An Accessible Coded Input Method for Japanese Extensive Writing

Takeshi Okadome **Junya Nakajima** **Sho Ito** **Koh Kakusho**

Kwansei Gakuin University/ 2-1 Gakuen, Sanda-shi, Hyogo-pref. 669-1337 Japan.

tokadome@acm.org cmj88416@kwansei.ac.jp baj886106@kwansei.ac.jp kakusho@kwansei.ac.jp

Abstract

The orthogonal code, *O-code*, proposed here is an asscessible coded input method for Japanese text typing. To a kana, it assigns a romaji (literally "Roman letter") sequence corresponding to the kana and, to each of the 458 most-frequently-used kanji characters, it also assigns a successive two key stroke code which differs from any romaji representation for a kana. With the standard English keyboard or virtual one on a tabletop, the features of the method enable a novice user to input Japanese text using the kana-to-kanji conversion by inputting the corresponding romaji for a kana and a well-trained user to type text fast with less fatigue.

1 Introduction

Widespread mobile phones and tabletops have brought us a chance of using various methods for text input not through a physical keyboard. In particular, smart phones use interactive and/or stroke-based soft "keyboard" on a touch screen. Imagine, however, when you must input a vast amount of text on a small touch screen with one or two of the fingers. Then you will soon need a physical standard keyboard or virtual one on a tabletop. (For example, see Lewis, LaLomia, and Kennedy (1999) and Rick (2010) for virtual keyboards.)

On the other hand, some of the methods adopt even different modes such as speech, handwriting, and gesture recognition (Kölsch and Turk, 2002). They have, however, not achieved good performance in practice.

The situations hold for Japanese text input. This paper discusses input methods for a great amount of Japanese text such as academic and research articles, novels, and long blogs. It introduces a Japanese input method that, with the standard English keyboard or virtual one, enables a novice user to input Japanese text and a well-trained user to type text fast with less fatigue.

2 Japanese Input Methods: Brief Review

2.1 Japanese writing system

Almost all normal Japanese writing is a mixture of kanjis, kanas, and others. Kanjis are Chinese characters. (For detail of Japanese writing system and discussion on a variety of Japanese typing methods, see Yamada (1983).) The largest kanji dictionary today lists more than 50,000 distinct kanjis. They are used for nouns and the verb roots, adjectives, adverbs, and the like. The set of kanas consists of two subsets of syllabaries, hiraganas and katakanas, each of which in turn is made up of about 80 letters. Hiraganas are used to write inflections and other grammatical parts of sentences and katakana are used for the transcription of foreign words. Many different kanjis that have different meanings may have the same reading. That is, they are represented by the same kana strings. This is called *homophone problem.*

2.2 Kana-to-kanji conversion

The complexity of Japanese writing system has led to the development of a variety of Japanese typing methods. Among the typing methods, kana-to-kanji conversion with the romaji (literally "Roman letters") input for the kanas is the most popular. If all kanjis are coded exactly as they are pronounced, and if all homophone problems are solved by syntactic and/or semantic analyses, then kana-to-kanji conversion might to be best solution for the input problem.

2.3 Coded input method

Unfortunately, too many obstacles prevent kana-to-kanji conversion from being an ideal input method. The homophone problem is the most severe one and, in kana-to-kanji conversion systems, all homophonic kanjis are displayed at once

Proceedings of the Workshop on Advances in Text Input Methods (WTIM 2011), pages 31–37,
Chiang Mai, Thailand, November 13, 2011.

and the user must select the correct one by an appropriate designation such as keying or pointing. This interactive feature of kana-to-kanji conversion prevents us from touch typing and makes us irritated in Japanese typing task.

Although kanji dictionaries lists many distinct kanjis, for ordinary use, we Japanese deal with only about 1,000 and the size of the necessary kanji set increases, up to perhaps 1,500 or so for an office of 10 people (Yamada, 1983). Focussing on the ordinary usage of kanjis, some researchers have developed coded input methods in which each of the kanas and frequently-used kanjis is coded in successive two (or three) key strokes on the standard English keyboard. Unlike kana-to-kanji conversion, a coded input method such as the *T-code* (Hiraga, Ono, and Yamada, 1980) and the *TUT-code* (Ohiwa, Takashima, and Mitsui, 1983) permit users to input Japanese texts in touch typing after users practice and acquire the codes. The problem in coded input methods is that it takes us 400 to 1,000 hours to acquire them.

