Analyzing Student Assimilation of Japanese Phonological Transformation Rules

Yun-Sun Kang Anthony A. Maciejewski

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Abstract—**This work describes a method for statistically analyzing a student's proficiency at reading one of the distinct orthographies of Japanese, known as *katakana*. The result of the analysis is being applied to a Japanese language intelligent tutoring system for appropriately individualizing the student's instruction.**

I. INTRODUCTION

Interest in Japanese language instruction has risen dramatically in recent years, particularly for those Americans engaged in technical disciplines. However, the Japanese language is generally regarded as one of the most difficult languages for English-speaking people to learn. While the number of individuals studying Japanese is increasing there remains an extremely high attrition rate, estimated by some to be as high as 80% [4]. Much of this difficulty can be associated with the Japanese writing system. Japanese text consists of two distinct orthographies, a phonetic syllabary known as *kana* and a set of logographic characters, originally derived from the Chinese, known as *kanji*. The *kana* are divided into two phonetically equivalent but graphically distinct sets, *katakana* and *hiragana*, both consisting of 46 symbols and two diacritic marks denoting changes in pronunciation. The *katakana* are used primarily for writing words of foreign origin that have been adapted to the Japanese phonetic system although they are also used for onomatopoeia, colloquialisms, and emphasis. The *hiragana* are used to write all inflectional endings and some types of native Japanese words that are not currently represented by *kanji*. Due to the limited number of *kana*, their relatively low visual complexity, and their systematic arrangement they do not represent a significant barrier to the student of Japanese. In fact, the relatively small effort required to learn *katakana* yields significant returns to readers of technical Japanese due to the high incidence of terms derived from English and transliterated into *katakana*.

This work describes a method for statistically analyzing a student's proficiency at reading *katakana*. The results of this analysis are being applied to a Japanese language intelligent tutoring system [2] for appropriately individualizing a student's instruction. The remainder of this paper is organized as follows: Section II provides a brief introduction to how a student model is constructed by analyzing a student's responses. A method is then presented for statistically analyzing a student model assuming that all of the phonological rules that would be required to completely transform these *katakana* into English contributed equally to the student's failure to understand. With this assumption, the student model becomes a binomial distribution for which the well-known Bayes' theorem is used to estimate the student's current knowledge state. A variety of techniques for assessing prior information are then proposed. In Section III, the correlation between the probability of comprehension and the phonetic properties of transformation rules is addressed. It is shown that combining the binomial model with these factors allows the tutorial system to more accurately estimate a student's knowledge state and thus provide more efficient instruction. Finally, the conclusions of this work are presented in the final section.

II. STATISTICAL ANALYSIS OF THE STUDENT MODEL

In this section, a method is presented for statistically analyzing a student's responses to the tutorial system. The knowledge base which a student must acquire in order to be proficient at reading *katakana* consists of a set of phonological rules which characterize the transformation of Japanese *katakana* to its English origin [3]. Information about the student is gathered passively by simply noting the words for which he requests translations from the Japanese language tutoring system [2]. Analyzing a student's response is, therefore, relatively difficult for the tutor. The following presents a simple illustration of how the student model is formed by analyzing the responses of a student who was tested for his *katakana* reading proficiency. This particular student had no difficulty with the following words:

shisutemu	→	system
bideo	→	video
totaru	→	total
tasuku	→	task

which include the rules u → *, ʃ → s, i → ɪ, b → v, and r → l. However, the student could not comprehend the following *katakana* words:

Manuscript received August 1, 1992. This material is based upon work supported by the National Science Foundation under Grant No. INT-8818039 and in part by the NEC Corporation

