

## Net Increase? Cross-lingual Linking in the Blogosphere

Scott A. Hale, Oxford Internet Institute, University of Oxford

### Abstract

This research analyzes linguistic barriers and cross-lingual interaction through link analysis of more than 100,000 blogs discussing the 2010 Haitian earthquake in English, Spanish, and Japanese. In addition, cross-lingual hyperlinks are qualitatively coded. This study finds English-language blogs are significantly less likely to link cross-lingually than Spanish or Japanese blogs. However, bloggers' awareness of foreign language content increases over time. Personal blogs contain most cross-lingual links, and these links point to (primarily English-language) media. Finally, most cross-lingual links in the dataset signal a citation or reference relationship while a smaller number of cross-lingual links signal a translation. Although most bloggers link to other blogs in the same language, the dataset reveals a surprising level of human translation in the blogosphere.

*Keywords:* webometrics, link analysis, language, Internet, intercultural, cross-lingual

### Author's Note

The author would like to thank Dr. Eric T. Meyer, Dr. Sandra Gonzalez-Bailon, and Dr. Bernie Hogan of the Oxford Internet Institute, University of Oxford, for their help and input throughout the course of this research. I would also like to thank Jorge Luis Salcedo Maldonado, Universitat Autònoma de Barcelona, and Yuko Tokumori for their assistance with the qualitative coding.

### About the Author

Scott A. Hale (scott.hale@oii.ox.ac.uk) is a doctoral candidate and research assistant at the Oxford Internet Institute, University of Oxford, interested in language separation online and the effects of design upon the transmission of information between languages.

Hale, S. A. (Forthcoming) Net Increase? Cross-lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*.

### Net Increase? Cross-lingual Linking in the Blogosphere

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 two 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.<sup>1</sup> 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.

## Literature

### 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

---

<sup>1</sup>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.

(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, 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.<sup>2</sup>

### **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

---

<sup>2</sup>Gerrand (2007) provides an overview of further studies looking at the number of web pages in various languages.

(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 exchange 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 US political blogs. These studies (e.g. Adamic & Glance, 2005; Hargittai et al., 2007) showed bloggers in the US 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 inter-cluster 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 US 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.<sup>3</sup> 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 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

---

<sup>3</sup> More information about this pilot study is available on the author's homepage:  
[http:// www.scottahale.net/](http://www.scottahale.net/)

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.

### **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,<sup>4</sup> but how common it is and the nature of cross-lingual links

---

<sup>4</sup>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

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.

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 US 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

---

would, however, be a cross-lingual blog.

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).

## **Data and Methods**

### **Data Collection**

This research develops new methods, tests them in a pilot study, and then applies them to a larger dataset.<sup>5</sup> Search queries for “haiti” and “earthquake” in Japanese (ハイチ and 地震), English, and Spanish ([haiti or haiti] 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.<sup>6</sup> 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 non-blog 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

---

<sup>5</sup>Full source code is available from the author upon request.

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

UCINET.

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.

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<sup>7</sup> 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

---

<sup>7</sup>[http://src.chromium.org/viewvc/chrome/trunk/src/third party/cld/](http://src.chromium.org/viewvc/chrome/trunk/src/third_party/cld/)

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

### **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,  $\kappa$ , 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.

### **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 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).

Insert Table 1 Here

### **Links Between Language Groups**

The links between language groups, given in Table 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 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.<sup>8</sup> Spanish accounts for 37% of the score, and Japanese accounts for 21%.

Insert Table 2 Here

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 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.

---

<sup>8</sup>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.

### **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 1 shows the modularity score of the blog network over the 45-day collection period employing a 5-day smoothing window.

Insert Figure 1 Here

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.

### **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.

Insert Table 3 Here

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 non-governmental 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 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 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% 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.

