Improving Mandarin Chinese Searches Through Mechanical Division of Syllables

Vijay John
August 2007

Abstract

This paper presents an extension, *Làzhú* (蜡烛), of a previous method known as *Xiǎozhǐ* (小纸). Both methods search for several transliterations of an entered term of Chinese origin. In contrast to *Xiǎozhǐ*, however, this algorithm inserts a syllable break for every instance of the letters *g*, *n*, or *r* in the search term, as some of these were transliterated erroneously in the previous version. Thus, this method is relatively accurate in transliterating a higher proportion of Mandarin Chinese search terms than *Xiǎozhǐ*.

Introduction

The vast majority of keyboards are currently based on Roman letters. For this reason, a significant number of websites are written in Roman script as well. At times, they may include transliterations of words of Chinese origin. One is more likely to find English-language pages including terms of Chinese origin, e.g. names of Chinese cities, by entering the term in Roman script on a search engine than by entering the term in *Hànzǐ* (汉字) or Chinese characters.

In a paper presented at the tenth conference of the Texas Linguistic Society under the name *A Method for Enhancing Search Using Transliteration of Mandarin Chinese*, an algorithm called *Xiǎozhǐ* was proposed to enhance online search of terms of Chinese origin. It was assumed that a search term would be entered in *Hànyǔ Pīnyīn* (汉语拼音), the transliteration system officially used in the People's Republic of China (PRC)[1].

Xiǎozhǐ involved using a right-to-left method to break the search term into parts, each of which would be mapped to their equivalents in various transliteration systems for Mandarin Chinese (e.g. Wade-Giles, *Hànyǔ Pīnyīn* itself, etc.). The original method did not at all take *Hànyǔ Pīnyīn*'s rules for syllabic division into account. A later version included syllabic division only for words including the letters *i* or *u*, without fully considering occurrences of *ng* or of the consonants *n* and *r*.

Because the use of syllabic division was very limited under the original method, some terms were not broken into the correct parts. When any of the letters *g*, *n*, and *r* was used between two vowels in a search term, *Xiǎozhǐ* failed to transliterate the term correctly. The approach described below uses an interpretation of the rules for syllabic division in order to decompose the search term more accurately. Section 4 of this paper includes a

detailed explanation of the rules for defining a syllable in *Hànyǔ Pīnyīn*, as well as a technique for implementing these rules into *Xiǎozhǐ*.

Previous Work

Various researchers have dealt with the transliteration of foreign (i.e. non-Chinese) names into *Hànzǐ* (or "Chinese characters") and then to *Hànyǔ Pīnyīn*. The Romanized names are then transliterated into *Hànzǐ* from *Hànyǔ Pīnyīn*. Google® also includes a feature in its Simplified Chinese version facilitating search for pages written in Mandarin Chinese by transliterating the *Hànyǔ Pīnyīn* (entered by the user as a search term) into *Hànzǐ*. Other researchers have presented algorithms for transliterating other scripts (e.g. *naskh* as used in Arabic, *katakana*) into Roman characters mechanically.

[Stalls and Knight] mention an IBM system for transliteration of Arabic names. They then extend this previous algorithm in order to create a program transliterating the Arabic versions of the Western names back into Roman script. Their program deals with an aspect unique to languages using the Hebrew and Arabic scripts: most texts in Arabic (as well as in Hebrew, Persian, Yiddish, Urdu, Pashtu, and many other languages) substantially neglect to write vowels to represent their spoken equivalents.

[Stalls and Knight] also mention a method by Knight and Graehl that probabilistically constructs alternate terms, finally picking optimal search terms through graph algorithms.

Some other papers have used similar algorithms in order to solve such problems in Japanese and Chinese, i.e. to use accepted transliterations of names common among those who speak the particular language and convert the Romanizations into *kanji* and *hànzǐ*, respectively. Yet another paper [Mettler] uses a syllable-by-syllable transliteration method in order to transliterate foreign names from Roman letters into *katakana*, a writing system normally used in order to write foreign words in Japanese.

[Kuo] describes how to find transliterated pairs, i.e. terms in languages other than Mandarin and their transliterations into traditional *hànzǐ*, from the Web. The target of the transliteration is Chinese. It mentions various transliteration systems (Wade-Giles, *Tōngyòng Pīnyīn*, and *Hànyǔ Pīnyīn*). Terms are segmented from left to right. [Wan] introduces another left-to-right algorithm focused on transliterating English terms, especially English names, into Chinese. [Gao] creates Chinese terms out of English terms, such as names. It tries to find matching Chinese characters for English phonemes.

