Cross-language activity in Okinawa on Wikipedia

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

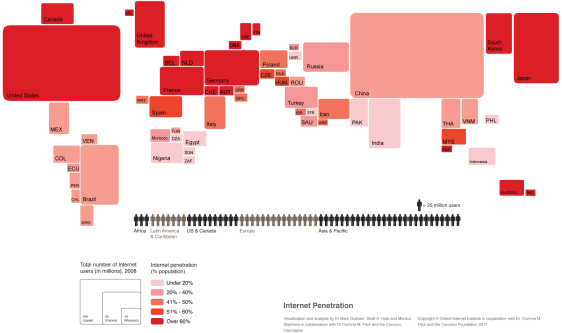
Oxford Internet Institute
University of Oxford

10 December 2013

Outline

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

$\Rightarrow$ 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

Past work

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

Hale, S.A. Multilinguals and Wikipedia Editing. `http://arxiv.org/abs/1312.0976`

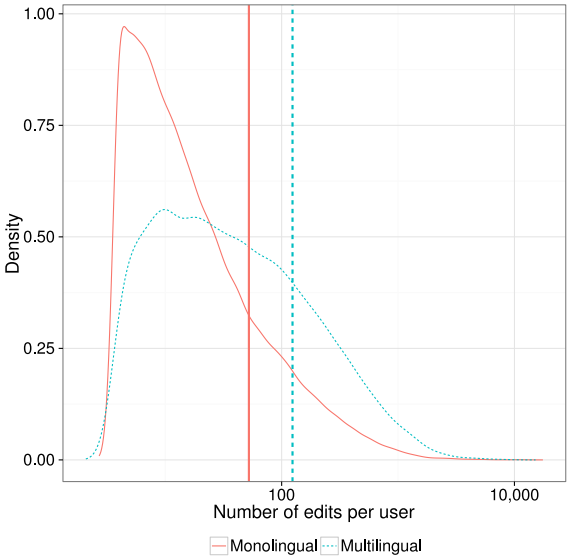
Data, Users by edition (global study)

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

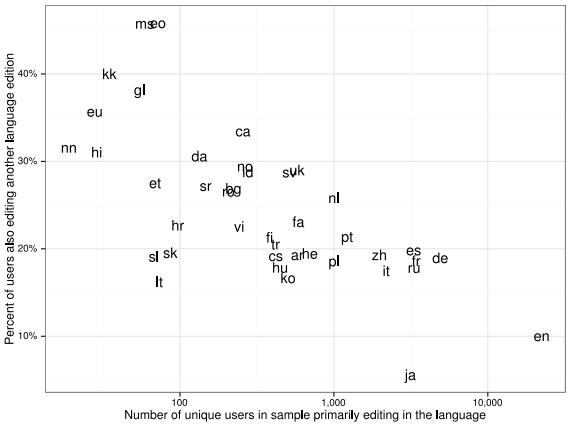
Language	User Count
English	**22,412**
German	4,920
French	3,430
Russian	3,330
Spanish	3,299
Japanese	**3,164**
Italian	2,202
Chinese	1,975
Portuguese	1,220
Polish	1,011
Dutch	1,007

Wikipedia: Multilinguals vs Monolinguals

- 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.

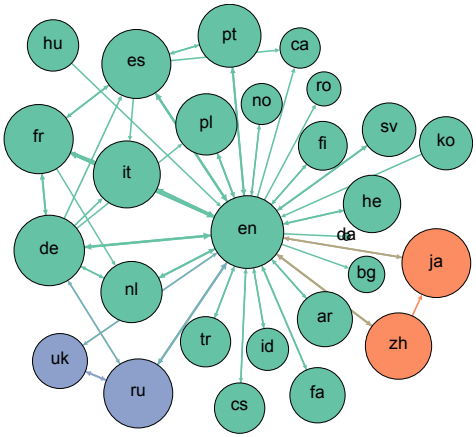


Multilingual user distribution

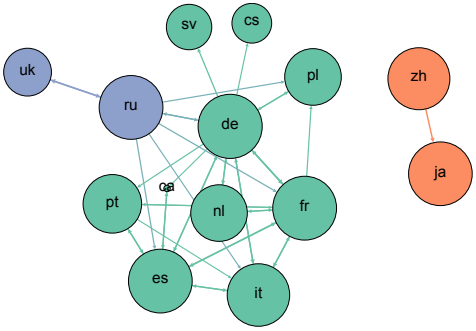


Number of users in each language compared to the percentage of these users classified as multilingual.

