Constructing a Database of Non-Japanese Pronunciations of Different Japanese Romanizations

Reiko Kaji, Hajime Mochizuki

Graduate School of Area and Culture Studies Tokyo University of Foreign Studies
3-11-1 Asahi-cho Fuchu-shi Tokyo, JAPAN 183-8534
kaji.reiko.g0@tufs.ac.jp, motizuki@tufs.ac.jp

Abstract

In this paper, we investigated how foreign language speakers pronounce Japanese words transliterated using two major Romanization systems, *Hepburn* and *Kunrei*. First, we recorded foreign language speakers pronouncing Romanized Japanese words. Next, Japanese speakers listened to the recordings and wrote down the words in Japanese *Kana*. Sets of each Romanized Japanese word, its correct *Kana* expression, its recorded reading, and the *Kana* dictated from the recording were stored in our database. We also investigated which of the two Romanization systems was pronounced more correctly by foreign language speakers by comparing the correctness of their respective readings. We also investigated which system's pronunciation by foreign language speakers was judged as more acceptable by Japanese speakers.

1. Introduction

The Roman alphabet has been adopted as an orthography by many languages, such as English, German, French and Italian among European languages; Swahili and Egyptian among African languages; and Indonesian and Vietnamese among Asian languages. There also exist languages which use the Roman alphabet subordinately, although they adopt a different type of orthography, such as Japanese, Chinese and Korean. Japanese-named entities that cannot be translated into a foreign language, such as a person's name, an organization name and a place name are transliterated into the Roman alphabet in order to express them in the context of foreign languages that have adopted the Roman alphabet. When Japanese characters, i.e., *Kanji* or *Kana*, are transliterated into the Roman alphabet, many Japanese speakers simply expect the words to maintain their same pronunciations between the Japanese characters and their transliterated target language Roman alphabet counterparts. However, even if the same Roman alphabet is used, depending on the language the pronunciation rules are may not be the same. Therefore if one does not know the pronunciation rules of a language, it is impossible to pronounce a Roman string correctly in that language. Japanese Romanization is no exception; if a person does not know the pronunciation rules for Romanized Japanese, then they will not be able to pronounce Romanized Japanese words correctly.

When we exchange information through text, the fact that the same Roman alphabet is pronounced differently in different languages is not a problem with respect to communication. It is still possible to identify such entities because their spelling is the same even if their pronunciation is different. The problem, then, occurs when we exchange information by oral communication without a text. If a Japanese speaker and a non-Japanese speaker communicate orally and without the aid of text, for example, it might be difficult for each of them to recognize the other person's pronunciation of some entities. In recent years, opportunities for oral communication between Japanese speakers and non-Japanese speakers have been increasing. This is partly the result of the increase in Japanese overseas tourists. Moreover, many foreigners who have visited Japan have had experiences in which the Japanese place names they pronounce when reading Japanese Romanized words are not understood by Japanese speakers. Though there are many systems currently in use to transliterate Japanese characters into the Roman alphabet, the major two transliteration systems are the *Kunrei* system and the *Hepburn* system. The *Kunrei* system was established by the Japanese government in 1937 (Unger, 1996) and is taught in Japanese elementary schools. On the other hand, the *Hepburn* system is taught in Japanese junior high schools, and became the more popular system after the General Headquarters/Supreme Commander for the Allied Powers (GHQ/SCAP) encouraged its use in representing railroad station names in Japan under the occupation after World War II (Kayanuma, 2000).

Until recently, the pronunciation of Romanized Japanese in other languages has not been given much attention. As a result, systematic research on the pronunciations of Romanized Japanese words by non-Japanese speakers has not been conducted, although language experts have long been aware that pronunciations for Romanized spellings are language dependent.

In this paper, we conduct experiments to investigate how Romanized Japanese words are pronounced by non-Japanese speakers. Japanese sounds consist of consonant and vowel pairs corresponding to *Kana* syllables. We first make a list of Roman substrings from the *Kana* expressions of Japanese sounds using the two transliteration systems, i.e., *Hepburn* and *Kunrei*. Each item on the list is then pronounced by the non-Japanese speakers, and their voices are recorded. Japanese speakers then listen to the recorded voices and back-transliterate the sounds into *Kana*. Finally, we construct a database by collecting the sets of Roman substrings, *Kana* expressions, pronounced voices and back-transliterated *Kana* from the experimental results.

Using this database, we then investigate how the Romanized Japanese substrings are pronounced by the non-Japanese speakers. From the viewpoint of equivalence of pronunciation between a foreign language and Japanese, we show the differences between the languages and also

compare the two Romanization systems.

Our work is related to previous studies on transliteration (Knight and Graehl, 1997; Xu et al., 2006) in the field of applied natural language Processing, including the areas of cross-language retrieval, information extraction, and machine translation. However, there is a difference with respect to the transliteration assumptions made in our work and those of others. Specifically, although transliteration is typically done according to the rules of the target language, in our study the Japanese transliteration to the Roman alphabet is done without any reference to a target language. Therefore, there is the possibility that speakers of a target language will be required to read a strange spelling that they have never seen before in their native language.

Our work also relates to the processing of out-of-vocabulary (OOV) in the text-to-speech domain (Divay and Vitale, 1997; Knight and Yamada, 1999). However, typical research in this domain would treat an OOV word as a misspelled word or a word not found in a dictionary, while we treat each substring as a correct Romanized Japanese word even if the substring seems to have a strange spelling by the non-Japanese speaker. We are not interested in which pronunciation is correct, but rather how each Roman substring is pronounced by non-Japanese speakers.

