THE UNIFIED PHONETIC TRANSCRIPTION FOR TEACHING AND LEARNING CHINESE LANGUAGES

Jiann-Cherng Shieh
Graduate Institute of Library and Information Studies
National Taiwan Normal University, Taiwan
jcshieh@ntnu.edu.tw

ABSTRACT
In order to preserve distinctive cultures, people anxiously figure out writing systems of their languages as recording tools. Mandarin, Taiwanese and Hakka languages are three major and the most popular dialects of Han languages spoken in Chinese society. Their writing systems are all in Han characters. Various and independent phonetic transcriptions have been thus developed to be as the mapping mechanisms between Chinese mother tongue languages and Han characters. For teaching and learning facilitation purposes, we really require a convenient phonetic transcription system between daily Mandarin, Taiwanese and Hakka to speed Han characters data processing applications. The Roman spelling system is a universal tool that owns the one and only one spelling rule. By studying and analyzing the Roman spelling system, we have disclosed that 4135 Romanized phonetic transcriptions can be adequately applied to handle Han characters' mappings of Mandarin, Taiwanese and Hakka spoken dialects. In this paper, we propose a minimal perfect hashing function to process unified 4135 Mandarin, Taiwanese and Hakka Romanized phonetic transcriptions to their corresponding Han characters simultaneously. The unified phonetic transcription can be used to promote Chinese mother tongue languages applications and developments. Furthermore, it can be applied as a mechanism to popularize digital learning and teaching of Chinese mother tongue languages.

INTRODUCTION

People generally recognize that it is valuable to teach and learn mother tongue languages in today societies. People anxiously figure out writing systems of their languages to record and preserve their distinctive cultures. Mandarin, Taiwanese and Hakka languages are the three major spoken dialects in Chinese society. There are many speakers of the languages in China, Malaysia, Singapore, Philippine, Thailand and Indonesia.

Mandarin is the widest spoken language in the world and there are about 1300 millions people worldwide. Pinyin, more formally Hanyu Pinyin, is the most common Standard Mandarin Romanization system in use. Hanyu means the Chinese language. pin means "together, connection, annotate" and yin means "sound". Pinyin uses the Latin alphabet to represent sounds in Standard Mandarin. Taiwan has adopted Tongyong Pinyin on the national level since October 2002. Tongyong Pinyin is a modified version of Hanyu Pinyin. Based on the Chinese remainder theorem, Chang and Wu (Chang & Wu, 1988) designed the hashing function to process 1303 distinct Mandarin phonetic transcriptions of Han characters.

Minnanyu refers to a family of Chinese languages which are spoken in southern Fujian and neighboring areas, and by descendants of emigrants from these areas in diasporas. It is usually called Taiwanese by residents of Taiwan, and Hokkien by residents of Southeast Asia. Taiwanese can be written with the Latin alphabet using a Romanized orthography which was developed first by Presbyterian missionaries in China and later by the indigenous Presbyterian Church in Taiwan; use of the orthography has been actively promoted since the late 19th century. Taiwanese is one of the most used dialects spoken in Taiwan, and evolved from the ancient languages of China, the Ho-Lo language family. According to the traditional but representative and authoritative Taiwanese dictionary (Shen, 2001), Shieh (2003) developed the hashing function of 3028 Taiwanese phonetic transcriptions of Han characters.

Hakka dialect is one of the seven major spoken dialects in Chinese Society. The Hakka language has numerous dialects spoken in southern provinces of China, Taiwan, Singapore, Philippine and Indonesia. It is the 32nd widest spoken language in the world and there are about 100 millions Hakka speakers worldwide. Hakka is not mutually intelligible with Mandarin, Cantonese, Minnan and most of the significant spoken variants of the Chinese language. The Hakka dialects across various China provinces differ phonologically, but the Meixian dialect of Hakka is considered the archetypal spoken form of the language. Shieh and Hsu (Shieh & Hsu, 2007) proposed a minimal perfect hashing function for the 1428 Hakka phonetic transcriptions of Han characters from authoritative Meixian Hakka dialect dictionary (Lee, 1995).

Many researchers have enthusiastically endeavored to study related the spoken languages subjects such as language curriculums in multicultural society (Kilimci, 2010), typologies of spoken language learning aids (Kartal, 2005), mappings between spoken language and its writing system (Chang & Wu, 1988; Shieh, 2003;

Shieh & Hsu, 2007) (as depicted in Figure 1), etc. They are striving to protect and promote their individual native cultures, and make them widespread utilization.

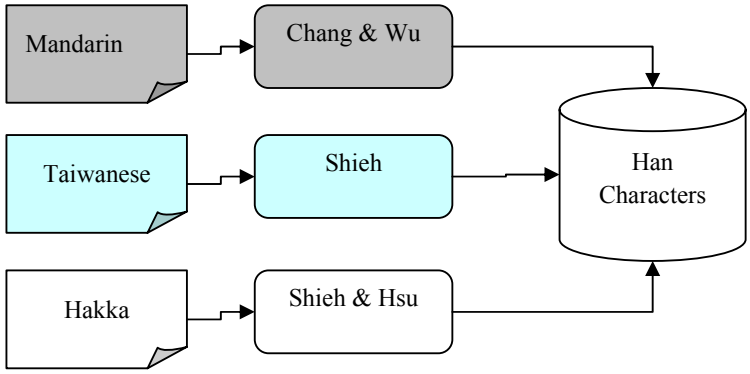


Figure 1: Various and Independent Han characters mappings for Chinese Languages

In Chinese societies, for teaching and learning facilitation purposes, we really require a convenient phonetic transcription system between daily Mandarin, Taiwanese and Hakka to speed Han characters data processing applications. These spoken languages are all with their respective Romanized phonetic transcriptions. Pleasantly surprised, the Roman spelling system is a universal tool that owns the one and only one spelling rule and can be generally and simultaneously applied to different languages applications. By studying and analyzing the Roman spelling system, we have disclosed that 4135 Romanized phonetic transcriptions can be adequately applied to handle Han characters' mappings of Mandarin, Taiwanese and Hakka spoken dialects. The 4135 integrated phonetic transcriptions are composed of 7 tones, 29 consonants, and 120 vowels at most. For language application purposes, it is much important for us to establish a mechanism to efficiently retrieve different Han characters and their corresponding pronunciations from its vocabulary repository, as illustrated in Figure 2. Many Chinese language learning applications, such as on-line or mobile dictionaries, translations, text-to-speech conversions, e-books, etc., can be further developed to help learners and teachers.

