Article

The use of probabilistic lexicality cues for word segmentation in Chinese reading

Zang, Chuanli, Wang, Yongsheng, Bai, Xuejun, Yan, Guoli, Drieghe, Denis and Liversedge, Simon Paul

Zang, Chuanli, Wang, Yongsheng, Bai, Xuejun, Yan, Guoli, Drieghe, Denis and Liversedge, Simon Paul ORCID: 0000-0002-8579-8546 (2016) The use of probabilistic lexicality cues for word segmentation in Chinese reading. The Quarterly Journal of Experimental Psychology Section B, 69 (3). pp. 548-560. ISSN 1747-0218

The Quarterly Journal of Experimental Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

The Use of Probabilistic Lexicality Cues for Word Segmentation in Chinese Reading

Chuanli Zang[a], Yongsheng Wang[a], Xuejun Bai[a], Guoli Yan[a], Denis Drieghe[b] & Simon P. Liversedge[b]

[a] Tianjin Normal University

[b] University of Southampton

Accepted author version posted online: 06 Jul 2015.

Publisher: Taylor & Francis & The Experimental Psychology Society
Journal: *The Quarterly Journal of Experimental Psychology*
DOI: 10.1080/17470218.2015.1061030

The Use of Probabilistic Lexicality Cues for Word Segmentation in Chinese Reading

Chuanli Zang[1], Yongsheng Wang[1], Xuejun Bai[1], Guoli Yan[1], Denis Drieghe[2] & Simon P. Liversedge[2]

[1] Tianjin Normal University

[2] University of Southampton

Send Correspondence to:

Simon P. Liversedge,

Centre for Visual Cognition,

School of Psychology, Shackleton Building,

University of Southampton,

Highfield, Southampton, SO17 1BJ, UK

Email: s.p.liversedge@soton.ac.uk

Phone: +44 23 8059 9399 Fax: +44 23 8059 4597

Running Head: Word Segmentation in Chinese Reading

Word Count: 7659 words

Abstract

In an eye tracking experiment we examined whether Chinese readers were sensitive to information concerning how often a Chinese character appears as a single character word versus the first character in a two character word, and whether readers use this information to segment words and adjust the amount of parafoveal processing of subsequent characters during reading. Participants read sentences containing a two-character target word with its first character more or less likely to be a single character word. The boundary paradigm (Rayner, 1975) was used. The boundary appeared between the first character and the second character of the target word and we manipulated whether readers saw an identity or a pseudocharacter preview of the second character of the target. Linear mixed-effects models revealed reduced preview benefit from the second character when the first character was more likely to be a single character word. This suggests that Chinese readers use probabilistic combinatorial information about the likelihood of a Chinese character being single-character word or a two-character word online to modulate the extent of parafoveal processing.

Keywords: Word segmentation, preview benefit, eye movements, Chinese reading.

It has been documented that words are the basic meaningful unit of spaced, alphabetic languages like English, and properties of words (such as word frequency and word length) influence *when* readers' saccades are initiated and *where* their eye movements are targeted during reading (Liversedge & Findlay, 2000; Rayner, 1998,

2009). This is a principal assumption of the most influential models of eye movement control such as the E-Z Reader model and the SWIFT model. The E-Z Reader model (e.g., Reichle, Rayner & Pollatsek, 2003; see Reichle, 2011 for a review) posits that attention is shifted serially and sequentially to only one word at a time, with the words in a sentence being lexically processed sequentially and serially. By contrast, the SWIFT model (e.g., Engbert, Nuthman, Reichter & Kliegl, 2005; see Engbert & Kliegl, 2011 for a review) assumes that two or more words in the perceptual span can be lexically processed (and potentially identified) in parallel. Both models differ in this assumption, however, they both assume that the lexical processing of a word, based on foveal and parafoveal processing, influences the decision to move the eyes forward in the text. The importance of parafoveal processing is evident from the finding that readers spend less time fixating a word when it is available prior to its fixation compared to when it is masked (or replaced) by other words, referred to as *preview benefit* (Rayner, 1975). Hence it is clear that the word unit is central to readers' eye movement control in these theories. Nevertheless, both models are primarily based on research in reading of spaced, alphabetic languages where the boundaries between words are demarcated using spaces, and they assume word-based processing and saccade targeting mechanisms.

In contrast to spaced, alphabetic languages like English, Chinese is an unspaced, character based language (e.g., Hoosain, 1991, 1992). There are no explicit visual markers to separate words in written Chinese, and the space between words has the same width as the space between individual characters; a single Chinese character can

be a word by itself, or can be a part of different multi-character words when combined with other characters. According to one corpus (Chinese lexicon, 2003), 3% of words are one-character words, 64% are two-character words, 18% are three-character words, 14% are four-character words, and 1% are longer than four characters (based on type frequency counts computed separately for words of a particular length, whereby a word's type frequency represents the proportion of all the words in the corpus that are of a particular length. For example, the number of 2 character words that exist in the Chinese corpus divided by the sum of the number of 1, 2, 3 & 4+ character words that exist in the corpus). In contrast, when word tokens are considered (token frequency is defined as the frequency of occurrence of a particular word in relation to all words in the corpus), 70% of words are one-character words, 27% are two-character words, 2% are three-character words, 1% are four-character words, and fewer than 0.1% are longer than four characters. These basic distributions of word length based on written text corpora are comparable to data reported in a recent corpus based on film subtitles by Cai and Brysbaert (2010) (Type frequency: 5%, 46%, 25%, and 12% of one, two, three and four characters, respectively; Token frequency: 64%, 34%, 2% and 0.5% of one, two, three and four characters, respectively). Cai and Brysbaert argue that the subtitle corpus data provide a better estimate of daily language exposure compared to corpora based on written materials. Overall, in written Chinese, the majority of characters can join others to form multi-character words, however, one-character words are used particularly frequently and therefore have token frequencies that are much higher than other types of words. These characteristics of Chinese lead one to

