Why One Dictionary Became Two Books:

Chinese Phonology and Script Reform in the Demopaedia Project

by Nicolas Brouard, chercheur émérite, Institut national d'études démographiques (Ined)
15 December 2025

Summary: This text attempts to summarize what we learned about the invention and revision of modern Chinese writing and phonology over more than 120 years that made it necessary to publish two variants of the Chinese volume of the unified edition of the ***Multilingual Demographic Dictionary****: one in* ***simplified characters*** *and the other in* ***traditional characters****.*

It was not initially planned to publish two books, but as the project progressed, it became evident that two versions would be necessary to produce printed versions of the dictionary. While the texts of the dictionary are identical, except for the character variants, the difficulty arose during the creation of the index. Due to the gradual abandonment of handwriting in favor of computers and smartphones, the traditional classification by stroke count has fallen into disuse. We had to find more appropriate classification methods. However, we discovered that each proposed classification method is specific to the character variant. In the process, we learned much about the development of two different systems, pinyin, mostly used in mainland China, and zhuyin, which is used by Chinese communities outside of mainland China.

This Chinese volume of the unified edition of the Multilingual Demographic Dictionary differs from the first (1982) and second (1992) editions in that the English translation of each Chinese demographic term or expression no longer appears in the text of the nine chapters of the dictionary but only in the index. Originally, and in accordance with the spirit of the United Nations Terminology Commission in the mid-1950s, each paragraph of a volume was to be freely written to introduce the various demographic terms or expressions for a particular edition of the dictionary (first in 1958, second in 1975, or unified in 2013).

The best example of independent writing is the German volume of 1987 where the authors wrote entirely new texts building on the 1960 edition. It was from this second German edition, which introduced new demographic terms borrowed from epidemiology, that the so-called unified edition was constructed, integrating all the terms omitted in the second editions in English, French, Spanish, and Arabic, among others. To date, the German edition, like the Korean, Japanese, Thai, Malay, Arabic, and Russian editions, are still not unified because of missing terms in other languages.

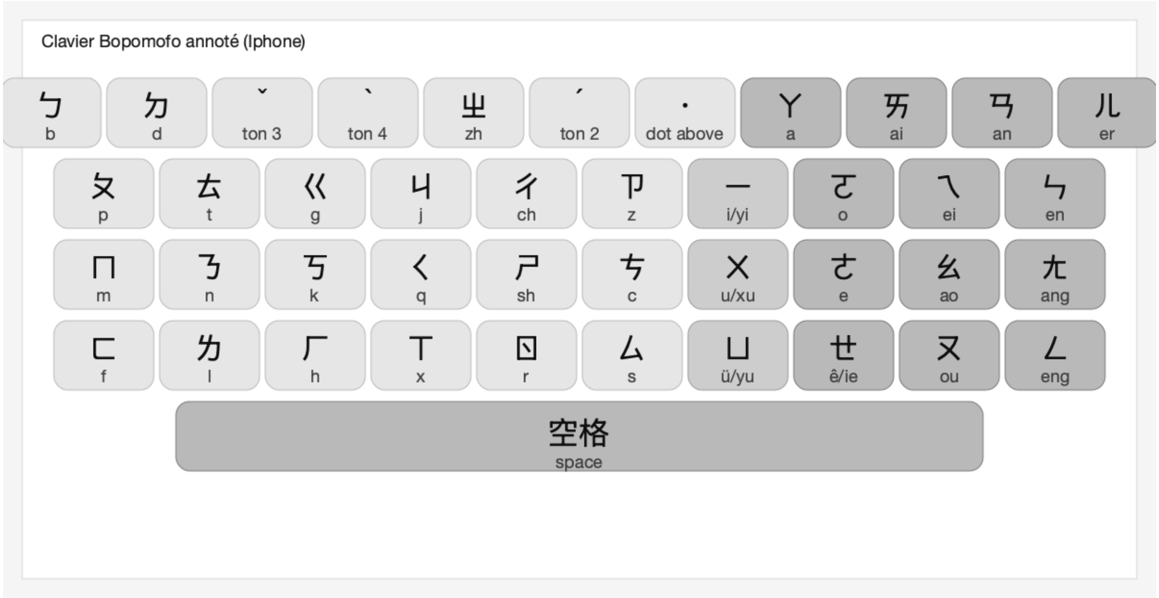
Demopædia offers internet users web access to all published dictionaries, allowing easy navigation from one language and edition to another. However, it is also important to preserve what has made this dictionary successful since the first editions of 1958: a **small, printed book with a comprehensive and easy-to-consult index** for searching demographic expressions. However, with digitization and the progressive abandonment of handwriting by younger generations, we encountered a major problem in producing the index for the Chinese volume. Many older Chinese dictionaries used indexes sorted by the number of strokes of the first character, then the second, and so on. But if you ask a young Chinese person to count the strokes of a character with more than ten strokes, there is a good chance they will make a mistake, as they may no longer know how to write it correctly. Surveys by the China Youth Daily ("Fewer and fewer opportunities," 2017) confirmed a crisis in handwriting among young Chinese, with figures close to 60–80% for difficulties related to forgetting characters.

