

A Corpus Analysis of Spelling Errors Made by Japanese EFL Writers

Takeshi Okada

Yamagata University (Birkbeck College, University of London)

Abstract

This corpus based analysis discusses Japanese phonological and orthographic interference to English spelling. After establishing that there are idiosyncratic properties in spelling errors generated by Japanese people who use English as a foreign language, I explore the reason why those errors occur. The investigation compares two corpora of English spelling errors, one from native speakers and the other from Japanese writers. A number of Japanese-specific errors observed in the corpora are discussed with special reference to substitution errors. The quantitative evidence indicating that Japanese spellers substitute l with r (as easily anticipated), and vice versa, i.e. r with l, suggests that when Japanese people do not have a reliable phonological (and perhaps visual) clue to select a desirable letter, they are apt to get greatly confused. I claim that 'romaji' plays a deleterious role in Japanese writers' spelling, and that Japanese spellers of English are hampered both by phonological differences between the two languages and by subsequent discrepancies in the way of using Roman letters (not kana).

1. Introduction

If the large majority of English spelling errors are shared by native speakers of English and Japanese people who write English as a foreign language (EFL), it suggests the problem lies in the irregularities of English orthography. If, on the contrary, spelling errors generated by Japanese people contain some idiosyncratic properties, it reflects an explicit interference from the Japanese writing system and the Japanese language itself.

Though it seems superficial, the analysis of spelling errors on a comparative linguistic basis will tell us the differences between the two languages on various linguistic levels and, more significantly, lead us to a more profound understanding of how one's native tongue affects his linguistic activities, especially in communicating via a foreign language. Japanese writers of English as a foreign language (hereafter abbreviated into JWEFL) sometimes get confused in selecting appropriate English letters because of the phonological differences between the two languages and the deficiency of the alphabetic transcription system called *romaji*.

This paper tries to explore peculiarities of spelling errors made by JWEFL by comparing them with errors made by native speakers/writers of English (hereafter NSWE). The exploration proceeds by paying special attention to the errors of selecting a wrong letter for the required one, the errors we call substitution errors. Section 2 presents an overall sketch of the error corpus and a number of theoretical problems concerning the comparison of these corpora. Section 3 is devoted to the analysis of substitution errors made by the two groups of writers. The serious and

deleterious effects of *romaji* on Japanese writers' English spelling are discussed there. The last section summarises the analysis and gives some concluding remarks.

2. Error Corpora

2.1 Description of the Corpora

Appendix 1 and 2 give descriptions of the original raw materials formatted into contrastive error corpora. JWEFL-made errors are obtained from various sources (two of them are not distributed on the web), and NSWEL-made errors are exclusively taken from Birkbeck Spelling Error Corpus (1986).

For the Japanese corpus, seven files were generated independently from the raw materials and merged into a single file. A corpus of spelling errors made by native speakers of British and American English, adults and children, was generated from the Birkbeck Spelling Error Corpus (compiled by Roger Mitton: A Machine Readable Edition (1986)), provided from Oxford Text Archive (<http://ota.ahds.ac.uk/>). After revising minor inadequacies and eliminating non-English target words, 13 data files were formatted into a uniform style. For further details of each source file, refer to its corresponding web site and accompanying documentation.

2.2 Inadequacies in the Corpora

A few words should be given here about some inadequacies in the original data. Some of those inadequacies are common to electronic corpora in general, and others are peculiar to these error corpora.

First I should point out some problems concerning electronic files. The present paper limits its range of analyses to non-word spelling errors taken from handwritten materials. In principle, there are at least four distinct ways of writing English. Two of them are traditional ways: handwriting and keyboarding. The other two ways of writing (or generating words, I should say) are children of the electronic age: they are realised via OCR (optical character recognition) and voice recognition technology. The latter two sometimes produce misspellings that are not produced by human beings, and errors and inadequacies generated by those systems should be handled within different frameworks.¹ The errors that occur in key-entering are divided into two types: genuine spelling errors, which I want to explore, and finger-movement errors (typos). Though we can sometimes separate genuine spelling errors from typos in key-entered materials, this is not always possible.

The next problems are the typos introduced by key-entering the handwritten materials into electronic files. Electronic data files, to which we have no alternative in computational analysis, inevitably allow the possibility of typos: I refer to typos of this sort as secondary key-entry errors.

In making error subcorpora for this research, considerable numbers of these secondary key-entry errors were detected and eliminated. For example, errors that are unlikely to be produced in handwriting such as *konds* for *kinds*, *waws* for *was*, *wehre* for *where*, *fromhere* for *from here*, etc were eliminated from the files, as were errors

like *ot* for *to*, *plat* for *play*, *pne* for *pen*, *sdked* for *asked*, *taht* for *that*, *thereis* for *there is*, etc.

The second problem arises when trying to compare different corpora—the method used to collect raw spelling errors. There are basically two distinct sources from which (handwritten) spelling errors are taken and stored: tests like spelling tests or dictation tests, and running texts such as free essays or compositions in examinations. From the former type of source we can obtain a number of spelling attempts—including correct spellings and errors—against a set of target words, while from the second type a small number of spelling errors per intended word is obtained.

In spelling tests the subjects are forced, or at least are expected, to spell out a word even if they are not confident of the correct spelling. By contrast, in free writing, a writer can choose a different word if he is not sure of one which he tries to spell for the first time; a writer can avoid the ‘risky’ word and stop writing any sentences or phrases containing a word potentially misspelled.

As a result, some target words have a lot of error attempts whereas others do not. This affects the overall texture of the corpus. There can also exist age differences and subsequent literacy differences among NSW and there are, in addition, differences of levels as foreign language learners among Japanese people. (Okada (2002) discussed in detail the problem of the level of spellers.) These differences must be born in mind when comparing the two corpora; we must make sure that the differences we find are represented across a number of the subcorpora and do not reflect an artefact of the data.

2.3 Overview of the Corpora

2.3.1 Quantitative Breakdown

I constructed two pairs of contrastive error corpora: one comprises two full corpora described in Appendix 1 and 2; another comprises subsets of these. Each of these two subcorpora contains 801 target words that appeared commonly both in the Japanese and in the English full corpora.

