Behavior Research Methods
2009, 41 (2), 575-583
doi:10.3758/BRM.41.2.575

An online database of phonological representations for Mandarin Chinese

Xiaowei Zhao
University of Richmond, Richmond, Virginia

and

Ping Li
Pennsylvania State University, University Park, Pennsylvania

A Web-based database is developed to provide psycholinguists with a large-scale phonological representation system for all Mandarin Chinese monosyllables. The construction of the system is based on the slot-based phonological pattern generator (PatPho), with an adequate consideration of the language-specific features of the Chinese phonology. Users can retrieve the relevant phonological representations through an interactive query system on the Web. The query outcomes can be saved in a number of formats, such as Excel spreadsheets, for further analyses. This representation system can be used for a variety of purposes—in particular, connectionist language modeling and, more generally, the study of Chinese phonology.

Researchers in connectionist modeling of language have for some time been concerned with the issue of phonological representations of the relevant linguistic input to the model. How to faithfully represent the phonological patterns of words and the differences between words in a language has been discussed since the pioneering work of Rumelhart and McClelland (1986) on the acquisition of the English past tense. Recent development in this field favors the approach in which a word's pronunciation is coded on a slot-based representation, while taking into consideration the articulatory features of phonemes in the word (Joanisse & Seidenberg, 1999; MacWhinney & Leinbach, 1991; Plunkett & Juola, 1999). In particular, the phonology of a word is encoded in terms of a template with a fixed set of slots; each phoneme of the word is assigned to a different slot, depending on which syllable it belongs to and at which position it appears in the syllable, such as the onset, nucleus, or coda.

Most recently, on the basis of this idea of syllabic templates, Li and MacWhinney (2002) introduced a phonological pattern generator (PatPho) for connectionist modeling. PatPho is able to represent English words with variable length (up to three syllables) in a syllabic template of CCCVVCCCVVCCCVVCCC, with Cs representing consonants, Vs representing vowels, and each CCCVVCCC representing one syllable. This system accurately captures the phonological features of English words and has been successfully applied in our connectionist models of child language development (Li, Farkas, & MacWhinney, 2004; Li, Zhao, & MacWhinney, 2007; Zhao & Li, in press).

The phonological representation of words is also an important issue in the connectionist study of other languages. For example, Chinese has an ideographic writing system, and it has always been a difficult problem for connectionist models to correctly represent the phonology of Chinese characters. To solve this problem, different researchers have developed different representational systems (e.g., Hsiao & Shillcock, 2004; Xing, Shu, & Li, 2004). Although these systems have greatly improved our understanding of language acquisition and language processing in Chinese, there are some problems with these systems—notably, in terms of their generalizability for computational models other than their own.

In Hsiao and Shillcock's (2004) work, the pronunciation of Chinese monosyllabic characters was represented by a 27-dimension binary vector. In their coding, the first 14 dimensions of the vector represent the phonetic features of an initial constant, the next 8 dimensions represent those of a nucleus vowel, 3 other dimensions represent the final constant, and the final 2 dimensions represent four tones in Mandarin Chinese. A significant advantage of their system is the parsimony of the binary codes (0 or 1), which allows their computational model to be tractable. The parsimony, however, introduces certain problems that may limit the accuracy of their representations. For example, only a single nucleus vowel can be represented in their system, which is inconsistent with Chinese phonology, which allows two or even three vowels to be clustered together (i.e., diphthongs or triphthongs). Hsiao and Shillcock's representations therefore cannot capture the vowel structure in Chinese. Another problem is related

X. Zhao, xzhao2@richmond.edu

to the tones in Mandarin Chinese. Because there are five tones (including a neutral tone) in Mandarin Chinese, the two-node binary representations in Hsiao and Shillcock's system are unable to represent all the five tones.

Xing et al.'s (2004) phonological representation of Chinese characters was based on PatPho. It splits Chinese monosyllables into three parts—initial, final, and tone—and uses six slots to represent the tone and the phonemes that can occur in different positions of the syllable. Each slot consists of five units, and each unit can be assigned a real value between 0.0 and 1.0 to represent a specific articulatory feature of the phoneme. In total, a 30-dimensional feature vector with real values can be used to represent the pronunciation of a Chinese character. This system, as compared with Hsiao and Shillcock's (2004), can successfully code the diphthongs and triphthongs in its representation and is able to capture the phonetic features of Chinese syllables.

One minor problem with Xing et al.'s (2004) system is that five units are used to represent a phoneme or a tone. However, as we will discuss below, three nodes are sufficient to represent the features of a phoneme, and a single unit with varying real numbers is able to represent all the five tones. As such, Xing et al.'s system has some redundancy, and there is room to reduce its computational complexity. This representation also heavily relies on the *Pinyin* system (the standard romanization system for Mandarin Chinese; Institute of Linguistics of the Chinese Academy of Social Sciences, 2002). The Pinyin system is simple and easy to learn, but its simplicity also causes the problem that many different phonemes have to be represented by the same letter. For example, the Pinyin letter "i" could represent three phonemes that are similar but different, according to its varying positions in a syllable. A similar situation holds for "a," "o," "e," and so on, since phonemic differences are not clearly represented in the system.

Although connectionist modeling of Chinese has become an increasingly important topic in psycholinguistic research, there has not yet been a convenient tool with which investigators can accurately generate large-scale phonological representations of Chinese characters. The issue is even more serious for researchers who are not familiar with the Chinese language but, nevertheless, want to do comparative studies, as well as for investigators whose native language is Chinese but who are not trained in the Pinyin system. It would be convenient for these investigators to obtain simple, easily accessible, and vector-based representations of Chinese pronunciations. Our online phonological database of Chinese characters is designed to help researchers to do just that.

