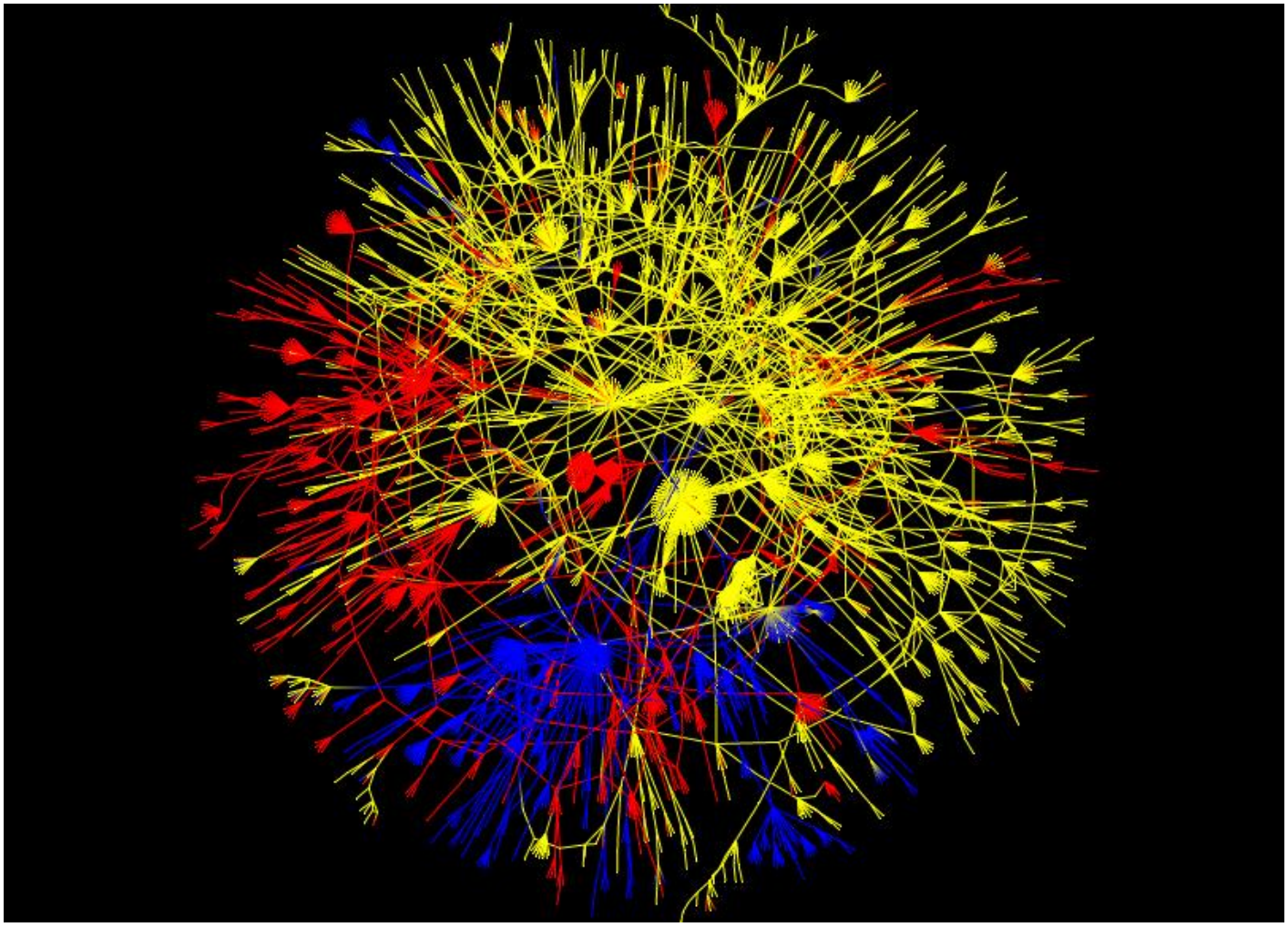
Online language bubbles: the last frontier?

Scott A Hale explores the effect of language in seeking and imparting information on the broader web.



The first draft principle of Free Speech Debate addresses the right to "seek, receive and impart information and ideas, regardless of frontiers". One of the most obvious, but least studied frontiers online is language. FSD recognises this and has committed to an impressive program of translating content into thirteen languages.

What, however, is the effect of language in seeking and imparting information in the broader web? Existing research does not fundamentally address this question, and it is not a question that can be addressed fully within this post. Yet, search engines provide one window into the differences in content between languages. When searching for images, search engines try to match query words to the text that appears near images in webpages as well as to the filenames of the images and to the words in links to the images. On the one hand, we might expect image results to be broadly

similar across languages as images can often be understood independent of descriptive text. Yet images are uploaded and annotated within specific cultural-linguistic settings. While Google is not the dominant search provider in all markets (Yahoo! Japan, Yandex and Baidu have more market share in Japan, Russia, and China respectively), it is still is the global search leader, indexes huge amounts of content, and presumably applies similar, if not the same, algorithms to searching content in different languages. This makes the differences in results between searches for different places very striking.

Figures 1 and 2 show results for Google Image searches for Tiananmen Square in English and Chinese. All search queries were conducted moments apart from the same computer in the UK using google.com; yet, the results are surprisingly different in sometimes innocuous and sometimes concerning ways. The results for Tiananmen Square, for instance, reveal a disparity in the number of images indexed in English versus in Chinese about the protests of 1989.

Figure 1: Google Image search results for Tiananmen Square in English

Figure 2: Google Image search results for Tiananmen Square in Chinese

More systematic study of the differences between language editions of the online encyclopedia Wikipedia also show "a surprisingly small amount of content overlap exists between language editions of Wikipedia" (Hecht & Gergle, 2010). Particularly interesting is that even the English edition, by far the largest edition of the encyclopedia, contains on average only 60% of the concepts discussed in any other Wikipedia edition in the study. (The highest overlap is between English and Hebrew at 75%.) In fact, English contains only about half of the concepts found in the second-largest edition, German, while German contains about 16% of the articles in English. Of course, even where "two language editions cover the same concept (with perfect clarity), they may

describe that concept differently", a point Hecht & Gergle (2010) analyse further in their paper and also address with new tools such as Omnipedia, which allows users to explore the differences between language editions.

This is not all doom and gloom, however. Several platforms have achieved truly global penetration (Facebook, Twitter, YouTube, and Wikipedia to name a few), and while communications on these platforms are primarily within language, these platforms also open the possibility of spreading information across frontiers at unprecedented speeds. My research on blogs discussing the 2010 Haitian earthquake (Figure 3) and on link sharing in Wikipedia and Twitter after the 2011 Japanese tsunami and earthquake as well as research by Irene Eleta on Twitter show instances where information is disseminated across languages and where multilingual users span multiple language groups and act as "bridge nodes" allowing the flow of information between language groups.

Enabling the flow of information between languages has both technical and social dimensions. Machine translation is not without errors, but it still enables engagement with other-language content at a high level and further research and new sources of training data will help to improve machine translation systems continually. In addition, research into the role design plays in helping users discover and diffuse information across languages, including my own research, is needed, and social media companies should use findings from this research and the broader social sciences when building new platforms. Finally, exciting new tools will leverage the skills of both computers and users to cross language frontiers. Duolingo and Monotrans2 are two examples that enable monolingual users to translate content, and, in the case of Doulingo, to also learn a new language at the same time. There will also always be a place for human translation and a role media organisations can play in identifying and verifying information about important events in other languages.

Scott A Hale is a research assistant and doctoral candidate *at the Oxford Internet Institute.*

Published on:July 25, 2012