A Student Model of Katakana Reading Proficiency for a Japanese Language Intelligent Tutoring System

Anthony A. Maciejewski, *Member, IEEE*, and Yun-Sun Kang

Abstract—**This work describes the development of a student model that is used in a Japanese language intelligent tutoring system to assess a pupil's proficiency at reading one of the distinct orthographies of Japanese, known as katakana. While the effort required to memorize the relatively few katakana symbols and their associated pronunciations is not prohibitive, a major difficulty in reading katakana is associated with the phonetic modifications which occur when English words which are transliterated into katakana are made to conform to the more restrictive rules of Japanese phonology. The algorithms described here are able to automatically acquire a knowledge base of these phonological transformation rules, use them to assess a student's proficiency, and then appropriately individualize 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% [12]. Much of this difficulty can be attributed to the Japanese writing system. Japanese text consists of two distinct orthographies (writing systems), a phonetic syllabary known as *kana* and a set of logographic characters (characters that represent words), 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[1] 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[2], colloquialisms[3], 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 the development of a system that is used to automatically acquire knowledge about how English words are transliterated into *katakana*. After the tutoring system has "learned" this domain knowledge it can be subsequently used to develop a model of a student's proficiency in reading *katakana*. This model is used by an intelligent tutoring system developed previously [10] which assists the student who is learning to read technical Japanese. The remainder of this paper is organized as follows: In Section II the structure of intelligent tutoring systems is overviewed by comparing existing intelligent tutoring systems to the Japanese language intelligent tutoring system. Section III provides a brief introduction to the *katakana* writing system and to the rules of Japanese phonology. This is follows by the description of an algorithm to automatically generate a *katakana* to English dictionary from an arbitrary Japanese text and its English translation. In Section IV, the *katakana* to English dictionary is used to develop a set of phonological rules that govern the inverse transformation from Japanese phonetics, represented by the *katakana* orthography, back to the original English pronunciation. This set of rules is treated as the knowledge base that the student must acquire in order to become proficient at reading *katakana*. The failure to recognize words that contain specific phonological rules is then used to build the student model, which is described in Section VI. 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 VII, 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 last section.

II. Overview of Intelligent Tutoring Systems

Intelligent tutoring systems (ITSs) are computer programs that can individualize their instruction based on inferences

Manuscript received August 29, 1992; revised September 10, 1993. This work was supported by the National Science Foundation under Grant No. INT-8818039 and in part by the NEC Corporation.

The authors are with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907-1285 USA.

IEEE Log Number 9403009.

[1] added to a symbol to alter its value.

[2] words that imitate sounds.

[3] expressions relating to conversation.

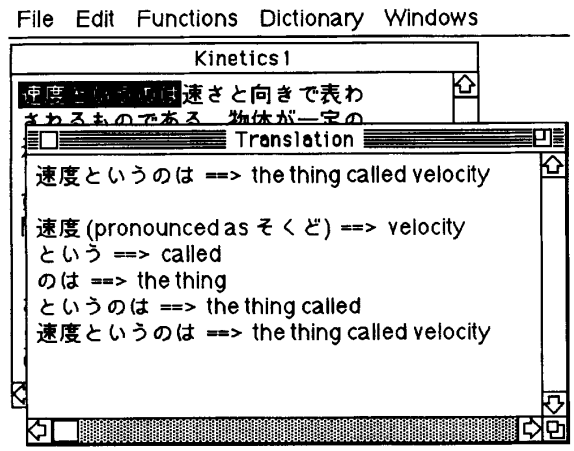


Fig. 1. A screen display from the Nihongo Tutorial System: The full translation menu option provides syntactic as well as semantc information to the student.

about a student's knowledge. While existing ITSs vary in architecture, they typically consist of at least four basic components [11], [20]; the expert knowledge module, the tutoring module, the user interface module, and the student model module. The expert knowledge module provides the domain knowledge that the system intends to teach. The tutoring module is the part of the ITS that designs and regulates instructional interactions with the student. The user interface module controls interactions between the system and the student. Finally, the fourth component of intelligent tutoring systems is the student model, which refers to the dynamic representation of a student's competence for the given domain. More details on ITS structure and previously developed prototypes are available in [9], [13], [14], [16].

The specific ITS being considered in this work is called the Nihongo Tutorial System and was designed to assist English-speaking scientists and engineers acquire Japanese reading proficiency in their technical area of expertise [10]. The architecture of this system closely resembles the general structure of ITSs outlined above. The domain knowledge database is prepared by a software module called the Parse Tree Editor that processes technical journal articles into instructional material by incorporating syntactic, semantic, phonetic, and morphological information into a representation known as an augmented parse tree. The instructional interactions between the system and a student are regulated by the Administrator module which matches the student's current level of Japanese proficiency and technical area of interest with the available instructional material produced by the Parse Tree Editor. The user interacts with the system through a graphical user interface to request information about the current instructional text or to obtain examples of material on the same or related concepts (see Fig. 1). The student model, which is updated based on a student's interactions with the user interface and subsequently used by the administrator to select an instructional strategy, is the focus of this work.

