Behavior Research Methods (2026) 58:51
https://doi.org/10.3758/s13428-025-02935-5

ORIGINAL MANUSCRIPT

A database of overlapping ambiguous strings in Chinese reading

Linjieqiong Huang[1,2] · Chenxi Li[1,2] · Xingshan Li[1,2]

Received: 21 February 2025 / Accepted: 18 December 2025 / Published online: 26 January 2026
© The Psychonomic Society, Inc. 2026

Abstract
In the absence of inter-word spaces, Chinese text sometimes presents word boundary ambiguity. One common case is the *overlapping ambiguous string* (OAS), a three-character string (ABC) where the middle character can form distinct words with both the character to its left (AB) and the character to its right (BC), creating segmentation ambiguity between AB-C and A-BC. This structure makes OASs a valuable tool for investigating the cognitive mechanisms of Chinese word segmentation. We introduce a comprehensive OAS database consisting of 952,497 OASs, each with 43 types of linguistic information at the character, word, and OAS levels. To illustrate how to use the database, we conducted an eye-tracking reading experiment manipulating whether the first character of the OAS (i.e., character A) could stand alone in sentences. Results showed that when character A could not stand alone, readers were more likely to group it with the next character B, leading to an AB-C segmentation. These findings validate the utility of the OAS database in understanding word segmentation during Chinese reading. The potential applications of the database in artificial intelligence, education, and writing system reform are discussed.

Keywords Chinese reading · Word segmentation · Lexical information · Overlapping ambiguous string

As a representative logographic writing system, the Chinese writing system differs from most alphabetic writing systems like English (Li et al., 2022; Perfetti & Harris, 2013), notably in the absence of spaces between words to mark word boundaries. Without inter-word spaces, word boundary ambiguity sometimes arises, with the same character strings segmented in different ways. One type of such ambiguity is known as the *overlapping ambiguous string* (OAS; Gan et al., 1996; Hsu & Huang, 2000a, b; Li et al., 2003; Sun & Zou, 2001). Typically, an OAS consists of three characters (ABC, denoting the characters from left to right), where the middle character can form distinct words with the characters on both its left (word AB) and its right (word BC), creating ambiguity about whether character B belongs to the words AB or BC. Taking the OAS "从小学" as an example, the left two characters form the word AB "从小" (meaning *since childhood*), and the right two form the word BC "小学" (meaning *primary school*). Each OAS can be segmented into an AB-C structure (e.g., "从小-学" meaning *learn since childhood*) or an A-BC structure (e.g., "从-小学" meaning *from primary school*). This is similar to the English ambiguous trimorphemic nonword like *milkteabag*, which can be segmented as *milkteabag* or *milkteabag*. OASs are not uncommon in Chinese texts, with an occurrence probability of approximately 3.6% (Yen et al., 2012). As a vital tool for the study of Chinese word segmentation, OASs enable researchers to explore the specific cognitive mechanisms underlying Chinese reading. However, preparing stimuli for such studies can be challenging. In this study, we present a database containing 952,497 OASs along with their related linguistic information.

The growing body of research utilizing OASs has provided valuable insights into cognitive processes underlying Chinese reading. Some studies have presented participants with isolated OASs, regardless of whether these OASs occur in natural text. For example, in the study by Ma et al. (2014), participants were asked to name the middle character B of OASs. Character B is a polyphone that is pronounced differently when it belongs to AB or BC. Taking the OAS "卫校订" (pronounced as WEIXIAO/JIAODING) for example,