question whether Chinese readers might use this information, particularly the likelihood that a character will appear as a single character word in the upcoming text, to facilitate word segmentation and eye movement control during reading.

Investigation of this issue will inform the understanding relating to how a Chinese reader segments an evenly distributed, continuous character string into words in order that each might be lexically identified. Perfetti and Tan (1999) proposed that Chinese readers have a default preference to segment character strings into two-character units, that is, they preferentially attempt to segment two characters into a single word rather than segment each single character into a word. Readers do this because most words in Chinese are two characters long (type frequency). In their experiments, participants were required to read sentences including a three-character target (ABC) region that according to the preceding context should be processed using an A-BC segmentation (i.e., "A" is a single character word and "BC" is a two-character word). However in an ambiguous condition (e.g., 经理同意**照顾客**的想法来设计产品), "A" (照) could also potentially form a word with "B" (顾), "AB" (顾客). Thus, in this condition, an ambiguity existed when readers initially read the sentence up to the "AB" characters. In a control condition (经理同意**按顾客**的想法来设计产品), "A" (按) was a character that had a similar meaning to that in the ambiguous condition (照), but was a character that could not form a word with "B" (顾), thus avoiding any potential lexical garden path. Perfetti and Tan found that reading times on the target region (ABC) were longer for the ambiguous condition than for the control condition, suggesting Chinese readers adopt a two-character

assembly strategy while initially segmenting character strings during reading.

In contrast to Perfetti and Tan's suggestion, Inhoff and Wu (2005) argued that the assignment of characters to words is not a serial, sequential process. Instead, they claimed that all of the possible words that can be formed from combinations of Chinese characters within the perceptual span (e.g., Inhoff & Liu, 1998; Yan, Zhou, Shu, & Kliegl, 2015), that is, the area from which meaningful information about words is available during a fixation in reading, are activated. The more words that are activated, the longer it takes readers to make word segmentation decisions. Li, Rayner, and Cave (2009) extended this argument and proposed that Chinese characters within the perceptual span are processed in parallel, with the characters nearer to the point of fixation being processed faster because they can be processed in high acuity vision and are more central with respect to visual attention. Within Li et al.'s model, the activation of characters feeds forward to activate word unit representations in the mental lexicon. This activation then feeds back to the characters belonging to the activated word. After a number of iterative cycles of activation, the system settles such that a single word is activated to such a degree that it is identified, and upon word recognition, the word boundary is determined. Note, however, as mentioned earlier, the majority of Chinese words are two or more characters long based on word type frequency, whereas, the mean token frequency of single character words within the language is much higher than that of multi-character words. Indeed, the Cai and Brysbaert corpus (2010) based on film subtitles showed that the top 10 most frequently used Chinese words (nine of which are single character words[1]) make up

26% of all words encountered, that is, one in every four words in the corpus. It seems a reasonable possibility, therefore, that Chinese readers might be able to process upcoming characters during reading such that they could use such probabilistic information to facilitate word segmentation processes. In other words, potentially, high token frequency single character words might form important "anchors" in the upcoming text that Chinese readers use to facilitate word segmentation processes.

Research in reading of other unspaced text like Japanese (Kajii, Nazir, & Osaka, 2001; Sainio, Hyönä, Bingushi, & Bertram, 2007) and Thai script (Kasisopa, Reilly, Luksaneeyanawin, & Burnham, 2013) has shown that some types of characters act as anchors in this way. For instance, Sainio et al. (2007) found that there was no benefit of word spacing when readers are presented with mixed Kanji-Hiragana text (ideographic-syllabic). And Hiragana characters were effectively identified as lexical units when they were surrounded by Kanji characters in the unspaced text. As they argued, the visually salient Kanji-characters (mostly derived from Chinese, representing morphological units) frequently occurred at the beginning of the words, and served as sufficiently strong segmentation cues like anchors, to signal word beginnings as well as more global word boundaries. In this case, Japanese readers could parse character-strings into words in parafoveal vision when a Kanji character appeared in the string and introducing word spacing did not result in a benefit. In addition, Kasisopa et al. (2013) found in Thai, an unspaced alphabetic language, that the positional frequency of characters within words (word-initial and word-final character frequency) influenced readers' initial landing position on a word. They

argued that Thai readers could use within-word positional information to compute word boundaries and thus aid readers' saccadic targeting.

Similar to research in Thai reading, Yen, Radach, Tzeng and Tsai (2012) investigated whether positional frequencies of Chinese characters are informative for readers. They manipulated the congruency of within-word character positions in relation to the end character of a target word. In the congruent condition, the end character was frequently used in this position. In contrast, in the incongruent condition, the end character did not usually occur in this position. They found that readers had longer gaze durations and made more refixations on words with incongruent than with congruent positional frequency characters, arguing that Chinese readers use within word character positional frequency information as a cue for word segmentation.