The student model currently used by the Nihongo Tutorial System falls into the general class of student models known as "overlay" models [3]. This commonly used type of model considers the student's knowledge to be a subset of the expert knowledge base. In the case of the Nihongo Tutorial System the model contains the Japanese characters, vocabulary, and the syntactic, morphological, and phonological transformation rules required to understand the Japanese text, along with a number that represents the probability that the student understands that particular piece of knowledge. An advantage of such an overlay model is that it is very simple for the Administrator module to compare the student's knowledge with that of the expert knowledge module in order to tailor the instructional material. A disadvantage of this approach, however, is that the system cannot deal with student errors that do not originate from incomplete knowledge. The concept of "buggy modeling", proposed by Brown and Burton [1], which would include a list of incorrect rules used by the student would be one technique for modeling these types of errors. The remainder of this work will only deal with the expert knowledge base and overlay student model associated with the phonological transformation rules required to understand Japanese text written in *katakana*.

III. Katakana and Japanese Phonology

The Japanese lexicon contains an extremely large number of words originating from foreign languages. Traditionally, these Japanese loan words are characterized as either *kango*, literally "Chinese word," which are of Chinese origin (also referred to as Sino-Japanese words) or *gairaigo*, literally "foreign word," for words from any other country of origin. While the proportion of Sino-Japanese words in the lexicon is extremely large due to the profound cultural influence of China, words of English origin have dominated the class of loan words since the late 19th century. In a study of Japanese publications performed between 1956 and 1964, over 80% of the *gairaigo* originated from English [8]. This process of adopting English words into the lexicon is particularly common for relatively new or specialized terms arising in technical literature.

When adopting a word of foreign origin into Japanese, the original pronunciation of that word is typically transliterated into *katakana* which graphically represents all of the possible phonetic sequences in the Japanese language. It is this process of modifying English phonetic sequences to conform to the rules of Japanese phonology which presents English-speaking readers of *katakana* with difficulty in identifying a word's meaning. This is due to the fact that the rules of Japanese phonology are quite different from those of English. In particular, Japanese has only five single vowel sounds /a i u e o/ in contrast to the large number of vowel sounds in English. These vowels, when combined with the nine Japanese consonants /k s t n h m j r w/ constitute 44 of the 46 basic sounds in Japanese which are traditionally organized as shown in Table I. This organization is referred to as *gozyuuoñzu*, which literally translates as "table of fifty sounds," since all fifty sounds were used at one point in time. The pronunciation of each *katakana* character in the table is represented using the International Phonetic Alphabet (IPA) symbols. In addition to these basic pronunciations, certain *katakana* may be written

TABLE I
KATAKANA CHARACTERS AND THEIR PHONETIC REPRESENTATIONS IN THE IPA SYMBOLS: BASIC SYLLABLES

ア	イ	ウ	エ	オ
/a/	/i/	/u/	/e/	/o/
カ	キ	ク	ケ	コ
/ka/	/ki/	/ku/	/ke/	/ko/
サ	シ	ス	セ	ソ
/sa/	/ʃi/	/su/	/se/	/so/
タ	チ	ツ	テ	ト
/ta/	/ʧi/	/tsu/	/te/	/to/
ナ	ニ	ヌ	ネ	ノ
/na/	/ni/	/nu/	/ne/	/no/
ハ	ヒ	フ	ヘ	ホ
/ha/	/hi/	/fu/	/he/	/ho/
マ	ミ	ム	メ	モ
/ma/	/mi/	/mu/	/me/	/mo/
ヤ		ユ		ヨ
/ja/		/ju/		/jo/
ラ	リ	ル	レ	ロ
/ra/	/ri/	/ru/	/re/	/ro/
ワ				
/wa/				

TABLE II
KATAKANA CHARACTERS AND THEIR PHONETIC REPRESENTATIONS IN THE IPA SYMBOLS: MODIFIED SYLLABLES

ガ	ギ	グ	ゲ	ゴ
/ga/	/gi/	/gu/	/ge/	/go/
ザ	ジ	ズ	ゼ	ゾ
/za/	/ʤi/	/zu/	/ze/	/zo/
ダ	ヂ	ヅ	デ	ド
/da/	/ʤi/	/zu/	/de/	/do/
バ	ビ	ブ	ベ	ボ
/ba/	/bi/	/bu/	/be/	/bo/
パ	ピ	プ	ペ	ポ
/pa/	/pi/	/pu/	/pe/	/po/

with one of two diacritic marks that modify their pronunciation (see Table II). In particular, the four rows in Table I that correspond to the consonants /k s t h/ may be written with a symbol consisting of two short parallel lines, known as *nigori*, which results in a voiced version of these consonants, i.e., /g z d b/. The second diacritic mark is a small circle referred to as *maru* which when combined with the *katakana* in the /h/ row results in pronunciations which include the bilabial[4] stop[5] /p/. The *katakana* in the second column, i.e. those associated with the vowel /i/, may also be combined with any of the *katakana* in the row associated with the consonant /j/ to produce additional single syllable pronunciations. The second *katakana* in this case is written slightly smaller in order to differentiate from a two syllable sequence, as illustrated in Table III. Finally, Japanese includes two moraic[6] consonants, /Q/ the mora obstruent[7] and /N/ the mora nasal[8], which are presented in Table IV.

[4] said of a consonant made with both lips.

[5] a consonant made by a complete closure in the vocal tract.

