

A New *Kana*

A Modest Proposal Based on Statistical Analysis of Chinese Loanword Pronunciations

ROBERT T. MYERS*

Even omitting radical proposals to replace Japanese entirely with some more congenial language such as French², a number of proposals for improving or modifying the Japanese writing system have been made.³ Several propose going to roman characters. Others suggest switching entirely to *kana*, abolishing *kanji*. This proposal outlines a new, third set of *kana* which may solve many of the problems with earlier proposals.

Introduction

It is assumed the reader is familiar with the Japanese writing system. Presently, *hiragana* are used for conjunctions, endings and other structural words. In addition, they are used to write Chinese loanwords with whose *kanji* it is expected the reader is not acquainted. *Katakana* are used primarily for foreign loanwords. *Kanji* are used for the great majority of native Japanese terms (in the case of verbs and adjectives, their stems), as well as for almost all Chinese loanwords.

Both *kanas* can be viewed as alphabets (although they are in fact syllabaries), and can be learned, written and processed quite easily. Not so for *kanji*. Japanese schoolchildren spend a large proportion of their school hours learning the first 2000 or so characters during the twelve years of primary school. Even so, it is common for Japanese adults to fail to read *kanji* and even more so for them to be stumped when it comes to trying to write one. Mistakes in writing *kanji* are commonplace.⁴ It appears that writing *kanji* in general takes more time than writing the equivalent word in *kana*, even though *kana* is not particularly optimized, as will be discussed below, for writing Chinese loanword pronunciations. Computer processing of *kanji*, although progressing rapidly, is still far less fluid than that of alphabets. In particular, inputting the same text in *kanji* as in roman characters is likely to take several times as long.

These are simple arguments, and one should not take lightly the prospect of abolishing or radically modifying a writing system which has been in use over one thousand years. Specifically, it is said that learning *kanji* constitutes not only learning a writing system, but also learning an entire etymological body of knowledge, or that learning *kanji* is equivalent to learning a language such as English plus its forebear Latin. There is anecdotal evidence that the comprehension ratio when reading *kanji* declines less rapidly as reading speed increases than when reading alphabets. Finally, *kanji* serve as a sort of pidgin common language between the Asian cultures of Japan, China and to a lesser extent Korea.

*Department of Information Science, Faculty of Science, University of Tokyo.

²"Japan might as well, at this juncture, adopt French as her national language" was the proposal made by Shiga Naoya in April, 1946. See Haruhiko Kindaichi, "The Japanese Language", translated and annotated by Umeyo Hirano, Charles E. Tuttle Co., Inc., 1978. Originally published as "Nippongo" by Iwanami Shoten, Tokyo, 1957. Also, Arinori Mori, a Meiji educator, is quoted as writing "All reasons suggest its disuse" concerning Japanese in his correspondence and suggested replacing Japanese by English.

³Including the well-known proposal by Allied GHQ after the war to adopt roman characters.

⁴"The Japanese writing system is without question, the most complicated and involved system of script employed today by any nation on earth; it is also one of the most complex orthographies ever employed by any culture anywhere at any time in human history." Roy Andrew Miller, "Nihongo", The Athlone Press, 1986, p.1.

However, in this paper we will take it as a given that it is a worthwhile goal to abolish the use of *kanji* in writing Japanese and proceed to consider how to accomplish this goal.

Function and Characteristics of *Kana* and *Kanji*

We first consider some aspects of the use of the three notations in modern Japanese. In particular, one important aspect of the commingled writing system is the way it accomplishes setting words apart. This is all the more critical since spaces are not generally used. Foreign loanwords are clearly delimited by virtue of being written in *katakana*. Chinese loanwords, most often composed of two *kanji*, are in general both preceded and followed by native Japanese particles written in *hiragana* and thus stand out. And native Japanese nouns, verbs and adjectives again are usually preceded by *hiragana* particles of some sort and followed by additional *hiragana* particles or endings. Thus the three writing systems provide important visual clues to the reader helping him or her to perform the first level lexical scanning of the sentence.

Thus the first problem raised by considering a new writing system is how to accomplish the same type of delineation. Converting entirely to *hiragana*, the simplest alternative, for Chinese loanwords and native Japanese words now written in *kanji*, would result in a sea of *hiragana* quite hard to decipher.

In addition, consider that currently approximately 5000 characters are successfully representable and distinguishable in the space of a single character. As long as the proportion of *hiragana* is relatively small, it is acceptable to use an entire character space to represent a single phonetic syllable, but to extent *hiragana* is used to a much greater degree, it becomes a waste of the potential information content of that character space. Another way of saying this is simply that all written materials would become much longer.

Finally, and related to the above point, is the fact that *hiragana* (or *katakana*, as both represent the same underlying alphabet) is poorly suited to representing the pronunciations of Chinese loanwords in particular. As will be shown in the analysis below, only 20% of characters used in Chinese loanwords can be written in a single *kana*. The majority require two, with some needing up to four. A word represented in two Chinese characters thus takes an average of five *kana*. This is a simple result of the fact that *kana* were designed with native Japanese in mind.

