

Article-12 (2003 년 씀)
(필자의 승인 없는 인용을 금함)

AhnMaTae Phonetic Keyboard for Chinese Hanzi

A Proposal for Shorthand Input of Hanzi with Hangul (安馬太中文標音鍵盤速記輸入法)

Matthew Y. Ahn

Abstract

AhnMaTae Phonetic Keyboard (hereafter call it APK, 安馬太 標音 鍵盤) shorthand input system for Hanzi is a fast and easy to use. It uses phonetic values of Hangul (朝鮮族標音文字) to call characters, and uses characteristics of radical shapes to select characters in case of many candidate characters or phrases pop up.

This proposal paper is to show the readers how it can process the most difficult writing system in the world in incredible speed and accuracy and explain why it is better than any other system.

Introduction

Chinese Hanzi(漢字) is the oldest writing system in the world and one of the most difficult to learn and to write. It is difficult to write because there are too many strokes in one character. It is also very difficult to input in the computer due to the fact that it has too many characters to process. No one knows exactly how many Chinese characters are there, and likewise, no one knows exactly how many computer input systems are there as of now.

Introducing all the Chinese input systems would be beyond the scope of this paper, therefore, only introduce a few systems on the market. Two most popular ones are described in detail, namely: Ubi(五筆) method and Pinyin(拼音) method. Ubi is based on the shape of radicals of Hanzi (structural and visual method) and faster than Pinyin method. Pinyin is based on the phonetic values of English alphabet and very slow. Both methods require pushing of several keys serially and very difficult to learn how to operate it.

What is Hanzi(漢字)?

Exactly when and how Hanzi was created is not clear and still debated by scholars. Archeological evidence, which was discovered at the end of the nineteenth century in the area of modern Anyang area in the bed of the Yellow River, shows that some pictographic writings inscribed on tortoiseshells and on ox and sheep scapulas, indicates that it may be the first Hanzi. These inscriptions on shells and bones(甲骨文) were divinatory texts of Shang rulers. Shang Dynasty (殷商時代) lasted between about 1,600 B.C and 1,046 B.C., and therefore, Hanzi is at least 3,000 years old writing system. It is undoubtedly oldest writing system in the world and it became an essential part of one of the greatest cultures of humankind.

These old inscriptions closely resemble their modern Hanzi. Out of more than 2,500 characters of Shang period, more than 1,400 could be identified as the models of standard Hanzi characters of modern day.¹ For instance, Shang period's Character 'ma' (馬) in the shape of an animal with large head, four legs and a tail, looks not much different from modern day calligraphy. No one knows exactly how many characters were there in that period, but the number of characters steadily increased since then by combining basic pictographic characters. It can be close to 100,000 characters about now, including all the archaic characters. However, modern day usage may be limited to a few tens of thousands of characters.

The most widely used and up to date dictionary in mainland China now a day is known to be Xinhua Cidian (新華詞典, 2001年修訂版)², which has 15,184 single character words and 32,047 multiple character words, totaling 47,231 words. Of the 15,184 single character words, 7,000 of them are ordinary Hanzi, 623 are new Hanzi and 2,261 are archaic Hanzi. 47,231 words are displayed in the Pinyin alphabet orders, and by subcategorizing by syllabic characters with tone marks, it has 1,318 categories of A to Z.

The Ubi method input dictionary (五筆字型編碼手冊)³ lists about 7,300 single characters, mostly simplified Hanzi characters

¹ Florian Coulmas, *The Writing Systems of the World*, 1989, Blackwell Publishers, Oxford, p.93

² 新華通信社, *新華詞典*, 北京, 2001

³ 黃文壽, 紫日淡 編著, *五筆字型編碼手冊*, 2002, 暨南大學出版社, 廣州

displayed in Pinyin alphabet, which had 400 categories with alphabetical orders.

These 400 phonetic syllable of Chinese characters are much larger than the Japanese syllables of less than 100, but much smaller than 2,350 of modern Korean phonetic syllables.

Original pictographic characters meant something. For instance the character 'ma' (馬) meant horse originally. By adding 'Nyu' (女, meaning woman) to this 'ma'(馬), meant 'ma' which means mother, adding 'kou' (口, meaning mouth) to 'ma', it becomes adverb 'ma'(碼), which turns out to be an adverb for questioning. Some characters are stacked up with several of these root pictograms, which complicates in spelling, writing and pronouncing it. Some complicated characters are consisted of more than six-dozens of strokes.