0-7803-0720-8/92 $3.00 ©1992 IEEE

aasu	→	earth
rengusu	→	length
saamaru	→	thermal

which use the phonological transformation rules: u → *, r → l , and s → θ. From analyzing these two sets of data, the tutoring system is able to correctly identify that the rule which the student has not mastered is the transformation s → θ. This is not particularly surprising since this is a rather radical change in pronunciation which occurs relatively infrequently. Indeed, this student is rather typical in that he has acquired the relatively straightforward rules such as u → * which occurs extremely frequently and is one of the primary mechanisms for dealing with the disparity in consonant clusters between English and Japanese. Likewise, this student has no trouble with the simple consonant substitutions b → v and r → l. Therefore, the tutoring system would tailor the instruction of this student with *katakana* words that contain the more obscure rules such as s → θ, hopefully being able to find occurrences in which this is the only rule present in order to provide more contextual information.

In the initial analysis, it is assumed that all of the phonological rules that would be required to completely transform these *katakana* into English contributed equally to the student's failure to understand. With this assumption the probability that a student understands x out of n words that require the rule R for transliteration back to their English origins becomes a binomial distribution with index n and π. This probability is associated with the conditional probability density $p(x|\pi)$. The student's current knowledge state can be estimated by the probability density $p(\pi|x)$. The posterior probability $p(\pi|x)$ function is then computed from the model density and the prior density $p(\pi)$ by using Bayes' theorem, i.e.

$$p(\pi|x) = \frac{p(x|\pi)\,p(\pi)}{p(x)}. \qquad (1)$$

The most common method for computing prior density is to approximate one's prior beliefs by a density which is a member of a mathematically convenient family. The prior density for the binomial distribution is, then, the well-known beta function. When assuming that the prior density function is chosen as the beta function of the parameters a and b, the posterior density function becomes the beta function of the parameters

$$p = x + a \qquad (2)$$

and

$$q = n - x + b. \qquad (3)$$

In order to compute the mean value of the beta function, one needs to determine only the two parameters of the beta function, p and q. The mean value of the posterior density for the binomial model is, therefore,

$$E(\pi|x) = \frac{p}{p+q} = \frac{a+x}{a+b+n}. \qquad (4)$$

If the only available evidence about a student's ability is the fact that he correctly understood x words on an n-word test; then, the tutor has no prior information whatsoever about this student. One possible approach is to express no prior information by considering all values of the prior density to be equally likely. This uniform prior is known as Bayes' postulate [5] and it corresponds to $a = b = 1$ in the beta prior. When the number of trials is extremely large, the effect of the prior information becomes relatively small. However, a non-uniform prior density results in a proper prior. One possible method is to give the student a pre-test in order to get prior information about the student's knowledge [1]. Unfortunately, since there are more than 130 phonological transformation rules in the knowledge base of the Japanese tutoring system, a simple test cannot cover all of the rules. There are a number of possible assumptions. When a student sees a rule for the first time, the tutor can assume that:

- The student does not have any knowledge of the rule (a = 0 and b = 1).
- The student's knowledge state is independent of the prior information (a = 0 and b ≈ 0).

A comparison of the results of using uniform and non-uniform priors for the student discussed above is presented in Table I. In the following section, the lack of prior information is compared to assumptions about its probable distribution based on such factors as the frequency of a rule or on the extent of the phonological transformation.

III. THE EFFECTS OF RULE FREQUENCIES

While the binomial model is shown to be reasonably effective in analyzing the difficulties which students encounter in comprehending *katakana*, there are also some significant limitations due to the assumption that all rules are equally responsible for the student's failure to understand. Clearly, a student may correctly identify the origin of a *katakana* without a mastery of all of the transformation rules required due to the redundancy in human language. Likewise, students may fail to comprehend words for which they know all of the transforma-

TABLE I
EFFECTS OF VARIOUS PRIOR DENSITY FUNCTIONS ON THE MEAN VALUE OF THE POSTERIOR FUNCTION IN THE BINOMIAL MODEL

Rule	a = 1, b= 1	a = 0, b = 1	a = 0, b ≈ 0
t → t	0.83	0.80	1.00
b → v	0.67	0.50	1.00
u → *	0.55	0.50	0.56
r → l	0.40	0.25	0.33
a: → ɚ	0.25	0.00	0.00
s → θ	0.25	0.00	0.00

tion rules due to such factors as unfamiliarity with the vocabulary or the sheer number and/or combination of rules required. For these reasons, it is relatively difficult for the tutoring system to classify the rules mastered and the rules that need more review based on the probability of comprehension as shown in Table II.