Table 1 Word-level breakdown of full and sub corpora

	no of target words	no of error types	no of error tokens
JWEFL	1184	5060	12478
NSWE	5624	34658	220043
total	6808	39718	232521
JWEFL	801	4135	10193
NSWE	801	7598	42660
total	1602	11733	52853

2.3.2 Frequency Problem

In the discussion so far, I have been using terms error types and error tokens without sufficient definition. Table 2 shows the attempts of 17 people at the target word *beautiful*; the table shows six error types, each of which occurred with a certain

frequency, making a total of 17 error tokens. As far as error letters are concerned, the result of the analysis is counted as in Table 3, taking frequencies into consideration.

Table 2 Part of original error source and its analysis

\$beautiful	(<-- target word)	\$beautiful	(<-- target word)
butiful	6(frequency)	butiful	6
beautiful	4	D e	(<-- Deletion of e)
biautiful	3	D a	
beautiful	2	beatiful	4
beautiful	1	D u	
beatifull	1	biautiful	3
		S i e	(<-- Substitution i for e)
		beutiful	2
		D a	
		bautiful	1
		D e	
		beatifull	1
		D u	
		I l	(<-- Insertion of l)

Table 3 Error counting

D a	8 (=6+2)
D e	7 (=6+1)
D u	5 (=4+1)
S i e	3
I l	1

Table 3 shows that the data in Table 2 contains five types of errors, including three types of deletion errors and one type of substitution and insertion error, with 24 total occurrences.

3. Error Analyses

3.1 Substitution Errors

3.1.1 Substitutions Common to Both Corpora

The commonest S-error in both of the two contrastive corpora is observed for the target letter *m*, the great majority of these being the substitution of *n* for *m*—94% of 235 JWEFL-made errors compared with 89% of 666 NSWE-made errors.

S-errors for *c* are also commonly observed both in JWEFL-made and NSWE-made errors, as show in Table 4. To take the first line of the Japanese-speakers' part of Table 4, this shows that the letter *s* was substituted for *c* 452 times, which makes 45.7% of all the substitutions of *c*. The final column shows that, out of 4297 occurrences of *c* that should have been written (if all the words containing *c* had been written correctly), 10.5% were substituted by *s*. In the native-speakers' part, *c* should have been written for 12239 times.

Table 4 Letters written in place of *c*

Japanese speakers				Native speakers			
other letter	freq	%1	%2	other letter	freq	%1	%2
s	452	45.7	10.5	s	785	58.3	6.4
x	324	32.8	7.5	x	195	14.5	1.6
t	150	15.2	3.5	t	127	9.4	1
k	36	3.6	0.8	k	93	6.9	0.8
r	6	0.6	0.1	q	42	3.1	0.3
g	6	0.6	0.1	r	23	1.7	0.2
e	5	0.5	0.1	g	19	1.4	0.2
d	3	0.3	0.1	n	14	1	0.1
n	2	0.2	0	l	10	0.7	0.1
a	1	0.1	0	f	9	0.7	0.1
l	1	0.1	0	h	8	0.6	0.1
y	1	0.1	0	a	4	0.3	0
f	1	0.1	0	v	3	0.2	0
q	1	0.1	0	e	2	0.1	0
total	989	(989)	(4297)	i	2	0.1	0
				d	2	0.1	0
				y	2	0.1	0
				b	2	0.1	0
				u	1	0.1	0
				p	1	0.1	0
				m	1	0.1	0
				w	1	0.1	0
				z	1	0.1	0
				total	1347	(1347)	(12239)

S-errors for *d* shows an interesting overlapping: despite the fact that there is no voiced and non-voiced distinction in Japanese phonology, substituting *t* for *d* is the most frequent error made by JWEFL as well as by NSW. What is peculiar about S-errors for *d* is the fact that while in NSW-made errors *b* is rarely misemployed for *d* (7.84% of total number of substituting errors for *d*), *b* is incorrectly used 26 times in JWEFL-made errors (25.49% of 102 S-errors for *d*). This relatively high percentage is mainly due to errors made by Japanese novice writers who are not good at distinguishing the letter *d* from *b*. Spelling errors made by junior high-school students contain *besk* for *desk* (10 times), *bon't* for *don't* (2 times), etc. However, as we will see shortly, an S-error using *d* for *b* such as *adout* for *about* is much commoner in NSW-made errors than they are in JWEFL-made errors.

Table 5 S-errors for vowel letters

Japanese speakers			Native speakers		
for a	freq	%	for a	freq	%
e	450	52.7	e	2352	56.5
o	174	20.4	o	996	23.9
u	156	18.3	i	536	12.9
i	60	7	u	226	5.4
g	4	0.5	n	23	0.6
total	844	(854)	total	4133	(4165)

Japanese speakers			Native speakers		
for e	freq	%	for e	freq	%
a	846	53.6	a	1723	43.5
i	602	38.1	i	1618	40.8
u	56	3.5	o	268	6.8
o	42	2.7	u	204	5.2
y	8	0.5	d	38	1
t	5	0.3	y	38	1
n	4	0.3	t	24	0.6
total	1563	(1579)	total	3913	(3961)

for i	freq	%
e	425	58.9
y	184	25.5
a	65	9
u	24	3.3
o	14	1.9
total	712	(722)

for i	freq	%
e	1733	51
a	669	19.7
y	645	19
u	168	4.9
o	145	4.3
total	3360	(3399)

for o	freq	%
e	356	55.2
a	222	34.4
u	34	5.3
i	25	3.9
total	637	(645)

for o	freq	%
a	819	39.8
e	573	27.8
u	426	20.7
i	222	10.8
total	2040	(2060)

for u	freq	%
a	332	51.6
e	109	17
o	108	16.8
i	77	12
n	6	0.9
w	6	0.9
total	638	(643)

for u	freq	%
o	721	42.7
a	320	19
e	301	17.8
i	155	9.2
w	96	5.7
n	64	3.8
total	1657	(1688)

S-errors for vowel letters are common to both groups of writers. Table 5 shows that, except for S-errors using *y* for *i*, there is a similar, clear tendency of substituting one of the four vowel letters where another vowel letter is required. In addition to the five vowel letters, the S-errors of semi-vowel letters such as using *w* for *y* and *vice versa*, show a common pattern across the two groups of spellers, i.e. *w* is most frequently substituted by *u*, while *y* is substituted by *i*.²

3.1.2 JWEFL-Peculiar Substitutions

The S-errors that JWEFL tend to make are presented in this subsection. The comparison of S-errors using *s* for *t* is shown as follows.