2. Japanese Romanization

2.1. Japanese Sounds

The sounds of Japanese are based on five vowels, *a*, *e*, *i*, *o*, *u*, which are used both alone or attached to either a consonant or a consonant plus a semi-vowel. Basic vowels change into long vowels by expanding the duration of their sound; that is, long vowels are distinguished in Japanese. The Japanese consonants are composed of both voiced and semi-voiced sounds in addition to the base unvoiced sounds. The phonology of Japanese adheres strongly to a consonant-vowel structure, which makes each pair a unit of sound. This unit is called a *mora*, which slightly differs from a *syllable* (Tsujimura, 1996). The phoneme of a single consonant is not recognized as a *mora* in Japanese except for the /N/ and glottal stop sounds. The *morae* of the /N/ and glottal stop sounds, however, never appear at the beginning of a word; they always follow another *mora.* Furthermore, except for loanwords, the glottal stop sound of modern Japanese is limited to consonants that begin with the letters *k*, *s*, *t* and *p*.

In summary, the patterns of Japanese sounds are as follows:

1. a vowel only (**V** structure),
2. a consonant and a vowel (**CV** structure),
3. a consonant plus a semi-vowel and a vowel (**CSV** structure)
4. V, CV or CSV plus the consonant /N/
5. V, CV or CSV plus the glottal stop sound of a consonant beginning with the letter *k*, *s*, *t*, or *p*.

2.2. Structure of *Kana*

Japanese uses three types of characters, *Kanji*, *Hiragana* and *Katakana*. *Kanji* refers to logographic characters of Chinese origin. *Hiragana* and *Katakana* are both syllabic characters that are commonly grouped together as simply *Kana.* We will not distinguish them here because they are used in almost the same way; instead, we use *Katakana* for all examples of *Kana.*

The Japanese *Kana* syllabary consists of about 80 symbols. As mentioned in Section 2.1., the *mora* is the basic Japanese sound unit. One or two *Kana* symbols are used to express one *mora*, as shown below:

1. V and CV structures are expressed by a single *Kana* symbol, such as ア for V and カ for CV.
2. A CSV structure is expressed by two *Kana* symbols, including small *Kana* symbols such as ャ, as in キャ.
3. The consonant /N/ is expressed by the symbol ン.
4. The glottal stop is expressed by the small symbol ッ.
5. A long vowel is composed of two *morae* which are expressed by either a *Kana* with the symbol ー, such as アー, or two of the same *Kana* symbols, such as アア.

2.3. Roman alphabets for Japanese

Roughly speaking, Japanese Romanization can be achieved by replacing the *Kana* strings with the Roman alphabet according to the transliteration rules of a Romanization system. In this study, we investigate the current two major Japanese Romanization systems, *Kunrei* and *Hepburn*. The two systems use different letters of the Roman alphabet for some consonants, but all of the same vowels are used.

Table 1 shows a list of the Japanese Roman alphabet used in the two systems.

Table 1: Inventories of the Roman alphabet

Vowels	a,e,i,o,u
Long vowels (A)	aa,ee,ei,ii,oo,ou,uu
Long vowels (B)	oh,ô,ō,û,ū
Consonants (Common)	b,d,g,h,k,m,n,p,r,s,t,y,w,z, by,gy,hy,ky,my,ny,py,ry
(Kunrei)	sy,ty,zy
(Hebpurn)	ch,f,j,sh,ts,tch

Vowels are expressed with five Roman letters, *a*, *e*, *i*, *o*, and *u*, in both systems. Consonants are composed of 31 strings, of which 22 are common between the two systems;3 strings are unique to *Kunrei*, and 6 strings are unique to *Hepburn.*

There are three variations in the expression of long vowels. One method is to use two vowels, such as the 'Long vowels (A),' shown in Table 1. This method is ambiguous because it is unclear whether the two vowels represent a long vowel or a two sequential vowels. However, any differences in the pronunciation of the two representations would not become a big problem when recognizing the words of named entities, because even if the two vowels are pronounced separately, they sound the same as a long vowel to most Japanese. Therefore, we treat the long vowels of 'Long vowels (A)' in Table 1 as two vowels in this study. The second method of expressing long vowels is to use any of the *macron*, *circumflex*, or *oh* marks, as shown in 'Long vowels (B)' in Table 1. We investigate the use of two marks to indicate long vowels in our experiment. Note that the *oh*, however, is ambiguous because it is unclear whether it indicates a long vowel or the vowel *o* plus the consonant *h*;

it is especially ambiguous when a vowel directly follows *h*, because *h* is also used as a consonant. Thus, we also investigate the *oh* with respect to a speaker's ability to distinguish the difference between a long vowel and a vowel plus a consonant. Although the *mora* 'ン' is normally transliterated as *n*, the *n* is changed to *m* in the *Hepburn* system when *b*, *m* or *p* follows it. This difference is also investigated in our experiment.
The glottal stop sound is expressed by doubling the consonant which follows the stop. In modern Japanese, with the exception of loanwords, this sound is limited to *morae* that begin with the consonants *k*, *s*, *t* or *p*. For example, the consonant *p* in *Sapporo* is the transliteration of サッポロ. As mentioned in Section 2.1., the consonant *n* and the glottal stop sound never appear at the beginning of a word. Therefore, we make temporary components by attaching a vowel to each for the purposes of our experiment. We explain these components in more detail in Section 4.2.