In order to resolve these problems, a statistical analysis was conducted on the data produced under the binomial model for 43 students ranging from 1 to 3 years of classical Japanese language instruction. In this analysis it is revealed that there is a strong correlation between the probability of comprehension and the extent of the phonetic modification of transformation rules. As is expected, Fig. 1 illustrates that a student can more easily comprehend the *katakana* that contain a large number of trivial consonant rules, i.e., those rules that do not represent a significant phonetic modification between consonants. Conversely, it is shown in Fig. 2 that students have more trouble understanding words that use large numbers of non-trivial consonant rules in the transliteration process. It is interesting to note, however, that similar data for the vowel rules, depicted in Fig. 3 and Fig. 4, do not exhibit this correlation. This is probably due to the large disparity between the number of Japanese and English vowels which greatly reduces their information content. The silent rules, i.e., the rules that transform a Japanese phoneme into the null phoneme in English, also seem to have little effect on a student's comprehension (see Fig. 5). The most dominant of the silent rules is the rule u → * which is the primary mechanism for dealing with English consonant clusters. Since this rule occurs more often than any of the others, it appears to be quickly assimilated by even first year students and therefore has little effect on overall comprehension. In summary, these results show that different types of transformation rules have different effects on a student's ability to comprehend *katakana*.

In order to account for the different effects of the various rule types, the assumptions used to calculate the probability of comprehension for the rules was modified. This involved the calculation of a scalar value $0 < w \leq 1$ associated with each rule. The value of w represents the likelihood that this rule will contribute to any difficulty with comprehension of words that contain this rule. This value of w is then used to modify (4) so that the probability of rule comprehension is calculated using

$$E_w(\pi|x) = \frac{a + wx}{a + b + wx + (1-w)(n-x)}. \qquad (5)$$

Thus rules with a large value of w are assumed to be trivial and their probabilities are not adversely affected for student who does not understand a word due to some other factors. Conversely, rules with a small value of w are considered to be more difficult and a student must demonstrate comprehension of such a rule by correctly identifying virtually every word in which it appears. This prevents an artificially high probability of comprehension for difficult rules due to a student being able to guess words with high degrees of redundancy. The two dominant factors that were empirically found to affect a rules difficulty were (1) the absolute frequency of a rule, i.e. the number of words that contain that rule divided by the total number of words that the student has read; and (2) the relative frequency, defined as the number of occurrences of a rule divided by the total number of all occurrences for all rules that govern the same Japanese phoneme. These frequencies were computed for all rules in the rule base used by the tutoring system with a representative sample presented in Table III. The value of w for a rule is then calculated as a linear combination of these two frequencies. Since it is not clear which of the two frequencies is dominant in determining a rules difficulty, the average of the two values is currently being used. Table IV shows the resulting probabilities of comprehension and illustrates a much closer correspondence to empirical evidence, particularly with respect to the trivial consonant rules

TABLE II
THE PROBABILITY OF COMPREHENSION
COMPUTED BY ASSIGNING THE SAME WEIGHTS
TO ALL OF THE RULES

Rule	Probability of comprehension		
	1st year	2nd year	3rd year
s → θ	0.07	0.09	0.38
b → v	0.23	0.32	0.65
a: → ɚ	0.25	0.36	0.59
r → l	0.30	0.46	0.78
u → *	0.33	0.48	0.77
e → e	0.35	0.44	0.75
t → t	0.41	0.53	0.84
k → k	0.55	0.64	0.90