An attepmt to combine a coded input method with ordinary kana-to-kanji conversion enables a novice user to input Japanese text if he/she at least learns the codes for the kanas. In particular, a coded input method with an enhancement of kana-to-kanji conversion called *kanji-kana mixture conversion* partly solves the homophone problem because, in the method, input strings may be partially represented in kanji, which reduces the number of homophones. (For kanji-kana mixture conversion, see Ono (1990).)

This paper introduces a new coded input method named the orthogonal code, *O-code*, in which, to each kana, a romaji sequence corresponding to the kana is assigned and, to each of the 458 most-frequently-used kanji characters, a successive two key stroke code which differs from any romaji representation for a kana is assigned. With the standard English keyboard, the features of the *O-code* enable a novice user to input Japanese text using the kana-to-kanji conversion by inputting the corresponding romaji (Roman letter) for a kana and a well-trained user to type text fast with less fatigue.

Today people ordinarily use kana-kanji conversion systems for inputting Japanese text. Because many of them can touch-type the romajis for the kanas, the *O-code* with kana-kanji conversion is more accessible than the other coded input methods in which even the kanas are coded in meaningless and non-associative two (or three) keystrokes.

3 Design Principles

3.1 Requirements

The design requirements of the *O-code* are its "accessibility" and "efficiency." That is, a code system are required to embody the following features:

1. Novice users who have learned English typing can easily use for Japanese text inputs together with kana-to-kanji conversion.
2. A high level of input speed is attainable.
3. Users suffer from less "fatigue." The fatigueness is indirectly measured by counting interactive choice of kanjis that are displayed on the screen, for example.
4. A high rate of accuracy may be maintained. That is, the codes should not be susceptible to erroneous finger motions.

The last three of them coincide with those in the *T-code* (Hiraga, Ono, Yamada, 1980).

These factors are closely correlated to each other. Of them, barrier-free (that is, less interactive) and speed are in a way the most decisive, and also the most appealing factors although they are somewhat contradictory.

A system design would be satisfied with a multi-stroke code system on the standard English keyboard, where we type a kana by the romaji input method and each of the most-frequently-used kanjis by two successive key strokes which differ from any romaji representation for a kana. This immediately implies that the system is usable by an untrained user and, for a trained user, he or she can type fast with less fatigue because of less interactive choices of kanjis.

3.2 Design

The *O-code* basically adopts the same strategy of the kanji-set selection and its assignment to two-stroke codes as the *T-code* does, because (1) the design policy of the *T-code* satisfies requirements 2-4 and (2) there are several multi-stroke systems currently implemented and put to use, and among them, the most famous one is the *T-code* that has the following specifications (Yamada, 1983; Hiraga, Ono, and Yamada, 1980).

1. Only 40 out of 48 printing key are used.
2. The shift key and space bar are not involved in character codes.
3. Out of 40 × 40 = 1,600 possible two-stroke pairs, some of those that involve the top-row keys are not used. In particular, the same-hand pairs involving the top-row keys are almost all unused. This gave us about 1,200 usable pairs, which should cover more than 95 percent of kanjis, or more than 98 percent of the total characters, in ordinary text. The coverage is almost 100 percent if the code set is tuned to the task of a specific group or individual.
4. There seem to be no scientific data indicating which is the better hand for starting alternate-hand stroking. The *T-code* chose to start with the right (and stop with the left); that is, the most frequently used characters are coded in RL pairs, then in RR pairs, and finally LR pairs, which makes the expected length of the alternating-hand sequences largest.
5. The actual assignment of codes to characters is made to optimize various parameters.

3.3 Design restrictions

Unlike the *T-code*, the *O-code* uses only 30 printing keys that are not the top-row keys. Out of 30 × 30 = 900 two-stroke pairs, only 458 pairs are used for the kanji coding. This is because,

1. to kanjis, the *O-code* cannot assign the two-stroke codes that match the first two letters in a romaji representing a kana or
2. it also cannot use those in which the first stroke is for a vowel: ‘a,’ ‘e,’ ‘i,’ ‘o,’ or ‘u,’ and a punctuation.

