Impact of Platform Design on Cross-language Information Exchange

Scott A. Hale
Oxford Internet Institute
University of Oxford
1 St Giles, Oxford UK
scott.hale@oii.ox.ac.uk

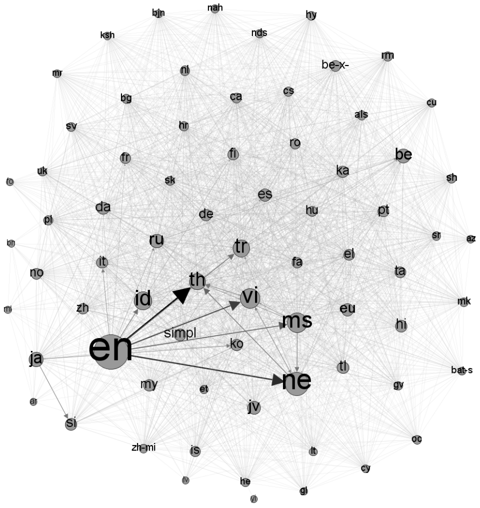


Figure 1: Relationships between articles about the Tohoku earthquake in different language editions of Wikipedia.

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract

This paper describes two case studies examining the impact of platform design on cross-language communications. The sharing of off-site hyperlinks between language editions of Wikipedia and between users on Twitter with different languages in their user descriptions are analyzed and compared in the context of the 2011 Tohoku earthquake and tsunami in Japan. The paper finds that a greater number of links are shared across languages on Twitter, while a higher percentage of links are shared between Wikipedia articles. The higher percentage of links being shared on Wikipedia is attributed to the persistence of links and the ability for users to link articles on the same topic together across languages.

Author Keywords

User Interfaces, Information Sharing, Localization, Internationalization, Interaction, Web-based Interaction, Wikipedia, Twitter

ACM Classification Keywords

H.5.3 [Information interfaces and presentation (e.g., HCI)]: Group and Organization Interfaces;

General Terms

Design, Languages, Human Factors, Measurement

Introduction

The percentage of webpages in English and the percentage of web users speaking English are in steady decline as more users creating more content in many different languages come online [9]. This trend is set to continue with the introduction of high-speed Internet in Africa, the increasing use of mobile phones for Internet access throughout the world, and the continued rising standards of living in much of the global South. This leads to pressing questions for Internet-based companies about the best way to adapt their platforms to accommodate users of multiple languages. These questions are particularly important for companies relying on user-generated content, which can now be expected to be posted and consumed in more languages. Platform providers must decide what language tools (machine translation, allowance of human/crowd translations, dictionary tools, linking of similar topics across languages, etc.) are appropriate. The important nature of this choice is indicated by the many very active user discussions on how monolingual communities like StackExchange and Quora might expand. Surprisingly, however, there is a lack of relevant academic and commercial research.

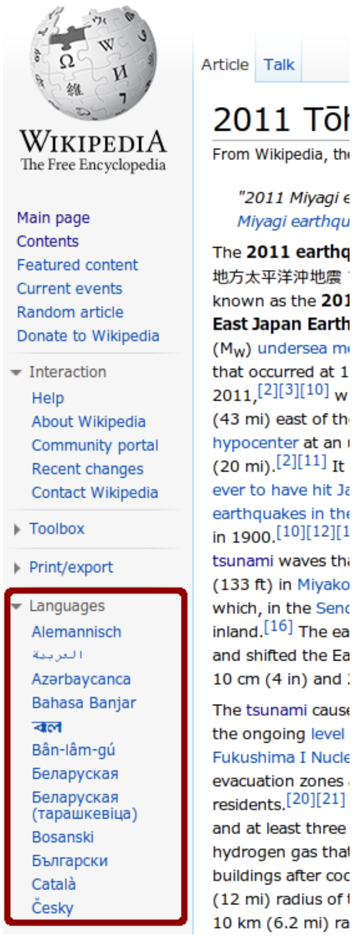
Sharing of information between languages on a platform may enable more efficient operations (e.g. less redundant information) and better outcomes (e.g. quicker discovery of good ideas). Higher amounts of cross-language information sharing mean a larger group of users see and potentially answer open questions, find missing references, or contribute needed content. This can lead to less duplication of content and effort through better coordination. In place of finding citations, images, and diagrams independently for each language version of a Wikipedia article, increased multilingual sharing could, for example, speed article creation and updates by better pooling user contributions. Within Twitter, cross-language information sharing may result in more information provided more quickly to users.

