Proceedings of the 29th North American Conference on Chinese Linguistics (NACCL-29). 2017. Volume 2. Edited by Lan Zhang. University of Memphis, Memphis, TN. Pages 463-480.

Language Policy, Dialect Writing and Linguistic Diversity[1]

Hongyuan Dong
George Washington University

This article studies the challenges encountered in the promotion of linguistic diversity in the context of Chinese dialects by examining the meta-data on Wikipedia sites written in major varieties of Chinese, with a focus on the type of writing systems used. The current language policy in China does not allow the explicit promotion of non-standard forms of Chinese in any official or national media. Therefore, online Wikipedia communities and sites of Chinese dialects have been flourishing. The choice of writing systems on these wiki sites to write Chinese dialects, including character-based and phonetic systems, is an important contributing factor to the success of these sites. I argue that the creation and practical use of an effective writing system conducive to literacy is a key issue in promoting dialects in the Chinese context.

1. Introduction

In this article, I study the effects of language policy and new collaborative technology on dialects from the perspective of the writing systems used by virtual linguistic communities. My focus here is on the different varieties of Chinese.[2]

In order to understand the current situation of linguistic diversity in terms of Chinese dialects and language policy making in China now, we need to take a historical perspective. The origins of modern language policy in China can be traced back to the year 1728 of the Qing Dynasty during the reign of Yongzheng Emperor, when an imperial edict was issued to order the establishments of local Mandarin schools in the Fujian and Guangdong areas (Dong 2014: 131; Wang 2014: 106). But this Mandarin Campaign was never met with any kind of enthusiasm from the local officials, and by 1775 during the reign of Qianlong Emperor the campaign was terminated (Deng 1994, Wu 2008, Dong 2015a). Consequently, the dialects in those areas were not affected at all.

Starting from the late 19th century until the founding of the People's Republic of China in 1949, another major wave of linguistic reform was implemented (Dong 2016,

[1] This paper benefitted from the discussions with the audience at NACCL-29, especially Miguel Cortiço dos Santos of The University of Tokyo.

[2] Here I will follow the traditional term "Chinese dialects" as a translation for "*Hànyǔ fāngyán*". Sometimes I refer to Chinese dialects as "varieties of Chinese". Many authors may prefer the term *topolects* or *Sinitic languages* (see e.g. Mair 1991).

Simons 2017). Although policies were made to promote Mandarin as the National Language, the implementations of these policies were not quite effective (Dong 2017). Thus, dialects were not affected much in this era either.

The new Chinese government after 1949 took a series of strong government measures to promote Putonghua as the national language (Zhou 2006, Zhou and Sun 2004). It is during this period up to the present time that usage of Chinese dialects has been gradually eroded. The situation resembles one of language loss. May (2006: 257–258) describes language decline and loss as occurring "most often in bilingual or multilingual contexts in which a majority language – that is, a language with greater political power, privilege, and social prestige – come to replace the range of functions of a minority language".

According to Baker and Jones (1998), and May (2006), there are three stages in the process of language shift. In terms of Chinese dialects, we may characterize these three stages as follows:

(1) Three Stages of Dialect Shift

- *Stage I*: increasing pressure on dialect speakers to speak the national language, particularly in formal language domains.
- *Stage II*: a decreasing number of fluent dialect speakers, especially among the younger generation.
- *Stage III*: replacement of dialects by the national language

Most varieties of Chinese, especially those in the south, are in the second stage of dialect shift as described above. This situation is directly related to the language laws in China. The most important one is the *Law of the People's Republic of China on the Standard Spoken and Written Chinese Language*, adopted at the 18th Meeting of Standing Committee of the Ninth National People's Congress on October 31, 2000. This law reflects various measures to promote Putonghua since 1949, and many of these measures are now officially codified to assume more power in its implementations. According to this law, "Putonghua and the standardized Chinese characters shall be used as the basic language in education and teaching in schools and other institutions of education, except where otherwise provided for in laws" (Article 10), "publications in Chinese shall be in conformity with the norms of the standard spoken and written Chinese language" (Article 11), and "Putonghua shall be used by the broadcasting and TV stations as the basic broadcasting language" (Article 12). Thus, dialects are restricted mostly to spoken forms in informal settings such as conversations at home.

Many scholars, dialect speakers, and dialect enthusiasts have started to try to preserve various dialects and, in some cases, oppose the promotion of Putonghua, e.g. resurgence of dialects in media (Liu 2013; Liu and Tao 2009, 2012), the campaign in Guangzhou to protect Cantonese from Putonghua erosion (Eng 2010), and etc. Much of such efforts to preserve dialects started in online communities, and the organizers made good use of social media. This leads to my interest in studying the use of new technology to promote linguistic diversity in the Chinese context.

In this article, I use the metadata on Wikipedia sites written in Chinese dialects to study the promotion of dialects on the Internet (see also Dong 2015b). This can be considered a kind of "virtual linguistic landscape" (Ivkovic and Lotherington 2009). Linguistic landscape studies language displayed in public space (Shohamy and Gorter 2008: 1). To some extent, the web is the global public space where multilingualism can be displayed at its best with minimal restrictions imposed by national language policies. This article studies the linguistic landscape on Wikipedia in the Chinese context.

The remaining part of this article is structured as follows. In section 2, I summarize the metadata from Wikipedia, and point out issues highlighted by the numbers. In section 3, I give examples of all the Wikipedia sites written in Chinese dialects to illustrate how these websites are promoting their own version of dialects. In section 4, I connect the issues in section 2 with the writing systems used to write these dialects, and show that writing Chinese dialects is a key component to promoting linguistic diversity. In section 5, I make further remarks in conclusion.

2. Metadata on Wikipedia

The reason for using Wikipedia as a tool for promoting linguistic diversity in the Chinese context can be phrased as follows.

