# Sublexical Units in the Processing of Chinese Characters

Cheng-hua Bai    Robert Schreuder
Donders Centre for Cognition, Radboud University Nijmegen, The Netherlands
c.bai@donders.ru.nl; r.schreuder@donders.ru.nl

**Abstract:** Two studies were conducted to address the question of what kinds of lexical information readers use for visual Chinese character recognition. The studies were investigating a specific type of Chinese character which is called “phonogram” (xing2 sheng1 zi4). This type of Chinese character which is composed by two units: one unit cues the meaning and another unit cues the pronunciation of that character. The aim of the present studies was to assess the relative contribution of information in Chinese characters which readers use during visual character recognition. The first study was an offline subjective familiarity rating test. The result suggested that subjective familiarity and lexical frequency are highly correlated. In addition, the more strokes a character has, the less familiar it is to the readers. The second study used a lexical decision task. The study employed a headstart paradigm, in which part of the target information was pre-exposed for a short period on the display. This study suggests that the pre-exposure of radicals prepares the process of recognizing the whole character. In addition, we discuss the qualitative and quantitative differences among semantic radicals and phonetic radicals. Overall, the preliminary results suggest that the lexical frequency, the number of strokes, and the type of radicals contribute functional information during the processes of Chinese characters.

**Keywords:** Phonograms, Radicals, Familiarity, Lexical Decision, Headstart

## Experiment 1: A familiarity rating study

Previous research has shown correlations between subjective frequency ratings and visual lexical decision reaction times. A high-frequent monomorphemic noun elicits higher subjective rating and shorter response latencies (Schreuder and Baayen, 1997). The present study examines whether the frequency of Chinese characters also correlates with higher subjective rating; additionally, whether the sublexical unit (i.e. radical) in Chinese phonograms serves meaning submorphemic functions to the processing of the whole character.

Taft, Zhu, and Peng (1999) argue that there is both a radical and a character level of representation in the recognition of Chinese characters. In addition, preceding the radical level there is a direct activation by featural information, such as positional features of the radicals.

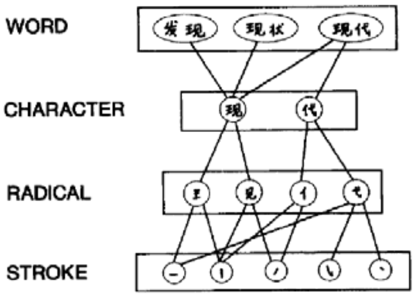

Figure 1. The multilevel activation framework of Taft and Zhu (1997)

In the first study, we examined the correlations between the familiarity ratings of Chinese written characters and the corpus frequency counts and the number of strokes of characters.

## Method

Materials

The materials for the rating study consisted of 120 Chinese characters and 55 fillers. Among all the arrangements, the left-right structure is the most frequent one which occurs in 72% of the phonograms. In addition, within the left-right structure, 90% of them have the semantic radical on the left and the phonetic radical on the right (Hsiao, Shillcock, & Lee, 2007). Thus, we group characters into a SP and a non-SP group for reflecting the statistical properties of written Chinese. Among the 120 characters, half of them were phonograms with a SP structure. Another 60 of them contain 20 characters with a PS structure, 20 $\genfrac{}{}{0pt}{}{s}{p}$ (semantic radical on the top and phonetic radical on the bottom) structure and 20 $\genfrac{}{}{0pt}{}{p}{s}$ structure respectively.

Our materials are common characters which are rather familiar to the readers. In order to allow the participants to make use of the full scale from 1 to 7 in the rating, an additional 55 characters with lower frequency, of which their token frequency is no more than 6 per million, were selected as fillers.

All the characters are nouns or are more frequently interpreted as a noun when presenting in isolation. In addition, the characters which were selected for this study are visually simple. They were mostly composed by only two units and there was no visual overlap between radicals. Characters which cannot stand alone as a meaningful noun were not selected either.

Participants

Eighteen native Mandarin Chinese readers in Taiwan.

Procedure

The material list in a form of questionnaire was rated by readers online. The readers were asked to rate their familiarity with the written characters on a scale of 1 to 7. Score one means rarely seen in the written language while seven indicates that a rated character is very often encountered. For the actual design of the questionnaire, see Appendix A.

**Results and Discussion**

The mean and the standard deviation for each character were calculated. Information of frequency and number of strokes were obtained from a Chinese character list for each of the 13,060 Chinese characters reported by Huang (1995) and Tsai (1996). The frequency count was based on a corpus of 171,882,493 tokens which consists of all the BIG-5 Chinese characters appeared on Usenet newsgroups during 1993-1994.

The results of familiarity rating is correlated highly with the log of frequency (per million) ($r$ = .88, $p$ < .001). The number of stokes also correlates with the frequency count significantly ($r$ = -.25, $p$ < .001). The familiarity rating and frequency were positively correlated while it was negatively correlated with the number of strokes. This result shows that a character with higher frequency has a higher score from the familiarity rating and less number of strokes.