TABLE III
KATAKANA CHARACTERS AND THEIR PHONETIC REPRESENTATIONS IN THE IPA SYMBOLS: CONSONANT PLUS /ja/, /ju/, or /jo/

キャ	キュ	キョ	ギャ	ギュ	ギョ
/kja/	/kju/	/kjo/	/gja/	/gju/	/gjo/
シャ	シュ	ショ	ジャ	ジュ	ジョ
/ʃa/	/ʃu/	/ʃo/	/ʤa/	/ʤu/	/ʤo/
チャ	チュ	チョ			
/ʧa/	/ʧu/	/ʧo/			
ニャ	ニュ	ニョ			
/nja/	/nju/	/njo/	ビャ	ビュ	ビョ
ヒャ	ヒュ	ヒョ	/bja/	/bju/	/bjo/
/hja/	/hju/	/hjo/	ピャ	ピュ	ピョ
ミャ	ミュ	ミョ	/pja/	/pju/	/pjo/
/mja/	/mju/	/mjo/			
リャ	リュ	リョ			
/rja/	/rju/	/rjo/			

TABLE IV
KATAKANA CHARACTERS AND THEIR PHONETIC REPRESENTATIONS IN THE IPA SYMBOLS: MORA CONSONANT

ン	ッ
/N/	/Q/

From the above description of the *katakana* orthography, it is clear that certain inherent limitations are imposed on phonetic sequences in the Japanese language. In particular, Japanese does not allow any consonant clusters except when a consonant is followed by a glide[9] or preceded by a moraic consonant [18]. In addition, consonants may not appear at the end of a sequence. These restrictions, together with the limited number of Japanese vowel sounds, result in the vast majority of phonological modifications which occur when transliterating an English word into *katakana*. By the same token, these resulting modifications are the source of difficulty for English-speaking readers of *katakana*.

It should also be noted that there are additional difficulties to comprehending *katakana* that are unrelated to the phonological processes involved. In particular, while it is true that the vast majority of loan words are created by the phonological process, foreign borrowing may also be modified by changes in form due to simplification, semantics, or Japanese coinage [17]. For example, simplification frequently occurs with polysyllabic words such as "television" and "word processor" which are shortened to the *katakana* words *terebi* and *waapuro*, respec-

[6] a minimal unit of rhythmical time equivalent to a short syllable.

[7] sounds made with a constriction.

[8] sounds made with the soft palate lowered, thus allowing air to resonate in the nose.

[9] a transitional sound made as the vocal organs move towards or away from an articulation.

tively. Changes in semantics have resulted in the *katakana* word *botan* being used to designate a touch tone type of telephone, whereas its phonetic origin is from the word button. Examples of coinage which result from combinations of existing loan words include *maikaa* (derived from my + car) and *maihoomu* (derived from my + home) which refer to privately owned cars and houses. All of these processes contribute to a student's difficulty in achieving reading proficiency in *katakana*, however, this work focuses on the phonological modifications.

IV. A Katakana to English Dictionary

Due to the relatively specialized and technical nature of a significant portion of the recently borrowed loan words, comprehensive *katakana* to English dictionaries are not readily available. Therefore, the first step in analyzing the phonetic modifications which occur when English words are transliterated into *katakana* is to develop an algorithm for automatically generating a *katakana* to English dictionary from Japanese text for which English translations are available. Fortunately, many examples of Japanese technical literature which contain significant amount of *katakana* are available in an electronic format, along with the English translation of the document. The primary source for the development of the *katakana* to English dictionary described here, is a set of 3,000 Japanese phrases from the telecommunications thesaurus obtained through the courtesy of NTT (Nippon Telephone and Telegraph). Of these 3,000 phrases, approximately one half of them contain at least one example of a word written in *katakana*.

The *katakana* words in an electronic Japanese document are easily identified by their unique values within the JIS coding system, the Japanese version of ASCII. The problem then becomes one of identifying the English word from which the *katakana* word was derived. Fortunately, there is typically a sentence to sentence correspondence between Japanese text and its English translation so that the number of possible English candidate words is restricted to a single sentence. The first step in identifying the corresponding English word or words, is to first convert the *katakana* sequence into *roomaji*, i.e., the traditional English letters. While several variations of transliterating *katakana* into English exist, the method used here is known as the Hepburn system. The English transliterations for *katakana* in the Hepburn system are presented in Table V, where they are arranged in the same order as in Tables I thru IV.

Once the Japanese *katakana* has been transliterated into English characters, the process of correlating this text string with the English text from which it originated can be simply considered as a string pattern matching problem. Since the transliterated *katakana* will not typically match the spelling of the English text from which it originated, an approximated string matching algorithm based on dynamic programming is employed [19]. Thus the *katakana* word that has been transliterated into English ASCII characters is considered as the pattern for which one would like to find the best approximate match in the corresponding English sentence. Presumably, this should be the English word or words from which the *katakana* was derived. Unfortunately, two factors prevent this straightforward approach from performing satisfactorily. The first is due to the fact that the original transliteration from English to *katakana* results in a significant change in pronunciation and the second is due to the highly irregular spelling conventions in English. To compensate for the first factor, two techniques are employed. The first is to apply a set of possible spelling changes to the transliterated *katakana* that attempt to account for some of the major differences between Japanese and English phonology. These possible spelling changes are listed in Table VI and roughly correspond to the suggestions given to students of Japanese when trying to decipher *katakana* [6]. The second technique used to improve

TABLE V
The Modified Hepburn Romanization System

a	i	u	e	o
ka	ki	ku	ke	ko
sa	shi	su	se	so
ta	chi	tsu	te	to
na	ni	nu	ne	no
ha	hi	hu	he	ho
ma	mi	mu	me	mo
ya		yu		yo
ra	ri	ru	re	ro
wa				
ga	gi	gu	ge	go
za	ji	zu	ze	zo
da	ji	zu	de	do
ba	bi	bu	be	bo
pa	pi	pu	pe	po
kya		kyu		kyo
sha		shu		sho
cha		chu		cho
nya		nyu		nyo
hya		hyu		hyo
mya		myu		myo
rya		ryu		ryo
gya		gyu		gyo
ja		ju		jo
bya		byu		byo
pya		pyu		pyo
n̄				

Note: The English transliterations for the *katakana* characters are arranged in the same order as Tables I thru IV. The mora consonant ·ソ does not have a unique transliteration but depends on the following consonant.