3. Selection of Languages

3.1. Comparison of Japanese and Other Languages

In theory, Japanese vowels can appear in succession without limitation. The pronunciation of each vowel basically does not change regardless of how many vowels are together in a string. For instance, three vowels often occur together, such as in the first name '*Aoi*,' for which the pronunciation is the same as '*A o i*' spelled separately. There are other languages, however, in which the pronunciations of vowels change when two or more vowels are connected. Such languages differ from Japanese in their tendencies of changing pronunciations. As mentioned in Section 2.3., in this study we investigate the continuousness of two vowels in our examination of long vowels. As for a greater number of vowels, it is difficult to investigate cases in which vowels continue more than three times using human subjects, because, for example, if we want to investigate a continuous string of n vowels out of five possible vowels, we have to consider 5^n variations. When n is three, then, the total number of possible spellings is 125. Thus, to construct our first database, languages that have tendencies of pronunciations relatively similar to Japanese are desirable. To this end, we first make a preliminary comparison of other languages with Japanese and classify them into the following three categories with respect to their vowel systems.

V0: Almost similar to Japanese
V1: A little different from Japanese
V2: Quite different from Japanese

We select languages which are classified into the categories V0 and V1 in this paper.
There also exist languages in which the pronunciations of the consonants change according to the characteristics of attached vowels. To investigate all such influences of vowels connected to consonants, therefore, we have to consider two types of vowel connections, that is, those both preceding and following consonants. The influence of vowels following consonants is relatively easy to investigate because all Japanese *morae* basically terminate with a vowel, as mentioned in Section 2.2. In contrast, the investigation of the influence of preceding vowels is more complicated because it requires the additional Roman spellings of phonemes different from those of minimum requirements in Japanese. We also classify languages by their consonants in our preliminary comparison; the classifications used for the pronunciation of consonants are as follows:

C0: No change
C1: Change influenced by the following vowel(s)
C2: Change influenced by the preceding vowel(s)
C3: Change depend on each word

Here, note that each language can be classified into both categories C1 and C2 because these classifications are not mutually exclusive.
We select the languages classified into category C0, C1 or C2 from among the languages belonging to the category V0 or V1. However, an additional experiment for consonants is needed for the languages classified into C2. Similarly, an additional experiment for vowels is also needed for the languages classified into V1. The languages classified into V2 or C3 are not considered in this paper.

3.2. Preliminary Comparison

We conduct a preliminary comparison on eight languages: English, French, German, Spanish, Italian, Czech, Swahili and Indonesian. Hereafter, these are described as *Eng*, *Fre*, *Ger*, *Spa*, *Ita*, *Cze*, *Swa* and *Ind*, respectively.
The procedure of the comparison is as follows:

1. Prepare basic words and their Japanese *Kana* expressions from an introductory level language text for each language. We prepare about 500 words each for *Swa* and *Ind*, and about 1500 words each for *Eng*, *Fre*, *Ger*, *Spa*, *Ita* and *Cze*. The *Kana* expressions are determined by the actual Japanese authors of each book for Japanese speakers who want to learn each language. For example the word 'pasta' in Italian is 'パスタ' in *Kana*, and the word 'amour' in French is 'アムール.'

2. Associate substrings of each word and the corresponding substrings of its *Kana* expression. For example, the words 'pasta' and 'パスタ' are each decomposed into three pairs of substrings, 'pa/パ,' 's/ス' and 'ta/タ.' The word 'amour' is decomposed into 'a/ア' and 'mou/ムー' and 'r/ル.'

3. Extract substring pairs that correspond to Japanese Romanization rules. For example, we extract the pairs 'pa/パ,' 'ta/タ,' 'a/ア' and 'mou/ムー.' The remaining pairs, 's/ス' and 'r/ル,' are not extracted because they do not have the CV structure mentioned in Section 2.1.

4. Analyze the consistency of pronunciations and spellings for each language. We make sure that there are pronunciation variations for some of the spellings in each language.

5. Classify all languages into the categories V0, V1 and V2 for vowels, and C0, C1, C2 and C3 for consonants.

Table 2: Result of preliminary comparison

Type	Eng	Fre	Spa	Ger	Ita	Cze	Ind	Swa
V0					○	○		○
V1			○				○	
V2	○	○		○				
C0	-	-	-	-	-	-	○	○
C1	-	○	○	-	○	○	-	-
C2	-	○	-	○	○	-	-	-
C3	○	-	-	-	-	-	-	-

Table 2 shows the results of the preliminary comparison. In this table, category **V0** includes *Ita*, *Cze* and *Swa*; **V1** includes *Spa* and *Ind*; and **V2** includes *Eng*, *Fre* and *Ger*. All of the languages in category V0 have the same vowel pronunciation tendencies as Japanese. In category V1, *Ind* has ambiguity of pronunciation of vowel *e*, and *Spa* has ambiguity of pronunciation for the sequenced vowels *ue* and *ui* when preceded by the consonant *g*. In category V2, all three languages have large differences with Japanese with respect to their vowel pronunciation tendencies. From these observations, we select *Spa*, *Ita*, *Cze*, *Ind* and *Swa* as the languages to be investigated in the rest of this paper.
Table 2 also shows the categories of consonants. Among the five selected languages, the category **C0** includes *Ind* and *Swa*; **C1** includes *Spa*, *Ita* and *Cze*; and **C2** includes *Ita*. In category C2, *Ita* has the ambiguity of consonant *s* tending to be unvoiced at the beginning of a word but voiced in other cases. An additional ambiguity of *Ita* is that the consonant letter *z* of the CV component *zi* tends to be voiced at the beginning of a word, although it is unvoiced when in the middle of a word. Furthermore, this CV component also tends to affect the pronunciation of the following vowels. Thus, this case is considered a combination of V1 and C1.
From all of the preliminary comparison observations, we design a basic experiment with human subjects for *Spa*, *Ita*, *Cze*, *Ind* and *Swa*, in addition to an additional experiment for the *e* of *Ind*, the *ue* and *ui* of *Spa*, and the *s* and *zi* of *Ita* in Section 4.