✉ Xingshan Li
lixs@psych.ac.cn

[1] State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

character B "校" is pronounced as XIAO when it belongs to AB, but it is pronounced as JIAO when it belongs to BC. Ma et al. (2014) manipulated the frequency of words AB and BC so that the frequency of one word was higher than that of the other word. They found that character B was more likely to be pronounced as that in the higher-frequency word than that in the lower-frequency word, suggesting the effect of word frequency on word segmentation. Additionally, in recent studies where participants were asked to report the word they identified first when reading isolated OASs, it was found that higher-level factors such as semantic information, emotional valence, and language experience could impact word segmentation (Chen et al., 2024; Huang et al., 2024; Liu et al., 2023). These studies demonstrate that OASs, even when they may not exist in natural text, are useful for studying word segmentation when presented in isolation. Other studies have embedded OASs within sentences to study how Chinese readers segment words in natural reading. For instance, in addition to manipulating the relative word frequency of AB and BC in OASs, Ma et al. (2014) embedded OASs into two types of sentence frames, where the post-OAS context disambiguated the segmentation as either AB-C or A-BC. Results showed that when the frequency of the word BC was higher, readers showed fewer regressions and shorter second-pass reading times on the OAS region when the post-OAS context supported the A-BC segmentation than the AB-C segmentation. The reverse pattern occurred when the frequency of the word AB was higher. On the basis of these results, Ma et al. proposed a two-stage segmentation strategy: readers first rely on local cues like word frequency to favor segmenting higher-frequency words, then use contextual information to confirm or revise initial segmentation when it conflicts with the context. In addition, by manipulating the pre-OAS context, researchers further suggest that words supported by prior context are more likely to be segmented as words (Huang et al., 2021; Huang & Li, 2020, 2024; Zhao et al., 2024). Furthermore, the words positioned on the left (i.e., AB) are more likely to be segmented than the words on the right (i.e., BC) when the other properties are equal, due to the left-to-right reading direction in Chinese (Huang et al., 2021; Huang & Li, 2020, 2024; Ma et al., 2014). Reading direction may cause more visual attention to be initially allocated to the characters on the left than on the right.

The findings on how Chinese readers segment OASs shown in isolation or embedded in sentences, as mentioned above, are suggestive for developing models of Chinese reading. The Chinese Reading Model (CRM; Li & Pollatsek, 2020) provides important insights into how Chinese readers address important challenges such as word segmentation without the aid of inter-word spaces. According to the CRM, all characters within the perceptual span are activated and processed in parallel, along with potential word candidates formed by these characters. A word is simultaneously identified and segmented once its activation surpasses a threshold, and each character is assigned to only one segmented word. In the case of OASs, where both AB and BC are word candidates sharing the character B, the two compete with each other for a single winner. CRM can successfully simulate the initial stage of processing of OASs, including the effects of word frequency and left-side word advantage.

To deepen our understanding of Chinese reading and apply the findings to fields beyond research on language cognition, a comprehensive OAS database is needed. On the one hand, due to the unique nature of OASs as special character strings in Chinese text, manually creating a sufficient number of OASs for experimental stimuli is a complex and time-consuming task. On the other hand, a comprehensive database that includes diverse linguistic information about OASs, such as details about the left-side word AB, right-side word BC, and constituent characters, would provide convenience and greater flexibility for manipulating and controlling various stimulus properties, which is crucial for experimental design and analysis.

In this article, we introduce an OAS database which contains 952,497 OASs and provides detailed information about each OAS, as well as its constituent words and characters. The information covers various attributes: at the OAS level, the database includes the OAS type, the ability to function as a three-character word, word frequency (if available), word class (also known as part of- speech), and its usage in real contexts; at the word level, it provides word frequency and word class; and at the character level, it covers character frequency, stroke count, number of neighbors, word class, the ability to function as a one-character word, and word frequency (when available). Overall, this comprehensive OAS database will assist in preparing stimuli for studies testing theoretical and computational hypotheses in Chinese word segmentation, thus contributing to a deeper understanding of the cognitive mechanisms underlying Chinese reading.

In the following sections, we first provide a detailed overview of the structure and contents of the OAS database. To illustrate how to use it, we then describe an eye-tracking reading experiment designed for investigating whether a character's ability to stand alone as a one-character word in real contexts influences word segmentation. Finally, we provide a general discussion of the potential contributions and applications of the OAS database.

The OAS database

A total of 952,497 OASs were generated for this database. Each OAS is formed by pairing two two-character words, where the last character of the first word matches the first character of the second word. The repeated character is

preserved only once, resulting in a three-character string. These two-character words are listed as words in the *Lexicon of Common Words in Contemporary Chinese* (hereafter LCWCC; Lexicon of Common Words in Contemporary Chinese Research Team, 2008), a standardized and widely used lexical resource published by the Commercial Press. The LCWCC contains 56,008 frequently used words in Modern Chinese, including one-, two-, and multi-character words that appear in contemporary social contexts. Words are organized by ascending frequency grades (without specifying exact frequency counts) and supplemented by phonetic information. Because of its comprehensive coverage and high representativeness, the LCWCC was used in the present study to identify valid words. The database provides 43 different types of information for each OAS. In the following section, we will provide a detailed overview of the information available at the OAS, word, and character levels. A summary of the data is presented in Table 1.

