Word Segmentation in Incidental Vocabulary Learning Through Reading: Differential Roles of Initial and Final Morpheme Positional Probabilities

Authors: Bai Xuejun, Liang Feifei, Feng Linlin, Liu Ying, Li Xin, Liang Feifei

Date: 2023-11-07T16:47:34+00:00

Abstract

This study employs two parallel experiments to investigate the changing patterns of how positional probability information of initial and final morphemes influences word segmentation during repeated learning of novel words. Using the reading-while-vocabulary-learning paradigm, two-character pseudowords serve as novel words. Experiment 1 manipulates the high versus low positional probability of the initial morpheme while holding the final morpheme constant; Experiment 2 manipulates the high versus low positional probability of the final morpheme while holding the initial morpheme constant. An eye tracker records eye movement trajectories of college students during reading. The results demonstrate: (1) The word segmentation effect of positional probability information for both initial and final morphemes gradually decreases as learning occurrences of the novel words increase during reading, exhibiting a "familiarity effect." (2) The "familiarity effect" of initial morpheme positional probability information manifests in two relatively late eye-movement measures—regression path duration and total fixation count—whereas the "familiarity effect" of final morpheme positional probability information begins from first-pass gaze duration, extends to regression path duration, and further persists into total fixation time. These findings indicate that both initial and final morpheme positional probability information affect word segmentation in reading-while-vocabulary-learning, but the initial morpheme's effect is more prolonged and stable, supporting the view that the initial morpheme possesses an advantage in two-character word processing.

Full Text

Word Segmentation during Incidental Word Learning in Reading: Different Roles of Initial and Final Character Positional Probabilities

LIANG Feifei[1],[2],[3], FENG Linlin[2], LIU Ying[2], LI Xin[1],[2],[3], BAI Xuejun[1],[2],[3]

[1]Key Research Base of Humanities and Social Sciences of the Ministry of Education, Academy of Psychology and Behavior, Tianjin Normal University, Tianjin 300387
[2]Faculty of Psychology, Tianjin Normal University, Tianjin 300387
[3]Tianjin Social Science Laboratory of Students' Mental Development and Learning, Tianjin 300387

Abstract

This study employed two parallel experiments to investigate how the roles of initial and final character positional probability information in word segmentation change during repeated exposure to novel words. Using the incidental word learning paradigm, we constructed two-character pseudowords as novel lexical items. Experiment 1 manipulated the positional probability of the initial character while holding the final character constant; Experiment 2 manipulated the positional probability of the final character while holding the initial character constant. Eye movements of university students were recorded during natural reading. Results revealed: (1) The word segmentation effects of both initial and final character positional probability information gradually diminished as learning trials increased, demonstrating a "familiarity effect." (2) The familiarity effect for initial character positional probability emerged in relatively late eye-movement measures (regression path duration and total fixation count), whereas the familiarity effect for final character positional probability began as early as gaze duration and persisted through regression path duration and total fixation time. These findings indicate that both initial and final character positional probabilities contribute to word segmentation during incidental word learning in Chinese reading, but the initial character's role extends over a longer and more stable time course, supporting the view that the initial character holds a processing advantage in two-character word recognition.

Keywords: character positional probability, word segmentation, incidental word learning, Chinese reading

1. Introduction

Words constitute the fundamental unit of processing in reading (Bai et al., 2008; Li et al., 2022; Li & Pollatsek, 2020; Radach & Kennedy, 2004; Rayner, 1998, 2009). In most alphabetic writing systems such as English and German, inter-

word spaces serve as natural segmentation cues that facilitate visual word segmentation, promote lexical identification, and guide saccade targeting (Clifton et al., 2016; Perea & Acha, 2009). In contrast, Chinese reading lacks such visual segmentation cues, making the word segmentation process considerably more complex (Bai et al., 2019; Liang et al., 2019; Bai et al., 2013; Blythe et al., 2012; Li et al., 2009; Li & Pollatsek, 2020; Zang et al., 2013). This complexity becomes particularly salient during incidental word learning, where readers must first segment novel words before they can infer their meanings through bottom-up lexical information and top-down contextual information, gradually constructing novel lexical representations and incorporating them into the mental lexicon. Understanding what cues Chinese readers use to segment novel words is crucial for elucidating the mechanisms of word segmentation during incidental word learning and provides theoretical support for developing efficient vocabulary learning methods.

Character positional probability information represents an effective statistical cue for word segmentation in Chinese reading. It refers to the probability of a character appearing in a specific position within multi-character words (e.g., word-initial, medial, word-final) (Lian et al., 2021; Liang et al., 2023). For example, among the 29 two-character words containing the character "各" (e.g., "各位" [everyone], "各自" [respectively], "各种" [various]), "各" always appears in the initial position, yielding a 100% probability of being word-initial. Conversely, among the 47 two-characters words containing "员" (e.g., "成员" [member], "演员" [actor], "员工" [staff]), "员" appears word-initially only in "员工" while occupying the final position in the remaining 46 words, making its positional cue strongly indicative of word-final position. Characters in specific positions within words thus provide segmentation information. For instance, in the sentence "学校定期举办｜各种活动丰富同学们的课余生活" (The school regularly holds various activities to enrich students' extracurricular lives), encountering "各" signals the end of the preceding word n-1 ("举办" [holds]) and the beginning of the current word n ("各种" [various]). Similarly, encountering "员" signals the end of word n and the beginning of word n+1.