[Google] is part of the help section of the Simplified Chinese version of Google. It allows the searcher to use a slightly modified version of *Hànyǔ Pīnyīn* (not yet included in *Xiǎozhǐ*) to search for Chinese characters. If a term in *Pīnyīn* is entered, Google will search only for the original search term. Towards the top of the page displaying search results, it will also provide Chinese characters that could be spelled similarly in *Pīnyīn* (or a variation thereof). However, if the user wishes to search for the *hànzǐ* transliterated as *lü*, *lüe* (or *lue*), *nü*, *nüe* (or *nue*), then "lv, lve, nv, nve," respectively, must be entered as the search term. Otherwise, Google will simply search for the original search term.

Like the references mentioned here, Google does not address the problem of enhancing search in romanized letters of terms from Mandarin, although it deals with transliterations.

(Temporary note: Include any info here about *improvements* of such systems. If the improvements allow more terms to be transliterated, that would be nice. Even better would be if the improvements are similar, though not identical, to *Làzhú*!)

Xiǎozhǐ

Xiǎozhǐ, the approach introduced in *A Method for Enhancing Search Using Transliteration of Mandarin Chinese*, enhances Google search involving terms of Chinese origin. Currently, Google has the tendency of searching only for the entered term if it is of Chinese origin. It rarely, if ever, suggests alternative spellings or transliterations. *Xiǎozhǐ*, on the contrary, does not only search for the original term (assuming that it is entered in standard *Hànyǔ Pīnyīn* as outlined in *Lonely Planet*'s *Mandarin Phrasebook*). It also searches for *Tōngyòng Pīnyīn* and Wade-Giles transliterations.

In addition, *Xiǎozhǐ* includes six other transliteration systems, which are all *modifications* of the three mentioned above. A modification is another transliteration system that differs only minimally, sometimes with only one difference with another system. (For example, the system *Modified Hànyǔ Pīnyīn #1* or MHP1 differs from *Hànyǔ Pīnyīn* only in the transliteration of the *Hànyǔ Pīnyīn* component *üe*, which is transliterated in MHP1 as ***ue***).

Xiǎozhǐ transliterates the search term(s) (or "string") entered using a right-to-left transliteration method. It also includes a table of "components," i.e. relevant strings of variable length that may be individually transliterated.

To account for some syllabic division, the method begins with the following steps:

1. Does searchterm include a vowel immediately after an "i" or "u"? If so, let "mid" be the "i" or "u" followed by another vowel. Then proceed to step 3. If not, decompose query.
2. Does searchterm include a vowel or an "n" immediately before mid? If so, proceed to step 5. If not, proceed to step 4.
3. Is mid immediately preceded by "ng"? If so, proceed to step 5. If not, decompose query.
4. Let "query" be searchterm up to the vowel or "n(g)" immediately before mid. Proceed to step 6.
5. Decompose query. Proceed to step 7.
6. Remove query from searchterm. Proceed to step 8.
7. Decompose searchterm.

Xiǎozhǐ then searches for the entire string in the first column of a table, which includes all components in standard *Hànyǔ Pīnyīn*. If the string is not to be found, the program implementing *Xiǎozhǐ* deletes the last letter of the string. This process is repeated until a

string is found within the first column of the table. After that, all of the above-mentioned steps (beginning with the search for "i" or "u") are repeated using the discarded portion of the string.

In *A Method for Enhancing Search Using Transliteration of Mandarin Chinese*, it was noted that there are still search terms that are incorrectly decomposed in *Xiǎozhǐ*. The Chinese city name *Jǐnán* (济南) was presented as a sample term, and the problem with the decomposition was explained as improper syllabic division. To improve the effectiveness of *Xiǎozhǐ* in dividing a search term into the correct syllables when necessary, it is necessary to consider the syllabic rules more carefully to account for all cases in which lack of correct division between syllables poses a problem.

Hànyǔ Pīnyīn's Rules for Mandarin Syllabic Division

Some of these words may include the letters *g*, *n*, or *r*. These three letters may be found either at the beginning or at the end of a syllable in *Hànyǔ Pīnyīn*.

In Standard *Hànyǔ Pīnyīn*, no syllable may begin with a vowel unless it is at the beginning of a word[2] or preceded by an apostrophe. Thus, all syllables in the middle of a word must begin with either a consonant or an apostrophe. For example, the word for "Europe" in Mandarin is spelled *Ōuzhōu* (欧洲), but the disyllabic Chinese city name *Xī'ān* (西安) must be spelled using an apostrophe since **xian* represents only one syllable, pronounced [**sjen**].