TABLE VI
LIST OF SPELLING TRANSFORMATION RULES USED AFTER TRANSLITERATION OF A KATAKANA WORD USING THE HEPBURN ROMANIZATION

Rule				Example	
				katakana word	English origin
u	→	*	/ C - (C|#)	*shisutemu*	system
o	→	*	/ (d|t) - (C|#)	*doraiba*	driver
i	→	*	/ C - (C|#)	*matchi*	match
howa	→	wh	/ - V	*howaito*	white
(u|uu)	→	w	/ - V	*uuru*	wool
i	→	y		*iesu*	yes
ee	→	yV		*eeru*	Yale
y	→	*	/ - V	*kyaburetaa*	carburetor
a	→	Vr	/ (oo|o|e|i) - #	*hea*	hair
aa	→	Vr		*misutaa*	mister
a	→	Vr	/ - #	*konpyuta*	computer
oo	→	Vr		*pooku*	pork
s	→	c	/ - e	*sero*	cello
s	→	th		*sumisu*	Smith
z	→	j	/ - e	*zerii*	jelly
z	→	th		*mazaa*	mother
j	→	(d|z)		*ejison*	edison
b	→	v		*banira*	vanilla
h	→	f		*haaiisuto*	Far East
r	→	l		*rengusu*	length
ts	→	(t|z)		*tsurii*	tree

Note: The rule format $A \rightarrow B/C_1\_C_2$ implies that the string C_1AC_2 may be replaced by C_1BC_2. All lower case letters in the table represent themselves. The upper case letters and special characters have the following meanings: C: any consonant, V: any vowel, #: sequence boundary, *: null character, $(x_1|x_2)$: either x_1 or x_2.

the pattern matching process is a nonuniform weighting on the ASCII symbols, with more importance placed on the matching of consonants. This is motivated by the poor correspondence between the five vowel sounds in Japanese with the numerous vowel sounds in English, as well as by the large variations in spelling.

With the modification of the approximate pattern matching algorithm to include spelling change rules and the consonant weighing, the *katakana* that appear in the sample of 1,500 telecommunications phrases can be matched to their correct corresponding English translation with 90% accuracy. The remaining errors are primarily a result of the irregularities in English spelling. This effect can be removed by performing the string matching based on the pronunciation of the strings rather than their spelling. When both the *katakana* and the English text are converted to strings of phonemes, the error rate is reduced to less then 0.1%. The exact details of the algorithms employed and an analysis of their success rates is available in [7], however, the following example illustrates some of the issues involved in the pattern matching scheme. Consider the *katakana* string レーリー which when transliterated into English using the Hepburn system becomes "reeri". This string must be matched with the English phrase "Rayleigh scattering loss" for which the correct match is "Rayleigh". If the straightforward approximate pattern matching algorithm is applied then the following matches will result

r	a	y	l	e	i	gh	sca	t	t	e	r	i	ng	loss
r	e	e	r	-	i			r	e	e	r	i		
√	×	×	×	×	√			×	×	√	√	√		

which illustrates that an incorrect match with the middle of the word "scattering" has a lower number of mismatched characters. Note that matches with partial words must be considered since abbreviations are extremely common. Applying all possible spelling transformation rules to the Hepburn

transliteration results in

r	a	y	l	e	i	gh	sca	t	t	e	r	i	ng	loss
r	e	e	l	-	i			r	e	e	r	i		
√	×	×	√	×	√			×	×	√	√	√		
C			C								C			

which improves the situation, however, the result is still incorrect if one considers the number of mismatched characters. Consonant weighing can improve the likelihood of a correct match but still results in some ambiguity. This ambiguity can be completely resolved by converting both strings to their phonetic representations and applying the phonological equivalent of the spelling transformation rules, resulting in

ɹ	e	l	i	skætəɹɪŋ	lɔs
ɹ	eː	l	i		
√	√	√	√		

which is a perfect match.

V. Phonological Rules for Katakana Conversion

The *katakana* to English dictionary generated in the previous section can now be used to determine the specific phonetic changes that occur in English text when it is transliterated into *katakana*. This section describes an algorithm for automatically determining a set of rules that govern these phonetic changes. These rules represent the domain knowledge that the tutoring system must "learn" in order to provide efficient instruction. The inputs into the algorithm are a string of *katakana* and the English text from which it was derived. These two inputs are then converted into IPA symbols using Tables I thru IV for the *katakana* and the UNIX version of Webster's Seventh New Collegiate Dictionary [5] for the English.

The phonological rules that are determined by the algorithm presented here consist of three elements: a source phoneme, a target phoneme, and the context. The source phonemes are those that are derived from the *katakana* string (since this is the text that the student will be reading) and the target phonemes are those that are derived from the English text. The following standard notation is used to represent a phonological rule:

$$S \rightarrow T/C_1 - C_2$$

where S is a phoneme of the source language, T is a phoneme of the target language, and C_1 and C_2 represent the context in which S may be replaced by T. The symbols C_1 and C_2 are either single phonemes of phonological categories such as vowels, consonants, etc. Thus this rule implies that if the string C_1SC_2 occurs in the source language it may be transformed into the string C_1TC_2 in the target language.