4. Construction of Database

To construct a database, we conduct experiments for the five languages of *Spa*, *Ita*, *Cze*, *Ind* and *Swa*. The database is formed in the following three steps:

1. Decide Roman spelling lists for investigation.
   We prepare two lists, a basic spellings list and an additional spellings list. The former is used for the two Japanese Romanization systems, *Hepburn* and *Kunrei*. The later is used for examining the languages categorized into V1 and C2.
2. Conduct experiments with human subjects.
   Pronunciations for each spelling in the lists made in step 1 according to native speakers of the five languages are recorded.
3. Update database.
   Japanese native speakers listen to the voices recorded voices in step 2; they then write down each word in *Kana* according to what they hear to the best of their ability. The database is updated by adding an entry composed of the Roman spelling, its correct *Kana* expression, its voice recording, and its dictated *Kana*.

4.1. Overview of Database

As an overview of the database, entry d_i is denoted by:

$$d_i = (R_i, K1_i, S_{f,i}, K2_{f,i}) \quad (1)$$

Here, R_i is the Roman spelling i, which is also expressed as 'an item' or 'a substring' in this paper; $K1_i$ is its correct *Kana* expression; $S_{f,i}$ is a link to the sound file of R_i as pronounced by a native speaker of foreign language f; $K2_{f,i}$ is the Japanese *Kana* expression as dictated by a Japanese native speaker who listened to the recording of $S_{f,i}$.
For example, when the foreign language f is Italian, the Roman spelling is 'ge,' its correct *Kana* is 'ゲ' the sound data is 'ge_{Ita}.mp3,' and the dictated Japanese *Kana* is 'ジェ.' The complete data entry d_i then becomes d_i =('ge','ゲ','$ge_{Ita}.mp3$','ジェ').
We can measure the similarity of pronunciation between the Romanized Japanese and language f by comparing all values of $K1_i$ and $K2_{f,i}$ in the database.

4.2. Decision of Roman Spelling lists for Examination

We prepare the following two Roman spelling lists:

- A basic spellings list that covers the two Japanese Romanization systems, *Hepburn* and *Kunrei*.
- An additional spellings list used to examine the languages categorized into V1 and C2.

4.2.1. Basic spellings list

The basic spellings list is composed of sets of Roman spellings that cover all variations of spellings used by the two Japanese Romanization systems, *Hepburn* and *Kunrei*. Based on the Japanese sounds described in Section 2.1. we make the basic spellings list as follows by combining the alphabets shown in Table 1,

- Basic Spelling Type 1: spellings of V, CV and CSV structures. This spelling type is consists of spellings in which the five vowels, *a*, *e*, *i*, *o*, and *u*, are used alone (V) or attached either to a consonant (CV) or a consonant plus a semi-vowel (CSV). The CV and CSV both have non-voiced, voiced and semi-voiced sounds. The total number of Type 1 basic spelling items is 114.
- Basic Spelling Type 2: spellings for all patterns of the sound /N/ that appear in Japanese. Although the sound /N/ is generally transliterated as *n*, the *n* is changed to *m* in the *Hepburn* system when followed by *b*, *m* or *p*. In order to include this point of view in our examination, we make spellings by adding each item of the Basic Type 1 spellings to the consonant *n* or *m*. Moreover, the sound /N/ never appears at the beginning of a word in Japanese, as mentioned in Section 2.1. Therefore, we attach the vowel *o* to the front of the items as a matter of convenience. The total number of Type 2 basic spelling items is 139.