It is clearly observed that Japanese errors using *s* where *t* is required is predominant. This is predictable since there is no / / sound and, of course, its corresponding letter or letter-combination in Japanese, substituting *t* (in the sequence *th*) with *s* is highly probable. However, we should be a little more careful before we determine that this S-error is actually reflecting JWEFL-made error quality, since the frequency of this error in the corpus is largely due to a particular target word that requires *th* and which is misspelled repeatedly by JWEFL in a spelling test, i.e. *threat* in `SAMANTHA-error.txt`.

There are 41 items containing the two letter sequence *th* in 801 common target words. In the JWEFL-made error file, there are two target words that were misspelled by substituting *th* with *s*: *anything* and *threat*, while in the NSWE-made error file, only one target word, *other* is misspelled as *usfer*. The problem is the number of error word types and their frequency. In the NSWE file *usfer* for *other* appears only once; and in the JWEFL file *anyising* for *anything* appears once, but the target word *threat* has 166 different error word types with 400 total occurrences.

This is because JWEFL-made errors for *threat* were obtained from a spelling test; the same target word was also used in one of the NSWE spelling tests but unfortunately the original report from this particular source does not present all of the error word types but only one or two of the most frequent error types. Therefore, I cannot conclude easily that JWEFL are inclined to make an S-error using *s* for *th*. But this inclination can still be supported even after the S-error of *s* for *th* is eliminated

from the table: JWEFL illegally use *s* where *t* (not *th*) is required 233 times (62.97%) out of the 370 total of S-errors for *t*, compared with for 112 times (25.81% out of 434) for NSWE.

Table 6 Letters written in place of *t*

Japanese speakers				Native speakers			
other letter	freq	%1	%2	other letter	freq	%1	%2
s	400	74.5	7.3	s	113	26	0.5
d	65	12.1	1.2	d	80	18.4	0.4
c	36	6.7	0.7	c	63	14.5	0.3
l	5	0.9	0.1	l	40	9.2	0.2
z	5	0.9	0.1	r	24	5.5	0.1
y	4	0.7	0.1	h	21	4.8	0.1
v	4	0.7	0.1	e	18	4.1	0.1
i	3	0.6	0.1	n	16	3.7	0.1
i	2	0.4	0	f	13	3	0.1
r	2	0.4	0	m	10	2.3	0
n	2	0.4	0	g	9	2.1	0
q	2	0.4	0	v	6	1.4	0
f	2	0.4	0	a	4	0.9	0
e	1	0.2	0	o	4	0.9	0
h	1	0.2	0	b	4	0.9	0
b	1	0.2	0	p	3	0.7	0
p	1	0.2	0	y	3	0.7	0
'	1	0.2	0	i	1	0.2	0
total	537	(537)	(5488)	u	1	0.2	0
				w	1	0.2	0
				k	1	0.2	0
				total	435	(435)	(21160)

The most striking difference about the S-error of *s* for *t* is observed in the spelling *sion* where *tion* is required. For five of the target words with *tion*, i.e. *accommodation*, *exhibition*, *pronunciation*, *station* and *vacation*, JWEFL incorrectly use *sion* with 59 different error word types with 221 in total frequency. The most frequent error is *exivision* for *exhibition* (54 times), where errors other than substituting *s* for *t*, i.e. deletion of *h* and substituting *v* for *b* are also involved; the simplest substitution *s* for *t* occurs six times in *exhibision* for *exhibition*. By contrast, though there are four target words with *tion* incorrectly spelled with *sion* in the NSWE corpus, i.e. *exhibition*, *position*, *satisfaction* and *station*, there are only 11 error word types corresponding to these four targets generated by substituting *t* with *s* such as *exibision* for *exhibition* or *satisfacson* for *satisfaction*, and the total frequency is only 13. From this quantitative comparison, it may be concluded that to use *s* incorrectly where *t* is required is a peculiar tendency among JWEFL-made errors.³

Next let us observe more striking results that reflect Japanese-peculiar S-error patterns. It is convenient to divide the following S-errors into two groups: one is a group of target letters *l* and *v* that do not exist in the *romazi* chart; and the other is of letters *b* and *r* that are employed in *romazi*.

It is clear that JWEFL make frequent S-errors for *l*. The letter *l* is required 3134 times in the JWEFL-made error corpus, and over 22% of these are incorrectly replaced by another letter. And, as can be expected, *l* is substituted with *r* overwhelmingly. In other words, when JWEFL make S-errors for *l*, the letter incorrectly chosen is mostly *r*. The following is a partial list of error words generated by JWEFL via S-errors of this sort and those generated by NSWE. The target word and the number of occurrence of each misspelling are indicated in parentheses.

