A Cross-Linguistic Study of Phonological Units: Syllables Emerge from the Statistics of Mandarin Chinese, but not from the Statistics of English

Train-Min, Chen (trainmin@alumni.ccu.edu.tw)
Department of Psychology, National Chung Cheng University
160 San-Hsing, Min-Hsiung, Chia-Yi 621 Taiwan, R.O.C.

Gary S. Dell (gdell@s.psych.uiuc.edu)
Beckman Institute, University of Illinois at Urbana-Champaign
405 N. Mathews Ave, Urbana, Illinois 61801 U.S.A.

Jenn-Yeu, Chen (psyjyc@ccu.edu.tw)
Department of Psychology, National Chung Cheng University
160 San-Hsing, Min-Hsiung, Chia-Yi 621 Taiwan, R.O.C.

Abstract

This study explored the statistical patterns of English and Mandarin Chinese sound sequences, by comparing their learning in a simple recurrent network. Experiment 1 showed that vivid syllable structure emerged from the sound sequence of Mandarin Chinese. Experiment 2 further demonstrated that the emerged syllable structure of Mandarin Chinese is considerably more salient than that of English. We claim that the more salient syllable structure in Mandarin Chinese inputs is one reason why syllable units are particularly emphasized in its processing in comparison to English.

Introduction

According to linguistic theory, the sound patterns of all languages are hierarchical. Segmental speech sounds (or phonemes) are concatenated into syllabic constituents (onset, rhyme), which join to form syllables, which, in turn, are the constituents of larger units such as feet and words. In psycholinguistic theories of production, each kind of unit plays a part, but some units are more salient than others. For example, standard theory (e.g. Levelt, Roelofs, & Meyer, 1999) holds that lexical items are stored as sequences of segments. The organization of these sounds into syllables, however, is not stored, but rather is computed during production. Evidence against stored syllables comes from two principal sources:

(1) *The absence of syllabic speech errors.* For example, exchanges of non-morphemic syllables are very rare. You would never hear "napkin" spoken as "kinnap."

(2) *The absence of syllabic priming effects that cannot be attributed to segmental units*. For example, naming a word is speeded when a masked orthographic prime syllable that matches the initial sounds of the word precedes it. In many such studies, this priming is unaffected by whether or not the prime syllable corresponds to a whole syllable in the target word (e.g. Schiller, 2000).

The conclusion that syllables are not units of storage has been based on studies of Germanic languages such as English and Dutch. Our own production research has demonstrated that this conclusion is not warranted for all languages. Here, we briefly review our studies of production in Mandarin Chinese, which show that the syllable is far more unitary than has been found in English and Dutch. Then, we present two computational studies involving the learning of Mandarin and English sound sequences. These studies suggest that cross-linguistic differences in the salience of the syllable in production emerge from the statistics of the sequences.

Speech Error Data Psycholinguists believe that the commonness of slips of units such as segments or words provides evidence that these units are psychologically real. Research has shown that whole syllables, in contrast, rarely move, at least in English and in related languages, and the few apparent cases can be otherwise explained as morphemic or segmental slips (Chen, 2000). However, this is not the case for Mandarin Chinese. Chen (2000) demonstrated that syllable movement errors indeed happen in the natural speech of Mandarin Chinese at a rate of 10,000 times greater than would be expected if these errors were the result of independent segmental slips. Importantly, these syllable errors "strand" tone; only the segmental part of the syllable moves. One such example is that the word [ching1zhuo2du4] (清濁度 'clarity') slipped to [ching1du2du4], an anticipation of the third syllable [du]. This stranding of tone rules out the explanation that these are slips of morphemes or characters.

Masked Priming Data Studies of masked priming of word naming in Mandarin Chinese also point to the syllable as a unit. Chen, Lin, and Ferrand (2003) found that when the segmental overlap between the target word to be produced and a preceding character prime constituted a complete syllable, the response time was faster than when it did not. This result was obtained for both CVC and CV-glide syllables (Lin & Chen, 2003).