Phonological rules of the type described above cannot be directly generated from the IPA symbols of the raw input strings because there is not a correspondence between the ith phoneme of the source string and the ith phoneme of the target string. This is primarily due to the different constraints on consonant clusters in Japanese as compared to English. As mentioned above, Japanese does not, in general, allow consonant clusters whereas English allows initial consonant clusters of two or three consonants, as in such words as "sky" and "spray", and either two, three, or four final consonants, as in "ask", "elks" and "glimpsed" [4]. To compensate for these phonological differences, the sequences of IPA symbols for the English text string is modified by inserting the symbol "*", which has no pronunciation, between any consecutive consonant phonemes and after sequences that end with a consonant. This modification greatly improves the correspondence between phoneme symbols in the *katakana* string with the symbols in the English string. However, it does create a difficulty with the mora nasal consonant cluster that is allowed in Japanese. It has been observed that the more nasal /N/ when followed by a consonant corresponds to either /m/, /n/, or /ŋ/ followed by a consonant in English. Therefore, in order to provide uniform treatment of these consonant clusters, the symbol "*" is also inserted into the *katakana* phoneme sequence between the mora nasal and the following consonant.

In addition to consonant cluster, one must also consider how to deal with non-identical vowel sequences in order to improve the correspondence between symbols in the source string and the target string. In Japanese, it is not obvious when to consider the second vowel of a sequence as a separate syllable as opposed to the second half of a diphthong[10]. This creates a difficulty when trying to match the English vowels. Therefore, the approach adopted here is that all consecutive vowel sequences are treated as a single unit. This is true even if it is known that the English vowel sequence represents two separate syllables. Therefore, after the appropriate modification of the input string by inserting "*" symbols, the strings are divided into units which consist of either a single consonant, a sequence of vowels or semi-vowels, or the symbol "*". If the number of partitions in the two strings are equal, then a set of phonological rules is generated.

The algorithm described above can be summarized by the following four steps:

1) Convert the *katakana* input string and the English input string into IPA symbols.
2) Modify the English IPA string by inserting a "*" symbol between any consecutive consonant phonemes and after a final consonant. Modify the Japanese IPA string by inserting a "*" symbol between the mora nasal /N/ and any following consonant.
3) Divide both symbol strings into partitions which include either a single consonant, a consecutive sequence of vowels and/or semi-vowels, or the symbol "*".

[10] a vowel in which there is a perceptible change in quality during a syllable.

TABLE VII
EXAMPLES OF THE PHONOLOGICAL RULES CREATED BY MATCHING THE IPA EQUIVALENT OF A KATAKANA WORD WITH THAT OF ITS ENGLISH ORIGIN

Rule†	Data		Example			
	Probability (%)	Number of Occurrences	*katakana* word	English word	IPA (JAP)	IPA (ENG)
* → w	63	5	シーケンス	sequence	ʃiːk(*)ensu	sikwens
a → ə	36	156	ディジタル	digital	diʤitaru	dɪʤətəl
a → ɚ	24	103	コンピュータ	computer	koNpjuːta	kəmpjutɚ
a → æ	34	145	セラミック	ceramic	seramiQku	səɹæmɪk
aː→ ɚ	69	43	サーマル	thermal	saːmaru	θɚməl
b → v	34	44	バルブ	valve	barubu	vælv
ʤ→ z	19	10	ビジー	busy	biʤiː	bɪzi
e → ə	28	58	ドキュメント	document	dokjumeNto	dakjəmənt
h → f	21	8	ヒューズ	fuse	hjuːzu	fjuz
i → ə	18	63	サービス	service	saːbisu	sɚvəs
i → ɪ	48	165	レジスタ	resistor	reʤisuta	ɹɪzɪstɚ
o → *	50	208	キーボード	keyboard	kiːboːdo	kiboɹd(*)
o → ə	20	81	セッション	session	seQʃoN	seʃən
o → a	15	64	プロセス	process	purosesu	pɹases
r → l	57	270	ライン	line	raiN	laɪn
s → θ	4	11	レングス	length	reNgusu	leŋθ
ʃ→ s	47	33	システム	system	ʃisutemu	sɪstəm
ʧ→ t	42	14	チューブ	tube	ʧuːbu	tjub
u → *	91	609	チェック	check	ʧeQku	ʧek(*)

† The context for these rules is not shown here for the sake of clarity.

4) If the number of partitions in the source string is equal to that of target string, generate a phonological rule which defines the transformation of a source partition into a target partition.

This process is illustrated in the following example, in which the *katakana* word "*shisutemu*" is compared with the word "system" from which it was derived:

システム → /ʃisutemu/ → /ʃ,i,s,u,t,e,m,u/

system → /sɪstəm/ → /s,ɪ,s,*,t,ə,m,*/.

The following eight rules are generated:

ʃ→ s/ # _ i t → t / u _ e

i → ɪ/ ʃ_ s e → ə/ t _ m

s → s / i _ u m → m / e _ u

u → * / s _ t u → * / m _ #.

Note that the somewhat obvious rules in which a phoneme is not changed are significant since the same phoneme may be modified in a different context. Additional information regarding the relative probability of occurrence for rules in an identical context is also maintained.