Here, we introduce a phonological representation system that attempts to accurately represent all the possible pronunciations of Mandarin Chinese monosyllables (spoken units corresponding to written characters). This system builds on the idea of a slot-based syllabic template, as adopted by PatPho (Li & MacWhinney, 2002), and is an improvement of the phonological representation system of Xing et al. (2004). Although the template includes only one syllable (because almost all the Mandarin characters are monosyllables), our system can be easily extended to represent Chinese words with more than one syllable. Our template can capture phonological similarities in Chinese characters more accurately than can previous systems and, at the same time, reduce the computational complexities. Specifically, each Pinyin code of a Chinese character is transcribed to its corresponding International Phonetic Alphabet (IPA) format, and the phonemes are aligned in a fixed-length template, with either real or binary values representing the articulatory features of the phonemes. To make the system available for research, we have provided our representations online at cogsci.richmond.edu. Like PatPho, our system is not a phonological learning system for Chinese, but a representational tool that allows researchers to compose input word representations that can capture the similarity structure of lexical phonology. It is our hope that our database provides a starting point for psycholinguistic and connectionistic models in the cross-linguistic comparative study of Chinese.

METHOD

Features of Chinese Phonology

PatPho was originally developed to characterize the phonological feature of English words. It uses a trisyllabic template with 18 phonemic slots to represent the pronunciation of an English word. The template is CCCVV-CCCVVCCCVVCCC, with Cs representing consonants and Vs representing vowels.

In the present study, we must revise PatPho so that it is able to adequately consider the language-specific phonological features of Chinese. First, a unique feature of the Chinese phonology is that all characters' pronunciation is made of only one single syllable (Huang & Liao, 2001), and therefore, we need space only for a monosyllable in our new system. Thus, instead of using a trisyllabic template with 18 phonemic slots, we used a template with 5 phonemic slots and 1 additional tonal slot in our representation to characterize a Chinese monosyllable. Second, in English, two or more consonants can occur together as a consonant cluster (e.g., /pl/ in *play*; /spr/ in *spring*), but such consonant clusters are not allowed in the Chinese phonological system. In addition, in Chinese, consonants can occur only at the beginning or the end of a single syllable, and the consonant at the end of a syllable can only be a nasal /n/ or /ŋ/. Given these phonological constraints in Chinese, we therefore replaced the consonant cluster representation of CCC in the original PatPho with a single C in our new system, one at the beginning and one at the end of the syllable, respectively. Third, in both English and Chinese, vowels can occur next to each other in a single syllable without intervening consonants (*diphthongs* or *gliding vowels*). However, the combination of three vowels (or *triphthongs*) is very rare in English[1] but happens frequently in Chinese (e.g., /uai/, /iɑu/, /uei/). So we replaced the cluster representation of VV in the original PatPho with VVV in our new system to leave enough space for Chinese triphthongs. Finally, Chinese is a tonal language, in which the pitch and duration of a syllable carry lexical meanings. In standard Mandarin Chinese, there are five tones: high-level (Tone 1), mid-

Table 1
Look-Up Table Between Pinyin Symbols and International Phonetic Alphabet (IPA) Symbols

Pinyin	IPA	Pinyin	IPA
a	a	ai	ai
o	o	ei	ei
e	ɤ	ui	uei
i	i	ao	ɑu
u	u	ou	əu
ü	y	uo	uo
b	p	an	an
p	p^h	en	ən
m	m	in	in
f	f	un	uən
d	t	ün	yn
t	t^h	ie	iɛ
n	n	üe	yɛ
l	l	er	əɻ
g	k	ia	ia
k	k^h	ua	ua
h	x	ang	ɑŋ
j	tɕ	eng	əŋ
q	$tɕ^h$	ing	iŋ
x	ɕ	ong	uŋ
z	ts	uai	uai
c	ts^h	ian	ian
s	s	uan	uan
zh	tʂ	üan	yan
ch	$tʂ^h$	iou	iəu
sh	ʂ	uang	uɑŋ
r	ʐ	ueng	uəŋ
y	i	iang	iɑŋ
w	u	iong	iuŋ
ng	ŋ		

Note—*i* after *z*, *c*, and *s* is pronounced as ɿ, and after *zh*, *ch*, *sh*, and *r* is pronounced as ʅ.

rising (Tone 2), low-dipping (Tone 3), high-falling (Tone 4) tones, and a neutral tone (no pitch pattern, usually shorter in duration). An extra slot T was thus used in our system to represent tones.

In sum, to represent a monosyllable in Chinese, we used a template with six phonemic plus tonal slots: CVVVCT, where each C represents a consonant, each V a vowel, and T the tone of the syllable.

From Pinyin to Phonemes

The Pinyin system uses Roman alphabets to spell the sounds of Chinese characters, but it was not designed as an accurate phonetic transcription system. The phonetic transcription system widely accepted by linguists is the IPA, a system designed for transcribing the phonemes of the world's languages. For example, the letter *b* in Pinyin represents the IPA voiceless consonant /p/, rather than the voiced consonant /b/; similarly, *d* represents the phoneme /t/, rather than voiced consonant /d/. The situation is even worse for vowels. For example, the Pinyin symbol *e* could represent four similar but different IPA phonemes according to its varying positions in a syllable: /ɤ/ as in a single vowel, /e/ as in diphthong *ei* or triphthong *u(e)i*, /ɛ/ as in *ie*, and /ə/ as in *en*, *eng*. Table 1 provides the translation table for the Pinyin and IPA symbols.