In this paper, we apply the Chinese remainder theorem to design a fast and efficient hashing function (Knuth, 1998) to map the unified 4135 phonetic transcriptions to corresponding Han characters of Mandarin, Taiwanese and Hakka languages. We also give a proof that the loading factor is more than 0.887, which is the best one when applying the Chinese remainder theorem to the design of hashing functions for the word sets.

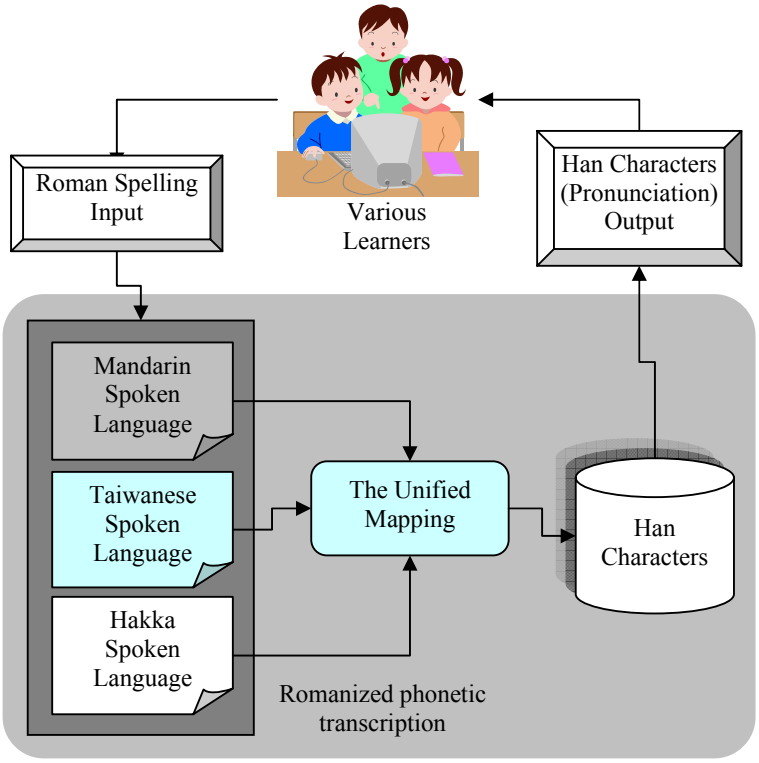


Figure 2: The Unified Mapping for Chinese Languages

Hashing Functions Based on the Chinese Remainder Theorem

In this section, we first introduce the Chinese remainder theorem and its application to hashing function designs of character data sets. Then we review the hashing function designs of Mandarin, Taiwanese and Hakka phonetic transcriptions based on the theorem.

The Chinese remainder theorem (Chang & Lee, 1986)

Theorem 1. Let $r_1, r_2, ..., r_n$, be n integers. There exists an integer C such that $C=r_1 \pmod{m_1}$, $C=r_2 \pmod{m_2}$, ..., $C=r_n \pmod{m_n}$, if m_i and m_j are relatively prime to each other for all $i \neq j$.

For example, let $r_1=1$, $r_2=2$, $r_3=3$, $r_4=4$ and $m_1=4$, $m_2=5$, $m_3=7$, $m_4=9$. Here m_i and m_j are relatively prime for $i \neq j$, $1< i, j< 4$. By the Chinese remainder theorem, there exists an integer C=157 such that $C \bmod m_1=157 \bmod 4=1=r_1$, $C \bmod m_2=157 \bmod 5=2=r_2$, $C \bmod m_3=157 \bmod 7=3=r_3$, $C \bmod m_4=157 \bmod 9=4=r_4$.

The following theorem results easily from the Chinese remainder theorem.

Theorem 2. Given a finite integer key set $K=\{L_1, L_2, ..., L_n\}$. If L_i and L_j are relatively prime to each other for all $i \neq j$, there exists a constant C such that $h(L_i)=C \bmod L_i$ is a minimal perfect hashing function (Chang & Lee, 1986).

Hashing scheme based on the Chinese remainder theorem

Based on the Chinese remainder theorem, Chang and Lee (1986) proposed a letter-oriented minimal perfect hashing scheme for a set of words. For a finite word set $K=\{L_1, L_2, ..., L_n\}$, it is heuristically assumed that there exist s_1 and s_2 such that the extracted letter pairs (L_{i1}, L_{i2}) are distinct, where L_{i1} and L_{i2} are the s_1-th and s_2-th characters of the word L_i, i= 1, 2, ..., n. Chang and Lee's hashing function is defined as $h(L_i) = H(L_{i1}, k_{i2}) = d(L_{i1}) + C(L_{i1}) \bmod p(L_{i2})$, where d and C are integer value functions, and p is a prime number function. Chang and Lee's applied the hashing scheme to 12 months and 9 major planets with 0.154 and 0.103 loading factors respectively.

When applying the Chinese remainder theorem to the design of letter-oriented minimal perfect hashing functions, we often encounter the intractable issue of extracting letters from the word sets to form distinct letter pairs, especially from large data sets. Chang and Shieh (1985) used a zero value rehash index to resolve the problem. They successfully applied the technique to rehash the 59 reserved words for data-flow language VAL,

the 65 Z-80 commands, and the 256 frequently used words. Furthermore, Chang and Wu (1988) utilized the characteristics of Mandarin phonetic symbols to cluster the word set and then produced 1303 distinct letter pairs. The hashing scheme is introduced in the next section.

