Okinawa in Japanese and English Wikipedia

Scott A. Hale
Oxford Internet Institute
University of Oxford
scott.hale@oii.ox.ac.uk

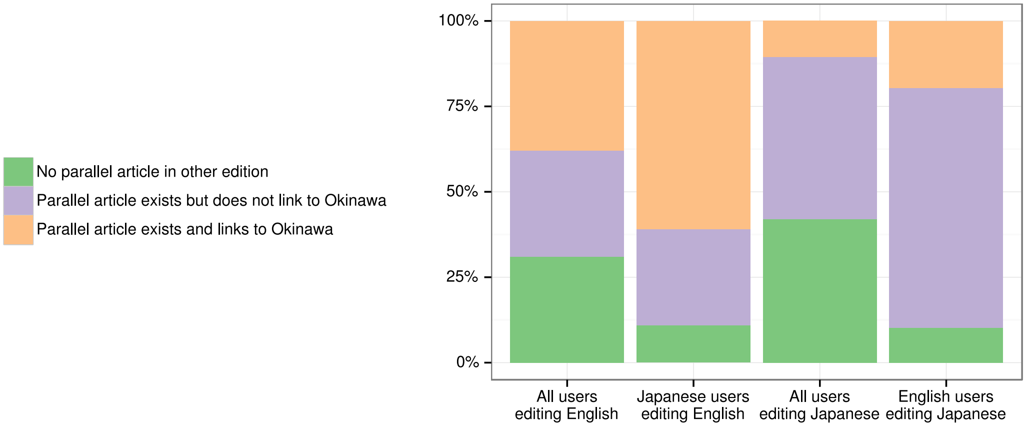

Figure 1: Users editing in a non-primary language are more likely than other users to edit articles that have interlanguage links to their primary language.

Abstract

This research analyzes edits by foreign-language users in Wikipedia articles about Okinawa, Japan, in the Japanese and English editions of the encyclopedia. Okinawa, home to both English and Japanese speaking users, provides a good case to look at content differences and cross-language editing in a small geographic area on Wikipedia. Consistent with prior work, this research finds large differences in the representations of Okinawa in the content of the two editions. The number of users crossing the language boundary to edit both editions is also extremely small. When users do edit in a non-primary language, they most frequently edit articles that have cross-language (interwiki) links, articles that are edited more by other users, and articles that have more images. Finally, the possible value of edits from foreign-language users and design possibilities to motivate wider contributions from foreign-language users are discussed.

Author Keywords

Social Media; Information Discovery; Information Diffusion; Cross-language; Wikipedia; Multilingual

ACM Classification Keywords

H.5.4 [Information Interfaces and Presentation (e.g. HCI)]: Hypertext/Hypermedia; H.5.3 [Information

Interfaces and Presentation (e.g. HCI)]: Group and Organization Interfaces

General Terms

Human Factors; Design

Introduction

Each language edition of Wikipedia captures unique information, much of which is not available in other languages [4]. Multilingual users who edit multiple language editions may play an important role in sharing new information between languages and thereby help to address some of the self-focus biases known to exist in the encyclopedia [2, 3]. This work-in-progress analyzes the full edit histories of nearly 20,000 Wikipedia articles related to Okinawa, Japan in the English and Japanese editions of the encyclopedia. It identifies the articles appearing in only one of the two language editions (e.g. the crash of a US helicopter into Okinawa International University in 2004 is covered only in Japanese) and examines what types of articles users edit when editing in a non-primary language (e.g. edits to the English edition by users who primarily edit the Japanese edition or vice versa). As an indicator of the similarity of articles in different editions, the research examines the percentage of images as well as hyperlinks to external sources that are found in common between the two editions. The amount of overlap in turn is compared to the number of users editing both editions.

Background and related work

In general, English and Japanese are among the most-used languages online, but speakers in each language play vastly different roles in interlanguage connections [1, 2]. A one-month study of edits to the top 46 language editions of Wikipedia found that approximately 15% of active Wikipedia users edited multiple language editions of the encyclopedia [2]. 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. Japanese was a major outlier in this respect, however, with only 6% of the primary editors of the Japanese edition editing a second edition. When non-English users did edit a second edition, that edition was most frequently English [2].