First, although there is content containing Chinese dialect elements on websites in China, such websites are nonetheless regulated by China's language laws, such as shown in the Introduction section. For example, the Chinese website *Bǎidù Bǎikē* 百度百科, which is the Chinese equivalent of Wikipedia, only allows content in the standard form of Chinese. There are no dialect versions of *Bǎidù Bǎikē*. Therefore, to fully promote dialects on the Internet, tools from outside China will be more effective because they are less subject to the laws within China.[3]

Second, Wikipedia has become the go-to site for information on any kind of topic. It is always listed on top of google search results. Therefore, by using Wikipedia, it can be guaranteed that the information will reach the widest audience and be used by the most readers, for purposes of gaining information, or simply learning a new language.

Third, the global reach of the Internet can make collaboration more easily achievable. The community of content contributors on Wikipedia consists of people from

[3] This is not to say that websites operated outside China are totally free from the influence of language policy in China. In effect, China's language policy has global reach in the linguistic standardizations adopted by international organizations and more recently in the establishments of language institutes around the globe. But indeed these websites are less restricted by language laws in China. For example, the Mandarin Wikipedia pages are often written with a mixture of simplified and traditional characters, likely due to the geographical regions of contributors. Such mixed use of Chinese characters is definitely not allowed by the linguistic laws in China.

different areas of expertise, not just linguists. Therefore, to my knowledge there is no other online tool or community that can compare to Wikipedia in its size and its power to pool resources globally to create content in a dialect.

Another important aspect about Wikipedia is that the content, including multi-media content, such as recordings and videos, creates a library, or a body of literature, of some sort in a language or a dialect. The existence of written documentation and other types of texts is the basis for the preservation and promotion of a language or a dialect.

Additionally, the official use of dialects is limited in China, but to create content on Wikipedia gives users and readers the practical opportunity to use the dialect. As shown in (1), one of the stages of language shift is the decreased use of dialects, and in this sense, to actually use dialects to do something is an important step towards preserving such dialects in the sense of increasing the use of such dialects.

Therefore, Wikipedia serves as the best model, so far, for bringing people in an online linguistic community to create a presence, or rather the virtual linguistic landscape, in order to preserve and promote linguistic diversity. Thus, studying these Wikipedia sites can tell us a great deal about how such efforts are faring and what challenges they encounter, so that we may better understand the promotion of linguistic diversity in terms of Chinese dialects. On a related note, the multi-language list for the same topic on Wikipedia can help us compare different languages or dialects easily. This is another advantage of using such data to study Chinese dialects on the web systematically.

Before discussing the meta-wiki data, let me introduce the major varieties of Chinese. According to the traditional classification of Chinese dialects, e.g. Yuan et al. (1960), there are seven major dialects of Chinese: Mandarin, Wu, Xiang, Gan, Min, Hakka, and Cantonese[4]. But the internal differences in each of these groups are still quite considerable, especially in the Min dialect, within which mutual intelligibility is the lowest of these seven groups. According to the *Language Atlas of China* (Wurm et al. 1987), the Min dialect can be further distinguished among the following subgroups in (2).

(2) Subgroups of the Min dialect

- Northern Min or Min Bei (Nanping Prefecture)
- Shaojiang Min (Shaowu, Jiangle, etc.)
- Eastern Min or Min Dong (Fuzhou, etc.)
- Central Min (Sanming Prefecture)
- Pu-Xian Min (Putian and Xianyou)
- Southern Min or Min Nan (Xiamen, Taiwan, etc.)
- Leizhou Min (Leizhou City)
- Hainan Min (Wenchang)

[4] The more accurate term here is the Yue dialect, instead of Cantonese.

The subgroups in (2) are arranged roughly from north to south. The place names in the parentheses are the representative versions of each subgroup.

A more recently recognized new group is the Jin dialect[5] spoken in Shanxi and the surrounding areas such as Hebei, Inner Mongolia, Henan and Shaanxi. It was included in the Mandarin group in the traditional classification. But in many newer classification systems such as in the *Language Atlas of China* (Wurm et al. 1987), the Jin dialect is a separate primary group on par with Mandarin.

Table 1 shows the relative proportion of each dialect among speakers of the major varieties of Chinese.

TABLE 1. Size of Chinese Dialects[6]

Chinese varieties	**% of L1 Speakers**
Mandarin	66.2%
Jin	5.2%
Min (all subgroups)	6.2%
Wu	6.1%
Cantonese	4.9%
Gan	4.0%
Hakka	3.5%
Xiang	3.0%
Other	0.9%

The percentage is the proportion of first-language speakers. The largest group in Table 1 is Mandarin at 66.2%. If we combine Jin and Mandarin it is almost ¾ of all speakers (71.4%). The second largest group is Min (6.2%), as one group including all the varieties in (2). The Wu dialect has more or less the same number of speakers (6.1%) as the Min dialect. Cantonese (4.9%) follows Wu. Then the next groups are Gan (4.0%), Hakka (3.5%) and Xiang (3.0%). The "Other" category includes smaller dialects such as Pinghua and Huizhou. Since there are no Wikipedia sites written in Pinghua, Huizhou and other lesser-known dialects, I will not discuss these dialects in the "Other" category in this current article.

Now let's see the data regarding the Wikipedia sites written in Chinese dialects. In my research, data were collected over two years. I look at two snapshots of Chinese dialect Wikipedia sites. Table 2 shows the data recorded on March 9, 2015. Table 3 shows the data recorded on May 18, 2017.

[5] Jìn Yǔ 晋语.

[6] Source: http://en.wikipedia.org/wiki/Varieties_of_Chinese [Retrieved on November 20, 2017], where the data are taken from the 2nd edition of *Language Atlas of Chinese* (Chinese version), edited by the Chinese Academy of Social Sciences, published by the Commercial Press in 2012.