A New, "Third" *Kana*

The above points lead naturally to the concept of a new writing system, or "third" *kana*, which packs more information into one character space than do *kana*, which is optimized with regard to Chinese loanword pronunciation, and which by definition will accomplish the goal of helping the delimit the component parts of a sentence.

The reader is most likely acquainted with the Hangul system used in Korea⁵. Originally hangul were developed in much the same way and for the same reason as *kana*. However, it appears that the intent with hangul was from the start to use them in representing both native Korean words as well as Chinese loanwords. Hence, hangul are already optimized for Chinese loanwords. In fact, any Chinese loanword character can always be written with exactly one hangul mark. Thus the proliferation of hangul occurred with relative ease, although not until after World War II. It is also worth pointing out that hangul contain some structural elements and some silent components and in that way is adapted specifically to Korean grammar and in particular the way verbs are inflected. With the exception of the fact that we do not propose writing native Japanese using the third *kana*, since *kana* are already well suited for that purpose, our third *kana* resembles hangul closely, in particular with regard to the information content of a single character — namely that

⁵Grant, Bruce K., *A Guide to Korean Characters*, Hollym International Corp., Seoul, 1979, or similar reference.

amount of information required to represent the Chinese loanword pronunciation of a single *kanji* character.

The proposal here then is as follows. Foreign loanwords are to continue to be represented in *katakana*. Japanese structure particles and endings will also continue to be written in *hiragana*. Chinese loanwords alone are to be written using the proposed third *kana*. The question remains of how to write native Japanese nouns and verb and adjective stems currently using *kanji*. Our proposal is to go to *hiragana* for such words. Of course this leads to the disadvantages raised above for switching to *hiragana* completely. However, these words are usually representable in a relatively few number of *kana*, one in a fair number of cases. They are often well distinguished in written materials by preceding object particles, for instance, or by characteristic endings. In fact, a trend is already visible to write such words, in particular generic ones such as "*iku*, go" or "*kuru*, come", in *hiragana*. This proposal is summarized in Fig. 1 below.

<u>Type of Word</u>	<u>Example</u>	<u>Current</u>	<u>Proposed</u>
Native Japanese Words	kaeru	Kanji	Hiragana
Chinese Loanwords	kitaku	Kanji	New Kana
Foreign Loanwords	rita-n	Katakana	Katakana
Endings, Prepositions, Particles de, masu		Hiragana	Hiragana

Fig. 1. Proposed Way to Write Various Word Types

A minor trend is already visible in modern Japan to use *kanji* as pseudo-phonetic elements. One may point to the everyday abbreviation used for the "ki" portion of "*kikai*, machine", which is composed of the tree radical *ki*-hen present in the original character followed by the *katakana* letter "ki". In another example, the "tai" of "*taifuu*", typhoon", which originally contained a wind radical, has now been simplified to another character with the same pronunciation. Finally, the character "*hatsu*" written with the hand radical *te*-hen has also been simplified as is now simply written with the *te*-hen-less version. In principle, all these cases represent use of *kanji* as information elements intermediate between the content present in *kanji* (full meaning plus pronunciation) and *kana* (pronunciation only).

Our goals in developing the new syllabary were as follows.:

- (1) Aesthetically, the characters should conform to the overall types of patterns and rhythms the Japanese are accustomed to
- (2) They should be clearly distinguishable from both *hiragana* and *katakana*
- (3) Of course, they should contain enough redundancy to be easily distinguished from each other and to survive some level of distortion à la fax machine
- (4) They must be able to represent the pronunciation of one Chinese loanword character in one character of the third *kana*.
- (5) The simplest characters should be reserved for the most common pronunciations and the more complex for the rarer ones
- (6) Statistical idealism should not be sacrificed to orthogonality. In other words, where initial sounds, vowels and final sounds can be represented by the same or similar symbol elements in a variety of

combinations, this is to be considered preferable to a minor gain in simplicity

Pronunciation Frequency Analysis

It turns out the most common pronunciation of such Chinese loanword characters is "kou." This is the result of the comprehensive frequency analysis we performed and describe in more detail below. This pronunciation would typically be written in *kana* as the character for "ko" followed by that for "u", amounting to a total of two characters. We see immediately, therefore, how poorly in fact *kana* are suited for writing such pronunciations, in addition to the fact that we must choose a very non-*kana* like approach for designing our new characters.