Given these findings, it is perhaps reasonable to suggest that the frequency of a Chinese character as a single character word, or as an initial constituent of a multi-character word might have a differential influence on word segmentation and eye movement control during Chinese reading. In the present study, a two-character Chinese word (C12) was embedded in a sentence as the target word. We manipulated whether the first character (C1) was likely to be a single character word, or the first character of a two character word, to investigate whether Chinese readers use probabilistic information in word segmentation and lexical identification. Furthermore, the boundary paradigm (Rayner, 1975) was employed to investigate whether Chinese readers can use such information about the first character of the target character string

in relation to parafoveal processing of the second character of the target prior to direct fixation of the second character. We, therefore, positioned the invisible boundary between the first (C1) and the second character (C2) of the target word. When the reader's eyes crossed the boundary, either an identity or a nonsense pseudocharacter preview changed to the target character C2. In this way we were able to determine whether the probability that the currently fixated character, C1, was likely to be a single character word modulated the extent to which readers preprocessed C2 (the upcoming character). We evaluated this possibility in relation to fixation times on the pre- and post- boundary characters. We predicted that if Chinese readers adopt the two-character word unit processing strategy as per Perfetti and Tan (1999), then the probabilistic information regarding the likelihood that a character is a single character word should not influence processing of both C1 and C2. Alternatively, Chinese readers might segment words in parallel with characters nearer fixation being processed faster than those further away (as per Li et al., 2009). If such processing occurred, then an influence of the probabilistic information should be observed. Specifically, when C1 is more likely to be a single character word, then activation of a two character word comprised of C1 and C2 should be reduced, and readers should parafoveally process C2 to a lesser degree (i.e. show reduced parafoveal preview effect) than when C1 is more likely to be the first character of a two character word.

Method

Participants

Forty-four undergraduate students at Tianjin Normal University were paid to

participate in the eye tracking experiment. They were all native speakers of Chinese with normal or corrected to normal vision.

Apparatus

Participants' eye movements were monitored using a SR Research Eyelink1000 system at a sampling rate of 1000 Hz. Viewing was binocular while only eye movements of the right eye were recorded. The sentences were presented on a 17-inch SAMSUNG SyncMaster 959NF monitor with a 1,024 × 768 pixel resolution and a refresh rate of 110 Hz. Stimuli were presented in black on a white background in Song font. Each character was approximately 27×27 pixels in size. The viewing distance was 65 cm, and at this distance each Chinese character subtended approximately 0.85° of visual angle.

Materials and Design

Two-character words were selected as targets. The probability of the first character (C1) of the target word (C12) being a single character word was manipulated. This probability was calculated as the frequency count of C1 used as a single character word, divided by the sum of frequency counts of words that contain C1 regardless of whether C1 was a single character word or a constituent of a multiple character word in the Cai and Brysbaert (2010) database that contains 46.8 million characters and 33.5 million words. The higher the C1 probability, the more likely it is used as a single character word rather than a constituent of a multiple character word.

Ninety-six two-character target words were selected from the database. Of these, half were in the high single character word likelihood condition, and the probability of

C1 being used as a single character word was higher than 70% (Mean = 84.5 %, SD = 8.0 %). The remaining half was in the counterpart low single character word likelihood condition, in which the probability of C1 being used as single character words was lower than 30% (Mean = 10%, SD = 7.4%). A *t*-test showed that words in the two conditions differed in the probability of C1 being used as single character words, t (94) = 33.4, $p < .001$. However the neighborhood size (i.e., the number of words sharing the same first constituent character) of C1 was matched in the high- (Mean = 8.7, SD = 5.0) and low-single character word likelihood conditions (Mean = 7.7, SD = 3.6), $t = 1.34$, $p > 0.05$. Furthermore, the number of strokes and frequency of C1, C2 and the whole two-character word were also matched (all *t*s < 1.2, all *p*s > 0.05; see Table 1).

Insert Table 1 about here

Forty-eight sentence frames were constructed for each pair of target words which were embedded in the middle part of each sentence and the context preceding the target words was neutral (see Figure 1). All the sentences were rated on a 5 point scale for their naturalness by 16 university students who did not take part in the eye tracking study. The mean score was 4.2 (where a score of 5 was "very natural"), and there was no difference between the high- and low-single character word likelihood conditions ($t < 1$). The contextual predictability of the target words was assessed by 19 college students who did not take part in the eye tracking experiment (10 participants conducted a cloze task and 9 conducted a sentence completion task). The mean predictability for the target word (C12) was very low (Sentence completion task:

0.2% and 0.6% in the high- and low-single character word likelihood conditions, respectively; Cloze task: 5.2% and 3.8% in the high- and low-single character word likelihood conditions, respectively), and was not different between conditions ($t < 1$).

Using the boundary paradigm (Rayner, 1975) the preview of the second character (C2) of the two-character target word was manipulated. The invisible boundary was placed between the two characters (C1 and C2) of the target. As soon as the eyes crossed the invisible boundary, an identical or pseudocharacter preview was replaced by the target character (it took approximately 10ms to complete a boundary change). The pseudocharacters were created using True Font software, they resembled real characters but were completely meaningless. Furthermore, the pseudocharacter previews did not contain any of the radicals of the target character, and the number of strokes of the pseudocharacter previews was matched with the targets in the high- and low-single character word likelihood conditions.