OAS-level information in the database

OAS To construct OASs, we began by selecting all two-character words from the LCWCC. For each two-character word, we then searched the entire list to identity other potential two-character words that could form a pair, where the first character of the second word matched the last character of the first word. Using this method, we constructed OASs by combining these two overlapping two-character words, with only one instance of the repeated character being preserved.

The database included both OASs that exist in natural text and those that do not. As demonstrated in recent studies (Chen et al., 2024; Huang et al., 2024; Liu et al., 2023), isolated OASs that do not appear in natural text can be used as stimuli in studies of Chinese word segmentation and help reveal how the properties of words or characters affect word segmentation. These findings highlight the value of using OASs, even when they may not exist in natural text, for testing specific hypotheses about Chinese word segmentation. By including both naturally occurring and artificially constructed OASs, the database supports a wide range of experimental designs.

OAS_isword OASs were classified into two groups based on whether they can form a valid three-character word listed in the LCWCC. A value of *0* indicates that the OAS does not form a three-character word, while a value of *1* indicates that it does. The ratio of OASs forming a valid three-character word is 0.002, indicating that most OASs are not three-character words, and hence OASs represent a distinct linguistic category different from typical words.

OASpos For OASs that are three-character words, word class annotation was performed using the Python implementation of THULAC (THU Lexical Analyzer for Chinese; Sun et al., 2016), a toolkit that integrates word segmentation and word class annotation. THULAC employs a comprehensive word class annotation set designed for Chinese, which has been widely adopted in natural language processing research due to its high accuracy and fast processing speed. For OASs not recognized by the toolkit, word class annotations were marked "#N/A" to indicate missing or unavailable situations. The main word class annotations are shown in Table 1. The majority of OASs that are three-character words are nouns (61.0%).

OAS_wordfreq If an OAS forms a three-character word, its word frequency was provided. Since the LCWCC does not provide exact word frequency count, word frequency information was retrieved from a lexicon database (Chinese Linguistic Data Consortium, 2003) which is a commonly used resource for word frequency estimates in Chinese psycholinguistic research. Even when an OAS forms a three-character word, it generally has a lower frequency of occurrence (M = 1.30, SD = 5.65 occurrence per million).

OAStype OASs were classified into different types based on whether the first character (A) and the third character (C) can function as independent one-character words in the lexicon. When a character cannot function independently as a word, it must combine with adjacent characters to form a valid word in the sentence contexts. The status of characters A and C was determined based on their presence as one-character words in the LCWCC (as described in the following section, *Character-level information in the database*).

When both characters A and C can function as independent one-character words (i.e., *char1_isword* = 1 and *char3_isword* = 1), the OAS is labeled as type *A+C+*. For instance, in the OAS "花生长", both characters A "花" (meaning *flower*) and C "长" (meaning *grow*) are valid one-character words. Depending on the sentence context, the OAS can be segmented as an AB-C structure (e.g., "花生-长" meaning *peanut grows*) or an A-BC structure (e.g., "花-生长" meaning *flower grows*). This type of OAS may remain ambiguous even in the sentence contexts. Taking the sentence "花生长在院子里" for example, when the OAS "花生长" is segmented as an A-BC structure, the sentence means *flower grows in the yard*, whereas the AB-C structure yields *peanut grows in the yard*; both interpretations are plausible.

When character A cannot be a one-character word but character C can (i.e., *char1_isword* = 0 and *char3_isword* = 1), the OAS is labeled as type *A–C+*. For example, in the OAS "啤酒会", character A "啤" cannot stand alone but character C "会" (meaning *will*) can. In the sentence context, character A should be combined with character B to form an AB-C structure (e.g., 啤酒-会 meaning *beer will*), or combined with the left-adjacent character to form