Mandarin phonetic symbols hashing scheme (Chang & Wu, 1988)
Chinese characters are constructed by 37 Mandarin phonetic symbols accompanied by one of the five tones. There are a total of 1303 distinct Mandarin phonetic transcriptions of Chinese characters. The phonetic symbols are divided into three categories: (1) the consonant, (2) the first vowel, and (3) the second vowel. For each symbol x in the symbol set, we have its order O(x). In the hashing scheme, Chang and Wu translate all the phonetic transcriptions to letter pairs of two phonetic symbols.

Chang and Wu (1988) then cluster all letter pairs according to the five tones. In each equal-tone cluster, letter pairs with the same leading character are further grouped together. We see that the maximum number of character pairs in one group might go up to 33. From the experiment, as applying the Chinese remainder theorem, this would make the constant C quite large. By dividing the character pairs into three sets, they thus can assign the least 11 prime numbers to the corresponding characters in each group of the three sets. The minimal perfect hashing function is defined as $H_j(L_{i1},L_{i2})=d_{jk}(L_{i1}) + C_{jk}(L_{i1}) \bmod p(L_{i2})$, where d_{jk} and C_{jk} are integer value functions of each L_{i1} in the k-th set of each j-tone cluster, and p is a prime number function of each L_{i2}. The total size of space used is 38*(3*(5*2+1)+3)+1303=2671, where 38 stands for 37 phonetic symbols and 1 dummy symbol; 3*(5*2+1) stands for d_{jk} and C_{jk} of 5 clusters and index k in three sets. The number 3 is for the functions O, p, and W. Thus, the loading factor is about 0.4878. If only the contiguous space is considered, the size of the space that is used becomes 38*3*14+1303= 2899; the loading factor is about 0.45.

Taiwanese phonetic transcriptions hashing scheme (Shieh, 2003)
The Taiwanese phonetic transcription system, referred to a traditional but representative and authoritative Taiwanese dictionary (Shen, 2001), is composed of 7 tones (Table 1), 15 consonants (Table 2), and 45 vowels (Table 3). Each Taiwanese phonetic transcription consists of a vowel, a consonant, and a tone. Theoretically there are a total of 4725 transcriptions. However, only 3028 of the transcriptions are associated with Han characters. Shieh (2003) takes these 3028 transcriptions as study word set.

Table 1: Taiwanese Seven Tones

Code k_{i1}	1	2	3	4	5	6	7
Tone	kun	kún	kùn	kut	kûn	kūn	kut̍

Table 2: Taiwanese Fifteen Consonants

Code k_{i2}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Consonant	liú	pinn	kiû	khì	tē	phó	thann	tsan	ji̍p	sî	ing	mn̂g	gí	tshut	hí
Assigned Prime $P(k_{i2})$	2	3	5	7	11	13	17	19	23	29	31	37	41	43	47

Table 3: Taiwanese Forty-five Vowels

Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel
1	kun	10	kuan	19	kam	28	ka	37	kong
2	kian	11	ko	20	kue	29	ki	38	m̀
3	kim	12	kiau	21	kang	30	kiu	39	muê
4	kui	13	ki	22	kiam	31	kenn	40	îng
5	ka	14	kiong	23	kau	32	kng	41	kiaunn
6	kan	15	kau	24	khia	33	kiô	42	tsim
7	kong	16	kai	25	kuè	34	kiunn	43	ngâu
8	kuai	17	kin	26	kam	35	kuan	44	kiann
9	king	18	khiong	27	ku	36	koo	45	kuan

Shieh handled 3028 distinct letter pairs of (k_{i1}, k_{i2}, k_{i3})'s, each with k_{i1} tones, k_{i2} consonants, and k_{i3} vowels. He sorted these letter pairs by their lexical orderings and then assigned each (k_{i1}, k_{i2}, k_{i3}) a unique address. According to k_{i1}, Shieh got seven groups and computed their starting addresses $d(k_{i1})$. For each group k_{i1}, based on 15 consonants, he produced 15 tone/consonant subgroups and computed their corresponding relative

subgroup starting addresses $d_{ki1}(k_{i2})$. For each subgroup, there are at most 45 letter pairs. He clustered the subgroup into 5 bunches by $b(k_{i3})$ and also calculated each relative starting address $d_{ki1,ki2}(b(k_{i3}))$, where each k_{i1} is associated with $b(k_{i3})$. Then he sequentially assigned the least 9 prime numbers $P(k_{i3})$'s to k_{i3} cyclically in each tone/consonant/vowel cluster. Finally, for every cluster, he applied the Chinese remainder theorem to compute constant $C_{ki1,ki2}(b(k_{i3}))$ such that $C_{ki1,ki2}(b(k_{i3}))$ mod $P(k_{i3})$ equals the relative address of the cluster. The corresponding minimal perfect hashing function is defined as $H(k_{i1}, k_{i2}, k_{i3}) = d(k_{i1}) + d_{ki1}(k_{i2}) + d_{ki1,ki2}(b(k_{i3})) + C_{ki1,ki2}(b(k_{i3}))$ mod $P(k_{i3})$. Totally, it takes 4235 spaces: 3028 for key words, 7 $d(k_{i1})$'s, 7*15 = 105 $d_{ki1}(k_{i2})$, 7*15*5 = 525 $d_{ki1,ki2}(b(k_{i3}))$'s, 7*15*5 = 525 $C_{ki1,ki2}(b(k_{i3}))$'s and 45 $P(k_{i3})$'s. The loading factor is 3028/4235 = 0.715.

Hakka phonetic transcriptions hashing scheme (Shieh & Hsu, 2007)
According the selected Meixian Hakka dialect dictionary, the Hakka phonetic transcription system is composed of 6 tones (Table 4), 17consonants (Table 5), and 72 vowels (Table 6). Each Hakka phonetic transcription consists of a tone, a consonant, and a vowel. However, only 1428 of the transcriptions are associated with Han characters. Shieh and Hsu took these 1428 transcriptions as study word set.