The experiment was a 2 (Likelihood that C1 was a single character word: High vs. Low) × 2 (Preview of C2: Identical vs. Pseudocharacter) within-participant design. Four files were constructed, with each file containing 48 sentences (12 sentences in each condition). Conditions were rotated across files according to a Latin square, each sentence was read by each participant only once. Sentences in each condition were presented randomly. Additionally, 6 practice sentences were presented at the beginning of the experiment. There were 18 comprehension questions that participants were required to try to answer correctly with a yes/no response.

An example sentence with the target word and the preview stimuli is shown in

Figure 1.

Insert Figure 1 about here

Procedure

Each participant was tested individually. Participants were instructed to read sentences for comprehension at their normal pace. They were informed that occasionally a comprehension question would appear after a sentence, and they should try hard to answer the questions correctly. Prior to the start of the experiment, a 3-point horizontal calibration procedure was completed with an average calibration error below 0.25 degrees. After a successful calibration, the sentences were presented in turn. During the experiment, each trial started with a fixation point presented at the location of the first character of the upcoming sentence. Participants pressed a response key on a button box to terminate the display once they finished reading a sentence. When a comprehension question appeared, participants gave answers to the questions by pressing response keys, and their answers were recorded by the computer. The experiment took approximately 15-25 min. The overall comprehension rate was 96% indicating that participants read and fully understood the sentences.

Results

Fixations less than 80 ms or greater than 800 ms were discarded. Trials were excluded due to (1) display changes occurred during a fixation, (2) tracker loss or blinks on or just before the target word during the first pass reading, (3) eye movement measures above or below three standard deviations from the participant's mean. This resulted in the removal of 10.9% of the data prior to conducting the

analyses.

Analyses were conducted for the first character (C1), the second character (C2) and the whole two-character word. For each interest area four first-pass measures were computed: first fixation duration (FFD, the duration of the first fixation on a region), single fixation duration (SFD, the fixation duration when only one fixation was made on the region during first pass reading), gaze duration (GD, the sum of all fixations on a region before moving to another region), and skipping probability (SP, the proportion of times a region was not fixated during first pass reading). The means and standard deviations for the eye movement measures are shown in Table 2.

Insert Table 2 about here

To analyze the data linear mixed models (LMM) were conducted using the lme4 package (version 1.1-7) in R (R Development Core Team, 2014). As fixed factors we included the Single Character Word Likelihood and Preview conditions and their interaction. A "full" random model including intercepts and slopes for the main effects and their interactions with participants and items as random factors did not converge for the dependent measures in all likelihood due to missing values related to the high skipping rates. Therefore we ran a model with intercepts and where possible slopes for the main effects with participants as a random factor and with intercepts for the items as random factors. Furthermore, two contrasts were programmed to test for preview effects in the two single character word likelihood conditions. The first contrast compared the identical and pseudocharacter previews in the high single character word likelihood condition, and the second contrast compared the identical

and pseudocharacter previews in the low single character word likelihood condition. The fixation times were analyzed using log-transformed data and the skipping rates were analyzed using logistic LMM's. Fixed effect estimations for the fixation times and skipping probability measures are shown in Table 3.

Insert Table 3 about here

The first character (C1)

We first considered measures on the first character (C1) as these might potentially reflect effects of the C2 preview prior to fixation. Some may argue that these reflect so-called parafoveal-on-foveal effects, though note that adjacent characters are both within foveal vision and are also strictly speaking within-word effects (Inhoff, Radach, Starr, & Greenberg, 2000; Zhou, Kliegl, & Yan, 2013). It is for this reason that we will refer to these simply as effects of the pseudocharacter preview that occur prior to the boundary change.

There was no reliable effect of the preview of the C2 mask on first and single fixation times on C1, though a marginal effect occurred on skipping probability with readers skipping characters more often for the identical preview. There was also a marginal effect of single character word likelihood in gaze duration, such that gaze durations were shorter in the high single character word likelihood condition (M = 270ms) compared to the low single character word likelihood condition (M = 287ms). More interestingly, there was an interaction between the single character word likelihood and preview conditions across all first pass fixation time measures. The planned contrasts showed that in the case of all measures this was due to increased

times on C1 when C2 was masked than when it was not when C1 was less likely to be a single character word (the low single character word likelihood condition, though this effect was marginal for SFD), but this effect did not occur (or was greatly reduced) when C1 was more likely to be a single character word (the high single character word likelihood condition). For first fixation durations, this effect was 18ms for the low single character word likelihood condition with a difference of -1ms for the high single character word likelihood condition; the respective differences for single fixation durations were 16ms and -3ms, and for gaze durations 35ms and 7ms. Thus, we obtained robust effects of the pseudocharacter preview for fixations on the C1 when it was less likely to be a single character word. It appears that whilst the C1 was fixated, probabilistic information associated with that character affected the extent to which C2 was processed. Clearly effects of the preview did not occur when the C1 was, probabilistically, a single character word, and thus signaled that the upcoming characters to the right were more likely to be part of a new word.

