson & Thelwall, 2012). This thesis concentrates on the role of language in structuring user experience, and argues that the users of many user-generated content websites operate within *language bubbles* akin to the filter bubbles of personalization that were suggested and popularized by Pariser (2011). Many of the findings are likely applicable in the related areas of personalization and geography, and these are discussed at the end of the thesis in Chapter 6. In contrast, this thesis has relatively little to say about differences created through censorship and copyright restrictions. Rather, censorship and copyright restrictions introduce limitations on this research that are difficult to workaround. For example, even though there are more native Chinese users online than native English speakers (Graham et al., 2012), Chinese plays a limited role in most of the work presented in this thesis. Access restrictions on platforms like Twitter and Wikipedia in mainland China, the largest location of native Chinese Internet users, result in many fewer

Chinese-language users on these platforms. The access restrictions have also created the opportunity for Chinese "clone" platforms (e.g., Sina Weibo and Baidu Baike) to thrive, which separates Chinese users and content from users and content in most other languages. Copyright restrictions such as IP address filtering also present a limitation for the research of this thesis. It seems likely that at least some content with IP address restrictions might diffuse more widely if those restrictions were removed as one can hardly expect users in another location to further propagate content that they themselves cannot view due to copyright/IP address restrictions. Nonetheless, it remains extremely difficult to quantify the effects of these restrictions on the diffusion patterns occurring within user-generated content platforms due to the nature of emergent effects created through the interaction of a large number of users together (see Section 2.2.3).

The contents of this thesis are more applicable to personalization and geographic limitations on information availability. The dangers of personalization were popularized by Eli Pariser (2011) in his book *The Filter Bubble: What the Internet is hiding from you*. Personalization itself is not entirely bad—personalization can help users find the most personally relevant content and avoid information overload and decision fatigue (Churchill, 2013). The danger, however, is in hyper-specialization or over-personalization that can lead to "safe" recommendations within a narrow space that the user is most-likely to appreciate. This can reduce the chance of serendipitous information discovery (Zuckerman, 2013) and disrupt the ability for people of contrasting opinions to interact online (Liljeblad, 2012). Social network theory discussed below (Section 2.3) finds individuals most often cluster into similar groups even in the absence of personalization; however, Liljeblad (2012) argues personalization removes the a user's choice to pursue alternative viewpoints through its narrowed presentation of results.

In Pariser's conceptualization, personalization is largely an issue created by and to be addressed through how companies design information retrieval and ranking

algorithms. These algorithms use many latent factors to influence the results a user sees (location, past search activity, etc.) (Pariser, 2011). The majority of studies in personalized information retrieval, however, have been monolingual (Ghorab, Zhou, Steichen, & Wade, 2011). By examining user-behavior in relation to language, this thesis is able to make additional suggestions that extend those of Pariser (who advocates for greater transparency, choice, and awareness). Specific suggestions are made for search as well as content and friend recommendation algorithms within the empirical chapters of this thesis. The levels of multilingual use found in the empirical chapters—supported by survey findings (e.g., Eurobarometer, 2011; Steichen, Ghorab, O'Connor, Lawless, & Wade, 2014)—provide ample evidence that the monolingual user models at the heart of most personalization algorithms need to change to better serve multilingual users (Ghorab et al., 2011).

Geography is often a key element of personalization (Pariser, 2011), but even in the absence of personalization, distance is a limitation itself on the likelihood of content spreading. Tobler's First Law of Geography captures this, stating that "everything is related to everything else, but near things are more related than distant things" (Li, Sen, & Hecht, 2014). While geographic divides are very real (Wilkinson & Thelwall, 2012), non-language geographic differences are comparatively easier to overcome than language divisions. Content in another location, but in the same language, can still be found through traditional information retrieval (search) techniques and monolingual users can read the content directly. Zuckerman (2013) highlights that often a user will need additional contextual information to understand international information and that there are motivation and awareness issues in discovering international, same-language content. These challenges, however, are not reduced by considering cross-language content exchange. Rather, through considering the relatively harder case of designing for cross-language use online we may very likely uncover techniques and approaches that also aid in discovery and accessibility of international, same-language content. This thesis references past studies

on geography where appropriate, but makes its main contributions to the role of language in online communication.

Language is a difficult and often under-acknowledged limitation on access to information. Information varies greatly between languages (Hecht & Gergle, 2010b) as users often contribute local knowledge in local languages (Hecht & Gergle, 2009). A user can more easily seek out news in his or her primary language from another country and keyword searches in the user's primary language can match content regardless of its geographic origin. In contrast, however, finding other-language information often requires switching to a different language version of an information portal or conducting a separate search in that language. In addition, language divides (and bridging) can occur not only between countries but also within a single country. Even within a small geographic area, language divides can still be very strong as shown in the final empirical chapter on Okinawa (Chapter 5).

Machine translation is not a panacea to this challenge as even if machine translation were perfect, a user would still need to identify the content of interest to be translated. As it is, machine translation holds much promise, but is far from perfect (Wilks, 2009; Mikolov, Le, & Sutskever, 2013). The first machine translation systems were rule-based, and used linguistic and symbolic knowledge to translate text. In contrast, newer systems use statical methods inspired by work using statistics for speech recognition (Wilks, 2009). The future of machine translation likely lies in a hybrid approach uniting statistical and non-statical methods and data although finding the best way to unite the two approaches is an open research question (Wilks, 2009).

User-generated content is challenging for both statical and non-statistical machine translation systems. There are a large number of languages spoken in the world and reflected on social media platforms, and for many language pairs, linguistic rules and sufficiently-sized parallel text corpora are not available (Wilks, 2009; Carter, Tsagkias, & Weerkamp, 2011). Some user-generated content platforms like

Wikipedia contain relatively well structured prose, but many user-generated content platforms like Twitter consist of content written in an informal style, may combine multiple languages in a single post, and may use non-standard lexicography (e.g., writing Arabic with Latin characters). All of these aspects are very challenging to machine translation systems—and also to algorithms seeking to determine the language of a given text as discussed in Appendix A and Carter et al. (2011). The sheer volume of user-generated content and the speed at which it is produced (and then obsolete) also present challenges for machine translation.

While many researchers continue to pursue machine translation, this thesis takes a different approach by studying human behavior with regards to language on user-generated content platforms. Multilingual users may be able to signal what pieces of foreign-language content are most relevant to monolingual users on the platform. These monolingual users can then act (or not) to try to comprehend the content with machine translation, other language tools (e.g., dictionaries), and/or ask the multilingual user for more details.

The study of human behavior on user-generated content platforms can not only inform how multilingual user-generated content platforms are designed (e.g., where, when, and how should machine translation and cross-language information retrieval algorithms be used) but can also contribute to back to the development of machine translation systems themselves. Recent efforts by machine translation researchers have started to mine the text of user-generated content platforms (especially Wikipedia) to create larger corpora for traditionally under-represented languages by searching for similar sentences across language pairs (Adafre & De Rijke, 2006; Mohammadi & GhasemAghaee, 2010; Smith, Quirk, & Toutanova, 2010). The study of user behavior on user-generated content platforms presented in this thesis takes a step back to establish first how much cross-language work is even happening on user-generated content platforms and between what language pairs. These findings can be used in efforts to create machine translation training corpora by,

for example, understanding which language pairs users most commonly cross between and thus the language pairs that are more likely to have a higher number of translated sentences that could be mined.

Beyond design, language limits the information a user can find, access, and comprehend. This thesis argues that speakers of different languages with access to different information represent a form of diversity from which users of global, multilingual platforms may benefit in the same way individuals, organizations, and other collectives have benefited from diversity (see Section 2.3). Studying user behavior at scale in user-generated content sites through the trace data left by users (Section 2.2.3) offers the possibility of better understanding computer-mediated communications on user-generated sites. These insights may then be used used to understand content flows and user connections (Section 2.3.3) as well as to suggest design changes that may further influence user behavior (Section 2.2).

2.2 Designing for multilingual users

The platforms users interact on are not merely neutral conduits of communication. Rather, users both shape technology and are also shaped by it (Thurlow, Lengel, & Tomic, 2004; Hughes, 1994; Kling, 1996). Computer-mediated communication (CMC) is the broad, interdisciplinary field encompassing the study of communication via computers (Herring, 1996). Within CMC, this research has a specific interest in online communication as well as the closely related field of human factors and computing (HCI).

Communication has been a core use of the Internet, with email predating the World Wide Web (Thurlow et al., 2004). While early websites generally broadcast information in one-way manner, multiple trends have driven web and mobile platforms to a new renaissance of large-scale, many-to-many user interaction and content creation on user-generated content platforms. These new platforms have

been driven by an increase in access to the Internet and a lowering of the technical expertise required to post content online. Not only has the number of people online increased (Graham et al., 2012), but the duration of the time spent online and the contexts of the online interactions have changed as the Internet has become more ubiquitous with mobile phones, tablets, and other devices (Keyes & Hale, 2014). In addition, the diversity of users online has increased as a greater percentage of the world's population has come online. This is reflected in a corresponding drop in the percentage of online content in English (Pimienta et al., 2009). Simultaneously, new platforms including blogs, social networking sites, question and answer forums, and wikis have greatly lowered the barriers for users to author content online.

Prior to the appearance of user-generated content sites, web users were often viewed as passive consumers of content authored by larger companies and organizations. The increasing multilingual composition of Internet users drove many scholars to investigate the differences between users of different regions/languages and the benefits to site owners for translating content into different languages (internationalization/localization). Early work by Russo and Boor (1993) focused on the importance of not just translating text, but also considering issues for colors, images, flow (e.g., for right-to-left languages), and date, time, and currency formats. Empirical work analyzed website differences across regions and languages (e.g., Cyr & Trevor-Smith, 2004) and also differences in social media platforms (e.g., B. Fogg & Iizawa, 2008, compares Facebook and Mixi, an early Japanese social network). At the same time, many studies, often from industry, reported that consumers spent more time on and were more likely to buy from sites localized to their countries and languages (e.g., DePalma, Sargent, & Beninatto, 2006).

Another strand of work at the intersection between language and the Internet sought to understand how individuals used multiple languages online. For example, Warschauer, Said, and Zohry (2002) studied language choice in Egypt, finding many Egyptians wrote in English due to technological limitations (lack of proper

input methods or concern that the recipient would not have Arabic fonts available) as well as cultural reasons (local dialectics of Arabic were generally only spoken and not written). Durham (2003) studied the languages medical students used on a mailing list in Switzerland and found English emerged as a *lingua franca* on the list. Gandal (2006) studied Internet use among native English and native French speakers in Quebec, Canada, and found speakers of both languages accessed content in the other language. Of their time spent on English- and French-language websites, native French speakers spent 64% of their time on English-language websites and native English speakers spent 13% of their time on French-language websites (Gandal, 2006). Wei and Kolko (2005) using survey methods to study the languages of Internet users in Uzbekistan, found users were willing to use Russian or English at times despite reporting low confidence in their foreign language skills. Similar findings seem to apply across the European Union, where a EU Commission report states that "[a]lthough 9 in 10 Internet users in the EU said that, when given a choice of languages, they always visited a website in their own language, a slim majority (53%) would accept using an English version of a website if it was not available in their own language" (Eurobarometer, 2011). None of these studies, however, addresses language choice in user-generated content platforms, a gap that this thesis fills by examining the number of users contributing content in multiple languages on Twitter and Wikipedia. The thesis further assesses the relative number of users connecting different pairs of languages on these two platforms. One limitation, however, of this study is that users read and watch content in another language more frequently than they contribute content in another language (Eurobarometer, 2011). Data on content viewing, however, is not generally available outside of the company operating a given platform.

Aside from language choice, a small number of studies have examined cross-language connections with hyperlinks. These studies have been focused mostly on blogs and suggests different languages have different interlinking patterns. In an

examination of Japanese, Spanish, and English blog posts about the 2010 Haitian earthquake as a pilot study for this thesis, I found limited cross-language hyperlinks, and the cross-language hyperlinks that were found were centered on English. Most cross-language hyperlinks indicated a reference relationship, rather than direct translation. Interestingly, 13.8% of posts linked across two languages included a common video or image, and the single largest destination for cross-language links in that study was to a photo blog by the Denver Post (Appendix B). A non-event specific project by the Berkman Center mapping the Arabic blogosphere found no hard division between English and Arabic blogs (Etling, Kelly, Faris, & Palfrey, 2010, p. 19), while a similar study by the Berkman Center did find a clear division between Farsi and English blogs (Kelly & Etling, 2008). The Arabic project (Etling et al., 2010) found several large national clusters as well as two clusters linking more to foreign language blogs: one to English and one to French. In an exception to blog-based hyperlink studies, Thelwall, Tang, and Price (2003) investigated hyperlink patterns between academic websites in Western Europe. They found interlinking throughout Europe generally occurred in English, although regional linking between countries sharing a common non-English language was also present. In the closest work to date, Herring et al. (2007) examined friend connections between a sample of users on LiveJournal, a blogging site. They found that English was used as a *lingua franca* and that among non-English languages, languages with more users had more persistent, dense, and centralized networks (Russian and Portuguese were examples of this in the study). Bridges between users of different languages where often created by students studying foreign languages as well as expatriates. Furthermore, as in the pilot study for this thesis (Appendix B), non-text content (photos, music) played important roles in connecting users of different languages on LiveJournal (Herring et al., 2007). This thesis extends this work on cross-language activity beyond blogs to other types of user-generated content sites.

2.2.1 Affordances

With the birth of user-generated content platforms, it is clear that users are not merely passive consumers of a pre-defined system. Rather, users are active on multiple platforms, are embedded in various environments and social contexts, and have a range of interactions with other individuals both online and offline (Lamb & Kling, 2003). This insight applies across a range of platforms and systems, but it is particularly important for user-generated content platforms where users are the producers of content and at times even define entirely new use-paradigms for the system. An example of this is seen in the adoption of hashtags on Twitter as a means to increase the discoverability of content or tag it as part of a larger conversation (Huang, Thornton, & Efthimiadis, 2010).

While users are central to the process of adopting and adapting new technologies, the design of the technology also influences the actions users take. Even subtle design decisions have been found to have an effect on how participants behave in online platforms (e.g., B. J. Fogg, 2002; Kraut & Resnick, 2012).

In the extreme, the perspective that design affects users' behaviors leads to a (hard) technological determinism: that technology is the primary cause of user behaviors online (Thurlow et al., 2004). On the other extreme, social constructivism argues that social and culture forces drive how users interact and ultimately use technology. Between these two extremes, the concept of affordances highlights the role of both sociocultural as well as design forces in driving user behavior. First used by perceptual psychologist Gibson (1979), affordances relates to how individuals "perceive the environment directly in terms of its potentials for action" (Gaver, 1991). The concept has been applied to the design of everyday objects (Norman, 1988) as well as to the design of computer software (Gaver, 1991) and CMC more broadly (Hutchby, 2001, 2003).

The benefits of the concept of affordances has been challenged by those within social constructivism (e.g., Rappert, 2003), who state the constructivist metaphor

of "technologies as text" (Grint & Woolgar, 1997) is sufficient. The technologies as text metaphor holds that technologies are like "texts" written by their developers (designed/configured in certain ways), but ultimately read and interpreted by users, who may seek to adapt the technology for another purpose. As used within this thesis, the idea of affordances is not in conflict with the technologies as text metaphor, but is rather complimentary. Hutchby (2001) states that while both the writing of a technology and its reading by users are open, interpretive processes, the idea of affordances is useful to "constrain the ways that [the texts] can possibly be 'written' or 'read' " (p. 447). The "technologies as text" metaphor can be used to understand how design decisions are made, how certain technologies are chosen, etc. in sociology of technology and other fields. In some cases, the concept of affordances is less than helpful (Rappert, 2003); however, affordances is particularly apt for understanding general use and adoption of technology by ordinary society members (Hutchby, 2003), which is the focus of this thesis. The design of software and web interfaces may highlight certain actions or affordances of the software and in so doing influence what users perceive as possible actions and norms of action (e.g., Kraut & Resnick, 2012; Gaver, 1991).

2.2.2 Language affordances

The flow of information between languages may occur in many ways—automated means (e.g., machine translation), professional translation, multilingual contributors on sites with user-generated content, etc. Key to how likely users of an online platform are to engage in multiple languages is how well users perceive the existence of related content in another language and the affordances users perceive to engage with that content through machine translation or other tools. Different design choices on user-generated platforms may encourage or discourage information brokerage by multilingual users and/or xenophiles—monolingual users who nonetheless are "lovers of the unfamiliar, ... people who find inspiration and creative energy in

the vast diversity of the world" (Zuckerman, 2013, ch. 6).

Language affordances can take many different forms. Websites may display content from multiple languages side-by-side (as in Twitter and Facebook) or websites may separate content in different languages into different pages (as when LinkedIn allows users to create separate profile pages for different languages). When content in different languages is separated across pages, those pages with similar content may or may not be linked directly. Articles on Wikipedia about the same concept, but in different languages, are linked together through interlanguage links shown to the user. The websites of many multinational companies, on the other hand, often ask a user to choose a region and language when first entering the site and then do not link similar pages across these different regions or languages. In the most linguistically isolating design, the presence of other-language content may be well hidden from the user. The Google Play app store exhibits this design as detailed in the previous chapter by only showing user reviews written in the user-interface (UI) language the device is configured to use and providing no method to access other-language reviews.

Beyond the prominence of other-language content, sites make design assumptions about how users will interact with the site that afford users different possibilities of finding and interacting with other-language content. Wikipedia, Facebook, and Google Play generally assume users will interact in one dominant language. Facebook offers the possibility to machine translate content, but that option is only available to translate content not in the user's UI language and the machine translation only translates into the user's UI language. Thus, a user with an English-language UI who also reads Spanish is only given the option to translate Portuguese into English even though translating to Spanish might produce a better result for the user given the linguistic similarities between Spanish and Portuguese. Wikipedia similarly makes many monolingual design assumptions: users may only search one language edition at a time. A user is told there are no results for a term (in the lan-

guage they are searching) even if the search would have produced results in another language edition. In contrast to this, Twitter infers users' preferred languages for search results, and allows for a user to have multiple preferred languages, although the user cannot view or specify these languages (M. Hardt, personal communication, October 18, 2013).

Importantly, cross-language information transfer can occur between groups composed entirely of monolingual speakers of different languages given a sufficient set of affordances. Monolingual translation systems (e.g., MonoTrans2 described in Hu, Bederson, Resnik, & Kronrod, 2011) that use only monolingual speakers and machine translation have been able to achieve high fluency and accuracy in translating content. In the case of MonoTrans2, 65% of sentence translations were rated as having both a fluency and accuracy of 5 out of 5 by bilingual evaluators in a pilot project. This compares with a baseline of 10% of sentence translations having both high fluency and accuracy with Google Translate alone. MonoTrans2 is one of a few examples of applications designed specifically for cross-language interaction. Additional examples include Omnipedia (Bao et al., 2012), which shows what concepts are covered in articles in multiple editions of Wikipedia, and Duolingo (Garcia, 2013), which allows users to learn a language while simultaneously translating content. Another exciting design strategy is the use of multilingual users to translate applications and user interfaces (Ellis, 2009).

2.2.3 Trace data for studying user behavior

This thesis makes extensive use of social media data as digital trace data—data left by users as a by-product of their interaction with a digital platform. While users of user-generated content primarily interact with the site to find or write content or to interact with other users, each interaction on the platform leaves a small trace of information about the interaction and the context of the interaction. A new interdisciplinary field of data science is emerging to build tools and techniques to

extract knowledge from this and other data. Such knowledge can take many forms and often includes attempts to understand larger, offline phenomenon.

There are a large number of examples using trace data, but one often cited example is Google Flu Trends.[1] Using a variety of search terms for common flu-like symptoms, Google Flu Trends is able to estimate the level of flu outbreak in near real time, where as official government statistics from the Centers for Disease Control and Prevention (CDC) in the US are available only with a 1–2 week delay. The estimates from search data often match official government figures closely (Figure 2.1).

While an often cited example of the possibilities of trace data, Google Flu Trends also highlights the limitations of such data. Prediction is difficult in and of itself, and calibrating a model that exclusively uses trace data from online platforms to predict offline events is even more difficult. Changes in the Google Search engine itself and search behavior by users have made Google Flu Trends less accurate over time (Lazer, Kennedy, King, & Vespignani, 2014; Olson, J., Marc, Cecile, & Lone, 2013). Changes in the nature of influenza outbreaks such as the non-seasonal H1N1 influenza pandemic also proved very difficult to predict (Lazer et al., 2014; Cook, Corrie, L., & H., 2011). Google Flu Trends, like other uses of trace data, is often a compliment to and not a replacement of other data sources: much more accurate predictions of future flu outbreaks can be made by combining real-time search data together with the less timely reports from the US CDC (Lazer et al., 2014).

These limitations are not directly relevant to this thesis, which purposely avoids predicting any offline events. As used in this thesis, trace data reveals the role language plays in connections between users on user-generated content platforms. This thesis analyzes the current state of the platforms and does not attempt to predict how the role of language will change in the future. Indeed, a prime supposition of this thesis is that the future role of language on user-generated content platforms

[1] See `http://www.google.org/flutrends/`.

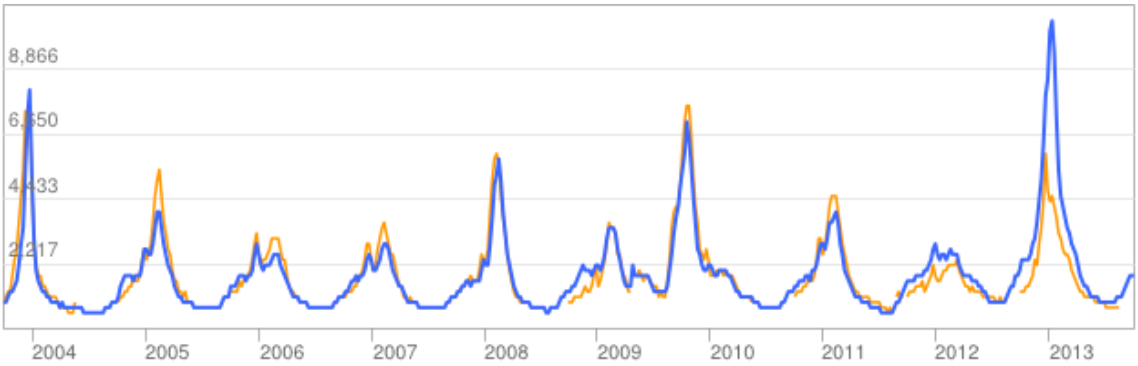


Figure 2.1. Flu trend estimates from search data (blue) plotted against official CDC statistics (orange). Reproduced from Google Flu Trends (`http://www.google.org/flutrends/`).

will be shaped, at least in part, by the design of the platforms, which can affect the behavior of (multilingual) users and the flow of information between users of different languages.

This thesis leverages trace data to better understand user behavior—findings regarding user behavior can then be refined with A/B testing and used in (re)designing online platforms. Trace data has a long history within CMC and HCI research traditions, such as analyzing log data from web servers (e.g., Drott, 1998). Several factors have given trace data increased importance. Increasingly, users are concentrated on a small number of global platforms. The user-generated content platforms under study in this thesis allow for users to interact on these platforms, record these interactions, and, critically, make much of this data available publicly. This enables new understandings of emergent user behavior and interaction that only occurs when users interact together in large numbers.

The use of trace data is complementary to the other CMC/HCI methods of experiments, surveys, diary studies, case studies, interviews/focus groups, and ethnography. Compared to these other methods, examination of trace data allows for analysis of the actions users actually do on a site, as opposed to the actions they remember and report doing or think that they will do. The data is furthermore available in near real-time and often offers whole-of-population data rather than a sample of the

population. Trace data also allows for analysis of behavior in conditions that would be extremely difficult to reproduce in a controlled laboratory setting.

Trace data, however, also has a number of limitations. The contexts of users' actions are often missing, and it is often difficult to match users across different platforms. Thus, it is usually not possible to know the wider environment in which the user performed the action: physical location (home/work, traveling, etc.), other technologies in use simultaneously (did a user see information on one platform and report it on another?), and co-location/communication (was one user in communication with another user via a different platform/technology [or even face-to-face] when performing the action?). Finally, in most cases (and within this thesis) analysis of trace data will only find correlations and not be able to determine causation. The exception to this is large-scale A/B testing or natural experiment setups where data around an intervention can be analyzed in laboratory like conditions (e.g., Bakshy, Rosenn, Marlow, & Adamic, 2012).

2.2.4 Case selection

A number of different definitions for user-generated content are in general use (Organization for Economic Co-Operation and Development, 2007; Hecht, 2013). The term is often used to describe websites or services where the majority of the content is contributed by end users. Common examples of user-generated content sites include photo-sharing sites (e.g., Flickr, Instagr.am), wiki-based sites (e.g., Wikipedia), social media sites (e.g., Twitter, Facebook), review sites (e.g., Yelp), and question and answer sites (e.g., Stack Overflow, Quora). At times only part of a site is user-generated, but this part of the site can often be treated as an example of user-generated content. Examples of this include user reviews for products on Amazon or app reviews in the Google Play store. Generally not included as user-generated content are private communications (e.g., instant messaging, email), personal websites, and other trace data (e.g., search logs, etc.). This broad defini-

tion is typically adapted implicitly by many authors and is the definition used for this thesis.

This thesis examines two user-generated content platforms with very different sets of affordances: Twitter and Wikipedia. Both sites are large, global platforms, popular with users from a wide variety of languages. In addition to their large sizes and international reaches, each platform began with a very different approach to internationalization and the language affordances available. Twitter started with a very passive approach to language. It began as one international system that did very little to structure the flow of information based on language. All users joined a common system, search operated irrespective of language, and "trending topics" were the same across locations and languages. As the platform has grown, this has changed slightly. Separate, local versions of the trending topics were first made available for different major cities. Later the trending topics were personalized for each user. In addition, the current search function initially preferences results from the user's interface language and other inferred language preferences. Nonetheless, clicking to see "all results" or clicking to show the newest tweets will still show tweets from a variety of languages for popular topics.

Wikipedia, in contrast, started with a very active approach to language. It began with separate, independent systems for each language, and users had to register separate accounts for each language edition they wished to edit. Over time, Wikipedia has began to unite these disparate editions by providing for the ability to link together articles on the same topic in different languages. Unified login further allows users to unite their accounts across editions, and Wikimedia Commons provides for a common location to store images and other media that can be used on multiple editions. Nonetheless, users are still asked to choose a language edition on Wikipedia's homepage, and the search facility only operates within one language edition at a time.

These two platforms while perhaps moving towards a common point still differ

in their fundamental approaches to internationalization. Twitter is a good example of a platform designed with high level of language integration *a priori*. Conversely, Wikipedia is a good example of a platform designed with a low level of language integration *a priori*. These differences make the two platforms good choices on which to examine information diffusion, global connectivity, and the role of multilingual users.

2.3 Networks and diversity

Diversity is a key condition for innovation, and computational models show groups of diverse problem solvers can outperform groups of high-ability problem solvers (Hong & Page, 2004). S. E. Page (2007) argues cognitive diversity brings unique perspectives and approaches to problem solving and that identity diversity, which includes racial, ethical, and cultural diversity, contributes to cognitive diversity as our experiences contribute to how we frame a problem (S. E. Page, 2007). This thesis further argues that linguistic diversity contributes to cognitive diversity in a similar way: differing linguistic abilities may open different information sources and enable different actions as in the final vignette in Chapter 1 about locating a car museum in Okinawa.

Within a large population of agents, choosing the top-X best performing agents results in a set where "its members have more ability [but] is less diverse." (Hong & Page, 2004, p. 16,386). This can led to worse performance on future problems in comparison to a more diverse set of agents because all the selected agents have similar knowledge and perspectives, and thus attempt similar solutions. The model has several limitations, including that it ignores issues of communication and learning. Communication is of particular interest, as diverse users may be less able to communicate their ideas. The networks literature, particularly work by Aral and Alstyne (2011), and the diversity literature address poor communication as one of

the conditions which can impede or even decrease performance.

Team diversity also has parallels in city population diversity: the economic growth of cities has been tied to ethnic diversity. An economic study by Ottaviano and Peri (2006) found that greater cultural diversity correlated with (and the authors argue likely caused) higher growth rates in wages and rents for US-born workers in US cities. The authors state, "[t]he foreign born conceivably have different sets of skills and abilities than the US born, and therefore could serve as valuable factors in the production of differentiated goods and services" and, "the skills and abilities of foreign-born workers and thinkers may complement those of native workers and thus boost problem solving and efficiency in the workplace" (Ottaviano & Peri, 2006, p. 10). S. E. Page (2007, p. 331) surveys the literature on diversity within cities and writes, "[m]any historical cases and some recent evidence demonstrate that cities with greater identity diversity can be more productive."

Diversity, however, can also bring miscommunication, conflict, and additional problems. Putnam (2007), looking at ethnic diversity in cities, finds that "in ethnically diverse neighborhoods residents of all races tend to 'hunker down.'" (p. 137). In these diverse neighborhoods, trust is lower (even between individuals of the same race), community cooperation rarer, and friends fewer. S. E. Page (2007) discusses how fundamental differences in preferences can create conflicting goals and incentives for group members to misrepresent their true positions. Such conflicts may lead to "edit wars" on Wikipedia (Yasseri, Sumi, Rung, Kornai, & Kertész, 2012) or flaming (O'Sullivan & Flanagin, 2003). S. E. Page (2007) distinguishes between two types of preference diversity: instrumental and fundamental. Instrumental preferences deal with the means, while fundamental preferences are concerned with the end goals. Instrumental preference diversity can be positive by providing several potential paths to an end for a group to consider. While fundamental differences in preferences can create conflict, fundamental diversity also indirectly drives cognitive diversity and may therefore nevertheless lead to better ideas and higher performance.

S. E. Page (2007) asserts the ultimate value of diversity depends on the task at hand and the management of that diversity. This is echoed by studies of diversity within firms, which suggest that the management of diversity is key to whether diversity aids or hinders performance (e.g., Thomas & Ely, 1996). No study within human-computer interaction has directly addressed the management of diversity between users on sites. However, as shown previously, the design of a site may influence user behaviors and thus influence to the extent to which diversity can be harnessed beneficially on the site.

2.3.1 Linguistic diversity

Even if other forms of diversity matter, why should linguistic diversity? One possible, although controversial, answer to this lies in the principle of linguistic relativity on the interconnection between language and thought (Whorf, 1940). While formulated and popularized by linguist Benjamin Lee Whorf, the idea that language may foster different world views goes back at least to the 17th century (Lee, 1996). Whorf likely did not intend the principle as a formal, testable hypothesis and never stated it as such (Lee, 1996), but the principle was nonetheless formalized in 1953 by Eric Lenneberg as two interrelated hypotheses that are now generally known collectively as the "Sapir-Whorf hypothesis." These hypotheses are:

1. Structural differences between language systems will, in general, be paralleled by non-linguistic cognitive differences, of an unspecified sort, in the native speakers of the two languages.

2. The structure of anyone's native language strongly influences or fully determines the world-view he will acquire as he learns the language (Brown, 1976, p. 128).

If true, linguistic diversity could have a large, positive impact on problem-solving performance. The cognitive differences in the first hypothesis align closely with the

differences that S. E. Page (2007) talks about being important for solving difficult problems, and the concept of world-view is certainly related to the different "perspectives" used in the computational models of Hong and Page (2004). While the so-called "hard" or deterministic version of the Sapir-Whorf hypothesis is generally not accepted, the "weak" version that language *influences* thought and certain kinds of non-linguistic behavior is widely accepted now (Kramsch, 1998).

Another, less controversial, answer for the benefit of linguistic diversity lies in the simple fact that different information is available in different languages (online). As illustrated in the introduction chapter, locating the car museum in Okinawa, Japan, was difficult using English while locating it using Japanese was comparatively straightforward. Although there is no doubt that much overlap exists in the information available in different languages, there is also a surprising diversity of information not shared between languages. Although the English-language edition of Wikipedia has more articles than any other edition, it is not true that the English edition is simply a superset of other language editions. That is, the English edition does not have all the articles in other languages plus more. In fact, the English edition contains only about half of the articles in the second largest edition, German (Hecht & Gergle, 2010b). Similarly, about 40% of the articles in the Japanese edition have a corresponding article in the English edition (Hecht & Gergle, 2010b). These very crude measures highlight only a sliver of the diversity in topics and information contained within this one platform.

Content diversity across languages may arise for a number of reasons, but at least one reason for user-generated content platforms is that users often contribute content local to them (Hecht & Gergle, 2010a). For example, over 50% of Flickr users contribute content that, on average, is generated within 100km or less from their home locations (Hecht & Gergle, 2010a). While this might be expected for spatial content production models like that of Flickr where a user must physically be present to contribute a photo to the repository, "distance still matters a great

deal on Wikipedia's 'flat earth'" (p. 231) model that allows anyone anywhere in the world to contribute to any article (Hecht & Gergle, 2010a). This local focus means that different regions have different depths of coverage in different languages (Hecht & Gergle, 2009).

The diversity of information across languages and the contribution of local content to user-generated content platforms is neither inherently good nor bad. However, it does mean that users speaking different languages have access to different information online. It also follows that a multilingual user speaking two or more languages has easier access to a wider amount of information online than does a monolingual user with easy access to information in only one language.

2.3.2 Homophily

Prevailing theory in social network analysis suggests individuals tend to group together with those similar to themselves, a property known as homophily (Lazarsfeld & Merton, 1954) and commonly expressed by the adage, "birds of a feather flock together." This common property "structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange..." (McPherson, Smith-Lovin, & Cook, 2001). Homophily makes connections between similar individuals more likely. This in turns leads to a network structure where similar individuals are connected in dense clusters with relatively few ties between these clusters. These clusters come to define the social groups in which information is sought and shared. As Burt (2005, p. 15) explains, "the higher density of relations within groups mean that information circulates more within than between groups."

In addition to the many social foci that lead to a clustered network structure within a single language group, past research suggests users also cluster in online platforms more broadly by language. Users on LiveJournal, for instance, were more likely to friend other users who write in the same language (Herring et al., 2007) and bloggers writing about the Haitian earthquake in Japanese, English, or Spanish

were more likely to link to other blog posts in the same language (Appendix B).

In contrast to these arguments from social network analysis, early rhetoric about the Internet characterized it as a global system that would make previous geolinguistic divides less relevant (Negroponte, 1995). Early studies of the Web supported this idea with Adamic (1999) finding any two websites in her sample were, on average, only 4.2 between-site links away from one another—an incredibly small distance for over 65,000 sites. Similarly, an early study of Twitter in July 2009 of 41 million users found any two users were separated by approximately 4.12 follower–following relationships (Kwak, Lee, Park, & Moon, 2010). The size of the Web and lack of a sampling frame as well as the limits on the Twitter API make newer estimations of the average path lengths in these networks very difficult to compute. Some models predicted that the power-law distribution of links on the Web[2] would mean the average path length would remain small despite large growth (Albert, Jeong, & Barabasi, 1999). This however has not been empirically tested, and it is unclear whether the average path lengths have remained small as the Web, Twitter, and other platforms have grown and diversified with users coming from more languages.

2.3.3 Bridging clusters and innovation

The idea of small-world networks addresses the two competing ideas of high clustering and low average path lengths (average distance between any two nodes). Watts and Strogatz (1998) formalized a model of a small-world network as the intermediate zone between a regular (e.g., ring lattice) network and a random network (Figure 2.2). The model depends on only one parameter p, the probability of an edge being rewired. It begins with a ring lattice where every node is connected with its k nearest neighbors. Each edge is then rewired to a random destination with probability p. For $p = 0$, no rewiring occurs and the ring lattice remains highly clustered with large

[2]Power-law distributions are a specific type of fat-tailed (or more generally heavy-tailed) distributions. Empirical power-law distributions are difficult to verify and often are only approximate or over a limited range (Clauset, Shalizi, & Newman, 2009)

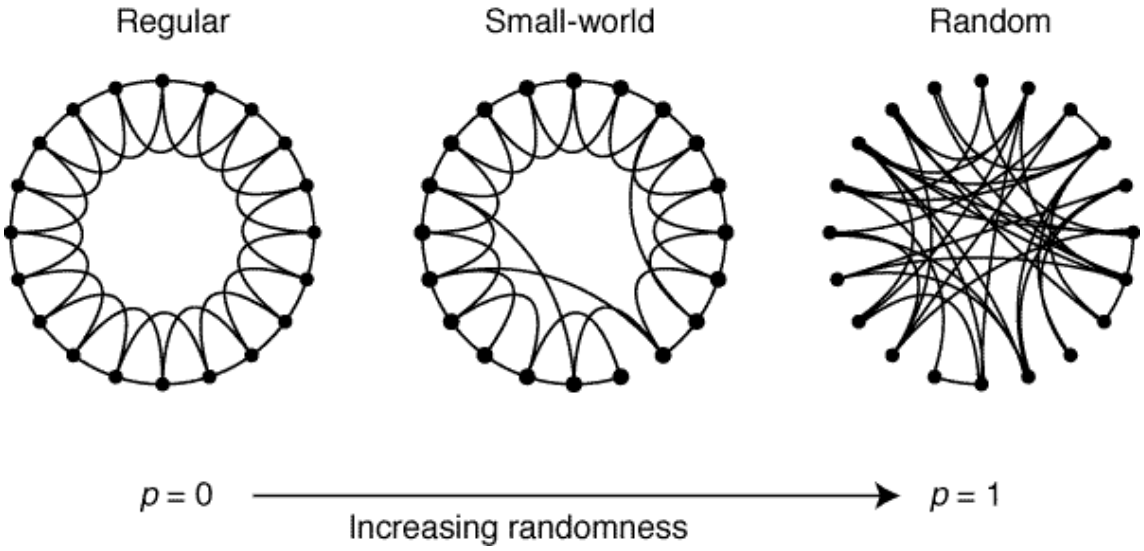


Figure 2.2. From Watts and Strogatz (1998) showing the effect of parameter p on a simple example network with 20 nodes each connected to their 4 nearest neighbors. Reprinted by permission from Macmillan Publishers Ltd: Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi:10.1038/30918, copyright 1998.

average path lengths. For $p = 1$, all edges are rewired and the resulting network is simply a random network with low clustering and small average path lengths. Small-world networks occur for the intermediate values of p ($0 < p < 1$). These small-world networks exhibit both relatively high clustering and also low average path lengths.

Several networks are known to be small-world networks including the power grid of the western United States and the collaboration network of film actors (Watts & Strogatz, 1998), the network of creative artists making Broadway musicals from 1945 to 1989 (Uzzi & Spiro, 2005), and the network of large US companies as connected by common board members (Davis, Yoo, & Baker, 2003). More broadly, Milgram (1969) popularized the idea of small-world networks through an experiment showing low average path lengths between individuals living in Nebraska and Boston. This suggested that general interpersonal networks also exhibit this small-world property (at least within one country and language).

Dodds, Muhamad, and Watts (2003) replicated the work of Milgram in an online

experiment where participants were asked to send a message to a target user by emailing an acquaintance at each step. More than 60,000 e-mail users took part in attempting to reach one of the 18 target persons in 13 different countries. The study found the average completed chain had a length of 4. However, like Milgram's study (Kleinfeld, 2002), the drop out was extremely high (less than 2% of started chains were completed). Assuming the primary reason for incompletion was lack of motivation (rather than inability to reach the target), Dodds et al. (2003) compute an adjusted median length of 5 for chains in which the target was in the same country as the starting user and 7 for chains in which the target was in a different country from the starting user. The study does not specifically mention language, but over half of all participants "resided in North America and were middle class, professional, college education, and Christian" (p. 827).

Previous research has shown half of the small-world network idea applies online with regards to language: high clustering by language. If communications in online platforms really are small-world, there must be users with interlanguage ties that bridge clusters of users of different languages. Individuals who speak multiple languages (multilinguals) are a natural class of users who may fulfill this requirement. Work examining personal networks on Twitter suggests that some multilingual Twitter users do act as local bridges between users of different languages (Eleta & Golbeck, 2012); however, the amount of bridging and its effects at a large-scale level remain unclear.

More generally, past research suggests many benefits for connecting clusters in social networks at both a large, network level as well as at an individual level. Uzzi, Amaral, and Reed-Tsochas (2007) in a review article of small-world research explained that in small-world networks, "the many separate clusters enabled the incubation of a diversity of specialized ideas while short paths allowed ideas or resources to break out of their chambers and mix into new and novel combinations (Uzzi & Spiro, 2005; Fleming & Marx, 2006)."