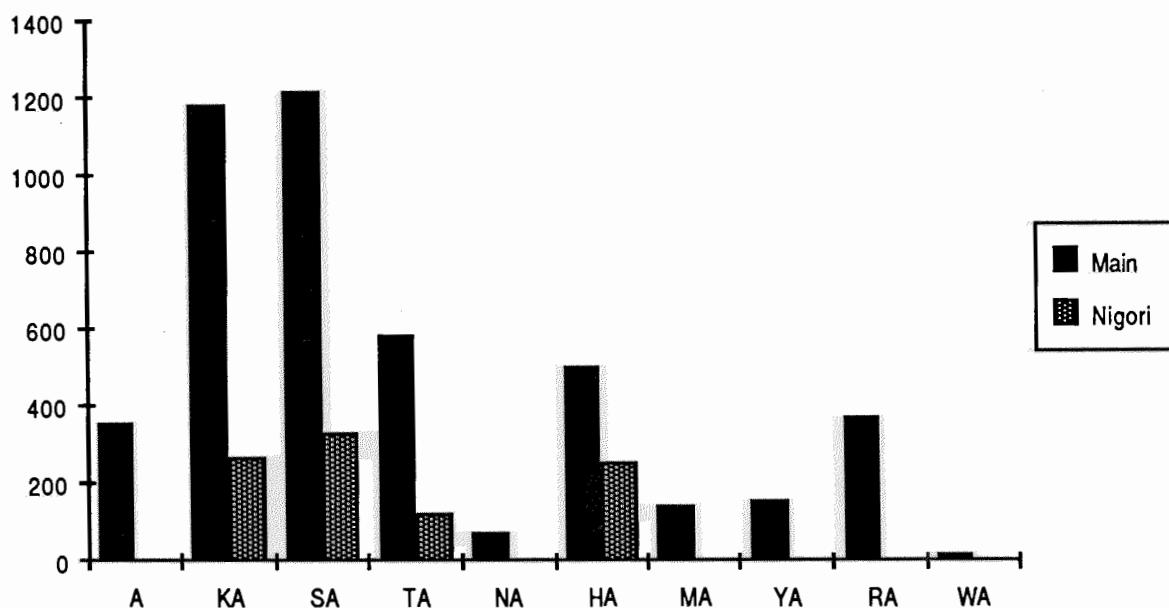
Before proceeding to giving the results of the analysis, we describe its process. We took as the universe all characters in a well known Japanese-English character dictionary.⁶ This dictionary is essentially equivalent in its coverage to popular Japanese-Japanese character dictionaries. Adding more characters to the universe would not only probably not change the frequencies we derived substantially, and to the extent the additional characters are extremely rare would also be irrelevant. This raises the question of the relative usage frequency of the characters that were included in the universe. Our results may be biased to the extent a large number of extremely rare characters have the same of closely related pronunciations. We assume this is not the case, and practical experience with the new writing system seems to confirm that assumption.

The pronunciation of each character, or each of its pronunciations if more than one, was treated as a single data point. The total number of pronunciations analyzed was 5641. The pronunciations were categorized based on the initial consonant, the presence or absence of the equivalent of a "y" following the consonant, the vowel, and the final sound if any. The final sounds analyzed were "u" (the elongation that may occur when the vowel is either o or u), n, ku, tsu, ki and chi, the last two being relatively rare.

Initial Consonant Frequency

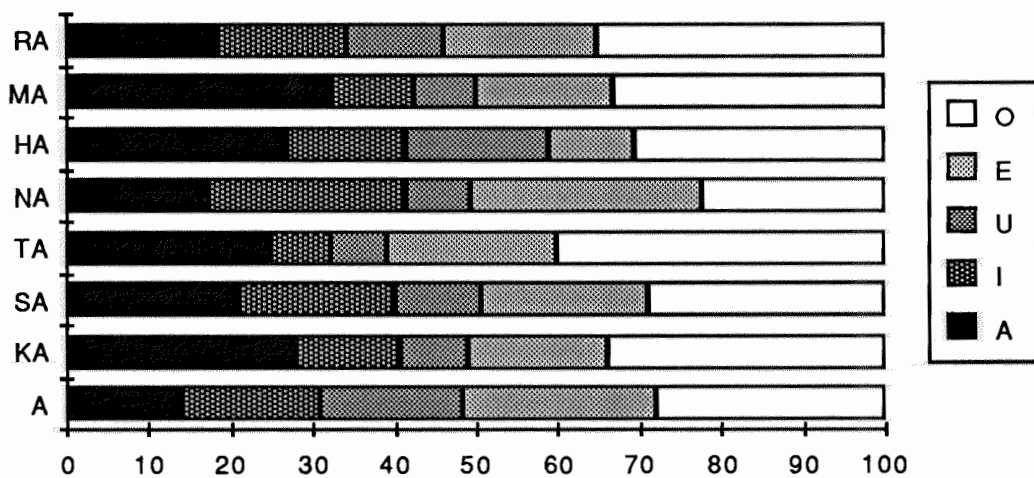
We start with the frequency analysis of initial consonant. This is shown in Dia. 1. This diagram shows both the frequency count for the standard form of the consonant as well as, in the shaded bar, that for the "nigori" form -- that is, "G" instead of "K", "Z" instead of "S" and so on.

⁶Nelson, Andrew Nathaniel Ph.D., *The Modern Reader's Japanese-English Character Dictionary*, Charles E. Tuttle Company, Tokyo, 1962..