Implicit Priming Data The implicit priming paradigm is a production task that is useful for discovering relevant phonological units. Participants learn several cue-target word pairs, and later must say the target member of the pair as quickly and correctly as possible upon seeing its paired cue word. Using this task in Dutch, Meyer (1991) showed that when target words in a set shared their initial portions,

responses were faster than when they did not. Moreover, when the shared initial portions did not constitute a syllable, priming was still observed and correlated positively with the number of shared segments. Hence, the observed priming effects seem to derive from shared segments. Implicit priming in Mandarin Chinese was quite different from that in Dutch (Chen, Chen, & Dell, 2002). Crucially, priming was only found when target words shared the segments of entire first syllable, or the segments and the tone together. Accordingly, the priming effects in Mandarin Chinese derive from syllable rather than segment sized units.

Why are Mandarin Chinese speakers more sensitive to syllable units than other speakers?

We believe that the production system reflects the language. Relative to English, Mandarin Chinese has few syllable types. English has more than 10,000 syllables, and Mandarin only around 400 (without counting the tone) or 1,200 (when tones are considered). In addition, re-syllabification commonly occurs in English speech, e.g. 'demand it' becomes 'de-man-dit', but not in Mandarin (Kuo, 1994). Moreover, English, but not Mandarin, involves ambisyllabicity, the apparent membership of a consonant in more than one syllable e.g. 'hap-py'. These properties make the syllable a more efficient planning unit for Mandarin Chinese. In this paper, we explore the hypothesis that these properties lead to sound sequences whose statistical structure favors syllabic, as opposed to word and segmental, units. More generally, we claim that the relative importance of various unit types is a product of experience and test this claim by adopting the computational approach of Elman (1990) to phonological sequences.

Analysis of Sequences by Simple Recurrent Network

The Simple Recurrent Network (Elman, 1990)

A simple recurrent network (see Figure 1) is a three-layer feedforward network, in which input, hidden and output layers are linked by forward trainable connections in a distributed fashion, i.e. fully connected. Crucially, the network includes another input layer, the context layer, which serves as the dynamic memory of the network. The connections from context layer to hidden layer are trainable and distributed, but the recurrent connections from hidden layer to context have fixed weights of 1.0 and are one-to-one. Functionally, the recurrent connections behave much like a copier, which duplicate the activation pattern of hidden layer at a particular time step on the context layer. Hence, output at any given time step is shaped by the network's previous internal state together with its current input. These properties make simple recurrent networks useful models of how people implicitly learn the structure of sequences.

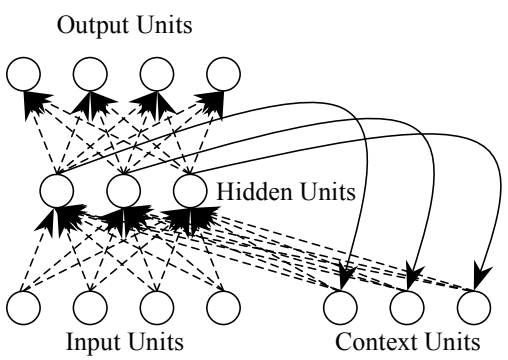

Figure 1. Elman's (1990) simple recurrent network, in which activations are copied from hidden layer to context layer on a one-for-one basis, with fixed weight of 1.0. Dotted lines represent trainable connections.

Word Structure in a Letter Sequence

Elman (1990) examined the statistical structure of English letter sequences by having the simple recurrent network predict the letter that follows the current input letter. Trainable weights were changed to the extent that the prediction was incorrect. The degree of prediction error was highly correlated with word boundaries. Error tended to spike up for word-initial letters, and declined as a function of the serial position for letters within words (see Figure 2). Hence, the relatively higher error for word-initial letters successfully demonstrated that the simple recurrent network discerns the word structure in the letter sequence without providing it with any word boundary cue during training. It appears that the word is the dominant unit, at least in English letter sequences. Next, we perform the same analysis on spoken Mandarin Chinese input.

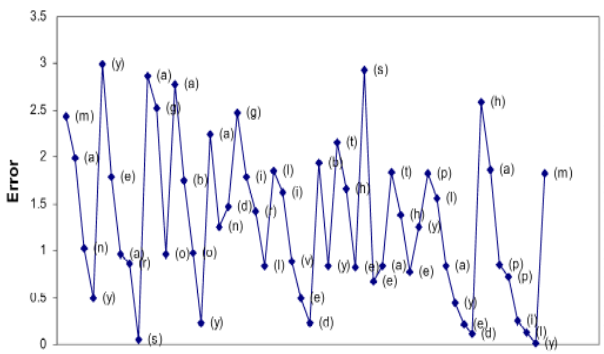

Figure 2. Graph of root mean squared error plotted over time in the letter prediction task (Elman, 1990, Figure 6). The letter to be predicted each time is shown in parenthesis.