Most Mandarin words can be transliterated into other systems using the right-to-left transliteration method that is described in detail in the paper *A Method for Enhancing Search Using Transliteration of Mandarin Chinese*. However, this system falsely assumes that a syllable in the middle of a word may begin with a vowel without being preceded by an apostrophe. For this reason, words including the letters *g*, *n*, and/or *r* are incorrectly transliterated by this system, unless the letters are in word-initial position. **(Provide examples? e.g. *dàngāo* "cake" (□ □))**

The problem posed by words containing these letters in syllable-initial, word-medial position could easily be solved by a left-to-right method. However, the differences in transliteration for different letters vary depending on phonological context. For example, "e" in *Hànyǔ Pīnyīn* is generally transliterated "e" in Wade-Giles, but "h**e**" in *Hànyǔ Pīnyīn* is "h**o**" in Wade-Giles. Because of differences like these, a right-to-left method is preferred for transliterating words of Chinese origin.

The method proposed in this paper implements the syllabic rules of *Hànyǔ Pīnyīn* while continuing the *Xiǎozhǐ* right-to-left transliteration method. In this new method, we begin by looking for any instances in the search term of "n," "r," or "ng" between two vowels. We then insert a syllable break immediately before "n," "r," or "g" (depending on whether we find between two vowels the consonants "n," "r," or "ng," respectively). Finally, we apply *Xiǎozhǐ*, so that the part of the search term before the break and the part after the break are decomposed separately.

Errors in *Xiǎozhǐ*

In some cases, *Xiǎozhǐ* could transliterate certain words that included the prerequisites correctly. For example, the province name `jinan` (i.e. *jǐnán* or 济南) is correctly transliterated, though the original paper cites the division of the word as needing improvement because the components are erroneously generated. Instead of finding the correct components `j`, `i`, `n`, and `an`, *Xiǎozhǐ* divides the word into the components `j`, `in`, and `an`.

It must be noted that it is mainly because of the small number of transliteration systems included that the term was correctly transliterated. Other systems distinguish between `i` + `n` and `in`; they would thus produce incorrect results. For example, if Yale Transcription were included in the tables of *Xiǎozhǐ*, *Hànyǔ Pīnyīn* `jinan` would be incorrectly transliterated as *`j`**`w`**`unyan` based on Yale Transcription (instead of `junyan`, i.e. `j` + `u` + `n` + `yan`). The syllable transliterated as *zhu* in *Hànyǔ Pīnyīn* is transliterated as *ju* in Yale Transcription. The syllable *zhun*, in contrast, is pronounced somewhat like the letters "-dge one" in the phrase "nu**dge one**." Thus, it is spelled *jwun* in Yale Transcription.

Although only nine transliteration systems (i.e. three transliteration systems and variations thereof) were included, some words were transliterated incorrectly. The word `binguan` (in the *Hànyǔ Pīnyīn* system) is incorrectly transliterated into Wade-Giles as *`pin`**`g`**`uan` and not as `pinkuan`, because *Xiǎozhǐ* classifies the "g" in this word as a part of the component `ing`. In reality, the word is divided into the two syllables `bin-guan`. Because the component "g" is transliterated as "k" in Wade-Giles, the original search term should be changed to `pin`**`k`**`uan`.

Additional Transliteration Knowledge

More transliteration systems have been added to the transliteration table. The original version already included *Tōngyòng Pīnyīn* and Wade-Giles in addition to *Hànyǔ Pīnyīn*. *Tōngyòng Pīnyīn* is the counterpart of *Hànyǔ Pīnyīn* in the Republic of China (ROC), sometimes known as Taiwan. Wade-Giles was the official system in the ROC before 1998 and is still used by many Taiwanese.

Standard *Hànyǔ Pīnyīn* is represented in the first column of the table under the abbreviated heading "HP." However, the table also includes the columns MHP1, MHP2, and MHP3. (MHP stands for **M**odified ***H**ànyǔ **P**īnyīn*). *Tōngyòng Pīnyīn* is represented in two columns, TP1 and TP2. Finally, Wade-Giles is included in the columns WG1, WG2, and WG3. The differences between the variations of each transliteration system are minute. Nevertheless, they are included in separate columns; otherwise, it would be necessary to have the program itself recognize many conditions (e.g. one system uses *lve* instead of Standard *Hànyǔ Pīnyīn*).

Làzhú also includes columns pertaining to Yale Transcription and Gwoyeu Romatzyh. Yale Transcription was developed by Yale University during World War II so that American soldiers could pronounce Mandarin more accurately. Gwoyeu Romatzyh was the official transliteration system of Taiwan until 1986. Neither Yale Transcription nor Gwoyeu Romatzyh is as popular as *Hànyǔ Pīnyīn* today, but they are used in some publications (particularly in the ROC).