Dia. 1. Frequency of Initial Consonant

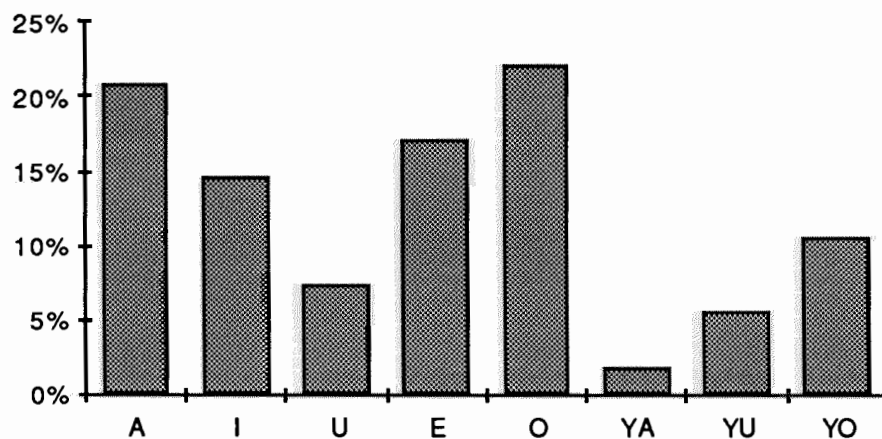
Before simply assigning a representatio to each initial consonant, we need to analyze the correlation between vowel and the initial consonant, to ensure, for example, that it would not be more optimal to assign elemental marks to some common consonant-vowel combinations. Doing so, we arrive at the results shown in the following diagram. Apparently the above distribution of vowels is relatively independent of the initial consonant. This means that the two can be treated as orthogonal, with signs chosen independently for the initial consonant and the following vowel.



Dia. 2. Distribution of Vowel by Initial Consonant

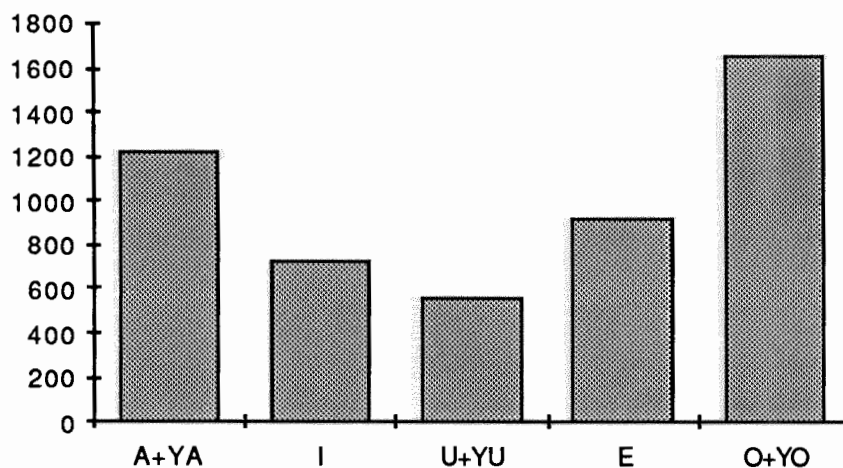
Frequency of Vowel

Next we proceed to the analysis of the vowel. This is shown in Dia. x. below, where, the bars marked "ya", "yu" and "yo" refer to the presence of these characters in their miniature form after the "i" form of the initial consonant when written in *kana*.



Dia. 3. *Distribution of Vowels*

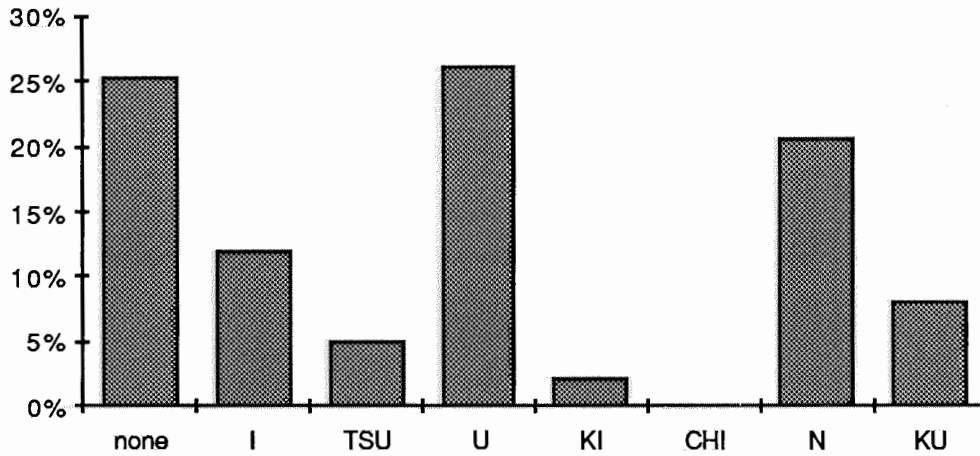
It is apparent that the "ya", "yu" and "yo" forms are relatively infrequently. However, when considered together with the standard forms of these vowels, we get a more interesting distribution as shown in Dia. 4.



Dia. 4. *Distribution of Vowels (II)*

Namely, the dominance of the "o" sound becomes apparent.