Table 3: Basic spellings list

Basic 1: V, CV and CSV structures (114)	
Common	a,i,u,e,o,ka,ki,ku,ke,ko,sa,su,se,so,ta,te,to,na,ni,nu,ne,no, ha,hi,he,ho,ma,mi,mu,me,mo,ya,yu,yo,ra, ri,ru,re,ro,wa,ga,gi,gu,ge,go,za,zu,ze,zo,da,de,do, ba,bi,bu,be,bo,pa,pi,pu,pe,po,kya,kyu, kyo,nya,nyu,nyo,hya,hyu,hyo,mya,myu,myo, rya,ryu,ryo,gya,gyu,gyo,bya,byu,byo,pya,pyu,pyo
Kunrei	si,ti,tu,hu,zi,sya,syu,syo,tya,tyu,tyo,zya,zyu,zyo
Hepburn	shi,chi,tsu,fu,ji,sha,shu,sho,cha, chu,cho,ja,ju,jo
Basic 2: All sound /N/ patterns (139)	
Common	on,ona,oni,onu,one,ono,onka,onki,onku,onke,onko,onsa, onsu,onse,onso,onta,onte,onto,onna,onni, onnu,onne,onno,onha,onhi,onhe,onho,onya,onyu,onyo, onra,onri,onru,onre,onro,onwa, onga,ongi, ongu,onge,ongo,onza,onzu,onze,onzo, onda,onde,ondo, onkya,onkyu,onkyo,onnya,onnyu,onnyo, onhya,onhyu,onhyo,onrya,onryu,onryo,ongya,ongyu, ongyo
Kunrei	onshi,onchi,ontsu,onfu,onji,omma,ommi,ommu,omme, ommo,omba,ombi,ombu,ombe,ombo,ompa, ompi,ompu,ompe,ompo,ommya,ommyu,ommyo,ombya,ombyu, ombyo,ompya,ompyu,ompyo
Hepburn	onsi,onti,ontu,onhu, onzi,onma,onmi,onmu,onme, onmo,onba,onbi,onbu,onbe,onbo,onpa, onpi,onpu,onpe,onpo,onmya,onmyu,onmyo,onbya,onbyu,onbyo, onpya,onpyu,onpyo
Basic 3: All glottal stop sound patterns (41)	
Common	okka,okki,okku,okke,okko,ossa,ossu,osse, osso,otta,otte,otto,oppa,oppi,oppu,oppe, oppo,okkya,okkyu,okkyo,oppya,oppyu,oppyo
Kunrei	ossi, otti,ottu,ossya,ossyu,ossyo,ottya,ottyu,ottyo
Hepburn	osshi,otchi,ottsu,ossha,osshu,ossho,otcha,otchu,otcho
Basic 4: Two sequential vowels (25)	
Common	aa,ai,au,ae,ao,ia,ii,iu,ie,io,ua,ui,uu, ue,uo,ea,ei,eu,ee,eo,oa,oi,ou,oe,oo
Basic 5: Long vowels *o* and *u* with macron (2)	
Common	ō,ū
Basic 6: Long vowels *o* and *u* with circumflex (2)	
Common	ô,û
Basic 7: All patterns of long vowels *o* expressed by *oh* (115)	
Common	oha,ohi,ohu,ohe,oho,ohka,ohki,ohku,ohke,ohko,ohsa, ohsu,ohse,ohso,ohta,ohte,ohto,ohna,ohni,ohnu, ohne,ohno,ohha,ohhi,ohhe,ohho,ohma,ohmi,ohmu,ohme, ohmo,ohya,ohyu,ohyo,ohra,ohri,ohru,ohre, ohro,ohwa,ohga,ohgi,ohgu,ohge,ohgo,ohza,ohzu,ohze,ohzo,ohda, ohde,ohdo,ohba,ohbi,ohbu,ohbe, ohbo,ohpa,ohpi,ohpu,ohpe,ohpo,ohkya,ohkyu,ohkyo, ohnya,ohnyu,ohnyo,ohhya,ohhyu,ohhyo,ohmya, ohmyu,ohmyo,ohrya,ohryu,ohryo,ohgya,ohgyu, ohgyo,ohbya,ohbyu,ohbyo,ohpya,ohpyu,ohpyo
Kunrei	ohsi,ohti,ohtu,ohhu,ohzi,ohsya,ohsyu,ohsyo,ohtya,ohtyu,ohtyo, ohzya,ohzyu,ohzyo
Hepburn	ohshi,ohchi,ohtsu,ohfu,ohji,ohsha,ohshu,ohsho,ohcha,ohchu,ohcho, ohja,ohju,ohjo

- Basic Spelling Type 3: spellings for all patterns of the glottal stop sound. The glottal stop is limited to consonants that begin with the letters *k*, *s*, *t* and *p*, as mentioned in Section 2.3. We choose items that begin with these letters from the Basic Type 1 items. Next, we make additional spellings by repeating their first letters. Similar to the sound /N/, the glottal stop sound never appears at the beginning of a word in Japanese. Therefore, we attach vowel *o* to the front of these items as a matter of convenience. The total number of Type 3 basic spelling items is 41.
- Basic Spelling Type 4: two sequential vowels. The total number of Type 4 basic spelling items is 25.
- Basic Spelling Type 5: long vowels *o* and *u* with a macron. The total number of Type 5 basic spelling items is 2, $\bar{o}$ and $\bar{u}$.
- Basic Spelling Type 6: long vowels *o* and *u* with a circumflex. The total number of Type 6 basic spelling items is 2, $\hat{o}$ and $\hat{u}$.
- Basic Spelling Type 7: spellings for all patterns of long *o* vowels expressed by *oh*. This spelling type is composed of *oh* and *oh* followed by each Type 1 basic spelling. The total number of Type 7 basic spelling items is 115.

The total number of basic spelling items is 438, as shown in Table 3.

4.2.2. Additional spellings list

The basic spellings list alone is not sufficient to investigate languages categorized into V1 or C2; thus an additional spellings list is necessary. The result of the preliminary comparison described in Section 3.2. is summarized as follows:

- *Ind* and *Spa* are classified as V1. *Ind* has ambiguity of pronunciation for vowel *e* and *Spa* has ambiguity of pronunciation for the sequenced vowels *ue* and *ui* when preceded by consonant *g*.
- *Ita* is classified into C2. *Ita* tends to affect the pronunciations of the consonant *s* according to its position in a word.
- *Ita* has an additional ambiguity. Consonant *z* of *zi* tends to change pronunciation according to the position of *zi* in a word and the following vowel.