VI. STUDENT MODEL

The phonological rules determined in the previous section are used to define the domain knowledge that a student must acquire in order to become proficient in reading *katakana*. Some examples of these rules resulting from processing the 1,500 entries in the *katakana* to English telecommunications dictionary are illustrated in Table VII[11]. For each student that uses the intelligent tutoring system, a database is maintained which keeps information on which of these rules the student has mastered and which ones need further review. This information is obtained both by passive monitoring of the student requests for information while using the tutoring system [10] as well as through active testing of the student. The *katakana* words with which a student has difficulty are analyzed to determine which phonological rules are present and this is compared to the rules which occur in *katakana* that are easily comprehended. The tutoring system then attempts to assist the student in acquiring the unfamiliar rules by presenting lessons which include *katakana* that use these rules.

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.

[11] The context for these rules does not appear to have a significant effect on the student model and so is not included.

This particular student has no difficulty with the following words:

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

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

aasu → earth
rengusu → length
saamuru → 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 (see Table VII). 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 or 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 order to calculate such a student model, the tutoring system statistically analyzes a student's responses to the system. The knowledge base which a student must acquire in order to be proficient at reading *katakana* consists of the set of phonological rules which characterize the transformation of Japanese *katakana* to its English origin as discussed in Section V. Information about the student is gathered passively by simply noting the words for which he requests translations from the Japanese language tutoring system [10] (see Fig. 1). Analyzing a student's response is, therefore, relatively difficult for the tutor. 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, where π is the probability that the student knows the rule R. The student's current knowledge state can be estimated by the posterior probability density, $p(\pi|x)$, which is computed from the model density, $p(x|\pi)$, and the prior density, $p(\pi)$, by using Bayes' theorem, i.e.,

$$p(\pi|x) \propto p(x|\pi)p(\pi). \tag{1}$$

The mean value of this $p(\pi|x)$ is calculated for each rule in the knowledge base and is used by the tutor as an estimate of the student's knowledge of these rules.

The most common method for computing prior density is to approximate one's prior belief 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 distribution. When assuming that the prior density function is chosen as the beta distribution with parameters a and b, the posterior density function becomes the beta distribution with parameters of

$$p = x + a \tag{2}$$

and

$$q = n - x + b. \tag{3}$$

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

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

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 [15] 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 pretest in order to get prior information about the student's knowledge [2]. 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 ($a \approx 0$ and $b = 1$).
- The student's knowledge state is independent of the prior information ($a \approx 0$ and $b \approx 0$).

A comparison of the results of using uniform and non-uniform priors for the student discussed above is presented in Table VIII. In the following section, the lack of prior

TABLE VIII
EFFECT OF VARIOUS PRIOR DENSITY FUNCTIONS OF THE MEAN VALUE OF THE POSTERIOR FUNCTION IN THE BINOMIAL MODEL

Rule†	$a=1$ $b=1$	$a\approx 0$ $b=1$	$a\approx 0$ $b\approx 0$
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
aɪ → ɚ	0.25	0.00	0.00
s → θ	0.25	0.00	0.00

† The context for these rules is not shown here for the sake of clarity.

TABLE IX
THE PROBABILITY OF COMPREHENSION COMPUTER BY ASSIGNING THE SAME WEIGHTS TO ALL OF THE RULES

	Probability of comprehension		
Rule†	1st year	2nd year	3rd year
s → θ	0.07	0.09	0.38
b → v	0.23	0.32	0.65
aː → ɑː	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

† The context for these rules is not shown here for the sake of clarity.

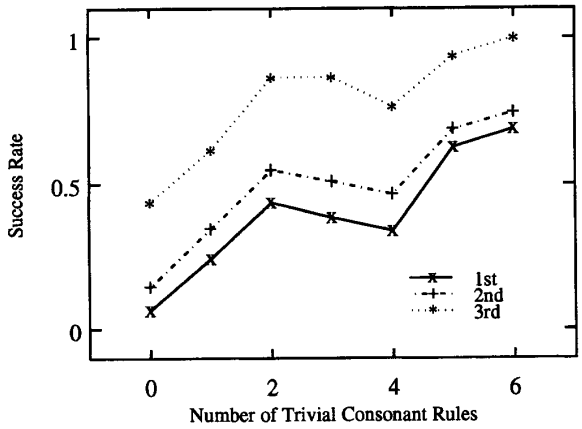


Fig. 2. The effect of trivial consonant rules on the probability of comprehension of a katakana word for 1st, 2nd, and 3rd year Japanese students.

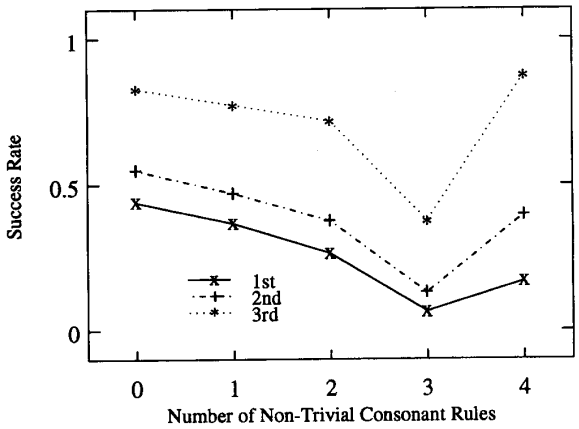


Fig. 3. The effect of non-trivial consonant rules on the probability of comprehension of a katakana word for 1st, 2nd, and 3rd year Japanese students.