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 in a relatively small geographic area.[1] 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 again separated from Japan at the end of World War II and administered by the United States until 1972. 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 [5] with base facilities occupying over 18% of the land area of the largest island [6].

Data

Using the Wikimedia Labs[2] infrastructure, a list of all articles linking to an article starting with "Okinawa" for the English edition or to an article starting with "沖縄" (Okinawa) for the Japanese edition was created. All edits to each article were then downloaded from the date the

[1] The author has lived in Okinawa for considerable time and can speak Japanese and English.

[2] https://www.mediawiki.org/wiki/Wikimedia_Labs

English title	Japanese title
Japan	日本
Taiwan	中華民国
Kana	仮名 (文字)
Guam	グアム
Saipan	サイパン島
Kyushu	九州
Karate	空手道
Tofu	豆腐
Tinian	テニアン島
Burakumin	部落問題

(a) Ordered by PageRank scores in the English network.

English title	Japanese title
Okinawa Prefecture	沖縄県
Japan	日本
1972	1972年
April 1	4月1日
Kagoshima Prefecture	鹿児島県
1945	1945年
Shōwa period	昭和
Kyushu	九州
Naha, Okinawa	那覇市
NHK	日本放送協会

(b) Ordered by PageRank scores in the Japanese network.

Table 1: Top articles appearing in both editions and referencing Okinawa.

article was created until October 2013 using the Wikipedia API.[3]

The list of articles was filtered to only include articles in the main, article namespace (i.e. not talk pages, user pages, etc.). The articles were also filtered to only include those 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).

Corresponding articles in the two editions were found using the October 2013 database dump from WikiData.[4] WikiData is a new initiative to centralize all interwiki references and category information (and, in the future, statistics and other structured data) in one location. It avoids some previous issues with out-of-date or conflicting interlanguage links. For each user, the Central Authorization database was queried with the username to determine if the username was a global account. If it was, the database for each language edition was queried to get the total number of edits per language that the user made since creating the account.[5]

Results

Article overlap

The data collection process found 14,825 articles edited by 510,488 unique users in the Japanese edition linking to an article beginning with Okinawa. In contrast, 5,441 articles edited by 346,544 unique users were found in the English edition. Consistent with the global findings of low information overlap between language editions [4], there was only moderate overlap between the articles related to Okinawa in the two editions. As Figure 2 shows, 37% of the English articles and 33% of the Japanese articles appeared in both the English and Japanese editions. Of these articles, however, only a smaller subset linked to Okinawa in both languages (1,304 articles or 24% of English articles and 9% of Japanese articles linking to Okinawa). This indicates that even when articles exist in both languages, often only the article in one of the two languages is connected to Okinawa.

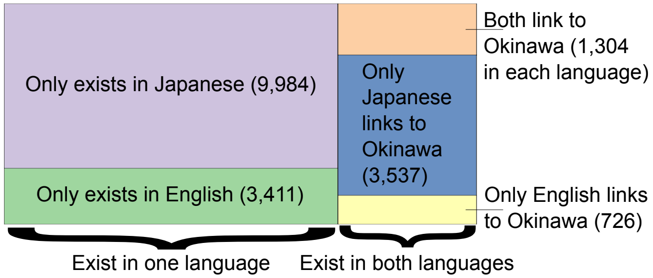

Figure 2: Article overlap. The area of each rectangle is proportional to the number of articles in that classification. Both the English and Japanese editions include many articles not found in the other edition referencing Okinawa, Japan.

In addition to interlanguage links across language editions, each Wikipedia article often includes links to other articles within the same language edition in the body text of the article. Using these links, two networks were constructed. One for Japanese with each Japanese article in the sample as a node, and one for English with each English article as a node. Edges in both networks were the links between the articles in the same language edition. Nodes were then ranked using the PageRank method [7]. This method, 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.

[3] `https://www.mediawiki.org/wiki/API:Main_page`

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

[5] The code used is available at `http://www.scotthale.net/pubs/?chi2014src`

The top-ranked nodes (articles) appearing in both language editions and linking to Okinawa are shown in Table 1a ordered by their ranks in the English network and in Table 1b ordered by their ranks in the Japanese network. The top-ranked articles in the English edition contain many surrounding geographical locations: Taiwan, Guam, Saipan, and Tinian. Japan and Kyushu (the southern most island of mainland Japan) feature in both lists, while the Japanese list also includes Kagoshima Prefecture, the prefecture neighboring Okinawa to the north. 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), Shōwa period (1926 to 1989 in the Japanese calendar). It also includes Naha, the capital of Okinawa, and NHK, the Japanese public broadcaster.