4 The Determination of the Kanji Set

The kanji set selection is based upon the frequencies of the usage of kanjis in texts on the web. For this, we use the 1-gram data in the Web Japanese N-gram data in the set of about 200,000,000 sentences that Google collected from the web and analysed (Kudo and Kazawa, 2007). The top ten kanjis are

1. 人
2. 日
3. 方
4. 者
5. 中
6. 的
7. 私
8. 一
9. 年
10. 時.

Considering a statistical fact that about 500 kanji and 150 kana characters are used by an average person for his daily use (Hiraga, Ono, and Yamada, 1980), although the set may change gradually, our selection of the kanji set seems reasonable.

5 Coding of the Kanji Set

The kanji coding is based on the efficiency of finger movements just like *T-code*. The coding method maps kanjis, arranged in the order of the frequency of usage, to the key pairs arranged in a suitable ordering as defined below. Second order adjustments will be made afterwards.

The method obtains the ordering of key pairs by assigning certain weights to certain characteristics of hand motions and using their linear sum for each key pair. It gives a larger weight to characteristics that seems to have a greater importance. The assignment will bring us the ordering of 458 key pairs on a 30-key keyboard.

This ordering, however, is not immediately usable to fix the assignment of characters directly, because the typing process is not a collection of isolated key pairs, but their continuous sequence. If, for example, key pair (h, f) is with a high score, then (f, h), its reverse, would also be with a high score, but the frequent appearances of these two key pairs would result in the frequent tapping motion of key pairs (h, h) and (f, f) in the interval of consecutive (h, h)s and (f, f)s, or vice versa, which are known to be less preferred. This would also be adverse to alternate hand stroking as well.

Through such considerations, Hiraga, Ono, and Yamada (1980) concludes that the desirable keyboard characteristics may be itemized as follows:

1. The whole typing procedure is to keep as much keying rhythm as possible. Fluent rhythm, as well as high average of typing speed, is best realized by alternate stroking by both hands. Thus, it would be our principal objective to let the code system be such that it would allow alternate hand stroking as much as possible.

2. Hands should not be moving up and down incessantly on key rows, but stay in the same row as much. Thus, strokes on the home row should be used as much as possible and excursions to other rows should be held minimum. Comparing between the upper and the bottom rows, all evidences point out that hands are more fluent on the upper row, so the ranking of rows should be in the preference order of the home, the upper, and the bottom.

3. Fingers should be loaded in proportion to their dexterity. In typing motions, fingers are divided into the stronger ones (index and middle fingers) and the weaker ones (ring and little fingers). Index and middle fingers are not so much different in their capacity and functions. However, we must keep in mind that each index finger must cover two inner columns. The difference between ring and little fingers is also not so obvious. Although a ring finger is superior in its stroking force in typing motions, a little finger may have the advantage of the twisting motion of the wrist. (In the *T-code*, little fingers will be given more emphasis than the ring.)

4. The number of awkward keying sequences must be decreased as much as possible. Almost all awkward key pairs are of one-handed stroking, again attesting to the importance of alternate hand stroking. The major awkward key pair sequences, in the order of their disadvantages are:

   (a) Hurdling: the stroking from the upper to the bottom row vice versa, jumping over the home row.

   (b) Reaching: the stroking of different keys with the same finger.

   (c) Tapping: the stroking of the same key.

   (d) Rocking: stroking with adjacent fingers, especially from an inner to an outer one.

The *T-code* weighting process starts by accommodating for condition (1) above. The key pairs are divided into 4 blocks, namely RL, RR, LL, and LR blocks, where symbols L and R stand for the hands that stroke the keys of the pair. The blocks are given preference in the order above, and key pairs in each of the blocks are then ordered by taking further conditions into account. Within individual blocks, conditions (2), (3), $\cdots$, are evaluated and weighed accordingly, and the whole ordering is decided. Awkward sequences are deliberately given negative weights in order to bring down their ranking, thus decreasing their occurrences when the codes are used.

In the *O-code*, we assgin a key pair in the order to a kanji in the occurrence-frequency order in the selected kanji set described in the previous section unless the pair matches the first two letters in a romaji representation for a kana. The entire code table for the kanji set is given in the appendix.