**Experiment 2: Lexical decision task and the headstart effect**

The second study investigated whether the semantic and phonetic radical in one phonogram play different functional roles in processing the character. The nature of phonograms is theoretically interesting for the architecture of lexical representation and processes. We investigate two models of visual word recognition and test how well these two models reflect the hypothesized contribution from semantic radical and phonetic radical. In Figure 1, we discussed a multilevel model proposed by Taft and Zhu. We incorporate a part from another visual word recognition model (Grainger, & Holcomb, 2009) and emphasize the semantic processing of a single character. We assume that there is a direct link from the semantic radical to the semantic reading of the character, which is called an S-unit. Thus, the semantic radical would be more helping in retrieving the lexical semantics of the given character than the phonetic radical.

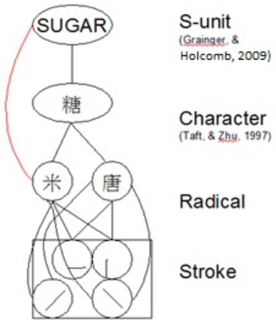

Figure 2. A hypothetical hierarchical model of Chinese visual character recognition

In experiment 2, participants were presented with part of the whole character as a short (100 ms) pre-exposure (Eriksen, & Eriksen, 1974). In some trials, participants saw either a semantic radical or a phonetic radical appear before the whole character. In the other trials, there is pre-exposition of a nonfunctional symbol (#) which served as the baseline. Pre-exposition of the semantic or phonetic unit of a character may differ in their effect on the processes of character recognition and are studied by measuring the reaction time.

**Method**

Participants

Twenty native speakers of Taiwan Mandarin currently residing in the Netherlands. They all learned Zhu-yin in school (Zhu-yin are non-alphabetic phonetic symbols speakers learned in primary school for annotating the phonemes in Taiwan). Meanwhile, traditional Chinese character is the first written language they used for reading since childhood.

Materials

The materials consisted of 64 existing Chinese characters and 64 pseudocharacters. The characters were selected from the rating study reported above. Thus, all the selection criteria were identical to the one for the rating study. In addition, in order to minimize repetition effect, we decided that each radical could only appear at most twice in the materials.

Existing characters

There are 64 existing characters in the material list. These characters are divided into two groups according to the positional information of their radicals. Thirty-two of the existing characters have the semantic radical on the left and the phonetic radical on the right, which is a SP character. Another 32 characters have a PS structure.

Pseudocharacters

All pseudocharacters were made up by combining an existing semantic radical and one phonetic radical. We used the same characters from the existing list to make up the pseudocharacters for the following reason. By using the existing characters from the materials list to make up pseudocharacters, we can use the identical semantic radicals and phonetic radicals in existing and pseudocharacters. Thus, we have controlled the repetition of semantic radicals and phonetic radicals. In addition, the amount of SP and PS pseudocharacters is identical to the SP and PS structure for existing characters. We controlled the visual complexity of radicals in the pseudocharacters since they come from the existing character list.

Procedure

All the participants were tested individually. They sat in front of a computer screen where the target characters were presented. The size of each character is 2 cm height and 2 cm width and in KaiU font. The participants were seated in front of the screen at a distance of approximately 50 cm. Each trial was preceded by a fixation cross that appeared 300 ms before the stimulus followed by a 200 ms blank on the display.

There were three conditions, baseline headstart, semantic radical as headstart or phonetic radical as headstart. The headstart was presented 100 ms before the target. The participants were asked to decide whether the characters exist in written language as quickly and as accurately as possible. The trial ended when a participant responded. And after 900 ms, the next trial started.

## Results and Discussion

We measured reaction time from the onset of target characters and error rate of the lexical decisions to investigate how functional the radicals are to the processing of the whole character. For the analysis of the existing characters, all trials which the participants made errors were excluded. A main headstart effect was found.

The main headstart effect suggests that the pre-exposure of radicals facilitates the lexical decision task. Pre-exposure of a radical (M = 576, SD = 186) facilitates the reaction to the lexical decision task compared with a non-functional baseline headstart (M = 589, SD = 168), (*t*(2408) = 1.70, *p* < .05 (one-tailed)). In order to examine further, we split up the data for SP and PS structures. That is because SP character are the most dominate type of phonograms in written Chinese language. We carried out a linear mixed-effect regression model, with participants and characters are the random factors, on the log RT's of the correct responses for existing SP characters, as predictor type of headstarts and the score of familiarity rating. The effect of rating is highly significant: *ß* = -0.04, *t*(1215) = -4.1, *p* < .0005 (one-tailed). The effect of type of headstarts to the SP characters is approaching significant for semantic radicals: *ß* = -.025, *t*(1215) = -1.60, *p* < .055 (one-tailed). For phonetic radicals, the headstart effect is non-significant: *ß* = -.018, *t*(1215) = -1.14, *p* < .10. As for the PS structure, the effect of type of headstarts to is highly significant for phonetic radicals: *ß* = -.061, *t*(1215) = -4.06, *p* < .001 (one-tailed). For semantic radicals, the headstart effect is non-significant: *ß* = -0.016, *t*(1215) = -1.09, *p* < .14 (one-tailed).