However, the vowel cannot be considered in isolation, but must rather be analyzed in combination with the following additional vowel or ending consonant. This is where the key features of the pronunciation frequency distribution become apparent. The final sounds consist of none, "n", "ki" "chi", "ku", "tsu", and the elongating "u". Let's look at their distribution.



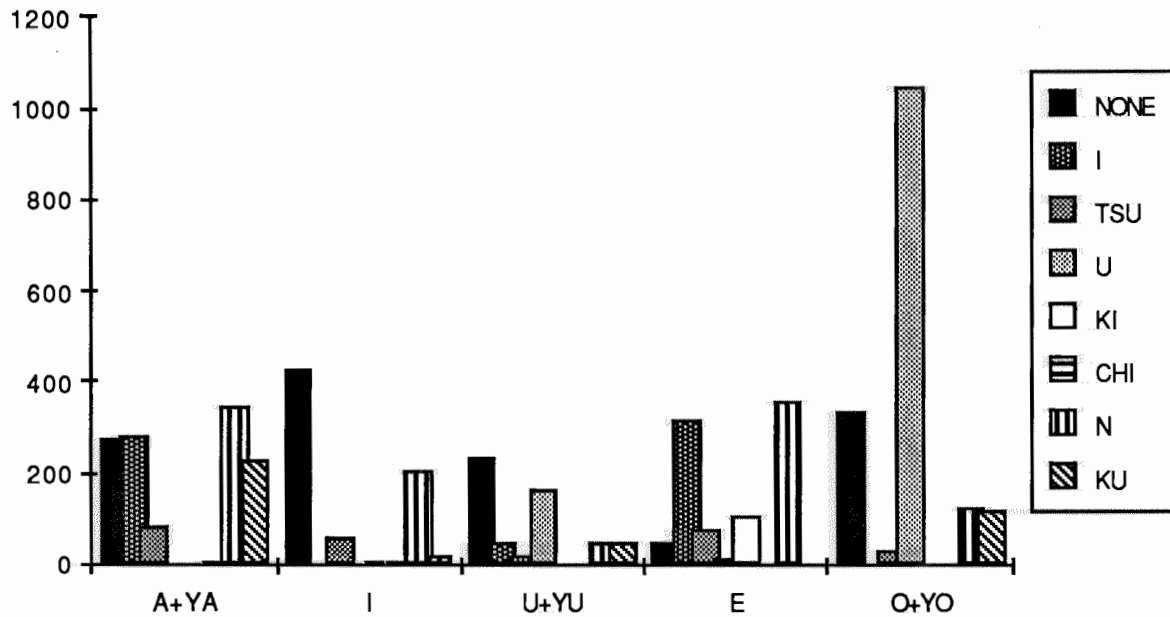
Dia. 5. *Distribution of Final Sound*

It is interesting to note that no final sound, the final elongating "u", and the final "n" occur with roughly equal frequencies. This type of analysis may have been behind the design of the "SKY" keyboard layout recently proposed by the NTT Communications Research Laboratory.⁷ This keyboard provides one-stroke access to all vowels followed by "n", as well as to the "ou", "uu" and "ei" combinations. In addition, it contains an "ai" key, which we did not find sufficiently interesting in the following analysis to make into a separate component.

Again, if it were the case that there were no particular correlation between the vowel and the final sound, then we could proceed to simply define symbol components for each vowel and each final

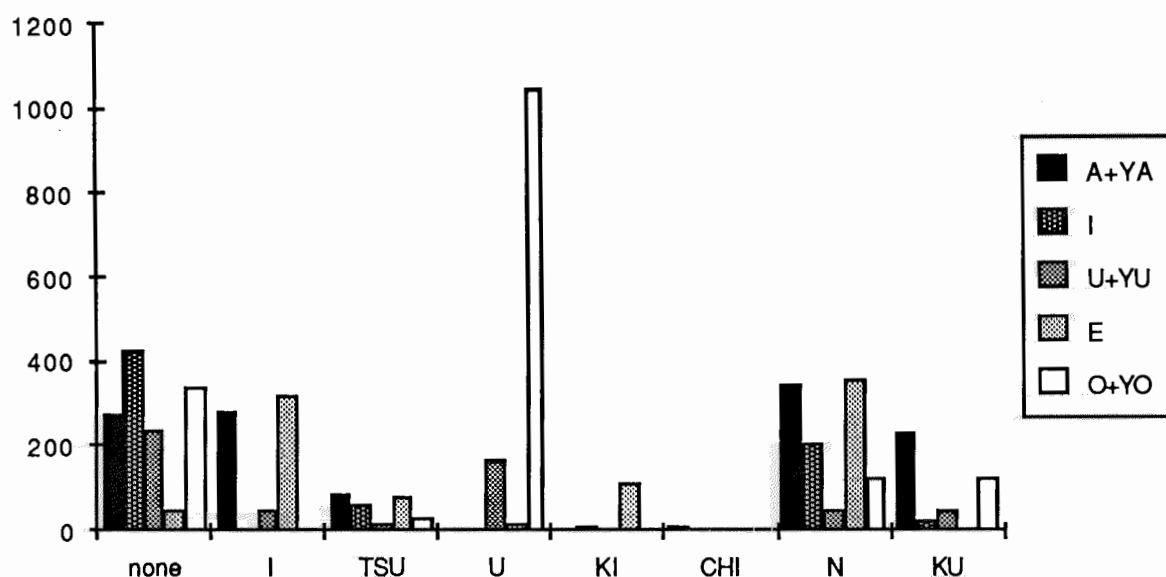
⁷Nikkan Kogyo Shimbun, February 17, 1987.

sound independently. However, there is such a correlation as shown dramatically in the following Dia. 6.