TABLE 2. Meta-wiki data of sites in Chinese dialects as of March 9, 2015

Rank	Dialect	Articles	Admins	Users	Active Users
15	Mandarin	814322	80	2007603	7949
79	Cantonese	35317	8	100829	167
119	Min Nan	12798	6	21324	38
143	Gan	6305	2	21862	24
161	Hakka	4512	0	13473	16
175	Wu	3536	3	31800	22
195	Min Dong	2518	1	8907	11

TABLE 3. Meta-wiki data of sites in Chinese dialects as of May 18, 2017

Rank	Dialect	Articles	Admins	Users	Active Users
15	Mandarin	941817	81	2375687	7363
39	Min Nan	208033	5	28898	66
76	Cantonese	53986	10	136487	239
147	Hakka	7423	0	18904	22
153	Min Dong	6432	3	11532	19
154	Gan	6388	2	26784	17
159	Wu	5812	3	49594	19

The data here were downloaded from the meta wiki webpage that can be easily retrieved from the follow address https://meta.wikimedia.org/wiki/List_of_Wikipedias. The different columns represent the overall ranking of the website among all Wikipedia websites in terms of total number of articles, the dialect used on the website, the total number of articles on that website, the total number of administrators in that specific wiki community, the total number of users, and the active users among them. According to the meta-wiki page, "Active Users" are defined as those that have registered and “have made at least one edit in the last thirty days” as of the date of the data collection. Thus “users” are those that have registered, being part of the relevant virtual linguistic community. The number of users is an indicator of the size of the virtual linguistic community, and the number of articles is an indicator of how well each site is doing generally.

Now let’s examine the numbers in Table 2 in detail first. The relative rankings of all Wikipedia websites of a variety of Chinese in terms of the total number of articles are Mandarin, Cantonese, Min Nan, Gan, Hakka, Wu and Min Dong. The Xiang, Min Bei and Pu-Xian versions of Wikipedia were being incubated at the time of data collection in

Table 2. Mandarin as the largest group of dialects (Table 1) has the largest Wikipedia site in terms of the number of articles, administrators, users and active users.[7]

Cantonese ranks second in both the number of users and the total number of articles, although in terms of speakers, Cantonese is behind Min and Wu. Some explanations for this relatively higher ranking of Cantonese can be found in the high internal homogeneity among all varieties of Cantonese, and the existence of a regional *lingua franca* based on the Guangzhou version of Cantonese. In this sense, the Cantonese linguistic community can pool the resources together more easily. Another reason might be due to the large number of overseas Cantonese speakers, e.g. in Europe and North America. In terms of Min, if we add the numbers of articles of Min Nan and Min Dong, their combined ranking is still third, right after Cantonese. Note that the size of Min in Table 1 is based on all varieties of Min. Thus the actual number of speakers of Min Nan an Min Dong should be much smaller, which can partially explain the ranking of Min Nan Wikipedia after Cantonese. The total number of users in the Min Nan and Min Dong virtual linguistic community ranks after Cantonese and Wu, but it is quite close to Wu.

The Gan and Hakka rankings on meta-wiki are more or less comparable to their real linguistic communities (Table 1). Xiang is the smallest among these major groups, and it is not surprising that its Wikipedia site was being incubated.

The only surprising fact from Table 2 is the low ranking of Wu in terms of total number of articles. But in terms of the total number of users, the virtual linguistic community of Wu ranks third, right after Cantonese. This is more in line with the size of the linguistic community in Table 1. This suggests that there are more people who are interested in the project of Wu Wikipedia than those who are actually contributing to the content creation.

To summarize the data in Table 2. The relative rankings of Wikipedia sites in major Chinese dialects are more or less comparable to their linguistic community sizes (Table 1). This shows that most of these linguistic communities are actively using Wikipedia as a way to promote their own dialects.

Now let’s compare the data from May 18, 2017 as shown in Table 3, with the data in Table 2 to see the growth of these Wikipedia sites. One trend is that most of these sites have higher rankings in Table 3 in terms of both the number of articles and number of users than their own rankings in Table 2, thus showing growth and maintenance of these sites over time. The Mandarin site has grown but maintains its ranking at 15. One

[7] As a comparison, English ranks No. 1 of all Wikipedia sites. As a global language, it is easy to see why English ranks No. 1 on Wikipedia. However, with the largest number of speakers, Mandarin’s ranking of No. 15 seems a little too low. There may be several reasons for this. For example, censorship within China intermittently blocks access to Wikipedia. Also there are Chinese equivalents of Wikipedia, such as *Bǎidù Bǎikē* 百度百科 and *Hùdòng Bǎikē* 互动百科, thus diluting the resources that users devote to one particular website. But since my focus is on Chinese dialects, instead of Mandarin in comparison to other major world languages, I will not go into any details here.

exception is the Gan Wikipedia, which dropped in its ranking from 143 to 154, although the number of articles and the number of users both increased. This shows a lack of momentum in the development of the Gan Wikipedia project. Those that were incubated in 2015 were still not up and running as of May 18, 2017, thus showing lack of growth.

The site that shows the most growth is Min Nan, which jumped from 119 in 2015 to 39 in 2017. Min Dong has also increased its ranking considerably as well. Although the Wu Wikipedia has also increased its ranking from 175 to 159, it is ranked last now among all these sites in terms of the total number of articles, although the number of users on the Wu Wikipedia is still third right after Mandarin and Cantonese. On the other hand, Cantonese has improved slightly in its ranking, and it seems that the Cantonese site is becoming quite stable and shows the highest number of administrators, users and active users after Mandarin.

To sum up the data in Table 3, we still see that the relative sizes of these Wikipedia sites are more or less proportional to those of their linguistic communities (Table 1), except in the case of Wu. Most of these sites have improved their overall rankings within the two years. Min Nan shows the largest growth, while Cantonese is stabilizing and becoming a more mature website.

By examining and comparing the data from Table 1, Table 2 and Table 3, we may give the following factors as contributing to the growth of a Wikipedia site written in a Chinese dialect.

First the internal homogeneity is a very important factor. Although officially speaking, Wu ranks higher than Cantonese in terms of the total number of speakers, the internal homogeneity of Cantonese is much higher than that of Wu. Some southern Wu dialects are actually not mutually intelligible with the northern Wu dialects. Even among the northern Wu dialects, Shanghainese as the prestigious variety can be understood by many speakers of Wu but they may not be able to contribute to creating content in Shanghainese.

The second major factor is the existence of overseas diaspora communities. In terms of both Cantonese and Min Nan, there are large linguistic communities in Europe, North America and Southeast Asia. These communities can help to bypass the restrictions on Internet access set forth within China. In this aspect, Wu dialect has much smaller overseas communities compared to Cantonese and Min.