Table 4: Hakka Six Tones

Tone	Yin Ping	Yang Ping	Shang	Qu	Yin Ru	Yang Ru
Code k_{i1}	1	2	3	4	5	6

Table 5: Hakka Seventeen Consonants

Consonant	p	p'	m	f	v	t	t'	n	l	ts	ts'	s	k	k'	ŋ	h	Ø
Code k_{i2}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Table 6: Seventy-three Vowels

Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel	Code k_{i3}	Vowel
1	i	13	eu	25	ok	37	e	49	ip	61	uo
2	au	14	uŋ	26	at	38	iap	50	əm	62	n
3	u	15	on	27	uk	39	iun	51	iai	63	uaŋ
4	oŋ	16	oi	28	ap	40	ot	52	ət	64	ep
5	a	17	en	29	it	41	iuk	53	ua	65	uat
6	o	18	iam	30	et	42	ŋ	54	uai	66	uet
7	ai	19	ui	31	ak	43	iet	55	uan	67	iut
8	un	20	aŋ	32	im	44	iak	56	ion	68	m
9	iau	21	iaŋ	33	iuŋ	45	iok	57	əp	69	iui
10	an	22	ien	34	ian	46	em	58	io	70	uen
11	in	23	ioŋ	35	ut	47	ən	59	uon	71	uak
12	am	24	iu	36	ia	48	iat	60	uoŋ	72	uok

They handled 1428 distinct letter pairs of (k_{i1}, k_{i2}, k_{i3})'s, each with k_{i1} tone, k_{i2} consonant, and k_{i3} vowel. Shieh and Hsu sorted these letter pairs by their lexical orderings and then assigned each (k_{i1}, k_{i2}, k_{i3}) a unique address. According to (k_{i1}, k_{i2}), they had 6*17 groups and compute their starting addresses $d(k_{i1}, k_{i2})$'s. Then, they assigned appropriate prime numbers $P(k_{i3})$'s for k_{i3}. Finally, for every group, they applied the Chinese remainder theorem to compute constant $C(k_{i1}, k_{i2})$ such that $C(k_{i1}, k_{i2})$ mod $P(k_{i3})$ equals the relative address of character pair (k_{i1}, k_{i2}, k_{i3}) in group headed with (k_{i1}, k_{i2}). The corresponding minimal perfect hashing function is defined as $H(k_{i1}, k_{i2}, k_{i3}) = d(k_{i1}, k_{i2}) + C(k_{i1}, k_{i2})$ mod $P(k_{i3})$. It takes 1704 spaces: 1428 key words, 6*17 $C(k_{i1}, k_{i2})$'s, 6*17 $d(k_{i1}, k_{i2})$'s, and 72 $P(k_{i3})$'s. The loading factor is 1428/1704=0.838.

The Unified Phonetic Transcription Design
Hashing Function Design
The unified Mandarin, Taiwanese and Hakka Romanized phonetic transcription is composed of 7 tones (Table 7), 29 consonants (Table 8), and 120 vowels (Table 9) associated with a prime number $P(k_{i3})$. Each phonetic transcription (k_{i1}, k_{i2}, k_{i3}) consists of a tone ki1, a consonant k_{i2}, and a vowel k_{i3}. There are totally 24360 combinations of (k_{i1}, k_{i2}, k_{i3})'s. According to our further analysis, we worked out that we can use exactly 4135 phonetic transcriptions to associate their corresponding Han characters.

Table 7: Tones

Tone	1	2	3	4	5	7	8

Code k_{i1}	1	2	3	4	5	6	7

Table 8: 29 Consonants

Consonants	b	c	ch	chi	d	f	g	h	j	ji	k	kh	l	m	n
Code k_{i2}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Consonants	ng	p	ph	r	s	sh	shi	t	th	ts	tsh	tz	v	null	
Code k_{i2}	16	17	18	19	20	21	22	23	24	25	26	27	28	29	

Table 9: 120 Vowels

Vowel	Code k_{i3}	$P(k_{i3})$	Vowel	Code k_{i3}	$P(k_{i3})$	Vowel	Code k_{i3}	$P(k_{i3})$
a	1	2	iat	41	2	mh	81	2
ah	2	3	iau	42	3	ng	82	3
ai	3	5	iauh	43	5	ngh	83	5
ainn	4	7	iaunn	44	7	o	84	7
ak	5	11	ie	45	11	oh	85	11
am	6	13	iem	46	13	oi	86	13
an	7	17	ien	47	17	ok	87	17
ang	8	19	iet	48	19	on	88	19
ann	9	23	ieu	49	23	ong	89	23
annh	10	29	ih	50	29	onn	90	29
ap	11	31	ii	51	31	onnh	91	31
at	12	37	iim	52	37	oo	92	37
au	13	41	iin	53	41	ot	93	41
auh	14	43	iip	54	43	ou	94	43
aunn	15	47	iit	55	47	u	95	47
aunnh	16	53	ik	56	53	ua	96	53
e	17	59	im	57	59	uah	97	59
eh	18	61	in	58	61	uai	98	61
ei	19	67	ing	59	67	uainn	99	67
em	20	71	inn	60	71	uan	100	71
en	21	73	innh	61	73	uang	101	73
eng	22	79	io	62	79	uann	102	79
enn	23	83	ioh	63	83	uat	103	83
ennh	24	89	iok	64	89	ue	104	89
ep	25	97	ion	65	97	ueh	105	97
er	26	101	iong	66	101	uei	106	101
et	27	103	iou	67	103	uen	107	103
eu	28	107	ip	68	107	uenn	108	107
i	29	109	it	69	109	uet	109	109
ia	30	113	iu	70	113	uh	110	113
iah	31	127	iuan	71	127	ui	111	127
iai	32	131	iue	72	131	uih	112	131
iak	33	137	iuh	73	137	uinn	113	137
iam	34	139	iui	74	139	uk	114	139
ian	35	149	iuk	75	149	un	115	149
iang	36	151	iun	76	151	ung	116	151
iann	37	157	iung	77	157	uo	117	157
iannh	38	163	iunn	78	163	ut	118	163
iaong	39	167	iut	79	167	yu	119	167
iap	40	173	m	80	173	Null	120	173