At the network level Uzzi and Spiro (2005) and Fleming and Marx (2006) agree on the effect a small world structure ought to have on creativity, success, and innovation, but the two studies found conflicting results when analyzing data. Uzzi and Spiro (2005) analyzed the network of artists creating Broadway musicals from 1945 to 1989. Controlling for a wide-range of economic and social variables (most of which were non-significant), they found the small worldliness of the network was a significant predictor of a musical's artistic success (based on critics' reviews) and financial success (based on ticket sales). Three quantities were used to measure the network. First, a clustering coefficient ratio was formed by dividing the clustering coefficient of the network understudy by the clustering coefficient of a comparison graph with the same number of nodes and edges, but the edges drawn at random. Second, a path length ratio was similarly defined. Finally, the small worldliness of a network was measured as Q, the clustering coefficient ratio divided by the path length ratio of the network. Larger values of Q correspond more closely to a small-world network with smaller than expected path lengths and a larger than expected clustering coefficient.

This small-world effect was parabolic: performance increased to a threshold as the network became more small worldly; however, beyond this threshold the performance decreased with further increases in Q. Uzzi and Spiro (2005) state that with small levels of Q, creative materials remains cloistered in separate teams. However, beyond the ideal point, too much connectivity "homogenize[s] the pool of creative material, while repeated ties and third-party-in-common ties promote common information exchanges, decreasing artists' ability to break out of conventional ideas or styles that worked in the past but that have since lost their market appeal" (Uzzi & Spiro, 2005, p. 492).

In this study, the path length ratio had very little variance. So, the clustering coefficient ratio alone explains the data equally well when used in a separate model from the small world measure Q . It is thus unclear whether Broadway production

teams are simply in a small-world network within which clustering is the important variable; or, as the authors assert, it is actually the state of being in a small-world network itself that is important.

Fleming, King, and Juda (2007) predicted similar outcomes in their regional networks of American copyright patents, but found differing results. In their data, path lengths varied widely between regions as did the number of inventors in the largest connected component for each region. The authors found that these aspects of the networks (average path length and size of the largest component) were more important predictors of innovative success than either clustering or small worldliness.

At the individual level within a network, much research supports the importance of bridging clusters and connecting diverse individuals. Users who bridge different clusters are more likely to be exposed to different information from the clusters they bridge. These users can take information from one cluster and apply it in another cluster.

Granovetter (1973) first asserted the importance of bridging, weak-ties in his seminal work, "The Strength of Weak Ties." He theorizes that as an individual's (call him ego) network of close friends has much overlap and is very dense (that is ego's close friends are likely friends with each other). In contrast, ego's acquaintances are less likely to be socially involved with one another and thus form a less dense network with few, if any, overlaps. However, each acquaintance is likely to be part of his own dense network of close friends, which is likely different from that of ego. Thus, the weak tie between ego and his acquaintance is "a crucial bridge between two densely knit clumps of close friends" (Granovetter, 1983, p. 202). Granovetter's empirical work supports this theory: he finds that more individuals found new job information from their acquaintances more often than from their close friends. Weak ties have been found important to many aspects including the spread of ideas and innovations (e.g. Fine & Kleinman, 1979; Burt, 2004).

Similarly, spanning structural holes (bridging clusters) in large companies has

been associated with a "vision advantage" that the individual may leverage, resulting in better compensation, more positive performance evaluations, and promotions (Burt, 2004). As "opinion and behavior are more homogeneous within than between groups" these brokers whose networks span structural holes are "more familiar with alternative ways of thinking and behaving" (Burt, 2004, p. 394). This leads to a vision advantage that forms a type of social capital. He further finds, "[t]he between-group brokers are more likely to express ideas, less likely to have ideas dismissed, and more likely to have ideas evaluated as valuable" (p. 394).

More recently, Aral and Alstyne (2011) raised an important caveat to this theory. The important question is not just where novel information is, but how much of that information reaches a user per unit of time. By definition, we talk less frequently to weak-tie acquaintances than to strong-tie friends. Thus, even though an acquaintance has more novel information, Aral and Alstyne (2011) argue a person is less likely to receive that information given that the person interacts with the acquaintance less frequently.

Thus, when analyzing the amount of novel information available to individuals in a network, the structural diversity of ties alone is not sufficient to understand the novel information to which individuals in a network are exposed. Aral and Alstyne (2011) show that the amount of information actually exchanged across a tie (the authors refer to the amount of information shared across a tie as bandwidth) and the amount of overlap in information between different individuals in the network are important aspects to measure in addition to the structural diversity of ties and other tie-strength attributes (e.g., emotional closeness). Aral and Alstyne (2011), looking only at dyadic ties, showed that often bandwidth is much greater between strong-tie friends than weak-tie acquaintances. Close friends have more opportunities to exchange information and are more likely to know what information is of interest to one another. In addition, the amount of information overlap between different alters in the network is important to how valuable structurally diverse ties (i.e., bridging

structural holes) will be. The greater the information overlap among alters the less valuable structural diversity should be in providing access to novel information (Aral & Alstyne, 2011).

An important limitation of the work of Aral and Alstyne (2011) is that it only examined communication between pairs of individuals and did not consider the broadcast nature of most social media platforms. By definition, strong-tie friendships require more time and effort to maintain the relationship as compared to weak-tie acquaintanceships. This means that a larger proportion of the friends/followers users have on social media platforms are actually weak-tie acquaintances and not strong-tie friends. Bakshy et al. (2012) test the importance of weak ties in diffusing novel information in a large-scale online experiment with Facebook. Following the work of Aral and Alstyne (2011), they find that any one strong tie is more influential when compared to one weak tie. However, since weak ties are much more abundant than strong ties on the platform, most information actually spreads via weak ties on Facebook (Bakshy et al., 2012). Thus, weak ties (defined in terms of interaction frequency) taken collectively are more important in diffusing novel information than the more limited number of strong ties.

The design of social media platforms changes the underlying assumptions of the work of Aral and Alstyne (2011) that strong-tie friends have more opportunities to exchange information than weak-tie acquaintances. The broadcast nature of social media platforms, in which messages are usually sent to all of one's friends/followers, give friends and acquaintances equal opportunities to exchange information (at least on the platforms themselves). To the extent to which cross-language ties act as weak ties, we might expect that any one cross-language tie will not yield a large amount of novel information. However, if a user has many cross-language ties, the information gained in aggregate from the many cross-language ties could be substantial.

The literature on weak-ties and diversity suggests cross-language connections could be incredibly important, and the few studies analyzing cross-language activ-

ity support this. When language boundaries are punctured the subset of content transferred into another language may be amplified much further than in the original language as is evident by China foreign correspondents being more likely to read and use information from English-language blog posts about China as compared to Chinese-language blog posts (MacKinnon, 2008). Multilingual users sit in a unique position with easier access to information in multiple languages than monolingual users. Active multilingual users may serve as brokers in the network connecting speakers of one language to content and users in another language. Multilingual users may signal interesting content in another language to their contacts, who may then try to use additional tools such as machine translation to understand more about the content (although contextual and cultural differences remain that may result in different interpretations of the same content). Photos, videos, and other multimedia materials may also be shared more frequently across languages given their broader intelligibility. These materials may lead users to further relevant content in a foreign language. The sharing of content between users in different languages can lead to discovery of novel information and help correct self-focus biases where more content is available about regions where a language is spoken than regions where it is not (Hecht & Gergle, 2009).

In summary, users on social media platforms are likely broadly clustered by language with a large diversity of unique content not shared across multiple languages. Multilingual users may act as brokers in this network moving information between speakers of different languages. The interlanguage ties of multilingual users may result in a small-world like network structure with low average path lengths despite the broad clustering of users by language. The movement of information between languages may in turn accrue benefits to the multilingual users themselves and to the network of all users as whole.

The next chapters examine the extent to which users cluster by language and the role of multilingual users as bridges between languages. Chapter 3 examines Twitter

at a global level, while Chapter 4 does the same for Wikipedia. Informed by the findings of these first two empirical chapters, Chapter 5 examines cross-language activity in more depth by considering both users and the content they contribute. It does so by studying users contributing content about Okinawa, Japan, a region with both English and Japanese speakers. The thesis concludes in Chapter 6 by analyzing the commonalities and differences found in the empirical chapters and using these to suggest possible design changes, which further work will investigate.

Chapter 3

Global Connectivity and Multilinguals in the Twitter Network

The limits of my language means the limits of my world. —Wittgenstein

3.1 Introduction

Given the wide differences in information available in different languages (Hecht & Gergle, 2010b; Hong, Convertino, & Chi, 2011), multilingual users of social media platforms may be able to share novel information with users of different languages and thereby help overcome in a small way traditional language divides (Eleta & Golbeck, 2012). This chapter examines the role that users engaging with content in multiple languages (referred to as multilingual users) play in the Twitter network in order to better inform the design of search and friend/follower recommendation systems on social media platforms. It analyzes the global connectivity of the Twitter

This chapter is adapted from the following publication:
Hale, S. A. (2014). Global Connectivity and Multilinguals in the Twitter Network. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14. New York: ACM. doi:10.1145/2556288.2557203.

mentions and retweet network to assess the extent to which language structures the network and asks whether multilingual users form cross-language bridges for information exchange that provide unique connections without which the network would break apart. The study finds language is a key force structuring ties in the network, but that multilingual users are an important case to design for: over 10% of users engage with content in multiple languages and these users are more active than their monolingual counterparts. There is thus value in a single multilingual system, and language should not be used as an absolute (i.e., only ever returning search results or recommending friends in the main language of the user).

English is the most used language on Twitter, but still accounts for less than half of the messages exchanged on the network (Hong et al., 2011). Unlike the different language editions of Wikipedia, which are only loosely connected, Twitter takes a more integrated approach to language. Users can follow users from multiple languages and have only one profile (unlike the one profile per language of LinkedIn). Understanding how users connect across languages is important in designing the search function and follower recommendation system. To what extent should language be used to restrict or prioritize results in cases where a term occurs in tweets of many languages (e.g., a company name)? Similarly, what role should language play in recommending other users one might follow? If language is used, but few results are available in the user's preferred language(s), what secondary languages might be most beneficial to include results from?

Past work in international telephony, television, and social media have often noted that language and country may be strong impediments to communication (e.g., Appendix B; Barnett & Choi, 1995; Hale, 2012; Herring et al., 2007; Nordenstreng & Varis, 1974; Takhteyev et al., 2012; Wilkinson & Thelwall, 2012). Exactly what information is shared between speakers of different languages on social media and to what extent, however, remains unclear. The handful of studies looking at language and social media have found language plays a large role in structuring the

hyperlink relationships between blogs (Appendix B; Herring et al., 2007) and the follower/following relationships between Twitter users (Takhteyev et al., 2012).

Nevertheless, the 'weightless bits' of digital communication may render barriers of language and geography less relevant online compared to the 'heavy atoms' of the physical world (Zuckerman, 2013). Online platforms must strike a balance between reinforcing within language tendencies and allowing for the serendipitous discovery of new content from speakers of other languages on the platform. Even if users do not understand the language of the content, awareness of the content allows users to apply machine translation and/or possibly understand the content in part (e.g., embedded pictures). On Twitter, users select other users to follow and see a feed of what these users have posted when logged in. There is no facility on the site to easily direct a message only to a subset of users, which means content authored by multilingual users is sent to all their followers irrespective of language. This inability to direct content to a subset of one's network stands in contrast to the 'circles' at the center of Google+ and the lists and groups features of Facebook. This may be positive and introduce users to new content as users may be more likely to try to understand or apply machine translation to foreign-language content from friends or acquaintances than from unknown users.

As we have seen in Chapter 2, prevailing theory in social network analysis suggests individuals tend to group together with those similar to themselves, a property known as homophily and commonly expressed by the adage, "birds of a feather flock together" (Granovetter, 1973). This leads to networks having many clusters or groups of nodes "within which network connections are dense, but between which they are sparser" (Newman & Girvan, 2004).

These clusters result from many factors (gender, race, age, etc.) including language (Appendix B; Herring et al., 2007). As these clusters come to define the social circles in which information is sought and shared, each cluster comes to contain unique information (Granovetter, 1973). Previous research on language is consistent

with this view, finding very different information is available in different languages. On Wikipedia, for instance, article overlap between language editions of the encyclopedia is low (Hecht & Gergle, 2010b). On Twitter, hashtags and links to different web domains often differ across languages (Hong et al., 2011). Given this difference in information, cross-language connections can be extremely important. Individuals who connect different clusters may be exposed to novel information that differs from the information in a single cluster.

Previous work has found language and geography structure the following–follower network on Twitter (Kulshrestha, Kooti, Nikravesh, & Gummadi, 2012; State, Park, Weber, Mejova, & Macy, 2013; Takhteyev et al., 2012), but has not examined the bridging potential of multilingual users. The closest work to date by Eleta and Golbeck (2012) used an ego-net approach, examining the follower–following networks of 73 multilingual Twitter users and found that 55 of these users connected groups of other Twitter users across language divides. In only 18 cases were two or more language groups tightly connected together. In 15 cases, a few gatekeepers connected the language groups, while in 22 cases there was one predominant language group with another language group in the periphery.

This chapter builds upon this work in two ways. First, it focuses on the emergent network of message sharing activity. This network is formed by Twitter users mentioning other users by username, replying to tweets, and retweeting (resending/forwarding) tweets. This higher bar for a network tie captures the dynamic, interaction patterns rather than the more static, following network on the platform. Consistent with previous work and corresponding with the low overlap in hashtag/link domains, it is predicted that *the mentions/retweet network will have many clusters composed of a single, dominant language* (H3.1). Second, this chapter examines the bridging role of multilingual users at a full network level rather than an ego-net level. It specifically tests the hypothesis that *multilingual Twitter users serve as bridges between different clusters in the network* (H3.2).

If language does structure the network but multilingual users serve as bridges between languages as predicted, then it would be useful to know the distribution of multilinguals and how they connect users across languages when designing search and friend recommendation approaches. Strongly connected languages are likely good languages to draw additional results from for search or friend recommendations when insufficient results are available in the preferred language(s) of the user.

In what languages are users more likely to cross language divides? Qualitative and survey work has suggested that users writing in less represented languages will more likely cross-language boundaries (Durham, 2003; Warschauer et al., 2002; Wei & Kolko, 2005). Importantly, survey respondents in Uzbekistan reported crossing language boundaries online even while simultaneously reporting low confidence in their foreign language skills (Wei & Kolko, 2005). This leads to the hypothesis that *users writing in less-represented languages will be more likely to cross language boundaries than users writing in highly-represented languages* (H3.3).

When users do cross languages, linguist David Crystal (2003) suggests these users will engage with content and users in larger languages, particularly English. Previous studies of language connectivity online have also suggested English plays a special, bridging role connecting speakers of other languages. Herring, et al. (2007) examined LiveJournal blogs and found language to be a strong factor in structuring 'friend' relationships on the site. English served as a bridging language, and "when non-English journals friend a journal in another language, that language is almost always English." Similarly, the pilot work for this thesis examining Japanese, English, and Spanish-language blog posts about the 2010 Haitain earthquake found significantly fewer links between Japanese and Spanish than either Japanese and English or Spanish and English (Appendix B). Both studies, however, had relatively small sample sizes and started with a small number of languages. English may not serve as strong a bridging role when more languages are included. A mapping of the Arabic blogosphere suggested French, for instance, may serve as an important bridging

language in North Africa (Etling et al., 2010). Similarly, survey work suggests Russian could serve a bridging role in the former Soviet bloc (Wei & Kolko, 2005).

This chapter will analyze the collective role of users in different languages in the Twitter mentions/retweet network to identify which languages serve more of a collective bridging role. It specifically tests if *English-language users as a whole form more bridges than users writing in other languages* (H3.4). The extent to which English is a natural bridging language in the network can help determine the extent to which English is a good default or fall-back language for search and friend recommendation results when little is known about the user.

Whereas H3.3 predicts the languages from which users will initiate cross-language activity, H3.4 predicts the languages to which these multilingual users will likely connect. The two hypotheses are complementary and together predict that users from less represented languages will cross language boundaries to engage with English-language content (and perhaps to a lesser degree content in other well-represented languages). These more-represented languages may then collectively form bridges between users in multiple less-represented languages.

3.2 Data

The data analyzed comes from the Twitter sample stream with 'spritzer' access, which gives a 1% sample of all tweets. Tweets were collected from June 11, 2011 to June 29, 2011.[1] From each tweet, the text of the tweet was extracted and analyzed with the Compact Language Detection (CLD) framework to determine the language of the tweet. The username of the author as well as any mention or retweet of another user were also extracted.

Language identification is difficult on such short text (Carter et al., 2011), but methodological work has found the CLD kit to produce acceptable results for a

[1]This was an uneventful period with no large geopolitical events, and the data thus aims to approximate the average background level of activity rather than examining a particular event as some past studies have done.

wide range of languages on Twitter (while recognizing limitations such as not being trained to recognize romanizations of languages with another traditional script) (Appendix A). CLD was developed by Google and is used within Google Chrome to detect the language of content. Urls, hashtags, and mentions were temporarily removed from the text of tweets for language detection per the findings of methodological work reported in Appendix A. CLD distinguishes between traditional and simplified Chinese as well as between Indonesian and Malay. However, given the similarity in these pairs, the two Chinese scripts were treated as one language (zh). Similarly Indonesian was included with Malay (ms).[2] Given the difficulties with shorter text it is useful to establish a threshold under which the detection of a language is more likely classifier error than authentic use of the language. For this study, a user was only considered to use a language if at least 20% of the user's tweets and at least two tweets were detected in that language.[3] Any user meeting this requirement for two or more languages was classified as a multilingual user, while the remaining users were classified as monolingual users. All multilingual users, therefore, authored at least four tweets in total (two tweets in each of two languages at a minimum). Users with less than four tweets were excluded entirely to avoid having any users in the sample with insufficient data to determine if they are monolingual or multilingual in their Twitter usage.

User mentions (@user) and retweets of another user's content were extracted to form weighted edges of the network. In the final network each node represents a Twitter user and each weighted, directed edge e_{ij} represents the number of tweets user i authored that mentioned or retweeted user j. Each node also has the total number of tweets in the sample authored by that user, the user's most-used language,

[2]Grouping these pairs together actually reduces the magnitude of the role multilingual users play, but is more linguistically appropriate.

[3]20% was chosen through manual examination of ad hoc subsets of multilingual users, which found the most prolific users had slightly less than 20% of their tweets misclassified as the wrong language in the worse cases. The 20% cutoff affects only 6% of all users. Without this bar all the results hold, often with larger effect sizes making multilingual users appear even more active and more responsible for the bridges in the network.

Language	User count	Tweets/user	(s.d.)
English (en)	375,474	8.43	(5.81)
Japanese (ja)	137,263	9.51	(8.38)
Portuguese (pt)	133,501	7.95	(5.18)
Malay / Indonesian (ms)	106,223	8.44	(5.51)
Spanish (es)	70,246	8.01	(5.18)
Dutch (nl)	31,035	8.81	(5.84)
Korean (ko)	16,123	10.46	(8.96)
Thai (th)	8,629	9.03	(6.48)
Arabic (ar)	7,679	8.30	(6.48)
French (fr)	5,769	9.06	(6.71)
Filipino / Tagalog (fil)	5,393	6.74	(3.64)
Italian (it)	4,795	9.03	(6.17)
Turkish (tr)	3,759	7.33	(4.49)
German (de)	2,299	8.15	(5.38)
Russian (ru)	2,282	7.72	(6.25)

Table 3.1. Languages with the most users and the average number of tweets per user. Each user is placed in the language he or she uses most frequently.

and the percentage of the tweets by that user in the user's most-used language.[4]

This study excludes the least-active users on Twitter due to the need to observe multiple tweets in order to classify users as either monolingual or multilingual and the limitations of the Twitter API. Raising the number of tweets required for each user improves the quality of language detection, but also decreases the number of users with sufficient data to remain in the sample since the distribution of the average number of tweets per user is heavy-tailed with most users writing only one tweet and a very small number of users writing a large number of tweets. This problem is aggravated by the fact that the Twitter API returns only a 1% sample of tweets, but is mitigated in part by the longer length of data collection. Nonetheless, excluding the least-active users is not a major limitation as these users send only a small number of tweets each, and they are unlikely, therefore, to have any sustained bridging role in the wider network. In contrast, the most active users have a greater probability of being at the core of the network, and the cross-language activity at

[4]The code used to record the stream (PHP), construct the network (Java / Hadoop), and perform the analysis (R / igraph) are available at `http://www.scotthale.net/pubs/?chi2014`

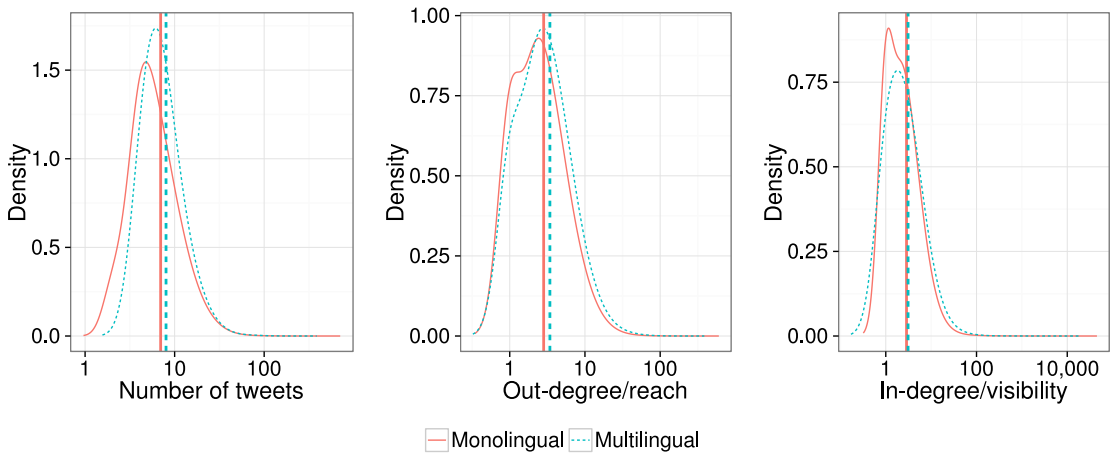


Figure 3.1. Density plots comparing tweet count, out-degree, and in-degree for multilingual and monolingual users. Vertical lines show mean values.

the core of the network is more likely to affect a wide number of users. Spam accounts are another challenge with Twitter data. To reduce the presence of these and focus on the role of humans in the network, the network is filtered to include only users receiving at least one mention/retweet ($indegree \geq 1$). Furthermore, only the largest weakly-connected component is selected.[5]

This results in a network with 916,836 nodes and 2,652,618 directed edges. The corresponding undirected network with every directed edge converted to an undirected edge and mutual edges combined has 2,380,675 undirected edges (i.e., there are 271,943 mutually connected users in the directed network).

3.3 Analysis

This section first compares multilingual and monolingual users. It then investigates the possible bridging role multilingual users play before looking at the isolation/ insularity of speakers of each language. It finishes with an analysis of the bridging role played by speakers of different languages and the specific role of English in the

[5]Other academics studying Twitter have independently arrived at the same thresholds as the best compromise between increased language detection accuracy and user inclusion (e.g., Ronen et al., 2014).

network. After this, the chapter discusses the results of the study and suggests areas for further research.

Of the over 916,000 Twitter users in the sample, 103,645 (11%) were observed to use more than one language and designated as multilingual users. The distribution of messages sent per user is highly skewed, reflecting a heavy-tailed distribution among both monolingual and multilingual users. A list of the languages with the most users is given in Table 3.1. This list is broadly in line with the previous work by Hong, Convertino, and Chi (2011) even though a different language classification approach was used (LingPipe and Google's Language API vs. the compact language detection kit). The list reflects varying uptake of the Twitter platform across the world. The conspicuous absence of Chinese, a very large language on the web, likely reflects the difficulty of accessing Twitter in mainland China as well as the availability and popularity of homegrown alternatives (Liao, 2014).

Multilingual users, on average, were more active than their monolingual counterparts sending a mean 8.0 (median 7, sd 5.2) messages per user on average compared to a mean 7.0 (median 5, sd 5.6) messages per user among monolinguals. Multilinguals also have a higher out-degree (mean 3.4 vs. 2.8; medians both 2, sd 3.9 and 3.3 respectively), and a slightly higher in-degree (mean 3.1 vs. 2.9; medians both 2, sd 14.1 and 21.5). Figure 3.1 shows these comparisons.[6]

3.3.1 Language and network structure

Many classic network clustering algorithms use betweenness to remove edges with the highest scores from a network until it breaks into unconnected components, and the number of components into which the network is optimally divided is usually measured with modularity, which is a goodness-of-fit measure comparing the density of edges within and between clusters (Newman & Girvan, 2004). These methods,

[6] All differences reported in this chapter are statistically significant with p-values less than 0.001 as established by t-tests on the sample means. These p-values are not reported in the main text for readability.

however, are computationally expensive and difficult to run on a network of this size. Raghavan, Albert, and Kumara (2007) developed a label propagation approach to detect clusters from the structure of the network alone. Each node is given a unique label and at each step changes its label to the label that most of its neighbors have at that time, with ties among multiple labels being broken randomly. The method does not consider edge direction or edge weight. The algorithm operates in near linear time and the authors claim that after five iterations most node labels are fixed regardless of the network's size. To achieve the best performance, the author wrote a custom implementation of this algorithm to parallelize it and make use of multiple processors/cores on modern computers.[7]

After 72 runs, 99.96% of the nodes had fixed their labels. Among the 17,480 clusters found by the algorithm, 12,740 (72.9%) of the clusters were formed of users who all shared one common most-used language. This is significantly higher than the 22 clusters (0.12%) that would be expected to share a common language if language assignment was independent of network structure.[8] Indeed, the number of languages in all clusters (mean 1.4, sd 1.1, median 1) was significantly lower than what would be expected randomly (mean 4.2, sd 2.4, median 3.7). Many of the clusters found were relatively small (mean size of 52, median size of 6), while a few were extremely large (see Figure 3.2).

Seven clusters had more than 10,000 users each, and collectively held 61% of all the users in the graph. Four of these seven clusters are heavily dominated by speakers of one language as shown in Table 3.2 and Figure 3.3. English users[9] are the most numerous language group in three of the seven largest clusters. Ninety-nine percent of the 114,826 users in the second-largest cluster use English most frequently.

[7]Code at `http://www.scotthale.net/pubs/?chi2014`

[8]Random expected values were formed by keeping the existing network and clusters, but shuffling/permuting language labels and assigning them to nodes randomly. This was done 100 times and the results averaged.

[9]An unfortunate limitation of the English language is that it often uses the same adjective to refer to both the people of a given country and people who speak the language of that country. That is, "English users" can refer to either users from England or English-language speaking users. In this chapter, all such terms refer strictly to languages and never to countries/locations.

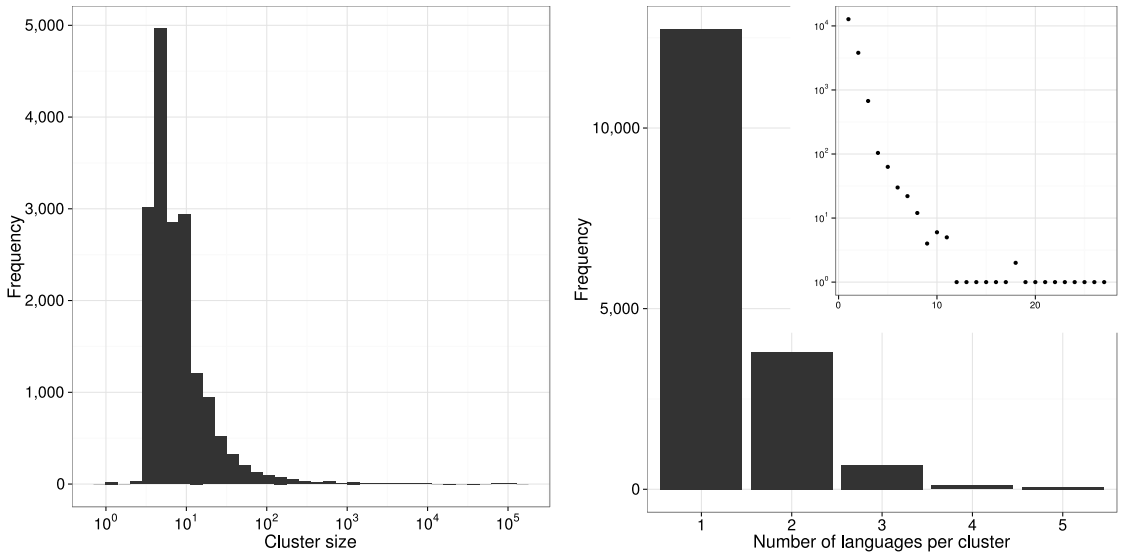


Figure 3.2. Histograms of the size of clusters found by the label propagation algorithm (left) and the number of languages within the clusters (right). The right histogram is truncated, while the insert displays the full distribution on a log-log scale.

In contrast, however, the other two English-language clusters while smaller are less dominated by English: only 75% and 55% of the users in these clusters use English most frequently, suggesting users in these clusters might be more cross-lingual in their communications.

The mentions/retweet network is more dynamic, representing only active communication patterns (compared to the more passive, static follower–following network), but it is also more volatile. Some of the smaller sized clusters found may have grouped together given a larger or longer data sample. Even so, most of these clusters would likely remain dominated by one language as most of the largest observed clusters are.

As mentioned previously, modularity is a measure of the goodness-of-fit of a given division of a network. A value of zero indicates the division is no better than random. (Theoretically the value has a lower limit of -1 with negative values indicating worse than random divisions, but such values are rarely observed.) Values approaching the maximum value, 1.0, indicate the network divides easily into densely connected

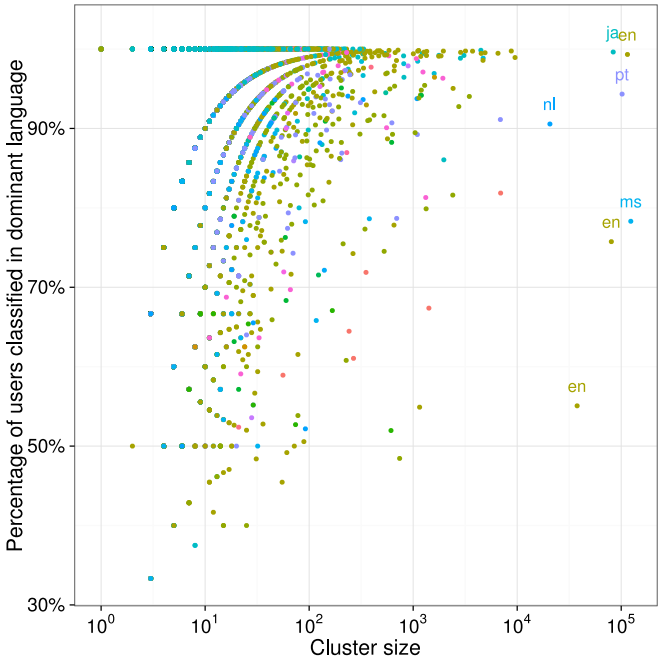


Figure 3.3. Scatter plot of cluster size and the percentage of users in the cluster most often using the most prevalent language. Details of the largest clusters are given in Table 3.2.

Most-used language	% users in most-used language	Number of languages	Number of nodes
Malay (ms)	78.3	41	123,616
English (en)	99.3	39	114,826
Portuguese (pt)	94.3	40	101,987
Japanese (ja)	99.6	19	83,785
English (en)	75.7	44	80,387
English (en)	55.1	42	37,688
Dutch (nl)	90.6	23	20,634

Table 3.2. Clusters with over 10,000 nodes found through the label propagation algorithm.

clusters with sparse connections between them. Newman and Girvan (2004) state typical values are usually in a range from 0.3 to 0.7. The divisions found through the label propagation algorithm have an extremely high modularity score of 0.81, indicating network is highly clustered. Simply dividing the network into groups *a priori* based on the majority language of users (i.e., all English users in one group, all German users in another, etc.) results in a modularity score of 0.68. Although not as strong a division as the groups found by the label propagation algorithm, the modularity score still indicates dividing the network by language alone captures much of the clustered structure in the network.

The structure found with the label propagation algorithm represents a strong division of the network, and the analysis shows that far more of the clusters found are composed of a single, dominant language than would be expected randomly, confirming H3.1. This claim is further strengthened by the high modularity score for a deductive division of the network based solely on language.

3.3.2 Bridging role of multilinguals

Do multilingual users form unique bridges connecting different clusters in the network? Depending on the fragility of the network and the position of users, removing a user may disconnect a large portion of users from the largest connected component. If multilingual Twitter users are more often than random situated in bridging positions, then removing multilingual users from the network will result in a smaller largest connected component than removing the same number of users randomly. Similarly, removing users in bridging positions should result in more disconnected components, and these components should be larger (i.e., more than isolates).

Figure 3.4 removes four different subsets of users from the network and records the size of the resulting largest, weakly-connected component, the number of components created, and the average size of these components (excluding the largest connected component). In the *Multilinguals* condition, all 103,645 multilingual users

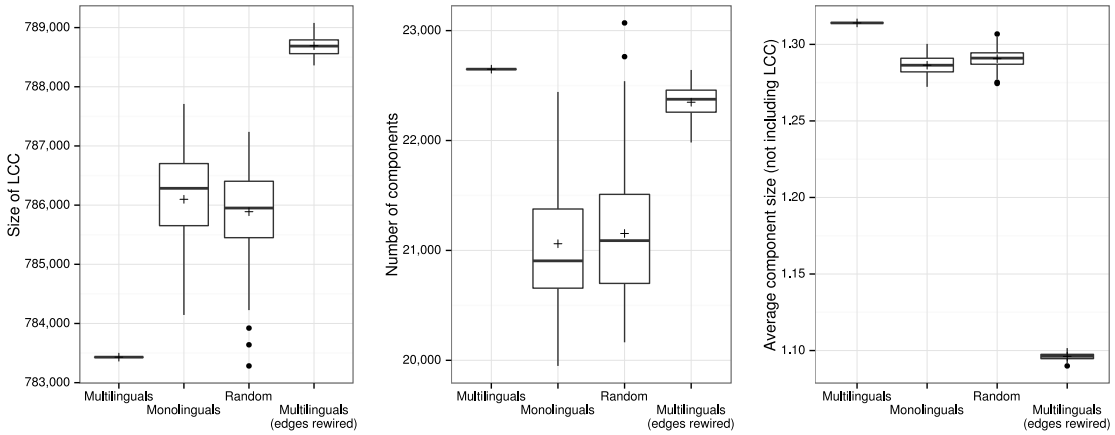


Figure 3.4. Size of the largest, weakly-connected component (left), total number of components (center), and average size of the components (right) created by removing all multilingual users, an equivalent number of monolingual users randomly, an equivalent number of all users randomly, and removing all multilingual users from a network with the same degree distribution but with edges randomly shuffled. Box plots show values from 100 realizations. Mean values are indicated with +.

are removed from the network. In the *Monolinguals* condition, an equivalent number of monolinguals are chosen randomly and removed from the network. In the *Random* condition, an equivalent number of users are chosen randomly without respect to their monolingual or multilingual classification (i.e., both monolinguals and multilinguals are potentially selected) and removed from the network.

Finally, the graphs also show a fourth condition, *Multilinguals (edges rewired)*, which provides a more stringent comparison to random. To ensure it is not just the higher degree of multilingual users responsible for a greater number of components, the statistics are calculated on a graph with the same degree distribution but with random edge wirings. This graph is formed according to the algorithm developed in Viger and Latapy (2005). The multilingual users removed from this graph have exactly the same (undirected) degree distribution as the multilingual users in the empirical network, but connect different nodes. Similarly, the remaining monolingual users have the same degree distribution as the empirical network. One-hundred realizations of the random selection of nodes are performed for all conditions, and

the distributions of the results are plotted in Figure 3.4.

All three comparisons suggest multilingual users are more often in unique bridging positions than monolingual users and than random. Removing multilingual users results in a significantly smaller largest connected component than removing the equivalent number of monolingual users or an equivalent number of users randomly. Removing multilingual users also results in a larger number of components even in comparison to the edge rewiring condition which controls for the difference in degree. Moreover, these components are on average larger than in any other condition.

This analysis confirms H3.2 and shows that multilingual users play a unique bridging role in the network and are critical to the global connectivity of the network.

3.3.3 Variations by language

There is variation in the percentage of users classified in each language as multilingual. Only 1% of users writing primarily in Japanese and 2% of users writing primarily in Korean also wrote in another language. On the other hand, over half of the users writing primarily in Tagalog/Filipino (fil) or in Italian (it) also wrote tweets in another language. These are the only two languages with over 1,000 users in the sample with a high level of multilingualism: see Figure 3.5, which shows a scatter plot of the logarithm of the number of users primarily using a language and the percentage of those users detected to also use another language. Contrary to the predictions that smaller languages would have more multilingual users, there is large variation in the percentage of multilingual users in languages with less than 1,000 primary users. Overall, there is only a weak correlation (-0.25) between the log of language size and the percentage of multilingual users.

Examining the languages of the users mentioned or retweeted reflects how much attention each user gives to other users within his/her language compared to users outside of his/her language. Averaged across all users of each language, this gives

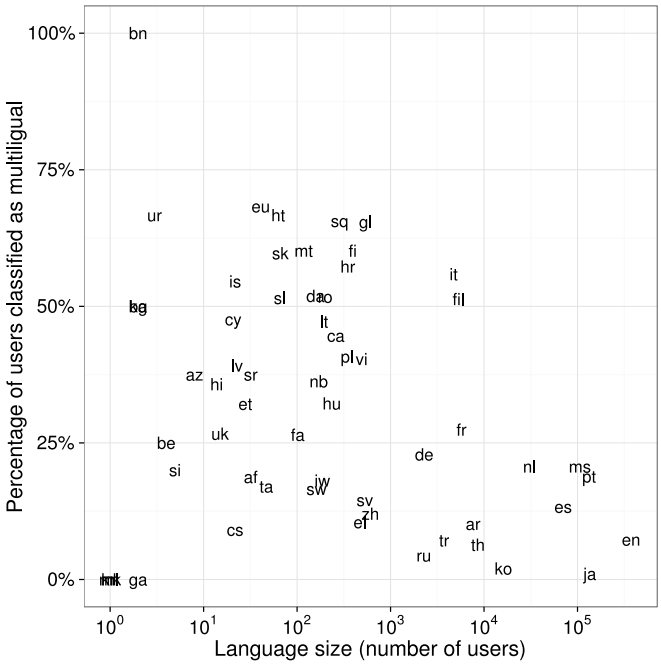


Figure 3.5. Number of users in each language compared to the percentage of these users classified as multilingual.

a measure of the collective isolation or insularity for users in each language. The expected percentage of links between users of the same language varies with the number of users writing in that language. If users mentioned other users without regard for their language, the percentage of all outgoing links from users of that language that pointed to other users in the same language would be proportional to the number of users writing in the language. For example, English represents approximately 40% of all users. If the destination of each edge originating from an English-speaking user were randomly chosen about 40% of the edges would terminate at another English-speaking user.

This analysis proceeds using the directed version of the network, but looks only at languages with at least 1,000 users identified in the language. This corresponds with a large drop in the number of users in each language between Russian, with nearly 2,300 users, and the next-most used language, Chinese, with about 600 users. As in the previous figure, languages with less than 1,000 users in the sample show great variability and are likely less representative of all Twitter users using that

Language	Language size (%)	Within language edges (%)
Japanese (ja)	14.97	99.49
Korean (ko)	1.76	98.70
Turkish (tr)	0.41	95.59
Thai (th)	0.94	94.33
Russian (ru)	0.25	93.94
Portuguese (pt)	14.56	93.10
Dutch (nl)	3.39	91.54
English (en)	40.95	90.32
Arabic (ar)	0.84	90.22
Spanish (es)	7.66	89.25
French (fr)	0.63	87.54
German (de)	0.25	85.87
Malay / Indonesian (ms)	11.59	80.48
Italian (it)	0.52	54.41
Filipino / Tagalog (fil)	0.59	21.34

Table 3.3. Language size and measure of insularity.

language.

H3.3 predicted the insularity of users would be proportional to the number of users writing in each language. Table 3.3 shows the fraction of all users represented by each language and the fraction of outgoing edges from users in each language that connect to a user of the same language. If the fraction of edges connecting users of the same language were proportional to the number of users writing in each language, the second and third columns in the table would vary similarly. Instead, a very different pattern emerges. For users in most well-represented languages, the number of outgoing edges connecting users of the same language is over 80% irrespective of the size of the language. Users in all languages are much more likely to mention/retweet users of the same language over users of different languages. Even so, users in two languages (Fillipino and Italian) do mention or retweet users writing in other languages more than users in most other languages.

