How Wiki Can Exploit Linguistic Diversity

Han-Teng Liao
Oxford Internet Institute
University of Oxford
Oxford, OX1 3JS
+44 (0)1865 287210
hanteng@gmail.com

ABSTRACT
Linguistic diversity can be a driver rather than an obstacle for collaboration over Wiki, as shown by the success of the cross-regional conversion mechanism developed in the Mandarin Chinese-language version of Wikipedia, even when the language politics of Mandarin Chinese remains one of the thorny issues among the Chinese-speaking population. This position paper discusses the ways wiki could be used in exploiting linguistic diversity for human-based computation and language learning. It then speculates how wiki may change the print-based language politics.

Categories and Subject Descriptors
I.2.6 [**Learning**]: Knowledge acquisition. H5.3 [**HCI**]: Web-based interaction.

General Terms
Design, Standardization, Languages, Theory

Keywords
Language learning, Linguistic diversity, Collaboration, Human-based computation.

1. Introduction
The idea of co-writing knowledge across linguistic (and usually thus political) communities is not new. Saint-Simon proposed in 1810 that the "common interest of France and England requires the Royal Society of London and the Imperial Institute of France to work together on a new encyclopedia" [1]. His proposal never materialized in the print era. What is materialized now, after almost two centuries, is a global online encyclopedia called Wikipedia, co-authored by individual volunteers on the Internet. The paper argues that linguistic diversity may offer rather than undermine incentives in creating user-generated content, if an appropriate socio-technical arrangement is managed as in the Mandarin Chinese version of Wikipedia.

2. Forming the De-Facto Digital Mandarin
Different from the mutually unintelligible spoken Sinitic language groups (e.g. Cantonese, Min, Wu), the linguistic diversity within written Mandarin Chinese is largely a product of modern development. The diversity within Mandarin Chinese does not hinder mutual intelligibility, similar to the national varieties of English. However, the Second World War and the Cold War have arguably created more divisions among Mandarin-speaking regions than in English-speaking ones. In terms of orthography, Mainland China and Singapore has adopted simplified Chinese characters whereas Taiwan, Hong Kong and Macau continue to use orthodox/traditional Chinese characters. Also, because of the different colonial experience and Cold War positions, standards of Mandarin are more diverse and politically-charged. It can be argued that now it is the Internet that makes them meet again.

It is why Chinese Wikipedia aims to accommodate such diversity by acknowledging four variants (1) mainland China, (2) Singapore, (3) Hong Kong and Macau, and (4) Taiwan. Policies such as "Avoid Region Centric" are thus established and enforced to respect differences. Language-wise, to facilitate the content presentation/storage processes, several tables of orthographic and lexical mappings are constructed and maintained by the contribution from users from different regions. Technology-wise, it aims to respect readers' different preferences for regional orthographic and lexical choice when presenting content, while preserving the underlying difference of each contribution when storing contributed content. Hence, I have argued elsewhere that Chinese Wikipedia is by far the most advanced Chinese-written website in tackling Chinese conversions in levels of orthography (character choice), vocabulary (word choice) and semantics (word meanings) with a user-generated conversion mechanism, thanks to its arguably the most authoritative mapping of terminology and orthography [3]. The mechanism facilitates the process of accommodating differences and works in effect on a common de-facto digital Mandarin which is not monopolized by any one.

Such a mechanism demands a fresher perspective to see Wikipedia projects as mere text corpus and software codes. Indeed, Chinese Wikipedia builds both software codes and orthographic and lexical conversion tables. These user-generated corpus and codes should be seen together as an ongoing perfection between computation and content, especially when the computation process in turn helps the development of the very same content. It could be seen as an example of human-based computation, defined here as an ongoing computational process which improves and is improved by user-generated content. This type of human-based computation appears to be a case of the chicken-egg paradox as shown in Chinese Wikipedia. Without

user-contribution from different regions, the mechanism alone does not work. Without the mechanism that opens to further users' input, some users may feel excluded when encountering Chinese characters or terms that are unfamiliar to them. It seems difficult to start definitely with one or the other but once started it can keep growing.

Thus, it is necessary to solve the paradox before harnessing the enormous potential of such semi-structured user-generated content. The paradox could be partially explained if the political dimension is considered along the desire for self-expression and the respect for diversity as the source of volunteer labor. Arguably, it is precisely the political and linguistic tensions between the national/regional varieties of Mandarin Chinese that demands such a mechanism that respects diversity and the need for self-expression. In other words, a de-facto digital Mandarin Chinese linguistic space is formed by accommodating diversity rather than oppressing varieties. Sources of problems are thus turned into main drivers of user-generated solutions.