Wikipedia: Cross-language connections



(a) Network graph with English



(b) Network graph English removed

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.

Okinawa

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?

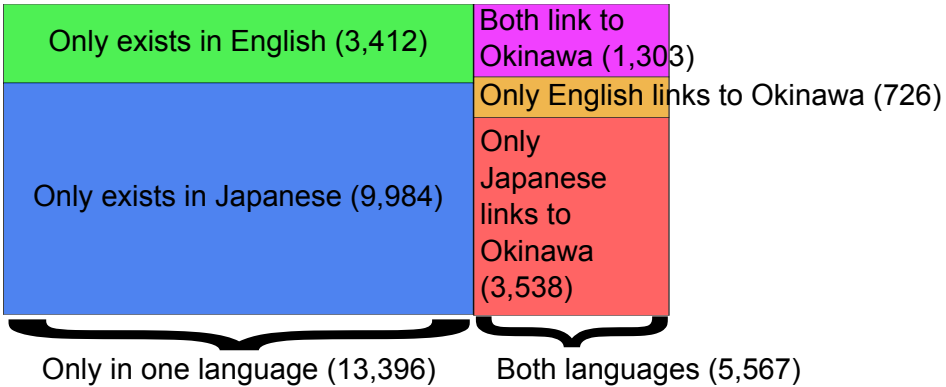
Data

- 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



Top Japanese only articles

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

Top English only articles

Article title	Description
Komainu	Broader category for shisa
Yukatchu	Ryūkyū Kingdom aristocracy
Karahafu	Japanese architectural style
Bunkai	Karate, Kata
Hagushi	Place in Yomitan
Isshin-ryū	Karate, Style
Matsubayashi-ryū	Karate, Style
Shōrinji-ryū	Karate, Style
Wanshū	Karate, Kata
Wankan	Karate, Kata