The second character (C2)

Measures for C2 reflect processing after the preview has been changed into its intended form. There was a significant effect of preview of C2 in all measures such that readers fixated C2 for less time and skipped it more often when they had received an identical preview (FFD = 259ms, SFD = 258ms, GD = 268ms, SP = 0.48) rather than a pseudocharacter preview (FFD = 303ms, SFD = 305ms, GD = 322ms, SP = 0.44). Unsurprisingly, this reflects the basic preview effect (e.g., Rayner, 1975, 1998, 2009). There was also a significant effect of single character word likelihood in all

fixation time measures, such that readers fixated for less time on the C2 when C1 was less likely to be a single character word (FFD = 276ms, SFD = 277ms, GD = 290ms) than when C1 was more likely to be a single character word (FFD = 287ms, SFD = 287ms, GD = 300ms). There was no significant interaction between the single character word likelihood and preview conditions across all of the measures, however, since we expected that there might be a difference between the high- and low-single character word likelihood conditions for identity previews, we undertook contrast analyses to examine this possibility. These analyses showed that there were marginal differences between the high- and low-single character word likelihood conditions for the identical previews on all fixation time measures (all ps < .07). For completeness, there were no effects for the pseudocharacter preview conditions (all *p*s > .05). For the identical preview conditions, readers fixated for less time on C2 when C1 was less likely to be a single character word (low single character word likelihood condition). This numerical trend is consistent with the suggestion that increased processing of the C2 preview when C1 was likely to be part of a two character word resulted in more efficient processing of C2 when it was ultimately fixated.

The whole two-character word

For the whole two-character word, there was a significant C2 mask effect in FFD, GD and SP, such that readers fixated the whole word for less time and skipped it more often in the identical preview condition (FFD = 262ms, GD = 333ms, SP = 0.16) than in the pseudocharacter preview condition (FFD = 277ms, GD = 411ms, SP = 0.12). Again this reflects the basic effect of a pseudocharacter preview. Furthermore, readers

skipped the whole word reliably less often in the high single character word likelihood condition (0.13) than in the low single character word likelihood condition (0.15). As with the C1 analyses, there were reliable interactions between the single character word likelihood and preview condition in FFD and SFD. The planned contrasts showed that for first fixation durations, the cost of a C2 mask was 25ms for the low single character word likelihood condition but only 4ms for the high single character word likelihood condition; for single fixation durations, it was 25ms for the low single character word likelihood condition and 2ms for the high single character word likelihood condition. Whilst these results are similar in pattern to the effects observed on C1, they are less robust due to the inclusion of the fixations on the second character, and of course, due to summation of fixations both before and after the boundary change. Presumably, this is also why the interaction was not robust for gaze duration.

Discussion

Since Chinese is an unspaced, character based language with no clear demarcation of word boundaries and since there is often ambiguity regarding which character strings comprise a word (Liu, Li, Lin & Li, 2013; Yan, Kliegl, Richter, Nuthmann, & Shu, 2010; Zang, et al., 2011), it is important to investigate how Chinese readers segment character strings into words as they read. In the present study we assessed whether Chinese readers were sensitive to information concerning how often a character appears as a single character word compared with the first character in a two character word, and whether such information is used to modulate

processing and word segmentation in relation to characters to the right of the current fixation. To investigate this question, we directly manipulated both the likelihood that the first character of a two-character Chinese target string would be a single character word, and the preview of its second character using the boundary paradigm (Rayner, 1975). We analyzed eye movement measures for the first character, the second character and the whole target word. Our analyses showed standard preview benefit effects (Rayner, 1975) for all reading times measures and the skipping probabilities associated with the second character, as well as complementary effects associated with the C2 mask in most measures for the whole word region. These results are not surprising and reflect the degree to which readers benefit from an identity preview of a word to the right of fixation relative to a preview of a pseudocharacter clearly showing that readers preprocess Chinese characters prior to their direct fixation. Note that because readers skipped the first character of the target character string more often when an identical versus pseudocharacter preview of its second character was presented, we have evidence that the preview affected decisions of where to target the eyes even when that information lay to the right of the current fixation. All of these effects replicated findings previously reported in the literature (see Li, Zang, Liversedge, & Pollatsek, 2015 and Zang, Liversedge, Bai, & Yan, 2011 for reviews of studies investigating saccadic targeting in Chinese). These findings clearly indicate that the preview manipulation that we achieved using the boundary paradigm was effective.

Of greater theoretical importance were the interactions between preview type and

single character word likelihood condition on the first character of the target string for the eye fixation measures (as well as on the entire two-character target string for the first and single fixation duration). We only obtained robust effects of the second character mask when the first character was likely to be the first character of a two character word, and not when it was likely to be a single character word. (see Cui, Drieghe, et al., 2013; Cui, Yan, et al., 2013; Drieghe et al., 2010 for similar findings). Furthermore, this increased preview processing when the first character was likely to be part of a two character word resulted in more efficient processing of the second character when it was subsequently fixated. This suggests that Chinese readers use probabilistic information about the likelihood of a Chinese character being a word to modulate the extent to which they processed the character to the right prior to fixation.