To accurately capture Chinese phonology in our system, we have to first retranscribe Pinyin symbols to IPA-based phonemic symbols. A free software called *py2ipa* was used for this purpose (Xu, 2008; py2ipa.sourceforge.net). The rules used by the software are based on the textbook *Modern Chinese* (Wang, Lu, Fu, Ma, & Su, 2004).

Phonological Features

In total, we have 34 phonemes in our representation of the Mandarin phonology. The phonemes and their articulatory features can be seen in Table 2. We used as our basis phonetic features similar to those in the PatPho system (Li & MacWhinney, 2002, which was based on Ladefoged, 1982), while also considering certain language-specific characteristics of Mandarin Chinese, as outlined in *A Course in Phonetics* (Lin & Wang, 1992).

In Table 2, the first column represents the ASCII symbols of the phonemes, and the second column the IPA symbols. The third through fifth columns are the three proposed dimensions (D1–D3). Here, D1 represents whether the phoneme is a vowel or a consonant. If it is a vowel, this dimension also represents the lip roundness of the vowel. If the phoneme is a consonant, the dimension also differentiates voicing (i.e., voiced vs. voiceless).

Table 2
Representation of 34 Mandarin Phonemes by Three Articulatory Dimensions (D1–D3)

Phoneme*	IPA	D1	D2	D3
A	ɑ	unrounded vowel	back	low
a	a	unrounded vowel	central	low
E	ɛ	unrounded vowel	front	mid-low
e	e	unrounded vowel	front	mid
o	o	rounded vowel	back	mid-high
Y	ɤ	unrounded vowel	back	mid-high
@	ə	unrounded vowel	central	mid-high
u	u	rounded vowel	back	high
y	y	rounded vowel	front	high
i	i	unrounded vowel	front	high
~	ɿ**	unrounded vowel	tip-front	high
^	ʅ**	unrounded vowel	tip-back	high
t	T	voiceless	alveolar	stop
T	t^h	voiceless/aspirated	alveolar	stop
p	p	voiceless	bilabial	stop
P	p^h	voiceless/aspirated	bilabial	stop
k	k	voiceless	velar	stop
K	k^h	voiceless/aspirated	velar	stop
n	n	voiced	alveolar	nasal
m	m	voiced	bilabial	nasal
N	ŋ	voiced	velar	nasal
l	l	voiced	alveolar	lateral
s	s	voiceless	alveolar	fricative
f	f	voiceless	labiodental	fricative
X	ɕ	voiceless	palatoalveolar	fricative
r	ʐ	voiced	retroflex	fricative
S	ʂ	voiceless	retroflex	fricative
x	x	voiceless	velar	fricative
z	ts	voiceless	alveolar	affricative
c	ts^h	voiceless/aspirated	alveolar	affricative
j	tɕ	voiceless	palatoalveolar	affricative
q	$tɕ^h$	voiceless/aspirated	palatoalveolar	affricative
Z	tʂ	voiceless	retroflex	affricative
C	$tʂ^h$	voiceless/aspirated	retroflex	affricative

*ASCII symbols for phonemes. **ɿ and ʅ are two vowels in Mandarin but are not listed in standard International Phonetic Alphabet (IPA) symbols. ɿ sounds like a prolonged *zzz*, and ʅ sounds like a prolonged American *r* sound.

Table 3
Conversion of Phonological Dimensions (D1–D3, Tones) to Numerical Representations

D1		D2		D3		Tones	
Unrounded vowel	.100	tip-front	.030	high	.100	neutral (0)	.000
Rounded vowel	.175	tip-back	.060	mid-high	.185		
		front	.100	mid	.270	high level (1)	.075
		central	.175	mid-low	.355		
		back	.250	low	.444	mid-rising (2)	.150
Voiced	.750	bilabial	.450	nasal	.644	low dipping (3)	.225
Voiceless/aspirated	.925	labiodental	.560	stop	.733		
Voiceless	1.000	alveolar	.670	fricative	.822	high falling (4)	.300
		retroflex	.780	affricative	.911		
		palatoalveolar	.890	lateral	1.000		
		velar	1.000				

Unlike English, Mandarin Chinese has several voiceless consonants with aspiration, such as /tʰ/, /pʰ/, and /kʰ/ (Lin & Wang, 1992), and this feature of aspiration is represented by D1 too. Like PatPho, D2 and D3 lump together two kinds of features: the tongue position for the vowels and the manner of articulation for the consonants. Mandarin Chinese also has the so-called apical vowels, which are very rare in other languages. For example, /ɿ/ and /ʅ/ are two apical vowels, and they are produced with the tip of tongue; we represent their tongue positions as tip-front and tip-back in D2. Finally, the tongue height for vowels and the place of articulation for consonants are represented in D3.

Feature Conversion

To convert the articulatory features above to numerical representations for each phoneme, we replaced the features with numerical real values, scaled between the range of 0 and 1. These numerical values were carefully chosen to adequately represent the similarities and differences among the articulatory features: The closer the numerical values are, the more similar the articulatory features should be, as shown in Table 3. For example, for D1, the values for the rounded vowel (.1) and unrounded vowel (.175) are close to each other, and the values for the voiced consonants (.75), voiceless consonants (.925), and aspirated consonants (1.0) are close to each other. By this numerical representation in D1, we can clearly distinguish the two groups of vowels and consonants. In addition, the interval between aspirated and voiceless consonants (1.0 − .925 = .075) is set to be smaller than that between voiced and voiceless consonants (.925 − .75 = .175), which takes into consideration the fact that all aspirated consonants in Chinese are voiceless. Finally, the five tones in Mandarin Chinese are replaced with numerical values (with a .075 interval): neutral (0), high level (.075), mid-rising (.15), low dipping (.225), and high falling (.3).