TABLE III
RELATIVE AND ABSOLUTE FREQUENCIES
FOR THE PHONOLOGICAL RULES

Rule	relative frequency	absolute frequency
s → θ	0.04	0.01
b → v	0.34	0.05
e → e	0.55	0.14
r → l	0.57	0.32
a: → ɚ	0.69	0.05
u → *	0.91	0.73
t → t	0.99	0.32
k → k	1.00	0.22

TABLE IV
THE PROBABILITY OF COMPREHENSION
COMPUTED BY ASSIGNING DIFFERENT WEIGHTS
TO EACH RULE

Rule	Probability of comprehension		
	1st year	2nd year	3rd year
s → θ	0.01	0.03	0.06
b → v	0.07	0.12	0.33
a: → ɚ	0.17	0.25	0.46
e → e	0.22	0.29	0.61
r → l	0.26	0.41	0.74
k → k	0.55	0.64	0.90
t → t	0.56	0.68	0.91
u → *	0.69	0.80	0.94

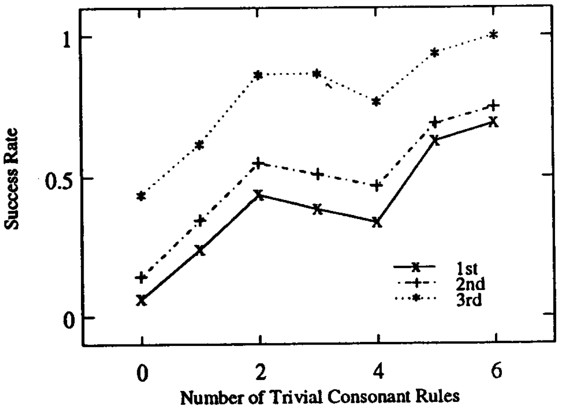

Fig. 1. The effect of trivial consonant rules
on the probability of comprehension of a katakana word
for 1st, 2nd, and 3rd year Japanese students

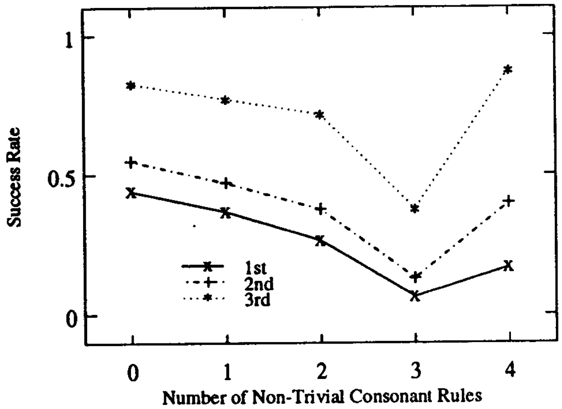

Fig. 2. The effect of non-trivial consonant rules
on the probability of comprehension of a katakana word
for 1st, 2nd, and 3rd year Japanese students

and the silent rule, as compared to Table II. The higher accuracy in the estimation of these probabilities as well as their wider distribution allows the tutorial system to more effectively select lessons that review the specific weaknesses of individual students.

IV. CONCLUSIONS

The goal of this work was the development of a model for representing a student's proficiency in reading *katakana*. This model is used to individualize the instruction of an intelligent tutoring system that is designed to assist scientists and engineers acquire a reading knowledge of technical Japanese. This is illustrated with data that shows a strong correlation between the probability of comprehension and both the relative and absolute frequency of a rules occurrence, as well as the extent of the phonetic modification. It is shown that combining such