This finding is inconsistent with the proposal put forward by Perfetti and Tan (1999), who argued that Chinese readers have a default preference to segment two characters into a single word rather than segment each character into a word. If Chinese readers had adopted this word segmentation strategy then we would not have seen the modulation of preview effects by the likelihood that the first character of the target character string was a one character word, compared to the first of a two character word. In contrast, our finding is consistent with the Li et al. (2009) model of word segmentation and identification in Chinese reading. Li et al. argue that all words in the perceptual span are activated in parallel, with increased activation for those words closer to fixation. With continued activation and competition between

words, over time, a single word is identified, and it is at the point when the word is identified that the word boundary is determined.

In the present study, when the first character of a two-character target word is more likely to form a single character word in its own right, it acts like an "anchor" to signify that there is a word boundary. Consequently, additional characters to the right of fixation are not required for the formation of an entire lexical unit, and therefore, those characters are not processed to the same degree prior to fixation as they would be if they were likely to join the first character to form a word. By contrast, when the first character is more likely to be part of a two-character word (and, therefore, less likely to be a single character word), then this signals that processing of the upcoming character is likely to be beneficial to the identification of the word. To this extent, in this situation processing of the first character licenses processing of the upcoming character(s), in order to facilitate lexical identification of the entire multi-character word. The consequence of this is a reliable C2 mask effect in this situation (see Cui, Drieghe, Bai, Yan, & Liversedge, 2014).

It may also be the case that our results have implications for models of eye movement control during reading such as E-Z Reader (Reichle et al., 2003; Reichle, 2011) and SWIFT (Engbert et al., 2005; Engbert & Kliegl, 2011). Currently in these models, probabilistic lexicality cues between the constituent characters of words, that is, the likelihood that a character is a single character word compared with the initial character of a multi-character word, do not modulate the degree to which an upcoming character is processed. Perhaps as empirical evidence for this kind of

effect builds, the models may need to be modified to reflect this constraint on processing. However, one factor that needs to be considered carefully in relation to any such modifications concerns whether any such effects are driven by processing of characters to the right of fixation that fall outside of current foveal processing. As we noted earlier, the effects we report here may well be considered to be foveal rather than parafoveal, and it is for this reason that we have been careful to talk about the effects as reflecting processing of an unfixated character rather than a parafoveal character.

We have argued firmly that the present results indicate that probabilistic lexicality cues associated with Chinese characters exert a strong influence over how a word is segmented and processed during reading. It is also important to note that the current findings cannot be explained by neighborhood size (e.g., the number of words sharing the same first constituent character) (Tsai, Lee, Lin, Tzeng, & Hung, 2006). It might initially seem to be the case that when the first character of the target string is more likely to be a single character word it might combine with fewer other characters to form a word. In contrast, when the first character of the target string is less likely to be a single character word it might potentially combine with many other characters to form a word. However, we foresaw this possibility, and as indicated earlier, we controlled the number of character neighbors associated with the first character across the two single character word likelihood conditions.

A final point of potential concern may be that while we have explained our results in terms of the role of probabilistic combinatorial information associated with

Chinese characters, the effects might actually arise due to the predictability of the second character on the basis of the first character. That is, the high predictability of the second character given the first character (when the two together form a two character word) might contribute to the lexical licensing process. This seems to us to be a fair concern. In order to formally assess this possibility, 22 participants were given sentence fragments up to and including the first character of the target character string, and were asked to complete them. The results showed that the second character was 31% and 46% predictable in the high- and low-single character word likelihood conditions, respectively. Given that the predictability of the two-character words from the global context was very low indeed (0.2% and 0.6% in the high- and low-single character word likelihood conditions, see Method section), we can be certain that any substantive effects of predictability on C2 must therefore have arisen from C1. That is, in terms of predictability, it is the first character of the target string that drives the effects, not the preceding sentential context. We can extend this argument to some degree by considering the sentence completion data in relation to the size of the C2 mask effects and the preview effects that we observed in our experiment. Let us consider again the fixation times on C1, that is, those fixations immediately prior to the boundary. We will also focus our attention on the reading time measure for C1 for which we obtained the largest C2 mask effects, namely, gaze duration. The sentence completion data show that at this character in the sentence, in the high single character word likelihood condition, participants produced C2 to complete the fragment (which included C1) on 31% of occasions. We also know that we obtained

a 7ms preview benefit effect at C1 for the high single character word likelihood condition. Next, if we consider the sentence completion data for the low single character word likelihood conditions at the same point in the sentence, we see that participants used C2 to complete the sentence fragment 46% of the time. Assuming a linear relationship between completion rates and preview effect sizes, we might therefore expect to see a preview effect that is approximately 150% of the magnitude of the effect observed in the high predictability condition (i.e., we might expect to see an effect in the order of 11ms). In fact, however, the size of the preview effect, at 35ms, was far greater than this[2]. Thus, on this basis, we might conclude that to produce preview effects of this magnitude at this point in the sentence, there is most likely an influence in addition to the effect of predictability that we have observed in our sentence completion data. We suggest that this additional influence is the information about the probabilistic likelihood that C1 is either a single character word, or instead the first character of a two character word. Of course, the idea that there are multiple sources of influence over the combinatorial possibilities that exist between Chinese characters in relation to the compositionality of words is not particularly novel. And of course, different sources of influence are not mutually exclusive. However, most importantly for the current results, based on these sentence completion analyses, it seems reasonable to conclude that predictability per se cannot account for the entirety of the preview effect we have obtained.

