Cross-language activity in Okinawa on Wikipedia

Scott A. Hale
scott.hale@oii.ox.ac.uk

Oxford Internet Institute
University of Oxford

10 December 2013
1 Introduction
2 Wikipedia at a global level
   - Data
   - Monolingual vs multilingual users
   - Language and structure
   - Bridging
   - Language size and introversion
   - Language connections
3 Okinawan content in Wikipedia
   - Data
   - Article overlap
   - User overlap
   - Content overlap
   - Next steps
Motivations: Practical

- Number of non-English speaking Internet users rising
  - (English speakers in relative decline)
  - New fiber-optic cables, mobile Internet accelerating this
Motivations: Practical

- Number of non-English speaking Internet users rising
  - (English speakers in relative decline)
  - New fiber-optic cables, mobile Internet accelerating this

- User-generated platforms internationalizing
  - Wikipedia
  - Twitter
  - Facebook
Motivations: Practical

- Number of non-English speaking Internet users rising
  - (English speakers in relative decline)
  - New fiber-optic cables, mobile Internet accelerating this

- User-generated platforms internationalizing
  - Wikipedia
  - Twitter
  - Facebook

- Monolingual platforms debating multilingual strategy
  - Quora
  - StackExchange (Stack Overflow)
Past work

Language clustering vs. small-worlds

- Users thought to cluster by language in most online platforms (Barnett & Choi, 1995; Hale, 2012a, 2012b; Herring et al., 2007; Nordenstreng & Varis, 1974; Takhteyev, Gruzd, & Wellman, 2011; Wilkinson & Thelwall, 2012)

- Large diversity in information between languages (Hecht & Gergle, 2009, 2010)

- Many online platforms thought to exhibit the ‘small-world’ phenomenon of small path lengths between users (despite high clustering)
Past work

Language clustering vs. small-worlds

- Users thought to cluster by language in most online platforms (Barnett & Choi, 1995; Hale, 2012a, 2012b; Herring et al., 2007; Nordenstreng & Varis, 1974; Takhteyev et al., 2011; Wilkinson & Thelwall, 2012)
- Large diversity in information between languages (Hecht & Gergle, 2009, 2010)
- Many online platforms thought to exhibit the ‘small-world’ phenomenon of small path lengths between users (despite high clustering)

Role of multilingual users

⇒ If users cluster by language and platforms are small-worlds, there must be brokers bridging different language groups (spanning structural holes)
- Multilingual users are possible brokers. Only one study investigating this—Ego-net level study on Twitter following–follower network structure (Eleta & Golbeck, 2012).
- No study multiplatform study, no study at large-scale level
Brokerage and small-world networks have benefits at network and individual levels

- consequences to bridging clusters within a network (Burt, 2004)
- “strength of weak ties” (Granovetter, 1973)
- performance related impacts (Page, 2007; Johnson, 2010; Uzzi, Amaral, & Reed-Tsochas, 2007)
Wikipedia at a global level

- Data and user counts by editions
- Multilinguals edit significantly more than monolinguals
- Percentage of multilinguals varies with edition size
- English very central in language–language network

Edits from top 46 language editions
8 July to 9 August 2013
3.5 million non-minor edits by 55,568 registered users

<table>
<thead>
<tr>
<th>Language</th>
<th>User Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>22,412</td>
</tr>
<tr>
<td>German</td>
<td>4,920</td>
</tr>
<tr>
<td>French</td>
<td>3,430</td>
</tr>
<tr>
<td>Russian</td>
<td>3,330</td>
</tr>
<tr>
<td>Spanish</td>
<td>3,299</td>
</tr>
<tr>
<td><strong>Japanese</strong></td>
<td><strong>3,164</strong></td>
</tr>
<tr>
<td>Italian</td>
<td>2,202</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,975</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1,220</td>
</tr>
<tr>
<td>Polish</td>
<td>1,011</td>
</tr>
<tr>
<td>Dutch</td>
<td>1,007</td>
</tr>
</tbody>
</table>
On Wikipedia, 15.4% of users (8,544) edited more than one language edition and were designated as multilingual users.

Density plot compares the number of edits made by monolingual and multilingual Wikipedia users. Size of edits does not differ significantly.

Only 2.6% of edits are from users writing in their non-primary languages on Wikipedia.
Number of users in each language compared to the percentage of these users classified as multilingual.
Co-editing patterns. Nodes represent language editions of the encyclopedia and the directed, weighted edges show the log of the number of users primarily editing one language edition who edited another edition as well. Only edges with weights over 1.96 standard deviations above the mean are shown.
Why?

- Japanese and English large languages online
- Okinawa is a small physical location with native speakers of both languages
- Levels of segregation in physical world

What?