Typing on a computer or smartphone replaces the need to know Chinese writing. That is why the youngest author[1] of this Chinese volume proposed sorting demographic terms or expressions according to their phonetic transcription in Latin characters, known as *pinyin*. Thus, the expression 701-1 'population growth' 人口增长, whose *pinyin* transcription is « rén kǒu zēng zhǎng », is classified before the expression 105-2 'population policy' 人口政策 « rén kǒu zhèng cè », because the letter « e » comes before « h » in the Latin alphabet.

However, when presenting the prototype of this dictionary to a Chinese student in demography and history, he told us that he was not very familiar with simplified characters and did not use the *pinyin* transcription. Instead, he used the *zhuyin* phonetic transcription, which is based on 37 ordered characters invented in 1913 in mainland China. The first four characters, ㄅㄆㄇㄈ, represent the sounds b, p, m, f, and are commonly called *bopomofo*.

Figure 1 shows such a keyboard, which can be easily defined on a smartphone (we have added Latin phonetics). To enter 人口, you would successively click on the four keys: « ㄖ|r », « ㄣ|en », « ㄎ|k », « ㄡ|ou ». After entering just three characters, the selected dictionary will propose 人口 (one might need to enter the tones to suppress ambiguity). Similarly, if you choose a *pinyin* keyboard, you would successively click on « renk », and without needing to click on « o » and « u », 人口 would be proposed.

Figure 1 Zhuyin or Bopomofo keyboard (iPhone) with Latin phonetics



Clavier Bopomofo annoté (Iphone)

Thus, we needed to consider the different users for simplified Chinese characteristics (with *pinyin* indexes) and traditional Chinese characteristics (with *zhuyin* indexes).

But since this phonetic transcription is only used for pronunciation and not for understanding a Chinese word or expression, we could have omitted them both in *pinyi*n and in *zhuyi*n. However, if the software makes a transcription error, the phonetic index will not be sorted correctly. Neither Google Translate nor Word, for example, offer transcriptions of sufficient quality. Moreover, although each character corresponds to a Unicode code, there is no universal sorting order for Chinese characters, as pronunciation depends on context.

[1] Feinuo Sun, currently an assistant professor at the University of Texas at Arlington, proposed using pinyin to Xiaochun Qiao, honorary professor of demography at Peking University and the main author of this volume who replaced Yeun-chung Yu, a Chinese demographer from the UN Population Division, after his death.

For example, 都会区 (307-4) was initially transcribed as « dōu huì qū », but the correct pronunciation is « dū huì qū » (ㄨ and not ㄡ in *zhuyin*), because the meaning of the character 都 depends on the context: « all, both » or, in this case, « metropolis, area ».

Chyong Fang Ko, an honorary Research Fellow at Academia Sinica in Taiwan, kindly agreed to verify and correct the phonetic transcriptions. These corrections were integrated into a very long Python program specially written to input all the WIKI texts entered by the authors in simplified Chinese and to output all the SQL tables to feed the indexes (see the open source program at the Demopaedia GitHub URL of Joseph Larmarange).