Insert Figure 2 Here

Insert Figure 3 Here

English-language media are the most central node in the network of cross-lingual links (Figure 3). The largest single destination of cross-lingual links is to a collection of photos

published by the Denver Post.<sup>9</sup> 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, non-news 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.

### **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 non-existent paths between blogs of different languages. Where these paths 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 under realized.

---

<sup>9</sup><http://blogs.denverpost.com/captured/2010/01/13/earthquake-in-haiti/>

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 non-professionals, 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.<sup>10</sup> 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 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<sup>11</sup> asking 21 questions about cross-

---

<sup>10</sup>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.

<sup>11</sup>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.

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.

### **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-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).

### **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 US 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.

### **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

- Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. *International conference on knowledge discovery and data mining: Proceedings of the 3rd international workshop on link discovery*, 36–43.
- Associated Press. (2010, February 10). Haiti raises earthquake toll to 230,000. *The Washington Post*. Retrieved 18 July 2010, from <http://www.washingtonpost.com/wp-dyn/content/article/2010/02/09/AR2010020904447.html>
- Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven, CT: Yale University Press.
- Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.
- Burt, R. S. (2004). Structural holes and good ideas. *The American Journal of Sociology*, 110 (2), 349–399.
- Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). The structure of broad topics on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web* (pp. 251–262). New York, NY, USA: ACM.
- Cozens, C. (2005, March 21). AFP sues Google over copyrighted content. *guardian.co.uk* . Retrieved 20 July 2010, from <http://www.guardian.co.uk/technology/2005/mar/21/media.newmedia>
- Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.
- Delwiche, A. (2005). Agenda–setting, opinion leadership, and the world of Web logs. *First Monday*, 10 (12).
- Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2009). *Mapping the Arabic blogosphere: Politics,*

- culture, and dissent*. Berkman Center Research Publication. Retrieved 12 January 2011, from [http://cyber.law.harvard.edu/publications/2009/Mapping the Arabic Blogosphere](http://cyber.law.harvard.edu/publications/2009/Mapping%20the%20Arabic%20Blogosphere)
- Federal Trade Commission. (2010, June). *How will journalism survive the Internet age?* [Workshop]. Retrieved 1 May 2011, from <http://www.ftc.gov/opp/workshops/news/index.shtml>
- Fine, G. A., & Kleinman, S. (1979). Rethinking subculture: An interactionist analysis. *The American Journal of Sociology*, 85 (1), 1–20.
- Fleming, L., King, C. I., & Juda, A. I. (2007). Small worlds and regional innovation. *Organization Science*, 18 (6), 938–954.
- Fujimura, K., Inoue, T., & Sugisaki, M. (2005). The eigenrumor algorithm for ranking blogs. In *WWW Workshop on the weblogging ecosystem*. Chiba, Japan.
- Gerrand, P. (2007). Estimating linguistic diversity on the Internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication*, 12 (4).
- Graham, M. (2009, Dec 2). Wikipedia's known unknowns. *guardian.co.uk* . Retrieved 22 Jan 2010, from <http://www.guardian.co.uk/technology/2009/dec/02/wikipedia-known-unknowns-geotagging-knowledge>
- Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology*, 78 (6), 1360–1380.
- Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, 1, 201–233.
- Hansell, S. (2008, June 16). The Associated Press to set guidelines for using its articles in blogs. *The New York Times*. Retrieved 20 July 2010, from <http://www.nytimes.com/2008/06/16/business/media/16ap.html>
- Hargittai, E., Gallo, J., & Kane, M. (2007). Cross-ideological discussions among conservative