Liang et al. (2015, 2017) conducted two experiments to examine whether children and adults utilize character positional probability information for word segmentation during incidental word learning. They constructed two-character pseudowords as novel words and manipulated the positional probabilities of both initial and final characters, creating three experimental conditions: (1) a consistent condition where the initial character frequently appeared word-initially and the final character frequently appeared word-finally, providing segmentation information consistent with character positional probabilities (e.g., "挑尔"); (2) an inconsistent condition where the initial character rarely appeared word-initially and the final character rarely appeared word-finally (e.g., "子左"), providing inconsistent segmentation information; and (3) a balanced condition where both characters appeared word-initially or word-finally with roughly 50% probability (e.g., "皮合"). Each novel word was embedded in six high-constraint contexts to facilitate lexical representation formation. Results showed that both children

and adults exhibited significantly longer gaze durations and total fixation times on novel words in the inconsistent condition compared to the consistent and balanced conditions, demonstrating that Chinese readers can use character positional probability information to segment novel words. Building on Li et al.'s (2009) basic assumptions about Chinese word segmentation and identification models, Liang et al. proposed a potential mechanism for how character positional probability information operates in Chinese reading. When all characters within the perceptual span are activated in parallel, their positional information is simultaneously activated. The higher the probability of a character appearing in a particular position, the stronger the activation of that character in that position. When this activated positional information is transmitted to the word unit, if the activated character's positional information matches its current position in the word unit, the word easily reaches threshold and becomes identified. Conversely, when the activated positional information conflicts with the character's current position, cognitive conflict arises, requiring additional time to resolve. Thus, character positional probability information is hypothesized to operate during the "character-to-word assignment" stage of word processing and identification. However, because Liang et al. simultaneously manipulated both initial and final character positional probabilities to maximize experimental control, their study could not determine whether the initial character, final character, or both contributed to the observed segmentation effects.

A series of studies have demonstrated that initial and final characters play different roles and undergo different processing mechanisms in Chinese word recognition. First, the visual complexity of initial characters affects both lexical identification and saccade targeting, whereas final character complexity influences only lexical identification, and to a lesser degree (Ma & Li, 2015). Second, while both initial and final character frequencies affect lexical identification, the effect of final character frequency is constrained by initial character frequency (Yan et al., 2006). Third, activation of initial character features (e.g., contextual diversity, semantic transparency) occurs earlier in the time course (0-100 ms), whereas activation of final character features begins later (100-200 ms) (Tsang & Zou, 2022; Wang et al., 2017). These findings suggest that initial characters hold a processing advantage in two-character word identification and processing, likely due to Chinese orthographic characteristics and left-to-right reading direction. Since visual word processing in Chinese proceeds from left to right, readers process initial characters before final characters, making initial characters critical for lexical identification. In alphabetic reading, the dominant role of initial letter combinations is also constrained by phonological form, as pronunciation is assembled left-to-right from all constituent letters (Milledge et al., 2022). Based on this evidence, researchers have incorporated the differential status of initial and final characters into lexical identification models. For example, the Self-Organizing Lexical Acquisition and Recognition (SOLAR) model posits that letter position activation decreases gradually from left to right across a word (Davis, 2001), and the Sequential Encoding Regulated by Inputs to Oscillations within Letter Units (SERIOL) model similarly proposes that letter

excitatory output forms a gradient that weakens from word-initial to word-final positions (Whitney, 2001). Given the distinct roles of initial and final characters in Chinese two-character word identification, it is necessary to clarify the specific mechanisms by which their positional probability information contributes to incidental word learning.

Two recent studies examining initial and final character positional probability information in Chinese reading have yielded contradictory results. Liang et al. (2023) investigated initial character positional probability in Experiment 1 by manipulating its probability while keeping the final character identical and positionally ambiguous (e.g., "湖水/泉水" [lake water/spring water]). Experiment 2 examined final character positional probability by manipulating its probability while keeping the initial character identical and positionally ambiguous (e.g., "包括/包含" [include/contain]). Results indicated that final character, but not initial character, positional probability information affected word segmentation in Chinese reading. In contrast, Cao et al. (2023) examined the role of both initial and final character positional probabilities in high-frequency (Experiment 2a) and low-frequency word (Experiment 2b) processing. They simultaneously manipulated both factors and found that neither affected high-frequency word processing, whereas initial character, but not final character, positional probability influenced word segmentation during low-frequency word processing. These conflicting results may stem from two primary differences.