6 Evaluation

The code system test would be to actually measure its typing speed and error rate on a real system. The test, however, requires much time and cost.

Hence, we show some statistical figures about *O-code* derived, again, from the Web Japanese N-gram data in the set of about 200,000,000 sentences that Google collected from the web and analysed (Kudo and Kazawa, 2007).

Figures 1 and 2 illustrate the finger loading and row distributions of the *O-code*. From the figures, we see for our code that:

1. Hands are evenly loaded, slightly lighter for the weaker left hand.

2. About 47 and 35 percents of strokes fall on the upper and home rows.

3. The loading of fingers is slightly different from the conjectured strengths of the fingers. In particular, the little finger loading of the left hand is relatively higher; Also the ring finger loading of the right hand is relatively higher.

In the *T-code*, about 24 and 56 percents of strokes fall on the upper and home rows and, in the *TUT-code*, about 34 and 56 percents of strokes do. Furthermore, in both the *T-code* and *TUT-code*, the loading of fingers is in a qualitative agreement with the conjectured strengths of the fingers. The

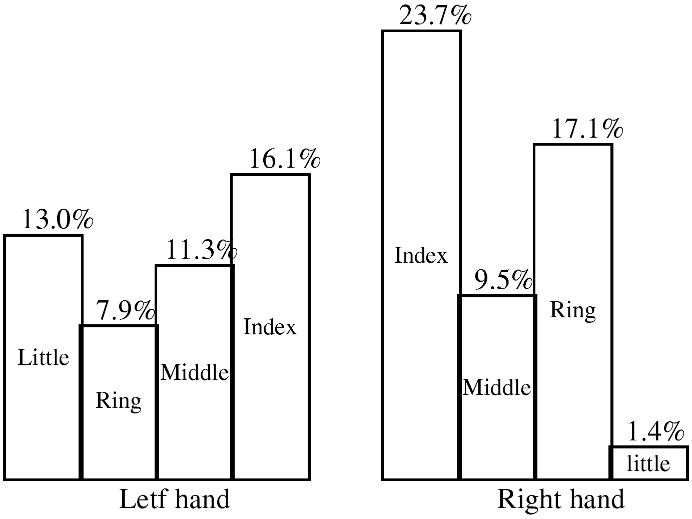

Figure 1: Finger loading distributions

results of these analyses may lead us to a conclusion that the *O-code* is less efficient than the *T-code* and *TUT-code*. This is because four out of the five vowels are on the upper row and vowel 'a' is typed by the little finger in the standard English keyboard.

7 Concluding Remarks

This paper proposes a new coded input method named the orthogonal code, *O-code*, for the input of Japanese texts. In the *O-code*, to each kana, a romaji sequence corresponding to the kana is assigned and, to each of the 458 most-frequently-used kanji characters, a successive two key stroke code which differs from any romaji representation for a kana is assigned. With the standard English keyboard, the features of the *O-code* enable a novice user to input Japanese text using the kana-to-kanji conversion by inputting the corresponding romaji for a kana and a well-trained user to type text fast with less fatigue. The principal goal has been to realize an input system that would allow a high degree of touch typing. The *O-code* is semi-optimal for experts, but it provides their easier accessibility by beginners.

We have implemented a *O-code* system using the Kanchoku Win system (Kanchoku Win, 2006) that enables us to input Japanese texts by a coded input method on the Windows. We have also developed the $O\text{-}code_{DSK}$ for the Dvorak simplified keyboard (Dvorak, 1943). The $O\text{-}code_{DSK}$ is more efficient that the *O-code*, because the five vowels are on the left-hand home row and the high-frequency consonants on the right-hand home row. Furthemore, the restriction of the romaji system into the kunrei system that is an optimal romaji system permits us to encode more kanjis in both the *O-code* and the $O\text{-}code_{DSK}$.

Furthermore, we have constructed about 1,500 sentences for training the *O-code* using the eelll/JS (eelll/JS, 2005) that is a system for practicing a coded input method on a web browser.

Acknowledgments

This research was partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 23300073, 2011.