Though only five more columns are included, the inclusion of these last two systems complicates the results of incorrect syllabic division. If these systems were added without the *Làzhú* algorithm, there would be more drastic changes between the transliterations of search terms. The search term `zhunian` would be converted into Yale as *`jwunyan` and (depending on the tones of each syllable) into Gwoyeu Romatzyh as:

1. *`juen_`,
2. *`juen_`,
3. *`juen_`,
4. *`juen_`,
5. *`jwen_`,
6. *`jwen_`,
7. *`jwen_`,
8. *`jwen_`,
9. *`joen_`,
10. *`joen_`,
11. *`joen_`,
12. *`joen_`,
13. *`juenn_`,
14. *`juenn_`,
15. *`juenn_`, and
16. *`juenn_`.

Làzhú obtains the correct combinations, which are listed below:

1. `junhian`,
2. `junian`,
3. `junean`,
4. `juniann`,
5. `jwunhian`,
6. `jwunian`,
7. `jwunean`,
8. `jwuniann`,
9. `juunhian`,
10. `juunian`,
11. `juunean`,
12. `juuniann`,
13. `juhnhian`,
14. `juhnian`,
15. `juhnean`, and

16. `juhniann`.

Implementation of Syllabic Division

Algorithm:

1. Let "word" be equal to searchterm. Proceed to step 2.
2. Let *i* be equal to 0. Proceed to step 3.
3. If the letter at position *i* is "n," proceed to step 4. Otherwise, proceed to step 12.
4. Is the letter before "n" a vowel? If so, proceed to step 5. Otherwise, proceed to step 11.
5. Is the letter after "n" also a vowel? If so, proceed to step 6. If the letter after the "n" is a "g," proceed instead to step 9. Otherwise, proceed to step 13.
6. Add the word upto position *i* to an array "divisions." Proceed to step 7.
7. Repeat steps 1-4 (and any other steps as needed, e.g. step 10). Proceed to step 8.
8. Let "word" be the rest of word (i.e. word after position *i*). Proceed to step 2.
9. Is the letter after "g" a vowel? If so, proceed to step 10. Otherwise, proceed to step 14.
10. Add the word upto position *i*+1 to an array "divisions." Proceed to step 11.
11. Repeat steps 1-4 (and any other steps as needed, e.g. step 5 or step 12). Proceed to step 8.
12. Increase *i* by one. If *i* is equal to the length of searchterm, stop. Otherwise, proceed to step 2.
13. If *i*+1 is equal to the length of word, add word to divisions. Proceed to step 16.
14. If *i*+2 is greater than or equal to the length of word, then proceed to step 15.
15. If the length of word is greater than 0, add word to divisions. Proceed to step 16.
16. Print divisions.

Shorter algorithm:

```
Function lazhu(word):
    For each position i in word:
        If there is an "n" or "r" in position i:
            If "n" or "r" is preceded by a vowel and followed by a vowel:
                Then break the word at position i
                Store the word upto position i in an array called "divisions"
                Call lazhu(remaining part of word)
                Return
            Else if "n" is preceded by a vowel and followed by "g":
                If "g" is followed by a vowel:
                    Then break the word at position i+1
                    Store the word upto position i+1 in divisions
                    Call lazhu(remaining part of word)
                    Return
    If the function did not return anywhere in the "for" loop:
        Store word in divisions
```

Return

Further Research

Include other transliteration systems. Figure out 邮政式拼音.
Maybe see whether this is applicable to other languages? Or at least, whether something like this can be done with other transliteration systems in Chinese?
Create 人人用, i.e. make 小纸 work so that a program can figure out the base transliteration system.

Conclusion

小纸 was effective, but it did not decompose syllables in *Hànyǔ Pīnyīn* as well as this method does. This method follows the rules of *Hànyǔ Pīnyīn* transliteration more precisely. Even more search terms can be found using this method than in 小纸. (Include statistic as a percentage to be more precise?).

Notes

[1]In fact, currently, *Hànyǔ Pīnyīn* is becoming increasingly popular internationally for transliterating Mandarin Chinese.

[2]Neither of the two letters in *Hànyǔ Pīnyīn* representing high vowels (*i* and *u*) may occur at the beginning of any syllable. The only possible exception to this rule is the syllable 唔, which may be transcribed *u* in some variants of *Hànyǔ Pīnyīn*. (In Cantonese, this syllable means “not” and is pronounced *m*, 4th tone. In Mandarin, it is most often seen as an equivalent of “Hmm…” in comic books.)