3.3.4 Bridging languages

To determine which language pairs are connected more than random, it is necessary to first compute the expected number of edges from a given language to another language. The graph is collapsed so that there is one node per language. Each directed edge $i \rightarrow j$ captures the total number of mentions/retweets by users in language i of users in language j. Given the high level of insularity for each language, the analysis proceeds by only looking at edges between users of two different languages ($i \neq j$). If edge destination were simply random, each language group would receive links in proportion to the number of users writing in that language. This value is treated as the expected value and the percent error is calculated as $\frac{observed-expected}{expected}$ to compare the differences between these expected values and the observed values, which cover several orders of magnitude. The percent error is less than zero if fewer than the expected number of edges connect two languages, and it is greater than zero if more than the expected number of edges are found between two languages. Table 3.4 gives the top ten pairs of languages with more mentions/retweets than expected. There are 24 directed language pairs with no connections between them at all, and for seven language pairs there is no edge is either direction. These are indicated in Table 3.5, with mutually absent edges in *italics*. A force-directed graph with one node per language group and edges weighted according to the percent error is shown in Figure 3.6.

The network diagram and Table 3.4 show a small Asian-language cluster with Korean, Japanese, and Thai users more tightly connected than expected. They also show Italian and Filipino users more mentioned than the number of speakers would suggest. This corresponds with the higher level of multilingualism found for users writing in these two languages.

Similar to the analysis examining the bridging role of multilingual users as a whole, it is also possible to remove users from only one language at a time. In Figure 3.7, users who write in each language most frequently are removed and the

Source language	Target language	Percent error
ja	ko	1,474%
th	ko	1,091%
pt	it	1,068%
ms	fil	784%
ko	th	550%
ja	th	536%
es	it	529%
en	fil	494%
tr	de	365%
en	it	350%

Table 3.4. Top language pairs with more than the expected number of edges.

Source language	Target language(s)
ar	de, pt, *ru*
de	*ko*, ru, *th*
fil	de, ru
ja	*tr*
ko	ar, *de*, *nl*, ru
nl	*ko*, *th*
ru	*ar*, nl, th, tr
th	*de*, *nl*, *tr*
tr	*ja*, *th*

Table 3.5. Language pairs with no connections. Mutually absent pairs indicated in *italics*.

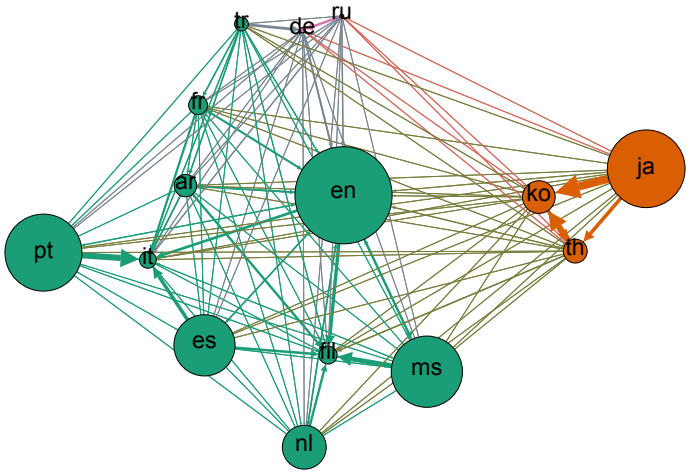


Figure 3.6. A collapsed network graph with users grouped to nodes representing the primary language used. Edges are weighted by the percent error in the expected vs. the actual number of mentions and retweets between language groups. Node size is proportional to the number of users primarily using each language, and node color is the result of a modularity-maximizing community detection algorithm.

number of components created is shown. Users are removed in order by the percentage of their tweets in the given language. That is, users who write in the language exclusively are removed first, followed by users who write in the language less frequently in decreasing order. Ties are broken randomly and the lines show the average of 100 realizations. For each curve, users writing in a different language more frequently are left untouched. Removing users from most languages did not create more components than removing users at random.

Most curves follow a general pattern. The curves rise and reach a maximum beyond which they begin to fall. The fall in the curves corresponds to the elimination of components made entirely of speakers of the language being removed. Each curve then stabilizes on a value for the number of components created by removing all users from the given language. (The graph omits the fall of the English curve, which falls and then plateaus at approximately 11,300 components.)

Three interesting exceptions to this general pattern emerge. The lines corresponding to the removal of Japanese, Thai, and Korean users drop almost immediately reflecting a high level of insularity and few internal divisions within each

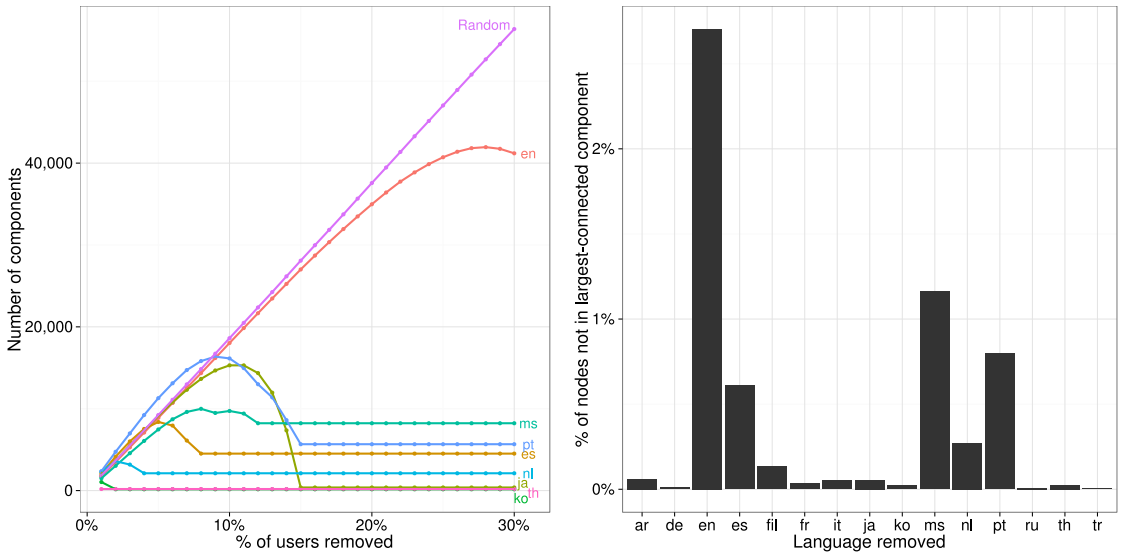


Figure 3.7. Number of components (top) and the percentage of remaining users not in the largest connected component (bottom) created by removing users in different languages.

language. The line corresponding to Malay users has almost no decline before plateauing. This suggests that the components created by the removal of Malay users almost always contain at least one user of another language. Finally, the line corresponding to the removal of Portuguese users initially produces more components than removing an equivalent number of users randomly. Like other languages, most of these additional components are made entirely of Portuguese users and the number of components ultimately decreases as further users are removed. The initial rise beyond the random curve suggests there are a number of subdivisions within Portuguese users themselves (possibly a Brazil–Portugal split, but this is not specifically tested).

It is clear that English serves the largest bridging role in an absolute sense, but it is also the language with the largest number of users. To calculate the relative extent of bridges created by speakers in a given language, all users from a given language are removed from the graph. Then the percentage of remaining other-language speakers not in the largest connected component is calculated. If speakers of a language do not collectively serve much of a unique bridging role, then almost no nodes will be

separated from the largest connected component. Conversely, removing users from a language that collectively serves a large bridging role will leave a larger percentage of users separated from the largest connected component. The bottom graph of Figure 3.7 shows these percentages after removing all users from a given language. Although all the values are relatively small, removing English-language users also leaves a larger percentage of users disconnected from the largest connected component than removing users from any other language. Users writing in English, therefore, play more of a bridging role than users writing in other languages, confirming H3.4.

Removing Portuguese users leaves a relatively high percentage of users disconnected, but this is somewhat expected given the large number of users writing in Portuguese. This is contrasted by the case of Japanese, however, which again reveals the insularity of Japanese speakers in the data. Despite the high number of Japanese speakers in the sample (second in size only to English) completely removing all of these Japanese users has minimal effect on the connectivity of the network.

Overall, removing all the users from any one language never disconnects as many nodes as removing all multilingual users. English users, for instance, represent three and a half times as many users as the total number of multilingual users from all languages; yet, removing this smaller number of multilingual users disconnects more nodes from the largest connected component than removing all English users (3.7% vs. 2.7%). Thus, no one particular language (not even English) appears to create as many bridges as multilinguals from multiple languages. It is the communication dynamics of multilinguals from many languages that create the network bridges found earlier.

3.4 Discussion

The analysis of this sample reveals the important role played by multilingual users of Twitter. The network of retweets and mentions is heavily structured by language

with most users retweeting and mentioning only users authoring content in the same language as themselves (H3.1). Nonetheless, users authoring content in multiple languages play a unique bridging role that is not duplicated by other users in the network (H3.2). This suggests that language is likely a useful feature to be used in search and friend recommendation algorithms. Even so, a balance must be struck as there are clusters of users spanning different languages. In addition, over 10% of the users in the sample wrote in multiple languages. Any use of language, therefore, should allow for each user to have a set of multiple preferred languages and not restrict the user to one. This is especially important as multilingual users are more active than their monolingual counterparts authoring more tweets and replying/ mentioning more users.

In using language for search results, it would be useful to infer a set of possible secondary languages to draw results from when there are insufficient results in the primary languages. Although there are some clear geo-linguistic patterns in the language–language network (e.g., an Asian language grouping), the level of overall multilingualism for users in each language does not vary straightforwardly with the number of users in each language as predicted by H3.3. Users from less-represented languages may discuss and link to information originally from other-language sources outside of Twitter, but simply do so in isolation from users of other languages on the platform. Equally possible, these users may use a non-native language more frequently than their native tongue and thus be classified by the methods used in this chapter as primary users of a different, larger language. Further research will need to look at the content and links authored by users in less-represented languages and the distribution of languages used by these users to better separate these possibilities.

Users writing most frequently in the top two represented languages, English and Japanese, play vastly different roles in the global connectivity of the network. Users in most languages direct most of their tweets to other users writing in the same lan-

guage. English users are no exception, but still direct about 10% of their mentions/ retweets to users in other languages. Japanese users, on the other hand, mention and retweet users from other languages only about 0.5% of the time. Of the mentions/ retweets Japanese users do direct to other-language users, a disproportional number go to Korean users likely due to the geographic proximity of Japan and Korea. This indicates the importance given to same language results should be weighted differently across language groups and probably even across users primarily using the same language.

It is the unique combination of multilinguals across many language pairs that results in the global connectivity of the network. Removing speakers from one language at a time never resulted in more components than removing an equivalent number of users randomly from the network. Nonetheless, in comparison to other languages, English speaking users collectively do play a much larger role in bridging speakers of different languages than do speakers of any other language. This confirms H3.4, which predicted this role for English given its global prevalence and the number of second-language speakers.

While this study shows multilinguals occupy a unique place in the global retweet and mentions network, it does not examine whether foreign language content is propagated further by the monolingual followers of a multilingual user. Further research will be needed to determine whether content sent by multilingual users is further propagated by their followers in different languages. Although the Twitter platform does not offer the ability to selectively filter friends' tweets by language, users may mentally filter these tweets and stop their propagation by not engaging with them. Future research should examine specific diffusion paths (e.g., diffusion cascades as in Bakshy, et al. (2011)) with reference to the languages of the users involved and the content of the tweets themselves.

Although much future work remains, this study is the first to examine the global connectivity of the Twitter retweet and mentions network with reference to lan-

guage, and to establish the unique role that multilingual users play in this connectivity. While the amount of cross-language bridging is small, its presence indicates there is value in one large, multilingual system over several separate, non-connected systems for each language. Experimental work could shed light on design features that could better enable multilingual users to bridge users of different languages in the network and further enhance the unique value of the single, multilingual system. In addition, systems similar to Omnipedia for viewing content across different Wikipedia language editions (Bao et al., 2012) could better enable monolingual users to directly discover interesting content in other languages on Twitter.

Twitter, of course, is only one of a plethora of different user-generated content platforms. In order to understand how the findings of this chapter in regards to Twitter may apply to other platforms, the next chapter examines multilingual user behavior on a very different user-generated content platform: Wikipedia.

Chapter 4

Multilinguals and Wikipedia

Language and knowledge are indissolubly connected; they are interdependent.

—Annie Sullivan, tutor of Helen Keller

4.1 Introduction

Wikipedia, the free, peer-produced online encyclopedia, contains a large collection of human knowledge. The foundation behind Wikipedia has characterized the encyclopedia as trying to provide access to "the sum of all human knowledge."[1] If any one language edition of Wikipedia were to achieve the goal of "all human knowledge," then that language should contain (at a minimum) all the information found in other language editions of the encyclopedia. Studies comparing content across language editions, however, have found a "surprisingly small amount of content overlap between languages of Wikipedia" (Hecht & Gergle, 2010b, p. 295). No one edition contains all the information found in other language editions, and the largest language edition, English, contains only 51% of the articles in the second-largest edition, German (Hecht & Gergle, 2010b, p. 295). Nonetheless, there clearly is some

This chapter is adapted from the following publication:
Hale, S. A. (2014). Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Annual Conference on Web Science*, WebSci '14. New York: ACM. doi:10.1145/2615569.2615684.

[1] `http://www.theatlantic.com/technology/archive/2011/05/is-wikipedia-a-world-cultural-repository/239274/`

overlap in content between languages, and a greater sharing of information between the language editions would enable monolingual readers of the encyclopedia to access a larger variety of content.

This chapter examines one month of all edits to the top 46 language editions of Wikipedia. This comprises all editions with at least 100,000 articles at the time of data collection in July 2013.[2] It identifies users who contribute to multiple language editions (these users are referred to as multilingual users in this chapter) and compares their contributions to that of users who edit only one language edition of the encyclopedia (monolingual users). It asks if multilingual editors play a unique bridging role diffusing information between different language editions akin to the bridging role multilingual Twitter users were found to play in the previous chapter.

4.2 Related work

Previous research about Wikipedia has tended to focus on the English-language edition. These studies have found that the scientific articles in the English edition compared favorably with Encyclopedia Britannica (Giles, 2005). However, studies have also suggested the edition suffers from issues of coverage and bias (Halavais & Lackaff, 2008; Hecht & Gergle, 2009; Holloway, Bozicevic, & Börner, 2007).

Among these biases, multilingual studies of Wikipedia have shown that there is low overlap in the articles present in multiple language editions of Wikipedia and that each language edition exhibits a self-focus bias (Hecht & Gergle, 2009). This bias manifests itself in the articles users choose to write (and not write) such that articles about places, people, and events where the language of an edition is spoken are more prominent than those in other regions (Hecht & Gergle, 2009). Self-focus bias is also present within the content of individual articles where, for example, the article on psychology in the Spanish-language edition has a section about contributions to the field from Latin America that other language editions do

[2] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

not (Hecht & Gergle, 2010b).

No study to date has used Wikipedia log data to determine what percentage of users contribute to multiple language editions of the encyclopedia. Language, however, is a large factor in the network structure of communication patterns on many platforms including telephone communications (Barnett & Choi, 1995), Twitter messaging (Chapter 3), and blog linking (Appendix B; Herring et al., 2007). Studies of other platforms and the low overlap in content between different language editions of Wikipedia suggest that *most editors will edit only one language edition* (H4.1).

In contrast to this hypothesis, a 2011 non-representative, survey of Wikipedia editors found that just over "half of Wikipedia editors contribute to more than one language Wikipedia, and an overwhelming majority (72%) read Wikipedia in more than one language" (N=4,930).[3] In addition, Yasseri et al. (2012) found registered users from many different timezones contribute to many language editions of Wikipedia. For example, 25% of edits to the Arabic and Persian editions likely came from users in North America. This suggests diaspora, language learners, or other speakers of these languages play an important role in editing the encyclopedia. Furthermore, the location of these users in North America suggests many of them might speak another language in addition to Arabic or Persian. If so, these users could introduce new information from other language sources and reduce the amount of self-focus in the edition.

Self-focus results were not reported for Arabic or Persian by Hecht and Gergle (2009), but the Dutch and Swedish editions were found to exhibit less self-focus than most of the other editions studied.[4] The authors speculated that high bilingualism with English in Dutch and Swedish societies could explain why the Swedish

[3] `https://meta.wikimedia.org/w/index.php?title=Editor_Survey_2011/Location_%26_Language&oldid=8409990`

[4] Portuguese was the only edition with a lower self-focus ratio, which the authors suggest might be a peculiarity with the contributions to the Portuguese edition or the result of a bot. English, Japanese, and German exhibited the greatest amount of self-focus among the editions studied.

and Dutch editions exhibited less self-focus in their study. They write that users contributing to the Dutch and Swedish editions, "may have gained significantly more guidance from the English Wikipedia, muting their spatial self-focus effect" (Hecht & Gergle, 2009, p. 17). This idea, however, is not specifically tested in their paper.

The literature therefore suggests that multilingual users who edit multiple language editions of Wikipedia could play a unique role in diffusing content between different language editions. From seemingly small changes like updated population numbers or new website addresses to large, fast-breaking news developments (e.g., the Japanese tsunami and earthquake discussed by Hale (2012)), multilingual users may help keep content in sync and reduce self-focus bias by introducing new content, updating old content, and correcting errors across multiple language editions. This chapter examines this idea in two ways. First, the articles edited by multilingual users are compared to the articles edited by monolingual users. It is expected that *multilingual users will edit different articles than monolingual users* (H4.2). Second, this chapter compares the articles edited across language editions by the same user to the network of interlanguage links that link articles on the same concept across language editions. If multilingual Wikipedia users serve as information bridges contributing similar information across multiple editions, then it is expected that *when a user edits an article in another language that same user will usually also edit the corresponding article in his native language* (H4.3).

The idea of network effects from network studies or positive externalities from economics may explain in part the reason editors of Wikipedia would contribute to a foreign language edition of the encyclopedia. Network effects suggest that larger-sized platforms or networks have more communicative value than similar, smaller networks. This is obvious in the trivial observation that if only one person in the world had a telephone it would be utterly useless to that person as he would have no one to call. More generally, a social media platform, like a telephone, is only valuable if one's social contacts also use the platform. For without this, a user

would have no one with whom to communicate. With each additional social media user, the value of the network grows for the existing users because each person now has a wider array of individuals who they are able to contact through the network. Crystal (2003) relates this network effect to languages arguing that the more individuals who use a common language, the more valuable it is for additional individuals to also learn that language. He speculates this effect might account in part for the growth and staying power of English as a global language. This idea is also suggested by Zuckerman (2013). Similarly, editions of Wikipedia written in more widely spoken languages have the possibility of reaching larger audiences, and past research has suggested an important factor motivating content production is the extent to which authors believe there is an audience to engage with the content (Zuckerman, 2008).

These ideas of network effects related to language size suggest two related hypotheses. The users who cross-language boundaries will *come from* smaller, less-represented languages and will *cross to* larger, more-represented languages. More specifically, *users writing primarily in smaller-sized language editions will be more likely to cross-language boundaries than users writing primarily in larger-sized language editions* (H4.4). When these users cross languages, they will most likely cross to a larger-sized language edition (e.g., English, German, French). As a consequence of this, *larger-sized language editions, English chief among them, will be more likely to have contributions from editors of different languages than smaller-sized language editions* (H4.5).

4.3 Data

Edits to Wikipedia are broadcast in near real-time over Internet Relay Chat (IRC).[5] Each edit to any Wikipedia edition is broadcast on the irc.wikimedia.org server on an IRC channel with a name in the format of #lang.wikipedia (e.g., #en.wikipedia

[5] `http://meta.wikimedia.org/wiki/IRC/Channels#Raw_feeds`

for the English edition of Wikipedia, #de.wikipedia for the German edition, etc.). Each entry contains the username (or IP address for anonymous users), the title of the article edited, comments written by the user about the edit, the size of the edit (how many bytes larger or smaller the result of the edit is compared to the previous version), and a link to the differences from the previous version. The date and time of the edit is not included, but this information was added by consulting the system clock. Similarly, the IRC channel of the message was recorded to know which language edition the user edited.[6]

All edits for the 46 language editions with 100,000 or more articles[7] were recorded through IRC from July 8, 2013, to August 9, 2013. Edits to the Simple English edition are excluded for most of the analysis and the role of the Simple English edition is addressed separately in Section 4.4.4. In addition to the main, article namespace, Wikipedia has separate namespaces for other content including user pages, portals, and administrative activity. This paper focuses on the main namespace to which the majority of the edits (63%) were directed. Consistent with prior research, many of these edits (15% of non-minor edits) were created by bots—automated scripts editing the encyclopedia for consistency, fixing common mistakes, and detecting and reverting vandalism (malicious edits). A number of edits were also from anonymous users without an account (28% of non-minor edits). Since IP addresses change over time and multiple users may edit from the same IP address, these edits were removed from the dataset. In order to focus on the activity of human editors, only non-minor edits from registered users, who were not listed as bots were considered for further analysis.

Initial analysis of the data suggested that there were many bots operating on the encyclopedia without being officially declared as bots. These suspected bots had very high edit counts across a large number of languages, and human examination

[6]The code used to record the IRC streams (Java), construct the network (Java/Hadoop), and perform the analysis (Python/R) are available at `http://www.scotthale.net/pubs/?websci2014`.

[7]`http://meta.wikimedia.org/wiki/List_of_Wikipedias`

of their contributions and user pages suggested most were indeed bots. A number of ideas drawn from the literature were examined and ad hoc subsets of users were manually inspected to arrive at a method to filter these unregistered bots. The most successful approach found was to examine the maximum amount of time between two successive edits from the same user. In accordance with past research, edits for most users (registered bots excluded) were bursty: that is, the edits were clustered such that many edits occurred in small amounts of time separated by comparatively longer absences of edits (Geiger & Halfaker, 2013; Yasseri, Sumi, Rung, et al., 2012). Looking at the length of the longest break between bursts of edits revealed that many users without the bot flag set never had a rest of more than a couple of hours over the entire 32-day data collection window. As most human editors would need to break longer than this for sleep—and editing activity has previously been shown to follow circadian cycles (Yasseri, Sumi, & Kertész, 2012)—these users are likely undeclared bots.

Through manual examination of different thresholds, six hours was chosen. Overall, 114,376 accounts did not have any break in editing of more than six hours over the course of the 32 days in the sample. These users were assumed to be bots (or humans with only one editing burst) and excluded from further analysis.

One edit is insufficient to determine whether a user edits in multiple languages, while with two edits in two different languages it is unclear which language is the user's primary language. To be certain multilingual users were identified as such and to be able to identify users' primary languages, all users with less than four edits overall (21%) or less than two edits in their most-edited language (0.6%) were excluded. As a result, this study focuses on the most active users. This is not, however, a major limitation, as past work has shown the most active users produce a disproportionate majority of the content in the encyclopedia (Ortega, Gonzalez-Barahona, & Robles, 2008).

4.3.1 Cross-language alignment

Previous cross-language studies on Wikipedia relied on the interlanguage links found in each edition of the encyclopedia. These links were maintained by a mixture of humans and machines (bots), but nonetheless contained a number of errors (Hecht & Gergle, 2009). The issues were often compounded by having dumps of each language from slightly different dates.

This study uses a new source of inter-language information, WikiData.[8] This new initiative centralizes all interwiki references and category information (and, in the future, statistics and other structured data) in one location. This avoids some previous issues with out-of-date or conflicting interlanguage links. Further study of the impact of the WikiData project on Wikipedia and its editors is not within the scope of this chapter, but would be a fruitful area for future research.

When Wikipedia began, each language edition was run independently. User accounts were created separately on each language edition, and thus the same username on different editions may refer to two different persons. As Wikipedia matured, a central authorization system was built to provide for unified login. Unified login allows users to unite their accounts across multiple language editions (and other projects: Wikitionary, Wikiquote, Wikibooks, etc.) and be able to login to all projects and editions at one time. Users who have unified their accounts have "global accounts" and information about the user is available from the *Global account manager*.[9]

There was an announcement in April 2013 that any remaining conflicts where different persons had accounts with the same usernames on different editions would be resolved and the accounts renamed.[10] This was to take place in May 2013, but was delayed first to August 2013 and then to an unspecified future date. Once this step is taken it will be trivially easy to determine if one user edits multiple language

[8] `http://www.wikidata.org/`
[9] `http://meta.wikimedia.org/w/index.php?title=Special:CentralAuth`
[10] `http://meta.wikimedia.org/wiki/Single_User_Login_finalisation_announcement`

editions. At the point of data collection, however, it still remained technically possible for one person to have two differently named accounts on different editions, or for two persons to have accounts of the same name on different editions.

The publicly-available data makes it difficult to identify one person with multiple accounts (false-negative monolingual). It is possible, however, to check whether a given account is a global account. If it is a global account, it is possible to get a list of all the language editions on which the user is active. This makes it possible to avoid any false-positive multilingual user classifications.

For this study, all usernames were first assumed to be unique across the editions. The usernames editing multiple language editions were identified and classified as possible multilingual users. Each of these usernames was checked against the *Global account manager* to ensure that the user was a global user registered with all the language editions the user was recorded as having edited during the data collection period. Users who were not registered as global users or whose global username was not associated with all the editions the user was recorded as having edited were treated as separate users.

There were very few false-positive matches found. Only 572 usernames found to be editing multiple editions were not global accounts. In a further 50 cases, a global username existed but was not associated with all language editions where the username was used. These local, non-global users were treated as separate users. The small number of these matches means that this correction has minimal effect on the results.

The available data does not easily allow the discovery of false-negatives—one user with different usernames on different editions. In addition, it is not possible to know if a user reads multiple language editions, while editing only one edition. Therefore, the number of multilingual users presented in this chapter is a lower bound on the actual amount of multilingual activity happening on Wikipedia.

Language	Edits	Articles	Users	NP users	NP edits
English	1,389,647	518,405	27,476	18%	3%
German	256,495	125,647	5,967	18%	2%
French	250,828	106,027	4,549	25%	3%
Spanish	191,934	66,848	4,338	24%	3%
Russian	239,267	92,326	3,961	16%	1%
Japanese	106,848	56,406	3,551	11%	2%
Italian	160,191	69,534	2,919	25%	2%
Chinese	112,888	42,937	2,309	14%	1%
Portuguese	67,505	32,753	1,730	29%	4%
Dutch	80,535	39,463	1,500	33%	3%
Polish	67,038	37,393	1,454	30%	3%
Swedish	42,390	25,269	904	43%	4%
Ukrainian	54,241	22,537	898	36%	3%
Hebrew	37,889	13,224	832	16%	2%
Arabic	43,924	15,993	729	20%	3%

Table 4.1. Statistics for the top 15 language editions in the sample. The *Users* column includes all users who edited the edition during the data collection period. A percentage of these users (*NP users*) are non-primary users who edited a different language edition more frequently. *Edits* and and *NP Edits* are defined similarly.

4.4 Analysis

Excluding the Simple English edition, 55,568 registered, human users edited at least one edition of Wikipedia two times or more and had at least four edits across all editions during the 32-day data collection period. This resulted in a total of 3,518,955 edits. Most of these edits (39%) were to the English language edition. Similarly, most users (40%) edited the English-language edition of the encyclopedia more than any other edition (Table 4.1).

Consistent with H4.1, a relatively small number of users (8,544 or 15.4%) edited multiple editions of the encyclopedia. These users were categorized as *multilingual* while all remaining users were classified as *monolingual*. Multilingual users were significantly more active than their monolingual counterparts. Multilingual users made a mean 124 (median 32, sd 299) edits overall, while monolingual users made only a

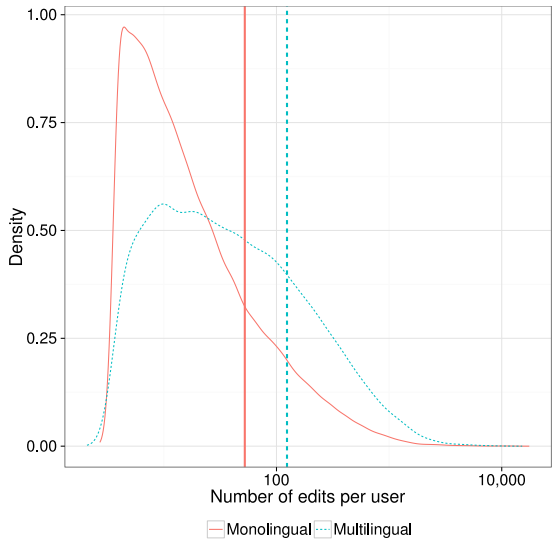


Figure 4.1. Density plot comparing the number of edits made by monolingual and multilingual Wikipedia users.

mean 52 (median 13, sd 192) edits overall (Figure 4.1). These additional edits by multilinguals are not only in other language editions but also in each user's primary language edition that the user edited most frequently. Multilinguals made a mean 113 (median 26, sd 285) edits to their primary language editions of the encyclopedia. Indeed, while only 15.4% of all users, multilingual users were responsible for 30.1% of all edits captured during the month.

Multilingual users were not just editing the same articles more, but also edited a wider number of articles. Multilingual users edited a mean 69 (median 16, sd 191) articles while monolinguals edited a mean 27 (median 5, sd 133) articles. Logically following from the fact that multilingual users were more active in their primary languages than monolingual users, it is clear multilinguals were not more active simply because they had more articles across more languages they could have edited. As discussed in the next subsection, multilinguals only directed a small percentage of their edits to their non-primary languages. Multilingual users were still more active than their monolingual counterparts after collapsing together articles in different languages on the same concept as determined by interlanguage article links. In this

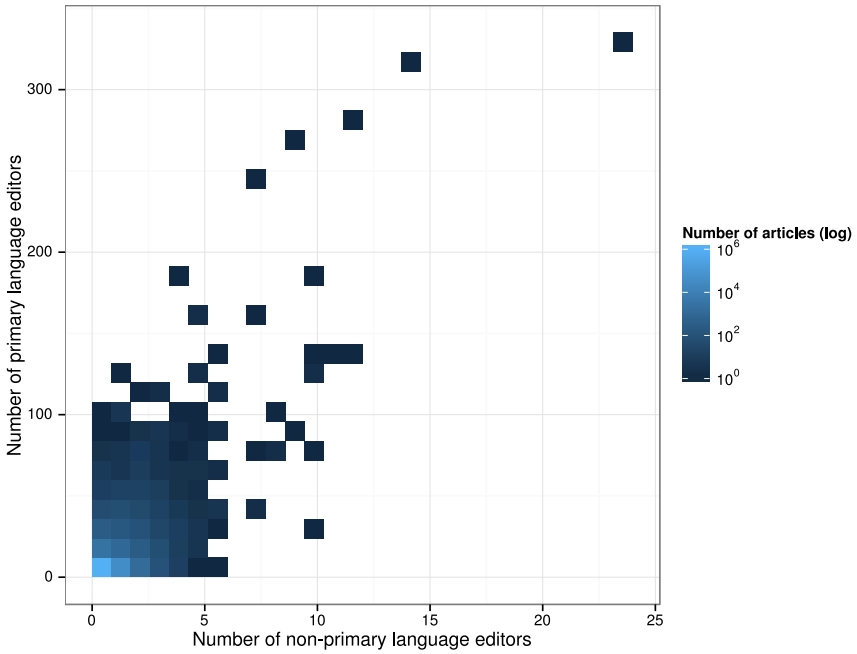


Figure 4.2. 2D density plot of the number of multilingual users editing articles in a non-primary language against the number of monolingual users editing the articles.

way, for example, editing United States (English) and Estados Unidos (Spanish) only counted as editing one "concept" since the two articles are linked together by interlanguage article links. Monolingual users edited the same number of concepts as articles since they only edited one language, while multilingual users edited a mean 65 (median 15, sd 185) concepts. All of these differences are significant as established with two-tailed t-tests ($p < 2.2 \times 10^{-16}$).

In contrast, the size of the edits made by multilinguals and monolinguals do not differ significantly. Edits by multilinguals had a mean size of 331 bytes (median 143, sd 912), while edits by monolinguals had a mean size of 339 bytes (median 125, sd 1327).

4.4.1 What do multilinguals edit?

Given the low overlap in content between language editions of Wikipedia, multilingual users may offer unique contributions to the editions they edit. This section

examines the edits of multilingual users to their non-primary language editions (that is, to editions other than the editions they edited most frequently).

Edits from multilingual users writing in their non-primary language are an extremely small fraction of all edits to Wikipedia. Only 2.6% of edits are from users writing in their non-primary languages. To some extent, multilingual users edit similar articles in their non-primary languages as do monolingual users. The 2D density plot in Figure 4.2 shows that the articles with the largest number of non-primary users also have a large number of primary users. There's a positive correlation of 0.25 between the number of multilingual users editing an article in a second language and the number of monolingual users editing the article. The most dense region is near the origin where most articles are edited by a small number of users. Within this region, however, multilingual users are often editing articles not edited by other users: 44% of the articles edited by multilingual users in their non-primary languages were not edited by any monolingual user during the data collection period.

Using the WikiData information for interlanguage article links, it is possible to connect articles across languages (e.g., the English-language *United Nations* article is on the same concept as the Spanish-language *Organización de las Naciones Unidas* article). This makes it possible to check when a multilingual user edited an article in a non-primary language, if that user also had edited the equivalent article in his or her primary language. Overall, 44.5% of the edits to non-primary languages by multilingual users were to articles where the user had edited the same article in his or her primary language. The underlying distribution per user, however, is bimodal at the two extremes (Figure 4.3). 43% of multilingual users did not edit the equivalent articles in their primary languages.[11] On the other hand, 25% of multilingual users always edited the equivalent articles in their primary languages.

Part of this behavior is explained by the fact that some of the articles edited

[11] My final empirical chapter looks at all edits to articles from the creation of each article onward and finds similar results, which suggests that this finding is not a result of unobserved edits outside of the data collection period.

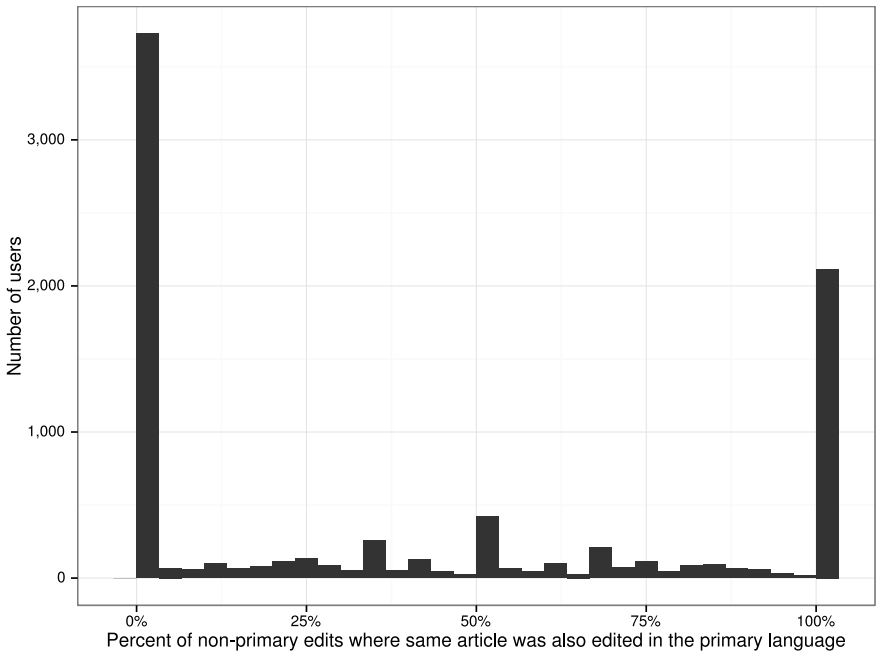


Figure 4.3. Histogram showing the distribution with which multilingual users edited articles in other languages that they also edited in their primary languages. The distribution is bimodal. A large number of users did not edit any of the same articles in their primary languages, but a large number of users always edited the same articles in their primary languages.

by multilingual users in their non-primary languages did not exist in their primary languages. Overall, 73% of the articles edited by multilingual users in a non-primary language existed in the primary languages of those users. Ignoring the instances where an equivalent article did not exist in the users' primary languages, 59.8% of edits to non-primary languages by multilingual users were to articles the user had edited in his or her primary language. The distribution remains bimodal with 34% of users not editing the equivalent articles in their primary languages, and 37% of users always editing the equivalent articles in their primary languages.

While the size of edits made by multilingual and monolingual users did not differ significantly, the size of edits multilingual users made in their non-primary languages were significantly smaller than the edits they made in their primary languages. Considering only edits with a positive size (i.e., not edits that removed more text than was added) multilingual users made edits with a mean size of 569 bytes (median 260, sd 1327) in their primary languages and a mean size of 468 bytes (median 83, sd 2156) in their non-primary languages. Nonetheless, 25% of multilingual users actually made larger positive-sized edits in their non-primary languages as compared to positive-sized edits in their primary languages.

Comparisons of edit sizes across languages is difficult for two reasons: first, different characters require a different number of bytes to store, and second, the information content contained in one character differs across languages. One standard English character is usually one byte, while a special or accented character (e.g., á) is usually two bytes, and a character from a more complex language like Japanese, Chinese, or Korean is generally three bytes. In contrast, however, one English character usually contains less information content than one Japanese or Chinese character, which could represent a full word. An information-theoretic approach, using entropy, has previously been employed to compare the information content per character across different languages on Twitter (Neubig & Duh, 2013), and a similar approach could be employed to compare Wikipedia edit sizes across

languages. Such an approach, however, would require the content of the edits rather than the meta-data about edits used here.

Overall, these findings support H4.2 that multilingual users would make unique contributions to the encyclopedia by editing articles less edited by monolingual users. The data is mixed for H4.3 which suggested multilingual users would often edit the same article in their primary and non-primary languages. For a quarter of users this was always true. However, just over two-fifths of multilingual users did not edit the equivalent articles in their primary languages. Data on the articles users view is not available to know whether these users viewed the equivalent articles in their primary languages before editing in another language.

4.4.2 Variations by language

The percentage of users classified as multilingual varied across the language editions studied. Previous research suggested this variation would correlate with the total number of users and/or the number of articles in each edition. Figure 4.4 shows the percentage of users primarily editing each language edition that also edited a second language edition compared with the total number of users primarily editing the language edition. Consistent with the suggestions of prior research, there is a strong correlation between the two variables. Looking only at languages with at least 10 users in the sample to avoid small number issues, the log of the number of users primarily editing each edition and the percentage of users editing multiple editions are correlated with a coefficient of -0.69. Similar results hold for comparing the percentage of users who are multilingual to the number of articles in each language edition, where the correlation coefficient is -0.46. (These two measures are interdependent as the total number of articles per edition and the number of users in the sample per edition have a correlation coefficient of 0.90.)

Among the smallest-sized editions, Esperanto (eo) and Malay (ms) stand out as two languages with high levels of multilingualism among their primary editors. It

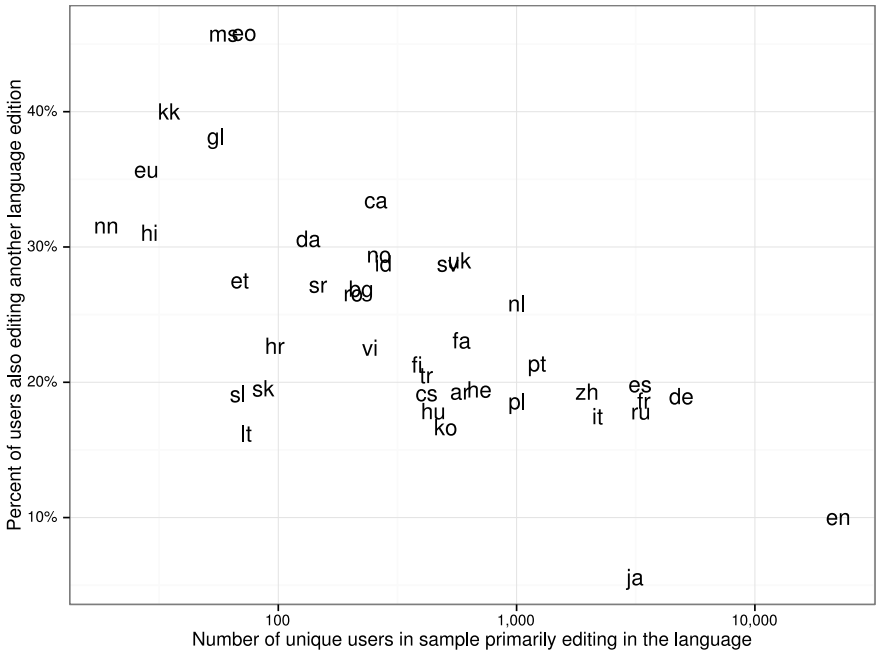


Figure 4.4. Scatter plot of language size (number of unique users) and percentage of users who are multilingual (edit more than one language edition). The three editions with less than 10 users in the sample are omitted (Uzbek, Cebuano, and Waray-Waray).

is surprising that Esperanto was not higher given that it is a constructed language and thus has no native speakers. Nonetheless, nearly 46% of the editors of the Esperanto edition edited that edition more than any other. Italian (it), Slovenian (sl), and Slovak (sk) are similarly sized but with far lower levels of multilingualism.