Thus, the first volume includes an appendix with an index of dictionary words in simplified Chinese, sorted by *pinyin*. To help the reader pronounce the terms, we have added their tonal *pinyin* transcription. Since there is a strict correspondence between the two phonetic transcriptions, it would have been sufficient to add the *zhuyin* transcription after the *pinyin*. However, the alphabetic *pinyin* sorting, which only sorts on letters, is totally different from the phonetic *zhuyin* sorting. For example, « population growth » 人口增长 is classified before « population policy » 人口政策 in *pinyin*. In z*huyin*, “population growth” is classified after “population policy” because the sound ㄓ (zh) comes before ㄗ (z).

Bopomofo characters are sorted in columns starting from the left of the keyboard (and using Latin transcription instead of *zhuyin* characters):

- Initials: bpmf, dtnl, gkh, jqx, zh-ch-sh-r, zcs
- Finals: a o e ê; ai ei ao ou; an en ang eng; er
- Medials: i, u, ü

Zhuyin always writes medials explicitly (兄 brother x+ü+eng:ㄒ ㄩ ㄥ → initial + medial + final), while *pinyin* often merges them into the final (e.g., xiōng written as “x” + “iong” hides medial ü ㄩ). This is also due to *pinyin* being restricted to 26 Latin letters, compared to 37 *zhuyin* symbols.

One solution would have been to create a second index sorted by *zhuyin*, but the Chinese volume had to remain small. Moreover, because *zhuyin* users often read and prefer traditional Chinese characters, it seemed preferable to “manufacture” and publish a second book in traditional characters. We say « manufacture » because, while it is possible to specify the variant (Cyrillic vs. Latin for Serbo-Croatian, or traditional vs. simplified for Chinese) for better on-screen reading, by adding the Wikipedia option “&variant=_zh_hant” at the end of each URL, “_zh_hans” being the default (see for example http://zh-ii.demopaedia.org/w/index.php?title=10&variant=zh-hant), such a solution does not exist for a printed book. And since we were making a second, longer book intended for a much smaller audience of Chinese speakers, we also added a « traditional » index, sorted by stroke count. The maximum number of strokes in this dictionary is 24 for traditional characters, compared to 15 for simplified characters.

In the stroke-count-sorted index, « population policy » 人口政策 comes before « population growth » 人口增長, because although both expressions are grouped in the two-stroke class due to the first character 人, the third characters 政 and 增 have 9 and 15 strokes, respectively. It should also be noted that the traditional character 長, which contains 8 strokes, has been simplified to 长, with only 4 strokes.

With 15 strokes, Chinese speakers can miscount, which is why the *zhuyin* index will, we hope, be useful. Finally, and similarly to the two previous editions (1982 and 1992) of the Chinese volume, we have added an index sorted in English. Due to the correspondence between the numbers and terms of a unified edition in our database, it is possible to propose the index of the unified English edition and attach the Chinese translation for each term. We could equally create an index sorted in French, Spanish, or Italian for this unified Chinese volume, as these three volumes are also unified.

The English index of the traditional dictionary contains both *zhuyin* and *pinyin* phonetic transcriptions, while the simplified dictionary index contains only the *pinyin* phonetic transcription. Although *pinyin* and simplified writing are widely used in mainland China, traditional writing and the *zhuyin* phonetic alphabet are taught from primary school in Taiwan. There are, of course, *pinyin* courses for students in major universities, such as Academia Sinica, for those who need to communicate with mainland China. Conversely, young Chinese from mainland China rarely read traditional characters and are mostly unaware of the *zhuyin* phonetic system or its origins, even though it was invented on the mainland and not in Taiwan.

A short history of modern Chinese writing

We wanted to know if technological developments would bring the two writing systems closer and if our dictionary, published in both variants, would remain useful in the future. Let us revisit the history of two fundamental reforms: modern Chinese phonology (1910–1920) and character simplification (1950-1955).