We handle 4135 distinct letter pairs of (k_{i1}, k_{i2}, k_{i3})'s, each with k_{i1} tone, k_{i2} consonant, and k_{i3} vowel. We sort these letter pairs by their lexical orderings and then assign each (k_{i1}, k_{i2}, k_{i3}) a unique address. According to (k_{i1}, k_{i2}), we have 7*29 groups and compute their starting addresses d(k_{i1}, k_{i2})'s. There are at most 120 characters k_{i3} in each group (k_{i1}, k_{i2}). Then, we heuristically assign appropriate prime numbers $P(k_{i3})$'s for k_{i3}. Finally, for every group, we apply the Chinese remainder theorem to compute constant C(k_{i1}, k_{i2}) such that C(k_{i1}, k_{i2}) mod

$P(k_{i3})$ equals the relative address of character pair (k_{i1}, k_{i2}, k_{i3}) in group headed with (k_{i1}, k_{i2}). The corresponding minimal perfect hashing function is defined as $H(k_{i1}, k_{i2}, k_{i3}) = d(k_{i1}, k_{i2}) + C(k_{i1}, k_{i2}) \bmod P(k_{i3})$.

The hashing function design of the unified phonetic transcription is summarized as follows:
Step 1: Using tone k_{i1}, consonant k_{i2} and vowel k_{i3} , we can have 4135 distinct letter pairs (k_{i1}, k_{i2}, k_{i3}). We sort them in their lexical orders and assign each a unique address.
Step 2: We allocate each (k_{i1}, k_{i2}) group a $d(k_{i1}, k_{i2})$, the first address of the letter pairs headed with (k_{i1}, k_{i2})'s.
Step 3: Associated with each group (k_{i1}, k_{i2}), we assign each (k_{i1}, k_{i2}, k_{i3}) a relative address.
Step 4: We assign appropriately the prime numbers $P(k_{i3})$'s to k_{i3}.
Step 5: Consider the letter pairs (k_{i1}, k_{i2}, k_r), $1 \le r \le n$, with the same (k_{i1}, k_{i2}), that is they are in the same group (k_{i1}, k_{i2}), and the corresponding assigned prime numbers of k_r's are $P_1, P_2, \ldots, P_n$, where $P_1 < P_2 < \ldots < P_n$. Applying the Chinese remainder theorem to find a constant $C(k_{i1}, k_{i2})$ such that $C(k_{i1}, k_{i2}) \equiv 1 \bmod P_1$, $C(k_{i1}, k_{i2}) \equiv 2 \bmod P_2, \ldots$, and $C(k_{i1}, k_{i2}) \equiv n \bmod P_n$. Our proposed minimal perfect hashing function is simply defined as $H(k_{i1}, k_{i2}, k_r) = d(k_{i1}, k_{i2}) + C(k_{i1}, k_{i2}) \bmod P(k_r)$. The values of all $C(k_{i1}, k_{i2})$'s are illustrated in the Appendix.

Loading Factor Comparisons

Loading factor is used to measure the efficiency of memory usage in hashing design. It is defined as a ration of the number of data and the total size of memory used. The loading factor of the hashing function designed in this paper is derived as follows: (1) Used spaces, 4135 key words, 7*29 $C(k_{i1}, k_{i2})$'s, 7*29 $d(k_{i1}, k_{i2})$'s, and 120 $P(k_{i3})$'s. We take 4661 spaces in total. (2) The loading factor is 4135/4661=0.887. The following table (Table 10) shows the loading factors of various minimal perfect hashing functions designed for diverse word sets by the Chinese remainder theorem. Obviously, it can be shown that our hashing function is superior to others.

Table 10: Comparisons of Loading Factors

Word Sets	Loading Factor	
Names of 12 months	0.154	Chang & Lee, 1986
59 VAL reserved words	0.312	Chang & Shieh, 1985
65 Z-80 commands	0.263	Chang & Shieh, 1985
256 frequently used words	0.472	Chang & Shieh, 1985
1303 Mandarin phonetic transcriptions	0.448	Chang & Wu, 1988
3028 Taiwanese phonetic transcriptions	0.715	Shieh, 2003
1428 Hakka dialect phonetic transcriptions	0.838	Shieh & Hsu, 2007
4135 unified Mandarin, Taiwanese and Hakka phonetic transcription	0.887	This paper

Number C Analysis

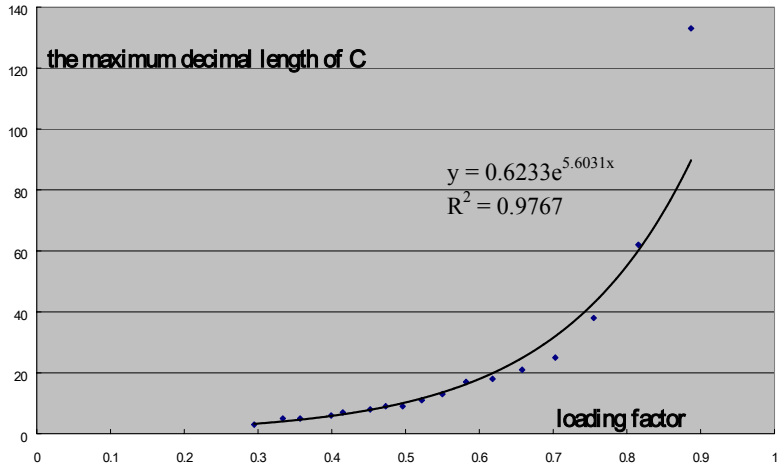
The numbers C's are the most intractable ones as applying the Chinese remainder theorem to design hashing functions for data sets. On observing variations of C's resulted by the experimental designs for the unified Romanized phonetic transcriptions, we have concluded that the size of C is dependent on the number of associated primes that we have used in each vowel group. In fact, during the hashing design, we can group k_{i3}'s vowels for each (k_{i1}, k_{i2}) in different sizes to have alternate C's results. The smaller size each vowel group has, the smaller constant C we result. However, what we should pay for is loading factor. There is a tradeoff between the size of constant C and the loading factor. The following table (Table 11) shows the experiments.