Top-ranked articles only appearing in one language edition are shown in Table 2 for the English edition and Table 3 for the Japanese edition. The English edition includes many articles on specific karate styles and kata (forms/patterns) that have no parallel article in Japanese. Okinawa is the birthplace of karate, and karate spread widely after gaining popularity among the American military members stationed in Okinawa after World War II. The presence of detailed articles on kata and styles in English, but not in Japanese, reflects this history.

Among the articles found only in Japanese, the Japanese edition includes historical references to "Okinawa Reversion" (Okinawa returning to Japan in 1972 after the US Administration) and "Okinawa under US Administration" (an overview article about the period not paralleled in English although both editions have articles about more specific aspects of the US Administration of Okinawa). The Japanese edition also includes many companies based partly or entirely in Okinawa that do have an article in English.

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 2: Top articles only in English ranked by the PageRank method on the page–page network.

User overlap

Most language editions of Wikipedia are known to suffer from self-focus bias where articles about places, people, and events where the language of the edition is spoken are more prominent than those in other regions [3]. Edits from non-primary language users have been suggested to be important in expanding coverage and addressing this self-focus bias [2]. Only a small percentage of users in each edition edited a different edition (Table 4). 660 primary editors of the Japanese edition edited articles in the English edition within the sample, and 1,014 primary editors of the English edition edited articles in the Japanese edition. When editing in their second language, users primarily edited articles with corresponding articles in their primary language (see Figure 1). The articles they edited tended to have more edits/editors, have higher

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 3: Top articles only in Japanese ranked by the PageRank method on the page–page network.

Japanese		
Total users	510,488	100.00 %
Anonymous	448,295	87.82 %
Local accounts	11,352	2.22 %
Primarily English	747	0.15 %
Primarily Japanese	48,581	9.52 %
Primarily Other	1,513	0.30 %

English		
Total users	346,544	100.00 %
Anonymous	255,969	73.86 %
Local accounts	18,084	5.22 %
Primarily English	65,783	18.98 %
Primarily Japanese	629	0.18 %
Primarily Other	6,079	1.75 %

Table 4: User distribution. The primary language is the language of the most-edited edition of Wikipedia.

PageRank scores computed as described earlier, and have more images. 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 (see Table 5).

Articles existing in both editions and referencing Okinawa shared 20-25% of the same links and images across the two editions. The percentage of edits by non-primary language users was only weakly correlated with the overlap in links (0.10 for English and 0.04 for Japanese) or images (0.24 for English and 0.18 for Japanese).

Discussion

Large differences remain between language editions, and contributions by non-primary language users are an extremely small proportion of all edits in both editions. Design changes might increase the number of non-primary language editors and/or increase the breadth of articles they edit. Currently, for example, it is only possible to search one language edition at a time, which makes it difficult to know what topics are not covered in a given language edition. When users in this study did edit in a non-primary language, they often edited only articles with interlanguage links. Therefore, approaches similar to the methods used here (gathering all articles linking to a given article and computing the PageRank scores) could be used to suggest central topics related to a specific theme in one language edition that have no equivalent article in another language edition. This could increase the breadth of articles users edit in their non-primary languages. Similarly, the search interface could check other language editions if no matches are found in the language edition being searched.

Further work is needed to elucidate the roles users editing in non-primary languages are playing for different topics/regions and in different language editions. This will be accomplished through a mix of further quantitative, qualitative, and experimental methods.

	# 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 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: 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.

References

[1] Hale, S. A. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the 2014 ACM Annual Conference on Human Factors in Computing Systems*, ACM (Montreal, Canada, 2014).

[2] Hale, S. A. Multilinguals and Wikipedia editing. http://arxiv.org/abs/1312.0976, 2014.

[3] 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.

[4] 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.

[5] Ministry of Foreign Affairs of Japan. 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, 2008.

[6] Okinawa Prefecture. 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, 2013.

[7] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, Nov. 1999.