According to the coding scheme in Table 3, we can represent each of the 34 phonemes in three numerical real values, as in Table 4. To test the accuracy and effectiveness of our representation system, we conducted a cluster analysis on the values of the phoneme representations, as shown in Figure 1. The dendrogram clearly indicates that our representation system captures the phonological similarities and differences among these phonemes: Phonetically similar phonemes were grouped together, whereas vowels and consonants were clearly classified into two large categories. Within each category of vowels or consonants, similar phonemes were clustered in adjacent branches, which means that the Euclidean distances between their numerical values in the representations are short. For example, the only difference between /p/ and /pʰ/ is the absence or presence of

Table 4
Three-Dimensional (D1–D3) Representation of Mandarin Phonemes

Phoneme*	IPA	D1	D2	D3
A	ɑ	.100	.250	.444
a	a	.100	.175	.444
E	ɛ	.100	.100	.355
e	e	.100	.100	.270
o	o	.175	.250	.185
Y	ʏ	.100	.250	.185
@	ə	.100	.175	.185
u	u	.175	.250	.100
y	y	.175	.100	.100
i	i	.100	.100	.100
~	ɿ**	.100	.030	.100
^	ʅ**	.100	.060	.100
t	t	1.000	.670	.733
T	tʰ	.925	.670	.733
p	p	1.000	.450	.733
P	pʰ	.925	.450	.733
k	k	1.000	1.000	.733
K	kʰ	.925	1.000	.733
n	n	.750	.670	.644
m	m	.750	.450	.644
N	ŋ	.750	1.000	.644
l	l	.750	.670	1.000
s	s	1.000	.670	.822
f	f	1.000	.560	.822
X	ɕ	1.000	.890	.822
r	ɻ	.750	.780	.822
S	ʂ	1.000	.780	.822
x	x	1.000	1.000	.822
z	ts	1.000	.670	.911
c	tsʰ	.925	.670	.911
j	tɕ	1.000	.890	.911
q	tɕʰ	.925	.890	.911
Z	tʂ	1.000	.780	.911
C	tʂʰ	.925	.780	.911

*ASCII symbols for phonemes. **ɿ and ʅ are two vowels in Mandarin but are not listed in standard International Phonetic Alphabet (IPA) symbols. ɿ sounds like a prolonged *zzz*, and ʅ sounds like a prolonged American *r* sound.

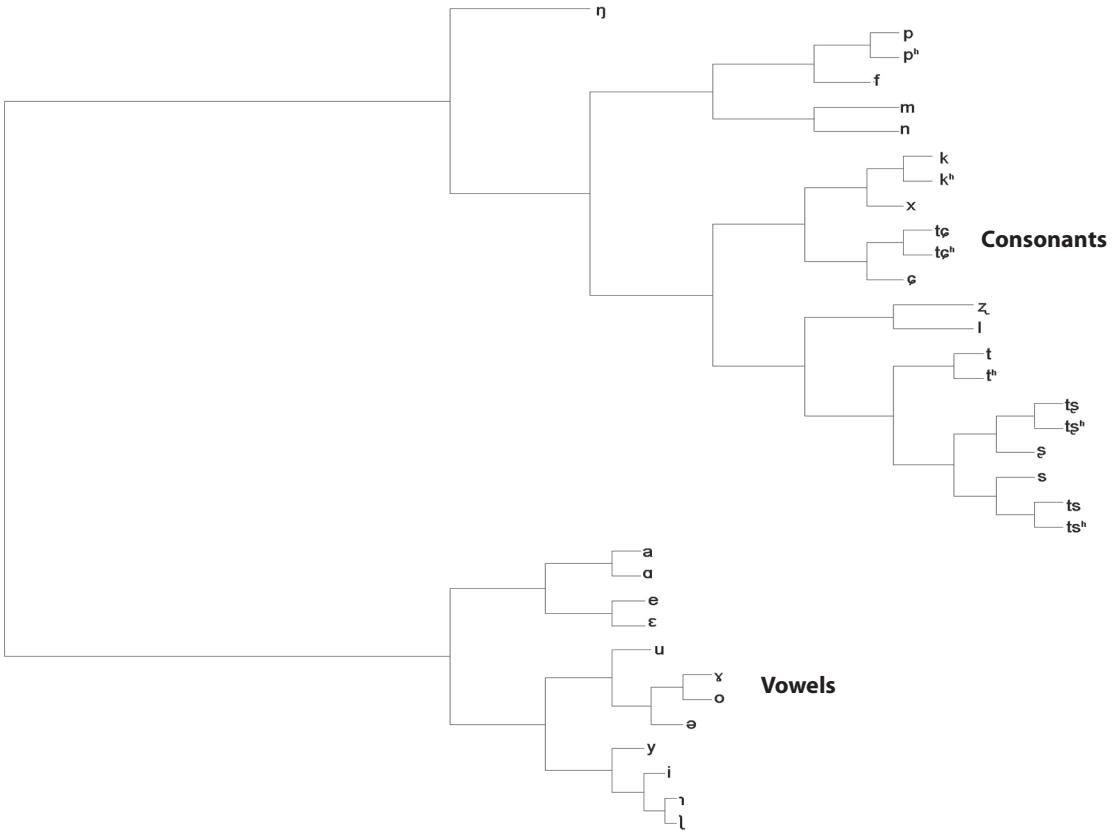

Figure 1. Dendrogram of the phonemic distance in a hierarchical cluster analysis. Clear clusters can be discerned on the dendrogram's upper and lower branches (consonants vs. vowels) and within each branch (e.g., aspiration vs. no aspiration).

aspiration, and their numerical values in the representation are very similar; therefore, /p/ and /pʰ/ appeared in the same branch of the dendrogram.