First, the core independent variable manipulation differed across studies. Cao et al. employed a classic 2$×$2 factorial design, simultaneously manipulating both initial and final character positional probabilities to create four conditions: high initial-high final (e.g., "遗憾" [regret]), high initial-low final (e.g., "享受" [enjoy]), low initial-high final (e.g., "责任" [responsibility]), and low initial-low final (e.g., "想念" [miss]). In contrast, Liang et al. manipulated initial character positional probability while keeping the final character identical and uninformative (Experiment 1: high initial-identical final [e.g., "湖水"] vs. low initial-identical final [e.g., "泉水"]; Experiment 2: identical initial-high final [e.g., "包括"] vs. identical initial-low final [e.g., "包含"]). This suggests that initial and final character positional probabilities may interactively influence word segmentation. To understand how they jointly contribute to segmentation during incidental word learning, we must first clarify their individual contributions.

Second, the target word frequency ranges differed. Cao et al. used high-frequency words ranging from 46-56 per million and low-frequency words ranging from 1.57-2.37 per million. Liang et al. used target words with an average frequency of 38 per million, falling between Cao et al.'s high- and low-frequency ranges—effectively medium-frequency words. This suggests that the roles of initial and final character positional probabilities may be modulated by word frequency, consistent with Yu et al.'s (2021) Chinese E-Z Reader model, which proposes a familiarity-based segmentation mechanism. According to this model, characters within the perceptual span are activated to varying degrees, and readers determine which characters constitute the next word

based on the familiarity of unrecognized characters and the words they form. Character positional probability, as statistical information about a character's morphological productivity in specific positions, may influence character and word familiarity calculations, thereby affecting segmentation decisions. For example, high-frequency words tend toward whole-word access, rendering both initial and final character positional probabilities irrelevant; medium- and low-frequency words tend toward morpheme-based access, making character positional probabilities functional, though the specific mechanisms require further investigation. Novel words represent an extreme case of low-frequency words that rely heavily on bottom-up morpheme representations. How, then, do initial and final character positional probabilities contribute to word segmentation during incidental word learning? Moreover, given the cumulative nature of incidental word learning (Joseph et al., 2014; Joseph & Nation, 2018; Pagán & Nation, 2019), researchers typically embed novel words across multiple contexts to help readers gradually form lexical representations. During this process, novel words transition from unfamiliar to familiar, effectively shifting from low-frequency to medium- and high-frequency status. Examining whether initial and final character positional probabilities operate similarly during incidental word learning can thus address these questions from a continuous word frequency perspective.

The present study therefore conducted two experiments manipulating initial and final character positional probabilities separately to address two primary questions. First, we examined the independent contributions of initial and final character positional probabilities to word segmentation during incidental word learning. Experiment 1 manipulated initial character positional probability while keeping the final character identical and uninformative; Experiment 2 manipulated final character positional probability while keeping the initial character identical and uninformative. Based on evidence that novel words, like low-frequency words, undergo morpheme-based processing (Coltheart et al., 2001) and that initial characters hold a processing advantage in two-character word recognition (Ma & Li, 2015; Tsang & Zou, 2022; Wang et al., 2017; Yan et al., 2006), we predicted that both initial and final character positional probabilities would influence incidental word learning, with the initial character's effect being stronger. Second, by including learning trials as a continuous variable, we examined how the roles of initial and final character positional probabilities change as novel words become increasingly familiar. Based on the modulatory effect of word frequency on character positional probability processing (Cao et al., 2023), we predicted that the segmentation effects of both initial and final character positional probabilities would gradually diminish with increased learning exposure.

2. Experiment 1: The Role of Initial Character Positional Probability

2.1 Method

2.1.1 Participants Sixty-four Tianjin Normal University students participated in the experiment. All were native Chinese speakers with normal or corrected-to-normal vision and were unaware of the experimental purpose. Participants received monetary compensation after completing the study. Sample size was determined based on Liang et al. (2015, 2017), with an effect size of 0.48 and alpha level of 0.01. G*Power analysis indicated a minimum sample size of 55 participants; our sample of 64 exceeded this requirement.

2.1.2 Design We employed a single-factor two-level (initial character positional probability: high vs. low) within-subjects design. Additionally, learning trials were included as a continuous variable in the model to examine the "familiarity effect" in initial character positional probability processing.

2.1.3 Materials Based on the SUBTLEX-CH corpus (Cai & Brysbaert, 2010), we selected 111 characters as morphemes for constructing novel words. These included 37 characters with high word-initial probability ($ 85 $15%, e.g., "望"). In the high positional probability condition (hereafter "high probability"), novel words combined a high word-initial probability character with an ambiguous character (e.g., "勾席"); in the low positional probability condition ("low probability"), novel words combined a low word-initial probability character with the same ambiguous character (e.g., "望席"). This yielded 37 pseudoword pairs. To ensure these were truly novel, 15 university students who did not participate in the main experiment were asked to write real words corresponding to the pinyin of these items. We selected 14 pairs that no participant correctly identified as target words.