(1) JWEFL-made S-errors using *r* for *l*

negrect(neglect:62) sorid(solid:49) crean(clear:24)
 poritician(politician:7) frower(flower:4)
 frowers(flowers:4) raughter(laughter:4) cleaned(cleaned:3)
 sincerery(sincerely:3) crass(class:2) probrem(problem:2)

NSWE-made S-errors using *r* for *l*

appre(apple:1) gralulary(gradually:1)
 imeditery(immediately:1) particurally(particularly:1)
 prombre(problem:1) reentry(recently:1)⁴

Table 7 S-errors for *l*

Japanese speakers

other letter	freq	%1	%2
r	665	95.1	21.22
t	6	0.9	0.19
n	4	0.6	0.13
o	4	0.6	0.13
h	4	0.6	0.13
s	3	0.4	0.1
d	2	0.3	0.06
b	2	0.3	0.06
y	2	0.3	0.06
e	1	0.1	0.03
i	1	0.1	0.03
a	1	0.1	0.03
c	1	0.1	0.03
w	1	0.1	0.03
v	1	0.1	0.03
z	1	0.1	0.03
total	699	(699)	(3134)

Native speakers

other letter	freq	%1	%2
r	35	23	0.3
n	30	19.7	0.2
t	25	16.4	0.2
d	13	8.6	0.1
h	7	4.6	0.1
s	6	3.9	0
y	5	3.3	0
e	4	2.6	0
g	4	2.6	0
i	3	2	0
c	3	2	0
p	3	2	0
w	3	2	0
k	3	2	0
a	2	1.3	0
b	2	1.3	0
v	2	1.3	0
u	1	0.7	0
x	1	0.7	0
total	152	(152)	(13004)

Table 8 S-errors for *v*

Japanese speakers

other letter	freq	%1	%2
b	44	64.7	12.9
n	7	10.3	2.1
r	5	7.4	1.5
w	4	5.9	1.2
m	2	2.9	0.6
f	2	2.9	0.6
t	1	1.5	0.3
s	1	1.5	0.3
d	1	1.5	0.3
g	1	1.5	0.3
total	68	(68)	(341)

Native speakers

other letter	freq	%1	%2
f	186	75.9	7.01
t	16	6.5	0.6
b	11	4.5	0.41
r	7	2.9	0.26
w	6	2.4	0.23
d	4	1.6	0.15
n	3	1.2	0.11
s	3	1.2	0.11
y	3	1.2	0.11
c	2	0.8	0.08
l	1	0.4	0.04
p	1	0.4	0.04
h	1	0.4	0.04
m	1	0.4	0.04
total	245	(245)	(2653)

This tendency is also observed clearly among S-errors in the word-initial position where *l* is required. There are 28 target words that require word-initial *l* with 281 attempts in the JWEFL corpus and 1265 attempts in the NSWE corpus. JWEFL succeed in beginning with correct *l* with 86.48% accuracy (243 attempts), while

NSWE begin with correct *l* with 99.53% accuracy (1259 attempts). More interestingly, while 86.84% of the word-initial letter that is incorrectly chosen (33 out of 38 incorrect letters) by JWEFL for *l* is *r*, NSWE never make this error, i.e. NSWE do not make an S-error for *l* with *r* in the word-initial position.

A similar inclination can be observed in S-errors using other letters for *v*. There are two points to note: first while JWEFL frequently employ *b* at the position where *v* is required regardless of their phonological contrast, NSWE seldom did so; second, whereas NSWE frequently misemployed *f*, the non-voiced counterpart of /v/ sound, for *v*, JWEFL rarely did so. These facts would suggest that for NSWE, whose phonology does distinguish both /v/ and /b/ on the one hand and /v/ and /f/ on the other, the pair of letters with the voiced and unvoiced contrast, i.e. *v* and *f*, is the more difficult to distinguish.

(2) JWEFL-made S-errors using *b* for *v*

busiter(visitor:7) bisiter(visitor:5) beliebe(believe:1)
 libing(living:1) twelbe(twelve:1) begiter(visitor:1)
 bisitar(visitor:1) bisitor(visitor:1) bister(visitor:1)
 biziter (visitor:1)

NSWE-made S-errors using *f* for *v*

abofe(above:2) comparifly(comparatively:1) gif(give:2)
 gife(give:1) inconfeniant(inconvenient:1)
 twelf(twelve:20) fery(very:1) fisited(visited:1)

On the contrary, for JWEFL, who do not have strict distinction between these two sounds, as pointed out in the discussion of *romazi* and Japanese sounds, choosing between *v* and *b* is a troublesome and confusing task.

The word-initial *v* is also frequently substituted with *b* by JWEFL (36 out of 173 attempts, i.e. 20.81%), while NSWE scarcely ever misemploy *b* for word-initial *v* (1 out of 507 attempts).

Table 9 S-errors for *b*
 Japanese speakers

other letter	freq	%1	%2
v	228	68.9	11.3
d	68	20.5	3.4
p	8	2.4	0.4
g	8	2.4	0.4
h	4	1.2	0.2
s	3	0.9	0.1
l	3	0.9	0.1
z	2	0.6	0.1
a	1	0.3	0
t	1	0.3	0
n	1	0.3	0
o	1	0.3	0
y	1	0.3	0
f	1	0.3	0
w	1	0.3	0
total	331	(331)	(2023)

Native speakers

other letter	freq	%1	%2
p	89	42.4	1.5
d	80	38.1	1.4
v	14	6.7	0.2
l	10	4.8	0.2
m	3	1.4	0.1
f	3	1.4	0.1
t	2	1	0
r	2	1	0
h	2	1	0
q	2	1	0
n	1	0.5	0
o	1	0.5	0
w	1	0.5	0
total	210	(210)	(5784)


```
acculate(accurate:1)   olange(orange:1)   reglet(regret:1)  
spilits(spirit:1)  surprised(surprised1)  tleat (treat:1)
```

3.2 Interference of *Romazi*

English makes any word forms with a small set of letters—the alphabet—that have no individual meanings, while in some other languages things are quite different. Chinese, for example, instead of using a small set of meaningless letters and their combinations, uses characters possessing inherent meaning.

What makes Japanese orthography particularly unique and therefore makes its speakers have persistent trouble in learning any Western languages, is the fact that it has a complicated, hybrid writing system in which at least four distinct writing codes go side by side. They are *hiragana*, *katakana*, *kanzi* and *romazi*.