Experiment 1: Exploring the Syllable Structure in Mandarin Chinese Inputs

In this experiment, Mandarin Chinese input was assessed by inspecting the relative performance of the network on predicting (1) word-initial sounds, (2) syllable-initial sounds that are not also word initial, and (3) the sounds within the syllable (hereafter, within-syllable sounds). If predicting syllable-initial sounds is harder than predicting within-syllable sounds, syllable boundaries will be protruded, that

is, syllable units will show up. Besides, if predicting the syllable-initial sound is as difficult as predicting the word-initial sound, it suggests that the syllable is the sole emergent unit. This is because word-initial sounds themselves are also syllable-initial, and the syllable unit alone could explain the pattern without postulating a word level.

Method

Simulation Materials and Sound Representation The simulation material came from a 30-minute stretch of a children's radio broadcast program (for ages 6 and upward) downloaded from the "National Education Radio" website at **http://www.ner.gov.tw.** It contained 5,394 sounds (sounds differing only in nasal features were regarded as different sounds), comprising 2,072 syllables and 1,300 words. For the simulation, each sound was represented as a 52-bit binary vector, 47-bit for the segment (because 47 different segments were involved) and 5-bit for the tone (because there are five lexical tones in Mandarin Chinese).

Simulation Design and Network Training The performance of the network on predicting the word-initial, syllable-initial, and within-syllable sounds was examined under 18 conditions created by crossing three factors: (1) tonal information (Syllable condition: tone distributed to each sound of the syllable; Rhyme condition: tone only distributed to sounds of the rhyme; Without condition: no tonal information), (2) the number of training epochs (20, 40), (3) the number of hidden units (25, 100, 200). Performance of the network was evaluated by two kinds of scores: (1) the Error Rate, calculated by regarding the output vector incorrect if the proposed target vector was not its closest vector, and (2) the Euclidian Distance between the actual output vector and its target vector. For both ways of scoring, the larger scores index greater unpredictability and, hence, a more salient boundary.

Before training, the connection weights were initialized randomly in the range of ±0.5. Training began with presenting the network a sequence of input vectors one at a time, and having the network learn to predict the next by adjusting the connection weights with the backpropagation algorithm. Leaning rate and momentum were set to 0.3 and 0.9 respectively.

Results and Discussion

Throughout the study, the Error Rate and Euclidian Distance displayed the same pattern. We present only the latter measure in Figure 3, which illustrates the results of single, but typical, condition. Sounds at word and syllable boundaries were much more difficult to predict than within-syllable sounds. The difficulty of predicting the syllable-initial sounds, however, was quite similar to that of the word-initial sounds. This pattern held no matter how many hidden units the network was equipped with, how many epochs of training had passed, and how (or even whether) the tone was represented. These findings strongly suggest that the syllable is functioning as a unit. Predictability is very high within the syllable, and the word boundary does little to increase uncertainty beyond that associated with the syllable boundary. Whether *any* word structure exists will be statistically examined in Experiment 2 when we directly compare Mandarin Chinese with English.

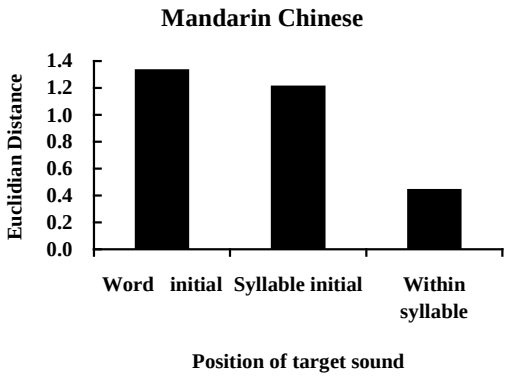

Figure 3. Average Euclidian Distance for predicting sounds in different positions. (Simulation condition: 100 hidden units, 20 epochs, tone was distributed to each sound of the syllable)