Table 11: The C's and Loading Factors of Different Groups

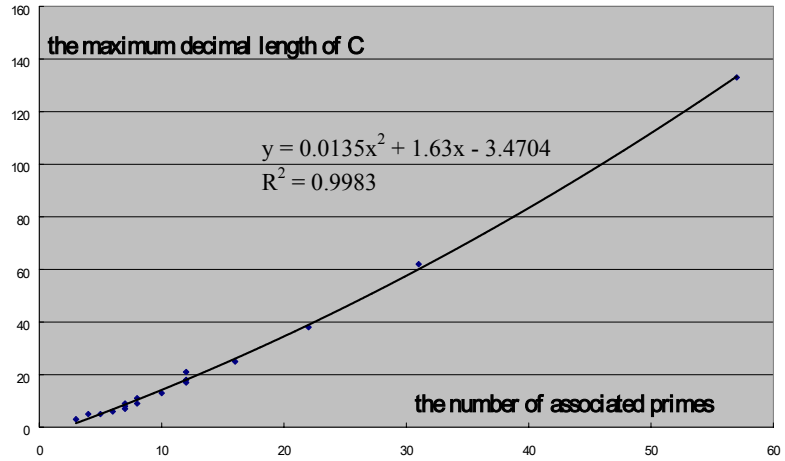
Number of vowel groups	Number of associated primes	Loading factors	The maximum length of C's
1	57	0.887	133
2	31	0.816	62
3	22	0.755	38
4	16	0.703	25
5	12	0.658	21
6	12	0.618	18
7	12	0.582	17
8	10	0.550	13
9	8	0.522	11
10	7	0.496	9
11	8	0.473	9
12	7	0.452	8
14	7	0.415	7
15	6	0.399	6

18	4	0.357	5
20	5	0.334	5
24	3	0.295	3

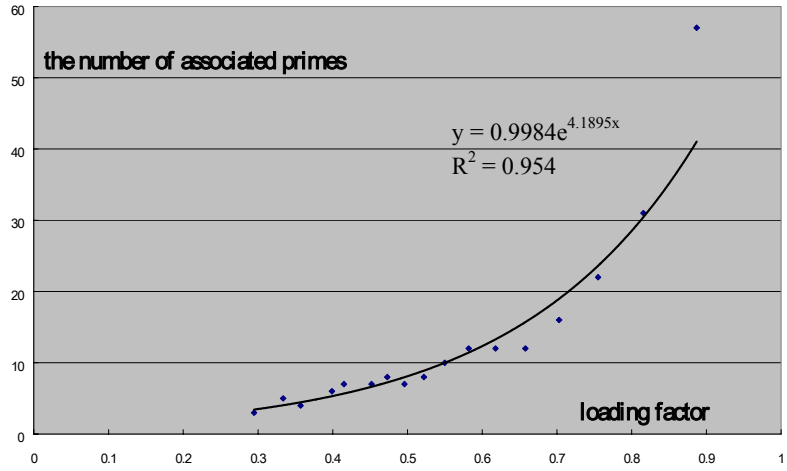
Next, we apply the statistical regression analysis to the experimental data to profile the correlations between the above parameters: loading factor vs. the decimal length of constant C, the number of associated primes vs. the decimal length of C and loading factor vs. the number of associated primes. The results are shown in the following Figure 3, where R^2 is the coefficient of determination and its value is between 0 and 1.



(a) Loading factor and the maximum decimal length of C



(b) The number of associated primes and the maximum decimal length of C



(c) Loading factor and the number of associated primes

Figure 3: The Regression Analysis of Effective Factors .

What the preferred situation is to have high loading factor with short decimal length of C. However, from the regression analysis, we have the fact that there is a tradeoff between two parameters. This will give us the concrete suggestion as we apply the unified phonetic transcription on diverse learning and teaching devices.

CONCLUSIONS

With the unified phonetic transcription system for mapping Mandarin, Taiwanese and Hakka mother tongue languages to their Han characters, people will be convenient to promote their learning and teaching activities, as well as to record and preserve their particular cultures. In this research, we have successfully applied the Chinese remainder theorem to design a novel minimal perfect hashing function for 4135 integrated Mandarin, Taiwanese and Hakka Romanized phonetic transcriptions of Han characters. We have achieved significant results in terms of loading factors. We further give experimental investigation and mathematical regression analysis for considered factors of hashing effectiveness. We get the conclusion that the size of number C is dependent on the number of associated primes. For the unified Romanized phonetic transcriptions case, we propose the grouping technique to promote the effective applications as concerning the practicability of accessing constants C's. However, we have explored that there is a tradeoff between the loading factor and the size of C. The unified phonetic transcription can be used to promote Chinese mother tongue languages applications and can be applied as a tool to popularize digital learning of the languages further.

REFERENCES

Chang, C. C., & Lee, R. C. T. (1986). A letter oriented minimal perfect hashing scheme. *Computer Journal, 29*, 277-281.

Chang, C. C., & Shieh, J. C. (1985). On the design of letter oriented minimal perfect hashing functions. *Journal of Chinese Institute Engineers, 8,* 285-297.

Chang, C. C., & Wu, H. J. (1988). A fast Chinese characters accessing technique using mandarin phonetic transcriptions. *International Journal of Pattern Recognition and Artificial Intelligence, 2,* 105-137.

Kartal, Erdoğan. (2005). The internet and autonomous language learning: A typology of suggested aids. *The Turkish Online Journal of Educational Technology, 4*(4), 54-58.

Kilimci, Songül. (2010). Integration of the internet into a language curriculum in a multicultural society. *The Turkish Online Journal of Educational Technology, 9*(1), 107-113.

Knuth, D. E. (1998). *The Art of Computer Programming III: Sorting and Searching 2nd ed.,* MA: Addison-Wesley, Reading.

Lee, R. (1995). *Meixian Dialect Dictionary,* Jiangsu Education Publish.

Shen, J. F. (2001). *Hui Im Po Kam. 46th ed.* Taiwan: Wen-I Publish.

Shieh, J. C. (2003). An efficient accessing technique for Taiwanese phonetic transcriptions. *ACM Transaction on Asia Language Information Processing, 2*(1), 63-77.