Based on the results of the preliminary comparison, we make an additional spellings list as follows.

- Additional Spellings 1 (*Spa*): *g* plus *ue*, and *g* plus *ui*.
- Addition Spellings 2 (*Ita*): all spellings of *s* preceded by each of the five vowels.
- Addition Spellings 3 (*Ita*): all spellings of *zi* followed by each of the five vowel. In order to investigate *zi* in the middle of a word, we attach vowel *o* to the front of each word as a matter of convenience.

We prepare the additional spellings list as shown in Table 4.

Table 4: Additional spellings list

	Lang	Spellings
Addition 1	Spa	gui, gue
Addition 2	Ita	asa,isa,usa,esa,osa
Addition 3	Ita	ozi,ozia,ozie,ozii,ozio,oziu

There is no entry for *e* for *Ind* in Table 4 because the pronunciation of vowel *e* would be changed according to each word, and there is no simple method for investigating the tendency of changing pronunciations. Fortunately, it can be expected that the ambiguity of *Ind* is limited to this vowel *e*. Therefore, we ask the subjects for *Ind* their mental pronunciation rules for *e*.

4.3. Experiments

To construct the database, we conduct two experiments using human subjects. Experiment 1 uses the basic spellings list covering all five target languages. Experiment 2 uses the additional spellings list covering the *Ita* and *Spa* items categorized as V1 and C2.

4.3.1. Experiments for the basic spellings list

We conduct experiment 1 for the basic spellings list in the five target languages, *Spa*, *Ita*, *Cze*, *Ind* and *Swa*. The procedure of the experiment is as follows.

1. Each native speaker of each language is shown the basic spellings list in Table 3. We use one speaker for each language, for a total of five speakers.
2. Each subject reads each entry of the list three times. The subjects are allowed to make two or more different pronunciations, if they feel to be able to pronounce a spelling. To avoid the pronouncing of items as a foreign language, each subject is instructed to pronounce the items according to their general knowledge and conventions of their own language. The subjects' pronunciations are recorded.

4.3.2. Experiments for the additional spellings list

We conduct experiment 2 for the additional spellings list in the two languages of *Spa* and *Ita*. The procedure of the experiment is as follows.

1. A Spanish native speaker and an Italian native speaker are shown each related entry in the additional spellings list in Table 4. The subjects are the same as those of experiment 1.
2. Each subject reads aloud each related entry of the list three times. This procedure is the same as that of experiment 1. The subjects' pronunciations are recorded.

4.4. Updating the database

The recorded voices from experiment 1 and experiment 2 are converted to their *Kana* expressions by a Japanese native speaker (one of the authors). The procedure for this step is as follows. The Japanese native speaker:
• listens to the recorded voices,
• tries to hear each spoken sound as a Japanese sound,
• writes each sound down as a *Kana* string.

After this procedure is performed for each word, it is possible to make a data set for each item consisting of its Roman spelling, its corresponding *Kana* representation, a recorded voice file of the human subject's pronunciation of the item, and its *Kana* as heard by a native Japanese speaker. We then update our database by adding each such set to the database as d_i of equation (1).
If a subject reads a Romanized spelling using two or more different pronunciations, then we record all pronunciations in the same sound file. Then, the *Kana* expressions for each of the pronunciations are stored in $K_{f,i}$ as a list.

5. Evaluation

In this section, we investigate how the non-Japanese speakers pronounce Japanese transliterated by the two major Romanization systems, *Hepburn* and *Kunrei*. We also investigate which of the two systems better yields a pronunciation acceptable by a native Japanese speaker. To this end, we use the database constructed in Section 4. and compare $K1_i$ and $K2_{f,i}$ of equation (1) using the two evaluation criteria described below.

5.1. Evaluation criteria

Each entry of the database d_i is evaluated according to whether the pronunciation of R_i in d_i by language f can be judged as an acceptable Japanese pronunciation by comparing $K1_i$ and $K2_{f,i}$. We judge the acceptance by the following two evaluation criteria:

- Criterion 1 (**CT1**): If $K1_i$ and $K2_{f,i}$ are the same, the d_i of language f is judged to be acceptable.
- Criterion 2 (**CT2**): If $K1_i$ and $K2_{f,i}$ are the same, the d_i of language f is judged to be acceptable. Otherwise, d_i is also judged to be acceptable, assuming $K2_{f,i}$ can be considered to be a kind of dialect of $K1_i$.

5.2. Evaluation of the basic spellings list

For the evaluation of the basic spellings list, we should distinguish the long vowels of the Basic Types 5, 6 and 7 from the others. Because they are the variations in the Basic Type 4 items, and are never used at the same time, items of these types should not be included. We thus show these results separately. The results of the Basic Types 1 to 4 are shown in Table 5. For CT1, in the table, each cell shows the number of acceptable items divided by the total items judged, although only the number of acceptable numbers are shown for CT2 because of lack of space. C, H and K refer to the *C*ommon part, *Hepburn* only part, and *Kunrei* only part, respectively. The ratio of acceptance is also shown in Table 6. Each ratio is calculated from Table 5 as the total ratio of an acceptance for a common part, and for each system.