Third, political factors also play a major role. For example, the growth of Min Nan Wikipedia is likely supported by the linguistic movements in Taiwan. The stabilization of Cantonese Wikipedia is likely supported by the fact that the majority language in Hong Kong is Cantonese, not Mandarin or English. The Taiwan government and the Hong Kong government, together with the local linguistic communities, have also taken measures to standardize aspects of Min Nan, Cantonese and Hakka.

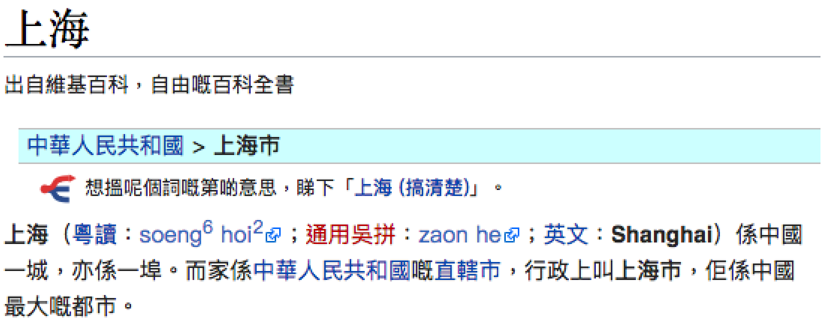
Another factor is writing systems. This will be the main focus of this article. In the next two sections, I will show examples of the type of writing systems in each of the

Wikipedia sites in Chinese dialects, and then I will compare these writing systems to how the Wikipedia sites in these writing systems are faring.

3. Writing Chinese Dialects

A Chinese dialect can be written in either a character-based system or a phonetic writing system. The Wikipedia sites that are written in a character-based system include Mandarin, Cantonese, Wu and Gan. Let's take a look at a snapshot of these websites by using the article on the city of Shanghai as an example, as shown in Figures 1, 2 and 3. I omit Mandarin because the writing system is standardized and well-known.

Figure 1 shows the article from the Cantonese Wikipedia site.



上海

出自維基百科，自由嘅百科全書

中華人民共和國 > 上海市

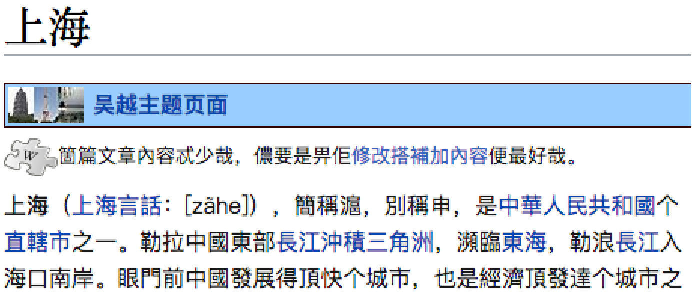
想搵呢個詞嘅第啲意思，睇下「上海 (搞清楚)」。

上海（粵讀：soeng^6 hoi^2；通用吳拼：zaon he；英文：**Shanghai**）係中國一城，亦係一埠。而家係中華人民共和國嘅直轄市，行政上叫上海市，佢係中國最大嘅都市。

FIGURE 1. Wikipedia page about Shanghai written in Cantonese

Cantonese is the only Chinese dialect that has developed a stable popular writing system which has been standardized to a greater extent than other dialects. According to Snow (2004: 6), written Cantonese can be traced back to the late Ming Dynasty (1368-1644), when books of verse were printed. Cantonese opera scripts were written down in characters in the early 20th century. Nowadays, although written Cantonese in many cases may contain elements from standard Chinese and Classical Chinese, the writing system is nonetheless capable of writing down spoken Cantonese (Snow 2004: 60).

Figure 2 shows the article from the Wu Wikipedia.



上海

吴越主题页面

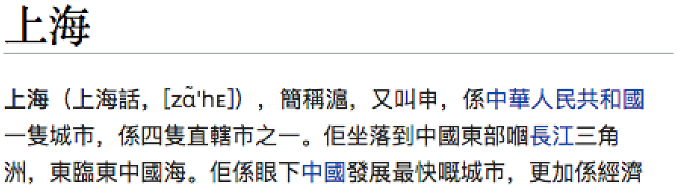
箇篇文章內容忒少哉，儂要是畀佢修改搭補加內容便最好哉。

上海（上海言話：[zãhe]），簡稱滬，別稱申，是中華人民共和國个直轄市之一。勒拉中國東部長江沖積三角洲，瀕臨東海，勒浪長江入海口南岸。眼門前中國發展得頂快个城市，也是經濟頂發達个城市之

FIGURE 2. Wikipedia page about Shanghai written in Shanghainese

Traditionally the representative version of Wu is that of Suzhou. Vernacular writing based on the Suzhou dialect can be traced as far back as early Qing Dynasty (1644-1912). There are texts of fiction and opera written in mixed Classical Chinese and Suzhou dialect by using characters. In the formation of the Shanghai dialect, one important contribution is Suzhou dialect. Therefore even though the contemporary representative version of the Wu dialect is that of Shanghai, the tradition of writing Wu dialects has been present in Shanghai as well. According to the texts cited by Qian (2003: 357–394) from the mid-19th and early 20th centuries, colloquial Shanghainese could be written down with characters. The degree of popularity and standardization of written vernacular Shanghainese is to a much lesser degree compared to Cantonese.

Figure 3 shows the article from the Gan Wikipedia.