References

1. Dvorak, A. (1943). There is a better typewriter keyboard. *National Business Education Quarterly*, 12, 51-58;66.

2. eeel/JS (2005). `http://www.sato.kuis.kyoto-u.ac.jp/~yuse/tcode/eljs/` (accessed on 13 May 2010).

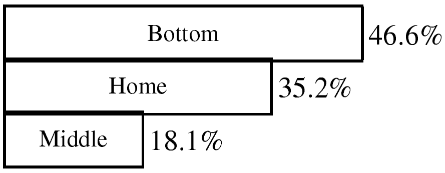

Figure 2: Row distributions

3. Kanchoku Win (2006). http://www.sato.kuis.kyoto-u.ac.jp/\~yuse/tcode/kw/ (accessed on 13 May 2010).

4. Kölsch, M. and M. Turk (2002). Keyboards without keyboards: a survey of virtual keyboards. *Technical Report 2002-21*, UCSB, Santa Barbara, CA.

5. Kudo, T. and H. Kazawa (2007). *Web Japanese N-gram Version 1*, Gengo Shigen Kyokai.

6. Hiraga, Y., Y. Ono, and H. Yamada (1980). An assignment of key-codes for a Japanese character keyboard. *Proceedings of the 8th International Conference on Computational Linguistics*, 249-256.

7. MacKenzie, I. S., S. X. Zhang, and R. W. Soukoreff (1999). Text entry using soft keyboards. *Behaviour and Information Technology*, 18(4), 235-244.

8. Rick, J. (2010). Performance optimizations of virtual keyboards for stroke-based text entry on a touch-based tabletop. *Proceedings of the Twenty-Third Annual ACM Symposium on User Interface Software and Technology (UIST'10)* 77-86.

9. Ohiwa, H., T. Takashima, O. Mitsui (1983). A Method of touch-typing Japanese text. *IPSJ Journal*, 24(6), 772-779. (In Japanese)

10. Ono, Y. (1990). Auxiliary input methods for T-Code system : a kanzi-form combination and a kanzi-mixed conversion. *IPSJ Journal*, 31(3), 404-414. (In Japanese)

11. Yamada, H. (1983). Certain problems associated with the design of input keyboards for Japanese writing. In: *Cognitive Aspects of Skilled Typewriting*, edited by W. E. Cooper, 305-407, Springer-Verlag.

Appendix: *O-code* table

Total 458 Kanjis

kd 人	jd 日	kf 方	jf 者	hd 中
kg 的	hf 私	jg 一	hg 年	ks 時
ld 前	js 店	lf 見	hs 何	lg 等
ls 月	kr 今	jr 本	hr 第	jt 円
ht 事	kq 上	jq 後	hq 気	kw 目
jw 性	lr 分	hw 会	lt 大	lq 市
lw 県	yd 他	yf 出	yg 名	pd 次
pf 家	pg 度	ys 数	ps 二	yr 回
yt 用	pr 話	yq 化	pt 笑	pq 件
yw 車	pw 内	kc 来	jc 様	kv 万
jv 駅	hc 順	kb 先	hv 力	jb 法
hb 別	kz 手	jz 点	hz 品	kx 料
lc 心	jx 新	lv 所	hx 版	lb 夏
lz 物	lx 系	md 三	mf 花	mg 町
mc 国	mv 色	mb 誰	mz 水	mx 得
yc 各	yv 場	yb 屋	pc 曲	pv 十
yz 声	pb 下	pz 間	yx 部	px 感
mr 子	mt 僕	mq 書	mw 際	jk 顔
kj 君	hk 元	kh 俺	hj 朝	jh 型
k; 全	j; 社	h; 体	kl 権	lk 達
jl 歳	lj 地	hl 高	lh 金	l; 頃
kp 風	jp 夜	hp 号	lp 氏	yk 男
yj 味	yh 彼	pk 集	pj 約	y; 道
ph 区	p; 条	yl 位	pl 代	yp 枚
k, 音	j, 外	km 夢	jm 酒	h, 表
kn 旅	hm 頭	jn 超	hn 街	k/ 為
j/ 女	h/ 海	k. 機	l, 室	j. 村
lm 同	h. 再	ln 線	l/ 生	l. 当
mk 雨	mj 主	mh 番	m; 猫	ml 語
m, 式	m/ 個	m. 足	y, 御	ym 側
yn 米	p, 犬	pm 長	y/ 員	pn 量
p/ 帯	y. 館	p. 形	mp 戦	df 観
fd 空	dg 初	gd 台	fg 株	gf 級
sd 昔	sf 面	fs 絵	sg 小	gs 愛
dr 口	fr 山	dt 着	ft 記	gr 局
gt 姿	dq 四	fq 率	gq 寝	dw 例
fw 学	sq 編	sw 未	rd 光	rf 倍
td 娘	rg 宿	tf 木	tg 食	qd 五
qf 春	qg 帳	qa 星	wd 母	wf 肌
rs 費	wg 図	qs 似	ws 付	rt 森
tr 券	qe 百	qr 器	rq 右	qt 税
tq 最	wr 便	rw 歌	wt 黒	tw 種
qw 無	wq 左	dc 欄	dv 白	fc 千
fv 行	db 親	gc 役	fb 都	gv 値
gb 島	dz 作	fz 裏	gz 項	sc 箱
dx 魚	sv 安	fx 額	sb 訳	gx 身
sz 薬	sx 桜	cd 汗	cf 楽	vd 商
vf 逆	cg 低	bd 馬	vg 赤	bf 章
bg 士	zd 業	zf 文	zg 組	xd 神
cs 肉	xf 冬	vs 雪	xg 六	bs 八