Table 5: Numbers of acceptable Basic Type 1 to 4 items for Criteria 1 and 2

		Criterion 1					Criterion 2				
		Basic 1	Basic 2	Basic 3	Basic 4	Total	Basic 1	Basic 2	Basic 3	Basic 4	Total
C	C	85/86	57/63	23/23	25/25	190/197	85	57	23	25	190
z	H	2/14	26/38	1/ 9	-	29/ 61	2	26	29	-	29
e	K	1/14	25/38	0/ 9	-	26/ 61	13	37	8	-	58
I	C	86/86	47/63	23/23	25/25	181/197	86	47	23	25	181
n	H	13/14	37/38	9/ 9	-	59/ 61	13	37	9	-	59
d	K	7/14	31/38	3/ 9	-	41/ 61	14	38	9	-	61
I	C	77/86	39/63	23/23	25/25	164/197	77	39	23	25	164
t	H	10/14	31/38	5/ 9	-	46/ 61	10	31	5	-	46
a	K	0/14	24/38	0/ 9	-	24/ 61	13	36	9	-	58
S	C	69/86	35/63	23/23	18/25	145/197	69	35	23	18	145
p	H	8/14	30/38	9/ 9	-	47/ 61	9	30	9	-	48
a	K	1/14	25/38	0/ 9	-	26/ 61	10	34	3	-	47
S	C	81/86	54/63	4/23	25/25	164/197	81	54	4	25	164
w	H	14/14	38/38	4/9	0/0	56/61	14	38	4	-	56
a	K	1/14	25/38	0/9	0/0	26/61	14	38	1	-	53
A	C	79.6/86	46.4/63	19.2/9	18.6/25	163.8/197	79.6	46.4	19.2	23.6	168.8
V	H	9.4/14	32.4/38	5.6/9	0.0/0	47.4/61	9.6	32.4	5.6	-	47.6
E	K	2.0/14	26.0/38	0.6/9	0.0/0	28.6/61	12.8	36.6	6.0	-	55.4

Table 6: Ratios of accepted pronunciations

	Cze	Ind	Ita	Spa	Swa	ave
Criterion 1						
H	84.9	93.0	81.4	74.4	85.3	83.8
K	83.7	86.0	72.9	66.3	73.6	76.5
Criterion 2						
H	84.9	93.0	81.4	74.8	85.3	83.9
K	96.1	93.8	86.0	74.4	84.1	86.9

For example, the acceptance ratio of H of *Cze* for CT1 is 84.9%, which is calculated as $(190+29)/(197+61)$. The numbers of acceptable long vowels for Basic Types 5 to 7 for CT1 are shown in Table 7. The result for CT2 was the same.

Table 7: Long vowel results

	Cze	Ind	Ita	Spa	Swa	ave.
Basic 5(ō,ū)	2	0	2	0	2	1.2/2
Basic 6(ô,û)	1	0	0	0	2	0.6/2
Basic 7(*oh*)	0	0	0	0	0	0/115

5.2.1. Comparison of Hepburn and Kunrei systems

As can be seen in Table 6, for CT1, the *Hepburn* system has a higher percentage of acceptance than the *Kunrei* system for all five languages. The *Hepburn* acceptance rate was 83.8%, while that of *Kunrei* was 76.5% on average. These results suggest that pronunciations according to the *Hepburn* system are more acceptable as Japanese pronunciations than those made according to the *Kunrei* system. In contrast, for CT2, the *Kunrei* system shows a higher percentage than the *Hepburn* system for *Cze*, *Ind* and *Ita*. *Kunrei* gives an acceptance rate of 86.9%, while *Hepburn* yields only 83.9% on average. Further, if Japanese speakers accept some pronunciations as a dialect variation, the ratio of understandable Japanese produced based on the *Kunrei* system is higher than that based on the *Hepburn* system on average. However, the degree of difference between the two systems in this case is not large. Additional results from the experiments are summarized below:

- There was no difference observed between *m* in the *Hepburn* and *n* in the *Kunrei* systems when followed by *m*, *b* or *p*.
- Pronunciations of *ti* in the *Kunrei* and *chi* in the *Hepburn* systems were both not acceptable in *Cze*.
- Pronunciations of *zi* in the *Kunrei* and *ji* in the *Hepburn* systems were both not acceptable in *Spa*.
- Pronunciations of *ssya*, *ssyu*, *ssyo*, *ttya*, *ttyu* and *ttyo* in the *Kunrei* system did not become a glottal stop sound in *Spa*.
- Pronunciations of *ssya*, *ssyu* and *ssyo* in the *Kunrei* system became different sounds in *Swa*.

5.2.2. Comparison among the five languages

The best scores of each language are shown in Table 8.

Table 8: Best results among the five languages

CT1	Ind_H 93.0	Swa_H 85.3	Cze_H 84.9	Ita_H 81.4	Spa_H 74.4
CT2	Cze_K 96.1	Ind_K 93.7	Ita_K 86.0	Swa_H 85.3	Spa_H 74.8

For CT1, the *Hepburn* Romanizations in *Ind* (Ind_H) were the highest, at 93.0%; Swa_H and Cze_H were relatively high, at greater than 80%; and Spa_H was the worst, at 74.4%. For CT2, the *Kunrei* Romanizations in *Cze* (Cze_K) were the highest, at 96.1%; Ind_K were also very high, at 93.7%; Ita_K and Swa_H were relatively high, at more than 80%; and Spa_H was the worst, at 74.8%. The acceptance ratio of *Spa* was the lowest for both CT1 and CT2.