To summarize, we wished to investigate whether Chinese readers were sensitive to information about how often a Chinese character appears as a single character word

compared with the first character in a two character word, and whether this information facilitates word segmentation and processing of upcoming character strings. On the basis of Li et al., (2009), we predicted a reduced preview benefit when the first character of a two-character target string was more likely to be a single character word than the first character of a two character word. We consider that this hypothesis was confirmed by our findings, demonstrating that Chinese readers use probabilistic information about the likelihood of a Chinese character being a word online to modulate the extent of parafoveal processing.

References

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitle. *PLoS ONE, 5(6), e10729.* DOI: 10.1371/journal.pone.0010729.

Chinese Linguistic Data Consortium. (2003). *Chinese lexicon* [现代汉语通用词表] (CLDC-LAC-2003-001). Beijing, China: Tsinghua University, State Key Laboratory of Intelligent Technology and Systems, and Chinese Academy of Sciences, Institute of Automation.

Cui, L., Drieghe, D., Yan, G., Bai, X., Chi, H., & Liversedge, S.P. (2013). Parafoveal processing across different lexical constituents in Chinese reading. *The Quarterly Journal of Experimental Psychology, 66,* 403-416.

Cui, L., Drieghe, D., Bai, X., Yan, G., & Liversedge, S.P. (2014). Parafoveal preview benefit in unspaced and spaced Chinese reading. *The Quarterly Journal of Experimental Psychology*, DOI: 10.1080/17470218.2014.909858.

Cui, L., Yan, G., Bai, X., Hyönä, J., Wang, S., & Liversedge, S.P. (2013). Processing of compound-word characters in reading Chinese: An eye-movement–contingent display change study, *The Quarterly Journal of Experimental Psychology, 66*, 527-547.

Drieghe, D., Pollatsek, A., Juhasz, B. J., & Rayner, K. (2010). Parafoveal processing during reading is reduced across a morphological boundary. *Cognition, 116*, 136-142.

Engbert, R. & Kliegl, R. (2011). Parallel graded attention models of reading. In

Liversedge, S.P., Gilchrist, I.D. & Everling, S. (Eds.), *The Oxford Handbook of Eye Movements* (pp.787-800). Oxford University Press.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*, 777-813.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: A case study of Chinese*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoosain, R. (1992). Psychological Reality of the word in Chinese. In H.-C. Chen & O. J.L. Tzeng. (Eds.), *Language processing in Chinese* (pp. 111-130). Amsterdam, Netherlands: North-Holland.

Inhoff, A. W., & Liu, W. (1998). The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 20-34.

Inhoff, A. W., Radach, R., Starr, M., & Greenberg, S. (2000). Allocation of visuo-spatial attention and saccade programming during reading. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 221–246). Oxford, UK: North-Holland/Elsevier.

Inhoff, A., & Wu, C. (2005). Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & Cognition, 33*, 1345-1356.

Kajii, N., Nazir, T.A., & Osaka, N. (2001). Eye movement control in reading unspaced text: The case of Japanese script. *Vision Research, 41*, 2503-2510.

Kasisopa, B., Reilly, R. G., Luksaneeyanawin, S., & Burnham, D. (2013). Eye movements while reading an unspaced writing system: the case of Thai. *Vision Research, 86*, 71-80.

Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology, 58*, 525-552.

Li, X., Zang, C., Liversedge, S.P., & Pollatsek, A. (2015). The role of words in Chinese reading. In A. Pollatsek & Treiman, R. (Eds.), *The Oxford Handbook of Reading* (pp.232-244). Oxford University Press.

Liu, P., Li, W., Lin, N., Li, X. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading? *PLoS ONE 8(2): e55440.*

Liversedge, S.P., & Findlay, J.M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science, 4*, 6-14.

Perfetti, C.A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In Wang, J., & Inhoff, A. W. (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 115-134). Hillsdale, NJ: Lawrence Erlbaum Associates.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*, 65-81.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rayner K. (2009). The thirty-fifth Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*, 1457-1506.

Reichle, E. D., (2011). Serial-Attention Models of Reading. In Liversedge, S.P., Gilchrist, I.D. & Everling, S. (Eds.), *The Oxford Handbook of Eye Movements* (pp.767-786). Oxford University Press.

Reichle, E.D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*, 445-476.

Sainio, M., Hyönä, J., Bingushi, K., & Bertram, R. (2007). The role of interword spacing in reading Japanese: An eye movement study. *Vision Research*, *47*, 2575-2584.

Tsai, J.L., Lee, C.Y., Lin, Y.C., Tzeng, O.J.L., & Hung, D.L. (2006). Neighborhood size effects of Chinese words in lexical decision and reading. *Language and Linguistics, 7*, 659-675.

Yan, M., Kliegl, R., Richter, E., Nuthmann, A., & Shu, H. (2010). Flexible saccade target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology, 63,* 705-725.

Yan, M., Zhou, W., Shu, H., & Kliegl, R. (2015). Perceptual span depends on font size during the reading of Chinese sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 209-219.

Yen, M. -H., Radach, R., Tzeng, O. J. -L., & Tsai, J. -L. (2012). Usage of statistical

cues for word boundary in reading Chinese sentences. *Reading and writing*, *25*, 1007-1029.