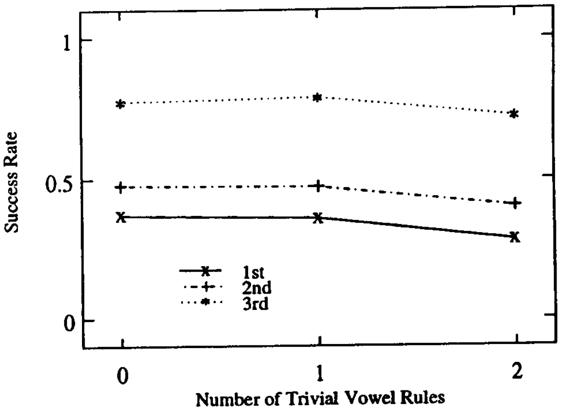

Fig. 3. The effect of trivial vowel rules
on the probability of comprehension of a katakana word
for 1st, 2nd, and 3rd year Japanese students

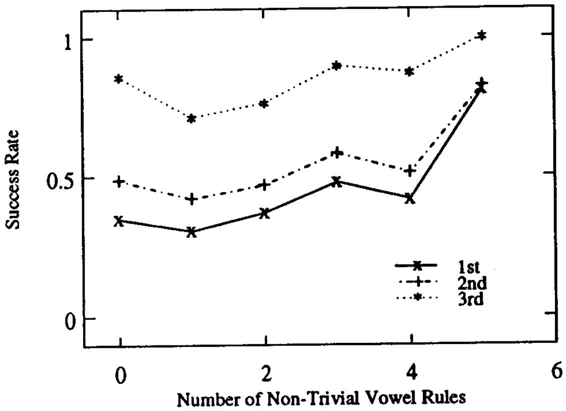

Fig. 4. The effect of non-trivial vowel rules
on the probability of comprehension of a katakana word
for 1st, 2nd, and 3rd year Japanese students

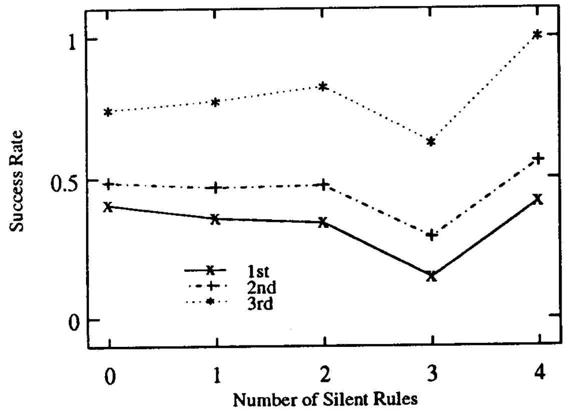

Fig. 5. The effect of silent rules
on the probability of comprehension of a katakana word
for 1st, 2nd, and 3rd year Japanese students

factors with the binomial model allows the tutorial system to more accurately estimate a student's knowledge state and thus provide more efficient instruction. This technique has proven very effective in analyzing the difficulties which students encounter in comprehending *katakana*.

REFERENCES

[1] R. B. Burton, "Diagnosis of errors in basic mathematical skills," in D. Sleeman & J. S. Brown (Eds.), *Intelligent Tutoring Systems*, New York: Academic Press, 1982, pp. 157-183.

[2] A. A. Maciejewski and N. K. Leung, "The Nihongo Tutorial System: An intelligent tutoring system for technical Japanese language instruction," *Journal of the Computer Assisted Language Learning and Instruction Consortium*, Vol. 9, No. 3, 1992.

[3] A. A. Maciejewski and Y.-S. Kang, "The student model of *katakana* reading proficiency for a Japanese language intelligent tutoring system," in *Proceedings 1991 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1871-1876, Charlottesville, Virginia, October 14-17, 1991.

[4] D. O. Mills, R. J. Samuels, and S. L. Sherwood, *Technical Japanese for Scientists and Engineers: Curricular Options*, a report to the National Science Foundation, Massachusetts Institute of Technology, MITJSTP WP 88-02, 1988.

[5] S. J. Press, *Bayesian Statistics: Principles, Models, and Applications*, New York: John Wiley & Sons, 1989.