Table 1 Summary of information by column name and content/description

Level	Column name	Content/description
OAS	OAS	Overlapping ambiguous string
	OAS_isword	Whether OAS is a three-character word; 0 = no, 1 = yes
	OASpos	Word class of OAS if it was a three-character word Annotations: **n** – noun; **np** – proper noun (person name); **ns** – proper noun (place name); **ni** – proper noun (organization name); **nz** – other proper nouns; **m** – numeral; **q** – quantifier; **mq** – numeral + quantifier; **t** – time word; **f** – direction word; **s** – location word; **v** – verb; **a** – adjective; **d** – adverb; **h** – pre-modifier; **k** – post-modifier; **i** – idiom; **j** – abbreviation; **r** – pronoun; **c** – conjunction; **p** – preposition; **u** – auxiliary word; **y** – modal particle; **e** – interjection; **o** – onomatopoeia; **g** – morpheme; **w** – punctuation; **x** – others #N/A = missing or unavailable annotations
	OAS_wordfreq	Word frequency of OAS if OAS is a three-character word
	OAStype	A+C+: both characters A and C can function as one-character words A–C+: character A cannot function as a one-character word but character C can A+C–: character C cannot function as a one-character word but character A can A–C–: neither character A nor character C can function as a one-character word
	CorpusSentence	Example sentence for each OAS in the BCC; NA = OAS not found in this corpus
	CorpusSentenceNum	The total number of sentences containing each OAS in the BCC
Word	word1	The left-side word of OAS (word AB)
	word2	The right-side word of OAS (word BC)
	word1pos	Word class of word AB (annotations defined as in *OASpos*)
	word2pos	Word class of word BC (annotations defined as in *OASpos*)
	word1freq	Word frequency of word AB
	word2freq	Word frequency of word BC
Character	char1	The first character of OAS (character A)
	char1_charfreq	Character frequency of character A
	char1strokes	The stroke count of character A
	char1initneig	The number of neighbors of character A as the first character of a word
	char1endneig	The number of neighbors of character A as the end character of a word
	char1initneigfor2charword	The number of neighbors of character A as the first character of a two-character word
	char1endneigfor2charword	The number of neighbors of character A as the end character of a two-character word
	char1_isword	Whether character A can be a one-character word; 0 = no, 1 = yes
	char1pos	Word class of character A if it is a one-character word (annotations defined as in OASpos)
	char1_wordfreq	Word frequency of character A if it can be a one-character word
	char2	The second character of OAS (character B)
	char2_charfreq	Character frequency of character B
	char2strokes	The stroke count of character B
	char2initneig	The number of neighbors of character B as the first character of a word
	char2endneig	The number of neighbors of character B as the end character of a word
	char2initneigfor2charword	The number of neighbors of character B as the first character of a two-character word
	char2endneigfor2charword	The number of neighbors of character B as the end character of a two-character word
	char2_isword	Whether character B can be a one-character word; 0 = no, 1 = yes
	char2pos	Word class of character B if it is a one-character word (annotations defined as in *OASpos*)
	char2_wordfreq	Word frequency of character B if it can be a one-character word
	char3	The third character of OAS (character C)
	char3_charfreq	Character frequency of character C
	char3strokes	The stroke count of character C
	char3initneig	The number of neighbors of character C as the first character of a word
	char3endneig	The number of neighbors of character C as the end character of a word

Table 1 (continued)

Level	Column name	Content/description
	char3initneigfor2charword	The number of neighbors of character C as the first character of a two-character word
	char3endneigfor2charword	The number of neighbors of character C as the end character of a two-character word
	char3_isword	Whether character C can be a one-character word; 0 = no, 1 = yes
	char3pos	Word class of character C if it is a one-character word (annotations defined as in *OASpos*)
	char3_wordfreq	Word frequency of character C if it can be a one-character word

Note. The unit of frequency is the number of occurrences per million. The BCC refers to the BLCU Corpus Center (see *CorpusSentence* section for details; Xun et al., 2015, 2016)

a word (XA-BC structure; e.g., 扎啤-酒会 meaning *draft beer party*).

When character C cannot be a one-character word but character A can (i.e., *char3_isword* = 0 and *char1_isword* = 1), the OAS is labeled as type *A+C–*. For example, in the OAS "爱国企", character A "爱" (meaning *love*) can stand alone but character C "企" cannot. In the sentence contexts, character C should be combined with character B to form an A-BC structure (e.g., 爱-国企 meaning *love state enterprise*), or combined with the right-adjacent character to form a word (AB-CX structure; e.g., 爱国-企业 meaning *patriotic enterprises*).