zs 発	xs 板	cv 差	vc 南	cb 階
bc 北	vb 横	bv 期	zc 億	cz 美
zv 巻	vz 席	zb 七	bz 派	xc 割
cx 秋	xv 恋	vx 涙	xb 冊	bx 川
zx 賞	xz 科	rc 剤	rv 師	tc 良
rb 郡	tv 可	tb 製	qc 不	qv 紙
rz 父	qb 服	tz 石	qz 材	wc 症
wv 杯	rx 週	wb 重	tx 耳	qx 脳
wz 非	wx 葉	ce 著	cr 制	vr 論
ct 激	vt 類	br 変	bt 皆	cq 東
zr 糸	vq 園	zt 病	bq 茶	zq 夫
cw 実	xr 湯	vw 末	bw 団	zw 多
xq 産	dk 暇	dj 技	fk 九	fj 旧
gk 軍	fh 士	gj 髪	gh 熱	d; 壁
f; 王	g; 客	sk 毛	dl 丸	sj 々
fl 西	gl 術	s; 柄	sl 緑	fy 麺
dp 嫌	fp 胸	gp 半	sp 和	rk 妻
rj 奴	tk 塩	rh 居	tj 袋	qk 天
qj 鍋	r; 牛	qh 窓	t; 程	q; 選
wk 鳥	wj 火	rl 歩	wh 祭	tl 昼
ql 死	w; 堂	wl 質	qi 真	qu 敵
rp 急	qy 城	tp 匹	qp 青	wy 旬
qo 歯	wp 層	d, 命	dm 誌	f, 秒
fm 院	dn 隣	g, 卵	fn 船	gm 玉
gn 謎	d/ 盤	f/ 字	g/ 庭	s, 史
d. 奥	sm 首	f. 姫	sn 流	g. 総
s/ 状	s. 隊	ck 展	cj 塾	vk 入
vj 指	bk 太	vh 橋	bj 某	bh 職
zk 府	c; 銀	zj 省	v; 靴	zh 角
b; 弾	z; 運	xk 穴	cl 対	xj 両
vl 腕	xh 豚	bl 界	zl 波	x; 友
xl 強	c, 枠	cm 爆	v, 案	vm 虫
cn 副	b, 泊	vn 豆	bm 辺	bn 輪
z, 噂	c/ 油	zm 経	v/ 証	zn 血
b/ 毎	z/ 幅	x, 英	c. 票	xm 肩
v. 諸	xn 州	b. 列	z. 即	x/ 有
x. 魂	r, 又	rm 縁	t, 里	rn 酢
tm 計	tn 妹	q, 座	qm 炎	r/ 腰
qn 骨	t/ 圏	q/ 連	w, 校	wm 要
r. 皮	wn 龍	t. 画	q. 詩	w/ 問
w. 床	ci 悪	vy 通	cp 雲	vp 林
bp 草	zp 世	xp 鶏		