The first step was to demonstrate that a Chinese vernacular language could replace classical Chinese, just as Dante Alighieri (1304–1321) demonstrated in the Divine Comedy that Tuscan Italian could supplant Latin, the scholarly and religious language. Chinese vernacular languages began to replace classical Chinese in the early 20th century. The Baihua movement (see Gao 2018, Weng 2020) promoted the use of vernacular written Chinese, which developed and spread through magazines and books from 1910 to 1920, largely based on the northern Mandarin dialect. A second step followed: the standardization of pronunciation. Drawing on the earlier Wade–Giles phonological tradition (Giles 1892), the *zhuyin* phonetic system aimed to clarify and simplify the structure of Mandarin syllables by explicitly marking the initial, medial, and final components.

In 1919, the May Fourth Movement became a major nationalist and anti-imperialist force in modern Chinese history. It was sparked by public outrage over the terms of the Treaty of Versailles, which gave Japan control of German concessions in Shandong Province rather than returning them to China. In doing so, the Entente countries failed to recognize China's role in WWI. China had provided over 140,000 laborers to support the Entente's military efforts. The Chinese viewed their being sidelined as an affront to their sovereignty and did not sign the Treaty. In Beijing, student protests erupted on May 4th, 1919, centered in Tiananmen Square, leading to calls for political and social change. The May 4th movement is seen as a pivotal moment in Chinese history laying the foundations for China's turn towards socialism (Boissoneault 2017).

As part of this movement, *Baihua* and *zhuyin* transformed education and unified the Chinese language, laying the foundation for modern Mandarin. However, the *zhuyin* phonetic system is today almost unknown to young Chinese in mainland China. Although its influence was limited due to the abundance of dialects, both *zhuyin* and *pinyin* have proven their effectiveness in Taiwan and mainland China in promoting Mandarin in the educational system.

The legacy of *zhuyin* is recognized in many works and interviews, particularly from 1999 onward, by Zhou Youguang (who died in 2017 at the age of 100), the main author of the *pinyin* transcription in 1955, who appears to have wanted to remind mainland China of *zhuyin*'s importance and even acknowledges its adoption in Taiwan: " *Pinyin* was not created from nothing, *zhuyin* was one of its predecessors "*(Zhou Youguang, interview, February 10, 2005)*. "Without the *zhuyin* symbols, China would not have been able to rely on decades of experience in phonetic teaching; *zhuyin* established the principle of a clear separation between initials and finals; the success of *zhuyin* shows that phonetic annotation can coexist with Chinese characters and does not threaten their existence (Zhou 2003). Taiwan has preserved *zhuyin*, which is a precious cultural tradition. They use it very effectively." (Zhou Youguang interview, CCTV, 2003).

However, *pinyin* faced the difficult task of completing the Latin alphabet, where three consonants that are not heard in 'Latin' countries (except for the German 'ich') must not be confused with the sounds already present in *pinyin*: zh, ch, sh. Instead of inventing new characters foreign to the Latin alphabet or diacritical signs like č, š, ž, Zhou Youguang preferred to draw on the Soviet experience of the 1920s–1930s, where ***Latinxua Sin Wenz** (Latinized New Writing System)*, invented by Russian linguists like N. A. Semenov and Chinese linguists like Qu Qiubai, was successfully tested with hundreds of thousands of Chinese workers in the Soviet Union and communist-influenced areas. Latin characters were at that time considered as the international script of the proletariat, not Cyrillic.

At first sight, it may seem audacious to reuse Latin consonant letters to represent sounds that are specific to Mandarin Chinese. Yet this is precisely what **LatinXua** did—successfully and systematically—by assigning new phonetic values to three letters: **J, Q, and X**.

In LatinXua, **J** represents the Zhuyin initial ㄐ, a "voiceless alveolo-palatal affricate", comparable to the initial sound in *Dieu* in French, or to *dj* in *adjacent* and *j* in *jump* in English. **Q** corresponds to ㄑ, its aspirated counterpart, similar to *ti* in French *tien* or *question*, and close to the sound of *ch* in *cheese* in English. **X** represents ㄒ, a voiceless alveolo-palatal fricative, comparable to *ch* in French *chat*, *ssi* in *mission*, or *sh* in English *sheep*.