Dia. 6. *Ending Sound by Vowel*

The key correlations shown in this graph are as follows. First, clearly the vowel "o" is followed with overwhelmingly high frequency by the elongating "u". Less markedly, the vowel "e" is followed by "i" or "n" in the great majority of cases. The vowel "a" is followed in roughly equal parts by nothing, by "i", by "n" and by "ku". It will be instructive to view this same relationship from the reverse standpoint, namely which vowels tend to precede which final sounds. This is shown in Dia. 7.



Dia. 7. Vowel by Ending Sound

Besides the dominance of the "o" vowel elongated by the "u" final sound, which was mentioned above, we can conclude the following from this graph. First, the absence of any final sound is distributed relatively evenly across the vowels, with the exception of "e", which almost always takes some final sound. The final sound "i" occurs mostly in conjunction with the vowels "a" and "i". "tsu" is infrequent and occurs with nearly equal frequency after all vowels. As is well known, the elongating "u" occurs only after "o" and "u". "chi" is extremely rare, occurring for instance in the word "hachi". Finally, "n" occurs fairly regularly after all vowels except "u". And "ku" is most common coming after "a" and to some extent "o".

To summarize these results:

- initial sound distribution correlates poorly with vowel or final sound
- the *nigori* phenomenon ("ka" going to "ga" etc.) is largely independent of other aspects of a character's pronunciation
- the frequency of presence of small "ya", "yo" and "yu" is independent of other factors
- "n" and elongating "o" are dominating as final sounds
- elongating "u" follows vowel "o" with great frequency

Criteria for the Design of the New Characters

The implications for design of our character set are straightforward. We need a mark for each initial consonant and a way to represent the *nigori* form of each. We choose require a representation for the small "ya", "yo" and "yu" sounds. In choosing marks to represent vowels and final sounds, we clearly choose one, simple mark for the "o" elongated with "u" combination. This requires a different mark to represent the "o" alone; due to its relative rarity, we choose to represent this by an addition to the mark giving the "ou" sound. To the extent we have introduced the "u" elongation as a standard element of the writing system and also a way to remove it, we

apply this also, in the interests of orthogonality, to the vowel "u" as well, the only other vowel to take the elongation.

Considering the vowel "e", the standard, simplest version should represent one of the combinations "ei" and "en", which occur with approximately equal frequency. However, we note that whereas "i" as a final sound occurs quite often after "e" and beyond that only after "a", whereas "n" follows almost any vowel, we choose to represent the combination of any vowel with "n" by a common symbol element. This leaves us with "ei" as the standard form involving the vowel "e". A mechanism to "remove" the final sound has already been introduced to handle the vowel "o" by itself without the elongating final sound "u", so that mechanism can well be used here as well to "remove" the final "i" sound from the "ei" combination, producing the simple, but rare "e" with no final sound. Since most final sounds follow "a" with equal frequency, we make a simple "a" with no final sound the standard form here. Finally, since "i" has no particular dominant final sound, that is made the standard form as well.

Thus the repertoire of marks becomes:

- marks for initial consonants
- notation for *nigori*
- marks designating the present of a small "ya", "yu" or "yo"
- a mark for "ou"
- a mark for "uu"
- a mark for "ei"
- a mark to remove the final sound from the above three
- a mark for final sound "n"
- marks for "a" and "i"
- and marks for the other final sounds "i", "ku", "ki", "tsu" and "chi"

In addition, we need to consider, from the standpoint of writing convenience, common ways in which the above marks will be combined. The mark for "n" should fit cleanly into any combination not involving another final sound, since it occurs frequently with all vowels and initial consonants. The "final sound remover" mark should obviously flow smoothly from those for "ou", "uu" and "ei". Final sound "i" is already built into the standard mark for "ei", but needs to be able to be easily written after vowel "a" and to a lesser extent vowel "u". In addition, the mark for final sound "i" should be related to that for vowel "i" if possible. Finally, final sound "ku" follows vowel "a" with some frequency and so this should be considered.

Example Writing System

The writing system proposed below is nothing more than one possible interpretation of the above constraints. It is meant only as an example of the final application of this statistical analysis.

Based on the initial consonant frequency distributions, we arbitrarily chose standard shapes to represent them shown in Fig. 2 below. "K" and "S" are written most simply due to their dominance in the distribution. Other considerations in choosing these shapes were, in the case of "M", to have the two horizontal bars remind one of those in *hiragana* "ma"; and in the case of "T", to have the single horizontal bar to the right remind one of that in *katakana* "to".