Notably, unlike the *A+C+* type, where OASs may remain ambiguous between AB-C and A-BC structures even within a sentence, the *A−C+* and *A+C−* types do not involve ambiguity in the sentence contexts. However, they are still helpful for studies on Chinese reading. As demonstrated in the following section (*Using the OAS database: An experimental investigation*), we manipulated whether character A of the OASs could stand alone (i.e., comparing *A−C+* with *A+C+*) to investigate the role of character-level properties in Chinese word segmentation, as well as the revision process that occurs when the initial segmentation is incorrect. Therefore, while these OASs do not involve real ambiguity, they are valuable for understanding the mechanisms underlying Chinese reading.

Finally, when neither character A nor C can function as a one-character word (i.e., *char1_isword* = 0 and *char3_isword* = 0), the type of OAS is categorized as *A−C−*. For example, in the OAS "承受挫", neither character A "承" nor character C "挫" can function independently. Thus, in the sentence contexts, characters A and C should be grouped with character B to form a word AB or BC, or be grouped with adjacent characters to form a word. Different types of OASs have varying proportions, with the type *A+C+* comprising the majority, approximately 73.1% (see Fig. 1 for full details).

CorpusSentence To indicate whether an OAS occurs in a natural corpus and how it is used in natural contexts, we searched the BLCU Corpus Center (BCC, http://bcc.blcu.edu.cn; Xun et al., 2015, 2016) for each OAS and retrieved its first example sentence. The BCC is a large-scale online corpus of approximately 9.5 billion words, representing contemporary linguistic usage across diverse domains such as newspaper, literature, general texts, classical texts, and dialogue. It features a user-friendly search interface that returns sentence segments containing the search term. Because the LCWCC does not involve sentence contexts or example sentences in natural usage, the BCC was adopted to extract instances of OASs in natural contexts. We used a customized C++ program (developed with VS2010) to automate the search and extract a sentence example for each OAS. If an OAS was found in the BCC, the first complete sentence containing it was returned. If not found, the value was recorded as "NA." The percentage of OASs found in the BCC was 24%.

CorpusSentenceNum To reflect how often each OAS appears in natural contexts, we report the total number of sentences in which each OAS occurred in the BCC. The number of retrieved sentences varied by OAS type. Specifically, the *A+C+* type appeared in the most sentences ($M = 27$, $SD = 717$), followed by *A−C+* type ($M = 13$, $SD = 259$), *A+C−* type ($M = 10$, $SD = 277$), and *A−C−* type ($M = 4$, $SD = 166$). This pattern suggests that OASs with greater segmentation ambiguity occur more frequently in natural language usage. Notably, for OASs where neither character A nor character C can function as a one-character word (i.e., the *A−C−* type), sentences containing such OASs can still exist naturally. In these sentences, if character A does not form a word with character B, it combines with the character to its left; if character C does not form a word with character B, it combines with the character to its right. For example, in the OAS "承受挫", neither character A "承" nor character C "挫" can stand alone. As shown in the column ***CorpusSentence***, this OAS can be embedded in a sentence "学会承受挫折" (meaning *learn to withstand setbacks*), where the OAS is segmented as AB-C structure "承受-挫" and character C "挫" combines with the right-adjacent character to form the word "挫折" (meaning *setbacks*).