By contrast with earlier romanization systems, LatinXua preserved the four tonal diacritics inherited from Zhuyin, allowing tones to be marked explicitly without altering the spelling of the syllable. This stands in contrast to *Gwoyeu Romatzy***,** officially adopted by the Republic of China in 1928 as the official romanization system for writing standard Chinese, which encoded tones through spelling alternations—often by doubling vowels—resulting in greater phonological precision but reduced transparency. For example, the syllables *mā, má, mǎ, mà* (妈, 麻, 马, 骂) were rendered as *mha, ma, maa, mah*.

In this respect, LatinXua anticipated one of the key strengths later adopted by *Hanyu Pinyi*n: the economical reuse of the Latin alphabet to represent Mandarin phonology clearly, consistently, and with minimal visual complexity—demonstrating that a Latin-based system could be both pedagogically effective and linguistically robust.

After describing the genesis of these two versions of the unified edition of the Multilingual Demographic Dictionary and the reasons why we deemed it necessary to publish a short dictionary in *simplified characters* and a longer one in *traditional characters*, we can ask ourselves about the future of these two writing systems in an increasingly digitized world. It seems likely that both systems, due to their structural proximity (language, phonetics), will probably endure.

Simplified Chinese writing dates back sixty years, while traditional writing has existed for several thousand years. However, illiteracy rates are now very low in both regions, proving the efficacy of phonetic scripts. Among the illiterate, the new possibility of transcribing most dialects spoken into Mandarin via voice on a smartphone to exchange an SMS or on social networks like WeChat eliminates the need to know how to write *pinyin* or *zhuyin*!

Political barriers remain at the state level but also at the level of smartphone manufacturers. Indeed, it is possible to converse in simplified Chinese on many Android smartphones while entering text on a *zhuyin* keyboard (like the one represented above), but this is strangely impossible on an iPhone!

The **Demopædia dictionaries**, available both in web format and as printed books, are released under the **Creative Commons ShareAlike license**. The printed versions can be accessed through **Lulu.com**, a "print-on-demand" service, by searching for Demopaedia. Additionally, you can download the **PDF or EPUB versions** of the dictionaries in Chinese as well as in other languages by visiting the **Download URL link** of demopaedia.org. If your browser automatically changes the URL from **http to https**, please manually edit the URL back to **http** to ensure proper access.

References

Boissoneault, L. (2017). *The surprisingly important role China played in WWI*. *Smithsonian Magazine*. Special report: *World War I: 100 years later*. https://www.smithsonianmag.com/history/surprisingly-important-role-china-played-wwi-180964268/

Gao, Y. (2018). *The Baihua Movement and ideological revolution* (pp. 73–90). In *The birth of twentieth-century Chinese literature*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-55936-4_4

Giles, H. A. (1892/1982). *A Chinese–English dictionary*. Kelly & Walsh.

Weng, J. (2020). *Vernacular language movement*. In *Oxford bibliographies in Chinese studies*. Oxford University Press. https://doi.org/10.1093/OBO/9780199920082-0180

Zhou, Y. (1999). *Hanyu pinyin fang'an de lishi* [《汉语拼音方案的历史》; *History of the Hanyu Pinyin scheme*]. Commercial Press.

Zhou, Y. (2002). *Wenzi gaige liushi nian* [《文字改革六十年》; *Sixty years of script reform*]. Beijing Language and Culture University Press.

Zhou, Y. (2003). *Zhou Youguang wenji* (Vol. 3) [《周有光文集》第三卷; *Collected works of Zhou Youguang*]. Zhonghua Book Company.

Fewer and fewer opportunities to write—Do you still care if your words look good? (2017, March 14). *People's Daily Online*. http://edu.people.com.cn/n1/2017/0314/c1053-29144147.html