Hiragana and *katakana* are perfect syllabaries and they correspond to each other, having 46 basic characters and some of their combinations. In other words, they have a perfect one-to-one phoneme-grapheme correspondence (PGC). In addition *hiragana* is usually used to transcribe original Japanese words including function words, whereas *katakana* is used to express foreign (borrowed) words, and the great majority of onomatopoeias are transcribed with it. Although, like many other languages, Japanese has a great number of homophones, over 3000 *kanzi* letters, Chinese characters borrowed and adopted through 1500 years, serve efficiently to disambiguate them.⁵ These three Japanese writing systems, *hiragana*, *katakana* and *kanzi* complement each other quite efficiently. Serious trouble is evoked, however, by the last writing code, i.e. *romazi*.

The term *romazi* can be divided into two components: *roma* (Roman) and *zi* (letter). Although the letters are all taken from the Roman alphabet, the *romazi* system uses only 19 to 21 of them. The difference in the number of letters between *romazi* and the English alphabet produces some of the JWEFL-peculiar spelling errors. *Romazi* was originally introduced to Japan (more exactly invented) over a century ago in order to transcribe Japanese with Romanized letters. However, when Japanese people try to write English words, *romazi* has serious, even deleterious effects. This writing code has two properties that are pernicious to foreign language learning, especially spelling.

First, although it uses letters of the alphabet, *romazi* is a transcription of Japanese sounds that are vastly different from English sounds. The *romazi* system was basically coined in order to represent Japanese *hiragana* and *katakana* letters that are perfect syllabaries. In the *romazi* system, only the five Japanese vowels are represented by a single letter, i.e. *a*, *i*, *u*, *e* and *o*; 64 other sounds written in a single *kana* and 49 written in combined *kana* are represented by combinations of 19 or 21 *romazi* letters. As there do not exist in the Japanese language a number of English sounds including diphthongs, semi-vowels or schwas, there are no *romazi* counterparts for them. In addition, and more importantly, Japanese lacks some phonological distinctions between certain phonemes such as /l/ and /r/, /b/ and /v/, /s/ and /z/, and, sometimes, even /f/ and /h/. However, the lack of *l* and *v* in *romazi* does not mean that Japanese does have actual /r/ and /b/ sounds instead. The truth is that there are purely Japanese sounds described as らりるれろ and ぱびぶべぼ in *hiragana* (or in ラ

リルレロ and バボブベボ in *katakana* respectively). The problem occurs because *romazi* employs only *r* and *b*, ignoring *l* and *v*, and transcribes those Japanese sounds as *ra, ri, ru, re, ro* and *ba, bi, bu, be, bo* respectively.

Observe the following table, where Hepburn style stands for a set of *romazi* letters decreed in 1945, whereas *Kun-rei* style stands for another set promulgated by Japanese cabinet order (= *kun-rei*) in 1937.

Table 11 English alphabet and *romazi* letters

English	a	b	c	d	e	f	g	h	i	j	k	l	m
Hepburn style	a	b	/	d	e	f	g	h	i	j	k	/	m
<i>Kun-rei</i> style	a	b	/	d	e	/	g	h	i	/	k	/	m

n	o	p	q	r	s	t	u	v	w	x	y	z
n	o	p	/	r	s	t	u	/	w	/	y	z
n	o	p	/	r	s	t	u	/	w	/	y	z

In addition to *l* and *v*, we notice vacant slots for English *c, q* and *x*; and for English *f* and *j* in the *Kun-rei* row.

What is inconsistent in the *romazi* system is the fact that although the Hepburn system is more popularly used in everyday transcription of Japanese sounds as can be seen in signs of transportation companies or in tourism, the *Kun-rei* style is the only single ‘officially approved’ *romazi*. Therefore, though ‘*romaji*’ provides a better orthographical clue for many non-native speakers at least to imagine and retrieve its Japanese correspondent sound, public educational institutes such as elementary schools which have to obey the cabinet order have for years taught the pupils to spell ‘*romazi*’ exclusively in the *Kun-rei* style.⁶

In spite of the original ambition of the ‘inventors’ of *romazi* in the Meiji Era, there is more than one *romazi* system. To be more accurate, there is the third set of *romazi* inventories, the so-called *Nippon-siki* style, which is similar to the *Kun-rei* style. This means that, even though the *Kun-rei* style is officially taught in educational institutes, Japanese people do not have a single, uniform guideline for the use of the alphabet in order to represent their language sounds.

The more serious problem about *romazi*, as far as foreign language learning is concerned, is the fact that *romazi* is introduced into the elementary school curriculum in a school year when pupils do not start learning English as a foreign language (EFL). In the Japanese obligatory education system, EFL teaching starts in junior high-schools, and this means that every 10-year-old pupil in elementary schools encounters *romazi* without any prerequisite knowledge of Western languages. They know nothing about the phonological differences between their mother tongue and English, and, therefore, they are naturally misled into feeling that *romazi* letters can represent English sounds without adaptation.

One of the instructions printed on a *romazi* training board accompanying a Japanese language textbook encourages pupils to get accustomed to *romazi* writing insisting that it is useful in spelling English. One reason why such a careless instruction is officially distributed is the widespread popularity of Japanese word processors. In order to use Japanese word processors many people do not bother to learn the unique arrangement of Japanese *kana* keys; instead they enter Japanese

words using the usual QWERTY alphabet keys or their combinations corresponding to Japanese *kana*, and afterwards, if necessary, transform a series of *kana* letters into appropriate *kanji*. For example, *ro-mazi nyuuryoku* is transformed into ろーまじにゅーりょく and then into ローマ字入力.⁷ The brain of a person key-entering this Japanese phrase is filled only with Japanese sounds and he is just key-entering them following *romazi* spelling rules; there is no correspondence to the English alphabet.

Although the necessity of *romazi* education in recent IT society has to be admitted, its deleterious influence on EFL teaching in Japan cannot be dismissed. Not just young junior high-school pupils who have just begun to learn English but many adult Japanese who try to write English are inclined to spell English words utilising the Japanese *romazi* system, resulting in peculiar spelling errors seriously different from those produced by NSWE. This is my basic contention: JWEFL generate English spelling errors with idiosyncratic features, some of them are so unexpected that ordinary commercial spellcheckers accompanying English word processors sometimes cannot cope with them successfully.