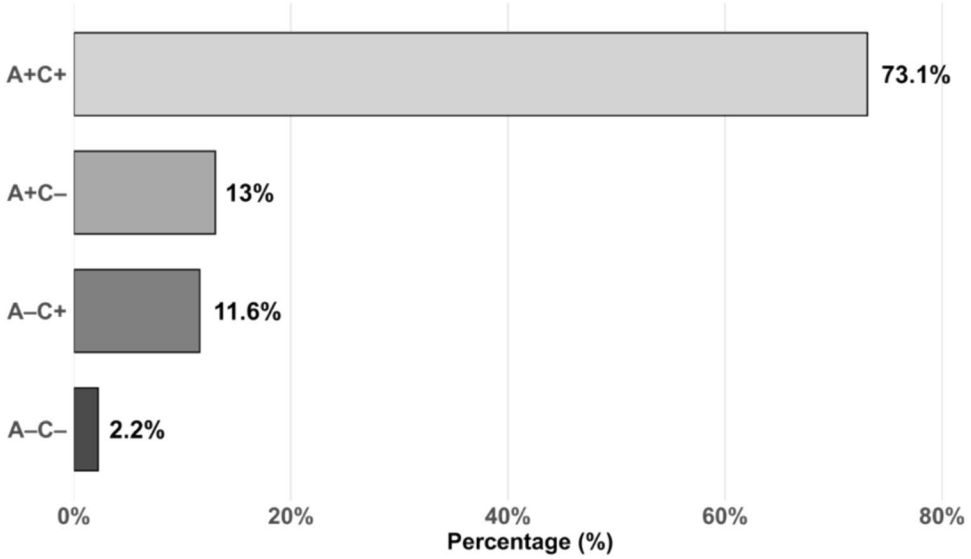

Fig. 1 Proportions of different types of OASs. *Note*. The figure illustrates the proportions of OASs that fall into one of four types: (**1**) *A+C+* indicates that both characters A and C can function as one-character words; (**2**) *A+C−* indicates that character A can function as a one-character word but character C cannot; (**3**) *A−C+* refers to cases where character C can function as a one-character word but character A cannot; (**4**) *A−C−* indicates that neither character A nor character C can function as a one-character word. Percentages are shown at the end of each bar. The figure was generated using the *ggplot2* package in R (Wickham, 2016). As shown, the *A+C+* type comprises the majority, approximately 73.1% of the total; the type *A+C−* represents 13% and the type *A−C+* represents 11.6%; the remaining 2.2% is classified as the *A−C−* type

Word-level information in the database

Word1 For each OAS, this denotes the left-side word (i.e., word AB) composed of the left two characters.

word2 For each OAS, this denotes the right-side word (i.e., word BC) composed of the right two characters.

word1pos/word2pos We used the same method described in ***OASpos*** to annotate word class information for each word AB and word BC. Nouns and verbs are the most frequent word classes: for word AB, 39.5% are nouns and 33.1% verbs; for word BC, 54.4% are nouns and 25.8% verbs.

word1freq/word2freq.[1] Based on the same lexical database (Chinese Linguistic Data Consortium, 2003) described in the ***OAS_wordfreq***, we calculated the word frequency of word AB ($M = 7.49$, $SD = 30.07$) or word BC ($M = 8.08$, $SD = 32.22$), measured in occurrences per million.

Character-level information in the database

Char1/2/3 This denotes the first character (A), second character (B), and third character (C) of the OAS, respectively.

char1_charfreq/char2_charfreq/char3_charfreq We generated a character frequency list providing frequency information for each character. For each character, we identified all words in the word frequency list (Chinese Linguistic Data Consortium, 2003) that contain this character, regardless of whether the character appears as a one-character word or as part of a multi-character word. We then summed the frequencies of all these words to obtain the total frequency of the character. Based on the generated character frequency list, this information provides the frequency of each character, measured in occurrences per million (character A: $M = 963.99$, $SD = 1{,}593.36$; character B: $M = 1{,}766.48$, $SD = 2{,}023.76$; character C: $M = 960.17$, $SD = 1{,}446.97$). In the OAS, the middle character B can serve as the first character of a word and also as the final character of a word, so it appears in more words overall. As a result, under this calculation method, the frequency of character B is higher than that of character A or character C.

[1] Word frequency and contextual diversity are highly correlated lexical variables. Word frequency reflects how often a word appears in a corpus, while contextual diversity refers to the number of distinct contexts in which a word occurs (Jones et al., 2017). Existing resources already provide contextual diversity measures for Chinese words, such as the W-CD column of the SUBTLEX-CH-WF file in the original database of Cai and Brysbaert (2010). Users can retrieve the corresponding contextual diversity values for each word in the present database using Excel functions such as *INDEX* (=INDEX(array, row_num, [column_num])) and *MATCH* (=MATCH(lookup_value, lookup_array, match_type)).

char1strokes/char2strokes/char3strokes This denotes the number of strokes of character A (M = 8.35, SD = 3.28), B (M = 7.09, SD = 3.01), or C (M = 8.39, SD = 3.30), respectively.