This paper investigates the levels of cross-language information sharing on Wikipedia and Twitter following the 2011 Tohoku earthquake and tsunami in Japan. The international attention of this event makes it a good candidate for measuring cross-language information sharing, and the different approaches of Wikipedia and Twitter to internationalization and localization (as described below) along with their international user bases make these two platforms good cases for comparison.

Related Work

Only a few studies have analyzed differences between languages on user-generated content sites, and even fewer have analyzed the sharing of information between languages. Studies have found that the uses of platforms differ across languages [7] and that there is a self-focus bias in the topics written about. That is, the topics written about are more likely to be related to regions where the language is spoken [4]. This self-focus bias leads to different content being available in different languages. Hecht and Gergle [5] found little overlap between article coverage and article content in different languages on Wikipedia. Tweets of different languages on Twitter also use different hashtags and URLs although there is some overlap [7]. Language and location have been shown to be strong determining factors of the following–follower relationships on Twitter [10], and Yamashita el al. [11] have described language as the "biggest barrier to intercultural collaboration."

Nevertheless, studies of the flows of information between languages online suggest connections between languages



2011 Tōł
From Wikipedia, the

"2011 Miyagi e
Miyagi earthqu

The **2011 earthq**
地方太平洋沖地震
known as the **201**
East Japan Earth
(M_W) undersea m
that occurred at 1
2011,[2][3][10] w
(43 mi) east of th
hypocenter at an
(20 mi).[2][11] It
ever to have hit Ja
earthquakes in the
in 1900.[10][12][1
tsunami waves th
(133 ft) in Miyako
which, in the Senc
inland.[16] The ea
and shifted the Ea
10 cm (4 in) and

The tsunami caus
the ongoing level
Fukushima I Nucle
evacuation zones
residents.[20][21]
and at least three
hydrogen gas that
buildings after coo
(12 mi) radius of
10 km (6.2 mi) ra

Figure 2: Interlanguage links on Wikipedia. A red box has been added to the screen capture to indicate the location of interlanguage links on Wikipedia articles. These links connect articles on similar topics across languages.

are few but important [1, 2, 6, 8]. In particular, MacKinnon [8] shows a reliance of foreign correspondents in China upon English-language blogs about China, which "suggests that English-language 'bridge blogs' about China have greater direct influence on China correspondents than Chinese-language blogs" (p. 19). Past studies show content introduced in English is likely to spread to other languages, but conflict on how likely foreign-language content is to be used in English. While Ford [1] found English webpages with a high page rank were likely to link to pages of other languages, Hale [2] found bloggers writing in English were less likely than those writing in Spanish or Japanese to link to another language.

Case selection and methods

This paper analyzes the sharing of hyperlinks about the 2011 Tohoku earthquake and tsunami in Japan. This large event held global attention and interest for an extended period of time and is a logical topic for which to expect the sharing of information across languages. Link sharing on Twitter and Wikipedia are compared as these large international platforms both carried recent news about the event and yet have very different designs particularly in regards to internationalization and localization. Twitter has a single data store which is searched irrespective of language, while Wikipedia language editions are more independent. Wikipedia, however, allows related articles to be linked across languages with interlanguage links.

This research will investigate two main questions related to cross-language link sharing. First, to what extent are links shared across languages on Twitter and Wikipedia related to a specific event, and what is the impact of platform design on this level of sharing? Second, what is the relationship between languages sending and receiving links? That is, are there languages that primarily only receive or only send links? The first question is important to understand how information is shared between languages and therefore to what extent communication occurs across languages versus being confined within languages. The second question analyzes what percentage of the overall information contained in links is present in a given language. The level of cross-language information sharing is important to the overall amount of information and the diversity of information available to users considering content from only one language. This is especially pertinent for news gathering and following fast-breaking news.

The English Wikipedia article, "2011 Tohoku earthquake and tsunami," was used as a starting point, and all interlanguage links to parallel Wikipedia articles in other language editions were collected. These interlanguage links are added by both automated bots and humans and are usually relatively complete. For each language edition, all revisions of the article from its creation until January 4, 2012, were downloaded through the Wikipedia API. Each time a new hyperlink was added to the article, the link itself and the timestamp of the revision were recorded. This resulted in a database table containing the hyperlinks in each language edition and the timestamp at which each hyperlink was first added to the given language edition.