An important side result of this experiment is that the pattern shown in Figure 3 was, for the most part, independent of how we represented tone, or even whether we represented it at all. For example, if tone was not represented, the condition corresponding to the one illustrated in Figure 3 led to distances of 1.00, 0.94, and 0.56 for word-initial, syllable-initial, and within-syllable sounds, respectively. Thus, the segmental pattern alone is more than enough to motivate dominant syllable-sized units. In fact, the speech error study of Chen (2000), the implicit priming study of Chen, Chen, and Dell (2002), and the masked priming study of Chen, Lin and Ferrand (2003) all suggested that tone-less or segmental syllables as well as syllables with tone function as important production units in Mandarin. The findings of Experiment 1 are quite consistent with these data.

Experiment 2: Comparing Sound Patterns in English and Mandarin Chinese

The second experiment compared the sound distributions in English and Mandarin Chinese inputs directly. This was achieved by replicating the prior experiment using English and Mandarin Chinese versions of comparable simulation materials. The experiment also manipulated the nature of the representation of diphthongs, that is, whether they are considered to be one or two sounds. Because the prior experiment had shown that the supra-segmental (tone) information, the number of the hidden units, and the number of training epochs did not matter, the present experiment was conducted without supra-segmental information (tone or stress), and with a constant 100 hidden-unit network trained for 20 epochs.

Method

Simulation Materials The simulation materials consisted of 10 short English-Mandarin Chinese bilingual children's stories, downloaded from the "Mandarin Daily News" website at **http://www.mdnkids.org.tw/**. The English versions contained a total of 6,511 sounds when diphthongs were counted as two sounds and 6,243 sounds when diphthongs were counted as one sound, 2,482 syllables and 1,949 words. The Mandarin Chinese versions contained 7,116 sounds when diphthongs were counted as two sounds and 6,472 sounds when diphthongs were counted as one sound, 2,743 syllables and 1,860 words. The principle of representing the sounds for simulation was identical to the prior experiment.

Simulation Design Four conditions were created by crossing two factors: (1) the number of sounds that diphthongs denote (one, two), and (2) languages (English, Mandarin Chinese). Aside from the Error Rate and Euclidian Distance, another score, Syllabic Saliency, was created for representing the degree of saliency of syllable structure. It was defined as where, in percentage terms, the performance of predicting the syllable-initial sound locates on a scale that is maximal at the performance of predicting the word-initial sound and minimal at the performance of predicting the within-syllable sound.

Results

The results with Euclidian Distance and Error Rate were statistically indistinguishable and so we continue to report only the distance measure. The main finding from the experiment was that, as expected, Mandarin Chinese differed considerable from English. Treating the 10 stories as "subjects" in an analysis of variance with language, type of boundary, and diphthong representation as independent variables yielded a strong interaction between language and boundary type, $F(1,18) = 197.2$, $p < .0001$. Figure 4 shows the findings from the diphthong-as-two-sounds condition. In Mandarin the predictability at syllable and word boundaries was nearly identical. The Syllable Saliency here was 96%, that is, predicting sounds at syllable boundaries was almost as inaccurate as predicting them at word boundaries. For English, the Syllable Salience in this condition was 35%, much lower than in Mandarin, $F(1,9) = 25.9$, $p < .001$. In fact, for English, syllable-initial prediction accuracy was actually closer to within-syllable than word-initial accuracy. The results were similar, but less dramatic, when the diphthong was treated as a single sound. The Syllable Saliencies for Mandarin and English were 88% and 43%, respectively, $F(1,9) = 19.4$, $p < .002$. Pooling across the diphthong treatment yielded a strong effect of language on this measure, $F(1,9) = 24.6$, $p < .001$.

Clearly, in Mandarin, the predictability was close for word-initial and syllable-initial sounds. Were word-initial sounds any less predictable than syllable-initial ones? For English, they definitely were, $F(1,9) = 28.7$, $p < .001$. For Mandarin, the very small effect was not significant in the diphthong-as-two-sounds condition, but it was in the diphthong -as-one-sound condition, $F(1,9) = 4.4$, $p < .03$.