In 1959, Premier Chou En-Lai proclaimed the reform program of Chinese script in the People's Republic of China, which is; (1) the standardization and simplification of Chinese characters, (2) the creation of Romanized orthography, and (3) the nationwide promotion of the standard language.⁴ Since then, this reform carried on significantly. Simplified Chinese characters are widely used in the mainland China. Pinyin is taught from the primary school levels and most of dictionaries are categorized by Pinyin system and 普通話(the Mandarin dialect) is used throughout this vast continent.

Chinese Pinyin (中國語併音)

Adoption of Pinyin was a historical necessity due to the complication of Chinese writing system. However, it created more problems, as Roman alphabet itself is not a good medium to scribe Chinese characters, than solving the problem.

1. Roman alphabet is not a good phonetic symbol.

⁴ Florian Coulmas, The Writing Systems of the World., p.245

Roman alphabet was developed from the Greek alphabet, which was originally adopted from pictographic Phoenician writing system. Name 'alphabet' came from the first two letters of Greek 'Alpha' and 'Beta'. Greek 'Alpha' was originally from Phoenicians first letter 'Alef', which meant 'ox'. It resembled tri-angled ox head with two horns on the head, directing side ways. When Greek adopted it (A) as their first letter, two horns directed toward ground. The second letter of Greek, 'Beta' (B) was adopted from the Phoenician's 'Beth' (𐤁), which meant a house and looked as a one room house, but Greek changed it to two room house (B), one for man and the other for woman. Original pictographic letters eventually became symbol of phonetics. However, English alphabet has so many phonetic values that even the native English speaking people have the difficulties of pronouncing some words.

Worst of all phonetics of English are vowels. 'A' can be pronounced e in the case of at, o in the case of ball, ei in the case of baby, a in the case of father. 'E' can be pronounced as e in the case of hen, i in the case of he, eo in the case of her. 'I' can be pronounced as i in the case of king, ai as in the case of identify. 'O' can be pronounced a in the case of lot, o in the case of old. 'U' can be pronounced as u in the case of use, oo as in the case of bull, a in the case of up. 'Y' can be pronounced e as in the case of yes, i in the case of system, ai in the case of psychology. Combination of these vowels can make all sorts of pronunciations, and there are no set rules how they are pronounced. It is a total chaos.

Also equally confusing are English consonants. 'C' can be pronounced s as in the case of city, k as in the case of cola. 'X' can be pronounced as ks in the case of taxi, s in the case of xylophone, 'G' can be pronounced as g in the case of girl, j as in the case of george. 'S' can be pronounced as s in the case of say, sh in the case of sure.

So confusing is the pronunciations of these alphabets that it requires phonetic symbols to pronounce it correctly. There is such phonetic symbol, called International Phonetic Alphabet (IPA). However, IPA is applicable only to Roman alphabets, and not suitable for Asian languages, such as Chinese, Korean and Japanese.

Adopting these Roman alphabet to Chinese was absolute necessity in that time, but it created lots of confusion as I described above. Let me give an example. To look into the Roman alphabet

phonetic value of 新華詞典, I typed 'sin hwa s dyan' which is a correct way of scribing it in English phonetics. Not a single Hanzi character showed up typing in that way. So I had to look into the dictionary. Surprisingly 新 is categorized under 'X', and 詞 was categorized under 'C'. S was not used at all. 華 was spelled hua and 典 was spelled dian.

Whether you are Ubi method user, or Pinyin method user, you have to have Hanzi dictionary, categorized by Pinyin. Unless you memorize all of the Chinese words how they were spelled them in Pinyin, you can not use computer for processing Hanzi.

Latin alphabet is taught from early ages of kindergarten and primary school levels.⁵ It has 23 vowels and 24 consonants.

Vowels(韻母) are; a, o, e, i, u, u(with umlaut), ai, ei, ui, ao, ou, iu, ie, u(with umlaut)i, er, an, en, in, un, u(with umlaut)n, ang, eng, ing, ong.

Consonants(聲母) are; b, p, b, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w.

Chinese use four (4) tone marks in addition to these forty seven (47) Pinyin characters,

They are: - mark, which has a flat tone,
/ mark, which is from down to up tone,
V mark, which is from up to down and up again tone,
| mark, which is from up to down tone,

Almost all modern day Chinese dictionaries use these Pinyin with tone marks as indexes of its characters(併音音節索引表), instead of numbers of strokes(部首檢字表) in old days. Indexes of stroke numbers are used only when you can not search the right character in the Pinyin indexes.