char1initneig/char2initneig/char3initneig We calculated the number of words in which character A (M = 83.47, SD = 101.41), B (M = 126.69, SD = 113.05), or C (M = 58.61, SD = 68.66) is the first character. Characters that frequently appear at the beginning of words are more likely to start a word, which may signal the boundary between the preceding character and itself. Therefore, such characters provide cues for identifying potential word boundaries, as they often mark the start of a new word in continuous text.

char1endneig/char2endneig/char3endneig We calculated the number of words in which character A (M = 26.44, SD = 40.65), B (M = 58.37, SD = 57.87), or C (M = 32.36, SD = 43.90) is the last character. Characters that frequently appear at the end of words are more likely to combine with preceding characters to form a word. Thus, these characters help indicate the word boundary between themselves and the subsequent character.

char1initneigfor2charword/char2initneigfor2charword/char3initneigfor2charword We calculated the number of two-character words in which character A (M = 60.30, SD = 81.61), B (M = 149.34, SD = 171.01), or C (M = 101.45, SD = 152.61) is the first character.

char1endneigfor2charword/char2endneigfor2charword/char3endneigfor2charword We calculated the number of two-character words in which character A (M = 39.00, SD = 46.16), B (M = 90.27, SD = 86.28), or C (M = 60.28, SD = 77.50) is the last character.

char1_isword/char2_isword/char3_isword We classified characters A, B, and C based on their presence in the LCWCC to determine whether they can function as one-character words. A value of *0* indicates that the character cannot be a one-character word, while *1* indicates that it can. The probability of each character being a one-character word is as follows: character A (M = 0.86, SD = 0.35), character B (M = 0.96, SD = 0.20), and character C (M = 0.85, SD = 0.36). These probabilities reflect how likely each character is to appear as a standalone word in the Chinese lexicon.

char1pos/char2pos/char3pos For characters that are one-character words, word class annotation was performed using the method described in ***OASpos***. Nouns and verbs are the most frequent word classes: for character A, 32.9% are verbs, 16.1% are nouns, and 14.9% are adjectives; for character B, 28.5% are nouns, 23.8% are verbs, and 11.5% are quantifiers; for character C, 27.7% are nouns, 22.7% are verbs, and 13.3% are morphemes.

char1_wordfreq/char2_wordfreq/char3_wordfreq For characters A, B, and C that are classified as one-character words, we provided their word frequency, measured in occurrences per million. The word frequency for each character is as follows: character B has a mean frequency of 1,229.73 (SD = 1,397.16), character A has a mean frequency of 767.80 (SD = 1,199.81), and character C has a mean frequency of 753.96 (SD = 1,123.18).

As shown in Fig. 2, character B, which can form words both with characters A and C, exhibits distinct properties compared to characters A and C: higher character frequency, higher word frequency (when it can function as a one-character word), fewer strokes, and more neighbors ($ps < .001$). These properties suggest that characters with higher frequency and simpler visual complexity might be more likely to combine with neighboring characters to form meaningful word units, thus demonstrating greater word-forming potential.

Using the OAS database: An experimental investigation

Without low-level visual cues like inter-word spaces, Chinese readers need to rely on other available information and knowledge to aid in word segmentation. Understanding what information Chinese readers employ for word segmentation is a critical issue in the study of Chinese word segmentation. Previous studies have suggested that Chinese readers can use information about the constituent words, such as word frequency (Ma et al., 2014) and emotional valence (Huang et al., 2024). Additionally, information about the constituent characters may also play a role. For example, readers can use within-word position information of characters (i.e., the likelihood of certain characters appearing at the word beginning or word ending) for word segmentation (Liang et al., 2023; Yen et al., 2012). Furthermore, as indicated by the *OAStype* in the OAS database, the ability to be used alone in real contexts is another key property of constituent character that may influence segmentation.