Zang, C., Liversedge, S.P., Bai, X., & Yan, G. (2011). Eye movements during Chinese reading. In S.P. Liversedge, I. Gilchrist, & S. Everling. (Eds.), *The Oxford handbook of eye movements* (pp. 961-978). Oxford University Press.

Zhou, W., Kliegl, R., & Yan, M. (2013). A validation of parafoveal semantic information extraction in reading Chinese. *Journal of Research in Reading, 36 (S1)*, S51-S63.

Acknowledgments

We are grateful for support from the Recruitment Program of Global Experts (1000 Talents Award from Tianjin); Natural Science Foundation of China Grants (31100729, 81471629); a postgraduate scholarship from the China Scholarship Council. Denis Drieghe and Simon Liversedge were supported by Research Grant RPG-2013-205, and Simon Liversedge was supported by Research Grant F/00 180/AN from the Leverhulme Trust. We also wish to thank Juan Liu and Fang Li for assistance with stimuli construction and participant testing.

Footnote

1. In the corpus based on written text by Chinese Linguistic Data Consortium (Chinese lexicon, 2003), the top 100 frequently used Chinese words are all one-character words, making up 30% of all words encountered.

2. In order to further investigate the possibility that the predictability of C2 on the basis of C1 could contribute to our effects, we undertook a further set of LMM analyses in which predictability was included as a fixed factor. These analyses produced an identical set of results for the other factors indicating that this variable did not cause our effects.

Figure Caption

Figure 1 An example stimuli used in the experiment. The vertical black line represents the position of the invisible boundary. As the eyes crossed the boundary, the preview was replaced by the target.

C1	Preview	Sentences
High Single Character Word Likelihood	Identical	近年来当地的群众已经**砍|伐**了大批的珍稀林木。
	Pseudocharacter	近年来当地的群众已经**砍|犰**了大批的珍稀林木。
	Translation	*In recent years the local people have **cut down** a large number of rare trees.*
Low Single Character Word Likelihood	Identical	近年来当地的群众已经**拯|救**了大批的珍稀林木。
	Pseudocharacter	近年来当地的群众已经**拯|旂**了大批的珍稀林木。
	Translation	*In recent years the local people have **saved** a large number of rare trees.*

Table 1 The number of strokes and frequency (per million) of the first character (C1), the second character (C2) and the whole two-character word in high- and low-single character word likelihood conditions. Standard deviations are provided in parentheses.

Single Character	C1		C2		The whole word	
Word Likelihood	Strokes	Frequency	Strokes	Frequency	Strokes	Frequency
High	9.1 (1.3)	142 (198)	8.6 (1.9)	831 (1602)	17.6 (2.2)	5.0 (10.8)
Low	8.8 (0.8)	105 (144)	8.7 (1.7)	512 (932)	17.4 (2.0)	7.6 (12.3)

Table 2 Eye movement measures for the first character (C1), the second character (C2) and the whole two-character words. Standard deviations are provided in parentheses.

	High Single Character Word Likelihood		Low Single Character Word Likelihood	
	Identical Preview	Pseudocharacter Preview	Identical Preview	Pseudocharacter Preview
		C1		
FFD	258(78)	257(100)	261(88)	279(109)
SFD	259(79)	256(99)	261(88)	277(109)
GD	266(92)	273(125)	269(100)	304(149)
SP	0.47(0.50)	0.44(0.50)	0.49(0.50)	0.45(0.50)
		C2		
FFD	269(95)	304(114)	249(81)	302(126)
SFD	268(90)	305(115)	248(81)	305(127)
GD	277(105)	322(133)	258(99)	322(137)
SP	0.48(0.50)	0.38(0.49)	0.48(0.50)	0.39(0.49)
		The whole word		
FFD	266(84)	270(99)	258(88)	283(111)
SFD	267(84)	269(102)	257(86)	282(110)
GD	336(172)	411(257)	330(192)	410(244)
SP	0.14(0.35)	0.11(0.32)	0.17(0.38)	0.13(0.34)

Table 3 Fixed effect estimates for the first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD) and skipping probability (SP) across all regions.

	Single Character Word Likelihood	Preview	Single Character Word Likelihood × Preview	Contrast1[a]	Contrast2[b]
			C1		
FFD	0.03	0.01	0.10*	-0.04	0.06*
SFD	0.03	0.01	0.10*	-0.04	0.06§
GD	0.05§	0.04	0.12*	-0.02	0.10 **
SP	0.06	-0.16§	-0.02		
			C2		
FFD	-0.05*	0.13***	0.03		
SFD	-0.04§	0.13***	0.04		
GD	-0.04§	0.16***	0.05		
SP	0.02	-0.42***	0.08		
			The whole word		
FFD	0.003	0.04*	0.07*	0.03	0.07 **
SFD	0.002	0.03	0.08*	-0.01	0.07*
GD	-0.01	0.16**	0.05		
SP	0.28§	-0.47**	-0.05		

[a] Refers to the comparison between the identical and pseudocharacter preview in high single character word likelihood condition; [b] Refers to the comparison between the

identical and pseudo- character preview in low single character word likelihood condition.

*** $p < .001$, ** $p < .01$, * $p < .05$, § $p < .10$