3. Redefining Cross-lingual Wiki

Although not an exact example of cross-lingual Wiki practices, Chinese Wikipedia's effort in accommodating linguistic diversity, using additional user-generated conversion tables and codes, provides valuable lessons. Within global de-facto English or "macrolanguages" defined in ISO 639 such as Arabic and Chinese, linguistic diversity can and should be harnessed for cross-regional, cross-national or even cross-lingual collaboration.

Indeed, Wiki practices have been friendly to linguistic diversity, as shown in the other Sinitic language versions of Wikipedia. They have different strategies in written forms: some only in simplified Chinese characters (Wu), some only in traditional Chinese characters (Cantonese and Gan), some in dual system of Latin-alphabets and traditional Chinese characters (Hakka and Min Dong), and some only Latin-alphabets (Min Nan). Without further discussion on their language politics, Wikipedia platform has provided at least separate but equal linguistic space for these dialects. It is a relatively new socio-political development in Chinese-speaking settings because mandarin Chinese has monopolized the written forms of Chinese language for centuries. Accordingly, though these versions might not be as big as the Mandarin Chinese version, the Cantonese and Wu versions have been growing partly due to the development of costal cities such as Hong Kong and Shanghai. How these language (or dialect) versions interact with the Mandarin Chinese version remains an open question. The cross-regional mechanism developed in the Mandarin Chinese version might not be suitable everywhere, but some human-based computation mechanism based on user-generated content and agreed by necessary participants is essential. It further raises the pressing question about the way the human-based computation mechanism is designed to elicit, include or exclude certain users' contribution. Such a mechanism needs further socio-political analysis, which requires nuanced contextual knowledge on specific cases.

A language version of Wikipedia should thus be regarded as a socio-political project with the linguistic community involved. It could be seen as cross-lingual inside if it is shared across several states or regions, such as Mandarin Chinese, German, and English, the desire for self-expression and thus respect for diversity must be taken seriously. Otherwise the energy and tension will result in edit wars [5]. For instance, which of these spellings is correct for the German name of "2006 FIFA World Cup": the German spelling of "Fußball" or Swiss German spelling of "Fussball"? Which spelling of "color" and "colour" should be preferred? If the desire for group self-expression is directed in such a way the linguistic diversity is respected or even valued, as shown in the case of Chinese Wikipedia, it can generate valuable resources for intra-lingual and cross-lingual computation and communication.

4. How Wiki May Change the Print-based Language Politics

The resources generated by Wiki for linguistic diversity may foster a new type of learning. For example, Chinese Wikipedia provides valuable resources in regionally-correct Chinese terms of "taxi", "sandwich", etc. used in regions such as Singapore, Taiwan, Hong Kong, and Beijing. Although it may not replace the existing standard language learning materials, they are often relevant and rich in cultural nuances, and thus provide valuable resources for active learning and contribution. Imagine English Wikipedia has similar variants of Indian English, Singlish, or even Euro-English. Such potentials can be enabled by Wiki and volunteers, which coincide with the new models of lingua franca for language planning, policy and education [2].

Such new type of learning may challenge the current language learning paradigm based on "native" standards (e.g. Queen's English, Beijing Mandarin, and Paris French). As a student of communications studies and nationalism studies, I believe Wikipedia projects provide alternatives to the print-based language-learning, and thus may initiate a socio-political re-configuration of standard language. Anderson explains the socio-political origin of print-based political communities by describing how the early European vernacular lexicographers, with the help from the print industries' need for new market, created diverse reading masses that were different to those of the previous pan-European Latin-reading elites [1]. Such a socio-political configuration has persisted in national dictionaries, grammar books, and pronunciation guidelines. The global Wikipedia project may reconfigure it if a meaningful socio-political project can be constructed to meet some of the existing demands in user-generated content.

5. ACKNOWLEDGMENTS

This work is financially supported by the scholarships offered by National Science Council, Taiwan (NSC-095-SAF-I-564-028-TMS) and the Oxford Internet Institute.

6. REFERENCES

[1] Anderson, B. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

[2] Ferguson, G. 2006. *Language Planning and Education*. Edinburgh: Edinburgh University Press.

[3] Liao, H. 2008. Conflictual Consensus in the Chinese Version of Wikipedia. In . Fredericton, New Brunswick, Canada: IEEE Society on Social Implications of Technology, June 26.

[4] Saint-Simon, H. 1975. *Henri Saint-Simon (1760-1825): Selected Writings on Science, Industry, and social organisation*. Ed. Keith Taylor. London: Croom Helm.

[5] Wikipedia English Version. 2008. *Wikipedia:Lamest edit wars*. http://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_war