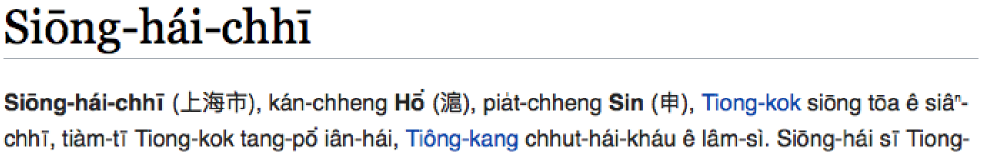
上海

上海（上海話，[zɑ̃'hɛ]），簡稱滬，又叫申，係中華人民共和國一隻城市，係四隻直轄市之一。佢坐落到中國東部嘓長江三角洲，東臨東中國海。佢係眼下中國發展最快嘅城市，更加係經濟

FIGURE 3. Wikipedia page about Shanghai written in Gan

The representative version of the Gan dialect is that of Nanchang. The internal homogeneity of the Gan dialect is relatively high. Although the Gan dialect can be written with a character-based writing system, e.g. as in the dictionary by Xiong (1995), there has not been a tradition of a popular vernacular writing in the Gan dialect.

All of the other Chinese dialect Wikipedia sites are currently written in a phonetic writing system. Figure 4 is the Min Nan page about Shanghai.



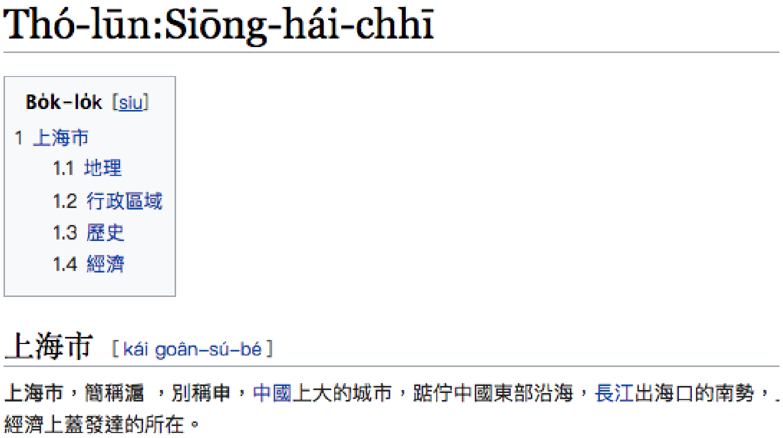
Siōng-hái-chhī

Siōng-hái-chhī (上海市), kán-chheng **Hō͘** (滬), pia̍t-chheng **Sin** (申), Tiong-kok siōng tōa ê siâⁿ-chhī, tiàm-tī Tiong-kok tang-pō͘ iân-hái, Tiông-kang chhut-hái-kháu ê lâm-sì. Siōng-hái sī Tiong-

FIGURE 4. Wikipedia page about Shanghai written in Southern Min

As with all of the other southern Chinese dialects, Southern Min can be written with characters. The earliest known written vernacular Southern Min is an opera script titled *The Tale of the Lychee Mirror* [Lì Jìng Jì 荔镜记] dated 1566 in the Ming Dynasty. According to Lin (1999), the development of written Taiwanese using a character-based system has not been up to the degree of Cantonese, and there are more issues with standardization as well, although speakers of Taiwanese nowadays do use the character-based writing system, especially in popular culture, e.g. song lyrics, film subtitles, etc. The Taiwan government has taken measures to standardize the character set used for Taiwanese Southern Min since 2007.

On the other hand, Southern Min has a long tradition of phonetic writing, such as those designed by early missionaries. Some of these systems were once quite popular and had a basis of literacy among speakers who might not know how to write Chinese characters. One system is the POJ system (Pe̍h-ōe-jī 白话字), or Church Romanization, designed by the Presbyterian Church in the 19th century. It has a sizable literature as well. Apart from political reasons that might disfavor using a character-based system, the practical usefulness of the phonetic writing system does seem to show the choice is reasonable. However, as shown in Figure 7, on the discussion page the contributors also use the character-based system almost exclusively.


Thó-lūn:Siōng-hái-chhī

Bo̍k-lo̍k [siu]

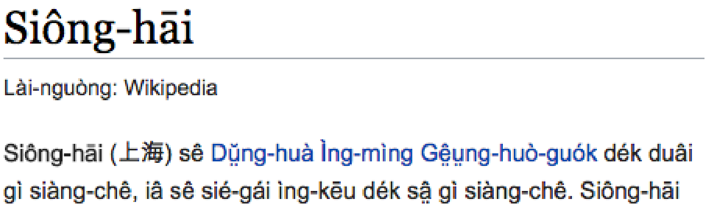
1 上海市
1.1 地理
1.2 行政區域
1.3 歷史
1.4 經濟

上海市 [kái goân-sú-bé]

上海市，簡稱滬 ，別稱申，中國上大的城市，踮佇中國東部沿海，長江出海口的南勢，.
經濟上蓋發達的所在。

FIGURE 5. The discussion page in Southern Min

Figure 6 shows the article about Shanghai writing in Min Dong based on Fuzhou.


Siông-hāi

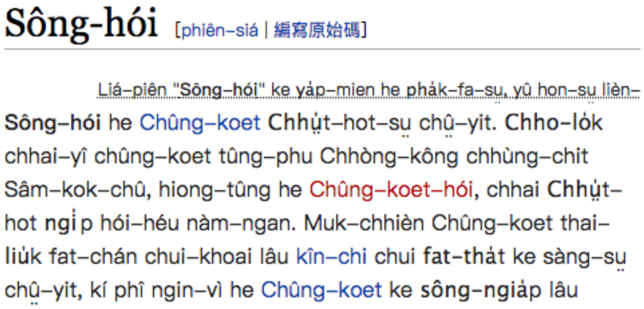
Lài-nguòng: Wikipedia

Siông-hāi (上海) sê Dṳ̆ng-huà Ìng-mìng Gê̤ṳng-huò-guók dék duâi gì siàng-chê, iâ sê sié-gái ìng-kēu dék sâ̤ gì siàng-chê. Siông-hāi

FIGURE 6. Wikipedia page about Shanghai written in Min Dong

The character-based writing of Fuzhou can be traced back to the 16th century. The early records include the rime book *Qī Lín Bāyīn* [戚林八音 The Book of Eight Tones], and the fiction writing *Mǐn Dū Bié Jì* [闽都别记 Alternative Records of the Capital of the Min] from the mid-Qing Dynasty. However the writing tradition in characters in Eastern Min has not been as popular as in Southern Min. Consequently practice of writing Eastern Min in characters is confined to a limited group of people. The once popular form is the BUC system (Bàng-uâ-cê 平话字) designed by missionaries in the 19th century.