2. Chinese has too many homonyms.

Since Chinese has only about four hundred (400) phonetic syllables for several tens of thousands of characters, there are too many characters with same sound value, some have as many as a

⁵ See 幼兒啓蒙併音掛圖, 云南科技出版社, 2001 年, 1 版

few dozens of them, if not a few hundred. Therefore, Chinese word processor with Pinyin system requires selection process, out of so many candidates. Fortunately, some advanced processor displays candidate characters by the frequencies of usage to shorten the time to look up. However, less frequently used characters require lots of time to look up and select the right one. Usually selection is done by dragging the prompt and clicking on the right character or typing the right number displayed on that character.

If you do not know the Pinyin of particular character, you have to look up in the stroke number indexes and find the right Pinyin for that particular character. If you mis-typed or mis-spelled, you have to try other spelling. Therefore, this method is awfully slow and inefficient. Average speed of input by this method is less than thirty (30) Characters per minute.

There is another method of more efficient and faster input system. It is called Ubi shurufa (五筆輸入法) developed by Professor Wang, Yongmin(王永民). This method uses the characteristics of Character radicals and shapes of these radicals in simple term. Each character uses several strokes, mostly three or four strokes, according to the given rules. However, the problem of this method is that it requires long period of training, mostly some years, to be an efficient typist. Most of trained typist can input about one hundred (100) characters per minute.

Another problem is that it can use only about seven thousand (7,000) characters of modern use. It is far insufficient numbers even to type some rare names of people and name of places.

There are more than a dozen other input methods, but the end results are about the same. Some method uses three corner shapes, invented by Huang, Kidong(黃克東). Some uses other structures of Chinese character such as Changjie Shurufa (倉頂輸入法), developed by Zhu Bangfu(朱邦復), Zheng Code Method (鄭碼輸入法), developed by Zheng Yili (鄭易里)and Zheng Long(鄭龍). Some use combination of sound and structure of Chinese characters, such as Zilai Shurufa (子來輸入法), developed by Yang Zilai (楊子來) and Renzima Shurufa (認知碼輸入法). Some use Dianbaoma(電報碼), using telex codes.

More simplified Ubi method is called 101 Input Method (101 輸入法)⁶, which has only one shape of stroke on each key, instead of Ubi Method's several shapes of strokes assigned on each key. Twenty four (24) shapes in total and each character needs one to four strokes to extract one character. It may be easy to locate the keys by the stroke shapes, but you need to remember several thousand character codes to master this method.

Fairly recently, there is a more improved and more efficient system called Shinfangma (新方碼, 新方部首輸入法)⁷, invented by Zheng, Shinfang (鄭新芳), is on market. It is one of the most efficient and fast method of input system than any other method. It claims that it can input as many as 280 characters a minutes, which is more than ten times faster than Pinin system (拼音輸入法).⁸ It uses Shinwha dictionary (新華詞典) Pinyin indexing for classification of characters, but for inputting purpose, it uses Ubi method as well as English abbreviations such case as in USA=美國. Instead of extracting single characters, it extracts phrases. For example, to extract 中國, it types only four strokes of KSKW, instead of total of eight strokes. So is the 解放軍. Instead of each character's three or four strokes, it types only JFGC. In the case of 中華人民共和國, it requires only four strokes of KRPK, instead of three or four stroke each of seven characters. To reach the speed of 280 characters per minute, it requires several years of training to memorize all of these abbreviations .

Hangul (朝鮮語 文字)

Hangul was invented in 1443. Before it was finalized, King Sejong dispatched his court scholar to China to learn Chinese phonetics. It is not clear how many times these Korean scholars visited as King Sejong dispatched them in secrecy due to some of his court scholars were objecting to his project.

Hangul was invented by learning Chinese phonetics.

⁶ 中國廣東字原科技有限公司, 101 輸入法, 2002 年.12 月版

⁷ 張新芳著, 新方碼, 新方部首輸入法, 珠海出版社, 珠海, 2002 年 12 月

⁸ Ibid, 新方碼宣言書, No. 2.

Though Hangul was invented by King Sejong and his court appointed scholars in fifteenth century Korea (朝鮮), Chinese Phonetics (Yin Yun Xue, 音韻學) played a great influence on it.

Recently Professor Han, Tae Dong of Yonsei University of Seoul(漢城,延世大學校), discovered that the Tang Dynasty music has also played an important role in invention of Hangul.⁹ He has proved his theory by comparing the Chinese musical instrument of flute of Tang Period(唐代) and five basic consonants of Hangul. By measuring the five musical notes of flute and five basic vowels with sonogram, he discovered the frequencies of these five notes are identically the same.