Among larger-sized languages, Catalan (ca), Swedish (sv), Ukrainian (uk), and Dutch (nl) all had relatively high levels of multilingualism. In contrast, the lowest level of multilingualism is found among users primarily editing the Japanese (ja) language edition, where only 6% of the users edited another edition.

While some exceptions emerge, the findings support H4.4: in general a larger percentage of the users primarily editing smaller-sized editions are multilingual.

4.4.3 Language crossings

Hypothesis H4.5 predicted that when users did edit a second edition, that edition would almost always be English or, to a lesser degree, another large edition. In order of the number of users active in the last 30 days, the largest editions of Wikipedia are English (129,900 active users), German (20,300), Spanish (15,800), French (15,500), Japanese (11,400), and Russian (10,800).[12]

The bipartite network of users and articles was collapsed to a network of language relationships. Each article was assigned to the node corresponding to the language edition to which it belonged. Similarly, users were grouped with the node representing the language they edited most frequently. Each directed, weighted edge e_{ij} records the log of the number of editors primarily editing language edition i that also edited language edition j. This network is shown in Figure 4.5a.

The number of users represented by each edge ranged from 1 to 775 with a mean value of 15.3 and a standard deviation of 50.7. Only edges with a log value at least 1.96 standard deviations above the mean of all log values are shown on the graph. This corresponds to edges with 60 or more users. Isolates (languages unconnected to any other language) are removed from the network diagram. Note that in contrast to the previous section, the network graph shows the logarithm of the number of users editing multiple editions and not the percentages of users editing each edition that also edit another edition.

The network reveals the English edition (en) does receive a large amount of attention from multilingual users in other languages. Every node in the graph is connected to English. Most of these edges are reciprocal, but in three-quarters of the cases more users from another language edited English than users from English edited the other language. Despite the very large size of English, this also holds globally as 4,659 users from other languages edited English while only a total of 3,673

[12]The latest information on the number of users and articles for all editions of Wikipedia is online at `http://meta.wikimedia.org/wiki/List_of_Wikipedias`. Accessed on 19 September 2013.

primary users of English edited another language. When users primarily editing the English edition did edit another language edition, the largest number of users edited the Spanish (es), German (de), and French (fr) editions.

There are only four languages that have a directed edge from English that is not reciprocated. These languages are Romanian (ro), Danish (da), Bulgarian (bg), and Catalan (ca). Each of these four languages is quite small, and while a sizable percentage of the users primarily editing these language editions also edited English[13] they simply did not constitute sufficient volume to rise above the edge weight threshold and appear on the graph.

There are are some strongly connected language pairs not involving English. German users edit the French edition, and Russian and Ukrainian users edit each others' editions with some frequency. Figure 4.5b shows the same network, but with English removed. The edge weight distribution is recalculated and edges with 33 or more users are shown (corresponding to 1.96 standard deviations above the mean of the log values with English removed).

Even with the English edition removed, editions with a larger number of active users continue to structure the network. The second-largest edition, German, is connected to every node except Ukrainian (uk), Japanese (ja), and Chinese (zh).

The infomap community detection algorithm (Rosvall, Axelsson, & Bergstrom, 2009; Rosvall & Bergstrom, 2008) finds the same community structure with and without English as shown with node color in Figure 4.5. The largest community is centered around the largest language, English or German. A strong relationship is present between Ukrainian and Russian (ru) where Ukrainian users edit Russian and English but rarely another language edition. Similarly, Chinese users edit Japanese and English but rarely any other edition. Unlike the Russian/Ukrainian relationship, the edge from Chinese to Japanese is one way. Indeed, apart from editing English,

[13] 27–33% of the users in each of these four languages edited multiple editions. Most of these multilingual users edited English in the case of Romanian (95%), Danish (88%), and Bulgarian (88%). Of the multilingual users that primarily edited Catalan, 50% also edited English.

Japanese users rarely edit any other language.

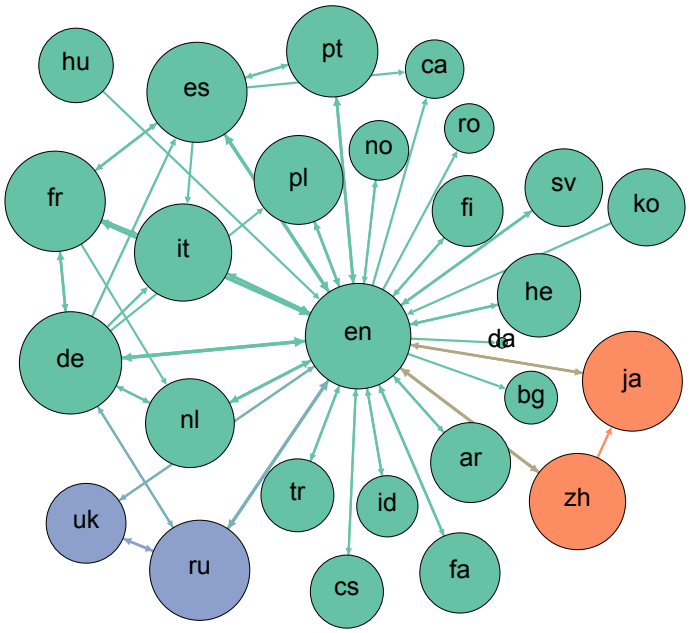
With English removed, it is also worth noting that all the romance languages (Spanish [es], Italian [it], French [fr], and Portuguese [pt]) have mutual edges between them. The only exception is Catalan (ca), which is only connected to Spanish and German. Nonetheless, the many links to German from other language editions overshadow these connections in the community detection algorithm.

These findings support hypothesis H4.5 that multilingual users from smaller languages would mostly cross language boundaries to edit larger-sized languages. English receives edits from users in almost every other edition. Even with English removed, the second-largest edition, German, receives edits from users in a large number of other editions. Nonetheless, regional and linguistic patterns are also evident in the co-editing network.

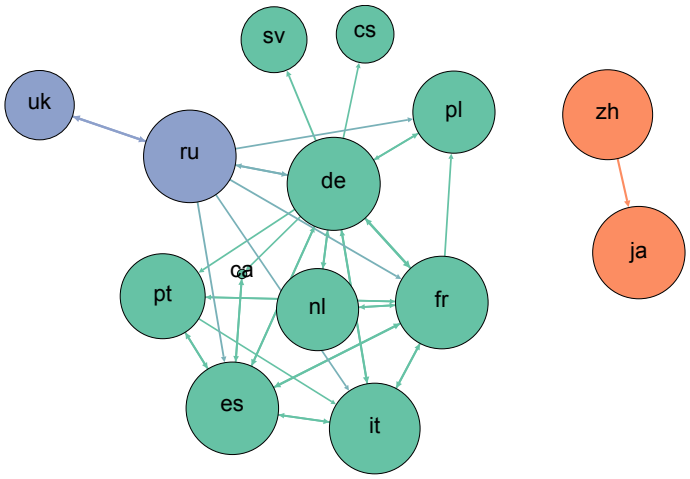
4.4.4 The role of Simple English

The Simple English edition of Wikipedia (hereafter Simple) is written in English, but aims to use simpler grammar and shorter sentences. While intended to be primarily read by children, adults with learning difficulties, and second-language learners of English, it may also be edited by many speakers of other languages (Yasseri, Sumi, & Kertész, 2012).

The findings presented so far in this chapter have excluded all edits to Simple. Its inclusion, however, makes little difference to the findings. There is a strong link between Simple and the English edition. Of the 221 users in the data sample who edited Simple, over half (124 users) primarily edited English. Users editing Simple and primarily editing a non-English edition were spread thinly over 26 other languages. The four largest of these languages were German (13 users), French (10), Dutch (9), and Russian (8). Two-thirds of the users editing Simple were multilingual users, who already edited at least two other language editions. Of the 76 users editing Simple who were previously classified as editing only one edition,



(a) Network graph with English



(b) Network graph with English removed

Figure 4.5. Network graphs of co-editing patterns. Nodes represent language editions of the encyclopedia and the directed, weighted edges show the log of the number of users primarily editing one language edition who edited another edition as well. Both graphs show only edges with weights over 1.96 standard deviations above the mean. The top graph shows all language editions. In the bottom graph, the English edition is removed and the distribution of edges recalculated. Colors indicate communities found by the infomap community detection algorithm.

66 were primarily editors of the English edition. Simple makes little difference to the structure of the co-editing networks in Figure 4.5. Enough English users edit Simple to have a small directed edge from English to Simple, but no other edges are added.

Like the Esperanto edition, there appears to be a cohort of users dedicated more to Simple than to their native languages. There are 21 users in the dataset who edited Simple more than any other edition (14 of these users edited the English edition second-most). In addition, there are a further 44 users not in the dataset, who edited Simple more than two times but did not have at least four edits across all editions when their edits to Simple were excluded. Just under half (45%) of all the users who edited Simple most often did not edit any other edition at all. Of those users that did edit a second edition less frequently than Simple, that edition was English for all but 9 users.

There have been two proposals in the past to close Simple, both of which have failed. Whatever the utility of the edition to readers, it does have a dedicated community of editors. In this respect, it is very similar to the Esperanto edition, where 54% of the users that primarily edited Esperanto edited no other language edition.

4.5 Discussion

By far, most Wikipedia users edited only one language edition, confirming H4.1. However, just over 15% of users also edited multiple language editions. These multilingual users were found to be more active than their monolingual counterparts making more than 2.3 times as many edits per user on average. Most of this additional activity occurred in the users' primary languages, with only 2.6% of all edits being made by users in their non-primary languages. It is important to note that this is a correlation between multilingualism and activity and not causation. It may

be that the most dedicated and active users of Wikipedia contribute to multiple language editions regardless of how great their foreign-language skills really are. Survey work, for instance, has shown many Internet users in Uzbekistan engaged with foreign-language content even while simultaneously reporting low comfort with foreign languages (Wei & Kolko, 2005). Regardless of the direction of this relationship, it will be important to keep these multilingual users in mind when considering design changes to Wikipedia.

The percentage of users editing multiple languages on Wikipedia is similar to the 11% of users found to tweet in multiple languages on Twitter (Chapter 3). On the other hand, the percentage of users editing multiple language editions is far less than the 50% of users that self-reported editing multiple language editions in the 2011 Wikipedia editors survey.[14] This could perhaps follow from the idea that the most dedicated users are multilingual and thus more likely to take the time to respond to a survey about Wikipedia when given the opportunity. Alternatively, given that multilinguals only made a small fraction of their edits to their non-primary language editions, it is possible that more users would be observed editing multiple language editions if they were observed for a longer period of time.

Multilingual users editing more than one edition of Wikipedia can bring information, sources, and perspectives from the primary edition they edit to other editions. A large portion (44%) of the articles edited by multilinguals in their non-primary languages were to articles that no monolingual users in that language edited during the month of study. A similar percentage of all edits by multilinguals in their non-primary languages were to articles that the same multilingual user had edited in his or her primary language. This suggests that multilingual users are making unique contributions not duplicated by monolingual users and that in many cases multilingual users are working on the same article in multiple languages.

Hecht and Gergle (2009) have previously suggested that users crossing between

[14] https://meta.wikimedia.org/w/index.php?title=Editor_Survey_2011/Location_%26_Language&oldid=8409990

different languages like this might reduce the amount of self-focus bias in Wikipedia. They found the Dutch and Swedish editions to be less self-focused than other editions. The research presented in this chapter supports their conjecture that this is likely due to higher levels of multilingualism among speakers of these languages. This research shows that a relatively higher percentage of users primarily editing the Dutch or Swedish editions also edit another language edition. Hecht and Gergle (2009) also found the Portuguese edition to be less self-focused. The rate of multilingualism among users primarily editing the Portuguese edition in this study is slightly above the mean, but is mostly explained by the size of the Portuguese edition. Overall self-focus bias among the 15 editions studied by Hecht and Gergle (2009) is negatively correlated with the level of multilingualism found in this study. That is higher levels of multilingualism are generally associated with less self-focus bias. The correlation coefficient between the measures is -0.67, although this drops to -0.33 if English and Japanese are excluded. Multilingualism is one of perhaps several factors affecting the level of self-focus bias in different editions of Wikipedia, and this study has only been able to observe cross-language editing and not cross-language reading. Further study should identify additional factors affecting self-focus bias and their relative roles.

Multilingual users are found in all language editions. Generally, however, a higher percentage of users primarily editing smaller-sized language editions are multilingual compared to users primarily editing larger-sized editions, supporting H4.4. This is also consistent with prior qualitative and survey work. Of the outliers found, Esperanto and Malay had higher percentages of multilingual users than their sizes would predict, while Japanese had a much lower percentage than its size would predict. Malay users have previously been found to be among the most multilingual user groups on Twitter, while Japanese users were similarly found to be the least multilingual group on Twitter in Chapter 3. This points to the importance of language-specific factors, which are also shown in the rather simple case of Esperanto

being a constructed language with no native speakers.

Differences between the results on Twitter (Chapter 3) and those found here on Wikipedia also suggest platform-specific characteristics affect the levels of multilingualism users exhibit. For example, in the Twitter study Italian users were more multilingual on Twitter than their size suggested, while the opposite was found here. In addition, the correlation in the Twitter study between language size and levels of multilingualism was very weak whereas the correlation on Wikipedia was much stronger. Further research will be needed to untangle the role of design and platform-specific characteristics affecting the levels of multilingualism on different platforms.

When users did edit a second language edition, that edition was most often English, supporting H4.5. English users did edit many other language editions, but these users were a much smaller percentage of the English user total than the percentage of users primarily editing other languages that also edited English. Even with English removed, the network of language crossings was centered around German, the second-largest edition. Some regional and linguistic groups were also apparent, pointing towards the importance of geo-linguistic factors (Liao & Petzold, 2010) in the cross-language activity of Wikipedia users.

Including or excluding the Simple English edition of Wikipedia had little impact on the findings of this chapter. Many users editing Simple already edited two other editions and were classified as multilingual. Past research analyzing the location of users through the circadian rhythms of their edits found more editors of Simple were in Europe and the Far East/Australia compared to the English edition (Yasseri, Sumi, & Kertész, 2012). This raised speculation that English as a second language (ESL) speakers might edit Simple more than the English edition. This research finds, however, that the English edition is edited much more by users of other languages than is Simple. Thus, the difference in geographic spread between the two editions is more likely one of awareness and commitment to editing among

English speakers rather than a native/ESL divide. Indeed, this research has shown that like Esperanto, there is a dedicated editing community for Simple. Many users edit Simple (or Esperanto) more than any other edition despite no one being a native speaker of Simple (or Esperanto).

Overall, this study shows multilingual users play a unique role on Wikipedia editing articles different to those edited by monolingual users. Multilingual users may further transfer information between language editions and thereby reduce the levels of self-focus bias in the encyclopedia. The correlation between self-focus bias and multilingualism is present, but noisy, and further research is needed to identify other factors that also affect self-focus bias. Finally, differences between the levels of multilingualism by language previously found on Twitter with the levels found in this research on Wikipedia indicate design and platform-specific factors that future research should explore.

This chapter and the previous chapter on Twitter have shown that language divides are very strong on each platform, but that multilingual users are in positions to bridge these divides. The global scope of these two chapters, analyzing many languages quantitatively, has given a broad understanding of how users connect across languages. The chapters have not, however, specifically investigated the content contributed by multilingual users on user-generated content sites. The final empirical chapter adds this depth by specifically examining the editing behavior of English and Japanese multilingual users and the content they contribute about Okinawa, Japan, on Wikipedia.

Chapter 5

Okinawa

If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them.

—Genesis 11:6, "The Tower of Babel" (NIV)

5.1 Introduction

Allowing users to contribute content in multiple languages on user-generated content platforms results in a large difference in the content available in different languages. Within Wikipedia, for example, over 74% of concepts have an article in only one language edition, and more than 95% of concepts appear in six or fewer languages (Hecht & Gergle, 2010b). This finding is not unique to Wikipedia: on Twitter there is also surprisingly little overlap between the top domain names and hashtags used in tweets of different languages (Hong et al., 2011). User-generated content platforms face a trade-off on allowing the use of multiple languages. On the one hand, increased language diversity increases the potential number of contributors

This chapter is adapted from the following publication:
Hale, S. A. (2015). Cross-language Wikipedia Editing of Okinawa, Japan. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15. New York: ACM. doi:10.1145/2702123.2702346.

Additional material from the following extended abstract has also been included:
Hale, S. A. (2014). Okinawa in Japanese and English Wikipedia. In *Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14. New York: ACM. 10.1145/2559206.2579413.

as monolingual individuals from multiple languages can join. On the other hand, increased language diversity may also increase the risk of fragmenting content and users across languages, particularly if multilingual users who would have used the site in a large, international language move exclusively to smaller language editions. On question and answer platform Stack Overflow, the risk of fragmentation—that is, the risk that multiple language editions of the site would result in fewer users to any one particular language edition and therefore less high-quality answers—has been cited as one of the reasons for the platform to remain monolingual.[1]

Key to the trade-off between the potential increase in other-language users and the risk of fragmentation across languages are the roles technology and users play in facilitating the flow of information between languages. The first two empirical chapters of this thesis and previous research have suggested multilingual users who read and contribute content in multiple languages may share novel information between languages and broaden the scope of information available to monolingual users online (Eleta & Golbeck, 2012; Hecht & Gergle, 2010b). Research to date, however, has either examined differences in content between languages (Bilic & Bulian, 2014; Callahan & Herring, 2011; Hecht & Gergle, 2010b) or, as in the first chapters of this thesis, examined user behavior (Eleta & Golbeck, 2012), but not both. Thus, while 15% of the active editors on Wikipedia contribute to multiple language editions (Chapter 4), it remains unclear what content multilingual users contribute, how much content they contribute, and how valuable the content they contribute is. These questions are important in understanding the current roles multilingual users play in transferring information between languages as well as gaining insight for designing multilingual platforms in ways that maximize cross-language information transfer. This chapter starts to address this gap by examining the content contributed by multilingual users in their primary and non-primary languages on one of the largest multilingual user-generated content platforms online, Wikipedia.

[1] `http://blog.stackoverflow.com/2009/07/non-english-question-policy/`

The findings lead to a more in-depth understanding of the role multilingual users play on multilingual user-generated content platforms and suggest platform design strategies.

5.2 Background and related work

Each language edition of Wikipedia exhibits a self-focus: that is, each edition has, in general, more information about the regions where the language of the edition is spoken and less information about the regions where the language is not spoken (Hecht & Gergle, 2010b). Even when corresponding articles for a common concept exist across multiple language editions, there is a surprising diversity in the topics covered in each language edition's article on that concept (Bao et al., 2012). For example, the Spanish-language article on psychology contains a section on important contributions from Latin America that is not found in other language editions (Hecht & Gergle, 2010b).

The global nature of the Internet allows for the possibility that expatriates, language-learners, and other-culture/language enthusiasts—who Zuckerman (2013) has termed xenophiles—can spread information between languages on user-generated content platforms. In the one-month study of edits to the top 46 language editions of Wikipedia reported in Chapter 4, I found that approximately 15% of active Wikipedia users edited multiple language editions of the encyclopedia. These multilingual users were distributed across all language editions, but smaller-sized editions with fewer users had a higher percentage of multilingual users compared to larger-sized editions. The percentage of multilingual users primarily editing each language edition in Chapter 4 was found to negatively correlate with the self-focus of the 15 editions studied by Hecht and Gergle (2010b): that is, editions with more multilingual users exhibited less self-focus.

Multilingual users are well situated to act as bridges or gatekeepers and transfer

content between languages; however, past work points to a more nuanced picture of the extent to which multilingual users actually fulfill this role. Studies of multilingual users on Twitter show they are in structural positions to act as bridges across language groups (Chapter 3; Eleta & Golbeck, 2012) and that approximately 11% of active Twitter users write in multiple languages (Chapter 3). However, a study of multilingual users on Twitter in Switzerland, Qatar, and Quebec using LDA topic modeling found that multilingual users often focused on different topics in different languages (Kim, Weber, Wei, & Oh, 2014). This suggests these multilingual users may not be bridging across languages as much as their structural positions suggest. A similar situation may be present in Wikipedia, where 43% of the multilingual users studied in Chapter 4 edited articles about different concepts in their primary and non-primary languages.

The first research question of this chapter, therefore, simply asks *what articles do multilinguals edit in their non-primary languages?* (RQ5.1). This chapter compares the articles multilingual users edit in their non-primary languages with the articles edited by other users to understand the scope within which multilingual users may transfer information between languages. The results show that multilingual users edit a narrower set of articles in their non-primary languages and suggest design interventions such as cross-language content recommendation systems could broaden the scope of articles users edit in their non-primary languages.

Beyond what articles users edit in a non-primary language, this chapter also asks *what types of edits do multilingual users make in their non-primary languages?* (RQ5.2) in order to understand the nature and the extent of the information that multilingual users transfer between languages. A first dimension by which to compare contributions is size. Using the meta data available from Wikipedia, Chapter 4 analyzed the difference in the size of articles before and after each edit in bytes, but that measure is not the most reliable as an edit that adds a large block of text but also removes a different block of text could result in a size difference very close to

zero bytes. Using the content of edits, this study calculates a more nuanced measure of edit size using the number of words added, the number of words removed, and the amount of reorganization performed. A second dimension by which to compare contributions is the types of content changes users make in their non-primary languages. The study of 2010 Haitian earthquake blogs found that the sharing of images and videos was a large motivation for crossing language boundaries among bloggers (Appendix B), which suggests adding, removing, or otherwise modifying images might be more prevalent in users' non-primary languages.

The final research question posed in this chapter asks *how valuable are the contributions by multilingual users in their non-primary languages?* (RQ5.3). Given the diversity in information between languages on Wikipedia (Hecht & Gergle, 2010b) as well as online more generally, edits by multilingual users have the potential to introduce truly novel and valuable information and sources from one language into articles in another language. In my literature review (Chapter 2), I suggested connections across languages may be similar to the concept of "weak ties" in social network analysis, where ample scholarship has shown weak ties to be of critical importance to the spread of information with benefits ascribed to both the individual and the system/network as a whole (Burt, 2004).

Examining multilingual users on Wikipedia in Chapter 4, I also found a positive correlation between multilingualism and the number of edits users' made in their primary languages: users more active in their primary languages were more likely to edit in multiple languages. This suggests that multilingual users overlap to some extent with the group of very active elite or power users on Wikipedia, on which, like on many other platforms, much of the work is done by a small percentage of very active users (Priedhorsky et al., 2007; Kittur, Chi, Pendleton, Suh, & Mytkowicz, 2007). Monolingual studies have found that these users are responsible for a disproportionately large percentage of the content in the encyclopedia (Kittur et al., 2007) and are overwhelmingly responsible for the content that is viewed most frequently

(Priedhorsky et al., 2007). It remains unclear, however, how much content these users contribute and how long it persists when they are editing in their non-primary languages.

In order to investigate both users and content effectively, this chapter focuses on content and multilingual user contributions in English and Japanese in a relatively contained geographic area: Okinawa, Japan. Okinawa is an archipelago of small, sub-tropical islands home to a large number of native Japanese speakers and a large number of native English speakers.[2] Geographically closer to Taipei than Tokyo, the islands were once part of a prosperous independent kingdom built on trade in the region. After formal incorporation into Japan at the end of the 1800's, the islands were separated from Japan at the end of World War II and administered by the United States until 1972. Since that time, the US has maintained a strong presence, with half of all US personnel (military, contractors, dependents) in Japan under the US Status of Forces Agreement located in Okinawa. This accounts for just under 50,000 individuals (Ministry of Foreign Affairs of Japan, 2008) with military facilities occupying just over 18% of the land area of the largest and most populated island (Okinawa Prefecture, 2013).

Previous qualitative studies of Wikipedia have found differences in the editing behavior of users editing different language editions of Wikipedia—such as correlations with Hofstede's cultural dimensions (Pfeil, Zaphiris, & Ang, 2006). These studies have not examined the roles played by users who edit in multiple languages as is done in this chapter, but these previous studies do underscore the importance of studying users in multiple languages before making generalizations. Nonetheless, English and Japanese are good initial languages to study given that they are among the most-used languages online, not only on Wikipedia (Chapter 4), but also on Twitter (Chapter 3). Furthermore, speakers in each language play vastly different roles in interlanguage connections. In the one-month study of edits to Wikipedia,

[2]I have lived in Okinawa and can speak Japanese and English.

Japanese was a major outlier with only 6% of the primary editors of the Japanese edition editing a second edition. In contrast, English was very central in the cross-language movements of users: when non-English users edited a second edition, that edition was most frequently English.

5.3 Data

Finding the subset of all articles related to a particular geographic region on Wikipedia involves a trade-off between direct relevance and completeness (or, using the terminology of information retrieval, between precision and recall). Using the Wikimedia Labs[3] infrastructure, three article samples were extracted in October 2013. The *geotag sample* included all articles with geographic information (geotags) physically placing the articles in Okinawa.[4] The *category sample* included all articles in any category that contained the word "Okinawa" for the English edition or "沖縄" (Okinawa) for the Japanese edition. Finally, the *article link sample* included all articles containing a link to an article starting with "Okinawa" for the English edition or to an article starting with "沖縄" (Okinawa) for the Japanese edition. All edits to each article in each sample were then downloaded from the date the article was created until October 2013 using the Wikipedia API.[5]

The article samples were filtered to only include articles in the main, article namespace (i.e., not talk pages, user pages, etc.). The article link sample was also filtered to only include articles that mentioned Okinawa in the main body text of the article (i.e., not transcluded via a template to appear in a sidebar or footer). This was done so that the articles in the sample had a more substantial connection to Okinawa than just being part of a large group of articles linked together by a common portal or category such as "Regions and administrative divisions of Japan" or "USAAF Eighth Air Force in World War II."

[3] `https://www.mediawiki.org/wiki/Wikimedia_Labs`
[4] The bounding box used included articles between 25.75 and 26.9 E and 126 and 129 N.
[5] `https://www.mediawiki.org/wiki/API:Main_page`

Corresponding articles in the English and Japanese language editions were found using the October 2013 database dump from WikiData.[6] Launched in 2012, WikiData centralizes all interlanguage links (and, increasingly, statistics and other structured data) in one location and avoids some previous issues with out-of-date or conflicting interlanguage links when the links were separately maintained in each language edition of Wikipedia. As in Chapter 4, the Central Authorization database was queried with the username of each non-anonymous user to determine if the username was a global account connected to multiple language editions. If it was, the database for each language edition the user edited was queried to get the total number of edits per language that the user made since creating the account. The language of each user's most edited edition is referred to as the user's primary language throughout this chapter while the languages of any other editions edited by the user are referred to as the user's non-primary languages. Users belonging to the (ro)'bot' group or having the 'bot' template on their userpages as well as users that had been suspended for malicious editing were removed in order to focus on the behavior of good-faith, human editors.[7]

5.3.1 Measures of edit size and value

Users contribute value to Wikipedia in many ways. For example, Kriplean et al. (2008) found 42 different types of contributions to Wikipedia through an analysis of barnstars (personalized tokens of appreciation given to fellow users). The types of contributions they identified included programming tools/bots, designing templates, performing administrative functions, teaching, and leadership. Welser et al. (2011) similarly identified multiple user roles by analyzing the distribution of users' edits across the different namespaces on Wikipedia (articles, article talk pages, user pages, etc.).

[6] `http://www.wikidata.org/`

[7] The status of user accounts was checked one year after data collection in December 2014 to allow sufficient time for malicious editors to be reported and blocked. The data as well as the code used for data collection and analysis are available at `http://www.scotthale.net/pubs/?chi2015`.

In order to have one quantitative measure of the value of edits to articles, this chapter uses edit persistence, following past work analyzing Wikipedia in one language (Adler, Chatterjee, et al., 2008; Adler & Alfaro, 2007; Adler, Alfaro, Pye, & Raman, 2008; Priedhorsky et al., 2007). In particular, this chapter uses the algorithms developed by Adler et al. (2007) for their WikiTrust content-driven reputation system for Wikipedia to compute the extent to which each edit survives through the next six revisions to the article (edit persistence). WikiTrust accounts for some of the complexities of Wikipedia including differences created by rewording and rearranging text, and it also correctly attributes text that is deleted and then restored to the first author who contributed it (rather than to the author who restored it) (Adler & Alfaro, 2007). Edit persistence is calculated on a word-by-word basis for the next six revisions to each article so that users get partial credit for an edit even when part of their edit is subsequently changed or removed.

WikiTrust also computes a more detailed measure of edit sizes than is reported directly from Wikipedia. The edit sizes reported by Wikipedia simply measure the difference in page size (in bytes) before and after an edit. In contrast, the sizes computed by WikiTrust and used in this chapter are determined by the number of words added, deleted, changed, or moved. New words and deleted words contribute one point each, replacement words contribute 0.5 points each, and moving a word a fraction x of the normalized page length ($0 < x < 1$) contributes x of a point (Adler & Alfaro, 2007).

WikiTrust was developed and tested on languages with spaces between words; however, Japanese is written without spaces between words. Therefore, the Japanese text was first preprocessed to add spaces between words with mecab, an open-source library for text segmentation, part-of-speech tagging, and morphological analysis of Japanese text.[8] The WikiTrust algorithm was then run twice: once over the articles in the English article link sample and once over the articles in the Japanese article

[8] `https://code.google.com/p/mecab/` (in Japanese)

Sample	en-only	ja-only	Both
Geotag	52	185	152
Category	156	2,819	707
Article link	3,411	9,984	5,567

Table 5.1. The number of unique concepts in each sample. The majority of concepts have an article either only in the English edition or only in the Japanese edition (en-only or ja-only), while a smaller number of concepts have articles in both the English and Japanese editions (Both).

link sample. The reputation scores were computed only on the basis of users' edits in the data samples, and not in the full multi-terabyte dumps of all of Wikipedia articles.

5.4 Results

This section begins by describing the three article samples in order to understand the article landscape within which Wikipedia users were editing. It then addresses each of the three main research questions in turn:

1. *What articles do multilinguals edit in their non-primary languages?* (RQ5.1)
2. *What types of edits do multilingual users make in their non-primary languages?* (RQ5.2)
3. *How valuable are the contributions by multilingual users in their non-primary languages?* (RQ5.3)

All three article samples (geotag, category, and article link) show differences in the concepts related to Okinawa covered in the Japanese and English editions of Wikipedia. Consistent with previous research (Hecht & Gergle, 2010b), there are more concepts with an article in only one language (either Japanese or English) than concepts with articles in both languages (Table 5.1). The three samples have substantial overlap—all but a small handful of articles in the geotag sample are

present in the category sample, and the article link sample contains all the articles in both the geotag sample and the category sample.

In order to investigate the differences between the editions further, the inter-article links that connect articles together within each language were used to construct two networks for each sample. One network for the Japanese edition with each Japanese article in the sample as a node, and one network for the English edition with each English article as a node. Edges in both networks were the inter-article links between articles in the same language edition. Nodes were then ranked using the PageRank method (L. Page, Brin, Motwani, & Winograd, 1999). This algorithm, also used by Google, ranks nodes by the number of links to them weighted by the PageRanks of the nodes from which the links originate.

Within the geotag sample, the top-ranked English-only articles were mainly about US military facilities, equipment, and historic battles, while the top-ranked geotagged Japanese articles included a variety of parks, tourist areas, ports/terminals, and a shrine, reflecting that Okinawa is a Japanese tourist hot spot.

The top-ranked articles within the category and article link samples were very similar to each other, and given the substantial overlaps between all the samples, the remainder of this chapter uses only the article link sample to investigate the roles of multilingual users in the spread of content between the Japanese and English language editions. The article link sample has the benefits that an article link sample can be formed for any seed set of articles and that article links have been better studied on Wikipedia (Milne & Witten, 2008) than categories. In particular, Bao et al. (2012) examined the frequency with which different language editions were missing article links (that is the frequency with which a concept was mentioned without a link to the article about the concept) and found that the English and Japanese editions were missing possible article links at similar frequencies.

The top-ranked concepts in the article link sample with articles in both language editions that linked to Okinawa are shown in Table 5.2 ordered by their ranks in

Rank order in Japanese			Rank order in English	
沖縄県	Okinawa Prefecture	1	Japan	日本
日本	Japan	2	Taiwan	中華民国
1972年	1972	3	Kana	仮名 (文字)
4月1日	April 1	4	Guam	グアム
鹿児島県	Kagoshima Prefecture	5	Saipan	サイパン島
1945年	1945	6	Kyushu	九州
昭和	Shōwa period	7	Karate	空手道
九州	Kyushu	8	Tofu	豆腐
那覇市	Naha, Okinawa	9	Tinian	テニアン島
日本放送協会	NHK	10	Burakumin	部落問題

Table 5.2. Top concepts from the article link sample appearing in both editions and referencing Okinawa ordered by their PageRank scores in Japanese and English. The two common concepts in the top-10 for each language, Japan and Kyushu, are highlighted with a gray background.

Article title	English translation
沖縄返還	Okinawa Reversion
琉球放送	Ryukyu Broadcasting Corporation
沖縄セルラー電話	Okinawa Cellular
日本プロサッカーリーグ	Japan Professional Football League
MBSテレビ	MBS (Mainichi Broadcasting System) TV
西日本	Japan West
落語家	Rakugo Story Teller (Comic story teller)
アメリカ合衆国による沖縄統治	Okinawa under US administration
南日本放送	Minaminihon Broadcasting Co
鹿児島テレビ放送	Kagoshima Television Station

Table 5.3. Top concepts from the article link sample with articles only in Japanese ranked by the PageRank algorithm on the inter-article link network.

Article title	Description
Komainu	Broader category for shisa
Yukatchu	Ryūkyū Kingdom aristocracy
Karahafu	Japanese architectural style
Bunkai	Karate, Kata
Hagushi	Place in Yomitan, Okinawa
Isshin-ryū	Karate, Style
Matsubayashi-ryū	Karate, Style
Shōrinji-ryū	Karate, Style
Wanshū	Karate, Kata
Wankan	Karate, Kata

Table 5.4. Top concepts from the article link sample with articles only in English ranked by the PageRank algorithm on the inter-article link network.

the Japanese network on the left and ordered by their ranks in the English network on the right. The top-ranked Japanese articles include many historical references: April 1 (the start of the Battle of Okinawa in World War II), 1945 (the year of the Battle of Okinawa after which Okinawa was separated from Japan), 1972 (the year Okinawa returned to Japan from US administration), and Shōwa period (1926 to 1989 in the Japanese calendar). The top-ranked articles also include NHK (the Japanese public broadcaster), Naha (the capital of Okinawa), and Kagoshima Prefecture (the prefecture neighboring Okinawa to the north). Both the English and Japanese lists include Japan and Kyushu (the southern most island of mainland Japan), while the top-ranked articles in the English edition also include many surrounding geographical locations within Asia, but outside of Japan: Taiwan, Guam, Saipan, and Tinian.

Articles only appearing in one language edition are shown in Table 5.3 for the Japanese edition and Table 5.4 for the English edition. The top-ranked articles found only in the Japanese edition again include historical articles such as "Okinawa Reversion" (Okinawa returning to Japan after US administration in 1972) and "Okinawa under US administration" (an overview article about the period not

paralleled in English).[9] The Japanese articles from the article link sample also include many companies based partially or entirely in Okinawa that do have articles in English.

The English edition includes many articles on specific karate styles and kata (forms/patterns) that have no corresponding articles in Japanese. The presence of detailed articles on karate kata and styles in English, but not in Japanese, reflects Okinawa's history as the birthplace of karate and that karate spread widely after gaining popularity among the American military members stationed in Okinawa after World War II.

Thus, in line with existing research on self-focus bias (Hecht & Gergle, 2009), this analysis shows that while there are overlaps in the coverage of Okinawa between the English and Japanese editions of Wikipedia, there is also a large difference in the content within each language. The English edition most reflects Okinawa as the birthplace of karate and as a location within Asia more generally. In contrast, the Japanese edition reflects more of Okinawa's history and Okinawa as a location within Japan rather than as a location within Asia more generally.

5.4.1 Article selection

This subsection examines what articles users edited, with a particular emphasis on what articles multilingual users edited in their non-primary languages in order to answer the first research question. Most editors were anonymous, had local accounts, or had global accounts that primarily edited the language edition in question (Table 5.5). The prevalence of edits by anonymous users is consistent with previous research analyzing Wikipedia and not a peculiarity of this sample (Anthony, Smith, & Williamson, 2009). It is difficult to calculate per-user statistics for anonymous users since IP addresses change over time and multiple users may edit from the same

[9]Both the Japanese and English editions do have articles about more specific aspects of the US administration of Okinawa. For example, both editions include more specific articles on the military and civilian governments and on the "B yen" currency used in Okinawa during part of the time.

	Total users		Total articles edited		Articles edited per user			Edit size per user (log)		
	Count	%	Count	%	Median	Mean	SD	Median	Mean	SD
English edition										
Anonymous	192,839	73.4%	216,840	46.15%						
Local account	15,008	5.7%	58,689	12.49%	1	3.91	21.79	1.79	1.97	1.93
Pri. English	50,038	19.0%	179,951	38.30%	1	3.60	20.74	1.94	2.08	1.97
Pri. Japanese	466	0.2%	1,488	0.32%	1	3.19	7.32	1.16	1.38	1.67
Pri. Other	4,341	1.7%	12,911	2.75%	1	2.97	16.44	0.47	1.13	1.71
Totals	262,692	100.0%	469,879	100.0%	1	3.62	20.67	1.84	2.00	1.96
Japanese edition										
Anonymous	372,852	88.4%	717,608	62.74%						
Local account	9,945	2.4%	109,765	9.60%	2	11.04	47.84	3.09	2.95	1.81
Pri. English	558	0.1%	5,531	0.48%	1	9.91	58.47	0.96	1.55	1.95
Pri. Japanese	37,191	8.8%	301,980	26.40%	1	8.12	43.97	3.00	2.91	1.83
Pri. Other	1,174	0.3%	8,954	0.78%	1	7.63	57.44	0.18	1.07	1.76
Totals	421,720	100.0%	1,143,838	100.0%	1	8.72	45.35	2.97	2.87	1.85

Table 5.5. User counts, articles edited, and edit sizes. The primary (pri.) language of a user with a global account is the language of the most-edited edition of Wikipedia.

	# of Japanese users editing English		# of English users editing Japanese	
	Estimate	(Standard error)	Estimate	(Standard error)
Exists in both languages	0.641***	(0.024)	3.285***	(0.034)
Total number of editors	0.001***	(0.0001)	0.003***	(0.0001)
PageRank	0.014***	(0.0005)	0.245***	(0.006)
Number of images	0.003***	(0.001)	0.054***	(0.002)
Number of external links	0.001***	(0.0003)	−0.0003	(0.0004)
Constant	0.008	(0.015)	0.029	(0.019)
Observations	5,441		14,825	
Adjusted R^2	0.348		0.572	
Residual Std. Error	0.849 (df = 5435)		1.828 (df = 14819)	

*p<0.1; **p<0.05; ***p<0.01

Table 5.6. Linear regression results fitting the number of primary Japanese users editing each English article and the number of primary English users editing each Japanese article.

IP address, but relevant statistics for anonymous users are presented where possible given the size of this group.

A small number of registered users primarily edited either the Japanese or English edition, but also edited the opposite edition: 558 primary editors of the English edition edited articles in the article link sample from the Japanese edition, and 466 primary editors of the Japanese edition also edited articles in the article link sample from the English edition.

The articles users choose to edit reflect how different groups of users distribute their time and energy across articles. Although the previous section showed that most concepts had articles in only one language, the majority of edits by all users were to concepts with articles in both languages (Figure 5.1). Even so, the edits by multilingual users writing in a non-primary language were significantly more concentrated on concepts with corresponding articles in both languages compared to the edits of other users writing in each language.[10]

The finding that multilingual users mostly edit articles with corresponding articles in their primary languages is confirmed by a linear regression (Table 5.6), which also shows the articles users edited in their non-primary languages tended to have

[10]The difference in means between any two groups within either edition is significant with $p < 0.001$.