Fig. 2. Representation of Initial Consonant Sound

In addition, due to the relatively low count of the *nigori* forms, we adopt the convention that they are written with the same two dots used in *katakana* and *hiragana*. For reference, we show those forms below.



Fig. 3. Nigori Forms of Initial Consonants

Here we digress to explain the way in which these marks will combined with those to be defined later. Each of the marks above ends with a vertical downstroke. In general, that downstroke is to be led directly into a downstroke which commences the mark representing the vowel. Additional marks, such as those for final sounds, are placed after that vowel mark, with the exception of the mark for terminal "n", which is arbitrarily placed at the top of the first initial sound stroke as will be discussed below. In general the intent is to appeal to the Japanese convention of writing characters from top to bottom and left to right.

Next we present the vowels and combinations "a", "i", "uu", "ei", "ou".

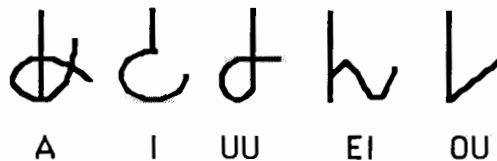


Fig. 4. Marks for Vowels and Combinations

Note here the intent is that the vertical downstroke commences from and is connected to the vertical downstroke in the marks for initial consonants defined above. Thus we have the following types of combinations as examples:



Fig. 5. Combinations of Initial Consonants and Vowels

The "final sound remover" is a vertical down stroke, used as follows:



Fig. 6. *Final sound Remover*

Next we introduce the representation for "ya", "yu" and "yo" in their combination forms. Note that these are formed by the addition of the small loop to the related vowel sounds. When combined with the final sound remover, it is proposed that the small loop go at its bottom.

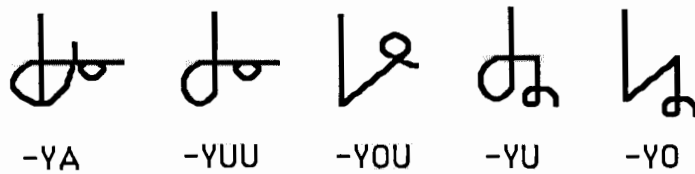


Fig. 7. *Marks for "ya", "yu" and "yo"*

So that the final sound "n" can be placed easily on any initial sound-vowel combination, it is placed on the top of the combined mark as follows:

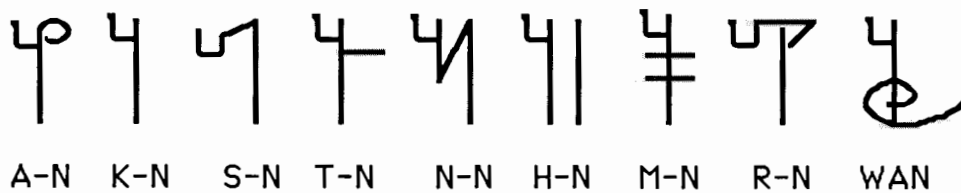


Fig. 8. *Representation of Terminal "n"*

Note that no sound beginning with "y" ends in "n", thus that combination is not defined. Furthermore, since there are no sounds such as "koun", we define the presence of the terminal "n" mark as automatically functioning as a final sound remover if used together with "uu", "ei" or "ou" vowel marks.

We next define characters for the final sounds, as follows:

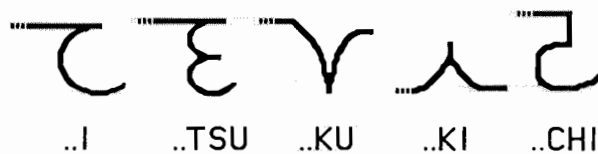


Fig. 9. *Representations of Final Sounds*

The tables following show some, but not all, of the characters derived from these rules.

Summary and Conclusion

In spite of their value from a variety of standpoints, *kanji* remain hard to learn and to process. It is worth investigating alternatives. Such alternatives should be based on a statistical analysis of *kanji* pronunciations. Such an analysis shows certain regularities and characteristics that can be used in designing an optimal writing system.