Court Annals of Yi Dynasty in Korea(李朝實錄), in King Sejong's period, recorded that King Sejong dispatched thirteen visits of 集賢殿 scholars 成三問, 申瀾舟 to 遼東 Province of China to see the Chinese phonetic scholar 黃璨, before the 訓民正音 was completed. This is the clear evidence that Hangul was invented by learning Chinese 音韻學.

These Korean scholars not only learnt the basic Chinese phonetics, but also greatly improved from the Chinese 音韻學. Let me describe how well these scholars have improved from Chinese phonetics by some examples.

In Chinese phonetics, five (9) nasal sound vowels (鼻韻母), (an, en, in, un, win, ang, eng, ing and ong), are treated as vowels in Chinese phonetics, but in reality, it is combinations of vowels and consonants, 'n' and 'ng' sound, 'N' sound is created by tongue tip touching the upper frontal part of palate, while the sound comes through the nose, like English word, navy is pronounced, or ton is pronounced. 'N' sound is created by the same tongue tips whether 'N' comes before the vowels or after the vowels. 'NG' sound, like singing, comes from the throat when it passes through the nose. As the Chinese phonology does not use much of consonants after the vowels, except in this case of 'N' and 'NG' sounds, these nine vowels plus 'N' and 'NG' sounds were lumped together as vowels. Korean scholars completely separated these consonants, which comes after vowels. Korean scholars created consonants of

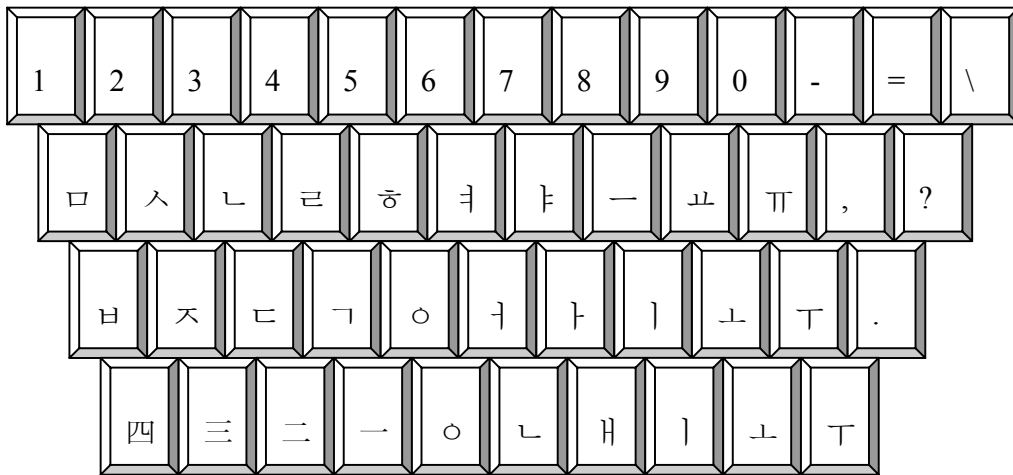
⁹ See Han Tae Dong's Phonetics of King Sejong Period 韓泰東 著, 世宗代の 音聲學, published by Yonsei University Press, 1998, Seoul. Chapter II, 훈민정음(訓民正音), pp. 27 -122.

distinctively different sound values which come after vowels, as following: g(ㄱ), n(ㄴ), d(ㄷ), r(ㄹ), m(ㅁ), b(ㅂ), s(ㅅ), tz(*), eu(ㅇ), ng(*), j(ㅈ), Ch(ㅊ), kh(ㅋ), t(ㅌ), P(ㅍ), H(ㅎ), and xh(*). (* in the parentheses are archaic Hangul consonants and no longer used in modern days.

Koreans use lots of these consonants after vowels and combination of them. These Hangul consonants after vowels are called Batchim(받침). It is also called Jongsung (終聲) as it comes at the tail of syllable. The total usage of these (終聲) is about nineteen (19) percent of total modern Hangul usage.

Since Chinese do not use much of the consonants after the vowels in the syllabic character, it is necessary to change the composition of these keys on APK as shown on the following key array.

AhnMaTae Phonetic Keyboard for Chinese (安馬太中文標音鍵盤)



(Patent No. US 6,462,678 B1)

Bottom left four numbers in Chinese are tone mark numbers and ㅇ(ng) and ㄴ(n) on the middle are consonants after vowels, and ㅈ(ae), ㅊ(ee), ㅏ(o) and ㅑ(oo) are the second vowels.