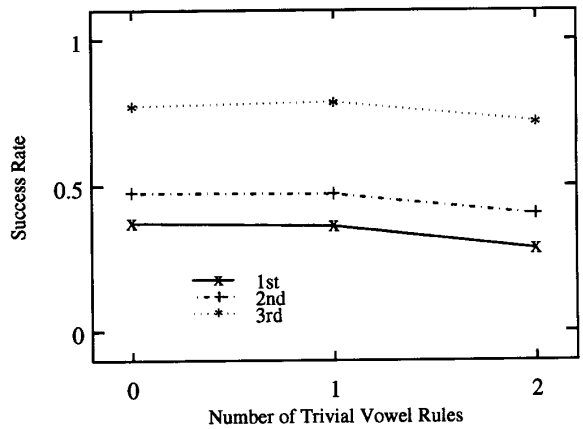


Fig. 4. The effect of trivial vowel rules on the probability of comprehension of a katakana word for 1st, 2nd, and 3rd year Japanese students.

information is compared to assumptions about its probable distribution based on such factors as the frequency of a rule and on the extent of the phonological transformation.

VII. THE EFFECTS OF RULE FREQUENCIES

While the binomial model is shown to be reasonalby 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 transformation 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 solely on the probability of comprehension as shown in Table IX.

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 limited correlation between the probability of comprehension and the extent of the phonetic modification of a transformation rule. As would be expected, Fig. 2 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. 3 that students have more trouble understanding words that use large numbers of non-trivial consonant rules in the transliteration process[12]. It is interesting to note, however, that similar data for the vowel rules, depicted in Fig. 4 and Fig. 5, do not exhibit this correlation. This is probably due to the large disparity

[12]The data point for four non-trivial consonant rules is not statistically significant since it is due to a single case (the word "inflation") which is more easily understood due to two occurrences of the rule $N \rightarrow n$ which is quickly learned because of its high frequency of occurrence.

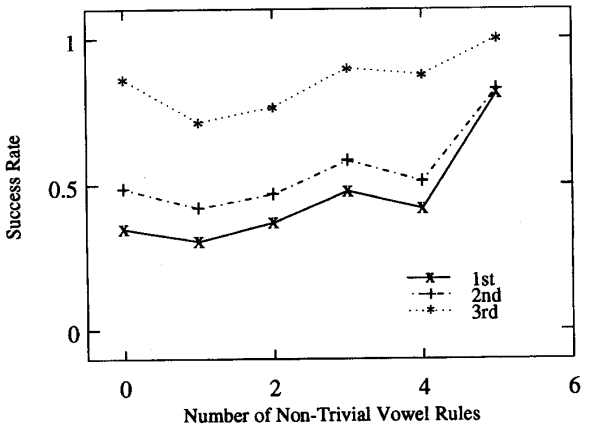


Fig. 5. The effect of non-trivial vowel rules on the probability of comprehension of a katakana word for 1st, 2nd, and 3rd year Japanese students.

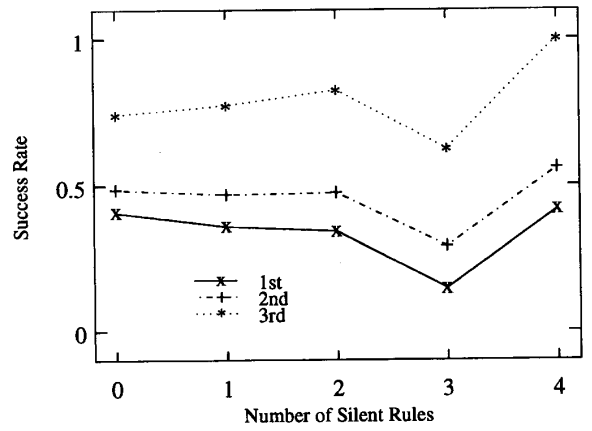


Fig. 6. The effect of silent rules on the probability of comprehension of a katakana word for 1st, 2nd, and 3rd year Japanese students.

between the number of Japanese and English vowel sounds 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. 6). The most dominant of the silent rules is the rule u $\rightarrow$ * 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 $0 \leq w \leq 1$ associated with each rule. The value of w represents the likelihood that this will contribute to any difficulty with comprehension of words that contain this rule. This value of w is then used to modify (4) so

TABLE X
RELATIVE AND ABSOLUTE FREQUENCIES FOR THE PHONOLOGICAL RULES

Rule†	relative frequency	absolute frequency
s $\rightarrow$ θ	0.04	0.01
b $\rightarrow$ v	0.34	0.05
e $\rightarrow$ e	0.55	0.14
r $\rightarrow$ l	0.57	0.32
aː $\rightarrow$ ɔː	0.69	0.05
u $\rightarrow$ *	0.91	0.73
t $\rightarrow$ t	0.99	0.32
k $\rightarrow$ k	1.00	0.22

† The context for these rules is not shown here for the sake of clarity.

that the probability of rule comprehension is calculated using

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

Thus rules with a large value of w are assumed to be trivial and their probabilities are not adversely affected for a 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 rule's 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 X. 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 rule's difficulty, the average of the two values is currently being used. Table XI shows the resulting probabilities of comprehension and illustrates a much closer correspondence to empirical evidence, particularly with respect to the trivial consontant rules and the silent rule, as compared to Table IX. 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.