Examples of items that were not judged as acceptable for CT2 are summarized as follows:

- *Ind*: *ng* plus *a*,*i*,*u*,*e*, *o*,*ya*,*yu*,*yo* and *tsu*
- *Cze*: *wa*, *ti*, *shi*, *chi* and *ji*
- *Swa*: a vowel was pronounced as a long vowel, and a glottal stop sound was not pronounced.
- *Ita*: *chi*; *h* plus *a*,*i*,*u*, *e*,*o*,*ya*,*yu*,*yo*; *g* plus *i*,*e* and *ns* plus *a*,*i*,*u*,*o*
- *Spa*: *y* plus *a*,*u*,*o*, *r* plus *a*,*i*,*u*,*e*,*o*; *g* plus *i*,*e*,*ya*,*yu*,*yo*; *z* plus *a*,*i*,*u*,*e*,*o*; *tsu* and *j* plus *a*,*i*,*u*,*o*

5.2.3. Comparison of long vowels

From these results, it can be concluded that the long vowel mark, *macron*, was more acceptable than the *circumflex* in *Cze* and *Ita* because Basic Type 5 item acceptance rates were higher than Basic Type 6 rates (Table 7). In *Spa*, *Ind* and *Swa*, the pronunciations did not change regardless of whether a mark was added, because these three languages do not use such marks. However, there was a difference among the three languages in that the *Spa* and *Ind* speakers pronounced a vowel as a short vowel, while the *Swa* speaker pronounced it as a long vowel by default. Therefore, the pronunciations of the *Swa* speaker were judged as acceptable as a long vowel by chance. The long vowels expressed by *oh* were not judged as acceptable for all 115 items of Basic Type 7 for all five languages.

These results suggest that when we Romanize a long vowel, we should use the *macron* in *Ita* and *Cze*, and two vowels in *Spa*, *Ind* and *Swa*; however, the *oh* should not be used in any of the five languages.

5.3. Evaluation of the additional spellings list

The results of the experiment for the additional spellings list for *Spa* and *Ita* is summarized below:

- Addition 1: In Japanese, *ue* of *gue* and *ui* of *gui* are both pronounced separately as two vowels. In *Spa*, although these vowel combinations were both treated as one vowel in the preliminary comparison, in the experiment their pronunciations were the same as their Japanese pronunciations.
- Addition 2: In Japanese, *s* is an unvoiced sound. In *Ita*, although it became both a voiced and an unvoiced sound in the preliminary comparison, in the experiment the pronunciations were the same as the Japanese pronunciation, that is, unvoiced.
- Addition 3: In Japanese, the *z* of *zi* is a voiced sound. In *Ita*, in both the preliminary comparison and in the experiment, it was an unvoiced sound.
- In *Ind*, *e* had ambiguity of pronunciation (two different pronunciations) in the preliminary comparison. In response to our questioning, the subject answered that if a word was recognized as a loan word, it would be pronounced differently, i.e., the same as Japanese *e*.

Finally, it should be noted that when Italian speakers pronounce the *z* of *zi*, their pronunciation is different from the normal Japanese pronunciation. The results of the other items in the additional experiments were contrary to the preliminary comparison; that is, they were pronounced the same as they are in Japanese in the experiments.

6. Conclusion

In this paper, we constructed a database to record the pronunciations of Romanized Japanese strings through experiments on non-Japanese speakers. Before the experiment, to decide the strings and languages that should be investigated, we conducted a preliminary comparison using the basic words of eight foreign languages and their corresponding *Kana* expressions. Based on the results of this comparison, we conducted experiments on *Ind*, *Cze*, *Ita*, *Spa* and *Swa*, and then constructed the database. Using the database, we investigated how non-Japanese speakers pronounced Romanized Japanese transliterated according to both the *Hepburn* and *Kunrei* systems. For the five languages, the order of the overall acceptance ratios, from highest to lowest, was Cze_K, Ind_K, Ita_K, Swa_H and Spa_H. When pronunciations approximating a dialect were also accepted, the average ratio of understandable Japanese was higher for the Romanizations based on the *Kunirei* system than those based on the *Hepburn* system. However, the degree of difference between the two systems was not large. The languages classified into the category V2, *Eng*, *Fre* and *Ger* were not investigated in this paper. Future research should conduct experiments in these languages. It is also necessary to conduct additional experiment to investigate languages classified into categories C2 and C3.

In our future research, we plan to construct a naming support system as an application of the database described in this paper. For example, using the database, the possible spellings of a new product name can be investigated from the viewpoint of international pronunciation equivalence.

7. Acknowledgments

This study was supported by the Global COE program "Corpus-based Linguistics and Language Education" at the Tokyo University of Foreign Studies.

8. References

M. Divay and A. J. Vitale. 1997. Algorithms for grapheme-phoneme translation for english and french: Applications for database searches and speech synthesis. *Computational Linguistics*, 23(4):495–523.

A. Kayanuma. 2000. *Research for Romanization of Japanese Orthography (in Japanese)*. Kazama-shobo.

K. Knight and J. Graehl. 1997. Machine transliteration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 128–135.

K. Knight and K. Yamada. 1999. A computational approach to deciphering unknown scripts. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 37–44.

N. Tsujimura. 1996. *An introduction to Japanese Linguistics*. Blackwell Publishers Inc.

J. M. Unger. 1996. *Literacy and Script Reform in Occupation Japan Reading between the Lines*. Oxford University Press.

L.L. Xu, A. Fujii, and T. Ishikawa. 2006. Modeling impression in probabilistic transliteration into chinese. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 242–249.