Top articles existing in both editions

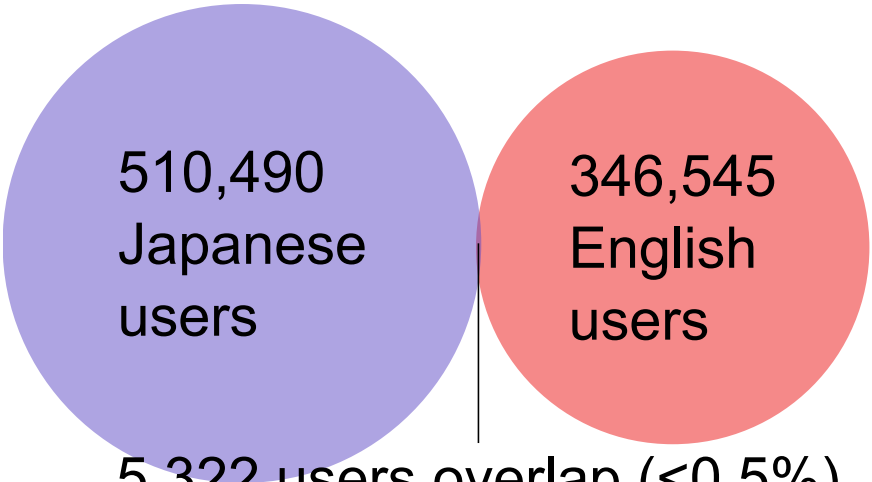
Ranked by order in English edition

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

Ranked by order in Japanese edition

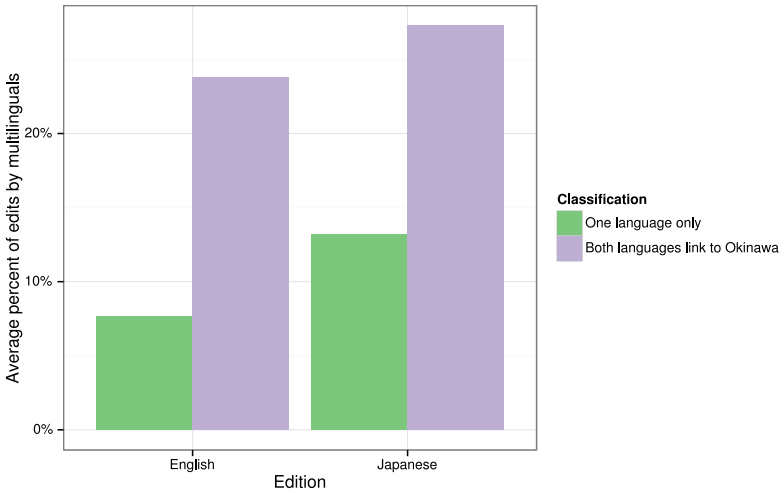
English title	Japanese title
Okinawa Prefecture	沖縄県
Japan	日本
1972	1972年
April 1	4月1日
Kagoshima Prefecture	鹿児島県
1945	1945年
Shōwa period	昭和
Kyushu	九州
Naha, Okinawa	那覇市
NHK	日本放送協会

User overlap



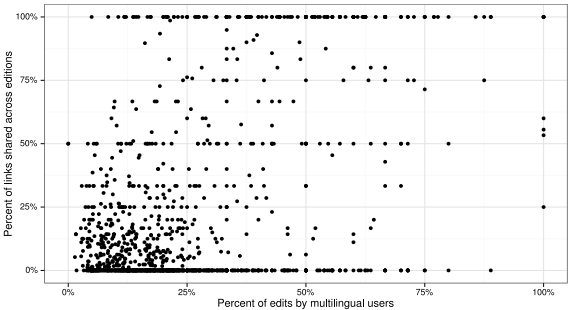
Less than 0.5% of users edit both the Japanese and English editions.

Where do multilinguals edit?



Links and images in common across editions

- 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

Barnett, G. A., & Choi, Y. (1995). Physical Distance and Language as Determinants of the International Telecommunications Network. *International Political Science Review*, *16*(3), 249–265. Available from http://ips.sagepub.com/content/16/3/249.abstract

Burt, R. S. (2004). Structural Holes and Good Ideas. *The American Journal of Sociology*, *110*(2), 349–399. Available from http://www.jstor.org/stable/3568221

Eleta, I., & Golbeck, J. (2012). Bridging Languages in Social Networks: How Multilingual Users of Twitter Connect Language Communities. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. Available from http://dx.doi.org/10.1002/meet.14504901327

Granovetter, M. (1973). The Strength of Weak Ties. *The American Journal of Sociology*, *78*(6), 1360–1380. Available from http://www.jstor.org/stable/2776392

Hale, S. A. (2012a). Impact of platform design on cross-language information exchange. In *Proceedings of the 2012 acm annual conference on human factors in computing systems extended abstracts* (pp. 1363–1368). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2212776.2212456

Hale, S. A. (2012b). Net Increase? Cross-Lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*, *17*(2), 135–151. Available from `http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2011.01568.x/full`

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies* (pp. 11–20). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1556460.1556463`

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, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1753326.1753370`

Herring, S. C., Paolillo, J. C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., et al. (2007). Language Networks on LiveJournal. In *Proceedings of the 40th annual hawaii international conference on system sciences.* Washington, DC, USA: IEEE Computer Society. Available from `http://dx.doi.org/10.1109/HICSS.2007.320`

Johnson, S. (2010). *Where good ideas come from: The natural history of innovation*. New York: Riverhead.

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).

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2011). Geography of Twitter networks. *Social Networks*, 1–26. Available from `http://www.sciencedirect.com/science/article/pii/S0378873311000359#FCANote`

Uzzi, B., Amaral, L. A. N., & Reed-Tsochas, F. (2007). Small-world networks and management science research: a review. *European Management Review*, *4*(2), 77–91. Available from `http://dx.doi.org/10.1057/palgrave.emr.1500078`

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology*, *63*(8), 1631–1646. Available from `http://dx.doi.org/10.1002/asi.22713`