JWEFL often ignore the fact that their native language is phonologically different from English because of the unique linguistic history of Japanese in which it borrowed and assimilated a great number of foreign words, mainly of English origin. Such borrowed words are transcribed with *katakana*, which to some extent acts as a signal to indicate that the word transcribed with it has a foreign origin. For the word *television*, the Japanese equivalent is not てれびじょん, which is *hiragana*, but テレビジョン in *katakana*. Moreover, if one tries to spell テレビジョン in *romazi*, it should be *terebizyon*, since both *l* and *v* are not included in the *romazi* chart.

What makes things more complicated is the fact that there is another Japanese word for *television*, i.e. a Japanese shortened form, テレビ and it should be spelled *terebi* in *romazi*. These peculiarities lead us to expect that a misspelling *terebi(zyon)* would appear frequently in JWEFL-made errors; however, things are not so simple. In daily Japanese, abbreviations and acronyms (both of Japanese and foreign words) are constantly used. Since *TV* is one of the most popular acronyms among Japanese people, such as in *TV Guide*, there is no S-error using *b* for *v* producing *ter(l)ebision* in the JWEFL-made error corpus.

Because of the heterogeneous characteristics of the Japanese orthographic system, or I should say the Japanese language as a whole, spelling errors made by JWEFL in their attempts to write English words have properties mysterious for human researchers and unpredictable for spellchecking facilities.

4. Discussion and Conclusion

The comparative analyses of S-errors produced by two groups of writers of English were developed in this paper on the presumption that there exist properties of spelling errors that are peculiar to JWEFL who have a different phonological/orthographical system. JWEFL-made spellings are observed to be strongly interfered with both by Japanese sounds and their self-sufficient encoding inventory, *romazi*. JWEFL are inclined to substitute a letter that is not included in *romazi* with another letter included in it, and, what is more important is the fact that they even substitute a target letter in the opposite direction. It seems unnatural that JWEFL use letters such as *l* and *v*

because they are not officially taught these letters in school. However, even though they are notorious for being bad speakers of English, the majority of Japanese people are eager to learn English, and, at the same time, are serious when trying to spell English words. They try to utilise every possible means in order to produce spellings that look most exotic, i.e. ‘English’ to them. Confusions occur at this stage: if a Japanese writer comes up to a word whose correct spelling he is not sure of, it is highly improbable that he could resort to the sound, i.e. phonological clue for the correct spelling, since in his native language there is no distinction, for example, between /l/ sound and /r/ sound, or between /v/ and /b/. In the case when a phonological clue is unavailable, the next clues for the correct word are visual clues and meta-linguistic knowledge (such as knowledge about the etymology of a word). And if all of these clues for correct spelling are not available, and this usually happens to him, a Japanese writer tries to spell English words in a *romazi* scheme. This makes him get confused in selecting one of the letters from a dubious pair for him—the results are, as we have observed, very peculiar and hard to predict.

We can assume that this kind of confusion is evoked by two distinct sources: the phonological difference between Japanese and English and structural gaps between the English alphabet and the Japanese *romazi*. It can be claimed that this confusion is a direct result of not knowing which to choose. But, as we have seen in Table 4, JWEFL make frequent S-errors for *c* just like NSW E do despite the fact that *c* is not involved in the *romazi* letters. This shows that the lack of a given letter in *romazi* does not necessarily lead only JWEFL to a labyrinth of spelling. They misemploy *s*, *t* and *k* alike, and, more significantly, JWEFL incorrectly use *x*, which is not a *romazi* letter. Both groups of writers seem to use similar phonological clues for the target spellings, although JWEFL resorted to Japanese sounds that are considerably different from English ones. I claim that especially when phonological clues are not available, such as in /l/ and /r/, or /b/ and /v/, JWEFL are influenced by the deficiency of the *romazi* inventory, resulting in producing peculiar errors.

This sort of confusion may occur on a relatively conscious level. However, there is another sort of S-error that seems to be produced unconsciously, due to the inconsistency among different *romazi* systems. Observe Table 12. In the left-most column there are 10 compound *romazi* units, each of them representing its corresponding Japanese (not exact English) sound. Four of them, i.e. *sya*, *syu*, *syo*, and *gya* are unfamiliar for English speaking people; the first three are *Kun-rei* style units and the last one is used in both the *Kun-rei* and the Hepburn style. The other six, on the contrary, are units employed only in the Hepburn style and are commonly used in English spellings. It is evident from this table that none of the *Kun-rei* style units appear in NSW E-made errors. While both JWEFL and NSW E commonly misemploy *sha* for (*politi*)*cia(n)*, it is likely that the sounds they have in mind and try to represent with *sha* are different from each other: NSW E try to represent /ʃə/, whereas Japanese シャ. The same things can be said about *shu* and *sho*.

This supports my presumption of interference from *romazi* in JWEFL-made errors. While Japanese people are aware that the Hepburn style is more suitable to represent English sounds than the *Kun-rei* style is, and, therefore, try to use Hepburn style units to spell English words more frequently, the *Kun-rei* transcriptions of Japanese *katakana* units such as シャ (*sya*), シュ (*syu*) or ショ (*syo*) still appear in

JWEFL-made errors. The officially taught *romaji* inventory, the *Kun-rei* style, which is a more faithful transplant of the *kana* inventory than the Hepburn style is, continues to have a deleterious influence on JWEFL's spellings, resulting in idiosyncratic spelling errors.