In the high probability condition, initial characters appeared word-initially in two-character words with a mean probability of 93.24% (range: 88.5%-100%); in the low probability condition, this mean was 8.84% (range: 0%-14%). Final characters were identical across conditions, with word-initial and word-final probabilities of approximately 50% (48%-52%). We matched initial characters across conditions on stroke count (high: $M = 6.33$, $SD = 1.91$; low: $M = 7.07$, $SD = 1.94$) and character frequency (high: $M = 444$ per million, $SD = 728$; low: $M = 261$ per million, $SD = 236$). Paired-sample t-tests revealed no significant differences in stroke count or frequency ($ts < 1$, $ps > 0.05$).

Each pseudoword was embedded in six high-constraint sentences describing it as a new member of a familiar semantic category (e.g., animals, plants, jewelry). Each category contained two new members corresponding to high and low initial character positional probability conditions. The experiment comprised 14 semantic categories and 168 contexts. To control for interference between paired words sharing the same final character within a category, we created eight bal-

anced blocks ensuring that participants read paired words with identical final characters across different semantic categories.

All sentences were 16 characters long, with target words positioned mid-sentence. Critically, neither the initial nor final character of target words formed a two-character word with adjacent characters, eliminating potential segmentation ambiguity. Ten university students who did not participate in the main experiment rated sentence naturalness and difficulty on 5-point scales (1 = very unnatural/very easy; 5 = very natural/very difficult). Mean naturalness was 3.93 (SD = 0.76) and mean difficulty was 1.98 (SD = 0.94), indicating that sentences were natural and easy to comprehend.

Within the six contexts for each novel word, 1-2 comprehension questions were randomly presented to ensure participants understood sentence meaning. To avoid influencing character positional probability processing, all comprehension questions appeared only after the third sentence. Additionally, after reading all six contexts, participants completed a semantic category selection task with four options: two from the experimental categories and two fillers. Experimental materials and procedures are illustrated in Table 1 .

2.1.4 Apparatus We used an EyeLink 1000 eye-tracker sampling at 1000 Hz, with a display resolution of $1024\$\times 768 pixels and refresh rate of 120 Hz. Viewing distance was 70 cm. Stimuli were p
point Song font, with each character subtending approximately $0.80°$ of visual angle ($25\times\$25$
pixels).

2.1.5 Procedure Participants were tested individually. After a three-point horizontal calibration with average error $< 0.25°$, participants read instructions and completed practice trials before proceeding to the main experiment. One sentence was presented per screen; participants pressed the spacebar to advance after reading each sentence. Comprehension questions were answered using a mouse left-click. After reading all six contexts for a novel word, participants selected its semantic category using the mouse. The experiment lasted approximately 30 minutes, with breaks provided to reduce fatigue.

2.1.6 Data Analysis Following previous research (Liang et al., 2015, 2017, 2023), we selected early eye-movement measures reflecting lexical identification (first fixation duration, gaze duration) and late measures (regression path duration, total fixation time, regression-out probability, total fixation count). Analyses were conducted using linear mixed models (LMM) and generalized linear mixed models (GLMM) in R (R Development Core Team, 2016) with the lme4 package (Bates et al., 2023). Temporal measures were log-transformed. Initial character positional probability and learning trials (as a continuous variable) were entered as fixed effects; participants and items were included as random effects. We used maximal random effects structures; if the model failed to converge, we gradually reduced complexity until successful convergence.

2.2 Results

Data were excluded based on established criteria (Bai et al., 2019; Liang et al., 2019; Liang et al., 2015, 2017): (1) fixation durations < 80 ms or > 1200 ms; (2) track loss; (3) fewer than three fixations per sentence; (4) outliers beyond 3 SD. Excluded data accounted for 0.3% of total data.

Mean accuracy was 97.77% for comprehension questions and 95.15% for semantic category selection, indicating that participants read sentences carefully and acquired novel word meanings. Fixation patterns for high and low probability conditions are shown in Figure 1 [Figure 1: see original paper], with model results summarized in Table 2 .

In all eye-movement measures, the main effect of learning trials was significant ($|t/z|s > 5.21$, $ps < 0.001$), with fixation durations decreasing and regression probabilities declining as exposure increased, replicating the cumulative nature of incidental word learning.

For first fixation duration, neither the main effect of initial character positional probability nor its interaction with learning trials was significant ($|t|s < 0.62$, $ps > 0.05$), suggesting that readers were not sensitive to initial character positional probability during early lexical processing stages.

For gaze duration, the effect of initial character positional probability was significant ($|t| = 2.11$, $p = 0.03$), with shorter gaze durations in the high than low probability condition, indicating that readers began processing initial character positional probability during relatively early processing stages. The interaction with learning trials was not significant ($t = 1.01$, $p > 0.05$), suggesting that initial character positional probability information exerted a stable influence throughout the entire word learning process.

For regression-out probability, neither the main effect of initial character positional probability nor its interaction with learning trials was significant ($|z|s < 1.61$, $ps > 0.05$). However, for regression path duration, both the main effect and interaction were significant ($|t|s > 2.04$, $ps < 0.05$; interaction shown in Figure 2a [Figure 2: see original paper]). The difference in regression path durations between high and low probability conditions decreased as learning trials increased. This indicates that initial character positional probability did not affect the likelihood of regressing to prior context but influenced the time spent re-reading that context.