To test if there is an interaction between character types (SP vs. PS) and the types of headstarts (semantic vs. phonetic), we carried a linear mixed-effect regression model on the combined dataset of SP and PS structures. We found that there is a marginally significant interaction between character types and the types of headstarts (*F*(2, 2456)) = 2.6, *p* < .07. So far, the result suggests that a semantic headstart is more helpful for processing a SP character while a phonetic headstart is more useful for processing a PS character. We will test further whether the left/ right position of a certain character is playing the functional role in the recognition of Chinese character instead of the type of sublexical information. We speculate that the left position is a functional node on the radical level which we can implement in the hypothesized model (Figure 2).

## References

[1] Eriksen, B. A., & Eriksen, C. W. (1974). The importance of being first "A tachistoscopic study of the contribution of each letter to the recognition of four-letter words. *Public Health*, 66-72.

[2] Grainger, J., & Holcomb, P. J. (2009). Watching the Word Go by: On the Time-course of Component Processes in Visual Word Recognition. *Language and linguistics compass*, *3*(1), 128-156. doi: 10.1111/j.1749-818X.2008.00121.x.

[3] Hsiao, J. H. -W., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: implications for orthographic processing. *Journal of psycholinguistic research*, *35*(5), 405-26. doi: 10.1007/s10936-006-9022-y.

[4] Hsiao, J. H.-W, Shillcock, R., & Lee, C.-Y. (2007). Neural correlates of foveal splitting in reading : Evidence from an ERP study of Chinese character recognition. *Neuropsychologia*, *45*, 1280-1292. doi: 10.1016/j.neuropsychologia.2006.10.001.

[5] Huang, S.K.(1995). Frequency counts of BIG-5 Chinese characters appeared on Usenet newsgroups during 1993– 1994. Retrieved November 17, 2010 from http://technology.chtsai.org/charfreq/.

[6] Lee, C.-Y. (2008). Rethinking of the Regularity and Consistency Effects in Reading, *Language and Linguistics*, *9*(1), 177-186.

[7] Taft, M., & Zhu, X. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 761-775. doi: 10.1037/0278-7393.23.3.761.

[8] Taft, M., Zhu, X., & Peng, D. (1999). Positional Specificity of Radicals in Chinese Character Recognition. *Journal of Memory and Language*, *40*(4), 498-519. doi: 10.1006/jmla.1998.2625.

[9] Tsai, C.-H. (1996). Frequency and Stroke Counts of Chinese Characters. Retrieved November 17, 2010 from http://technology.chtsai.org/charfreq/.

## Appendix A. Questionnaire for experiment 1

(A1. in Chinese; A2. in English)

### Appendix A1. 中文字在書面用語裡的使用

親愛的受訪者，您好:

感謝您撥空參加我的論文研究。在以下的問卷中，您將會看到 180 個正體中文字。請依據您認為一般台灣讀者在日常文書(如報章雜誌, 信件, 書籍, 網路文章等等)中有多常看到下列文字作答。
每一題都有一個 1 到 7 的量尺，請盡量使用整個量尺。量尺表示這個字常見的程度: 1 是指這個字在書面上十分罕見，7 是指這個字在書面上十分常見。

例如:
"時"在書面上十分常見，您應該給這個字 6 或 7。
"箭"在書面上不算常見不算罕見，您應該給這個字 4。
"鬃"在書面上十分罕見，您應該給這個字 1 或 2。

請依照一般台灣讀者在日常文書中看到這些字的原則作答。有些字很常在口語中使用，但較少在書面中出現，像是"酷"。在這種情況下，請給"酷"這給字較低的分數。

### Appendix A2. Chinese Characters in Written Forms

Dear Participants,

Thanks for participating in this study and helping me with my master thesis. In the following, you will see 180 Chinese characters. Please indicate how often do you think that general Taiwanese readers see each Chinese character in daily written language, such as in newspapers, letters, and books.

There is a seven point scale for you to use. Please try to use the whole scale, including the very rare (1) and very often (7) ones. One means you rarely see this character in written form while seven means that you often see this character in written form.

For instance,
時 appears so often in daily written language that you should give it a 6 or 7.
箭 appears in daily written language that you should give it a 3 or 4.
騌 appears rarely in daily written language that that you should give it a 1 or 2.

Please make the decision according to the written language you think the general public sees in daily life. If people say a word often like '酷,' but it occurs less in written language. This character should receive a low score (1 or 2).