Instead of typing serially as in the case of Pinyin and Ubi systems, APK is typing simultaneously, which means pushing

several keys all at once. For instance, to type 鍵盤, only once for 鍵 (鍵) and tone mark of 四, and another stroke for 盤(盤) with tone mark of 二, with no space bar in between will solve the problem. In case, several candidates show up, the shape of the radicals with the shift key for the first character will chose the right one. If no candidate is shown, only two (2) strokes are needed. Otherwise, three (3) strokes are required.

In Pinyin system, you have to type 'jian' and type the right number of the candidates or click on the right character will choose one character, and then you have to type 'pan' and then do the same as the first one. In this case you need at least nine (9) strokes. Typing the right number out of several candidate characters, and dragging the prompt to the right character takes lots of time, sometimes requires ten times more than just typing one key as you have to move your hands and fingers in different locations on the keyboard.

In Ubi system, you have to type the right key by the shape of the radicals. In this case , 'QVGP' for 鍵, and since there is no such character as '盤' in Ubi system, you have to substitute it with 'TUL' which is similar looking. If it is the right one, then you have to push space bar as the choice of the right character in each step. This one takes at least nine (9) strokes.

Conclusion.

AhnMaTae Phonetic Keyboard for Chinese Hanzi will be much faster than any other system. It can type at least four hundred (400) characters per minutes, which is the same speed as an ordinary person speaks.

Reasons for such a high speed of input with APK, can be found at my web page (<http://ai.kaist.ac.kr/ahnmatae>): 'AhnMaTae Phonetic Hangul Keyboard' and 'Hangul Speedy Input for ECJ and other languages' are written in English. If you can read Hangul and understand, there are more than half a dozen articles written by me in that web page. I am an invited researcher at the Artificial Intelligence Research Lab. of the KAIST (Korean Advanced Insitute of Science and Technology).

This new input system for Chinese hanzi will be available sometime in 2004.

It would be much better to adopt its own writing system, Hangul, which is the writing system of one of the ethnic groups in China (朝鮮族). Hangul (朝鮮族文字) is one of the Chinese Standard Coded Character Set GB 12052-89 (中國標準文字符號 GB 12052-89). GB means 國家標準 (*Guojia Biaozhun*) in Peoples' Republic of China (中國人民共和國). Moreover, Hangul is the perfect phonetic symbol for every language in the world. Computer word processing guru, particularly of Easter Asian Languages, Mr. Ken Lunde describes Hangul as 'one of the most scientific writing systems due to its extremely regular and predictable structure.'¹⁰

Before APK was introduced in Korea, no one believed that it could process Korean language in lightning speed of about 1,500 strokes in a minute (about 500 syllabic characters of Hangul) without using a separate shorthand machine. Likewise, until it will be introduced in China in 2004, no one would believe it can process so easily and so speedily.

References:

- Matthew Y. Ahn, AhnMaTae Phonetic Hangul Keyboard, 2000, KAIST
Matthew Y. Ahn, Hangul Speedy Input for ECJ and other languages, 2000, KAIST
Florian Coulmas, The Writing Systems of the World, 1989, Blackwell Publishers
Kim-Renaud, The Korean Alphabet, Its History and Structure, Hawaii Univ., 1997
Ken Lunde, CJKV Information Processing, O'Reilly, 1999
新華通信社, 新華詞典, 北京, 2001
黃文壽, 紫日淡 編著, 五筆字型編碼手冊, 2002, 南大學出版社, 廣州
幼兒啟夢併音卦圖, 云南科技出版社, 2001, 1 版
中國廣東字原有限公司, 101 輸入法, 2002 年 12 月版
張新芳, 新方碼 打字王, 珠海出版社, 珠海
韓泰東, 世宗代の 音聲學, Yonsei Univ., Press, 1998, Seoul
안마태, 한글 통일 글자판 시안, 1985 (<http://ai.kaist.ac.kr/ahnmatae>)
안마태, 안마태 소리글판의 제안, 1997 (<http://ai.kaist.ac.kr/ahnmatae>)
안마태, 초고속 입력자판 개발과 한글의 세계화, 2001 (<http://ai.kaist.ac.kr/ahnmatae>)
안마태, 컴퓨터에서의 과학적 한글 사용 방법, 2003 (<http://ai.kaist.ac.kr/ahnmatae>)
The Unicode Consortium, Unicode Standard Version, 3.0, Addison-Wesley

¹⁰ Lunde, Ken, CJKV Information Processing, O'Reilly, USA, 1999, p. 48.