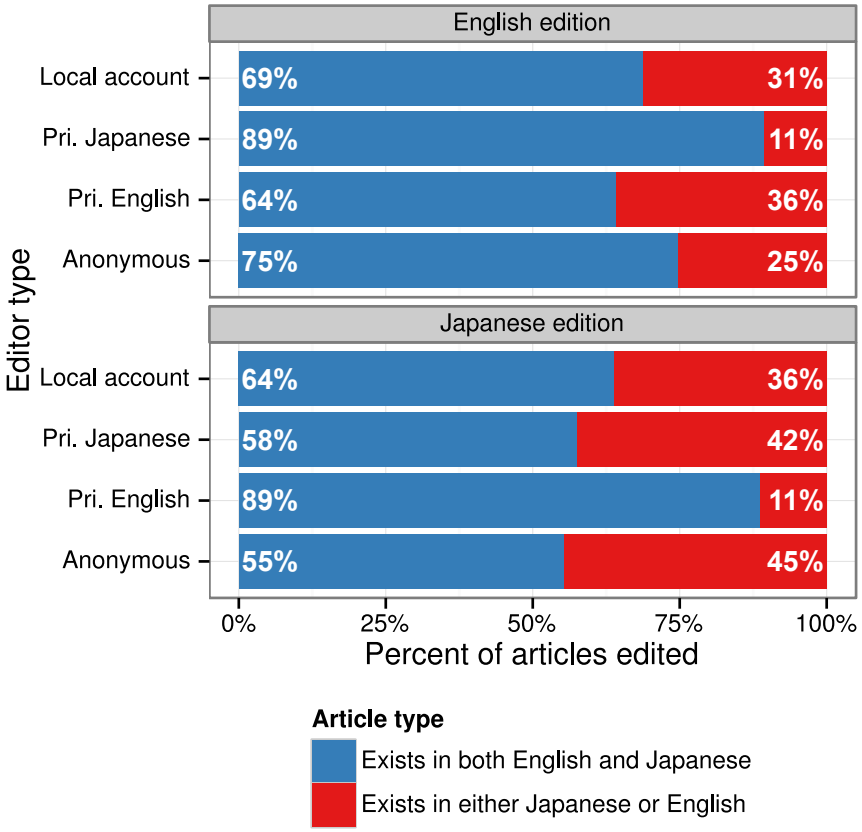


Figure 5.1. English users editing the Japanese edition are far less likely than other users to edit articles that only appear in Japanese. Similarly, Japanese users editing the English edition are far less likely than other users to edit articles that only appear in English.

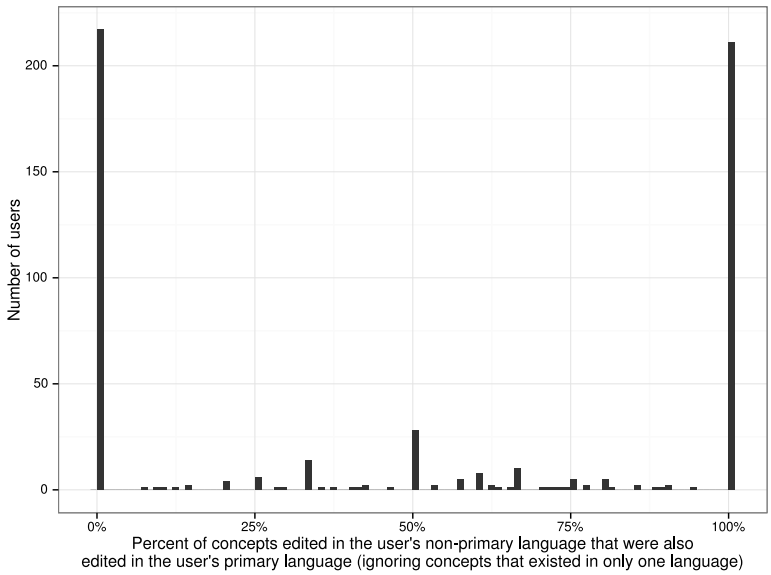


Figure 5.2. Histogram showing the percentage of concepts that multilingual users edited in their non-primary languages that they also edited in their primary languages. Users editing concepts only existing in one language are omitted.

more overall edits/editors, more images, and have higher PageRank scores computed as described earlier. The articles Japanese users edited in English also tended to have more links to external sources, but the number of links to external sources was not significantly associated with the number of English users editing an article in Japanese.

A per-user investigation of editing behavior reveals a more nuanced understanding. About half (47%) of all multilingual users only edited Okinawan-related concepts with articles in one language (the users in this group made only a small number of overall edits within the Okinawa sample). As in Chapter 4, the multilingual users who edited concepts with articles in both languages within the Okinawa sample fit into two distinct groups (Figure 5.2). For about 20% of multilingual users, every article the users edited in their non-primary languages was about a concept for which they had also edited the corresponding article in their primary languages. However, a slightly larger percentage of multilingual users (21%) edited different concepts in their non-primary and primary languages. That is, for these users every article the

users edited in their non-primary languages was about a concept for which they did not edit the corresponding article in their primary language. The behavior of these two groups of users was very similar with the only significant difference being that the users editing different concepts across languages made more image-related edits than the users editing the same concepts across languages. This finding is expanded upon in the relevant subsection below.

These results give a more detailed picture of the article selection behavior of multilingual users. Data on the articles that users viewed is not available, and thus it is not possible to say whether multilingual users read (but did not edit) articles in their primary languages before editing the corresponding articles in their non-primary languages. The editing data, however, clearly shows that while multilingual users in this dataset edit similar proportions as other user groups of one- and two-language concepts in their primary languages, they disproportionately edit a smaller amount of one-language concepts in their non-primary languages.

5.4.2 Types of contributions

This subsection addresses the second research question on the size and type of edits multilingual users make. Setting aside anonymous users and local accounts to compare users who have global accounts to one another, a finding in both editions is that those users who edited articles in both Japanese and English in the dataset were very active on their primary language editions. Considering users with global accounts who primarily edited the English edition, about one percent of these users also edited the Japanese edition. However, this group of users was responsible for six percent of all edits to the English edition made by English users. Similarly, Japanese users who also edited the English edition were also about one percent of all Japanese users with global accounts, but they made 13% of all edits by Japanese users in the Japanese edition.

While multilingual users were very active in their primary languages, a smaller

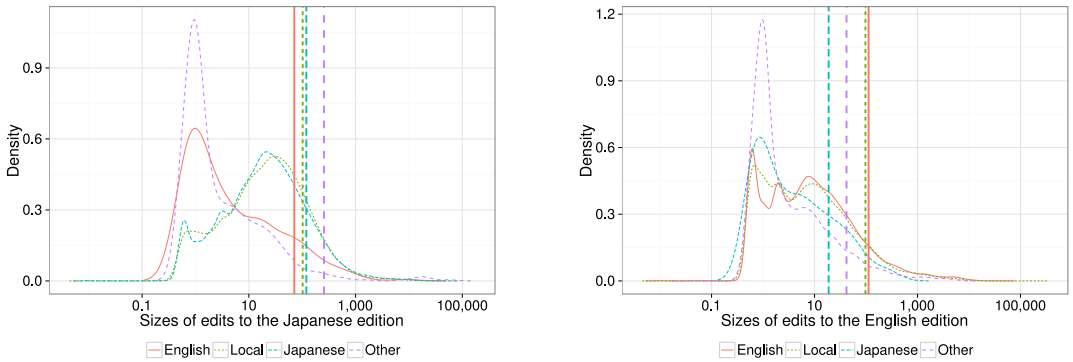


Figure 5.3. Density plots for non-anonymous users editing articles in the Japanese (left) and English (right) editions grouped by their primary language editions. Vertical lines indicate distribution means.

percentage of each user's total edits were to each user's non-primary languages. Overall, 14% of all edits by multilingual users were to their non-primary languages, which is higher than the work in Chapter 4 that found only 2.6% of edits across all language editions for one month were from users writing in their non-primary languages. This could be due to the focus on a geographic region with large numbers of English and Japanese speakers and/or the longer time frame of the analysis.

Edit sizes

Users who edited both the Japanese and English editions have two sets of scores for their edit sizes and edit persistence. One set for the English edition and one set for the Japanese edition. First looking at the scores for users' primary language editions, a consistent finding in both Japanese and English is that those users who edited both editions made significantly larger edits than the users who only edited one edition.[11] Japanese users who also edited the English edition made larger sized edits in Japanese compared to Japanese users who did not edit the English edition (median 3.9 vs. 3.0, mean 3.8 vs. 2.9, sd both 1.8, $p < 0.001$). Likewise, English users who also edited the Japanese edition made larger sized edits in English compared

[11] Given the heavy-tailed distribution of average edit sizes, the results are reported after being transformed with the logarithm.

Edit category	Pri. lang.		Non-pri. lang.		p-val†
Addition	97	31%	47	26%	0.25
Maintenance	103	33%	44	24%	0.04
Deletion/Reversion	37	12%	11	6%	0.03
Image-related	27	9%	32	18%	0.01
Interlanguage links	8	3%	32	18%	0.00
Change	65	21%	34	19%	0.62
Total edits‡	315		181		

Table 5.7. Exploratory, qualitative coding of edits in users' primary languages (pri. lang.) and non-primary languages (non-pri. lang.). †p-values are for two-tailed t-tests on difference of percentage means. ‡Some edits are assigned to multiple categories and, therefore, the column sums are greater than the total number of edits reported.

to English users who did not edit the Japanese edition (median 2.4 vs. 1.9, mean 2.5 vs. 2.1, sd 1.9 and 2.0, $p < 0.001$).

While multilingual users made larger sized edits in their primary languages, the analysis of their edits in their non-primary languages reveals a very different picture (Figure 5.3 and Table 5.5). The sizes of edits to the Japanese edition by English users were significantly smaller than the sizes of edits to the Japanese edition by Japanese users ($p < 0.001$). Similarly, the sizes of edits to the English edition by Japanese users were significantly smaller than the sizes of edits to the English edition by English users ($p < 0.001$).

Content changes in primary and non-primary languages

In order to better understand the gap between the edit sizes of users in their primary and non-primary languages, a small subset of edits was explored qualitatively. A random set of 70 users who had edited both editions was chosen: 35 users who primarily edited the English edition and 35 users who primarily edited the Japanese edition. Up to five edits from each edition were randomly chosen for a total of up to 10 edits from each user. Despite the measures to remove (ro)bots described in the data section, qualitative analysis revealed that one randomly chosen user was a bot:

this user was replaced with another randomly chosen user. Not all users had five edits in each edition; so, 496 edits were reviewed in total. This set included edits by English users to the English edition (145 edits) and the Japanese edition (96) as well as edits by Japanese users to the English edition (85) and the Japanese edition (170). The findings suggest that users made different types of contributions in their primary and non-primary languages, which may account for the differences in the computed size of their edits.

Edits made to articles (but not to talk pages, etc.) were examined in order to understand the contributions users made to articles in their primary and non-primary languages. After consulting previous literature (Pfeil et al., 2006; Kriplean et al., 2008), initial codes were created through an emergent coding of a subset of the data. These were refined into six (non-exclusive) categories, and the full sample was systematically coded using these categories. Each edit was classified as making an addition (adding new text or references to an existing article or creating a new article), as maintenance (adding, removing, or adjusting templates, categories, links in a "See Also" section, or whitespace changes that did not alter text), as deletion/reversion (reverting an edit or deleting text from an article), as image-related (adding, altering, or removing an image), as altering interlanguage links, and/or as change (edits that changed existing text such as correcting spelling errors or updating facts that had changed like the latest winner of an annual sports tournament).

There was a significant difference between the types of edits users made in their primary languages compared to the types they made in their non-primary languages ($\chi^2 = 48, p < 0.001$, Table 5.7). Users made significantly more deletions/reversions and maintenance edits in their primary languages compared to in their non-primary languages. On the other hand, users made significantly more image-related edits and added/removed significantly more interlanguage links in their non-primary languages compared to in their primary languages. The findings related to interlanguage links are no longer applicable to Wikipedia as these links are now maintained

separately within WikiData. Nonetheless, it is noteworthy that the task of locating a related article and linking it across languages was motivating enough for some users to edit a foreign language edition.

The proportion of addition edits and change edits did not differ significantly between users' primary and non-primary languages. Overall, 15% of the edits made by Japanese users to the English edition concerned fixing incorrect romanizations of Japanese words and/or adding Japanese characters for terms. These types of language-specific edits that are easy for native speakers but harder for non-native speakers illustrate both the value of cross-language collaboration and also why these users may have been making edits of different types in their primary and non-primary languages.

The percentage of image-related edits is the only significant difference between the group of multilingual users who edited different concepts in each language compared to the group of multilingual users who edited the same concepts in each language (Figure 5.2). Users who edited different concepts in their primary and non-primary languages made significantly more image-related edits in both languages (16% of all examined edits by these users) compared to users who edited the same concepts in both languages (4% of all examined edits by these users, $p = 0.006$).

5.4.3 Value of edits

As stated in the Data section, there are many ways in which users contribute value to Wikipedia, but a common, quantitative measure on which to compare the many different types of contributions is how much of each edit is retained through subsequent revisions to the article. Using the WikiTrust algorithms, the next six revisions after each edit were examined to compute how much of the edit was retained (persisted) through these revisions. Each edit was given a normalized score from -1 (edit completely removed) to 1 (edit completely retained). Comparing the mean edit persistence scores showed that the text from edits made by non-primary editors

survived at a similar rate to the text from edits made by users who primarily edited each edition.[12]

5.5 Discussion

The large differences in content observed between different editions of Wikipedia globally (Hecht & Gergle, 2010b) also applied to articles related to Okinawa despite the presence of large numbers of Japanese and English speakers living on the island. Similarly, the small percentage of users editing multiple editions of Wikipedia found in Chapter 4 applied to this dataset as well. In many ways, Okinawa is a hard case: Japanese and English use very different writing systems, and Japanese users have consistently been observed to engage less with other-language content not only on Wikipedia (Chapter 4), but also on Twitter (Chapter 3).

Nonetheless, this work has shed greater light on the selection of articles multilingual users in these two languages edit, the types of contributions they make, and the value of these edits. If further research confirms the patterns found in this chapter apply more broadly, then one key challenge for designers of multilingual platforms seeking to facilitate cross-language information exchange is increasing multilingual users' awareness of and contributions to related other-language content. The multilingual users in this study were far less likely to edit articles in their non-primary languages that did not have corresponding articles in their primary languages despite these articles being more numerous. The large difference in content between languages applies not only to the sample used for this study, but also overall on Wikipedia (Hecht & Gergle, 2010b), on Twitter (Hong et al., 2011), and likely to most other user-generated content platforms. The challenge of making users aware of

[12]The average score for Japanese users editing the English edition (median 0.44, mean 0.38, sd 0.58) was higher but not significantly different from the average score for English users editing the English edition (median 0.47, mean 0.35, sd 0.64, $p = 0.46$). Similarly, the average score for English users editing the Japanese edition (median 0.45, mean 0.44, sd 0.53) was marginally higher but not significantly different from the average score for Japanese users editing the Japanese edition (median 0.37, mean 0.43, sd 0.50, $p = 0.72$).

content available in their non-primary languages but not in their primarily languages is thus a challenge likely to be faced by designers of all multilingual user-generated content platforms.

Further research will be needed to understand the precise implications of design on user content selection, but it seems likely that the prominence of interlanguage links connecting related articles across languages on Wikipedia and the lack of other-language discovery tools are partially responsible for the narrow scope of multilingual users' edits in their non-primary languages. Currently, for example, there is no facility to search multiple language editions of Wikipedia simultaneously. So, if a user generally searches only in his primary language, that user may not discover the content that exists in another language if the content has no corresponding article in his primary language even if the user reads the other language. Very often the full text of articles in the Japanese edition includes an English translation of the article's concept, and likewise Japanese terms are often included in Japanese-themed articles in the English edition. If the search interface automatically checked a user's non-primary language editions when no matches were found in the user's primary language edition, that user might discover articles in his non-primary languages that have no article in his primary language.

Another possible design change would be to suggest articles related to a specific theme that exist in users' non-primary languages, but not in their primary languages. Such an approach could employ similar methods to those used here: gathering all articles linking to a given article and computing the PageRank scores or other methods like Latent Dirichlet Allocation (LDA) (Haruechaiyasak & Damrongrat, 2008). In practice, this might look very similar to the valuable SuggestBot (Cosley, Frankowski, Terveen, & Riedl, 2007) tool, which can recommend articles for users to edit based on the articles the users have previously edited in one language. Currently, separate versions of SuggestBot run independently in multiple languages, and a potentially useful (although certainly non-trivial) step would be to extend

SuggestBot to offer suggestions across user-selected (or inferred) languages for users who desire such suggestions. That is, based on the articles a user has previously edited in one language, the user could ask for articles in another language that either need work or that have no equivalent in the user's primary language.

A small, but dedicated group of users, who made large-sized edits in their primary languages, also edited articles in a second language. The edits in the users' non-primary languages were smaller in size, but were equally valued by the site's users persisting through subsequent revisions at a similar rate to the edits made by users editing only one language edition. An exploratory analysis of the edits users made in their primary and non-primary languages indicated that the differences in edit sizes were partially due to users making different types of edits in their primary and non-primary languages—multilingual users more frequently edited images and interlanguage links in their non-primary languages. In contrast, they made more maintenance and deletion/reversion edits in their primary languages.

Even if the edits in users' non-primary languages are smaller and of a different type, they still have value. The qualitative exploration of edits revealed many examples of users updating out-of-date information and correcting errors in their non-primary languages. Japanese users also frequently added or corrected relevant Japanese-language text in the English edition. There were also edits of addition and, occasionally, translation into users' non-primary languages.

The Language Engineering team of the Wikimedia Foundation[13] has been actively developing an (open-source) content translation tool to help users translate content between different language editions of Wikipedia.[14] While the integration of machine translation and bilingual dictionaries, the automatic conversion of article links, and the streamlined user-interface of the translation tool will no doubt assist would-be translators, this research indicates that helping users find articles they want to translate will be a major hurdle. Multilingual users in this study

[13] `https://wikimediafoundation.org/wiki/Language_Engineering_team`
[14] `https://www.mediawiki.org/wiki/Content_translation`

clearly made their largest-sized contributions in their primary languages, suggesting that platforms might be more successful in encouraging translation from users' non-primary languages to their primary languages rather than from users' primary languages to their non-primary languages. This would require surfacing content in users' non-primary languages that does not exist in the users' primary languages even while the users are viewing content in their primary languages.

Beyond translation, there are a range of contributions users can make on multilingual user-generated content platforms that require a varying level of cross-language proficiency. Offline data on multilingualism is imperfect and incomplete, but best estimates suggest around half of all humans speak two or more languages (Grosjean, 2010). Thus, it might be possible to encourage far more users to contribute content in multiple languages on user-generated content platforms. Survey work suggests that Internet users consume content in multiple languages more frequently than they contribute content in multiple languages (not only on Wikipedia,[15] but also more generally online (Eurobarometer, 2011)). The prevalence of image-related edits in this study and the apparent motivation images provide for cross-language linking in the blogosphere (Appendix B) suggest multimedia content is a low-barrier entry point for increasing multilingual contributions from users. Designers of multilingual user-generated content platforms wishing to increase multilingual activity could specifically consider and optimize cross-language multimedia content related tasks and other low-barrier entry points for multilingual contributions as one possible way to increase the number of users contributing in multiple languages on their sites.

Adler et al. (2008) advocate combining together the measures of edit size and edit persistence to form a reputation score for each user, which their WikiTrust work uses to predict the quality or trustworthiness of users' contributions to Wikipedia. If such a system evaluated the contributions of multilingual users separately for each

[15] `https://meta.wikimedia.org/w/index.php?title=Editor_Survey_2011/Location_%26_Language&oldid=8409990`

language, most multilingual users would have low reputation scores in their non-primary languages due to the smaller sizes and smaller number of their edits while they would have large reputation scores in their primary languages. The analysis in this chapter, therefore, shows the importance of evaluating multilingual users holistically across their multiple languages to accurately measure their contributions to user-generated content sites.

Successfully combating the risk of fragmenting users and content too thinly across multiple languages on user-generated platforms requires a more advanced understanding of localization and internationalization than traditional principals such as translating interfaces to "speak the user's language(s)" (Nielsen, 1993). Multilingual users need to be specifically considered in site design. This involves not only making existing cross-language connections visible, but also designing for the discovery of related foreign-language content not available in users' primary languages. The most active and dedicated users will reach across language boundaries online to contribute to other-language content, but the users in this study made their largest-sized contributions in their primary languages. Thus, successful multilingual user-generated content platforms need to nurture both active, dedicated monolingual communities and also encourage multilingual users to discover other-language content and serve as bridges between monolingual communities. Further research should analyze the extent to which multilingual search, cross-language content recommendation, and optimized low-barrier entry points for multilingual contribution can help multilingual users better understand the differences in content between languages, encourage these users to transfer more information between different languages, and thereby enable wider access for all users to the most interesting and important material that is not yet in their primary languages.

Chapter 6

Discussion and Conclusions

> *All of them were filled with the Holy Spirit and began to speak in other tongues... a crowd came together in bewilderment, because each one heard their own language being spoken. Utterly amazed, they asked: "... how is it that each of us hears them in our native language?"* —Acts 2:4–8 (NIV)

This thesis investigated the cross-language dynamics of Twitter and Wikipedia as two examples of user-generated content platforms. The two platforms differ in their designs, uses, and affordances (Section 2.2.4). Twitter, which began with a very passive approach to language, was selected as a platform designed with high level of language integration *a priori*. In contrast, Wikipedia, which began with independent sites for each language, was selected as a platform designed with a low level of language integration *a priori*. Despite these design differences, several commonalities were found on both platforms.

Apart from design differences and how content diffuses within a platform, it is clear that platforms themselves diffuse (are adapted) differently. The largest language communities present on Twitter (Table 3.1) and Wikipedia (Table 4.1) differ in meaningful ways. The most used language is English on both platforms (41% of Twitter users, 48% of Wikipedia users), but whereas German and French are more prominent on Wikipedia, they are less so on Twitter where Japanese, Portuguese, and Malay/Indonesian are more prominent.

Chinese is a special case given the government filtering in mainland China. The language is much more prominent on Wikipedia than on Twitter, despite both platforms facing competition from similar Chinese "clone" platforms as well as on-and-off access restrictions in mainland China. A recent exploratory study suggests one possible reason for the prominence of Chinese on Wikipedia is the perception among users that Wikipedia is more reliable than its Chinese clone (Baidu Baike, 百度百科)—although the same study also found Baidu was popular for its lax enforcement of copyright (Liao, 2014). The design of the Chinese edition of Wikipedia allows for both simplified or traditional Chinese characters with automatic conversation between them, and this permits editors both in mainland China and other locations (Taiwan, Hong Kong, Singapore) to contribute to the same articles in either script. Baidu Baike exclusively uses simplified characters and does not offer any automatic conversion to traditional characters, which likely reduces its appeal to users outside of mainland China.

6.1 Findings

Despite these design and userbase differences, several common characteristics emerged from the analysis of the two platforms (Table 6.1). These commonalities included that language had a strong role in structuring both platforms, multilingual users served as bridges between monolingual users in different languages, the number of speakers primarily using a given language affected the cross-language dynamics, and multilingual users were more active than their monolingual counterparts. Further research will need to assess these factors on other user-generated content platforms, but the fact that these commonalities were found on two very different user-generated content platforms is an encouraging indicator of more general applicability.

Twitter	Wikipedia	
✓	✓	Language has strong role in structuring the platform
✓	✓?	Users engaging with content in multiple languages (multilingual users) serve as bridges between different clusters/editions
X	✓	Users primarily writing in less-represented languages are more likely to cross-language boundaries than users writing in highly-represented languages
✓	✓	When users cross languages they cross to larger languages (e.g. English) and thus at a language level English forms more bridges than any other single language
✓†	✓†	Multilingualism is correlated with an increased amount of activity

Table 6.1. Comparable hypotheses from Twitter and Wikipedia. ✓ indicates the hypothesis was supported, ✓? indicates the hypothesis was partially supported, and X indicates the hypothesis was not supported. †The final correlation between multilingualism and activity was not hypothesized.

6.1.1 Language structures platforms

Language plays a large role in structuring user interactions on user-generated content platforms. Language is part of the design of Wikipedia, insofar as separate language editions each operate with a high degree of autonomy. The separation of languages in the design is reflected in user behavior in that only 2.6% of all edits to the encyclopedia were made by users editing the encyclopedia in a non-primary language. Furthermore, the edits that users make in their non-primary languages tend to be smaller in size than and of different types to the edits they make in their primary languages.

The mention and retweet behavior of users on Twitter is also heavily influenced by language even though the design of the platform is much more language-neutral (i.e., there are not separate editions of Twitter for different languages and users cannot selectively send a tweet to only a subset of their followers). Nonetheless, the comparison between the the languages of users and the network structure of users' mentions/retweets showed that the users' languages and the network structure are

very closely aligned. Grouping users together based solely on the most-used language of each user resulted in a division of the network with a very strong modularity score of 0.68. As modularity is a measure of goodness-of-fit that compares the density of edges within and between subsets of a network (Newman & Girvan, 2004), the high modularity score achieved solely by grouping users with the same primary language together shows a strong alignment between mention/retweet behavior and users' languages. This is further demonstrated in that over 70% of the clusters found using the label propagation algorithm were formed of users who all shared one common, most-used language. This is significantly higher than the less than 0.5% that would be expected randomly and occurs despite the fact that the label propagation algorithm works exclusively on the network structure to find clusters and has no information about users' linguistic abilities.

The finding that language plays a large role in structuring user interactions on platforms fits with other work showing a large diversity in content between languages (e.g., Hecht & Gergle, 2010b). The finding has a retrospective obviousness to it, but it should not be under-rated. In a world that feels increasingly globally connected with more immigration and travel, this finding is a reminder that the fragmenting effects of language are still very present and real (at least in online communications).

6.1.2 Multilingual users serve as bridges

At the same time that language structures interactions and divides users, users who contribute content in multiple languages (multilingual users) have the potential to bridge these divides in user-generated content platforms. Multilingual users on Twitter are more often in unique bridging positions than monolingual users and than would be expected at random (Section 3.3.2). Multilingual users had previously been shown to bridge languages at the individual (ego-net) level (Eleta & Golbeck, 2012), but this thesis expanded that previous work by performing analysis of cross-language bridging at the larger network-level. The analysis showed that in general the bridges

created by multilingual users connected meaningfully-sized clusters of monolingual users together rather than only connecting small, peripheral other-language clusters. Overall, the many bridges across different language pairs in aggregate resulted in the global connectivity of the network.

By the design of Wikipedia, multilingual users editing multiple language editions are in positions to bridge across languages, and the cross-language (interwiki) links that connect topics across languages on Wikipedia allow for a more in-depth analysis of the actual topics multilingual users contributed to in different languages. While multilingual users are in bridging positions, the analysis of edits across languages revealed a more nuanced picture in which only a subset of users edited articles on the same concepts across languages (Figure 4.3).

The extent to which and the reasons that multilingual users contribute to different topics across language divides is an enduring question that my future research will pursue. The case study of edits to Okinawa (Chapter 5) suggests one possible reason users would contribute to different concepts across languages is that they could be making different types of contributions in different languages. Another possible explanation is that users are reading, but not editing, content on a topic in one language before contributing content in a second language. In future work, I hope to analyze this possibility by working with the operator of a user-generated content platform to examine both the content users read and the content they contribute across languages on the platform. Finally, a possible linguistic explanation for the observed behavior lies in multilingual users having different domains of expertise in different languages (e.g., academic topics in one language and cultural topics in another) (Grosjean, 2010). In a small survey of multilingual users, Steichen et al. (2014) report multilingual users often search for and browse different topics in different languages.

Overall, the findings of this thesis establish the potential for multilingual users to bridge language divides and transfer information between monolingual users of dif-

ferent languages. The findings in this thesis suggest that cross-language transfer by multilingual users is occurring to some extent in user-generated content platforms. Several large, multilingual clusters were found on Twitter, and clear examples of introducing new information into a second language were seen qualitatively in the case study of Okinawa-related articles on Wikipedia. At the same time, the findings (particularly from the Okinawan case study) suggest that cross-language contributions are generally smaller in size and less frequent than same-language contributions. Compared to offline data on multilingualism, the data collected in this thesis shows the potential for greater cross-language contribution, which is discussed further below (Section 6.2.1).

6.1.3 Language-specific factors

Similar relationships exist in how users connect across different languages on the two platforms. These commonalities point to the role of language-specific factors affecting the geolinguistic dynamics (Liao & Petzold, 2010) of platforms that transcend individual platform differences. Japanese, as a major outlier on both platforms, is a prime example in this respect. In general, Japanese users are less likely to contribute content in multiple languages online (Figure 3.5 and Figure 4.4) even when taking into account the number of speakers using the language on a platform.

English is also an outlier given the very large number of users primarily writing in the language. Worldwide there are more native speakers of Chinese or Spanish than of English (Lewis, Simons, & Fennig, 2014). The large size of English, therefore, is likely a reflection of Internet diffusion patterns—more Internet access in English-speaking regions (Graham et al., 2012)—and also of the large number of people who speak English as a second language, including the use of English as a common language between speakers with different native languages (Crystal, 2003). The commonalities in the size and role of Japanese and English across the two platforms demonstrate wider language/cultural phenomena that, along with specifics of other

language groups, deserve further study.

The number of users contributing content in a language also has important effects on the position and role of the language in the network of language crossings on user-generated content platforms. Larger, established languages—particularly English, but also German on Wikipedia—have extremely central roles in the cross-language networks. Multilingual users most often cross between one of these larger languages and another language. However, regional, geolinguistic patterns are also clearly present. Likely driven by linguistic similarity and geographical proximity, the romance languages (Spanish, French, Portuguese, Italian, Catalan, etc.) are well connected, with multilingual users crossing between pairs of these languages frequently. Less well-defined on the two platforms, but nonetheless present on each, are crossings between south-east Asian languages involving to various degrees Chinese, Japanese, Korean, and Thai.

This work has examined medium- and larger-sized languages online, and has not looked at languages that have smaller presences currently (although many of these languages may grow quickly in the future as Internet access expands). Language crossings can be reasonably expected between other geolinguistically similar languages such as the Indic languages of the Indian subcontinent. Given the network effects that predicted the central role of larger languages (Crystal, 2003), however, it is difficult to see the central role of larger languages diminishing to any great extent in the near future.

Related to the finding of larger languages being more central, it was hypothesized from previous literature that a larger percentage of users from smaller languages would be multilingual online. This hypothesis was confirmed on Wikipedia, but not on Twitter. It may be that such a relationship does not exist on Twitter (pointing to platform-specific differences) or that methodological limitations prevented detecting the relationship.

In regards to methodological limitations, language classification on Wikipedia is

simple and straightforward: a user either edits a language edition or the user does not. Language classification on Twitter, by contrast, is more prone to error. The machine learning language detection algorithms used in this thesis have greatly increased in reliability and accuracy, but are still far from perfect. As explained in the chapter on Twitter (Chapter 3) and further investigated in Appendix A, care was taken to reduce the number of false classifications (removing usernames/urls, requiring more than one tweet in a language, etc.). Nonetheless, it is possible that the expected relationship between language size and multilingual percentage is present on Twitter but that it is obscured by noise from language classification error.

6.1.4 Multilingualism correlated with activity

Unexpectedly, the data from both Twitter and Wikipedia showed a positive relationship between activity and multilingualism. This can be read as either that multilingual users are more active (i.e., they have more to read, edit, and contribute) or alternatively that the most dedicated and active users of a platform are more likely to contribute content in multiple languages (perhaps without regard to their proficiencies in the languages). The correlation with activity and length of time active suggest that the group of multilingual users defined by and investigated in this thesis overlaps to some extent with the group of power users known to be more active on user-generated content platforms (Priedhorsky et al., 2007; Kittur et al., 2007).

In reality, the set of users defined by and investigated in this thesis as multilingual users is likely composed of multiple subgroups. There is no doubt that genuine proficiency in multiple languages was needed for many of the edits investigated in the Okinawan case study. It would be difficult to imagine an English user with little Japanese experience trying to make even a small correction in Japanese given the large differences in the writing scripts and grammars of the two languages. (In contrast, a Spanish user might feel it was easy enough to fix something small in

a closely related language like Italian or Portuguese.) Machine translation performance between Japanese and English is also still relatively poor, which suggests that the examples of translation and article creation found in the qualitative analysis are acts by users truly proficient in both languages.

At the same time, some of the tasks (such as adjusting interlanguage links[1] and adding or replacing images, or updating website urls) may require less foreign language skill. The most dedicated users may also rely on language tools (dictionaries, machine translation, etc.) to perform small edits with limited foreign language knowledge for some language pairs.[2]

6.2 Implications

6.2.1 Multilingualism offline

Before it is possible to discuss multilingualism offline, it is first necessary to define the term. Linguists generally rely on a similar definition to that used in this thesis: a multilingual individual is someone who uses two (or more) languages (Grosjean, 2010).[3] It is not necessary that such a person possess native fluency in each language. One of the fathers of multilingualism research, Einar Haugen, argued that multilinguals rarely, if ever, have equal and perfect fluency in their languages (Haugen, 1969). Indeed, Grosjean (2010) points out that "[if] one were to count as bilingual only those who can pass as monolinguals in each language, one would have no label for the vast majority of people who use two or more languages regularly but do not have native-like fluency in each" (p. 20).

[1] Interlanguage links exist in the Okinawa case study, but had been removed from Wikipedia and centralized in Wikidata before the time period considered in the global Wikipedia study (Chapter 4)

[2] Note, however, that Wikipedia policy specifically states that "an unedited machine translation, left as a Wikipedia article, is worse than nothing" (`https://en.wikipedia.org/wiki/Wikipedia:Translation`, accessed in January 2015).

[3] The term *bilingual* is often used within linguistics to denote a person who speaks two or more languages (Grosjean, 2010).

Unfortunately, offline data on multilingualism is generally poor (Grosjean, 2010). What is certain is that multilingualism "is present in practically every country of the world, in all classes of society, in all age groups" (Grosjean, 2010, p. 13), and across the world there are more multilingual individuals than monolingual individuals (Birner, 2005; Grosjean, 2010).

National censuses and surveys ask different questions in regards to multilingualism, but some data is available for Europe, the United Sates, and Canada. The most direct assessment of multilingualism was a survey conducted for a European Commission report published in 2012, which found 54% of Europeans spoke a second language well enough to be to hold a conversation (European Commission, 2012). In their questionnaires, the United States Census Bureau and Statistics Canada focus on the languages spoken *at home* (Erard, 2012; Statistics Canada, 2012). As of 2013, the US Census Bureau reported that 21% of Americans spoke a language other than English at home.[4] Statistics Canada reported that the use of multiple languages within the home had risen consistently from 2001 to 2011 and that approximately 18% of the population reported speaking at least two languages at home in Canada in 2011 (Statistics Canada, 2012). These figures for the US and Canada, however, do not include any multilingual individuals who speak a second language exclusively outside of the home, and so the true percentages of the populations that are multilingual are almost certainly larger (Erard, 2012). Based on the number of languages spoken in each country across the world, the highest percentages of multilingualism are probably in Asian and African countries, but data for these countries is most lacking (Grosjean, 2010).

Despite the prevalence of multilingual individuals offline, the online multilingual contributions studied in this thesis were much rarer. There are many possible reasons for this mismatch. First, some of the most linguistically diverse regions of the planet also have low Internet connectivity. For example, Papua New Guinea is the most

[4]USA quick facts from the US Census Bureau, `http://quickfacts.census.gov/qfd/states/00000.html`, accessed January 2015.

linguistically diverse country with 838 languages spoken and a probability of 99% that any two randomly chosen people from the country would have different mother tongues (Lewis et al., 2014). However, the World Bank reports that only 6.5% percent of the population of Papua New Guinea was online in 2013.[5] In addition, not all languages are written, while this thesis has focused on larger, written languages. Beyond these differences and as mentioned previously, the data used in this thesis also does not capture users who read content in one language and contribute in another.

Even so, the data suggests that a much larger number of individuals on user-generated content platforms could contribute in multiple languages. There are many examples of multilingual societies with high levels of Internet penetration (e.g., Quebec and Catalonia), but these offline examples of multilingualism do not appear to transfer easily to online settings: less than 35% of the primary editors of the Catalan edition of Wikipedia edited another edition even though almost all speakers of Catalan would also speak at least some Spanish offline.

Offline levels of multilingualism are driven by many factors, but one chief factor is contact between people of different language groups (Grosjean, 2010). The possibility of contact with people of different languages seems higher online than offline given the world-wide nature of the Internet and the global userbases cultivated by many user-generated content platforms. Nonetheless, this possibility of contact with other-language users and content is shaped in large part by the design of websites. Like Eli Pariser's filter bubbles of personalization (2011), platform design and algorithms may completely hide other-language content from users. Neither personalization nor language filtering is particularly harmful in and of itself—many users may prefer same-language content. Nonetheless, in the same way that over-personalization can reduce the possibility of serendipitous information discovery (Zuckerman, 2013) and remove user-choice (Liljeblad, 2012), so too can overly ag-

[5]The World Bank, Internet users (per 100 people), `http://data.worldbank.org/indicator/IT.NET.USER.P2`, accessed December 2014.

gressive language filtering reduce information discovery and user-choice (especially for individuals who speak multiple languages). Specific avenues for further research on the role of platform design are presented below (Section 6.2.3).

Beyond design, the findings of this thesis in regard to the behavior of multilingual users online also complement offline research on multilingualism. In particular, the finding that many multilingual users edit different different sets of articles in different languages fits well with the idea that multilingual individuals use their languages in different roles and contexts (Grosjean, 2010). Early findings in linguistics documented multilinguals having different language proficiencies in different domains (e.g., language related to work/school vs. home) (Cooper, 1969). Later work also found that multilinguals were able to recall memories and taught knowledge more easily when the language of encoding matched the language of retrieval (Marian & Neisser, 2000; Marian & Fausey, 2006). For example, Russian–English multilinguals remembered more experiences from the Russian-speaking period of their lives when interviewed in Russian and more experiences from the English-speaking period of their lives when interviewed in English than vice versa (Marian & Neisser, 2000). Thus, it may require (marginally) more effort for multilingual users to contribute content for a subject in a language different from which they first learned it.

Finally, studies from linguistics show the importance of online content and information to the future development of languages. Many languages spoken today are considered endangered offline (that is, many languages are likely to have no speakers in the future), and Kornai (2013) found that a substantially larger number of languages (perhaps up to 95% of all languages currently spoken) are at risk of never getting substantial traction online. Digital language prominence has many factors: languages require standardized written forms, input methods, tools like spellcheckers, and the translation of software/website interfaces (Kornai, 2013). The availability of content and information in user-generated content platforms is also critical: the existence of a Wikipedia edition in a language is a key indication of whether the

language will succeed online (Kornai, 2013). Thus, the continued study of multilingualism in user-generated content platforms (and Internet-based platforms more generally) is important to how speakers of smaller languages not yet well-represented online will experience the Internet when they come online for the first time.

6.2.2 Cross-language ties and diversity

Cross-language connections in user-generated platforms were theorized to be similar to weak ties in social networks (Section 2.3). There is a large difference in the information available in different languages online, and this diversity is most acutely felt in user-generated content platforms where users tend to contribute local information (Hecht & Gergle, 2010a). Empirical analyses on Wikipedia (Chapter 5; Hecht & Gergle, 2010b) and Twitter (Hong et al., 2011) quantify some of this diversity and show the differences between the information available in different languages are very large indeed.

Given this diversity in the information available in different languages and the comparatively easier access multilinguals have to information in different languages, multilingual users on user-generated content platforms can be critical links in the network of users transporting information between language clusters. From a network structure view, monolingual users generally cluster closely together and multilingual users are in bridging positions between these clusters. Beyond being in bridging positions, multilingual users do actively transfer information between languages as shown in the Okinawan case study (Chapter 5). In addition, a high rate of multilingualism within a language edition of Wikipedia was associated with less self-focus bias (Chapter 4).

In general, cross-language ties align well with the more general concept of weak ties from social network analysis. Like same language weak ties, the online cross-language ties studied in this thesis generally connect two users who are otherwise embedded in mostly non-overlapping clusters of same-language contacts. While the

overall network remains highly clustered (with most clusters composed of users all sharing the same primary language), cross-language ties provide bridges that connect otherwise distant parts of the network. These bridges can help information move more quickly and easily from one cluster of users to another cluster of users on a user-generated content platform.

The role of multilingual users, however, is also subject to limitations. Some multilingual users engage with completely different topics in different languages, and multilingual users make smaller-sized and less frequent contributions in their non-primary languages even though they make many large contributions in their primary languages. They also contribute more often to the limited set of common concepts that exists in multiple languages rather than to the much larger and more diverse set of concepts that exist in only one language.

Using the concept of bandwidth as a measure of the amount of information that flows between two people in a network from the work of Aral and Alstyne (2011), the limitations of multilingual users for transferring information between languages align well with the idea of cross-language ties (and weak ties in general) having a more limited bandwidth than other ties in a network. Any single cross-language tie is likely to provide less information per unit of time than one same-language tie. Aral and Alstyne (2011) argue that the benefits of having a weak tie versus a strong tie depend on the diversity of information available and how quickly the information changes. The specifics of this trade-off will depend on the particular individuals and their contexts, but the large differences in content available in different languages certainly means that speakers of different languages have a diversity of information available. This may mean that cross-language ties are even more valuable than same-language weak ties in yielding novel information. While the work of Aral and Alstyne (2011) focuses on what information a person knows, the value of cross-language ties is not only what information each person knows, but also to what information each person has access. Multilingual individuals can search out content in multiple languages

more easily than can monolingual individuals. This ability compliments any novel information gained from diverse (e.g., multilingual) information consumption and/or novel information arriving socially via friends as would be expected for all types of weak ties.