For total fixation time, both the main effect of initial character positional probability and learning trials were significant ($|t|s > 4.32$, $ps < 0.001$), but their interaction was not significant ($t = 1.37$, $p > 0.05$). For total fixation count, both the main effect and interaction were significant ($|t|s > 3.97$, $ps < 0.001$; interaction shown in Figure 2b [Figure 2: see original paper]), with the difference in fixation counts between conditions decreasing as learning trials increased.

2.3 Discussion

Experiment 1 examined whether Chinese readers use initial character positional probability information for word segmentation during incidental word learning. The first key finding was significant initial character positional probability effects in relatively late eye-movement measures (gaze duration, regression path duration, total fixation time, and total fixation count): novel words with initial characters that frequently appear word-initially were processed faster than those with initial characters that rarely appear word-initially. This finding aligns with our first prediction and demonstrates that initial character positional probability information contributes to word segmentation during incidental word learning. It contradicts Liang et al. (2023) but converges with Cao et al.'s (2023) findings for low-frequency words. Based on these discrepancies and differences in target word frequency manipulation, we infer that initial character positional probability effects are absent in high-frequency (Cao et al., 2023) and medium-frequency words (Liang et al., 2023) but present in low-frequency words and novel words. We will elaborate on the mechanisms underlying word frequency modulation of character positional probability processing in the General Discussion, integrating findings from both initial and final character manipulations.

The second finding was significant interactions between initial character positional probability and learning trials in regression path duration and total fixation count. The initial character positional probability effect diminished and eventually disappeared as learning trials increased. This result supports our second hypothesis, demonstrating a "familiarity effect" or "learning effect" in initial character positional probability processing. Given the cumulative nature of incidental word learning, this familiarity effect is intrinsically linked to the gradual construction of novel lexical representations and the transition from extreme low-frequency to higher-frequency status. Specifically, before a novel word's first appearance, readers have no lexical representation, making it a true extreme low-frequency word. Upon first encounter, readers must rely on contextual and initial character positional probability information for segmentation and identification, initially constructing orthographic, phonological, and semantic representations. As exposure increases and representations become more established, readers can increasingly rely on stored lexical representations for top-down segmentation, reducing dependence on bottom-up initial character positional probability information, which manifests as the diminishing initial character positional probability effect.

3. Experiment 2: The Role of Final Character Positional Probability

3.1 Method

3.1.1 Participants Sixty-four additional Tianjin Normal University students participated, selected using the same criteria as Experiment 1.

3.1.2 Design We employed a single-factor two-level (final character positional probability: high vs. low) within-subjects design, with learning trials included as a continuous variable to examine familiarity effects in final character positional probability processing.

3.1.3 Materials Based on the SUBTLEX-CH corpus (Cai & Brysbaert, 2010), we selected 132 characters as morphemes, including 44 characters with high word-final probability ($ 85 $15%, e.g., "吊"). Target words were constructed following the same logic as Experiment 1: in the high final character positional probability condition, novel words ended with high word-final probability characters; in the low probability condition, they ended with low word-final probability characters. Initial characters were identical across conditions, with word-initial and word-final probabilities of approximately 50%. To ensure pseudoword status, 15 university students who did not participate in the main experiment were asked to write real words from pinyin. We selected 15 pairs that no participant correctly identified as target words.

Manipulation and matching procedures mirrored Experiment 1, with descriptive statistics shown in Table 3 . Paired-sample t-tests revealed no significant differences in stroke count or frequency for final characters across conditions.

To eliminate potential effects of different sentence frames across experiments, Experiment 2 used identical sentence frames as Experiment 1. Materials and procedures are illustrated in Table 4 .

3.1.4 Apparatus and Procedure Identical to Experiment 1.

3.2 Results

Data exclusion criteria matched Experiment 1, with 0.2% of data excluded. Eye-movement measures and analytic approaches were identical to Experiment 1. Mean accuracy was 97.40% for comprehension questions and 94.79% for semantic category selection, confirming careful reading and successful semantic category acquisition. Fixation patterns are shown in Figure 3 [Figure 3: see original paper], with model results summarized in Table 5 .

In all eye-movement measures, the main effect of learning trials was significant ($|t/z|s > 6.42$, $ps < 0.001$), with fixation durations decreasing and regression probabilities declining as exposure increased, again demonstrating the cumulative nature of incidental word learning.

For first fixation duration, neither the main effect of final character positional probability nor its interaction with learning trials was significant ($|t|s < 0.97$, $ps > 0.05$), indicating that readers were not sensitive to final character positional probability during early processing stages.

For gaze duration, total fixation time, and total fixation count, both the main effects and interactions with learning trials were significant ($|t|s > 2.14$, $ps <$

0.05). Participants showed shorter gaze durations, shorter total fixation times, and fewer fixations in the high than low probability condition, demonstrating significant final character positional probability effects. Interaction analyses (shown in Figures 4a [Figure 4: see original paper], 4b, and 4c) revealed that these effects diminished and eventually disappeared as learning trials increased.