Twitter data was collected via the `statuses/filter` streaming API. All tweets mentioning "Japan" or its translation were recorded starting on March 11, 2011, at 16:24 UTC, slightly less than 11 hours after the earthquake struck Japan. Although more data was collected, this paper analyzes one week of Twitter data starting on March 11. The languages of users' profile descriptions (biographies) were identified with the

Compact Language Detection kit available as part of the Chromium open-source software project (and used within Google Chrome). This identification algorithm has been found to perform well on tweets [3] and on blog content [2]. Similar to Wikipedia, the timestamp, URL, and detected language of the user's profile description were recorded for every tweet containing a link.

The language of users' profile descriptions was selected so that even unaltered retweets containing a link could be tracked through users of different languages. In addition, shortened URLs (e.g. links from bit.ly, t.co, tinyurl.com, etc.) were not expanded as this paper is interested in how these links are shared within the Twitter ecosystem, and the links were not likely to change when being shared within Twitter.

The data from Wikipedia and Twitter allows for analysis of the total number of hyperlinks in each language and how many of these hyperlinks are shared with another language. Furthermore, the first language using a hyperlink can be determined by comparing the timestamps of when the link first appeared in each language. The actual diffusion path from one language to another is more difficult to determine (particularly in Wikipedia) and is the subject of ongoing research. This is further complicated by the possibility of a hyperlink being used in two languages independently and without knowledge of its use in the other language. Within Twitter, the use of shortened URLs reduces the chances of the same link being introduced independently. Even so, the current dataset represents the first attempt to relate design with cross-language information exchange on these user-generated content platforms. The data provides a rough picture and highlights areas for future work.

	Wikipedia	Twitter
Overall	1,179	1,221,972
Chinese	61	1,531
English	294	472,628
French	55	11,881
Indonesian	6	3,587
Japanese	139	176,272
Korean	26	5,215
Portuguese	15	26,524
Russian	93	2,138
Spanish	54	72,731

Table 1: Number of unique users on Wikipedia and Twitter contributing content in selected languages. Additional languages are available on the author's homepage.

Results

Overall, there were 75 language editions of Wikipedia with an article about the earthquake containing links. Collectively, these articles contained 4,174 unique links. Of these, 1,105 links (or 26%) were found in multiple languages. The top link (linked to by 53 language editions) was to the United States Geological Service's page listing the details of the earthquake. Most links that were shared between different language editions appeared first in the English article or, to a lesser extent, in the Japanese article. Table 2 shows details for several languages, and Figure 1 shows a network diagram of the relationships between the language editions.

Comparatively, a much greater volume of links were found in Twitter despite the shorter collection period. 929,436 unique links appeared in Twitter, with 7% of these appearing in multiple languages. Most links that were shared between languages were first tweeted by a user with a user description in English. Links tweeted by users with Spanish or Japanese user descriptions were also shared more than other languages, although to a lesser extent than those tweeted by users with English user descriptions.

Overall, a lower percentage of links was shared between users with different languages in their descriptions on Twitter (7%) than between language editions of Wikipedia (26%). However, the volume on Twitter was so great that even the 7% of links shared across multiple languages (a total of 63,274 links) is greater than the total number of unique links within articles from all language editions of Wikipedia combined. In addition, there are some important variations between languages: a greater percentage of links in Chinese and French were shared with another language on Twitter than on

	Wikipedia				Twitter			
	Unique links	Shared links†	Orig. links‡	Foreign links∗	Unique links	Shared links†	Orig. links‡	Foreign links∗
Overall	4,174	26.5%			929,436	6.8%		
Chinese	246	31.3%	1.7%	1.5%	3,628	44.4%	0.4%	0.2%
English	1,297	61.4%	64.3%	3.7%	672,351	8.7%	50.8%	10.9%
French	458	14.8%	0.8%	1.6%	28,092	27.9%	4.1%	0.6%
Indonesian	140	95.7%	0.1%	3.3%	8,396	47.7%	1.3%	0.3%
Japanese	715	32.3%	14.9%	2.1%	80,437	24.1%	10.4%	1.5%
Korean	259	58.7%	1.8%	3.4%	5,479	55.4%	0.7%	0.3%
Portuguese	67	68.7%	0.2%	1.1%	31,717	28.7%	4.3%	0.7%
Russian	349	36.1%	4.2%	2.1%	4,381	37.0%	0.4%	0.1%
Spanish	153	45.1%	1.1%	1.4%	106,844	22.9%	18.4%	1.6%