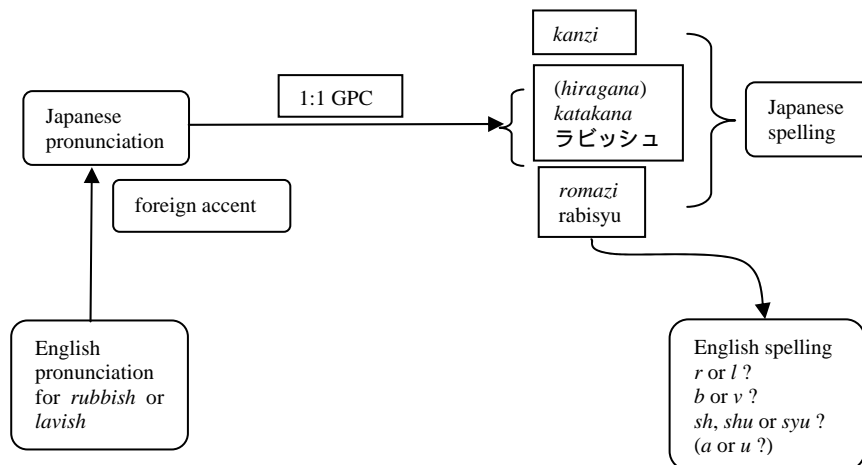
Since Japanese sounds are almost completely different from English sounds, Japanese pronunciation strongly interferes with Japanese people's pronunciation of English words, ending up with a foreign accent. For example Japanese people cannot distinguish even between *rubbish* and *lavish*, because they do not have /r/ : /l/ and /b/ : /v/ contrast in their native language. However, this sort of foreign, inaccurate pronunciation is not directly mapped onto English spelling.

Table 12 Distribution of the *Kun-rei* style errors

Japanese speakers				Native speakers				
sya	exvisyan	1	puronansyation	1				
	penlsya	2	sinsyary	2				
	poritisyian	1	cinsyary	1				
	pronansyation	2	sinsyally	1				
	pronuasyathion	1						
sha	poritishan	1			apreshate	4	esspeshally	1
	pronanshation	1			apreshat	2	expeshaly	1
	pronanshatione	1			apprshate	1	politishan	2
	sishary	1			apreashate	1	satifishan	1
	shar (\$sure)	1			aprishate	1	satishan	1
					apshate	1	satsha	1
					preshat	1	shaw (\$sure)	7
					bennyfishaly	1	sham (\$sure)	1
					penafishal	1	shar (\$sure)	1
					espeshaly	4	shau (\$sure)	1
				espeshaily	1			
syu	syuor (\$sure)	1						
shu	Engrshu	1	shuare (\$sure)	1	conshus	3	shuch (\$such)	5
	radishu	4	woshu	1	epeshuly	1	shure (\$sure)	17
	shure (\$sure)	2			satisfacshun	1	shur (\$sure)	3
syo	acomodesyon	1	pronansiesyon	1				
	exvisyon	1						
sho	acomodeashon	1			exabishon	2	stanshon	1
	acomodershon	1			effebishon	1	shor (\$sure)	26
	ekigivishon	1			eserdishon	1	shour	8
	pronanseeshon	1			exapishon	1	shore	81
	shou (\$saw)	1			excibeashon	1	show	6
	stashon	1			exebishon	1	shoe	2
					exepishon	1	shoor	2
					exvishoin	1	shoar	1
					politishon	1	shon	1
					satisfacshon	4	shoower	1
					satisfacshoin	2	shorch	1
					satiffashon	1	shorer	1
					satisfachshon	1	shou	1
				satisfackshon	1	shoure	1	
				satisfakshon	1	shouw	1	
				satisfashoin	1	shouy	1	
				satisfuckshon	1	shoy	1	
				stashon	3	vacashon	13	
				sashon	1			
cha	politichan	1	teachare	1				
	tchar	1	tichar	1				
	teachar	2	tirchar	1				
chu	cenchury	1	schuar (\$sure)	1				
cho	queschon	1	teachor	1				
gya	gyarary	5	gyaruly	1				
	gyaraly	2						

Observe the next figure, and notice the influence of inaccurate pronunciation (or I should say foreign accent) which, through the one-to-one correspondence of the Japanese sounds to *katakana*, eventually evokes an unhelpful spelling in *romazi*. JWEFL, especially weak learners of English who are inclined to get confused with *romazi* spelling and English spelling tend to make idiosyncratic spelling errors that reflect strong influence from *romazi*. For example, the ambiguous Japanese sound either for *rubbish* or *lavish* can be transcribed at least in two ways, i.e. ラビッシュ in *katakana* (since all Japanese take this sound to be a borrowed, foreign word and, therefore, there is no possibility this sound is transcribed in *hiragana*) and *rabisyu* in *romazi*. Many JWEFL who do not have a reliable phonological clue to the correct English spelling would get confused in choosing between *r* or *l* and *b* or *v*; and they even lose their way in the two distinct *romazi* styles.

(5) JWEFL's route to English spelling



Appendix 1

AEMH-error.txt

Collection of non-word errors occurring in English essays written by Japanese university students whose major is English. (Corpus of English by Japanese Learners: by Koujiro Asao at Tokai University, Japan) (<http://www.lb.u-tokai.ac.jp/lcorpus/>)

EXAMS-error.txt

Errors made by 49 Japanese writers in Cambridge Certificate in English examinations included in Birkbeck Spelling Error Corpus (<http://ota.ahds.ac.uk/>)

HELC-JR-error.txt

Errors made in translation attempts by Japanese junior high-school students (Hiroshima English Learners' Corpus No.1: by Shogo Miura at Hiroshima University, Japan) (<http://home.hiroshima-u.ac.jp/d052121/eigo1.html>)

HELC-SR-error.txt

A Corpus Analysis of Spelling Errors Made by Japanese EFL Writers

Errors made in translation attempts by Japanese senior high-school students (Hiroshima English Learners' Corpus No.2: by Shogo Miura at Hiroshima University, Japan) (<http://home.hiroshima-u.ac.jp/d052121/eigo2.html>)

SAMANTHA-error.txt

Errors collected from an original spelling test of 53 target words given to 333 Japanese university students and adults.

(SAMANTHA Error Corpus: by Takeshi Okada at Yamagata University, Japan) (<http://www.e.yamagatau.ac.jp/~t.okada/docs/olp/Samantha/Samantha-top.html>)

SUZUKI-error.txt

Collection of spelling errors made by Japanese high-school students in their classroom activities (by Michiaki Suzuki at Nan'yo High School, Yamagata, Japan)

FRGRI-error.txt

Errors made by 88 Japanese university students not majoring in English (in a list in an article written in Japanese, 'Furugouri, T. and K. Hiranuma (1987) 'Statistical Characteristics of English Sentences Written by the Japanese and Detecting and Correcting Spelling-Errors,' *Mathematical Linguistics*, 16: 16-26.)