For regression-out probability, neither the main effect nor interaction was significant ($|z|s < 0.75$, $ps > 0.05$). However, for regression path duration, both the main effect and interaction were significant ($|t|s > 2.18$, $ps < 0.05$; interaction shown in Figure 4d [Figure 4: see original paper]). Regression path durations were shorter in the high than low probability condition, and this difference decreased with increasing learning trials. These findings indicate that final character positional probability did not affect the likelihood of regressing to prior context but influenced the time spent re-reading that context.

3.3 Discussion

Experiment 2 examined whether readers use final character positional probability information for word segmentation during incidental word learning. First, similar to Experiment 1, significant final character positional probability effects emerged in relatively late eye-movement measures (gaze duration, regression path duration, total fixation time, and total fixation count): novel words with final characters that frequently appear word-finally were processed faster than those with final characters that rarely appear word-finally. This finding confirms our first hypothesis and aligns with Yen et al. (2012) and Liang et al. (2023), demonstrating that final character positional probability information contributes to word segmentation during incidental word learning.

Second, significant interactions between final character positional probability and learning trials emerged in gaze duration, total fixation time, regression path duration, and total fixation count. The final character positional probability effect diminished and disappeared as learning trials increased, consistent with our second hypothesis and demonstrating a familiarity or learning effect.

Comparing Experiments 1 and 2, both initial and final character positional probability effects appeared in relatively late eye-movement measures (except regression-out probability), indicating that both types of information contribute to word segmentation during incidental word learning in Chinese reading, with similar time courses. This aligns with findings from Thai reading (Kasisopa et al., 2013, 2016), another unspaced script lacking clear visual segmentation cues, where both adult and child readers use initial and final character positional probabilities as statistical segmentation cues to facilitate lexical identification and guide saccades to optimal viewing positions.

Crucially, the time course of these effects differed between initial and final characters. In Experiment 1, the interaction between initial character positional probability and learning trials appeared only in late measures (regression path duration and total fixation count), not in gaze duration, which reflects relatively

early processing. This suggests that during early processing stages, initial character positional probability information exerted a stable influence throughout the entire learning process (from first to sixth exposure), showing no familiarity effect. During later processing stages, however, its influence gradually diminished and disappeared, demonstrating a familiarity effect. In Experiment 2, the interaction between final character positional probability and learning trials began in gaze duration (a relatively early measure) and persisted through regression path duration, total fixation time, and total fixation count (late measures). This indicates that final character positional probability's segmentation role began diminishing during early processing stages as exposure increased. Examination of the interaction patterns reveals that the familiarity effect reflects readers' use of both initial and final character positional probability information during early learning trials for segmentation and identification. As novel lexical representations became established and strengthened, this information ceased to serve a segmentation function in later trials. Thus, initial character positional probability information exerts a longer-lasting and more stable segmentation role than final character information during incidental word learning, providing new evidence for the initial character advantage in two-character word processing (Ma & Li, 2015; Tsang & Zou, 2022; Wang et al., 2017; Yan et al., 2006).

4. General Discussion

Using two parallel experiments that independently manipulated initial and final character positional probabilities, this study investigated how these statistical cues contribute to word segmentation during incidental word learning in Chinese reading. Three main findings emerged: (1) Both initial and final character positional probability information affect word segmentation during incidental word learning. (2) The segmentation roles of both initial and final character positional probabilities show familiarity effects, diminishing and eventually disappearing as learning trials increase. (3) Compared to final characters, initial character positional probability information exerts a longer-lasting and more stable segmentation role. We discuss these findings in relation to the unique characteristics of Chinese text presentation and current Chinese reading models.

4.1 Differential Roles of Initial and Final Character Positional Probabilities

This study demonstrates that both initial and final character positional probabilities contribute to word segmentation during incidental word learning, but their roles change differentially across learning stages. During early learning trials, both cues are effective, producing significant effects in gaze duration, regression path duration, total fixation time, and total fixation count. During later trials, initial character positional probability continues to influence segmentation (evidenced by effects in gaze duration and total fixation time), though its influence

diminishes in regression path duration and total fixation count. In contrast, final character positional probability's segmentation role disappears completely in later trials. These results indicate that the segmentation functions of initial and final character positional probabilities are modulated by learning stage.

According to the mixed representational model for compound words (Caramazza et al., 1988), during early word learning stages, novel words are processed like low-frequency words (stored as morphemes in the mental lexicon), making their identification highly susceptible to morpheme-level representations and thus activating inherent positional probability information to facilitate segmentation. During later learning stages, as novel words transition toward medium- and high-frequency status (stored as whole-word units), whole-word representations increasingly dominate identification while morpheme-level representations decline, reducing sensitivity to character-level positional probability information and diminishing their segmentation roles.