Many well-known neural network models of language were originally developed to handle binary codes (Plunkett & Juola, 1999; Sejnowski & Rosenberg, 1988). In order to increase the generalizability of our system, we also provided an alternative for coding the phonemic features with binary values, as shown in Table 5. Here, D1 was represented by two binary units, whereas D2, D3, and the tonal feature were represented by three binary units to adequately represent all the features. In this scheme, a phoneme is represented by eight binary units. As was discussed earlier, binary codes can reduce the computational complexity but tend to generate less accurate representations, as compared with representations that contain real values.

Syllabic Vector Representations

In the present system, we used a syllabic template with five phonemic slots and a tonal slot (CVVVCT) to represent a monosyllable in Mandarin Chinese as a

Table 5
Conversion of Phonological Dimensions (D1–D3, Tones) With Binary Codes

D1		D2		D3		Tones	
Unrounded vowel	0 1	tip-front	0 1 0	high	0 0 1	neutral (0)	0 0 1
Rounded vowel	1 0	tip-back	0 1 1	mid-high	0 1 1		
		front	0 0 1	mid	1 1 1	high level (1)	0 1 1
		central	1 1 1	mid-low	1 1 0		
		back	1 0 0	low	1 0 0	mid-rising (2)	1 1 1
Voiced	1 0	bilabial	0 0 1	nasal	0 0 1	low dipping (3)	1 1 0
Voiceless/aspirated	1 1	labiodental	0 1 0	stop	0 1 1		
Voiceless	0 1	alveolar	0 1 1	fricative	1 1 1	high falling (4)	1 0 0
		retroflex	1 1 1	affricative	1 1 0		
		palatoalveolar	1 1 0	lateral	1 0 0		
		velar	1 0 0				

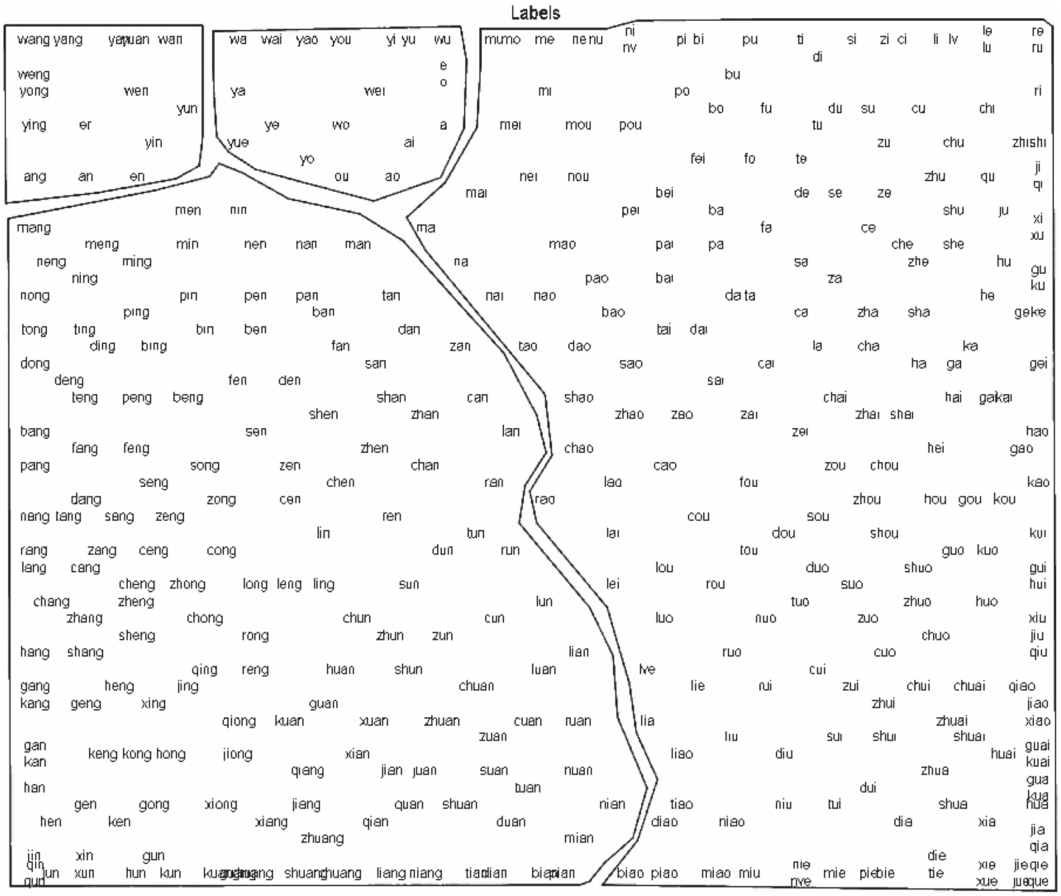

Figure 2. Emerged phonological structure in a self-organizing map model after 200 epochs of training on the numerical representations of 401 Mandarin Chinese monosyllables. The map size is 60 × 50. It can be seen that on the left side of the map, all the syllables ending with consonants /n/ and /ŋ/ (represented by Pinyin symbols *n* and *ng*) are grouped together; in the middle of the top side, the syllables without consonants (such as triphthongs /iɑu/ and /uei/ and the single vowels such as /a/, /o/, and /ɤ/) are clustered together. The right side of the map includes all the syllables that start with a consonant and end with a vowel. Within this area, the syllables that start with the same consonant (such as /m/) or end with the same vowels or diphthongs (such as /ɑu/) are often closely located.