Breakdown of the seven JWEFL-made error subcorpora

	no of target words	no of error types	no of error tokens
AEMH-error.txt	234	296	396
EXAMS-error.txt	151	162	213
HELC-JR-error.txt	431	1921	3366
HELC-SR-error.txt	187	346	673
SAMANTHA-error.txt	53	2071	7418
SUZUKI-error.txt	43	46	46
FRGRI-error.txt	324	366	366
total	1423	5208	12478

Appendix 2

Original data files from which NSWE-made errors are collected

CHES.frq FAWTH1.frq FAWTH2.frq GATES.frq MASTERS.frq
 NFER1.frq PERIN.frq PERIN2.frq PERIN3.frq PETERS1.frq
 PETERS2.frq UPWARD.frq WING.FRQ

Breakdown of the thirteen NSWE-made error subcorpora

	no of target words	no of error types	no of error tokens
CHES.frq	30	1364	2474
FAWTH1.frq	739	809	809
FAWTH2.frq	484	557	1084
GATES.frq	3390	4401	144179
MASTERS.frq	264	13020	43755
NFER1.frq	40	495	838
PERIN1.frq	61	640	807
PERIN2.frq	538	625	658
PERIN3.frq	40	901	1678
PETERS1.frq	290	10556	18304
PETERS2.frq	1618	2576	4147
UPWARD.frq	576	753	1073
WING.frq	185	191	237
total	8255	36888	220043

Notes

- 1 For a detailed discussion on OCR-generated error detection and correction, see Okada (2000).
- 2 To keep the table to manageable proportion, substitutions with frequency less than 20 in the native speakers' errors have been excluded from Table 5.
- 3 On the contrary, the S-error of using *t* for *s*, which occurs in errors spelling *tion* in place of *sion*, can be regarded as NSW-specific error. There are only two target words with *sion*, i.e. *extension* and *television*. For these target words JWEFL make S-errors for *s* only twice: *extention* and *televition* once each. NSW make 79 S-errors for *extension* producing *extention*. However, the evidence is not conclusive, since 77 out of 79 of these are taken from the spelling tests.
- 4 It is notable that, whereas in these JWEFL errors the *l/r* substitution was the only error, in the NSW it was one of many.
- 5 Standard commercial Japanese word processors based on JIS X O213 (Japanese Industrial Standard) are generally equipped with an internal *kanji* dictionary with 6355 characters including 2967 characters mentioned here.
- 6 Therefore the name of the most popular Japanese mountain should be spelled out as (Mt) *Huzi* in the *Kun-rei* style, not as *Fuji*, which is far more familiar both to Japanese and foreign people. But the reality is slightly different. There is no /f/ sound in Japanese, and the first consonant of Mt Fuji is pronounced with Japanese *f* sound, which is a kind of bilabial-fricative sound.
- 7 In key-entering *romaji*, the dash “-” is usually a signal to lengthen its preceding vowel sound.

References

- Barron, R. W. (1980) 'Visual and Phonological Strategies in Reading and Spelling,' in U. Frith (ed.) *Cognitive Process in Spelling*. Academic Press. 195-213.
- Burt, J. S. and M. B. Fury (2000) 'Spelling in Adults: The Role of Reading Skills and Experience,' *Reading and Writing: An Interdisciplinary Journal*, 31: 1-30.
- Burt, J. S. and C. S. Shrubsole (2000) 'Processing of Phonological Representations and Adult Spelling Proficiency,' *Australian Journal of Psychology*, 52.2: 100-109.
- Burt, J. S. and H. Tate (2002) 'Does a Reading Lexicon Provide Orthographic Representations for Spelling?' *Journal of Memory and Language*, 46: 518-543.
- Ehri, L. C. (1980) 'The Development of Orthographic Images,' in U. Frith (ed.) *Cognitive Process in Spelling*. Academic Press. 312-338.
- Fairbairn, G. J. and C. Winch (1996) *Reading, Writing and Reasoning: A Guide for Students: Second Edition*. Open University Press.
- Frith, U. (ed.) (1980) *Cognitive Process in Spelling*. Academic Press.

- James, C. (1998) *Errors in Language Learning and Use*. Pearson Education.
- James, C. and K. Klein (1994) 'Foreign Language Learners' Spelling and Proof-Reading Strategies,' *Papers and Studies in Contrastive Linguistics*, 29: 31-46.
- Katamba, F. (1994) *English Words*. Routledge.
- Kress, G. (2000) *Early Spelling*. Routledge.
- Logan, F. A. (1999) 'Errors in Copy Typewriting,' *Journal of Experimental Psychology: Human Perception and Performance*, 25.6: 1760-1773.
- Mitton, R. (1996) *English Spelling and the Computer*. Longman.
- Okada, T. (2000) 'Remarks on Spelling Errors in OCR Generated Electronic Texts: With Reference to Misrecognition of Word Delimiter,' *Yamagata English Studies*, 5: 47-63.
- Okada, T. (2002) 'Remarks on Japanese Poor Spellers of English,' *Yamagata English Studies*, 7: 21-41.
- Rapp, B., C. Epstein and M. Tainturier (2002) 'The Integration of Information Across Lexical and Sublexical Processes in Spelling,' *Cognitive Neuropsychology*, 19.1: 1-29.
- Shuren, J. E., L. M. Maher and K. M. Heilman (1996) 'The Role of Visual Imagery in Spelling,' *Brain and Language*, 52: 365-372.
- Stanlaw, J. (2002) "'Hire" or "Fire"?: Taking AD-Vantage of Innovations in the Japanese Syllabary System,' *Language Science*, 24: 537-574.
- Venezky, R. L. (2002) 'In Search of the Perfect Orthography,' abstract of paper read at Third International Workshop on Writing Systems. University of Cologne.