Comprehensive research suggests that acquiring a novel word requires 12-15 exposures during reading (Joseph et al., 2014; Liang et al., 2021; Nation et al., 2007; Tamura et al., 2017). Our novel words appeared only six times, insufficient for forming complete lexical representations and reaching high-frequency status—at best achieving low-medium frequency. At this stage, initial character segmentation effects persisted while final character effects disappeared. This pattern suggests that during Chinese incidental word learning, final character positional probability's segmentation role diminishes first, followed by that of initial characters. If learning continued and lexical representations became fully established, whole-word representations would dominate, replicating Cao et al.'s (2023) finding that both initial and final character positional probability effects disappear. We therefore propose that throughout the entire word acquisition process, the segmentation roles of character positional probabilities change as follows: both cues are effective during early learning; as learning progresses, final character effects diminish first, followed by initial character effects; during later learning stages, both disappear (see Figure 5 [Figure 5: see original paper]).

This differential pattern provides new evidence for the initial character advantage in two-character word identification (Cao et al., 2023; Ma & Li, 2015; Milledge et al., 2022; Tsang & Zou, 2022; Wang et al., 2017; Yan et al., 2006) and supports core assumptions of the SOLAR and SERIOL models developed for alphabetic reading—that letter activation decreases from word-initial to word-final positions (Davis, 2001; Whitney, 2001). Despite Chinese having less variable word length than alphabetic languages and being dominated by two-character words, evidence consistently supports an initial character advantage in Chinese reading. This may arise because: (1) left-to-right reading makes initial characters more important; (2) left-to-right visual processing makes initial character information more accessible; and (3) sequential left-to-right processing of initial and final characters is necessary for phonological assembly. These findings suggest that Chinese reading models should incorporate initial character

processing advantages to enhance explanatory power.

4.2 Implications for Understanding Word Segmentation Mechanisms in Chinese Reading

The absence of interword spaces makes word segmentation mechanisms in Chinese reading particularly complex. Recent Chinese reading models propose that word segmentation and identification are unified processes—when a word is segmented, it is identified (Li & Pollatsek, 2020). Liang et al. (2023) used this framework to explain why final character positional probability, but not initial character probability, affects segmentation. When reading a sentence (e.g., "快乐阅读是我们最美的教育追求" [Happy reading is our most beautiful educational pursuit]), the left boundary of the first word is certain; readers need only use final character positional probability to determine the word's right boundary to complete segmentation. Since adjacent words share boundaries in Chinese, the previous word's final character marks the next word's initial character. By cognitive economy principles, readers need not re-segment word beginnings using initial character positional probability but only need final character information to identify word endings, with the next word's beginning simultaneously identified.

However, Liang et al.'s (2023) account cannot explain our findings. Our discovery that initial and final character positional probability processing is modulated by word frequency has important theoretical implications: initial character positional probability may be invoked depending on lexical processing difficulty, potentially related to verification processes for segmentation accuracy. Although word beginnings are identified when recognizing the previous word, high-frequency words allow more parafoveal preprocessing of upcoming words, yielding higher initial segmentation accuracy and reducing the need for initial character positional probability verification. In contrast, low-frequency or novel words permit less parafoveal preprocessing, resulting in lower initial segmentation accuracy. When verifying segmentation correctness, readers may reactivate initial character positional probability information. Future research should investigate the conditions under which initial character positional probability affects Chinese reading to understand the trade-off between text presentation format and segmentation mechanisms.

Our finding that character positional probability processing is modulated by word frequency provides direct evidence for Yu et al.'s (2021) familiarity-based segmentation algorithm. Character positional probability represents statistical information about a character's morphological productivity in specific positions, likely influencing character and word familiarity calculations. If a character frequently appears word-initially in many high-frequency two-character words, its familiarity is high and segmentation is easier; otherwise, segmentation difficulty increases. Future research should clarify the relationship between character positional probability and character/word familiarity and incorporate this mechanism into models to explain familiarity-based segmentation in Chinese reading.

5. Conclusion

Under the present experimental conditions, we conclude: (1) Both initial and final character positional probability information contribute to word segmentation during incidental word learning in Chinese reading. (2) Compared to final characters, initial character positional probability information exerts a longer-lasting and more stable segmentation role during incidental word learning.

References

Bai, X. J., Liang, F. F., Blythe, H. I., Zang, C. L., Yan, G. L., & Liversedge, S. P. (2013). Interword spacing effects on the acquisition of new vocabulary for readers of Chinese as a second language. *Journal of Research in Reading, 36*(S1), S4–S17.

Bai, X. J., Ma, J., Li, X., Lian, K. Y., Tan, K., Yang, Y., & Liang, F. F. (2019). The efficiency and improvement of novel word's learning in Chinese children with developmental dyslexia during natural reading. *Acta Psychologica Sinica, 51*(4), 471–483.

Bai, X. J., Yan, G. L., Liversedge, S. P., Zang, C. L., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1277−1287.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). *lme4: Linear mixed–effects models using 'Eigen' and S4*. Retrieved July 4, 2023, from https://cran.r-project.org/web/packages/lme4/index.html

Blythe, H. I., Liang, F. F., Zang, C. L., Wang, J. X., Yan, G. L., Bai, X. J., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words: An eye movement study. *Journal of Memory & Language, 67*(2), 241–254.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX–CH: Chinese word and character frequencies based on film subtitles. *PloS One, 5*(6), e10729.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition, 28*(3), 297–332.