Figure 7 shows the article on Shanghai written in Hakka. Note there is one line of characters after the title, which gives a link to edit the article. But the article itself is written in a phonetic writing system.



Sông-hói [phiên-siá | 編寫原始碼]

Liá-piên "Sông-hói" ke ya̍p-mien he pha̍k-fa-sṳ, yû hon-sṳ lièn-

Sông-hói he Chûng-koet **Chhṳ̍t**-hot-sṳ chṳ̂-yit. **Chho-lo̍k** chhai-yî chûng-koet tûng-phu Chhòng-kông chhùng-chit Sâm-kok-chû, hiong-tûng he Chûng-koet-hói, chhai **Chhṳ̍t**-hot **ngi̍p** hói-héu nàm-ngan. Muk-chhièn Chûng-koet thai-liu̍k fat-chán chui-khoai lâu kîn-chi chui **fat-tha̍t** ke sàng-sṳ chṳ̂-yit, kí phî ngin-vì he Chûng-koet ke **sông-ngia̍p** lâu

FIGURE 7. Wikipedia page about Shanghai written in Hakka

Hakka can be written in Chinese characters, although there has not been much study on this topic. In terms of the phonetic systems, there have been systems designed by missionaries, e.g. Pha̍k-fa-sṳ (白話字) created by the Presbyterian church in the 19th century. The Taiwanese Hakka linguistic community and the Taiwan government also adopted the Taiwanese Hakka Romanization System in 2012.

Although the Wikipedia sites in Xiang, Min Bei and Pu-Xian Min are still being incubated, some pages exist nonetheless. The Xiang Wikipedia uses a character-based system, but has two side-by-side versions, one for Old Xiang, and one for New Xiang, which is due to the significant differences between these two versions of Xiang. In this sense, the Wu Wikipedia could also have multiple versions. The Min Bei and Pu-Xian Min Wikipedia sites use a phonetic system similar to earlier systems designed by missionaries in the 19th century.

The data here are summarized in Table 4. The dialects in parentheses are those Wikipedia sites still being incubated. Although in theory and in practice (to varying degrees) all Chinese dialects can be written with a character-based writing system, writing tradition and practical needs vary and therefore on these Wikipedia sites, different writing systems are used, among other reasons. Character-based systems are used on the Wikipedia sites of Mandarin, Cantonese, Wu, and Gan, and also on the preliminary pages of Xiang. In the Min dialects (i.e. the four Min Wikipedia sites), and in Hakka, a phonetic writing system is used, which mostly can be traced back to earlier systems designed by missionaries in the 19th century.

TABLE 4. Writing Chinese Dialects on Wikipedia

Character-Based	Letter-Based
Mandarin	Southern Min
Cantonese	Hakka
Gan	Min Dong
Wu	(Min Bei)
(Xiang)	(Pu-Xian Min)

In the next section, I look at the choice of writing system in connection with the development and growth of the Wikipedia sites.

4. Writing system and linguistic diversity

Systematic research on the writing systems used in Chinese dialects is quite rare. The practice of writing Chinese dialects has also been equally sparse for the most part of the history of the Chinese language. This can be explained by the following factors.

First, the *Law of the People's Republic of China on the Standard Spoken and Written Chinese Language* recognizes the use of languages of different ethnic groups within China. The minority languages, e.g. Mongolian, Zhuang etc., have the legal rights to use their own languages alongside Putonghua. For the minority languages that did not have a writing system, or in the case of the Zhuang language which has a character-based writing system[8], new phonetic writing systems were created to standardize the use of these languages by the Chinese government since 1949 (Zhou 2003). Despite the various issues with the language policy towards minority languages in China, the legal status of minority languages at least draws attention to the use and standardization of these languages both in the spoken form and in the written form. However, the various Chinese dialects are not recognized as such. Therefore, the standardization and the creation of a writing system for Chinese dialects were never formally considered. Even in Taiwan, the standardization of the writing systems for Taiwanese and Hakka is still quite recent, and these measures have limited effects outside Taiwan in the Southern Min and Hakka linguistic communities.

Second, the language laws in China also do not allow the explicit use of dialects in all official media. Although there have always been gaps between language laws and the implementation of such laws in language practices, in most cases dialect writings are not possible. Especially in primary education, no explicit teaching in writing dialects is

[8] Gǔ Zhuàngzì 古壮字 in Chinese, or Sawndip 𭨡𮦺(“saw + ndip”: writing raw) in Zhuang. It is a similar system to the Chữ Nôm 𡨸喃 used in Vietnam.

allowed, although some areas, e.g. Shanghai, have introduced classes of dialects outside the normal curriculum in elementary schools. More importantly, the language laws command economic incentives. Learning Mandarin means more economic and employment opportunities, and the use of writing in dialects is practically quite limited.

Third, traditionally the use of Chinese dialects mostly is confined to the spoken form, and this is true of most dialects even nowadays. Thus when people write, they tend to write standard Chinese. The need to write dialects is not strong enough to call for a full writing system for most dialects.

Fourth, all Chinese dialects share a core vocabulary to different extents (Wang 1994: 1448; Wang 1998: 530), and therefore writing Chinese dialects have always been possible with Chinese characters, with additional dialect characters[9] added. The need to create a dialect writing system has not been urgent for most dialects, because they can all be written somehow and to some degree for practical purposes. In cases of words for which the etymologically correct characters[10] cannot be determined, or are too specialist for the average speaker to use, homophonous characters can be used to write those words.

For all these reasons, the research and practice in writing dialects in the Chinese context have been quite rare. Now with the emergence of new technology and media such as Wikipedia, which gives Chinese dialects a channel to become fully functional in both the spoken form and the written forms, the lack of systematic research and practice in writing definitely is a major obstacle to the growth of these dialect Wikipedia sites.

But all dialects are not equal. As I have discussed in section 3, Cantonese has created and standardized the writing system to the most degree among all Chinese dialects. Writing Cantonese is not really an issue. This can be shown in the relative high ranking of the Cantonese Wikipedia as shown in Table 2 and Table 3. The Cantonese Wikipedia is relatively stable and has the largest user base after Mandarin Wikipedia.