Shieh, J. C., & Hsu I. Y. (2007). The study of Hakka Han characters retrieval based on Chinese remainder theorem. In *Proceeding of The Sixth International Conference on Information and Management Sciences* (pp.1-6). Lhasa, Tibet, China.

Appendix: The Values of $C(k_{i1}, k_{i2})$'s

Tonek$_{i1}$	Vowel k$_{i2}$	*d* Address	C(k$_{i1}$, k$_{i2}$) Value
1	1	0	64944529802520314615071967316445377
1	3	18	22072260041420641168670456121911469647357
1	4	37	23412621910698223737945114631640 75
1	5	51	17826075993273201357220661350819824012907
1	6	70	367502517688998026838303969 35
1	7	84	85342397197550106036904548080006 4457
1	8	102	32308882346973138357375801330549409340784759069506405134158841920335470218544507965940504191556559 11774608617063627
1	9	153	2811216111161579226740265389366581840581217
1	10	173	23412621910698223737945114631640 75
1	11	187	55806471626116507741506184193402139761763449148176938660322244270242937251421132853009341113616024428807332733392463007
1	12	239	59928711375561357894756570673864159244561695449833894341282590670541979389250807654745457
1	13	279	24197217353650163217771149596309809449153550858213283847565738348809919021 7
1	14	313	39774182781708941178658831746933487096784519221276767
1	15	339	69609765823880953326545861008353128876 7
1	16	359	83764549241795453708720944 0
1	17	371	92568168872506905741078035045853377912122516485340250930244207525476 47
1	18	403	53029176958893610960599612722517469788043181707822176722777
1	19	430	1
1	20	431	91393335755534353442175736813803360551177558488874096007881464715928500115628213403688313249448958961803878451096 17
1	21	482	29574185437495859353253340680215 7
1	22	498	23412621910698223737945114631640 75
1	23	512	71078163204701162494320127676529158963307127236180232942362905093585975451695572232441855417
1	24	554	10159740356428895920107347265669280117663964827424965345545544520122099986242077
1	25	590	25680831206694299788187281462979050127871314065913680434272401072566577907876566001440722639060786 2742466451591047
1	26	640	49649134073858853145464593131977400066439456533884309779754104491386396280253179164579079621 7
1	27	682	71711593239345115255790093017
1	28	696	12678327602740390681
1	29	705	25448666592997941375380470598812543677217331772605823322822721875158327900530386438590267244477609418106313557232288231552048755660 07
2	1	762	26596592538584851410864169534873984565642403 47
2	3	784	40814994482202470765822533630343 97

2	4	800	22492123753135128914677172498929
2	5	813	18918548605449
2	6	820	156563980539764048842400510252 93
2	7	835	32478880744933868806045498990713283027
2	8	852	45142872643313888348447679491289523860829234442446623984535147089043984659375527 11849399021517
2	9	893	55766522032380908973111891400 37
2	10	908	46991006860557
2	11	914	78393788072040545161479883632339493312614494804521543383818411446409126296404252400573506901816689336887
2	12	960	77127188180692132642624245238351589007994302185879060505979666007675816 07
2	13	993	15143168537809828377100419544541365821633570471544893275227408843124279085504667920 87
2	14	1031	46591301942195557247275902422152791239702631268433869874166755 37
2	15	1062	15976545590443589577964218020515479236044241387594606870299138871861
2	16	1094	18845927832216712715331052840599555831510 7
2	17	1113	77033515018191239085473949783811984382979369473661778398530898 87
2	18	1143	335515470975926798929702309321948 41
2	19	1159	29669456805767918039415 13
2	20	1170	95522625794052222725955378231009169583895363846720670887183208686293110648399890673471687
2	21	1210	80206105335638512818
2	22	1219	81001532852794987570887585688 09
2	23	1232	39670192748601628992463888678127284341690841451562239271884108498461027851680 3037
2	24	1270	72910906215906470414390722078141740653275638721384299989 46
2	25	1297	23608042164432274357986673436684790524936068861052117344284790568803024813899597787348158445339764 18337
2	26	1344	16235324522707821862813323850660155824254019382404065514604339945261215768313481429641647
2	27	1383	879162072665
2	28	1390	1610599946459446681057
2	29	1400	10832069208893924698671239425211479452495680405257462415907115669137260295506321932051196389762439741730608770078697 77
3	1	1451	17497636388696084120234526 87
3	3	1466	32224101304872050974634860 5415
3	4	1481	38996819222356748983769636
3	5	1492	12836548132994382932787795693977039 07
3	6	1510	12944896164105491213843598330327
3	7	1526	11568660202804135185055898659096319642708013 6
3	8	1546	33577371359585342015279827319285889470778117587641764645865158324866160558099879928133577

3	9	1585	1125191554831508114950600231003373 97
3	10	1602	737553263685840851839830834 4372
3	11	1615	168478332903063275473137794027122671888329952881 2796893082348548895450031281698617653759518975954 179726981687
3	12	1663	745269433664902894163359045994587857844499249663 5454164446302943434372609536040359022872257447637
3	13	1706	110798725514047045340559141674224070117345616947 7312521246649357372612324561 37
3	14	1741	1146716624527498163646366474429344250436465555139 7
3	15	1765	937838657188351229264604308451932560184198271752 2457
3	16	1790	1298576315038675938103983126134005621 27
3	17	1809	132424275448880028171435912212685164007984477009 59730454708332410405727
3	18	1841	130068064945759139729310554389815013511294378121 8234973749353742267
3	19	1871	1938060110162566238
3	20	1880	198867146973903763823940355422534807774355631616 0622194027739000596843164820524077676390009608643 56992536620915351 7
3	21	1930	266639658260349771068599529535437
3	22	1946	109377925939343038412220929
3	23	1957	588682143850527293157395802336827274576257740857 2 1736359074588233991728153404609140625608 67
3	24	1998	181103869605878506315549051576744048149460544128 6 5245534675173521753620476569 7
3	25	2033	215396644570665254776182588046653222958353173595 6 2709056680833772470328455473953405921584279597819 927
3	26	2078	121548619867212034241117715858155355043418139658 3 0646185662675177405797099461521832119009777
3	27	2119	377996006950881838549761642 07
3	28	2133	16543624983193568638265
3	29	2144	296914638627653455463317613427423810006739520813 1 7633140090885800771410499540926195098086014029692 493010838364551621171194109575 7
4	1	2199	799134382793824594814893642656107119025725 13
4	3	2222	7242217174804522687402910212 17
4	4	2237	6769896829431796783602782
4	5	2248	6980762466230080398861812412681587787
4	6	2266	17679299271350614000587
4	7	2278	133992049438710392013872191061021976666785 7
4	8	2298	295181752141937049376487343357222576766251561933 0 3361399592215954488978645347601785821665241124 17
4	9	2342	17949551164119179547091855128239434 37
4	10	2360	536409050938860000725975833906 9
4	11	2373	478089139064721563998096026852343453488753391575 8 7053623970562287297219909959516679155202188 42