Prior research also shows that the large number of (same-language) weak ties in aggregate ultimately transfer more information than that transfered by the smaller number of high-bandwidth strong ties on many-to-many communications platforms (Bakshy et al., 2012). On this point, cross-language ties differ from same-language weak ties in that cross-language ties were found to be much fewer in number than same-language ties on the user-generated content platforms studied in this thesis. Nonetheless, the research of Bakshy et al. makes clear the importance of considering the total amount of information collectively transfered through multiple cross-language ties. Although the bandwidth of each cross-language tie is limited, cross-language ties can still transfer a non-trivial amount of information when considered in aggregate.

Although most, if not all, scholarship on weak ties focuses on same-language weak ties, the research of this thesis shows that cross-languages ties are an important subset of the user connections on multilingual user-generated content platforms. Cross-language ties are generally similar to same-language weak ties, but they are also unique in the ability of multilingual users to search for content in multiple languages and in the fact that cross-language ties are smaller in number than same-language weak ties. However, the number of cross-language connections online (and network structures more generally) may be influenced by the design of a platform in a way that offline cross-language connections are not. These influences of platform design are the focus of the next subsection.

6.2.3 Designing for multilingual users

Within the field of human-computer interaction, researchers have long studied language and culture, most notably within the subfield of internationalization and localization. One guiding design heuristic or rule of thumb has been "speak the users' language" (Nielsen, 1993). This principle captures the need not only to translate user-interfaces, but also issues relating to colors, images, text direction, and date, time, and currency formats (Nielsen, 1993).

The formulation of this heuristic dates from the late 1980's, but it remains a key pillar of internationalization and localization. The original formulation of the heuristic, however, show its age: to "speak the users' language" suggests one language spoken by the group of target users. Most modern Internet-based platforms have the potential to acquire users from a large number of languages and countries simultaneously, and more recent research has focused on designing platforms to work efficiently for users from many backgrounds (e.g., Hsieh, 2014).

Work in computer-mediated communication and human-computer interaction has studied the development of websites and their use by different groups of people across the world even before the advent of user-generated content platforms (see Section 2.2). The rise of user-generated content platforms and the increasing linguistic diversity of Internet users, however, has led to new opportunities and challenges in multilingual design, which the findings of this thesis highlight. As mentioned earlier, there are more multilingual individuals than monolingual individuals offline (Birner, 2005; Grosjean, 2010), but many platforms force a user to select one language in which to use the site. This may explain in part why fewer users exhibit multilingual behavior in user-generated content platforms compared to offline data on multilingualism (Grosjean, 2010). The numbers presented in this thesis represent a lower bound on the amount of multilingual activity occurring given the inaccessibility of data related to reading behavior and the evidence from surveys that more users read/view content in multiple languages than contribute in multiple languages

(Eurobarometer, 2011). Nonetheless, the offline data on multilingualism suggests more users could be encouraged to contribute in multiple languages online, and the research undertaken in this thesis has clearly shown the importance of designing for multilingual users and the potential positive impacts of additional multilingual users.

This thesis has also highlighted the potential to deepen and broaden the contributions of multilingual users online. All the empirical chapters of this thesis have shown that the amount of activity multilingual users perform in their non-primary languages is far lower than the amount they perform in their primary languages. The Okinawa case study further shows that despite the tremendous diversity of information available in different languages, multilingual users are more likely to contribute to other-language content that is easily discoverable from their primary languages (e.g., through cross-language links on Wikipedia) compared to the content found exclusively in their non-primary languages. Design does not fully determine user behavior, but design can certainly influence behavior (Section 2.2). As such, the impact of design changes on the depth and breadth of contributions by multilingual users in their non-primary languages should be invested. This thesis has not performed the type of A/B testing that would be necessary to recommend specific design changes, but the general patterns of behavior observed suggest changes such as highlighting multilingual tasks, making related other-language content easier to find, and rewarding users for cross-language contributions.

The results of this thesis also suggest that multilingual users on user-generated content platforms are not a simple monolithic group. In designing platforms, it will be important to consider use cases and user models that involve power users wanting to perform small actions in another language, truly proficient multilingual users wanting to perform larger actions including authoring and translating content, and multilingual readers wanting to consume content in multiple languages but only contribute content in one language. The exact specific design applications

naturally depend on the particular platform under consideration and will require further research, but some multilingual-friendly design ideas might be ensuring the transparency and customizability of language filters, allowing for different roles of language in search results and friend recommendations systems, and the possibility of surfacing related content in other languages.

The differing roles, motivations, and contributions of multilingual users in user-generated content platforms merit further study. The various contributions made by multilingual users in this thesis suggest there might be a ladder of multilingual contributions similar to the ladders of political or citizen participation in political science (see, for example, Arnstein, 1969). At the base of this ladder would be small and simple actions that do not require high levels of foreign-language proficiency. These actions could include image manipulations and/or updating simple facts and sources across languages. This would fit well with the findings of the relatively larger number of image-related edits in the Okinawa case study (Chapter 5) and with the sharing of images and videos across languages by bloggers found in the pilot study for this thesis (Appendix B).

In the middle level of participation would be tasks that require reading proficiency to perform such as flagging information in one language as out-of-date in comparison to another language and/or identifying and maintaining links between related pieces of content across languages. Wikipedia's interlanguage links are the most obvious example of links between related pieces of content across languages, but arguably many platforms could benefit from similar connections such as linking tags together on a question and answer platform or flagging related groups on Facebook (e.g., the fan groups for a popular Japanese animation studio in English and Japanese on Facebook have no formal links between them currently).

Translation likely sits towards the advanced end of any ladder of multilingual participation for many language pairs currently, which would explain why only a handful of the edits explored qualitatively in the Okinawan chapter (Chapter 5)

were clear examples of translation. However, as machine translation improves and translations environments are better designed, the effort and skill required to translate content into a user's primary language is decreasing. Users may be able to translate content by correcting the output of machine translation algorithms and verifying the general accuracy of the translation. Motivated in part by the work of this thesis, the Language Engineering Team at Wikimedia is creating a content translation tool with exactly this goal. By starting with language pairs such as Catalan and Spanish where machine translation accuracy is generally high, the team hopes to push the effort/skill required for content translation down the ladder of multilingual participation (A. Sharma, Director of Engineering: Internationalization and Localization at the Wikimedia Foundation, personal communication, August 12, 2014). Crowd-sourced translation systems that allow groups of monolingual users to work together and in conjunction with machine translation systems to iteratively perform high quality translations are another example of lowering the level of foreign-language proficiency required for translating content (e.g., one such system is MonoTrans2 discussed in Section 2.2.2). Nonetheless, the poor performance of machine translation for many language pairs suggests that translation will continue to be at the high end of any ladder for some time (particularly translation from a primary language into a non-primary language given the high bar of foreign-language proficiency required).

As mentioned in the subsection on offline multilingualism, many languages are at risk never gaining traction online (Section 6.2.1). Platform design can play a large role in determining the languages in which users contribute content and thus could effect the ultimate success or failure of a language to survive online. My current work stemming from this thesis is investigating online review sites. The network-effects theory that predicted the presence of more multilingual users in smaller-sized languages in this thesis also suggests that if a speaker of Swahili and English is forced to see reviews in only one language, that speaker would likely opt to see reviews

in English if there are a smaller number in Swahili. This would create a feedback loop where the perceived lack of an audience to read user-contributed content in Swahili further discourages other speakers of Swahili and English from posting in Swahili. This is an artificial constraint, however. If multilingual users were allowed to see reviews in all their languages (e.g., in both Swahili and English), then the user considering writing a review in Swahili could be confident that her review would be seen by all Swahili speakers (including those that also speak another language).

6.2.4 Personalization, machine translation, and other algorithms

Multilingual information discovery is most prominent within the field of information retrieval in computer science. Approaches to cross-language information retrieval (CLIR), however, have focused mostly on monolingual users (Ghorab et al., 2011). Enabling the flow of more information across language divides is not an either–or choice between cross-language technologies and human multilingual users. Rather, the greatest flow of information across languages is likely to be accomplished through designing platforms that influence user behavior and create a symbiotic relationship between algorithms and users. The aim of this thesis was to understand the behavior of multilingual users on user-generated content platforms. Nonetheless, many of the design areas recommended for further investigation also apply more fundamentally to the algorithms used for content and friend recommendation and search personalization across a wide array of web platforms.

Just as interface designers need to consider language and multilingual users, algorithm designers too must consider these aspects. Multilingual user models have been considered in theoretical information retrieval work, but not yet put into practice (Ghorab et al., 2011). Personalized, multilingual information retrieval systems could take the interests and language abilities of users as input in ranking results. Thus, results in the language(s) (perhaps more than one) that the user understands

can be prioritized and translation can be offered only for the languages which the user does not know (Ghorab et al., 2011; Steichen et al., 2014). This thesis has shown that same-language activity is most common, but that a very active segment of users on user-generated content platforms is also contributing in multiple languages. Thus, there is a real potential demand for multilingual search on these platforms.

In line with the offline linguistics research mentioned earlier, the findings of this thesis also suggest that not all multilingual users are equally proficient in their non-primary languages. Open-ended questions on a survey completed by multilingual users also found that some users only used their non-primary languages for watching multimedia content (Steichen et al., 2014). Thus, personalized multilingual information retrieval systems may consider personalizing output based not only on users' languages, but also on users' proficiencies in each language. For example, it might be appropriate to only offer image or video results from one language for a given user.

Supporting the findings from linguistics that multilinguals have different areas of proficiency in different languages, the same survey also found that many users reported browsing and searching for different topics in different languages, although there were also topics they browsed and searched for in multiple languages (Steichen et al., 2014). This fits with the differences in concepts edited across language editions on Wikipedia by some users found in this thesis. However, this thesis also finds that some users edit the same concepts across languages, and thus further research is necessary to understand how topic should be incorporated into models of personalized multilingual information retrieval. Open-ended responses to questions in the survey also suggested that some users might engage with more topics in other languages, but that they lacked the language proficiency to formulate accurate queries in their non-primary languages at times (Steichen et al., 2014).

Beyond information retrieval, the results of this thesis also have implications

for efforts in natural language processing to train machine translation systems with user-generated content. Such efforts aim to discovery instances of human translation in user-generated content sites and use these examples of human translation as training data for machine translation systems (e.g., Potthast, Stein, & Anderka, 2008; Dalton, 2012). The initial results of this thesis are not overly promising for such efforts as multilingual users are a small percentage of users, but the findings of this thesis do give some hints as to approaches for finding translated content. Multilingual users were very active, and a larger percentage of users on smaller-sized language editions of Wikipedia were multilingual in comparison to larger-sized editions. These smaller-sized languages are often the languages for which traditional corpora are lacking, and thus there is potential in mining content in these languages. In accordance with the networks of language crossings, the most success will probably be found mining text between English and small languages with a high number of multilingual users. The findings of this thesis also suggest image captions and/or other text content related to multimedia items may be more likely to contain translated text as multilingual users often edited these items in their non-primary languages.

Overall, the findings of this thesis give an indication of the rich level of information that is available through studying user behavior. Combining user behavior information with traditional text similarity metrics has the potential to improve efforts to find translated text in user-generated content platforms. The improvement of machine translation algorithms using data mined from user-generated content platforms in turn offers the possibility of making improved machine translation features available to users on user-generated content platforms. Improvements in machine translation, text similarity, and other methods will also open further possibilities for studying multilingual user behavior on user-generated content platforms, as discussed in the next section.

6.3 Next steps

The research presented in this thesis has been made possible through novel, publicly accessible trace data about user behavior on Twitter and Wikipedia. Trace data is not particularly new to the field of human-computer interaction nor to computer-mediated communication more generally (Section 2.2.3). However, the public availability of this data and the increasing amount of activity occurring online have created new possibilities. For example, petition signing previously occurred entirely offline, while now most petition signing occurs entirely online: this has made it possible to analyze petitions that would not have left enough of a historical imprint to be included in research datasets previously (Hale, Margetts, & Yasseri, 2013; Yasseri, Hale, & Margetts, 2014; Margetts, John, Hale, & Yasseri, 2015)

The use of trace data to study of multilingualism online is heavily reliant on the algorithms from natural language processing, network/graph studies, and other areas of computer science. Increases in cross-language topic analysis, geolocation, and language detection will enable new opportunities for research on multilingual user behavior using trace data. Advances in cross-language topic analysis could allow for large-scale, quantitative study of the topics contained in the text contributed by users. Such advances would then allow for the investigation of language and topic beyond platforms that clearly connect similar concepts across languages (i.e., for platforms without the equivalent of the interlanguage links on Wikipedia used in this thesis). Given the short and informal nature of text on many user-generated content platforms, improved language detection methods are needed that work well on such text. Language detection methods are also needed to detect the mixing of multiple languages and creoles reliably (Appendix A).

Looking ahead within the field of human-computer interaction, the transition towards mobile devices, especially smartphones but also tablets and wearables (e.g., smart watches and optical head-mounted display like Google Glass) is resulting in a large shift. A first question for operators of user-generated content platforms is how (and

if) users will contribute from mobile devices and how mobile device contributions might change the nature of contributions and/or the userbase itself. Mobile devices hold the possibility of increased userbase engagement through use in more contexts when users are not at their desktop/laptop computers (Leung & Wei, 2000; Perry, O'hara, Sellen, Brown, & Harper, 2001; Humphreys, Von Pape, & Karnowski, 2013) and also of increased userbase diversity given the lower prices of and less skill required to use mobile devices (Stork et al., 2012). To date, the largest change has been felt in increased engagement, while the geographic diversity of userbases has remained relatively stable (Keyes & Hale, 2014).

As design principles are evaluated and adapted to mobile interfaces, the role of multilingual users will need further consideration and evaluation. While many websites offer the ability to relatively easily switch between languages using a menu on the site, many native mobile applications offer only the language chosen in the operating system for the whole mobile device. This language may not be the most appropriate for a multilingual user depending on the application. Likely even more of an issue for multilingual users, however, is when features or content are hidden simply because they are not available in the selected language (as in the case of other-language reviews of apps in the Google Play store mentioned in Chapter 1).

This thesis did not analyze the differences between mobile and desktop users specifically, although in general most Twitter messages are sent from mobile devices while most Wikipedia edits are performed with desktop/laptop computers (Keyes & Hale, 2014). Most of the findings from this thesis stem from the interactions of users on user-generated content platforms, and thus most of the findings likely apply across mobile and desktop interfaces. Nonetheless, the intersection of mobile interfaces and multilingual user-generated content platforms is an important area for further study.

As the first multiplatform investigation into the roles of platform design and social factors in the spread of information online between speakers of different lan-

guages, this thesis has contributed to both theory and practice in this emerging area. Further work will need to be undertaken not only in academia, but also in industry settings by private companies with access to user data (e.g., by ISPs, platform operators, and advertising-related companies). Beyond the use of trace data, online experiments offer an exciting opportunity to move beyond correlations into causation. A key goal of this study has been to show the importance of further research into multilingual activity and highlight areas for possible A/B experimentation, which can most easily happen in industry settings. The publications from this thesis have already inspired (and help justify internal funding for) the Language Engineering team at the Wikimedia Foundation to build a content translation tool to help users translate content between different language editions of Wikipedia. I am in close contact with the team and hope to collaborate on an A/B test involving one of the Wikimedia platforms in the future around the theme of cross-language content discovery. It is my sincere hope that the publications stemming from this thesis along with my future work will encourage more research in this area and also result in practical changes concerning the multilingual design of modern web platforms.

Appendix A

Language and Geographic Identification on Twitter

The work presented in this appendix chapter evaluates the performance of popular off-the-shelf language detection and geolocation services on Twitter data. The findings of this study influenced the language detection methods used within Chapter 3 studying Twitter. The results also underscored the challenges with geolocation, and reinforced my decision to focus exclusively on language within this thesis. I completed this work with Mark Graham and Devin Gaffney in 2012, and the materials in this appendix have appeared in publication as:

Graham, M., Hale, S.A., Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*. doi:10.1080/00330124.2014.907699.

All three authors designed the study, performed the qualitative coding, and wrote the paper. Mark secured funding; Devin collected the data; Devin and I led the evaluation of geolocation services; I led the evaluation of language identification algorithms.

Microblogging services such as Twitter allow researchers, marketers, activists, and governments unprecedented access to digital trails of data as users share information and communicate online. Patterns of information exchange on platforms that rely on user-generated content have been used recently in scholarly research about community (Gruzd, Wellman, & Takhteyev, 2011), information diffusion (Romero, Meeder, & Kleinberg, 2011), politics (Bruns & Burgess, 2011), religion (Shelton, Zook, & Graham, 2012), crisis response (Zook, Graham, Shelton, & Gorman, 2010; Palen, Vieweg, & Anderson, 2010), and many other topics. Such data are also important to governments and marketers seeking to understand trends and patterns ranging from customer and citizen feedback to the mapping of health pandemics (Graham & Zook, 2011). Twitter, in particular, with its large and international user base (there are now more than 350 million users on the platform), has been the source of much scholarly research.

Content passed through Twitter remains decontextualized, however, unless we find ways to reattach it to geography. In other words, we do not just want to know what is said, but we also want to know where it is said and to whom it is said. As such, the attributes of language and location are crucial for understanding the geographies of online flows of information and the ways that they might reveal underlying economic, social, political, and environmental trends and patterns. Yet, both language and location are challenging to deduce in the short messages that pass through Twitter, and no well-accepted methodology for their extraction and analysis has been articulated. This point is especially salient because of the increasing number of studies, journalistic accounts, and real-world applications that rely on harvested locational and language data from Twitter. Therefore, to provide a useful starting point for future research on Twitter (and indeed other microblogging platforms), this article compares several approaches to working with linguistic and geographic information in Twitter to better understand the strengths and limitations of each approach.

Figure A.1. Screenshot from Barack Obama's Twitter profile page.

The short size of posts (140 characters on Twitter) presents a challenge to accurate language identification due to the fact that most language identification algorithms are trained on larger sized documents (Carter et al., 2011). In addition, the style of writing on Twitter using abbreviations and acronyms complicates language classification. In many instances, researchers have simply relied on the user interface (UI) language of a user's account or used an off-the-shelf language detection package without consideration of its suitability for use on short, informal text phrases. The disagreement of several studies on the most used languages in Twitter (Honeycutt & Herring, 2009; Semiocast, 2010; Hong et al., 2011; Takhteyev et al., 2012) highlights the difficulty of language detection. All four studies agree that English is the most used language but give percentages ranging from 50 percent (Semiocast, 2010) to 72.5 percent (Takhteyev et al., 2012). The purpose of our work is not to study the prominence of different languages on the platform but rather to highlight important methodological issues related to language identification for future research to more critically engage in geolinguistic analyses.

Accurately determining location in messages sent through Twitter is also a significant challenge. The most apparent method is to consider the profile information that is directly provided by a user (e.g., the text "Washington, DC" in Figure A.1)

in response to an account setup question: "Where in the world are you?" This question, though, which allows users to input any text string to describe their location (referred to in this article as profile location), is often hard to geolocate correctly (the open-ended text could just as easily say "Edinburgh, Scotland," "Barad-dûr, Mordor, Middle-earth," or simply "here"). High error rates, missing data, and non-standardized text in profile locations have forced some researchers wishing to employ this geographic data to use smaller samples and labor-intensive manual coding of profile locations (e.g., Takhteyev et al., 2012).

An alternate approach that some researchers have adopted is to narrow their samples to only use geocoded tweets. Depending on the user's privacy settings and the geolocation method used, these tweets have either an exact location specified as a pair of latitude and longitude coordinates or an approximate location specified as a rectangular bounding box. This type of geographic information (referred to in this article as device location) represents the location of the machine or device that a user used to send a message on Twitter. More precisely, the data are derived from either the user's device itself (using the Global Positioning System [GPS]) or by detecting the location of the user's Internet Protocol (IP) address. Precise coordinates are almost certainly from devices with built-in GPS receivers (e.g., phones and tablets). Bounding boxes, however, can result from privacy settings applied to GPS data or from GeoIP data.

Irrespective of these limitations, device locations are challenging for users to manually manipulate and, because they are structured data, are easily interpreted by computers. Only a small portion of users publish geocoded tweets, however, and it is unlikely that they form a representative sample of the broader universe of content (i.e., the division between geocoding and nongeocoding users is almost certainly biased by factors such as socioeconomic status, location, and education). From a sample of 19.6 million tweets collected by the authors (these data were collected using Twitter's "statuses/sample stream" collection method with "spritzer access")

over nineteen days in June 2011, only 0.7 percent of tweets contained structured geolocation information. As such, the extremely low proportion of information with attached device locations means that researchers either have to work with data that are likely highly skewed or devise effective methods to work with the profile location that is attached to all of the tweets that do not contain explicitly geocoded device location information.

This article deals with these gaps of knowledge related to language and location in two primary ways. First, it explores the accuracy of a range of language detection methods on tweets, which, by definition, are short and often contain informal phrasings and abbreviations. It identifies common sources of errors and compares performance over four research locations, each including a large variety of languages. Second, it compares various location information within tweets (profile location, device location, time zone information) and the accuracy with which geolocation algorithms can interpret the free-form profile location information.

In performing this work, we are able to refine methods that can be employed to map and measure the geolinguistic contours of people's information trails on Twitter. Doing so will ultimately allow future work to build on this research to create more accurate and nuanced understandings of the clouds of digital information that overlay our planet.

A.1 Related work

A variety of methods have been employed in looking at Twitter's geolinguistic contours. Hong, Convertino, and Chi (2011) used two automated tools to determine the language of a tweet—LingPipe and the Google Language application programming interface (API)—whereas Semiocast (2010) used an internal proprietary tool. Carter, Tsagkias, and Weerkamp (2011) and Gottron and Lipka (2010) discussed several of the challenges with language identification of short texts, the largest being

that most language detection algorithms have been developed and trained on full documents that are longer and better formulated than the short text snippets that pass through Twitter. Carter, Tsagkias, and Weerkamp (2011) focused on microblog posts and developed two approaches (priors) to enhance performance: a link-based approach to consider the language of linked-to content and a blogger-based approach to aggregate tweets on a per account basis to form a larger document to classify. They found that both approaches improve accuracy but still leave room for further improvement. Pilot work for this thesis (Appendix B) used the Compact Language Detection (CLD) kit, part of Google Chrome, for detecting the language of blogs in conjunction with the presence of certain key words. He found these two methods in combination to be 95 percent accurate on a sample of 965 blogs about the Haitian earthquake. The CLD has since been used in creating visualizations of language on Twitter (Fischer, 2011), but its accuracy has not yet been evaluated for short posts passed through Twitter.

Whereas geographic metadata in device locations (i.e., precise coordinates) are unlikely to be subject to much debate about their validity, the self-reported profile location field in a user's profile is problematic because of its unstructured nature. It remains, however, that the usage of profile locations is often contemplated in papers that discuss the virtual data shadows to geographically bound situations such as the Arab Spring of 2011 or the Iranian election protests of 2009 (Gaffney, 2010; Hecht, Hong, Suh, & Chi, 2011; Lotan et al., 2011). Takhteyev, Gruzd, and Wellman (2012) attempted an automated coding of profile locations with an unnamed tool but ultimately decided to hand-code profile location details due to high error rates. Vieweg et al. (2010) also hand-coded profile locations but also manually used tweet content in addition to profile content in determining the user's physical location. Java et al. (2007) used the Yahoo! geocoding API, which attempts to assign a precise location to self-reported profile locations. The accuracy of such geocoding algorithms to profile location data on Twitter has not been previously determined,

however. Most important, Hecht et al. (2011) showed the need for great caution in finding that only 66 percent of the Twitter profiles they examined by hand had valid geographic information, 18 percent were blank, and 16 percent had nongeographic information, mostly consisting of popular culture references. As a result, geocoding APIs will likely struggle with this input. In contrast to the free-form nature of the profile location, Krishnamurthy, Gill, and Arlitt (2008) opted to use time zone (Universal Time Coordinated [UTC] offset) information in a user's profile to get a user's local time and thereby approximate longitude. Although it is impossible to determine latitude using this method, such a strategy can still improve our best guesses about profile locations. It is unclear, however, how many people actually set an accurate time zone. This is particularly a concern for users who employ third-party clients instead of visiting the Twitter Web site itself (within our sample fewer than 50 percent of tweets are created on Twitter's own Web site).

Newer research is developing methods to locate users based on the text content of their tweets, the time of day users tweet at, and the location of the users they are following or are followed by (Cheng, Caverlee, & Lee, 2010; Eisenstein, O'Connor, Smith, & Xing, 2010; Hecht et al., 2011; Wing & Baldridge, 2011; Mahmud, Nichols, & Drews, 2012; Sadilek, Kautz, & Bigham, 2012). All of these approaches, however, have only been developed and evaluated using tweets in the English language or geocoded tweets from the United States. This article does not consider these developing approaches but evaluates two off-the-shelf geocoding services and assesses their accuracy and performance across four different regions, only one of which is in the United States. In the manual examination of profile locations, the article also raises hints of possible challenges these newer approaches will have to overcome to be geographically and linguistically broader. The article also provides important insights into the disaccord between profile and device locations, which is important for the data used to develop and test these new approaches.

In sum, it is important to be aware of the myriad, overlapping, and complex

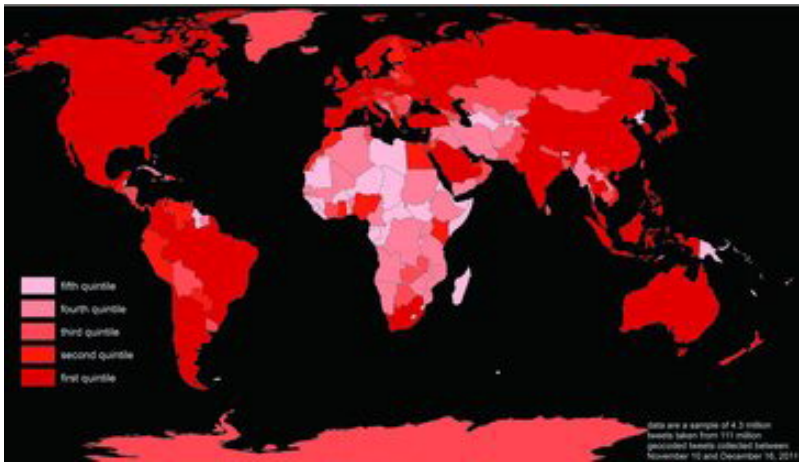


Figure A.2. Number of geotagged tweets per country between 10 November 10 and 16 December 2011.

ways in which location is ascribed to information in Twitter before attempting to employ it in any geographic analyses. The following section more closely examines the accuracy and sources of error in a range of methods used to extract location and language from Twitter to build on existing work and more clearly articulate how this information collected from Twitter might be of use for research in geography.

A.2 Methods

Between 10 November and 16 December 2011, 111,143,814 tweets were collected by using the statuses/filter method of Twitter's streaming API (these data are mapped in Figure A.2). The method allows tweets to be collected from within a user-specified bounding box, which was drawn as a 180- by 360-degree box in this study (or a bounding box that encompasses the whole planet). Tweets sampled this way from the streaming API only include tweets with an explicit GeoIP or GPS device location.[1] The search API, which might make guesses as to users' locations, was not used in this study. Although some rate limiting errors are met when a large box is built (Twitter defines a maximum data rate, and any additional data above that

[1]This method captures tweets that are geocoded by both IP addresses and GPS-enabled devices.

limit are dropped from the stream), this effect is measurable. Rate limiting was only noticed during times that were coincident with some North American weekday peak hours, and only 1.1 percent of our files were ultimately affected by any significant errors of data. Data were otherwise collected constantly with a few intermittent and brief crashes, and all downloaded information was stored in tab-separated files.

From our sample of 111 million tweets, 1,000 tweets were randomly selected from each of four metropolitan areas[2] (Cairo, Montreal, San Diego, and Tokyo). These areas were selected by the research team as research sites that are characterized by interesting geographic, linguistic, and cultural differences. To avoid overrepresentation by heavy users, we only include a maximum of one tweet per user in the sample. The location of every tweet was determined by the device location (GeoIP/GPS) recorded by Twitter.

Rather than relying on one particular algorithm, library, or toolkit to establish a higher degree of accuracy, validity, or certainty about the data, a central aim of this article is to compare existing solutions. As such, the research team reviewed existing available geolocation and language identification solutions and evaluated them on their ease of use, throughput, and thoroughness. Based on these results, three language detection and two geolocation services were ultimately selected (Table A.1).

Custom scripts employing each of the automated language identification algorithms (Alchemy, the CLD kit, and Xerox) and each of the geolocation solutions (Google and Yahoo!) were written and the results of these algorithms were stored in a database. Other services certainly exist and the services considered in this article are not exhaustive or authoritative, but these services currently constitute the more easily implemented off-the-shelf solutions that are available. In particular, since

[2]The bounding boxes that we used to define the four urban areas are as follows: Cairo (31.1, 29.95, 31.54, 30.28), Montreal (–74, 45.33, –73.35, 45.78), San Diego (–117.3, 32.43, –116.74, 32.9), and Tokyo (139.3, 35.4, 140.2, 35.9). In all cases, outer ring roads were used to determine the approximate extent of each city's conurbation. Although this approach is relatively imprecise, we deemed it equally problematic to establish a consistent bounding box size for all sample cities (due to the significantly different sizes of the four urban areas).

Service	Type	Use	Throughput	Thoroughness
Compact Language Detection kit	Language detection	C/C++ with a Ruby library and Python bindings available	No limits: Service is executed locally	161 languages
Alchemy API	Language detection	Web service	1,000 requests per day per API key	97 languages
Xerox open source	Language detection	Web service	No discernible limits	46 languages
Twitter UI language	User-selected value	Delivered with tweet	N/A	33 languages
Google Geocoding API	Geolocation	Web service	2,500 requests per day per IP address	
Yahoo! PlaceFinder	Geolocation	Web service	50,000 requests per day per API key	

Table A.1. Language and location services overview

Google switched its Language Detection API to a paid service, many researchers and companies have tried the CLD kit, but it has not been compared with other solutions. CLD also forms part of the language-detection augmentation offered by DataSift, a Twitter data reseller, and hence is likely used by many commercial companies working with Twitter data. All of the language detection algorithms surveyed are pretrained and immediately usable for any piece of text. This allows comparison apart from the specific data used to train the algorithm.

To test language detection tools, we also randomly selected 1,000 tweets from each of our four study regions. These messages were then manually coded by the study's authors for the primary language of the tweet and disagreements were resolved through discussion. The study's authors collectively have experience with Arabic, English, German, Japanese, Korean, Mandarin, Persian, Spanish, and Thai. Where words from multiple languages were found in a single tweet, the tweet was coded for the most abundant, primary language of the tweet.

	Humans		Algorithms	
Research Site	Coders	Fleiss' Kappa	Without UI Language	With UI Language
Cairo	3	0.814	0.525	-
Montreal	3	0.779	0.587	0.461
San Diego	3	0.868	0.513	0.440
Tokyo	3	0.724	0.485	0.329
Overall	3	0.888	0.630	0.579[†]

Table A.2. Human and algorithm agreement on language (Fleiss' kappa) with and without user interface (UI) language.
[†]Arabic was not available as a UI language choice at the time of the study; so, Cairo is not included in the overall statistics.

A.3 Findings

A.3.1 Language

When examining the manual coding of language in the 4,000 sample tweets, overall intercoder agreement was high between the human coders (as shown in Table A.2). Agreement was calculated with Fleiss' (1971) kappa. The measure ranges from -1, complete disagreement, to 1, perfect agreement, and, compared with percentage agreement, better accounts for agreement between coders that could occur by simple chance (Gwet, 2010). Conflicts between human coders were generally due to multiple languages being used within a single tweet. Tweets containing auto-generated text (e.g., automatically generated messages from Foursquare) often contained this mix of languages. In comparison to the human coders, the agreement between different language detection algorithms can be seen to be much lower (Table A.2).

If the agreed human coding is treated as a gold standard, the CLD kit and Alchemy matched human classifications most closely, although all methods are with one standard deviation of each other (Table A.3). Alchemy in general performed better than CLD with the significant exception of Tokyo. In the best case, Alchemy agreed with human coders on 91 percent of tweets in San Diego. Even this, however,

Research Site	Alchemy	Xerox	CLD	Twitter UI Lang
Cairo	0.510 (69.6%)	0.374 (55.9%)	0.464 (61.7%)	-0.215 (44.9%)‡
Montreal	0.653 (83.1%)	0.548 (73.9%)	0.550 (74.7%)	0.462 (75.6%)
San Diego	0.644 (90.9%)	0.501 (84.0%)	0.487 (82.1%)	0.565 (90.6%)
Tokyo	0.017 (56.2%)	0.029 (57.0%)	0.418 (87.2%)	0.278 (83.0%)
Overall	0.609 (75.0%)	0.534 (67.7%)	0.670 (76.4%)	0.714 (83.1%)†

Table A.3. Algorithm agreement with human coders on language: Fleiss' Kappa/Scott's PI with percent agreement in parentheses.
The standard deviation on percent agreement is about 0.5 in all cases.
‡Arabic was not available as a UI language choice at the time of the study; so, this value should be interpreted with caution.
†Arabic was not available as a UI language choice at the time of the study; so, Cairo is not included in the overall statistics.

only translates an intercoder agreement with a Fleiss' kappa of 0.644, which better accounts for agreement that could occur by simple chance as explained previously. The best overall score was achieved by the CLD, which agreed with the human classification in 76.4 percent of cases. This translates to an intercoder agreement with a Fleiss' kappa of 0.670.

Yet, both of these scores are much lower than the overall intercoder agreement between the human coders, which had a Fleiss' kappa of 0.888. The relatively high percentage agreement scores and lower kappa scores suggest that disagreement between the algorithm and human coders occurred more about less frequently appearing languages. Nevertheless, the CLD is more apt for large data sets, as it is the only local, fully offline method considered here. The code for CLD is also open source and could be adopted, although the training corpora used to create the language identification fingerprints are unknown and unavailable.

Analysis shows that the CLD performed particularly well in differentiating text in different Asian scripts. CLD and Alchemy did not do so well in Cairo, where a number of Arabic-language tweets were written in the "Arabic chat alphabet" (i.e., Arabic using Latin characters). Whereas CLD nearly always classified text in Japanese, Korean, Chinese, and Arabic correctly when these languages were written

		Yahoo!			Google		
Research site	Blank	In box	Out of box	Failed	In box	Out of box	Failed
Cairo	201	431	336	32	444	295	60
Montreal	138	469	356	37	561	224	77
San Diego	133	446	359	62	407	348	112
Tokyo	170	456	335	39	419	197	214
Overall	642	1,802	1,386	170	1,831	1,064	463

Table A.4. Results of the geolocation of 1,000 randomly selected user profiles from each research site.

in their usual scripts, it failed in all eighty-nine cases to classify Arabic written with Latin characters correctly. Indeed, all of the language identification algorithms considered here failed to accurately classify these messages written in the Arabic chat alphabet.

It should also be pointed out that the UI language of Twitter users was a useful indicator of language in some research sites. It corresponded with the human coding of language for more than 75 percent of the tweets from Montreal, San Diego, and Tokyo. At the time of data collection, there was not an option to set the Twitter UI language to Arabic, which likely explains why the UI language of users in Cairo only agrees with human coders for 45 percent of the tweets collected there.

Overall, language identification of tweets is difficult for human and machine coders alike. One preprocessing step that could improve results is to remove auto-generated text and non-language-specific text (e.g., emoticons). It will be important to train machine algorithms on informal scripts (e.g., Arabic chat alphabet) in addition to classical scripts. The suitability of off-the-shelf language identification packages and the appropriateness of the UI language setting vary by research site, so the best algorithm will likely depend on the specific research questions and study location.

A.3.2 Geolocation

To better understand how self-reported profile locations might be used to map the geography of information in Twitter, the Yahoo! and Google geolocation algorithms were applied to 1,000 randomly selected users from each research site. Ideally, this study would apply the geolocation algorithms to the entirety of the users in each site, but rate limitations (i.e., the number of allowed requests per minute) prevent this from completing in a reasonable time frame. We found that 16 percent of the location fields in our sample of 1,000 users from each research site were blank, which is similar to the 18 percent of profiles that Hecht et al. (2011) found blank in a general (geocoded and nongeocoded) sample of Twitter, although it is much lower than the 28 percent of profiles that (Cheng et al., 2010) found blank in their general sample of Twitter profiles. Overall, Yahoo!'s and Google's geolocation algorithms perform similarly (Table A.4). Excluding the blank profile locations, 94.5 percent of attempts using Yahoo!'s PlaceFinder and 86.2 percent of attempts using Google's Geocoder placed the user in some location. Many of these locations were outside of the bounding boxes defining each research site, however. On average, only 53.7 percent of attempts with Yahoo! and 54.5 percent of attempts with Google placed the user within the bounding box from which they originally tweeted. If blank profiles are included, these percentages drop to 45.0 percent for Yahoo! and 45.8 percent for Google, which would be closer to the likely upper bound on the actual percentage of users in a general sample from Twitter that could be placed correctly by geolocation algorithms alone.

Besides returning a geographic position, each geo-location service might report that it failed to geolocate the input. Comparing the values, it is clear that although Yahoo!'s algorithm geocodes more profile locations than Google's algorithm, many of these additional locations do not fall within the bounding boxes. Google's algorithm tends to fail to geocode a larger number of profile locations compared to Yahoo!. Of the locations that Google's algorithm does geocode, however, more of these locations

	Yahoo!	Google
Profile location blank	642	642
Geolocated within bounding area	1,802	1,831
Geolocated outside bounding area	1,386	1,064
Geolocated within bounding box by other algorithm	206	107
Identified as within bounding area by human coder	72	39
Not within bounding box (human coder)	481	462
More general (e.g., country, state, region)	282	269
Multiple locations including within bounding area	56	39
Latitude, longitude pair	21	10
Invalid/generic (e.g., la la land, earth)	268	138
Failed to geolocate	170	463
Geolocated within bounding box by other algorithm	5	75
Identified as within bounding area by human coder	4	37
Not within bounding box (human coder)	1	18
More general (e.g., country, state, region)	4	16
Multiple locations including within bounding area	1	18
Latitude, longitude pair	102	113
Invalid/generic (e.g., la la land, earth)	53	186

Table A.5. Human analysis of profiles failing to geolocate or geolocating outside of the relevant bounding box.

Research site	All users captured			Geocoding user sample		
	N	City-specific time zone	Correct UTC offset	N	City-specific time zone	Correct UTC offset
Cairo	1,952	54.6%	55.4%	1,000	55.5%	56.3%
Montreal	5,235	41.7%	57.0%	1,000	40.1%	57.3%
San Diego	9,292	57.1%	60.5%	1,000	58.6%	63.2%
Tokyo	55,573	68.2%	72.3%	1,000	70.1%	73.8%
Overall	72,052	64.4%	69.2%	4,000	56.1%	62.7%

Table A.6. Time zone information has been seen as another proxy for location; however, this information is not routinely provided by all users.

are within the bounding boxes of the research sites than Yahoo! (Table A.4). This is particularly apparent in the Tokyo research site, where Google declines to determine a location for many more profile locations than Yahoo!.

Profile locations outside of the relevant bounding box might be due to users tweeting from a different location than that written in their profiles or due to geocoder error. To resolve this ambiguity, the authors manually examined all user locations that failed to geolocate or that geolocated outside of the relevant bounding box with both Google and Yahoo! (Table A.5). The largest portion (35.6 percent) of these profile locations were legitimate geographical locations outside of the bounding boxes. This suggests that users do not update their profile locations with great frequency. The second largest portion (24.1 percent) was nongeographic text (e.g., "Neverland") or generic, nonspecific locations (e.g., Earth, a peach orchard). After this, a large portion (21.2 percent) was more general geo-graphic locations that included the relevant research site (e.g., Japan, California, or Kantō [the eastern half of Japan, including Tokyo]).

The analysis also suggested ways to improve geo-location accuracy. Of the 5.8 percent of locations that were actually within the bounding boxes, abbreviations of place names was the most common reason for the geolocator to fail. Beyond this, another 4.2 percent of profile locations had multiple locations, one of which was the relevant study area. Finally, 9.1 percent of the tweets actually had latitude and

longitude coordinates in the profile location field along with additional text (usually the name of an app placing the information in the profile location). Google and Yahoo! recognized latitude and longitude coordinates without additional text, but any additional text causes both geocoders to fail (or in a smaller number of cases to geolocate to a location that did not correspond with the coordinates in the profile). All three of these situations—abbreviations, multiple locations, and latitude and longitude coordinates—could likely be handled by preprocessing the data for these possibilities. This is especially applicable when targeting a single area, where a list of likely abbreviations might be more easily created.