numerical vector. Specifically, the phonemes are sequentially arranged in the phonemic slots according to their order of occurrence in a syllable and according to their status as consonants or vowels. For example, the monosyllable /lan/ would be encoded as *laVVn*, /tai/ as *taiVC*, and /pai/ as *paiVC*. Accordingly, the real-value vector representations of these syllables can be shown as follows:

/*laVVn*/: .75 .67 1.0 .1 .175 .444 0 0 0 0 0 0 .75 .67 .644;

/*taiVC*/: 1.0 .67 .733 .1 .175 .444 .1 .1 .1 0 0 0 0 0;

/*paiVC*/: 1.0 .45 .733 .1 .175 .444 .1 .1 .1 0 0 0 0 0.

Here, the empty phonemic slots (C or V in the symbol codes) are replaced by zeros in the numerical vectors. The vector representations should capture the overall similarities of the phonetic structure of words, as can be seen in the examples /tai/ and /pai/.

The vector representation as described here has the capacity to represent all the possible pronunciations of Mandarin Chinese characters, including pronounceable but nonexistent monosyllables (e.g., /tei/, /piəu/, which are phonologically possible combinations of consonants and vowels but are not used as Mandarin syllables). However, for simplicity, our online database includes only those monosyllables that are specifically used to represent the Mandarin lexicon (cf. the Mandarin consonant–vowel combination table, p. 95 in Huang & Liao, 2001). In total, there are 401 legal Mandarin monosyllables, not counting the tonal differences, that our system includes in the representation.

To further test the validity and reliability of our method in representing the Chinese phonology, we trained a self-organizing maps (SOM) model (Kohonen, 1982) on the real-value vector representations of all the 401 Mandarin monosyllables. The core of the SOM model is a two-dimensional square lattice consisting of a set of units, and every unit has the same number of input connections to receive external stimulus patterns (in this case, the numerical vectors representing Chinese monosyllables). A

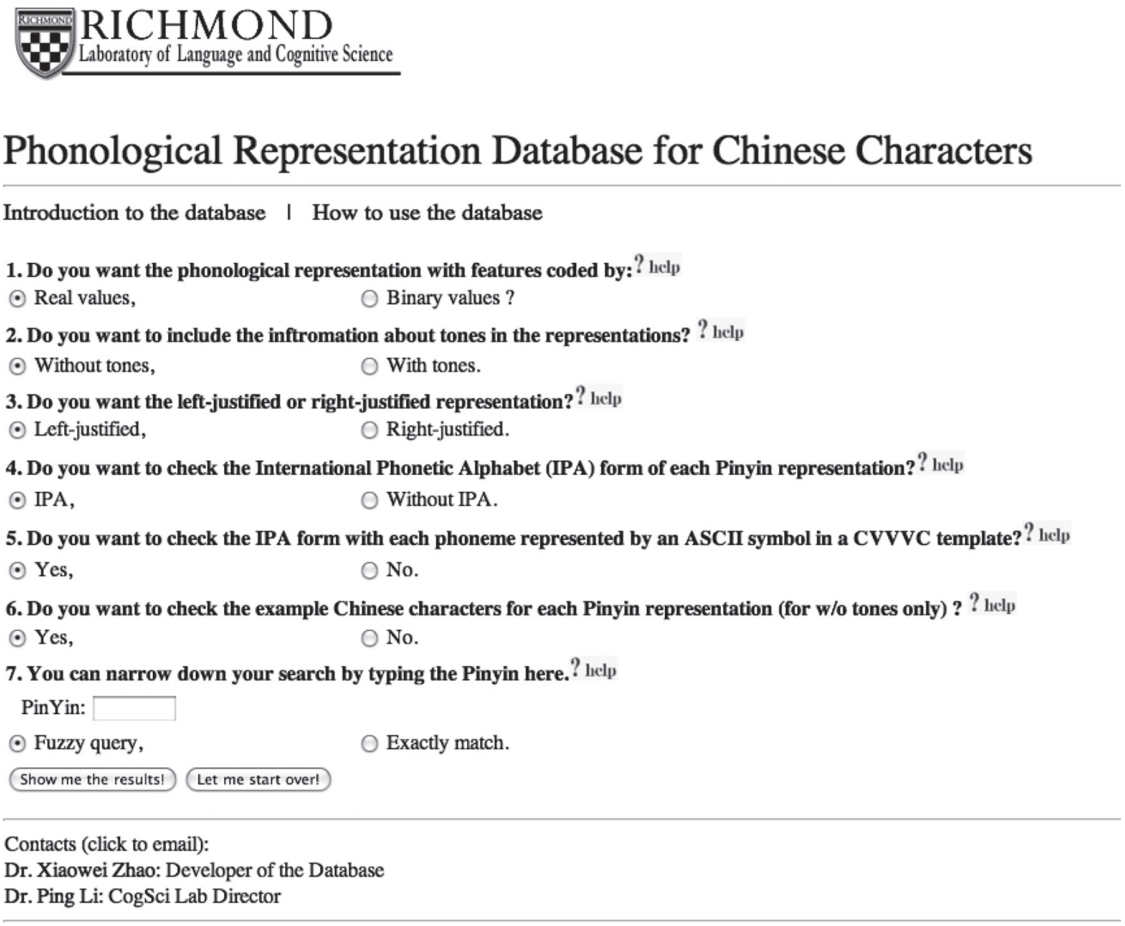

Figure 3. A snapshot of the Web-based interactive query system for our Mandarin phonological representation database.