Cao, H. B., Lan, Z. B., Gao, F., Yu, H. T., Li, P., & Wang, J. X. (2023). The role of character positional frequency on word recognition during Chinese reading: Lexical decision and eye movements studies. *Acta Psychologica Sinica, 55*(2), 159–176.

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language, 86*, 1–19.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256.

Davis, C. J. (2001). The self–organising lexical acquisition and recognition (SOLAR) model of visual word recognition. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 62*(1–B), 594.

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology, 166*, 190–211.

Joseph, H. S., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133*(1), 238–248.

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.

Kasisopa, B., Reilly, R. G., Luksaneeyanawin, S., & Burnham, D. (2016). Child readers' eye movements in reading Thai. *Vision Research, 123*, 8–19.

Li, X. S., Huang, L. J. Q., Yao, P. P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology, 1*, 133–144.

Li, X. S., & Pollatsek, A. (2020). An integrated model of word processing and eye–movement control during Chinese reading. *Psychological Review, 127*(6), 1139–1162.

Li, X. S., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology, 58*(4), 525–552.

Lian, K. Y., Ma, J., Wei, L., Zhang, S. W., & Bai, X. J. (2021). The role of character positional frequency on college and primary student in oral reading. *Studies of Psychology and Behavior, 19*(2), 179–185.

Liang, F. F., Blythe, H. I., Bai, X. J., Yan, G. L., Li, X., Zang, C. L., & Liversedge, S. P. (2017). The role of character positional frequency on Chinese word learning during natural reading. *PloS One, 12*(11), e0187656.

Liang, F. F., Blythe, H. I., Zang, C. L., Bai, X. J., Yan, G. L., & Liversedge, S. P. (2015). Positional character frequency and word spacing facilitate the acquisition of novel words during Chinese children's reading. *Journal of Cognitive Psychology, 27*(5), 594–608.

Liang, F. F., Gao, Q., Li, X., Wang, Y. S., Bai, X. J., & Liversedge, S. P. (2023). The importance of the positional probability of word final (but not word initial) characters for word segmentation and identification in children and

adults' natural Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(1), 98–115.

Liang, F. F., Ma, J., Bai, X. J., & Liversedge, S. P. (2021). Initial landing position effects on Chinese word learning in children and adults. *Journal of Memory and Language, 116*(1), 104183.

Liang, F. F., Ma, J., Li, X., Lian, K. Y., Tan, K., & Bai, X. J. (2019). Saccadic targeting deficits of Chinese children with developmental dyslexia: Evidence from novel word learning in reading. *Acta Psychologica Sinica, 51*(7), 805–815.

Ma, G. J., & Li, X. S. (2015). How character complexity modulates eye movement control in Chinese reading. *Reading and Writing, 28*(6), 747–761.

Milledge, S. V., Liversedge, S. P., & Blythe, H. I. (2022). The importance of the first letter in children's parafoveal preprocessing in English: Is it phonologically or orthographically driven? *Journal of Experimental Psychology: Human Perception and Performance, 48*(5), 427–442.

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self–teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology, 96*(1), 71–84.

Pagán, A., & Nation, K. (2019). Learning words via reading: Contextual diversity, spacing, and retrieval effects in adults. *Cognitive Science, 43*(1), e12705.

Perea, M., & Acha, J. (2009). Space information is important for reading. *Vision Research, 49*(15), 1994–2000.

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology, 16*(1–2), 3–26.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K. (2009). The thirty fifth Sir Frederick Bartlett Lecture: Eye movements and attention during reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Tamura, N., Castles, A., & Nation, K. (2017). Orthographic learning, fast and slow: Lexical competition effects reveal the time course of word learning in developing readers. *Cognition, 163*, 93–102.

Tsang, Y.-K., & Zou, Y. (2022). An ERP megastudy of Chinese word recognition. *Psychophysiology, 59*(11), e14111.

Wang, W. N., Lu, A. T., He, D. P., Zhang, B., & Zhang, J. X. (2017). ERP evidence for Chinese compound word recognition: Does morpheme work all the

time? *Neuroquantology, 5*(3), 142–152.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: the seriol model and selective literature review. *Psychonomic Bulletin & Review, 8*(2), 221–243.

Yan, G. L., Tian, H. J., Bai, X. J., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology, 97*(2), 259–268.

Yen, M. -H., Radach, R., Tzeng, J. L., & Tsai, J. L. (2012). Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and Writing, 25*(5), 1007–1029.

Yu, L. L., Liu, Y. P., & Reichle, E. D. (2021). A corpus-based versus experimental examination of word- and character-frequency effects in Chinese reading: Theoretical implications for models of reading. *Journal of Experimental Psychology: General, 150*(8), 1612–1641.

Zang, C. L., Liang, F. F., Bai, X. J., Yan, G. L., & Liversedge, S. P. (2013). Inter–word spacing and landing position effects during Chinese reading in children and adults. *Journal of Experimental Psychology Human Perception & Performance, 39*(3), 720–734.

Note: Figure translations are in progress. See original paper for figures.