One vital caveat to these results is that the researchers coded data in the most naïve form available. For both Yahoo!'s PlaceFinder and Google's Geocoder, additional options exist that might increase the accuracy of results. Yahoo!'s PlaceFinder returns multiple locations ordered by relevance for a given string (multiple locations can be, and are, returned routinely, such as Oxford, UK, and Oxford, Mississippi, when using a string of "Oxford"). This relevance, or confidence score, is between zero and one hundred and is shown with every returned result. This confidence score could be used to reject all results when every result is below a certain threshold. This would increase the number of locations that fail to geocode at all but would likely raise the overall accuracy of profile locations that do geolocate. Furthermore, Google's API allows researchers to set a location hint to better capture data (potentially) originating from the region of focus.

Given the significant difficulties associated with geo-locating profile locations, time zones have also been seen as a more reliable metric for approximating location (Krishnamurthy et al., 2008). Our data set, however, suggests that many users have incorrectly configured time zones in their profiles (Table A.6). Users select their time zone on the Twitter site from a predefined list. Several options have the same UTC offset (e.g., for UTC+9 a user can select Tokyo or Seoul) although the daylight savings or summertime rules might differ. It is likely that some users are traveling or

have purposefully set different time zones from the locations in which they are using the service; however, the fact that only 57 percent of users tweeting in Montreal had set an east coast time zone (much less the specific option of Eastern Time [United States and Canada]) indicates that many users likely do not set their time zone correctly. In addition, many users tweeting from Montreal had other UTC+5 time zones selected (e.g., 231 users in our sample of 1,000 users had set their time zone to Quito), suggesting that caution is needed in interpreting the time zone more specifically than the UTC offset. Across all research sites, 69.2 percent of users had selected a time zone with a UTC offset that corresponded to the device location information in the tweet (Table A.6). The low number of users correctly setting their time zone might be influenced partially by the large number of third-party client devices used to tweet. Overall, only 23.6 percent of tweets captured across all our research sites were published via the Twitter Web site, with the remainder sent using third-party applications. Geocoded tweets might be more likely to be sent from a third-party application than nongeocoded tweets, however; so the percentage of tweets sent from third-party applications in our sample of only geocoded tweets is likely higher than the percentage would be in a general sample of tweets including both geocoded and nongeocoded tweets.

Ultimately, our findings related to location point to the significant challenges associated with automatically identifying geographic references in unstructured text. It is important for researchers to be aware of these difficulties if they want to move beyond the limitations of only relying on the unrepresentative amount of information tagged with device locations.

A.4 Discussion and conclusions

Over 300 million users publish hundreds of millions of short messages every day on Twitter. As a result, this content has been used by researchers from fields as diverse

as epidemiology, politics, marketing, and geography to better understand, map, and measure large-scale social, economic, and political trends and patterns. Much of this analysis is carried out with only limited understandings of how best to work with the spatial and linguistic contexts in which the information was produced, however. As such, it has been necessary to study the reliability of key methods used to determine language and location of content in Twitter.

This article found that there are significant challenges to accurately determining the language of tweets in an automated manner. None of the language identification methods tested in this article is able to match the accuracy of human coding by multiple coders.[3] The informal writing style, short length of tweets, use of multiple languages within a single tweet, and the presence of non-language-specific content such as Uniform Resource Locators (URLs) and emoticons complicate the identification of language and limit accuracy.

The utility of the UI language setting varies across regions and languages. As of January 2013, Twitter has thirty-three UI languages available, which covers many major languages but misses many key African, Middle Eastern, Indian, and Asian languages (e.g., Afrikaans, Zulu, Bengali, Marathi, Persian, Vietnamese, and Javanese). The importance of these omissions will depend on the design of the study. It is important to note, however, that even when a UI language is present, users writing primarily in that language might still not use that setting. This could be the case when a new language setting is introduced and user adoption lags or for font and device compatibility concerns (Warschauer, Said, and Zohry 2002), among other reasons. The UI language setting also does not capture multilingual users who write in multiple languages on the platform (Eleta and Golbeck 2012).

Nevertheless, the CLD kit and Alchemy show useful promise as automated language identification packages. The former in particular has a great amount of flexibility, as it can both be run offline and be modified as it is open source. The best

[3]It is our hope that future work will also consider the possibilities of using crowdsourced labor to reference tweets spatially accurately.

language identification algorithm for a particular study will depend on a number of factors. One important factor is the languages that an algorithm is trained to identify and the scripts of these languages with which it is trained (e.g., Arabic in Latin characters). CLD performed much better than Alchemy in Japan but otherwise Alchemy performed better in our other research sites. Yet, neither recognized the Arabic chat alphabet, a key omission that is likely to be mirrored in other informal and transliterated alphabets. Cases such as San Diego, however, show relative success in language detection. We might then conclude that on some level, the context of the study matters when considering algorithmic approaches to language identification.

Always running multiple language detection algorithms and reviewing subsets of the results with human coders, as in the work with Twitter of Carter, Tsagkias, and Weerkamp (2011) and the work with the pilot work for this thesis with blogs (Appendix B), might provide insight into the biases and reliability of different automated approaches and flag potential issues on a specific sample of tweets. One method to potentially increase the accuracy of off-the-shelf language identification packages is to preprocess the tweets to temporarily remove emoticons, URLs, and other non-language-specific text (something not attempted in this article). Text automatically generated by third-party services (e.g., Foursquare) often resulted in a mix of languages within a single tweet, so identifying and temporarily removing this text could likely also increase the accuracy of off-the-shelf language identification packages. Removing such text temporarily might improve language identification, but the text itself might nonetheless be useful for further analysis (e.g., studies of link diffusion across languages; e.g., Hale (2012)), and so researchers might wish to retain a copy of the unaltered tweet. Studying the effects of these preprocessing measures and the effects of using link content or grouping several tweets by a single user together for language identification as in Carter, Tsagkias, and Weerkamp (2011) will be a useful avenue for further research. The linguistic and geographic

analysis of short, microblog texts is still an area of active research without any established best practices. Further studies to compare various methods and new approaches (e.g., crowdsourcing with Amazon's Mechanical Turk) are needed to identify concerns and possible future areas of improvement.

The article also compared open-ended profile locations within four research sites to better understand how useful profile locations might be for studying the geography of information. Importantly, it finds that the geolocation results of profile locations are not a useful proxy for device locations (i.e., the place in which the information was disseminated) and identifies several reasons for this discord. This is an important finding not only for the social science analysis of where users are or perceive themselves to be but also for computer science research, which often uses geocoded tweets to evaluate the performance of new location classification approaches. For instance, Sadilek, Kautz, and Bigham (2012) demonstrated that when the location of a subset of users is known, it is then possible to infer the location of the friends of these users. The work reported here suggests that there will be an important difference in where users are placed depending on whether profile location or device location is used to create the starting set of users with a known location. The subset of users who geocode and the subset of users with clear place names in their profiles are unique and, importantly, this article has found that even when both the profile and device location are valid, they do not always correspond. Similarly, there is a danger in relying on the device location as the baseline, true location of the user in training new geolocation algorithms based on text content (a practice used in, e.g., Eisenstein et al., 2010; Wing & Baldridge, 2011; Mahmud et al., 2012).

The article identifies three main reasons for the lack of correlation between profile and device location. First, commensurate with previous work (e.g., Hecht et al., 2011), a large number of profiles contained invalid, nongeographic text or simply larger geographic regions (countries, states). Second, adding to the literature, this

article finds a large number of users tweeting within the study areas had profile locations set to locations outside of the study area.This likely resulted from users who were commuting, traveling, or simply had not updated their profile locations. Finally, several users were within the relevant study site but wrote their profile location information in such a way that the geolocation algorithms used failed to correctly code it. In addition to the recommendations by Hecht et al. (2011) to preprocess profile location information for fictitious names, this study finds it is important to preprocess profile locations to handle abbreviations, lists of multiple locations, and latitude–longitude coordinates surrounded by other text. These steps should be investigated along with tweaking the available parameters to geolocation services. Several profiles had more general geographic boundaries (regions, states, countries), suggesting that the success at being able to place users within a geographic region will vary with the specificity of the region. Attempts to simply locate users within a country are more likely to be successful than trying to locate users to a specific city or metropolitan area. For city-level areas, local gazetteers might be useful (an approach not tested here), but the analysis in this article highlights the importance of supplementing such a list with common abbreviations, misspellings, other-language names, and transliterations of place names. Time zone information, specifically the UTC offset, although not perfect and showing differences in accuracy across study sites, seems to often correspond with the user's current location. UTC offsets also have the value of being more easily processed than the free-form profile locations; however, UTC offsets only give an indication of longitude and not latitude. In this way, time zone information has an interesting parallel to early ship navigation methods where longitude, but not latitude, was measurable (Halley, 1731).

Because of the significant challenges associated with geolocating content and profiles in Twitter, it is tempting to associate certain languages with an assumed geographic origin of content. This article, however, demonstrates the large need for caution in using language as a proxy for location. Within each of the four

research sites considered in this article, a mix of languages was found, suggesting that focusing on language as a proxy for location can lead to two issues. First, such a strategy would miss other-language users located within the location; second, it would likely capture users outside of the location of interest. Future work should look at the dispersion of various languages to determine to what extent language use clusters within certain geographic areas.

Although this article highlights the challenges associated with accurately understanding the geography of information in Twitter, this should not lead us to discount the usefulness of profile locations as a means of geolocating content. Profile locations tell us much about how users perceive, present, and place themselves, and this article has expanded on two methods that can be used to geolocate that unstructured information. Most important, the majority of the 300 million accounts on Twitter contain some type of profile location, whereas only a small proportion of tweets contain any structured device location. As such, further research and additional human coding of profile locations might be needed to accurately determine how well profile locations compare with device locations, how we might best geolocate profile locations, and the ways in which the geolocation of profile information might be linguistically or geographically contingent.

Appendix B

Cross-language Linking in the Blogosphere

The work presented in this appendix chapter is based on my masters thesis, which I rewrote during my first year as a DPhil Student and published as:
Hale, S. A. (2012). Net Increase? Cross-Lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*, 17: 135–151. doi:10.1111/j.1083-6101.2011.01568.x.
The findings of this study motivated the design and execution of my thesis. This appendix focuses on a specific event, the 2010 Haitian earthquake, and bloggers writing in one of three languages (English, Japanese, or Spanish). In contrast, I designed my thesis to analyze a larger number of languages without reference to a particular event. Nonetheless, the complementary findings between the thesis and this appendix (particularly around images) help to further indicate the broad applicability of the findings.

In mid-2009, the Japanese police arrested two foreign teachers on suspicion of importing drugs with the intention to distribute them. The story received national attention in Japan, but the discussion was entirely in Japanese with the names of the teachers written in Japanese characters. A few machine translations were posted

in forums, but the machine translations so muddled the teachers' names that there were no Google results about the story when searching with the teachers' names in English. This changed one week later when a blogger wrote about the story in English and correctly spelled the teachers' full names. His blog soon became the top search result for their names in English. When the charges were quietly dropped 2 months later, local papers reported the update, but the blogger—uninterested or busy—never translated news of the charges being dropped to English. His original blog post remains one of the top results for a search of either teacher's name.

As this anecdote makes clear, information can pass between language groups online, but what information is passed and by whom it is passed is less clear. In the anecdote, machine translation alone was not sufficient to move the information across the language boundary. A bilingual speaker bridged the gap, but he was selective and through him only news of the teachers' arrest and not their release was translated. Bilinguals who translate information have the opportunity to shape opinion on news and politics, and it may be difficult for monolingual speakers to accurately judge the reliability and completeness of the translated information.

This study explores the extent of interaction between speakers of different languages in the blogosphere using hyperlink analysis. It identifies the extent of cross-lingual interaction and the actors who play the largest role in moving information between different languages. The study does so by focusing on the Haitian earthquake of January 2010. This 7.0 magnitude earthquake near Port-au-Prince, Haiti, caused catastrophic damage and at least 230,000 deaths according to Haitian government estimates (Associated Press, 2010). It achieved global resonance and was widely discussed in traditional media and in the blogosphere. This research has limited its focus to interactions between bloggers writing in Japanese, Spanish, and English. These are three predominant languages in the blogosphere (Sifry, 2007), and the earthquake was likely of equal relevance to bloggers writing in all three

languages.[1] The Haitian earthquake also presents an opportunity to study communication patterns surrounding a specific event from the beginning rather than simply part of a conversation in progress about general topics.

The patterns of linking between languages in the blogosphere can shed light on how information is shared between different languages and who is responsible for that sharing. This research looks at the direction of links: Many links into a language may indicate a language with greater agenda-setting power (Delwiche, 2005), while many links out of a language indicates a high level of awareness about information in other languages. In addition, the research investigates the type of bloggers (corporate, personal, etc.) creating these links, how the linking patterns change during the 45-day study window, and what relationship a cross-lingual hyperlink signals.

B.1 Literature

B.1.1 Importance of linking patterns

While the Internet allows a user to view content from any server, the technical ability to view information does not correspond with the ability to understand that information. As content in languages other than English increases and the number of non-English users increases, information has become fragmented into different language groups. Pimienta, Prado, and Blanco (2009) found the percentage of English webpages fell steadily from 75% to less than 45% from 1996 to 2006. During the same time, the percentage of Internet users who were native English speakers also fell from 80% to less than 30%. While the other languages in the study—Spanish, French, Italian, Portuguese, Romanian, and Greek—made steady gains,

[1] Other languages such as French might have been included in this set; however, being an official language of Haiti along with Haitian Creole, French would have had unequal interest in the events in Haiti, a former colony of France. This research is interested in the general flow of information between bloggers in different languages, and the languages chosen for this study have equal potential interest in the events in Haiti with no one language overly dominant a priori. Future research will likely incorporate French and other languages as additional cross-lingual interactions are investigated.

each often accounted for less than 5% of webpages. While the number of pages in each language is important, an understanding of how the pages link together is also important and little work has examined links between language groups.[2]

B.1.2 How languages are connected and why it matters

Hyperlinks may be used as a proxy to measure the awareness of foreign-language content among bloggers. Although all the nuanced motivations for creating cross-lingual hyperlinks are not known, hyperlinks are among the best data available as they can be observed passively, are publically available, and possess a similarity to citations. While bloggers create hyperlinks for a multifaceted number of reasons, a hyperlink within a blog post at the very least signals the author's awareness of the content linked to. With this minimal definition, it is possible to measure to what extent bloggers are aware of content in languages other than the languages in which they write. This definition is well-supported by previous work, which has justified an even deeper meaning of interaction or communication (e.g. Adamic & Glance, 2005; Hargittai, Gallo, & Kane, 2007), and the awareness individuals have for information in other languages is important. Crystal (2003), discussing the dangers of unawareness as linguistic complacency (p. 17), states that a third of British exporters miss opportunities because of poor language skills according to a study by the UK-based Centre for Information on Language Teaching and Research.

Multilingual individuals creating content in peer-produced spheres (e.g. blogs, Wikipedia, open-source software) may create opportunities for information exchanges akin to Granovetter's (1973)"weak ties." Weak-tie acquaintanceships form "crucial bridge[s] between two densely knit clumps of close friends" (Granovetter, 1983, p. 202) and have been found to be important to many areas including the spread of ideas and innovations (e.g. Fine & Kleinman, 1979; Burt, 2004). In the same manner, cross-lingual hyperlinks may represent similarly crucial bridges in the ex-

[2]Gerrand (2007) provides an overview of further studies looking at the number of web pages in various languages.

change of information online. Human-produced translations, while not as ubiquitous as their machine-made counterparts, often better capture nuances in meaning and have the potential to translate cultural meaning in addition to linguistic meaning. This is especially true of more distant language pairs such as Japanese and English. These exchanges could present novel information as the content available in various languages may be very different: Hecht and Gergle (2010) found very little overlap in topics and article content between different language editions of Wikipedia, for example.

Many link analysis studies of the blogosphere have focused on the structure of links between U.S. political blogs. These studies (e.g. Adamic & Glance, 2005; Hargittai et al., 2007) showed bloggers in the U.S. political blogosphere were highly polarized by ideology and linked to blogs with similar political affiliations over those with different affiliations, demonstrating that homophily (Lazarsfeld & Merton, 1954), commonly expressed by the adage "birds of a feather flock together," truly does "structure [] network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, [etc.]" (McPherson, Smith-Lovin, & Cook, 2001). Perhaps unsurprisingly then, Internet websites have been shown to cluster by topic (Chakrabarti, Joshi, Punera, & Pennock, 2002) and language (Hale, 2010). However, even a small number of intercluster bridging ties in a highly clustered network can drastically decrease the path length between any two nodes (Watts & Strogatz, 1998). These large networks with comparatively small average path lengths are said to exhibit the small-world property, and their impact upon innovation and the spread of ideas (e.g. Fleming, King, & Juda, 2007; Uzzi & Spiro, 2005) demonstrates the importance of weak ties.

The limited prior research available suggests different languages have different interlinking patterns. A Berkman Center project mapping the Arabic blogosphere found no hard division between English and Arabic blogs (Etling et al., 2009, p. 19), while a similar study by the Berkman Center did find a clear division between

Farsi and English blogs (Kelly & Etling, 2008). The Arabic project (Etling et al., 2009) found several large national clusters as well as two clusters linking more to foreign language blogs: one to English and one to French. Thelwall, Tang, and Price (2003) investigated linking patterns between academic institutions in Western Europe. They found most interlinking throughout Europe occurred in English. Regional linking between countries sharing a common non-English language was also present. Notably, a typical academic site had about half of its pages in English, with the remaining half in the national language(s). Finally, Zuckerman (2008) states that Japanese language blogs are widely considered less political and more personal than U.S. blogs. Exploration of a set of 9.2 million Japanese blog posts by Fujimura, Inoue, and Sugisaki (2005) seems to confirm this by revealing remarkably few posts (about 1.25%) linked to other blog posts in the set. Indeed, only 16.3% of the blog posts linked to any other webpage at all.

Analysis of data from a pilot study led to three findings: First, the blogosphere demonstrated linguistic homophily with bloggers preferring to link to same-language content over foreign-language content. Second, most cross-lingual links were found to involve English as opposed to directly connecting Spanish and Japanese pages. Finally, the data suggested English might be used more to broadcast than to receive cross-lingual information; however, the number of cross-lingual links to English pages was higher than, but not significantly different from, the number of cross-lingual links from English pages.[3] The pilot study used a sample of 1,968 pages in Spanish, Japanese, and English about the Haitian earthquake at a single point in time. It began with a seed sample of 100 blogs in each language and expanded the set by following all off-site links to pages mentioning Haiti and earthquake. The present study allows analysis of how the network of links changes over time by collecting blogs over a longer period and capturing the date each blog was published. By aggregating a much larger initial set of blogs and not expanding outward from this

[3]More information about this pilot study is available on the author's homepage: `http://www.scotthale.net/`

set, better conditions are established to measure insularity, modularity, and other network properties.

This final result of the pilot study, while not significant, is consistent with findings about the diffusion of television and would suggest bloggers writing in English are generally less aware of foreign language content than bloggers writing in other languages. Nordenstreng and Varis (1974) found television content flow was generally one-way in that a small number of countries exported but did not import content, while many countries imported television content without exporting much content. This also stands in line with the fear of linguistic complacency that Crystal (2003, p. 17) identifies as a danger of English as a global language and with a 2002 European Business Survey by Grant Thornton (cited in Crystal, 2003), which found the percentage of businesses with an executive able to negotiate in another language was much lower in the UK than elsewhere in Europe.

This study compares the number of cross-lingual links from and to blog posts in English, Japanese, and Spanish. The literature suggests the following hypothesis, which this study tests: fewer cross-lingual links will originate from English language blogs than either from Spanish or Japanese language blogs (H1). In addition to looking at links for the full time period, this study will also look at how the distribution of hyperlinks changes over the 45-day period following the earthquake. In particular, this study will test the hypothesis that bloggers' awareness of foreign-language content as measured by the separation between language groups will increase with time (H2). This follows the conventional thinking that if one blogger bridges a language gap, and bloggers read one another, then other bloggers may cite the same foreign source or the blog referencing it. That is, for each translation or cross-lingual link made, there should be a ripple or knock-on effect as additional bloggers become aware of the foreign-language content and possibly link to it or a blog citing it.

B.1.3 Who connects languages?

Mainstream media sources and others at times rely heavily on cross-lingual blogs to monitor foreign events. A survey of foreign correspondents in China found that nearly three times as many survey respondents followed English-language blogs on a daily basis as compared with Chinese-language blogs (MacKinnon, 2008). MacKinnon (2008) writes that "this suggests that English-language 'bridge blogs' about China have greater direct influence on China correspondents than Chinese-language blogs" (p. 19). Thus, it is important to analyze who is creating the multilingual connections and the nature of these connections.

Within the blogosphere, multilingual bloggers may bridge language gaps by blogging about content in other languages. Qualitative evidence (e.g. Zuckerman, 2008) shows examples of cross-lingual or bridgeblogging,[4] but how common it is and the nature of cross-lingual links remain unclear. Nevertheless, where it occurs qualitative studies suggest cross-lingual blogging "play[s] an increasingly important role in connecting [culturally and linguistically] disparate spheres of conversation and argument together [online]" (Zuckerman, 2008, p. 47).

Global Voices (`http://globalvoicesonline.org/`), founded by MacKinnon and Zuckerman, seeks to aggregate bridgeblogs and encourage translation between languages. Other services such as Meedan (`http://news.meedan.net/`) and Mojofiti (`http://www.mojofiti.com/`) also seek to encourage cross-lingual blogging through a combination of machine and human translation. This study examines the impact and the importance of encouragement in such communities.

[4]Zuckerman (2008) specifically discusses "bridgeblogging," which is a concept closely related to, although not fully interchangeable with, cross-lingual blogging. Zuckerman distinguishes bridgeblogs by their intended audience. He states bridgeblogs are "intended to be read by an audience from a different nation, religion, or culture." This study concerns itself with cross-lingual blogs, which link to content in a language different from that of the blog. In some cases, these terms overlap as when Jeremy Goldkorn on his blog Danwei discusses Chinese news in English and links to the Chinese sources. However, a blogger may bridge gaps between different cultures that share a common language (e.g. the UK and India or Spain and Argentina) or without using any hyperlinks at all. Likewise, a Japanese blogger writing a blog "targeted to friends, family or countrymen" is not bridgeblogging according to Zuckerman's definition, even if he references material from foreign-language websites. Such a blog would, however, be a cross-lingual blog.

The meaning of cross-lingual hyperlinks in the blogosphere has not been previously studied. Links are often considered a form of citation, and Benkler (2006) suggests the prevalence and importance of linking to sources online is part of a "see for yourself" link culture. This culture and the general linking structure of the Internet, Benkler argues, mitigate against polarizing and fragmenting forces and actually form a more egalitarian landscape online than is possible with the skewing forces of capital investment required in the mass-media sphere. However, Benkler only cites examples from English language websites, and it is unclear to what extent a "see for yourself" culture can overcome fragmentation tendencies between languages online. Hargittai et al. (2007) found a diversity of types of links through a qualitative coding of cross-ideological links in the U.S. political blogosphere suggesting reasons for creating cross-lingual links may be equally varied.

Building upon the recognition of Zuckerman (2008) and Hargittai et al. (2007) that there are meaningful differences between bloggers and types of hyperlinks, this work examines all blogs posts with cross-lingual links qualitatively. The type of author creating the cross-lingual link and the nature of the link are classified by human coders in order to test the hypothesis that blogs with multiple authors, professional affiliation, and/or higher traffic are more likely to create cross-lingual links (H3). Categories are developed through analysis of data from a pilot study and refined through the coding process.

The next section will describe the methods used to test the three research hypotheses in reference to English, Spanish, and Japanese blogs about the Haitian earthquake. Based on the literature above, the three research hypotheses are that there will be fewer cross-lingual links from English than either from Spanish or from Japanese (H1), the awareness of foreign-language content will increase with time (H2), and blogs creating cross-lingual links will more likely have multiple authors, professional affiliation, and/or a high amount of traffic (H3).

B.2 Data and methods

B.2.1 Data collection

This research develops new methods, tests them in a pilot study, and then applies them to a larger dataset.[5] Search queries for "haiti" and "earthquake" in Japanese (ハイチ and 地震), English, and Spanish ([haiti or haití] and terremoto) were conducted on three search engines: blogs in all three languages were gathered from Google Blog Search and BlogPulse, and the Japanese results were further supplemented by results from Yahoo! Japan Blog Search given the service's extreme popularity in Japan.[6] As search engines limit the number of maximum results they return for any one query (Thelwall, 2008), separate queries were conducted for each language on each day of the 45-day window studied and the results combined. The sample was not expanded to include linked pages in order to reduce the chance of nonblog webpages entering the set and to reduce bias in community detection algorithms that would occur if expanding. All links between blog posts in the set were recorded and analyzed using igraph and UCINET (Borgatti et al., 2002).

Duplicate pages were identified and removed through an automated screening process. Data from the pilot study revealed duplicate entries were created by links to URL-shortening services like tinyurl.com and bit.ly. Therefore, all pages were expanded to their full and final URLs by following all HTTP redirection responses. The pilot work also indicated the necessity of considering the addition of an anchor tag name (#) or auxiliary query string parameters. As most blogging sites are database driven some query string parameters (e.g., post, p, id) are important while others (e.g., footer, style) are not. Through analysis of the 1,968 pages in the pilot study, a white list of important query string arguments to aid in the detection of duplicate pages was constructed.

[5]Full source code is available from the author upon request.

[6]Alexa ranks Yahoo! Japan as the highest traffic site in Japan (`http://www.alexa.com/topsites/countries/JP`).

The language of each blog post was detected in two ways. First, a simple count of the number of times "earthquake" appeared in each language on the page was conducted. Second, the compact language detection code released in the open-source Chromium project[7] and used in the Google Chrome web browser was adapted and run against all pages in the dataset. Manual review of random subsets found both methods had a tendency to classify ambiguous texts as English. Where the two methods disagreed, a result of Spanish or Japanese was preferred over a result of English: This occurred in only 12% of the blog posts. The languages of all blogs involved in cross-lingual links were manually verified as part of the qualitative coding discussed below. Where neither method could reliably identify the language of a page, that page was excluded. This resulted in 859 pages (0.75%) being excluded.

The data was collected over a 45-day period and the date of each blog post was recorded. The data collected for this research begins on the day of the earthquake, 12 January 2010, and ends on 25 February 2010, the day before two successive earthquakes in Okinawa, Japan, and Chile. These earthquakes caused an increase in earthquake related blogging as reflected in trend lines from BlogPulse and Yahoo! Japan, and likely influenced the three language groups differently.

Two issues of link validity need to be considered. Users may create links without understanding the source content (false positives) as well as not create links even when using content from a cross-lingual site (false negatives). The first issue is mitigated by the research design. All pages in the dataset discuss the Haitian earthquake, and this unity of topic limits the number of irrelevant false-positive links. Furthermore, any irrelevant cross-lingual links were identified during the manual coding of cross-lingual links. Using content without linking (false negatives) is a potentially more serious threat to the validity of the study. This concern is limited by the widespread practice of linking to source material as a core aspect of blogger culture (Benkler, 2006; Adamic & Glance, 2005; Hargittai et al., 2007). In

[7] `http://src.chromium.org/viewvc/chrome/trunk/src/thirdparty/cld/`

addition, anecdotal experience suggests many bloggers still link to source material even when it is in a different language.

B.2.2 Coding of blog attributes

Further qualitative coding was carried out manually on the cross-lingual links in the set to gain a more in-depth look and better interrogate the meaning of cross-lingual hyperlinks in the blogosphere. All blogs sending or receiving cross-lingual links were examined manually to confirm their languages, to classify their types (e.g. personal, group, professional), and to identify any obvious topics of focus. This data is used to determine which types of authors are most likely to create links to blogs in other languages. In addition to these characteristics about the blogs, the relationship (e.g. translation, excerpt, citation) between two blogs sharing a cross-lingual link was classified to determine the meaning of cross-lingual hyperlinks. Both sets of categories were developed through an iterative process while coding.

All cross-lingual links were coded by the researcher. To ensure the reliability of the qualitative coding, two independent coders (one for English–Japanese blog pairs and another for English–Spanish blog pairs) independently examined 50 cross-lingual links each based on the recommendation of Lombard, Snyder-Duch, and Bracken (2002) of calculating intercoder reliability on not less than 10% of the dataset or 50 links. Overall intercoder reliability was high: percent agreement for the author type was 0.82 and 0.84 for English–Spanish and English–Japanese pairs respectively (Cohen's unweighted kappa, κ, was 0.76 and 0.79). For the relationship between blogs, percent agreement was 0.85 and 0.90 for the same language pairs respectively with kappa values of 0.65 and 0.79. Disagreements between coders were resolved through discussion.

Language	Page Count
English	47.8% (54,053)
Spanish	31.9% (36,111)
Japanese	20.3% (22,953)
Total	100.0% (113,117)

Table B.1. Language Distribution
N.B. The dataset includes all blogs mentioning "Haiti" and "earthquake" in English, Spanish, or Japanese returned by searches on Google Blog Search, BlogPulse, and Yahoo! Japan for the 45-day period following the Haitian earthquake.

B.3 Analysis and results

The dataset consists of 113,117 blogs after aggregating the search results from the three blog search engines (Google Blog Search, BlogPulse, and Yahoo! Japan), removing duplicates, and excluding blogs in indeterminable languages. The distribution of detected languages within the dataset is given in Table B.1. Despite the size of the Japanese blogosphere (which one study, Sifry, 2007, found to be on par with English) many fewer blog posts in Japanese were found discussing the Haitian earthquake than blog posts in Spanish or English. This is consistent with qualitative findings about the Japanese blogosphere that many Japanese blogs resemble diaries with fewer links and are frequently about the authors' daily lives (Zuckerman, 2008).

B.3.1 Links between language groups

The links between language groups, given in Table B.2, show each language group is highly insular in its linking pattern as predicted by the literature and the pilot study. The diagonal of the table, which represents links within the same language group, contains 94% of the hyperlinks in the dataset, demonstrating the relevance of homophily to language.

Modularity (Newman & Girvan, 2004) is a measure of "the goodness of fit" of a given partitioning of a network and can be used as a measure of language group

	Destination						
Source	English		Spanish		Japanese		Total
English	98.5%	(6,844)	1.3%	(88)	0.2%	(16)	6,948
Spanish	10.6%	(408)	89.3%	(3, 425)	0.0%	(1)	3,834
Japanese	10.8%	(188)	0.3%	(6)	88.9%	(1,551)	1,745

Table B.2. Dataset Hyperlinks
N.B. The rows represent the source and the columns the destination of links to and from each language group within the dataset, which consists of a total of 12,527 hyperlinks. Percentages are row percentages.

insularity and an operationalization of the separation between language groups. Modularity measures how the network deviates from a network of the same number of nodes and community divisions but with random edges and has been justified previously as a measure of polarization (Waugh, Pei, Fowler, Mucha, & Porter, 2009). In the present context, the lowest possible modularity score (0.0) represents no separation between language groups (the language groups are linked together as much as in a random network), and the highest score (1.0) represents the most separation between language groups (i.e. no cross-lingual links). For this dataset, the modularity score for the entire network is 0.51, which indicates that language is a strong dividing force (Newman & Girvan, 2004). English is the most insular of the three language groups and accounts for 42% of the modularity score.[8] Spanish accounts for 37% of the score, and Japanese accounts for 21%.

There are 707 cross-lingual links in the dataset, which represent 5.6% of all links. Consistent with H1, English is the only group to receive more links than it sends (596 vs. 104). This difference is significant ($p < 0.0001$) as determined by a t-test

[8]Modularity is calculated by considering a division of a network into k communities. Let e denote a $k \times k$ symmetric matrix where each element e_{ij} is the fraction of links from vertices in community i to vertices in community j. Furthermore, let the row (or column) sums be defined as $a_i = \sum_j e_{ij}$, which represent the fraction of all links connecting to vertices in community i. Modularity is then defined as: $Q = \sum_i (e_{ii} - a_i^2)$. The observed fraction of edges connecting edges within community i is e_{ii}, while a_i is "the expected value of the same quantity in a network with the same community divisions but random connections between the vertices" (Newman & Girvan, 2004, p. 7). The contribution of community i to the overall modularity score is simply $Q_i = (e_{ii} - a_i^2)$. Expressed as a percentage of the entire network's modularity score, this is Q_i/Q. Further details are available in the pilot study (Hale, 2010) available on the author's website.

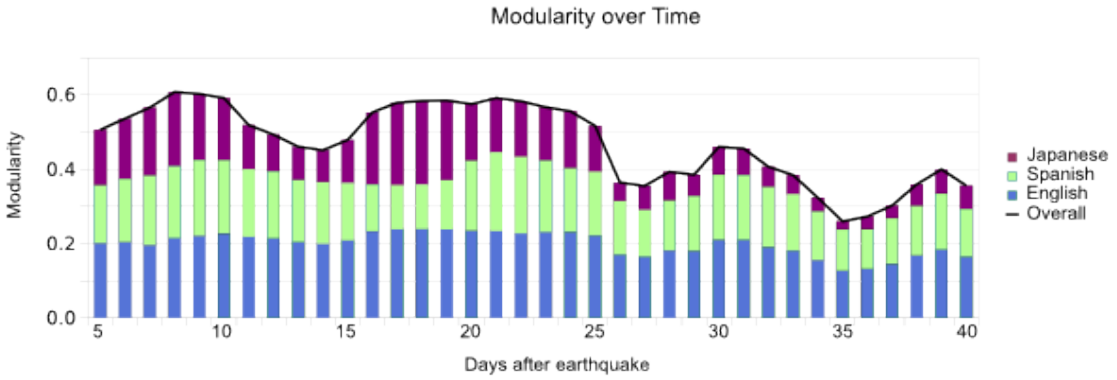


Figure B.1. Modularity Over Time

in UCINET comparing the in-degree and out-degree of English blogs receiving or sending cross-lingual links. Furthermore, the number of cross-lingual links originating on English blogs (104) is significantly lower than the number originating from either Spanish (409, $p < 0.0001$) or Japanese (194, $p < 0.0001$) blogs.

B.3.2 Changes over time

The network of hyperlinks changes over time, as new nodes (blog posts) and edges (hyperlinks) are added, and previous research has not accounted for changes in the separation between language groups. To examine H2 that awareness will increase over time, a modularity score is calculated for each day in the dataset. The modularity scores vary widely from day-to-day depending on the number of blogs published in each language on that day. Therefore, a smoothing window is advantageous to identify overall trends. The score cannot be calculated simply by cumulatively adding nodes and edges as this would cause modularity to decrease simply because the network would start with zero edges between language groups and steadily add them as they were created. Rather, old blogs must be removed and new blogs added each day. Figure B.1 shows the modularity score of the blog network over the 45-day collection period employing a 5-day smoothing window.

The modularity score undulates, but shows an overall downward trend from a peak score of 0.61 eight days after the earthquake to a low score of 0.26 thirty-five days after the earthquake. The undulation perhaps suggests a delay between the creation of foreign language content and its incorporation into blogs in other languages. Overall, the more than halving of the score is consistent with the hypothesis that awareness would increase over time.

B.3.3 Meaning of cross-lingual links

The qualitative analysis of the cross-lingual links in the dataset revealed seven distinct types of links: translation, quotation, inclusion, source, citation, blogroll, and comment. Each of these is discussed below. Cross-lingual links to and from blog posts were further classified by their author-type into the following categories: personal, group, professional, media, and government. All blog posts fit within these categories.

Links classified as translations provide nearly verbatim copies of the source content in another language. Quotation links translate only small portions of the linked-to content. Source links identify a foreign language blog as the source of the story, but include no obvious translation of the original content. These pages often have summaries or other adaptations of the original content in a different language. Inclusion links quote text from the foreign language page, but do so in the foreign language without translation. Finally, citation links serve as footnotes indicating the source of a fact or figure within the blog post or providing a destination for further reading.

Blogroll links and links in blog comments are qualitatively different from links created in the body of blog posts. Hargittai et al. (2007) notes that blogroll links, which are links in a sidebar common to all posts on a particular blog, are updated with differing regularity, and may signal different levels of engagement with the linked content. Links in blog comments are created by readers, and hence cannot

be used as an indicator of blog author engagement with cross-lingual sources.

Blogs classified as personal are written by one person, a couple, or a family, and often use the personal pronoun in their self-descriptions. Group blogs are written by multiple individuals and include charities and nongovernmental organizations. Professional blogs are authored by companies that are not primarily focused on news dissemination (e.g. search engine companies, professional music groups), while media blogs are written by companies with the primary focus of news dissemination. Finally, government blogs are written by national government entities.

The automated methods identified language and cross-lingual links well. The machine-detected language coincided with a human coding of language in 95% of the 965 blogs in the set of blogs with cross-lingual links. The misclassified blogs were often due to linguistically similar languages (e.g. Italian or Portuguese instead of Spanish) or font encoding issues. The incorrectly identified cross-lingual links, along with 76 blogroll links and 34 links in comments were excluded from further analysis leaving 541 cross-lingual links, representing 4.3% of all links in the dataset.

The qualitative coding revealed that citation links account for the majority of cross-lingual links (65.2%). Some form of translation occurs in 24.1% of links. Of these links, 17.6% are complete or nearly complete translations while another 6.5% are quotations, translating only excerpts of the original content. Table B.3 shows the percentage of each link type in the dataset.

There is a clear difference between the origins and destinations of cross-lingual links. In Figure B.2, while 29% of all cross-lingual links originate on Spanish-language personal blogs and a further 26% on Spanish-language group blogs, the destinations of these links are primarily to English-language sources: 29% of all cross-lingual link destinations are English-language media and another 29% are English-language group blogs.

Most blogs (54%) containing cross-lingual links were classified as personal. Personal blogs contain the largest number of cross-lingual links for both Spanish (29%

Relationship	Frequency	Valid %	Total %	Full Dataset %
Translation	95	17.6	14.6	0.76
Quotation	35	6.5	5.4	0.28
Inclusion	9	1.7	1.4	0.07
Source	49	9.1	7.5	0.39
Citation	353	65.2	54.2	2.82
Valid Total	541	100	83.1	4.32
Blogroll	76		11.7	0.61
Comment	34		5.2	0.27
Invalid Total	110		16.9	0.88
Grand Total	651		100	5.20

Table B.3. Relationship of Cross-Lingual Linking Blogs

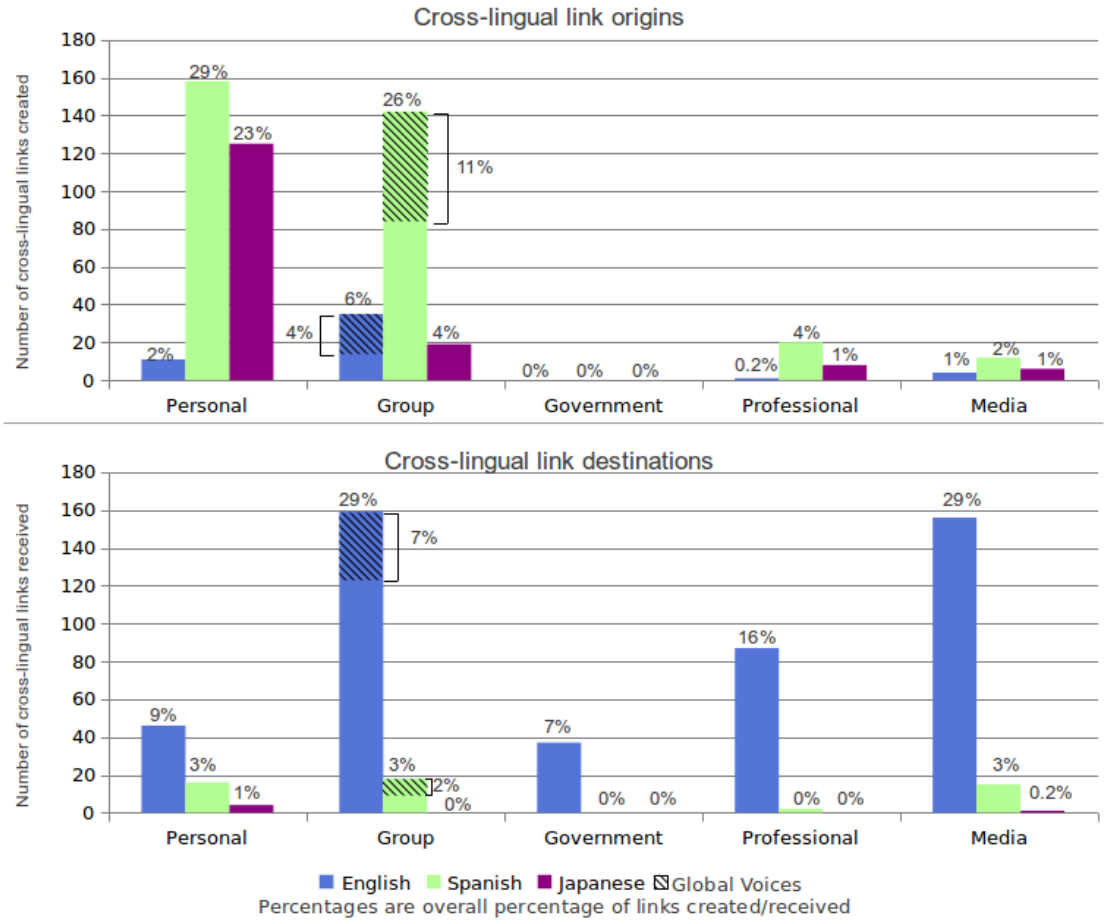


Figure B.2. Origins and Destinations of Cross-Lingual Hyperlinks

of all cross-lingual links) and Japanese (23% of all cross-lingual links); however, for English group blogs contain the largest number of cross-lingual links. English-language groups author 6% of all cross-lingual hyperlinks in the dataset—two-thirds of these links are created by Global Voices, a bridgeblogging community. Overall, Global Voices alone creates 15% of all cross-lingual links in the dataset. It accounts for half of the quotations and one-third of the translations in the dataset.