- and liberal bloggers. *Public Choice*, 134 (1 & 2), 67–89.
- Hecht, B., & Gergle, D. (2010). 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, New York: ACM.
- Johnson, S. (2010). *Where good ideas come from: The natural history of innovation* (1st ed.). New York: Riverhead Hardcover.
- Kelly, J., & Etling, B. (2008). *Mapping Iran's online public: Politics and culture in the Persian blogosphere*. Berkman Center Research Publication. Retrieved 12 January 2011, from [http://cyber.law.harvard.edu/publications/2008/Mapping\\_Irans\\_Online\\_Public](http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public)
- Lazarsfeld, P. & Merton, R. (1954) Friendship as social process: A substantive and methodological analysis. In *Freedom and Control in Modern Society* (pp. 18–16). New York: Van Nostrand.
- Ledbetter, B. C. (2008, June 16). AP: the Internet's big bully. *guardian.co.uk* . Retrieved 20 July 2010, from <http://pajamasmedia.com/blog/ap-the-internets-big-bully/>
- Lombard, M., Snyder-Duch, J., & Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28 (4), 587–604.
- MacKinnon, R. (2008). Blogs and China correspondence: lessons about global information flows. *Chinese Journal of Communication*, 1 (2), 242–257.
- Margetts, H., & Hale, S. A. (2010). User experiments. In *Phase 1: Life events report* (pp. 79–96). Study on user expectations of a life events approach for designing e-government services, European Commission.
- McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27 , 415–444.

- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69 (2).
- 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).
- Pimienta, D., Prado, D., & Blanco, A. (2009). *Twelve years of measuring linguistic diversity in the Internet: Balance and perspectives*. Paris: United Nations Educational, Scientific and Cultural Organization.
- Schrenk, M. (2007). *Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL*. San Francisco, CA: No Starch Press.
- Sifry, D. (2007). The state of the live Web, April 2007. Retrieved 20 April 2010, from <http://www.sifry.com/alerts/archives/000493.html>
- Thelwall, M. (2008). Extracting accurate and complete results from search engines: case study Windows Live. *Journal of the American Society for Information Science and Technology*, 59 (1), 38–50.
- Thelwall, M., Tang, R., & Price, L. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics*, 56 (3), 417–432.
- Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. *The American Journal of Sociology*, 111 (2), pp. 447–504.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393, 440–442.
- Waugh, A., Pei, L., Fowler, J., Mucha, P., & Porter, M. (2009). Party polarization in congress: A social networks approach. *arXiv:0907.3509v1 [physics.soc-ph]*.
- Zuckerman, E. (2008). Meet the bridgebloggers. *Public Choice*, 134 (1), 47–65.

Table 1: *Language Distribution*

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

*Note.* 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.

Table 2: *Dataset hyperlinks*

Source	Destination						Total
	English	Spanish		Japanese			
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	

*Note.* 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.