- Wikipedia (plans to add Twitter)
- Representation of Okinawa on Wikipedia (what articles exist in one/both language editions)
- Users editing both ja and en editions? To what effect?
- Wikilabs to query for all articles with a link to “Okinawa”
- API to pull all edits to each article from creation to Oct. 2013.
- Only articles in main namespace (not talk pages, user pages, etc.)
- Filter articles that don’t mention Okinawa in the main body text of the article
- Articles from the two editions connected with database dump from WikiData

**Resulting data**
- 22,810 articles; 510,490 users in the Japanese edition
- 7,180 articles; 346,545 users in the English edition
Article overlap

Only exists in English (3,412)
Both link to Okinawa (1,303)

Only exists in Japanese (9,984)
Only English links to Okinawa (726)
Only Japanese links to Okinawa (3,538)

Only in one language (13,396)
Both languages (5,567)
<table>
<thead>
<tr>
<th>Article title</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>沖縄返還</td>
<td>Okinawa Reversion</td>
</tr>
<tr>
<td>琉球放送</td>
<td>Ryukyu Broadcasting Corporation</td>
</tr>
<tr>
<td>沖縄セルラー電話</td>
<td>Okinawa Cellular</td>
</tr>
<tr>
<td>日本プロサッカーリーグ</td>
<td>Japan Professional Football League</td>
</tr>
<tr>
<td>MBSテレビ</td>
<td>MBS (Mainichi Broadcasting System) TV</td>
</tr>
<tr>
<td>西日本</td>
<td>Japan West</td>
</tr>
<tr>
<td>落語家</td>
<td>Rakugo Story Teller (Comic story teller)</td>
</tr>
<tr>
<td>アメリカ合衆国による沖縄統治</td>
<td>Okinawa under US Administration</td>
</tr>
<tr>
<td>南日本放送</td>
<td>Minaminihon Broadcasting Co</td>
</tr>
<tr>
<td>鹿児島テレビ放送</td>
<td>Kagoshima Television Station</td>
</tr>
<tr>
<td>Article title</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Komainu</td>
<td>Broader category for shisa</td>
</tr>
<tr>
<td>Yukatchu</td>
<td>Ryūkyū Kingdom aristocracy</td>
</tr>
<tr>
<td>Karahafu</td>
<td>Japanese architectural style</td>
</tr>
<tr>
<td>Bunkai</td>
<td>Karate, Kata</td>
</tr>
<tr>
<td>Hagushi</td>
<td>Place in Yomitan</td>
</tr>
<tr>
<td>Isshin-ryū</td>
<td>Karate, Style</td>
</tr>
<tr>
<td>Matsubayashi-ryū</td>
<td>Karate, Style</td>
</tr>
<tr>
<td>Shōrinji-ryū</td>
<td>Karate, Style</td>
</tr>
<tr>
<td>Wanshū</td>
<td>Karate, Kata</td>
</tr>
<tr>
<td>Wankan</td>
<td>Karate, Kata</td>
</tr>
</tbody>
</table>
### Ranked by order in English edition

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

### Ranked by order in Japanese edition

<table>
<thead>
<tr>
<th>English title</th>
<th>Japanese title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okinawa Prefecture</td>
<td>沖縄県</td>
</tr>
<tr>
<td>Japan</td>
<td>日本</td>
</tr>
<tr>
<td>1972</td>
<td>1972年</td>
</tr>
<tr>
<td>April 1</td>
<td>4月1日</td>
</tr>
<tr>
<td>Kagoshima Prefecture</td>
<td>鹿児島県</td>
</tr>
<tr>
<td>1945</td>
<td>1945年</td>
</tr>
<tr>
<td>Shōwa period</td>
<td>昭和</td>
</tr>
<tr>
<td>Kyushu</td>
<td>九州</td>
</tr>
<tr>
<td>Naha, Okinawa</td>
<td>那覇市</td>
</tr>
<tr>
<td>NHK</td>
<td>日本放送協会</td>
</tr>
</tbody>
</table>
510,490
Japanese
users

346,545
English
users

5,322 users overlap (<0.5%)

Less than 0.5% of users edit both the Japanese and English editions.
Where do multilinguals edit?

![Bar chart showing the average percent of edits by multilinguals in English and Japanese editions. The chart includes two categories: "One language only" and "Both languages link to Okinawa." ]
The percentage of edits by multilinguals does not predict the proportion of links or images in common across editions.

Mean 20% to 25% of images or links in common

Image correlation with percentage of edits by multilinguals is low (0.11 and 0.18 for English and Japanese)

Link correlation with percentage of edits by multilinguals is low (0.09 and 0.45 for English and Japanese)
Summary

- Different representations of Okinawa in Japanese and English
  - English has more focus on karate related articles; “Okinawa in Asia”
  - Japanese has more focus on corporations and US–Okinawa; “Okinawa in Japan”
- Small number of users edit articles about Okinawa in both editions
- These multilingual users edit articles that exist in both languages more than articles in only one of the two languages
- On average articles that exist in both languages share 20% to 25% of the same images and links to external websites
- The percentage of edits by multilingual users doesn’t correlate strongly with the percentage of images or links in common across editions.
Next steps

- Survival time of edits to articles by multi/monolingual users
- Qualitative analysis of edits by multilinguals
- Typology of roles / impacts of multilingual users
Cross-language activity in Okinawa on Wikipedia

Scott A. Hale
scott.hale@oii.ox.ac.uk

Oxford Internet Institute
University of Oxford

10 December 2013