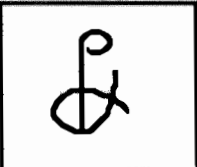

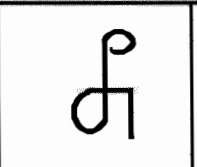
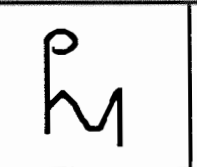
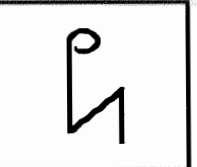
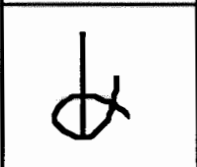
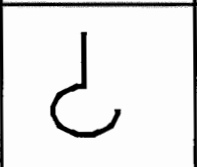
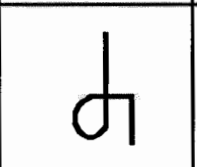
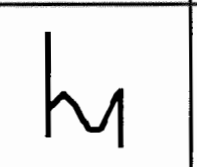
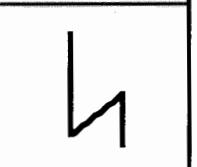
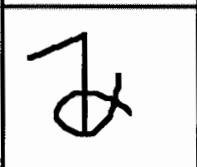
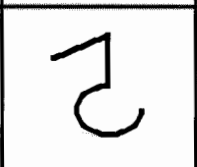
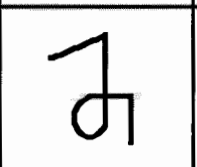
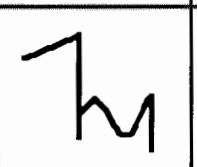
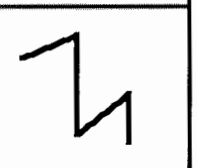
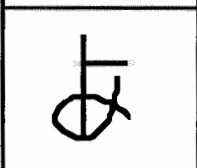
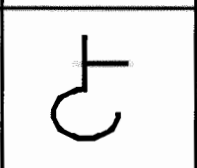
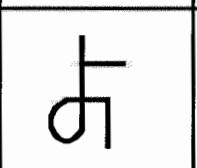
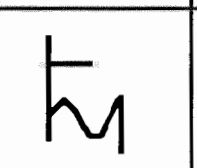
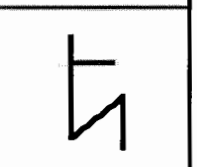
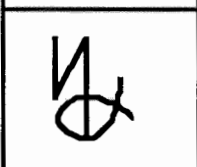
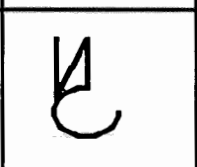
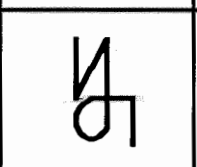
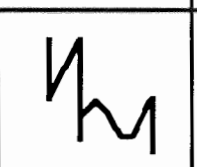
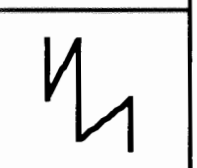
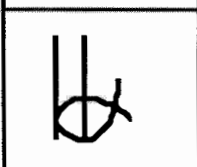
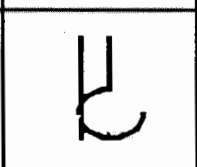
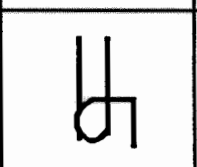
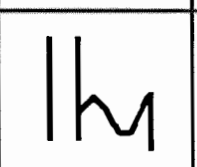
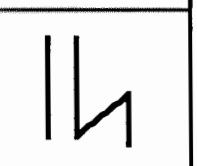
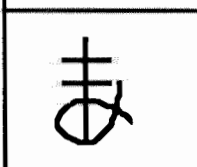
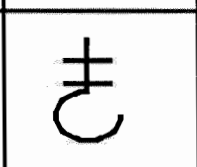
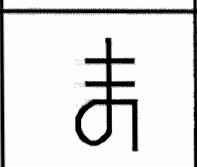
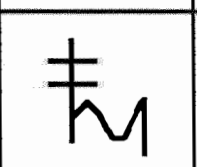
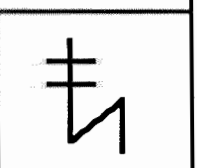
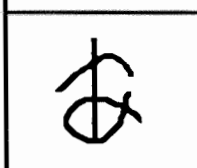
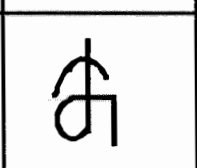
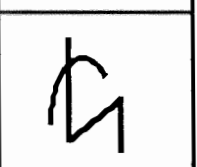
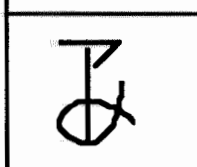
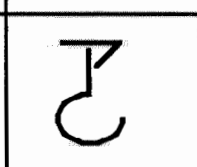
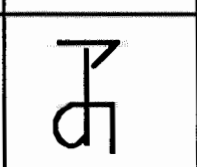
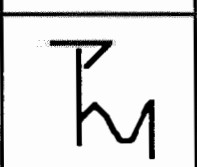
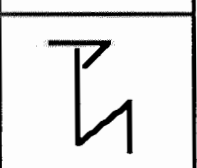
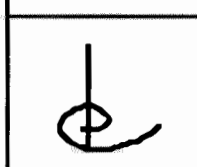
	A	I	E	U	O
A					
KA					
SA					
TA					
NA					
HA					
MA					
YA					
RA					
WA					

Table 1. Standard Combinations

	AI	UU	UI	EI	OU
A					
KA					
SA					
TA					
NA					
HA					
MA					
YA					
RA					

Table 2. Diphthong Combinations

	A	I	U	E	O
A					
KA					
SA					
TA					
NA					
HA					
MA					
RA					
WA					

Table 3. Combinations with Terminal "n"