Table 2: Sharing of links on Wikipedia and Twitter in selected languages. Additional languages are available on the author's homepage. †Percentage of links in this language shared with at least one other language; ‡Percentage of all shared links appearing first in this language; ∗Percentage of all links started in another language and appearing in this language

Wikipedia, for example. Most links shared between languages overwhelmingly originated in English on both platforms, but Spanish users originated and shared a greater portion of links on Twitter than on Wikipedia.

Discussion

Twitter had substantially more links being shared in each language compared to Wikipedia despite the more limited time frame for data collection on Twitter. However, a greater percentage of links was shared between languages on Wikipedia as compared to Twitter. The design and use of these platforms as well as the composition of their user bases may partially explain these differences.

The design of Twitter as a single platform with a single data store across all languages seems to facilitate the sharing of links. Search occurs irrespective of the majority language of the tweet or the language of the user. This results in users of Twitter more likely encountering foreign-language information and then reposting that information. However, the ephemeral nature of tweets and the large volume of links results in an overall lower percentage of links being shared across languages on Twitter. In contrast, the ability to link articles across languages, the clearer topic space, and the more persistent nature of content on Wikipedia allowed a larger percentage of links to be shared.

Conclusions and future work

Important variations occur in the amount of link sharing between languages on different platforms. This variation in link sharing is partially explained by variations in design and user composition, and further study is needed to better isolate the impacts of design more specifically.

A substantially larger percentage of new links on both Twitter and Wikipedia are introduced in English. In contrast to previous research with blogs [2], English also receives many links introduced first in another language. However, the overall percentage of links started in a foreign language and seen in English is low and suggests there are opportunities for discovery of different information in foreign languages. This is particularly true for users of non-English languages where an even smaller number of foreign-language links are seen.

How speakers of different languages interact with each other and with a platform given its design is an important area requiring further HCI research. This study is exploratory and many future avenues are being pursued (better tracking of diffusion paths, specific vs general events, changes in sharing over time, A/B experimental testing). Nevertheless, this study represents the first attempt to investigate the diffusion of information across languages on platforms with users of many languages. The results show cross-language information sharing is occurring on these platforms and that there are important variations in this sharing by language and by platform. Further HCI research should better isolate specific elements of design affecting the level of cross-language sharing on platforms.

References

[1] Ford, D., and Batson, J. Languages of the World (Wide Web). *Google Research Blog* (July 2011).

[2] Hale, S. A. Net increase? Cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication 17*, 2 (2012), 135–151.

[3] Hale, S. A., Gaffney, D., and Graham, M. Where in the world are you? Geolocation and language identification in Twitter. Working paper, 2012.

[4] Hecht, B., and Gergle, D. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09, ACM (New York, NY, USA, 2009), 11–20.

[5] Hecht, B., and Gergle, D. 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*, CHI '10, ACM (New York, NY, USA, 2010), 291–300.

[6] Herring, S. C., Paolillo, J. C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L. A., and Clark, B. Language networks on LiveJournal. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, IEEE Computer Society (Washington, DC, USA, 2007).

[7] Hong, L., Convertino, G., and Chi, E. Language matters in Twitter: A large scale study. In *International AAAI Conference on Weblogs and Social Media* (2011), 518–521.

[8] MacKinnon, R. Blogs and China correspondence: Lessons about global information flows. *Chinese Journal of Communication 1*, 2 (2008), 242–257.

[9] Pimienta, D., Prado, D., and Blanco, A. Twelve years of measuring linguistic diversity in the Internet: Balance and perspectives, 2009.

[10] Takhteyev, Y., Gruzd, A., and Wellman, B. Geography of Twitter networks. *Social Networks* (2011), 1–26.

[11] Yamashita, N., Inaba, R., Kuzuoka, H., and Ishida, T. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 679–688.