Table 3: *Relationship of Cross-lingual Linking Blogs*

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

Figure 1: *Modularity over Time*

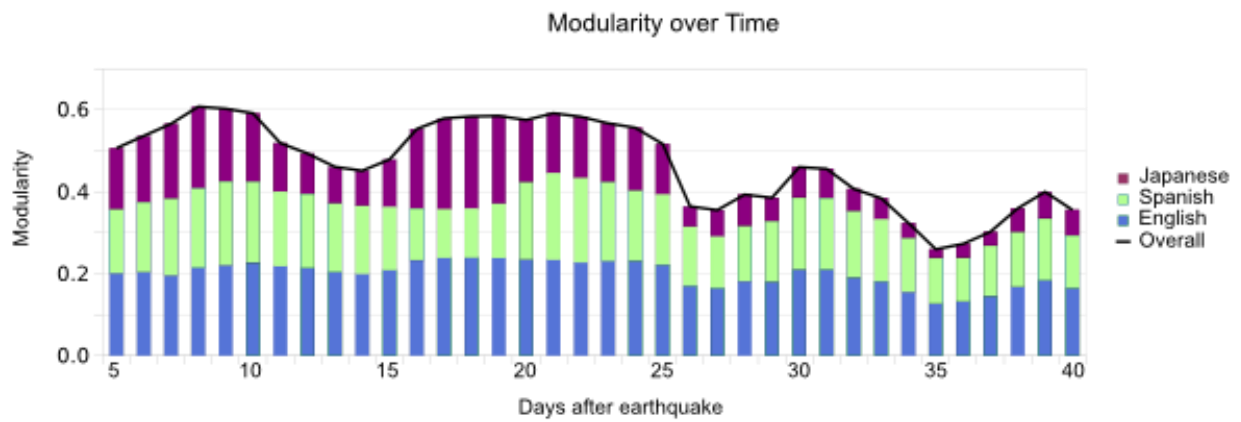


Figure 2: *Origins and Destinations of Cross-lingual Hyperlinks*

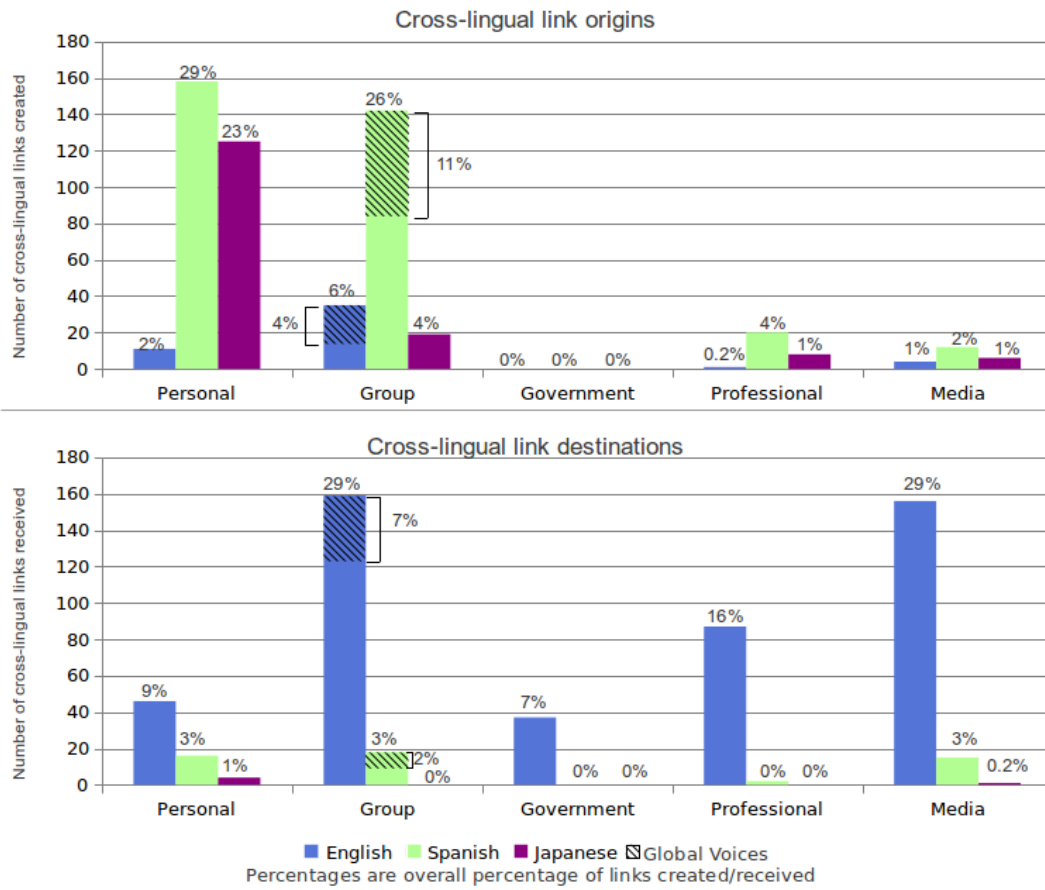
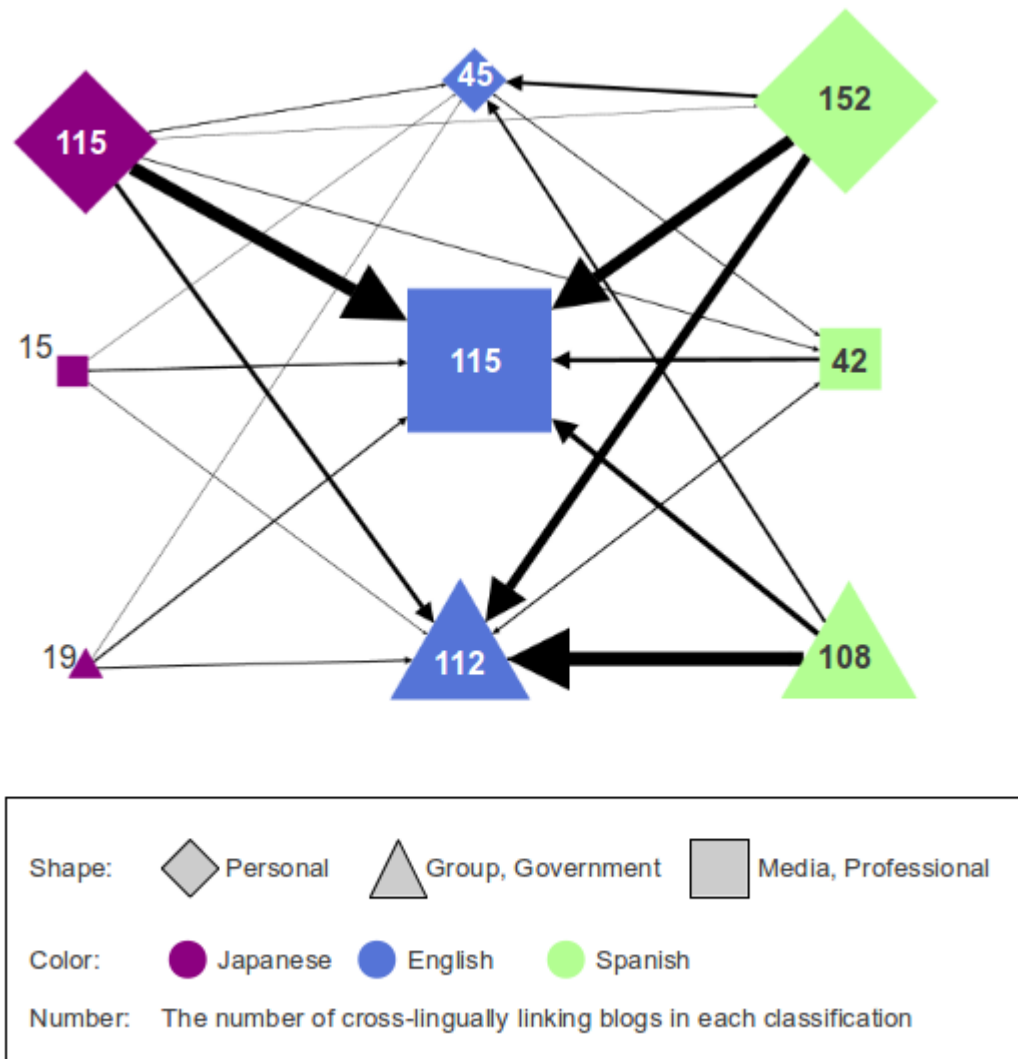


Figure 3: *Cross-lingual hyperlink network diagram*

*Note:* This network diagram shows all cross-lingual links by creator/destination type and language. Shape identifies the type of entity (personal, group, or media/professional) creating or receiving the link, and the arrow width is proportional to the number of links sent/received. Node size is proportional to the number of cross-lingually linking blogs in each classification, and this number is written on each node. The nodes are hand positioned as same-language links—which have been omitted—hold nodes of the same language tightly together. English receives the majority of cross-lingual links, while few links directly connect Japanese and Spanish.