In contrast, the Wu dialect has a large linguistic community but ranks last in Table 3 in terms of the number of articles, although the total number of users ranks right after Cantonese. Among the factors mentioned before, e.g. the actual speakers of Shanghainese being much smaller than all Wu dialect speakers, the lack of a standardized writing system and the lack of basic literacy education might also be factors.

Although the Gan Wikipedia is written in a character-based system, it is to an even lesser degree in terms of standardization and basic literacy education. Thus Gan Wikipedia is actually losing its momentum, as shown in the data in Table 2 and Table 3. Within the two years, there was little increase of the total number of articles and the ranking of the Gan Wikipedia dropped from 143 to 154. Similarly, in the Xiang Wikipedia, the same issues exist, in addition to the fact that the two versions of Xiang, i.e. Old Xiang and New Xiang, are so different that they call for two versions of the Xiang Wikipedia.

[9] Fāngyán zì 方言字

[10] Fāngyán běnzì 方言本字

Regarding Min Nan, people have been using characters to write in recent decades, especially in Taiwanese popular culture. However Min Nan Wikipedia uses a phonetic writing system. This might be due to three factors. First, the need for a unique identity as a political factor can lead some speakers to favor a phonetic system, since it looks radically different from Mandarin Chinese writing. Second, the Southern Min dialect is probably the most advanced among all Chinese dialects in terms of the phonetic writing system. Although phonetic writing systems were created by missionaries in the 19th century for many varieties of Chinese, the POJ system was the most successful in producing a large body of literature and in its literacy education. Third, the standardization that took place in Taiwan only has limited effects on Southern Min spoken outside Taiwan. Therefore to reach a larger readership, a phonetic writing system does seem to have its advantage given the high internal homogeneity among the major Southern Min speaker communities. As can be seen from Table 2 and Table 3, the growth of Min Nan Wikipedia within the two years was phenomenal! Although this has to be ascribed to the enthusiasm of a smaller number of contributors, as can be seen from the increase of the total number of articles from 12,798 to 208,033, a 15-time increase, while the total number of users only increased from 21,324 to 28,898. But there is no doubt the phonetic writing system facilitates the creation of articles.

Hakka has a similar situation in terms of its writing system compared to Min Nan, although the practice of writing Hakka in characters has not been to the same extent as in Min Nan. The Hakka Wikipedia grew tremendously, as can be seen by the 65% increase of total number of articles, and 40% increase in total number of users. The ease of the phonetic writing system is likely a contributing factor.

For the other two Min dialect Wikipedia sites, i.e. Min Bei and Pu-Xian, their choice of using a phonetic writing system is based on a lack of character-based writing. But the phonetic writing system is equally less popular in practical use. Therefore there is no actual momentum in bringing these sites out of the incubator. We see here the lack of a practical popular writing system does seem to be an obstacle to the growth of these sites.

In summary, I argue that a practical popular writing system is an important factor in the growth and maintenance of Chinese dialect Wikipedia sites. By "popular" I mean the actual use of the writing by the average speakers. For the most successful ones, i.e. Cantonese and Min Nan, both enjoy a popular writing system that has a large user base, and their virtual linguistic communities can build upon such a user base to promote these dialects. For the less successful ones, e.g. Xiang, Wu, Min Bei, Pu-Xian, and Gan, the lack of a practical popular writing system impedes the growth and maintenance of these sites, hence hampering efforts to promote these dialects. Compared to these two groups, the Hakka Wikipedia seems to be doing quite well, maybe more or less in the middle.

5. Conclusions

This article is part of my larger project to explore the creation of the standard form of modern Chinese, i.e. Putonghua, and its relation to nation-building. Here I have shown that Wikipedia is an important tool to promote linguistic diversity. A practical popular writing system is needed to guarantee the success of such sites. In connection to what writing systems to use, there are various other issues.

One issue is related to the classification of Chinese dialects. Although there are seven major groups, the actual mutually-unintelligible forms of Chinese can be much greater than seven. Even among the Mandarin group, speakers from different areas do not necessarily understand each other. Moreover, the Jin dialect has been recognized by many scholars as a separate group. Therefore there is the issue of how many Wikipedia sites of Chinese dialects should be recognized. As Ensslin (2011) points out, "Wikipedia defines itself as 'the biggest multilingual free-content encyclopedia on the internet', thus featuring an explicit language policy in its mission statement". Thus to be recognized as a language by Wikipedia is not an automatic process.

Another issue is internal homogeneity. Among many dialect groups, there are local speech forms that are not mutually-intelligible. For example, the distinction between Northern Wu and Southern Wu, and that between Old Xiang and New Xiang. Even among groups or subgroups that have greater internal homogeneity, which version should be regarded as the representative is a major issue, such as in the case of Wu. These two issues need to be sorted out before standardization on the form and writing of dialects can be carried out. Then after standardization, literacy education and content or literature creation need to be addressed.

Furthermore for the majority of Chinese dialects, there has never been a writing system, either character-based or phonetic. If one is to create a writing system, which way is to go? In terms of the advantages and disadvantages of these two types of writing, the character-based system is considered more authentically Chinese, and can be partially understood by speakers of other dialects. But for the uniquely local vocabulary, it is more difficult to write with characters. Moreover, the etymologically correct characters might be very rare characters that can be difficult to input. The unique dialect characters may also be difficult to input. The phonetic system can be considered less authentically Chinese, and the diacritics for tones and vowels can be overwhelming both typographically and in terms of readability. However a phonetic system is much easier to create and to learn for everyone, including people who do not know Chinese characters. Therefore a phonetic writing system is more efficient if one is to create a writing system for a dialect that has never been systematically written. Such systems can be very instrumental in promoting linguistic diversity, especially by using Wikipedia sites.

This paper has drawn attention to the importance of writing systems for Chinese dialects in the process of promoting linguistic diversity, especially with new technological tools and channels such as Wikipedia, given the context where language policy restricts the maintenance of dialects. It is my hope that more research will be conducted in this respect in the future to solve both the theoretical and practical issues.