English-language media are the most central node in the network of cross-lingual links (figure available in the supplemental materials online). The largest single destination of cross-lingual links is to a collection of photos published by the Denver Post.[9] After a link to this page was first posted on one Japanese blog, a wave of additional Japanese bloggers also linked to the same page. In fact, a large number of blogs (13.8%) share a photo or video directly on the page, while even more link to another blog specifically mentioning the multimedia content of the page.

Additional coding revealed several other aspects. Only 11.4% of cross-lingual links are between pages owned by the same organization. In addition, nonnews blogs sharing a cross-lingual link often also share a common topic or theme (e.g. technology, automotive, music). Finally, a large number of blogs making cross-lingual links (15.1%) appear to be very similar to another cross-lingually linking blog in the set. These blogs have nearly the same text, images, and/or links as another blog in the dataset.

B.4 Discussion

Human translation exists within the blogosphere. This translation has a superior potential to machine translation in that it can translate not only language but also cultural meaning. This dataset shows that cross-lingual linking is a small activity within the blogosphere, but that such linking increases over time and enables otherwise nonexistent paths between blogs of different languages. Where these paths

[9] `http://blogs.denverpost.com/captured/2010/01/13/earthquake-in-haiti/`

form, they potentially enable the flow of information and innovations across language divisions. Nevertheless, as MacKinnon (2008) finds regarding her survey of foreign correspondents' use of social media in China, this dataset too shows the current situation is more of a limited information exchange than a many-to-many global discourse. While the Internet presents the opportunity to consume information from far-flung corners of the world, our natural tendency to interact with others similar to us (homophily), linguistic barriers, and a lack of foreign-language awareness have caused the diversity-enhancing potential of the Internet to be underrealized.

Incomplete or poor translations can misrepresent reality. Selective translation, for example, has the possibility of giving distorted views, as in the opening anecdote where a blogger translated news about alleged drug distribution, but did not later translate information that the charges were dropped. Human translators can mistranslate information or misrepresent facts. Poor machine translation also may lead to misunderstandings. In both cases, readers who do not speak the language of the linked-to content cannot evaluate the information fully. This interferes with the "see for yourself" cultural ideal on the Internet (Benkler, 2006) where readers are free to evaluate sources themselves to determine an author's trustworthiness. The Wikipedia user community, for example, emphasizes the need for a citation for each claim in an article; however, where these citations are to foreign language content, Wikipedia readers often will not be able to evaluate the trustworthiness of the content themselves. Nevertheless, such cross-lingual links are valuable in promoting transparency and foreign language awareness, and also in enabling verification of the information by those who are able. Indeed, mathematical or computational skills limitations may prevent some from evaluating the trustworthiness of even same-language sources; yet, these links are useful (and encouraged) in order to promote transparency and enable those who can to evaluate the sources.

The research presented here has found that fewer cross-lingual links come from English language blog posts than either from Spanish or from Japanese language

blog posts (H1). Secondly, bloggers' awareness of foreign-language content as measured by the separation between language groups does increase with time (H2). Finally, the data do not support H3, but suggest that most cross-lingual links come from personal blogs with few authors and not from blogs with multiple authors, professional affiliation, or higher traffic. While being created by nonprofessionals, most cross-lingual links in the dataset point to professionally affiliated, high traffic blogs. How the Haitian earthquake as a media-driven event differs from other blogging topics or events will require further study. The dataset used in this study provides a lower bound on the amount of cross-lingual activity in the blogosphere as not all individuals translating foreign language content necessarily linked to the content.[10] In particular, traditional media, discussed further below, seem not to participate extensively in the "see for yourself" culture. The implications of each finding are discussed below.

English and Cross-Lingual Links: Promotion of Foreign Content Awareness This dataset underscores Crystal's (2003) view of English as a global language. Links directly between Spanish and Japanese are rare in this dataset while links to English are comparatively much more common. The pilot study, which expanded a smaller dataset by following all off-site links, showed that paths through English-language pages may connect Spanish and Japanese pages even where direct paths do not exist.

English's status as a global language comes at the cost of its bloggers being less aware, in general, of content in other languages. This stands in line with the danger of linguistic complacency that Crystal (2003, p. 17) identifies with a global language. Such linguistic complacency and lack of awareness of foreign language content may partially explain the lack of coverage of developing nations (particularly of those in Africa) that Zuckerman (2008) finds in English-language media and blogs and the

[10]Language reliability checks discovered one such site in this dataset. Masomi, who lives in Ecuador, translates local news articles from Spanish to Japanese on his blog `http://d.hatena.ne.jp/masomi1979`. Each post alternates between the original Spanish text and translation text paragraph by paragraph. The posts do not, however, link to the original sources. This site was discovered as it was the only site classified as Spanish by one language detection method and as Japanese by the other method.

similar coverage holes Graham (2009) identifies on Wikipedia.

Machine translation is not necessarily the solution to this issue based on the behavior of participants in an Oxford Experimental Laboratory experiment[11] asking 21 questions about cross-European life scenarios. Only 29% of the 130 participants used machine translation while browsing the Internet during the session; yet, 73% reported encountering foreign language webpages during the experiment in the post-experiment questionnaire. In addition, 45% indicated that in general they simply "hit the back button" when encountering non-English content online (Margetts & Hale, 2010).

Sites like Global Voices, which promote the translation of content, play a critical role in increasing information diffusion online. Global Voices is the largest and most successful of these sites in the dataset. It deserves special note for its ability to drastically increase the reach of authors through its community translation. It offers policymakers a good model to increase awareness of foreign-language content. Like other volunteer services, Global Voices is only as strong as its volunteer community. This dataset highlights that, as far as coverage of the Haitian earthquake, Spanish–English translations far outnumber Japanese–English translations in the overall dataset. This is also reflected within Global Voices, where noticeably fewer Haitian earthquake articles were translated to/from Japanese than to/from Spanish.

B.4.1 Cross-lingual links over time

The dataset demonstrates a decrease in modularity over the 45-day period. This is consistent with the idea of awareness increasing over time. The idea of a ripple or knock-on effect is confirmed empirically by the qualitative coding of blogs, which found 15.1% of the blogs were very similar to another blog in the set. It appears that often after one translation is made, that translation is cited or included directly in other blogs in the new language group. In addition, the largest node receiving cross-

[11] The experiment involved 130 participants answering 21 questions. The sample was not random; so, it is unclear to what extent these trends would hold in a larger, random sample.

lingual links—a photoblog by the Denver Post—accumulated links at an increasing rate. The use of photos and video seem an especially good way to encourage cross-lingual linking as they can often be understood independent of any written text. The percentage of blogs sharing a photo or video directly within the post (13.8%) is relatively high when one considers copyright issues. Bloggers may be reluctant to directly include a photograph from a media organization since news agency AFN's lawsuit against Google News (Cozens, 2005) and the Associated Press's threatened legal action against various bloggers using AP photos without explicit approval (Hansell, 2008; Ledbetter, 2008).

B.4.2 Authors and targets of links

The largest number of cross-lingual links were created by individuals or groups and pointed to traditional media and Global Voices. As traditional media adapt to online markets and seek to increase online profits, several media organizations have sought additional protection for content online (Federal Trade Commission, 2010). As media organizations advocate for expanded copyright protections, policymakers should specifically consider the issues of translation. Currently U.S. copyright law and international copyright treaties (e.g. WIPO) restrict translation as a derivative right; however, where media organizations are not translating content themselves, translations by individuals ought to be allowed. In addition, there should not be an overly complex or legalistic process for securing approval to translate media content as this could be a particularly acute burden to the small groups and individuals, who as this study reveals, author the majority of translations.

Outside of links to professional media, blogs sharing a cross-lingual link often also shared a common topic or theme. This suggests homophily is present in two ways within the dataset. First, bloggers prefer to link to other bloggers writing in the same language. Second, bloggers prefer to link to other bloggers writing about the same or similar topics. This cohesion of topic in some cases overcomes

language gaps. Benkler (2006) suggests that small clusters of blogs about similar topics serve as a collaborative filter. He suggests the best posts in each cluster are passed onto more prominent clusters, and eventually the best or most insightful posts may be referenced by authors of high-traffic blogs. This dataset suggests that in some of these small clusters, cross-lingual interaction is taking place; however, further work will need to analyze how the probability of any foreign language content being elevated to mainstream blogs compares with that of same-language content and how interpretations of foreign language content change as they are referenced by higher traffic blogs.

Newspaper foreign correspondents are largely absent from cross-lingual link creation in this dataset. However, the main role of foreign correspondents is to move information between different countries often with distinct languages. Indeed, most news stories about an event or development in Haiti moved information between languages. This dataset suggests news media organizations do not link cross-lingually with great frequency. Greater transparency, cross-lingual recognition, and value might be created if more media organizations linked to sources, even if they are in another language.

B.5 Conclusion

Human translation occurs in the blogosphere in a decentralized patchwork of mainly individuals and small groups. Communities that encourage translation are a particularly effective means to locate translations, avoid duplication, and provide support and encouragement. Bloggers seem to read one another and on occasion link to blogs referencing foreign content or the foreign content itself demonstrating an increasing awareness of foreign content that undulates and changes over time perhaps with the amount of content available. Bloggers writing in English link much less to foreign content than bloggers writing either Spanish or Japanese. Although there is

a substantial amount of content in English, the percentage of all Internet content in English is steadily declining, and human summarization and translation provide one way to communicate information between languages. This is particularly important for languages where machine translation performs poorly.

Individuals creating cross-lingual links select what content to translate and how to present that content. Given the reliance of readers and mainstream media on this content, it is important to study these authors further. This work has laid the foundation for such future work, which will analyze real-world outcomes and the role language plays in online communities such as Wikipedia, Twitter, and question and answer forums. Additional work could look at geographically closer and more linguistically similar languages, and also seek to determine how generalizable the results found here are to other, less media-driven events. Such a study might look at policymaking affecting a constituency with multiple languages, and how discussions influence, correlate, or diverge with final policy decisions, for example. Other future work might further investigate the presence of shared themes between blogs sharing cross-lingual links.

This work has developed the techniques to make such studies possible in the future. The blog recruitment strategy, the language classification method, and the use of modularity as a measure of the separation between language groups may easily be adapted to future studies. In addition, the qualitative coding of cross-lingual links will help inform future research designs as to the meaning and significance of cross-lingual links.

References

Adafre, S. F., & De Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the workshop on new text: Wikis and blogs and other dynamic text sources at the 11th conference of the european chapter of the association for computational linguistics* (pp. 62–69). Association for Computational Linguistics.

Adamic, L. A. (1999). The small world web. *Lecture Notes in Computer Science*, *1696*, 443–452.

Adler, B. T., & Alfaro, L. de. (2007). A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on world wide web* (pp. 261–270). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1242572.1242608`

Adler, B. T., Alfaro, L. de, Pye, I., & Raman, V. (2008). Measuring author contributions to the Wikipedia. In *Proceedings of the 4th international symposium on wikis* (pp. 15:1–15:10). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1822258.1822279`

Adler, B. T., Chatterjee, K., Alfaro, L. de, Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content. In *Proceedings of the 4th international symposium on wikis* (pp. 26:1–26:12). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1822258.1822293`

Albert, R., Jeong, H., & Barabasi, A.-L. (1999). Internet: Diameter of the World-Wide Web. *Nature*, *401*(6749), 130–131.

Anthony, D., Smith, S. W., & Williamson, T. (2009). Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia. *Rationality and Society*, *21*(3), 283–306. Available from http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf

Aral, S., & Alstyne, M. V. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, *117*(1), 90–171. Available from http://www.jstor.org/stable/10.1086/661238

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, *35*(4), 216–224.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the fourth acm international conference on web search and data mining* (pp. 65–74). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1935826.1935845

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on world wide web* (pp. 519–528). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2187836.2187907

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1075–1084). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2207676.2208553

Barnett, G. A., & Choi, Y. (1995). Physical distance and language as determinants of the international telecommunications network. *International Political Science Review*, *16*(3), 249–265. Available from http://ips.sagepub.com/content/16/3/249.abstract

Organization for Economic Co-Operation and Development. (2007). *Paticipativeweb*

and user-created content: Web 2.0, wikis and social networking. Available from http://www.oecd.org/internet/ieconomy/participativewebanduser-createdcontentweb20wikisandsocialnetworking.htm

Bilic, P., & Bulian, L. (2014). Lost in translation: Contexts, computing, disputing on Wikipedia. In M. Kindling & E. Greifender (Eds.), *iconference 2014 proceedings* (pp. 32–44). iSchools. Available from http://hdl.handle.net/2142/47309

Birner, B. (2005). *Bilingualism* (Tech. Rep.). Washington, DC, USA: Linguistic Socieyt of America. Available from http://www.linguisticsociety.org/files/Bilingual.pdf

Brown, R. (1976). Reference in memorial tribute to Eric Lenneberg. *Cognition*, *4*(2), 125–153. Available from http://www.sciencedirect.com/science/article/pii/0010027776900019

Bruns, A., & Burgess, J. E. (2011). #Ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics and Culture*, *44*(2), 37–56. Available from http://eprints.qut.edu.au/47816

Burt, R. S. (2004). Structural holes and good ideas. *The American Journal of Sociology*, *110*(2), 349–399. Available from http://www.jstor.org/stable/3568221

Burt, R. S. (2005). *Brokerage and closure: An introduction to social captial*. New York: Oxford University Press.

Callahan, E. S., & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, *62*(10), 1899–1915. Available from http://dx.doi.org/10.1002/asi.21577

Carter, S., Tsagkias, M., & Weerkamp, W. (2011). Semi-supervised priors for microblog language identification. In *Dutch-belgian information retrieval workshop (dir 2011)*. Available from http://wouter.weerkamp.com/downloads/

dir2011-lid.pdf

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 759–768). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1871437.1871535

Churchill, E. F. (2013, Sep). Putting the person back into personalization. *Interactions*, *20*(5), 12–15. Available from http://doi.acm.org/10.1145/2504847

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. Available from http://epubs.siam.org/doi/abs/10.1137/070710111

Cook, S., Corrie, C., L., F. A., & H., M. M. (2011, 08). Assessing Google Flu Trends performance in the United States during the 2009 influenza virus a (h1n1) pandemic. *PLoS ONE*, *6*(8), e23610. Available from http://dx.doi.org/10.1371/journal.pone.0023610

Cooper, R. L. (1969). Two contextualized measures of degree of bilingualism*. *The Modern Language Journal*, *53*(3), 172–178. Available from http://dx.doi.org/10.1111/j.1540-4781.1969.tb04585.x

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2007). Suggestbot: Using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on intelligent user interfaces* (pp. 32–41). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1216295.1216309

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.

Cyr, D., & Trevor-Smith, H. (2004). Localization of web design: An empirical comparison of German, Japanese, and United States web site characteristics. *Journal of the American Society for Information Science and Technology*, *55*(13),

1199–1208. Available from http://dx.doi.org/10.1002/asi.20075

Dalton, J. (2012). *Useful data from "pointless babble": Discovering sentence-aligned parallel texts from Twitter streams.* Unpublished master's thesis, Department of Computer Science, University of Oxford.

Davis, G. F., Yoo, M., & Baker, W. E. (2003). The small world of the american corporate elite, 1982-2001. *Strategic Organization*, *1*(3), 301–326. Available from http://soq.sagepub.com/content/1/3/301.abstract

DePalma, D. A., Sargent, B. B., & Beninatto, R. S. (2006). *Can't read, won't buy: Why language matters on global websites: An international survey of global buying preferences* (Tech. Rep.). Lowell, Masschusetts, USA: Common Sense Advistory, Inc.

Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, *301*(5634), 827–829. Available from http://www.sciencemag.org/content/301/5634/827.abstract

Drott, M. C. (1998). Using web server logs to improve site design. In *Proceedings of the 16th annual international conference on computer documentation* (pp. 43–50). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/296336.296350

Durham, M. (2003). Language choice on a Swiss mailing list. *Journal of Computer-Mediated Communication*, *9*(1). Available from http://dx.doi.org/10.1111/j.1083-6101.2003.tb00359.x

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277–1287). Stroudsburg, PA, USA: Association for Computational Linguistics. Available from http://dl.acm.org/citation.cfm?id=1870658.1870782

Eleta, I., & Golbeck, J. (2012). Bridging languages in social networks: How multilingual users of Twitter connect language communities. *Proceedings of the*

American Society for Information Science and Technology, *49*(1), 1–4. Available from http://dx.doi.org/10.1002/meet.14504901327

Ellis, D. (2009). A case study in community-driven translation of a fast-changing website. In N. Aykin (Ed.), *Internationalization, design and global development* (Vol. 5623, pp. 236–244). Springer Berlin Heidelberg. Available from http://dx.doi.org/10.1007/978-3-642-02767-3_26

Erard, M. (2012, January). *Are we really monolingual?* Available from http://www.nytimes.com/2012/01/15/opinion/sunday/are-we-really-monolingual.html

Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: Politics and dissent online. *New Media & Society*, *12*(8), 1225–1243. Available from http://cyber.law.harvard.edu/publications/2009/Mapping_the_Arabic_Blogosphere;http://nms.sagepub.com/content/12/8/1225.abstract

Eurobarometer. (2011). *User language preferences online: Analytical report* (Tech. Rep.). (Survey conducted by The Gallup Organization, Hungary upon the request of Directorate-General Information Society and Media, European Commission)

European Commission. (2012). *Europeans and their languages.* Available from http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf

Fine, G. A., & Kleinman, S. (1979). Rethinking subculture: An interactionist analysis. *The American Journal of Sociology*, *85*(1), 1–20. Available from http://www.jstor.org/stable/2778065

Fischer, E. (2011). *Language communities of Twitter.* Available from http://www.flickr.com/photos/walkingsf/6277163176/in/photostream

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fleming, L., King, C. I. I. I., & Juda, A. I. (2007). Small worlds and regional

innovation. *Organization Science*, *18*(6), 938–954. Available from http://orgsci.journal.informs.org/cgi/content/abstract/18/6/938

Fleming, L., & Marx, M. (2006). Managing creativity in small worlds. *California Management Review*, *48*(4), 6.

Fogg, B., & Iizawa, D. (2008). Online persuasion in facebook and mixi: A cross-cultural comparison. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, & P. Ø hrstrø m (Eds.), *Persuasive technology* (Vol. 5033, pp. 35–46). Springer Berlin / Heidelberg. Available from http://dx.doi.org/10.1007/978-3-540-68504-3_4

Fogg, B. J. (2002, December). Persuasive technology: using computers to change what we think and do. *Ubiquity*, *2002*(December). Available from http://doi.acm.org/10.1145/763955.763957

Gaffney, D. (2010). #iranElection: Quantifying online activism. In *Proceedings of web science 10: Extending the frontiers of society on-line*. Raleigh, North Carolina, USA: Web Science Trust. Available from http://journal.webscience.org/295

Gandal, N. (2006). Native language and Internet usage. *International Journal of the Sociology of Language*, *182*, 25–40.

Garcia, I. (2013, January). Learning a language for free while translating the web. does duolingo work? *International Journal of English Linguistics*, *3*(1), 19–25. Available from http://www.ccsenet.org/journal/index.php/ijel/article/view/24236

Gaver, W. W. (1991). Technology affordances. In *Proceedings of the sigchi conference on human factors in computing systems: Reaching through technology* (pp. 79–84). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/108844.108856

Geiger, R. S., & Halfaker, A. (2013). Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 conference on computer supported*

cooperative work (pp. 861–870). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2441776.2441873

Ghorab, M. R., Zhou, D., Steichen, B., & Wade, V. (2011). Towards multilingual user models for personalized multilingual information retrieval. In *Proceedings of the first workshop on personalised multilingual hypertext retrieval* (pp. 42–49). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2047403.2047411

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Houghton Mifflin.

Giles, J. (2005, December). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–1. Available from http://dx.doi.org/10.1038/438900a

Goldsmith, J., & Wu, T. (2006). *Who controls the Internet? illusions of a borderless world*. New York, NY: Oxford University Press.

Gottron, T., & Lipka, N. (2010). A comparison of language identification approaches on short, query-style texts. In C. Gurrin et al. (Eds.), *Advances in information retrieval* (Vol. 5993, pp. 611–614). Springer Berlin / Heidelberg. Available from http://dx.doi.org/10.1007/978-3-642-12275-0_59

Graham, M., Hale, S. A., & Stephens, M. (2012). Featured graphic: Digital divide: The geography of internet access. *Environment and Planning A*, *44*(5), 1009–1010. Available from http://www.envplan.com/epa/fulltext/a44/a44497.pdf

Graham, M., & Zook, M. (2011). Visualizing global cyberscapes: Mapping user-generated placemarks. *Journal of Urban Technology*, *18*(1), 115–132. Available from http://www.tandfonline.com/doi/abs/10.1080/10630732.2011.578412

Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology*, *78*(6), 1360–1380. Available from http://www.jstor.org/stable/2776392

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, *1*, 201–233. Available from http://www.jstor.org/stable/202051

Grint, K., & Woolgar, S. (1997). *The machine at work: Technology, work and organization*. Polity.

Grosjean, F. (2010). *Bilingual: Life and reality*. Harvard University Press.

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining twitter as an imagined community. *American Behavioral Scientist*, *55*(10), 1294–1318. Available from http://abs.sagepub.com/content/55/10/1294.abstract

Gwet, K. L. (2010). *Handbook of interrater reliability* (2nd ed.). Advanced Analytics.

Halavais, A., & Lackaff, D. (2008). An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication*, *13*(2), 429–440. Available from http://dx.doi.org/10.1111/j.1083-6101.2008.00403.x

Hale, S. A. (2012). Impact of platform design on cross-language information exchange. In *Proceedings of the 2012 acm annual conference on human factors in computing systems extended abstracts* (pp. 1363–1368). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2212776.2212456

Hale, S. A., Margetts, H., & Yasseri, T. (2013). Petition growth and success rates on the UK No. 10 Downing Street website. In *Proceedings of the 5th annual acm web science conference* (pp. 132–138). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2464464.2464518

Halley, E. (1731). A proposal of a method for finding the longitude at sea within a degree, or twenty leagues. *Philosophical Transactions (1683-1775)*, *37*, 185–195. Available from http://www.jstor.org/stable/104083

Haruechaiyasak, C., & Damrongrat, C. (2008). Article recommendation based on a topic model for Wikipedia selection for schools. In G. Buchanan, M. Masoodian, & S. Cunningham (Eds.), *Digital libraries: Universal and ubiquitous*

access to information (Vol. 5362, pp. 339–342). Springer Berlin Heidelberg. Available from http://dx.doi.org/10.1007/978-3-540-89533-6_39

Haugen, E. (1969). *The Norwegian language in America: A study in bilingual behavior*. Indiana University Press.

Hecht, B. (2013). *The mining and application of diverse cultural perspectives in user-generated content*. Unpublished doctoral dissertation, Northwestern University.

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies* (pp. 11–20). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1556460.1556463

Hecht, B., & Gergle, D. (2010a). On the "localness" of user-generated content. In *Proceedings of the 2010 acm conference on computer supported cooperative work* (pp. 229–232). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1718918.1718962

Hecht, B., & Gergle, D. (2010b). The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 291–300). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1753326.1753370

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 237–246). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1978942.1978976

Herring, S. C. (1996). *Computer-mediated communication: Lingustic, social and cross-cultural perspectives*. Amsterdam: John Benjamins.

Herring, S. C., Paolillo, J. C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S.,

et al. (2007). Language networks on LiveJournal. In *Proceedings of the 40th annual hawaii international conference on system sciences*. Washington, DC, USA: IEEE Computer Society. Available from http://dx.doi.org/10.1109/HICSS.2007.320

Holloway, T., Bozicevic, M., & Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, *12*(3), 30–40. Available from http://dx.doi.org/10.1002/cplx.20164

Honeycutt, C., & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *System sciences, 2009. hicss '09. 42nd hawaii international conference on system sciences (hicss-42)* (pp. 1–10).

Hong, L., Convertino, G., & Chi, E. (2011). Language matters in Twitter: A large scale study. In *International AAAI conference on weblogs and social media* (pp. 518–521). Available from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2856

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389. Available from http://www.pnas.org/content/101/46/16385.abstract

Hsieh, H. C. L. (2014). Evaluating the effects of cultural preferences on website use. In P. Rau (Ed.), *Cross-cultural design* (Vol. 8528, pp. 162–173). Springer International Publishing. Available from http://dx.doi.org/10.1007/978-3-319-07308-8_16

Hu, C., Bederson, B. B., Resnik, P., & Kronrod, Y. (2011). Monotrans2: a new human computation system to support monolingual translation. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 1133–1136). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1978942.1979111

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tag-

ging in Twitter. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (pp. 173–178). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1810617.1810647

Hughes, T. P. (1994). Technological momentum. In M. R. Smith & L. Marx (Eds.), *Does technology drive history?: The dilemma of technological determinism* (pp. 101–113). MIT Press.

Humphreys, L., Von Pape, T., & Karnowski, V. (2013). Evolving mobile media: Uses and conceptualizations of the mobile Internet. *Journal of Computer-Mediated Communication*, *18*(4), 491–507. Available from http://dx.doi.org/10.1111/jcc4.12019

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, *35*(2), 441–456. Available from http://soc.sagepub.com/content/35/2/441.abstract

Hutchby, I. (2003). Affordances and the analysis of technologically mediated interaction: A response to Brian Rappert. *Sociology*, *37*(3), 581–589. Available from http://soc.sagepub.com/content/37/3/565.short

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th webkdd and 1st sna-kdd 2007 workshop on web mining and social network analysis* (pp. 56–65). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1348549.1348556

Juma, C., & Moyer, E. (2008). Broadband internet for africa. *Science*, *320*(5881), 1261.

Kelly, J., & Etling, B. (2008). *Mapping Iran's online public: Politics and culture in the Persian blogosphere.* Available from http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public

Keyes, O., & Hale, S. A. (2014). *Mobile user interfaces and user-generated content platforms.*

Kim, S., Weber, I., Wei, L., & Oh, A. (2014). Sociolinguistic analysis of Twitter in

multilingual societies. In *Proceedings of the 25th acm conference on hypertext and social media* (pp. 243–248). New York, NY, USA: ACM.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). *Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie.* Presented at alt.CHI, ACM SIGCHI Conference, San Jose, CA, USA.

Kleinfeld, J. (2002). The small world problem. *Society*, *39*(2), 61–66. Available from http://dx.doi.org/10.1007/BF02717530

Kling, R. (1996). Hopes and horrors: Technological utopianism and anti-utopianism in narratives of computerization. In R. Kling (Ed.), *Computerization and controversy* (2nd ed., pp. 40–58). Orlando, FL, USA: Academic Press, Inc.

Kornai, A. (2013, 10). Digital language death. *PLoS ONE*, *8*(10), e77056. Available from http://dx.doi.org/10.1371/journal.pone.0077056

Kramsch, C. (1998). *Language and culture*. Oxford: Oxford University Press.

Kraut, R. E., & Resnick, P. (2012). *Building successful online communities: Evidence-based social design*. The MIT Press.

Kriplean, T., Beschastnikh, I., & McDonald, D. W. (2008). Articulations of wikiwork: Uncovering valued work in Wikipedia through barnstars. In *Proceedings of the 2008 acm conference on computer supported cooperative work* (pp. 47–56). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1460563.1460573

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about twitter. In *Proceedings of the first workshop on online social networks* (pp. 19–24). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1397735.1397741

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. P. (2012). Geographic dissection of the Twitter network. In *Proceedings of the sixth international conference on weblogs and social media (icwsm-2012)*. Dublin, Ireland: The AAAI Press. Available from http://www.aaai.org/Library/

ICWSM/icwsm12contents.php

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *WWW '10: Proceedings of the 19th international conference on World Wide Web*, 591–600.

Lamb, R., & Kling, R. (2003). Reconceptualizing users as social actors in information systems research. *MIS Quarterly*, *27*(2), 197–236. Available from http://www.jstor.org/stable/30036529

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18–66.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Lee, P. (1996). *The whorf theory complex: A critical reconstruction.* Amsterdam: John Benjamins Publishing Company.

Leung, L., & Wei, R. (2000). More than just talk on the move: Uses and gratifications of the cellular phone. *Journalism & Mass Communication Quarterly*, *77*(2), 308–320. Available from http://jmq.sagepub.com/content/77/2/308.abstract

Lewis, P. M., Simons, G. F., & Fennig, C. D. (Eds.). (2014). Ethnologue: Languages of the world. Online version: http://www.ethnologue.com.

Li, T. J.-J., Sen, S., & Hecht, B. (2014). Leveraging advances in natural language processing to better understand Tobler's frist law of geography. In *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems.* New York, NY, USA: ACM.

Liao, H.-T. (2014). What do Chinese-language microblog users do with Baidu Baike and Chinese Wikipedia? a case study of information engagement. In *Proceedings of the international symposium on open collaboration* (pp. 23:1–23:10). New York, NY, USA: ACM. Available from http://doi.acm.org/

10.1145/2641580.2641611

Liao, H.-T., & Petzold, T. (2010). Analysing geo-linguistic dynamics of the World Wide Web: The use of cartograms and network analysis to understand linguistic development in Wikipedia. *Cultural Science*, *3*(2). Available from http://cultural-science.org/journal/index.php/culturalscience/article/view/44

Liljeblad, J. (2012). The implications of personal internet search for theories of global civil society. *International Journal of Technology, Knowledge and Society*, *8*(1), 103–114.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & Boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*. Available from http://ijoc.org/ojs/index.php/ijoc/article/view/1246/643

MacKinnon, R. (2008). Blogs and China correspondence: Lessons about global information flows. *Chinese Journal of Communication*, *1*(2), 242–257.

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of Twitter users. In *International aaai conference on weblogs and social media.* Available from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4605

Margetts, H., John, P., Hale, S. A., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action.* Princeton University Press.

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, *20*(8), 1025–1047. Available from http://dx.doi.org/10.1002/acp.1242

Marian, V., & Neisser, U. (2000, Sep). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Ho-

mophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv:1309.4168*. Available from `http://arxiv.org/abs/1309.4168`

Milne, D., & Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *In proceedings of the aaai 2008 workshop on wikipedia and artificial intelligence.* Chicago, IL, USA: AAAI.

Ministry of Foreign Affairs of Japan. (2008). 米軍人等の施設・区域内外居住者の人数について*(*全国*)(*平成*20*年*3*月*31*日時点*) (Information on the number of people living on and off US military related facilities [Nationwide][March 31, 2008]).* `http://www.mofa.go.jp/mofaj/press/release/h20/6/1181033_910.html`.

Mohammadi, M., & GhasemAghaee, N. (2010, March). Building bilingual parallel corpora based on Wikipedia. In *Computer engineering and applications (iccea), 2010 second international conference on* (Vol. 2, pp. 264–268).

Negroponte, N. (1995). *Being digital.* New York: Knopf.

Neubig, G., & Duh, K. (2013). How much is said in a tweet? A multilingual, information-theoretic perspective. In *Aaai spring symposium series.* Available from `http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5698`

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, *69*(2). Available from `http://link.aps.org/doi/10.1103/PhysRevE.69.026113`

Nielsen, J. (1993). *Usability engineering.* Academic Press.

Nordenstreng, K., & Varis, T. (1974). Television traffic: A one-way street? A survey and analysis of the international flow of television programme material. *Reports and Papers on Mass Communication*(70).

Norman, D. (1988). *The design of everyday things.* Basic Books.

Okinawa Prefecture. (2013). 沖縄の米軍基地及び自衛隊基地（統計資料集）

(US military bases and Japan Self-Defense Force bases in Okinawa [Compiled statistics]). http://www.pref.okinawa.jp/site/chijiko/kichitai/toukeisiryousyu2503.html.

Olson, D. R., J., K. K., Marc, P., Cecile, V., & Lone, S. (2013, 10). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Comput Biol*, *9*(10), e1003256. Available from http://dx.doi.org/10.1371/journal.pcbi.1003256

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the inequality of contributions to Wikipedia. In *Hawaii international conference on system sciences, proceedings of the 41st annual* (p. 304).

O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing 'flaming' and other problematic messages. *New Media & Society*, *5*(1), 69–94. Available from http://nms.sagepub.com/content/5/1/69.abstract

Ottaviano, G. I. P., & Peri, G. (2006). The economic value of cultural diversity: Evidence from US cities. *Journal of Economic Geography*, *6*(1), 9–44. Available from http://joeg.oxfordjournals.org/content/6/1/9.abstract

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November). *The pagerank citation ranking: Bringing order to the web.* (Technical Report No. 1999-66). Stanford InfoLab. Available from http://ilpubs.stanford.edu:8090/422

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton, NJ: Princeton University Press.

Palen, L., Vieweg, S., & Anderson, K. M. (2010). Supporting "everyday analysts" in safety- and time-critical situations. *The Information Society*, *27*(1), 52–62. Available from http://dx.doi.org/10.1080/01972243.2011.534370

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* London: Viking.

Perry, M., O'hara, K., Sellen, A., Brown, B., & Harper, R. (2001, Dec). Deal-

ing with mobility: Understanding access anytime, anywhere. *ACM Trans. Comput.-Hum. Interact.*, *8*(4), 323–347. Available from http://doi.acm.org/10.1145/504704.504707

Pew Research Internet Project. (2014). *Mobile technology fact sheet.* Available from http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet

Pfeil, U., Zaphiris, P., & Ang, C. S. (2006). Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, *12*(1), 88–113. Available from http://dx.doi.org/10.1111/j.1083-6101.2006.00316.x

Pimienta, D., Prado, D., & Blanco, Á. (2009). *Twelve years of measuring linguistic diversity in the Internet: Balance and perspectives.* Paris: United Nations Educational, Scientific and Cultural Organization. Available from http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Advances in information retrieval* (Vol. 4956, pp. 522–530). Springer Berlin Heidelberg. Available from http://dx.doi.org/10.1007/978-3-540-78646-7_51

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 international acm conference on supporting group work* (pp. 259–268). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1316624.1316663

Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century the 2006 johan skytte prize lecture. *Scandinavian Political Studies*, *30*(2), 137–174. Available from http://dx.doi.org/10.1111/j.1467-9477.2007.00176.x

Raghavan, U. N., Albert, R., & Kumara, S. (2007, September). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, *76*(3), 36106. Available from http://link.aps.org/doi/10.1103/PhysRevE.76.036106

Rappert, B. (2003). Technologies, texts and possibilities: A reply to Hutchby. *Sociology*, *37*(3), 565–580. Available from http://soc.sagepub.com/content/37/3/565.short

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 695–704). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1963405.1963503

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, *111*(52), E5616–E5622. Available from http://www.pnas.org/content/111/52/E5616.abstract

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. Available from http://dx.doi.org/10.1140/epjst/e2010-01179-1

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. Available from http://www.pnas.org/content/105/4/1118.abstract

Russo, P., & Boor, S. (1993). How fluent is your interface?: Designing for international users. In *Proceedings of the interact '93 and chi '93 conference on human factors in computing systems* (pp. 342–347). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/169059.169274

Sadilek, A., Kautz, H., & Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth acm international conference on web search and data mining* (pp. 723–732). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2124295.2124380

Semiocast. (2010). *Half of messages on Twitter are not in English: Japanese is the second most used language.* Semiocast Press Release, Paris, France. Available from http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf

Shelton, T., Zook, M., & Graham, M. (2012). The technology of religion: Mapping religious cyberscapes. *The Professional Geographer*, *64*(4), 602–617.

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 403–411). Stroudsburg, PA, USA: Association for Computational Linguistics. Available from http://dl.acm.org/citation.cfm?id=1857999.1858062

State, B., Park, P., Weber, I., Mejova, Y., & Macy, M. (2013, February). The mesh of civilizations and international email flows [misc]. Available from http://arxiv.org/abs/1303.0045

Statistics Canada. (2012). *Linguistic characteristics of Canadians.* Catalogue no. 98-314-X2011001. Available from http://www12.statcan.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm

Steichen, B., Ghorab, M., O'Connor, A., Lawless, S., & Wade, V. (2014). Towards personalized multilingual information access - exploring the browsing and search behavior of multilingual users. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User modeling, adaptation, and personalization* (Vol. 8538, pp. 435–446). Springer International Publishing. Available from http://dx.doi.org/10.1007/978-3-319-08786-3_39

Stork, C., Calandro, E., & Gillwald, A. (2012). *Internet going mobile: Internet access and usage in eleven African countries.* (19th ITS Biennial Conference 2012, Bangkok, Thailand, 18 - 21 November 2012: Moving Forward with Future Technologies: Opening a Platform for All)

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of twitter networks. *Social Networks*, *34*(1), 73–81. Available from http://www.sciencedirect.com/science/article/pii/S0378873311000359

Thelwall, M., Tang, R., & Price, L. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics*, *56*(3), 417–432.

Thomas, D. A., & Ely, R. J. (1996). Making differences matter: A new paradigm for managing diversity. *Harvard Business Review*, *74*(5), 79–90.

Thurlow, C., Lengel, L., & Tomic, A. (2004). *Computer mediated communication: Social interaction and the Internet*. London: Sage.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*(4), 425–443. Available from http://www.jstor.org/stable/2786545

Uzzi, B., Amaral, L. A. N., & Reed-Tsochas, F. (2007). Small-world networks and management science research: a review. *European Management Review*, *4*(2), 77–91. Available from http://dx.doi.org/10.1057/palgrave.emr.1500078

Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. *The American Journal of Sociology*, *111*(2), 447–504. Available from http://www.jstor.org/stable/10.1086/432782

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on human factors in computing systems* (pp. 1079–1088). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1753326.1753486

Viger, F., & Latapy, M. (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In L. Wang (Ed.), *Computing and combinatorics* (Vol. 3595, pp. 440–449). Springer Berlin Heidelberg. Available from http://dx.doi.org/10.1007/11533719_45

Warschauer, M., Said, G. R. E., & Zohry, A. (2002). Language choice online: Globalization and identity in Egypt. *Journal of Computer-Mediated Communication*, *7*(4). Available from http://jcmc.indiana.edu/vol7/issue4/warschauer.html

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442. Available from http://dx.doi.org/10.1038/30918

Wei, C. Y., & Kolko, B. E. (2005). Resistance to globalization: Language and Internet diffusion patterns in Uzbekistan. *New Review of Hypermedia and Multimedia*, *11*(2), 205–220.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., et al. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iconference* (pp. 122–129). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1940761.1940778

Whorf, B. (1940). Science and linguistics: How words and customs affect reasoning. *MIT Technology Review*, *42*(6), 220–231.

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology*, *63*(8), 1631–1646. Available from http://dx.doi.org/10.1002/asi.22713

Wilks, Y. (2009). *Machine translation: Its scope and limits*. New Yrok: Springer.

Wing, B., & Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Acl* (pp. 955–964). The Association for Computer Linguistics. Available from http://

dblp.uni-trier.de/db/conf/acl/acl2011.html#WingB11

Yasseri, T., Hale, S. A., & Margetts, H. (2014). *Modeling the rise in internet-based petitions.* Available from http://arxiv.org/abs/1308.0239 (Under review.)

Yasseri, T., Sumi, R., & Kertész, J. (2012). Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS ONE*, *7*(1), e30091. Available from http://dx.doi.org/10.1371/journal.pone.0030091

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PLoS ONE*, *7*(6), e38869. Available from http://dx.doi.org/10.1371/journal.pone.0038869

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010, July). Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical & Health Policy*, *2*(2), 7–33. Available from http://www.psocommons.org/wmhp/vol2/iss2/art2

Zuckerman, E. (2008). Meet the bridgebloggers. *Public Choice*, *134*(1), 47–65.

Zuckerman, E. (2013). *Rewire: Digital cosmopolitans in the age of connection.* London: W. W. Norton & Company.