particularly useful feature of SOM for our purpose is its topography-preserving ability, in that the model will learn to project input patterns that are similar in structure onto neighboring units on a two-dimensional map. The larger the difference between two input patterns, the larger the Euclidean distance will be between the activated units on the map corresponding to the two input patterns.[2]

The size of the SOM used here was 50 × 60, with a total of 3,000 nodes in the network. We trained the network for 200 epochs. In each epoch, every input pattern was presented to the network once. After the training, we calculated and displayed on the map the locations of each monosyllable's best matching unit (BMU; the unit that maximally responds to an input pattern and thus is used to represent the input). As can be seen in Figure 2, a clear phonological structure has emerged in the map as a result of the network's learning of our phonological representations in Chinese, in that similar patterns are grouped together on the map, whereas dissimilar patterns are located further apart (e.g., all the syllables ending with consonant /n/ and /ŋ/, all the syllables ending with diphthongs such as /ai/, /ei/, etc.).

In sum, the SOM model clearly shows that our representational scheme is able to accurately capture the fundamental phonetic structures of Chinese monosyllables. The numerical vectors generated by the system can be applied in a computational model of language learning such as DevLex and DevLex-II and can serve as the phonological input or output of such a model. Elsewhere, we have successfully applied these representations in comparative computational studies of lexical development in Chinese and English with considerable accuracy and efficiency (Zhao & Li, 2007, 2008).

ONLINE INTERACTIVE QUERY SYSTEM

To make our representational system accessible to other investigators, we generated the numerical phonological representations of all Mandarin monosyllables and provided them online in a database. We further developed a Web-based interactive query interface so that users can easily and accurately retrieve phonological representations from our database. The database engine of this Web site is MySQL (www.mysql.com), and the interface works well with all the major Web browsers (e.g., Internet Explorer, Netscape, Mozilla, Firefox, and Safari). Figure 3 presents a snapshot of the interface. The address of the interface on the World-Wide Web is cogsci.richmond.edu/Mandarin_patpho.php.

The interface provides a user-friendly environment. It allows users to enter their query in a flexible way, with various search options available for user-defined specific output of the phonological representations. Users need to first answer seven questions regarding their selection options (see the Appendix) and then proceed to click on the "Show me the results!" button. A new page will pop up with the output numerical representations, along with the other features of the character that the user has selected. The contents of the output page can be saved by the user and be imported into a spreadsheet (e.g., Microsoft Excel) for further analyses.[3] However, for further computational modeling, investigators need to arrange the output numerical representations in a file with appropriate formats that fit the requirements of their modeling software (e.g., Tlearn, PDP++, SOM toolbox).

CONCLUSIONS

In this article, we introduced a phonological representation system and an online database for Mandarin Chinese monosyllables. The basic representation system is based on PatPho, a phonological pattern generator for English, with an adequate consideration of the language-specific features of the Chinese phonology. Our system accurately captures the phonological properties of Chinese characters and allows computational researchers to derive and use phonological representations of Chinese characters in their models. More important, our online database and the accompanying interactive query system can provide a large set of Chinese phonological representations for computational modeling. It also provides a convenient tool for linguists, psychologists, and other cognitive scientists who are interested in crosslinguistic and comparative studies of Chinese and other languages.

We expect some further developments of the online database, including more flexible query options. We also plan to add a new input textbox that will allow users to input Chinese characters to retrieve the corresponding phonological representations of the characters. This feature will be very useful to investigators who do not use the Pinyin system (e.g., scholars in Hong Kong or Taiwan). Such refinements to the database should allow for better and easier uses of the online system by a large number of researchers.

AUTHOR NOTE

Preparation of this article was made possible by a grant from the National Science Foundation (BCS-0642586) to P.L. The writing of the article was completed while P.L. was working for the NSF. The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the NSF. We thank Brian MacWhinney and two anonymous reviewers for their comments and suggestions. Please address correspondence to X. Zhao, Department of Psychology, University of Richmond, Richmond, VA 23173 (e-mail: xzhao2@richmond.edu or xiaoweizhao@gmail.com).

REFERENCES

GUSSMANN, E. (2002). *Phonology analysis and theory*. New York: Cambridge University Press.

HSIAO, J. H., & SHILLCOCK, R. (2004). Connectionist modeling of Chinese character pronunciation based on foveal splitting. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 601-606). Mahwah, NJ: Erlbaum.

HUANG, B., & LIAO, X. (EDS.) (2001). *Modern Chinese* (3rd ed.). Beijing: Higher Education Press.

INSTITUTE OF LINGUISTICS OF THE CHINESE ACADEMY OF SOCIAL SCIENCES (CASS) (2002). *Modern Chinese dictionary*. Beijing: Commercial Press.

JOANISSE, M., & SEIDENBERG, M. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences*, **96**, 7592-7597.

KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59-69.

LADEFOGED, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

LI, P., FARKAS, I., & MACWHINNEY, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, **17**, 1345-1362. doi:10.1016/j.neunet.2004.07.004

LI, P., & MACWHINNEY, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers*, **34**, 408-415.

LI, P., ZHAO, X., & MACWHINNEY, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, **31**, 581-612.

LIN, T., & WANG, L. (1992). *A course in phonetics.* Beijing: Peking University Press.

MACWHINNEY, B., & LEINBACH, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, **40**, 121-157. doi:10.1016/0010-0277(91)90048-9

MIIKKULAINEN, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. Cambridge, MA: MIT Press.

MIIKKULAINEN, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain & Language*, **59**, 334-366. doi:10.1006/brln.1997.1820

PLUNKETT, K., & JUOLA, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, **23**, 463-490. doi:10.1016/S0364-0213(99)00012-9

RUMELHART, D. E., & MCCLELLAND, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216-271). Cambridge, MA: MIT Press.

SEJNOWSKI, T. J., & ROSENBERG, C. R. (1988). NETtalk: A parallel network that learns to read aloud. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: Foundations of research* (pp. 661-672). Cambridge, MA: MIT Press.

WANG, L., LU, J., FU, H., MA, Z., & SU, P. (2004). *Modern Chinese* (2nd ed.). Beijing: Commercial Press.

XING, H., SHU, H., & LI, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science*, **5**, 1-49.

XU, Q. (2008). *Py2ipa*. Retrieved June 14, 2008, from py2ipa .sourceforge.net.

ZHAO, X., & LI, P. (2007). Bilingual lexical representation in a self-organizing neural network. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 755-760). Austin, TX: Cognitive Science Society.

ZHAO, X., & LI, P. (2008). Vocabulary development in English and Chinese: A comparative study with self-organizing neural networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1900-1905). Austin, TX: Cognitive Science Society.

ZHAO, X., & LI, P. (in press). The acquisition of lexical and grammatical aspect in a developmental lexicon model. *Linguistics*.

NOTES

1. Some pronunciations, such as [aʊ-ə] in *hour* and [aɪ-ə] in *fire*, may be thought of as triphthongs, but many linguists treat them as a diphthong plus a vowel in two syllables (Gussmann, 2002, p. 20).

2. For applications of SOM-based models in linguistics and cognitive science, see Miikkulainen (1993, 1997), Li et al. (2004), Li et al. (2007), and Zhao and Li (in press).

3. Investigators may need Unicode fonts to correctly display all the IPA symbols in a spreadsheet program. A good free Unicode font is *Charis SIL Font*, which is downloadable from scripts.sil.org/CharisSILfont. Another Unicode font is *Arial Unicode MS*. It comes with the Mac OS X v10.5 or higher for Mac users; for Windows users, it can be installed from the Microsoft Office Professional CD through the custom install option for international language support.

APPENDIX

There are several flexible options on our Web-based interactive query interface. As is shown in the snapshot of the interface (Figure 3), users can specify their unique searching options for specific phonological representations.

1. *Feature codes.* The representations can be coded by real values between 0.0 and 1.0, or by binary numbers (0 or 1). Users need to select one code. The default is the real value code.

2. *Tones.* Tones represent an important feature of the Chinese phonology but might not be the research goal for some investigators. Users can choose whether they want the output of the phonological representations with tonal information. There are a total of 401 monosyllables without tones in Mandarin Chinese.

3. *Template justification of phonemes*. As in PatPho, in our phonetic template for Chinese, the phonemes can be arranged starting from the leftmost slot (left justified) or the rightmost slot (right justified). For example, the syllable *kan* can be represented in the CVVVC template as [kaVVn] (left justified) or [kVVan] (right justified). The left-justified representations place emphasis on phonological similarities at the beginning of the words, whereas the right-justified representations place emphasis on the coda of the words.

4. *IPA*. The IPA symbols for the phonemes in every syllable can be shown in the output by selecting this option. These symbols, along with the example characters (see below), provide a guide on how to pronounce the specified Mandarin syllable.

5. *ASCII symbols in a template*. Some computers may not correctly display IPA. Therefore, ASCII symbols, instead of IPA symbols, can be used to represent the phonemes. Users can see how the phonemes are arranged in a CVVVC template if this option is checked. Note that the selection of different justifications in Option 3 above will cause different representations of the same word; the empty slots in the template are marked by "-" in the output.

6. *Exemplar characters*. For the convenience of users who are not familiar with Pinyin, we provide an exemplar character for each syllable in our database.

7. *Specified search*. Users can narrow down their search by typing Pinyin symbols (without tone) into an input text field next to "Pinyin" (e.g., *an*, *wang*, etc.). Also, if "with tones" (see Option 2) is selected, a new input field of "Tone" emerges; users can type the tone (0, 1, 2, 3, or 4) of the syllable here. If the "Tone" text box is left empty, all the five tones (if applicable) of the same Pinyin will be shown in the output. Finally, users can choose a fuzzy query or an exact-match query. The fuzzy query is extremely useful if the user wants to find all the possible syllables that include a certain combination of phonemes (e.g., all the syllables including *ang* [/ɑŋ/]). If the user selects this option while keeping the "Pinyin" field empty, all the possible syllables in the entire database will be displayed in the output table. The exact-match query is useful if the user wants to find the record with a syllable exactly as typed in the input field.

(Manuscript received November 24, 2008;
revision accepted for publication February 19, 2009.)