4	12	2415	267951971912278708562985116748582021686812323602699914
4	13	2439	116012062670677569241279653559194615454830456485972039861168855153743422550719234467876 3
4	14	2480	97373777951009638539236858346107326192463603722 7
4	15	2504	199094769690239464340817391969470531589161435176432687553 7
4	16	2532	28800703500190574616359876
4	17	2543	261075739342253514364471586900682524971668864333644163460843516469006 3
4	18	2577	125537595856724759979850633214533280 8
4	19	2594	55733162486284969083133
4	20	2604	436659130805557897285313263821328174746154946297815193568375467719643449810322027759065196202 57
4	21	2647	155599977486564813010145580993355477
4	22	2664	23412621910698223737945114631640 75
4	23	2678	564893727256931405815285688095364401594458847252846721181784752850438230459973900092334287 53
4	24	2721	619401890263600802564735186852281568028143620
4	25	2742	159944280687954138282181444177154981695113620615284476871746236698500443003838193599407111761 27
4	26	2784	1177951520918898812109627110561616339148328440148973 91
4	27	2808	12498487446992614625781047249 6
4	28	2822	5257164150129
4	29	2828	162357497274237810998381141782520030369256507221663993161724622599869585927305476691843454470266159129257952292083544120089 3
5	1	2883	168878377699201632388054139240250981910241757
5	3	2904	1
5	4	2905	1
5	5	2906	1
5	6	2907	33882697977350288868356526 7
5	7	2921	5463468561338543864619194035827898228211851622285950 17
5	8	2946	2005397379617378022596671315414152052591005045578311624782663720962125478217362238156938751723125927
5	9	2991	24038372962628398637203249 72
5	10	3003	20324972
5	11	3005	1526071016818703996330628204219975979083131452418117891912536510252620342305684332540851
5	12	3043	8623234986969407454565259816508746889575564401442671903791516900168098383256 31
5	13	3077	1276734450754302363880076783480142658077756426979375917678216015308263593781 97
5	14	3112	3053968034854532998664963183349871816999112456579828874327616 7
5	15	3141	97269015711185102301016876553495751270857992 87

5	16	3165	10587285660886522436235863509116832500140004354162318987888117 7
5	17	3194	16695211938013019296763102843627286117501831602 07
5	18	3217	39456471054531983303920781002795535348650486549280944903607
5	20	3244	49524240386872258022186861959067409982484913342076512653156269673699091920243665463317900 87
5	21	3285	31790087
5	22	3287	1
5	23	3288	17935302336013196289239950993422133721017962165637024468678081364282274312154997772 7
5	24	3326	18724593095632640498805225276783717969213395235474773694767317618416 87
5	25	3358	62601391713058698035238897750198085861141820152628497996105936522158303747192330346427
5	26	3398	54555409820036689249679439695096755516188405644461958769005361302758861117492153728081808450 7
5	27	3440	1
5	28	3441	6284602471488624036687
5	29	3452	10635120913382026068213653618211059095557541333041118183268416172585224888185959625486255845 87
6	1	3494	991291924043032694366940565068354227143
6	7	3512	18933561304517760405430740100927651 7
6	8	3529	15248177515452358181961225227790851030670821772451993234307854054 7
6	9	3559	77231424308070559
6	11	3566	59401758626585618088870537074886995150214030086871800571087
6	12	3592	35412161870767049414814
6	13	3602	15225256976747410801031769982689050233666025392871701937978 67
6	14	3630	10326977934158301764669 38
6	15	3641	289578503341955866322657
6	16	3653	3006263
6	17	3657	13859744833169121680225830534372541480386084855987
6	18	3680	98334985008736815942514736877516469074908177
6	20	3701	84946173039332342080760303650575783088154843129899 8996
6	23	3724	37534038965377134159836775506342237007907199020952066149586514250072 7
6	24	3756	63374710004052765862371153128592 81
6	25	3772	96270146771858984062822000501356658417545186937843951171786971032177
6	26	3803	801128368398242695123417
6	29	3816	51898131345074632363479570408106273294864038823095515659472936601273
7	1	3846	21464379162989113065268673575923 7
7	6	3861	182287063653
7	7	3866	50681484602829005943633225742 01

7	8	3879	736825915636088504727488653008292847317 33
7	9	3898	3452032766467
7	11	3903	1244034032674158815061603896 26
7	12	3916	4935409942089541429388313310860735781
7	13	3933	4482628447566173767329199931470555272675772565355 70497949
7	14	3959	25629913937977764131711391
7	15	3971	1661275737352881
7	16	3978	1456363052715719456
7	17	3986	12172994738807119694929487674
7	18	3999	16222681315494410513594737121487
7	20	4013	371889250621734835833464667490324642674470781429 6385
7	23	4036	71665238182396689761617365372974801810387668 48
7	24	4057	200218239419106128490513795004
7	25	4071	161345243787280869196121178765709531548367864303 581037
7	26	4095	33287520344187096554000336156494463862
7	28	4112	175858486
7	29	4116	109663261776479189304603073134346770927669 34