REFERENCES

BAKER, COLIN, and SYLVIA PRYS JONES (eds). 1998. *Encylcopedia of bilingualism and bilingual education.* Multilingual Matters.

DENG, HONGBO (邓洪波). 1994. Zhèngyīn shūyuàn yǔ Qīngdài de guānhuà yùndòng 正音书院与清代的官话运动 [Mandarin Academies and the Mandarin Campaign in Qing Dynasty]. Journal of East China Normal University 3: 79-86. Shanghai, China.

DONG, HONGYUAN. 2014. *A history of the Chinese language*. New York, NY, USA and Abingdon, UK: Routledge.

DONG, HONGYUAN. 2015a. Mandatory Mandarin: An archival study of Qing dynasty language policy. Manuscript, George Washington University.

DONG, HONGYUAN. 2015b. New media as channels for linguistic diversity: A case study of Chinese. Manuscript, George Washington University.

DONG, HONGYUAN. 2016. An archival study on linguistic reforms in pre-modern East Asia. Manuscript, George Washington University.

DONG, HONGYUAN. 2017. An historical comparative view on contemporary language policy in China. Manuscript, George Washington University.

ENG, ROBERT Y. 2010. Is Cantonese in danger of extinction? The politics and culture of language policy in China. Blog post August 20, 2010 on *China Notes: Superfluous Musings of a Chinese Historian*. Retrieved on November 20, 2017 from http://chinamusictech.blogspot.com/2010/08/is-cantonese-in-danger-of-extinction.html

ENSSLIN, ASTRID. 2011. What an un-wiki way of doing things: Wikipedia's multilingual policy and metalinguistic practice. *Journal of Language and Politics* 10(4): 535-561.

IVKOVIC, DEJAN., and HEATHER LOTHERINGTON. 2009. Multilingualism in cyberspace: Conceptualising the virtual linguistic landscape. *International Journal of Multilingualism*, 6(1): 17-36.

LIN, ALVIN. 1999. Writing Taiwanese: The development of Modern Written Taiwanese. *Sino-Platonic Papers* 89, Dept. of Oriental Studies, University of Pennsylvania.

LIU, JIN. 2013. *Signifying the local: Media productions rendered in local languages in Mainland China in the new millennium*. Leiden: Brill

LIU, JIN, and HONGYIN TAO. 2009. Negotiating linguistic identities under globalization: language use in contemporary China. *Harvard Asia Pacific Review*, 10(1): 7-10.

LIU, JIN, and HONGYIN TAO. 2012. *Chinese under globalization: Emerging trends in language use in China*. World Scientific.

MAIR, VICTOR H. 1991. What is a Chinese "dialect/dialect"? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers, 29*. Department of Oriental Studies, University of Pennsylvania.

MAY, STEPHEN. 2006. Language policy and minority rights. In Thomas Ricento (ed.) *An introduction to language policy: Theory and method.* 255-272. Blackwell.

QIAN, NAIRONG (钱乃荣). 2003. *Shànghǎi yǔyán fāzhǎn shǐ* 上海语言发展史 [A history of the development of the language of Shanghai]. Shanghai, China: Shànghǎi Rénmín Chūbǎnshè 上海人民出版社 [Shanghai People's Publishing House].

SHOHAMY, ELANA, and DURK GORTER (eds). 2008. *Linguistic landscape: Expanding the scenery*. Routledge.

SIMMONS, RICHARD VANNESS. 2017. Whence came Mandarin? Qīng guānhuà, the Běijīng dialect, and the national language standard in early Republican China. *Journal of American Oriental Society* 137, 1: 63-88.

SNOW, DON. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong, China: Hong Kong University Press.

WANG, HUI. 2014. *China from empire to nation-state*, translated by Michael Gibbs Hill. Cambridge, MA: Harvard University Press.

WANG, WILLIAM S.-Y. 1994. Glottochronology, lexicostatistics and other numerical methods. In the *Encyclopedia of language and linguistics*, edited by R. E. Asher and J. M. Y. Simpson. 1445-1450. Oxford and New York: Pergamon Press.

WANG, WILLIAM S.-Y. 1998. Three windows on the past. In *The Bronze Age and early Iron Age peoples of eastern Central Asia*, edited by Victor H. Mair. University of Pennsylvania Museum Publications. 508-534.

WU, YONGBIN (吴永斌). 2008. Shì xī Yōng-Qián nián jiān de guānhuà yùndòng 试析雍乾年间的官话运动 [An analysis of the Mandarin Campaign in the Yongzheng-Qianlong era]. *Mínzú Jiàoyù Yánjiū* 民族教育研究 [Journal of Research on Education in Ethnic Minorities]. 2:113-116. Beijing, China.

WURM, STEPHEN ADOLPHE; RONG LI; THEO BAUMANN; and MEI W. LEE (eds). 1987. *Language Atlas of China*. Hong Kong: Longman.

Xiong, Zhenghui (熊正辉). 1995. *Nánchāng fāngyán cídiǎn* 南昌方言词典 [A dictionary of the Nanchang dialect]. Nanjing, China: Jiāngsū Jiàoyù Chūbǎnshè 江苏教育出版社 [Jiangsu Education Publishing House].

YUAN, JIAHUA (袁家骅) et al. 1960. *Hànyǔ fāngyán gàiyào* 汉语方言概要 [An outline of Chinese dialects]. Beijing, China: Wénzì Gǎigé Chūbǎnshè 文字改革出版社 [Writing System Reform Publishing House].

ZHOU, MINGLANG. 2003. *Multilingualism in China: The politics of writing reforms for minority languages 1949-2002*. Walter de Gruyter.

ZHOU, MINGLANG. 2006. Theorizing language contact, spread, and variation in status planning: A case study of Modern Standard Chinese. *Journal of Asian Pacific Communication*, 16, 2: 159-174.

ZHOU, MINGLANG, and HONGKAI SUN (eds). 2004. *Language policy in the People's Republic of China: Theory and practice since 1949*. Springer.