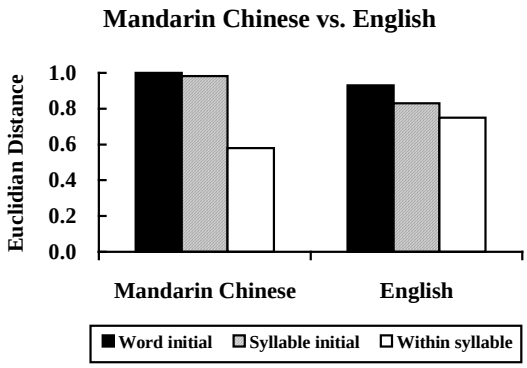

Figure 4. Average Euclidian Distance for predicting sounds in different positions in Mandarin Chinese and English (Simulation condition: diphthong as 2 sounds, 100 hidden units, 20 training epochs, without tonal information)

Discussion

To summarize, three major findings were obtained. First, vivid syllable structure emerged from the sound sequence of Mandarin Chinese. Second, the emerged syllable structure is more salient in Mandarin Chinese than in English. Third, equivalent syllable structure was found even when supra-segmental information was removed from the sound sequence. Implications of these results are discussed below.

As described in the introduction, psycholinguistic studies demonstrated that the role of the syllable is not equally emphasized in the production of English and Mandarin Chinese, a finding that hints that the sound patterns the language presents should reflect such difference. This is exactly what we demonstrated in this experiment. In Mandarin, the predictability of a sound was almost entirely determined by whether or not it is at a syllable boundary. In English, word structure was more salient, and the predictability within a syllable was not that much greater than that at syllable boundaries that are not word boundaries.

A stronger, but more speculative, interpretation of our findings makes reference to the particular kind of model that we used to assess predictability, the simple recurrent network. This network architecture has been offered as an account of phonological retrieval in production (e.g. Dell, Juliano, & Govindjee, 1993). One of the advantages of such an account is that one does not need to explicitly include or exclude particular kinds of units. Rather, the weights acquired through learning lead to activation states with greater or lesser correspondence to discrete units at several levels. Hence, the learner is not faced with the all-or-none decision as to whether to have a syllable level in the system. To the extent that different languages possess gradations in the salience of units such as the syllable, this connectionist approach may help explain the cross-linguistic variation

Another finding of note was that a strong syllable structure emerged in Mandarin Chinese even when supra-segmental (tone) information was not considered. This suggests that the segmental syllable, i.e. the syllable without the tone, has statistical support in the input, and may function as a processing unit. Psycholinguistic studies of Mandarin support this hypothesis. For instance, analysis of natural speech errors indicated that sometimes a syllable moves to a new location, leaving its tone behind (Chen, 2000). That is, the slipping unit was a segmental syllable. Pan, Chen, and Chen (1999) demonstrated this effect with experimentally generated slips. Furthermore, implicit priming and masked priming findings are robust both for syllables with tones and for segmental syllables (Chen, Chen & Dell, 2002; Chen, Lin & Ferrand, 2003).

References

Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43(1),* 15-26.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by implicit priming task. *Journal of Memory and Languages, 46,* 751-781.

Chen, J.-Y., Lin, W.-C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese speech production. *Chinese Journal of Psychology, 45(1),* 107-120.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and Content in Language Production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17,* 149-195.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14,* 179-212.

Kuo, F.-L. (1994). *Aspects of segmental phonology and Chinese syllable structure.* Ph.D. thesis, University of Illinois at Urbana-Champaign, Illinois.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-75.

Lin, W.-C., & Chen, J.-Y. (2003). Masked priming of the segmental syllable in Mandarin Chinese speech production: More evidence. Presented orally at the 42nd Annual Meeting of the Chinese Psychological Association, October 4-5, Taipei, Taiwan.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30,* 69-89.

Pan, H.-C., Chen, T.-M., & Chen, J.-Y. (2000). GuoYu YinJie de DengLu LiCheng (The role of syllable in phonological encoding in Mandarin Chinese). Presented orally at the First Taiwan Cognitive Science Conference: A Review and a Prospect of Cognitive Psychology in Taiwan, June 23-25, Chiayi, Taiwan.

Schiller, N. O. (2000). Single-word production in English: The role of subsyllabic units during phonological encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 512-528.