VIII. CONCLUSION

The goal of this work was the development of a model for representing a student's proficiency in reading *katakana*. This

TABLE XI
THE PROBABILITY OF COMPREHENSION COMPUTER BY ASSIGNING DIFFERENT WEIGHTS TO EACH RULE

	Probability of comprehension		
Rule†	1st year	2nd year	3rd year
s → θ	0.01	0.03	0.06
b → v	0.07	0.12	0.33
aː → 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

† The context for these rules is not shown here for the sake of clarity.

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. To accomplish this goal an algorithm was first developed to automatically generate a *katakana* to English dictionary from raw Japanese text and its English translation. This dictionary is then used as input into a second algorithm that is used to generate the phonological rules that govern the transformations that English words undergo when being transliterated into *katakana*. These rules are used as the domain knowledge base against which a student's performance is measured.

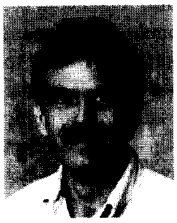
It was illustrated that the probability of a student's assimilation of any phonological rule is strongly correlated to both the relative and absolute frequency of that rule's occurrence, as well as the extent of the phonetic modification. It is shown that combining such factors with the binomial model 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*.

ACKNOWLEDGMENT

We would like to express our sincerest gratitude to Prof. Kazumi Hatasa and Prof. Yukiko Hatasa of the Foreign Languages and Literatures Department of Purdue University for their comments and many useful insights on Japanese phonology. We would also like to thank the anonymous reviewers for their suggestions which have improved this paper.

REFERENCES

[1] J. S. Brown and R. R. Burton, "Diagnostic models for procedural bugs in basic mathematical skills," *Cognitive Science,* vol. 2, pp. 155–192, 1978.
[2] R. R. Burton, "Diagnosing bugs in a simple procedural skill," in D. Sleeman and J. S. Brown, editors, *Intelligent Tutoring Systems,* pp. 157–183, New York, NY: Academic, 1982.
[3] B. Carr and I. Goldstein, *Overlays: A Theory of Modeling for Computer Aided Instruction,* Cambridge, MA: MIT, Artificial Intelligence Laboratory, 1977.
[4] J. C. Catford, *A Practical Introduction to Phonetics*, New York, NY: Oxford Univ. Press, 1988.
[5] P. B. Gove, editor, *Webster's Seventh New Collegiate Dictionary,* Springfield, MA: G. & C. Meriam Company, 1963.
[6] E. H. Jorden and H. I. Chaplin, *Reading Japanese,* New Haven, CT: Yale Univ. Press, 1976.
[7] Y.-S. Kang, "A Knowledge Base Acquisition for a Japanese Language Intelligent Tutoring System," *Ph.D. Thesis,* Purdue Univ., W. Lafayette, IN, 1994.
[8] Kokuritsu Kokugo Kenkyuujo, *Gendai-zasshi 90shu no yoogo yooji (3),* Report 25, 1964.
[9] R. W. Lawler and M. Yazdani, editors, *Artificial Intelligence and Education,* Norwood, NJ: Ablex, 1987.
[10] A. A. Maciejewski and N. K. Leung, "The Nihongo Tutorial System: An intelligent tutoring system for technical Japanese language instruction," *J. Computer Assisted Language Learning and Instruction Consortium,* vol. 9, pp. 5–25, 1992.
[11] H. Mandl and A. Lesgold, editors, *Learning Issues for Intelligent Tutoring Systems,* New York, NY: Springer-Verlag, 1988.
[12] D. O. Mills, R. J. Samuels and S. L. Sherwood, "Technical Japanese for scientists and engineers: Curricular options," *Technical Report MITJSTP WP 88-02,* MIT, Cambridge, MA, 1988.
[13] O.-C. Park, R. S. Perez and R. J. Seidel, "Intelligent CAI: Old wine in new bottles, or a new vintage?," in G. Kearsley, editor, *Artificial Intelligence and Instruction,* pp. 11–45, Reading, MA: Addison Wesley, 1987.
[14] M. C. Polson and J. J. Richardson, editors, *Foundations of Intelligent Tutoring Systems,* Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
[15] S. J. Press, *Bayesian Statistics: Principles, Models, and Applications,* New York, NY: John Wiley & Sons, 1989.
[16] J. W. Rickel, "Intelligent computer-aided instruction: A survey organized around system components," *IEEE Trans. Syst., Man, Cybern.,* vol. SMC-19, pp. 40–57, 1989.
[17] M. Shibatani, *The Language of Japan,* New York, NY: Cambridge Univ. Press, 1990.
[18] J. T. Vance, *An Introduction to Japanese Phonology,* Albany, NY: State Univ. of New York Press, 1987.
[19] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM,* vol. 21, pp. 168–173, 1974.
[20] E. Wenger, *Artificial Intelligence and Tutoring Systems,* Los Altos, CA: Morgan Kaufmann, 1987.



Anthony A. Maciejewski (S'82–M'87) received the B.S.E.E., M.S., and Ph.D. degrees in electrical engineering from the Ohio State University, Columbus, in 1982, 1984, and 1987 respectively. Since 1988 he has been with the School of Electrical Engineering at Purdue University, West Lafayette, IN, where he is currently an Associate Professor. His primary research interests center on the simulation and control of kinematically redundant robotic systems.

Yun-Sun Kang received a B.S. degree in control and instumentation engineering from Seoul National University, Korea, in 1985 and an M.S. degree in electrical engineering from North Carolina State University, Raleigh, NC, in 1987. He is currently a Ph.D. candidate at Purdue University, West Lafayette, IN. His primary research interests are intelligent tutoring systems and knowledge acquisitions.