Previous studies using the gaze-contingent boundary paradigm (Rayner, 1975) have shown that whether a character indicates the need for another character to form a word can influence word segmentation, thereby modulating parafoveal processing (Cui et al., 2013; Xie et al., 2025; Zang et al., 2016). In this paradigm, an invisible boundary is placed in a sentence, and the preview text to the right of that boundary changes once the reader's eyes cross it. Using this method, some studies have manipulated the lexical relationship between adjacent characters, and found that parafoveal

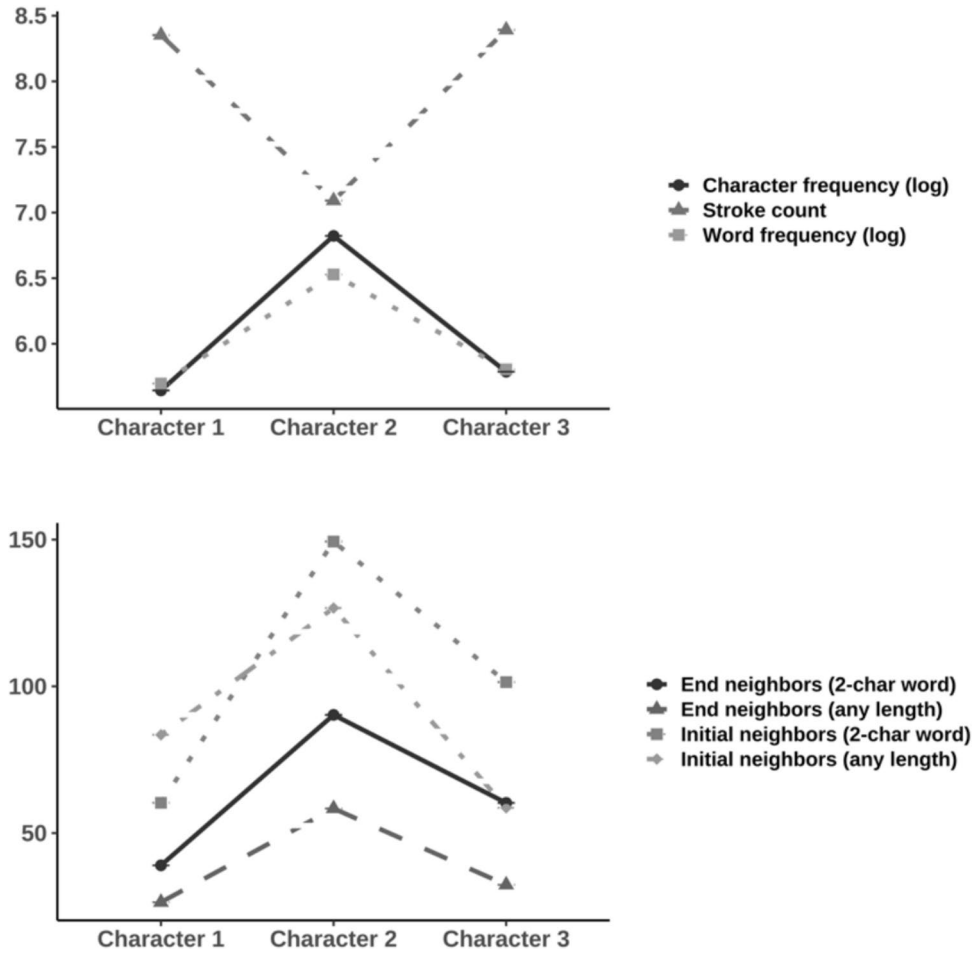

Fig. 2 Properties of characters in the OAS database. *Note*. Line plots show mean values for each measure by character position (characters 1, 2, and 3 correspond to A, B, and C, respectively). The top panel includes log-transformed character frequency, stroke count, and log-transformed word frequency (if a character can be a one-character word). The bottom panel presents neighbor-based metrics, including the number of neighbors for each character in initial or final positions, for both any-length and specifically two-character words. The figure was generated using the *ggplot2* package in R (Wickham, 2016)

processing increases when the character cannot form a word on its own and must combine with the next character. For example, Cui et al. (2013) compared compound words (e.g., "灯塔" meaning *lighthouse*) where each character corresponds to an independent word and they together convey the overall meaning of the whole word, with monomorphemic words (e.g., "玫瑰" meaning *rose*) where the individual characters do not directly reflect the overall meaning of the whole word. Thus, the first character of a monomorphemic word indicates the need to combine with another character to form a word, whereas the first character of a compound word can function as a standalone word. Cui et al. found that parafoveal-on-foveal effects (i.e., a nonsense preview of the second character influenced fixations on the first character) occurred for monomorphemic words, but not for compound words. Similarly, Xie et al. (2025) compared two types of two-character strings: a two-character word (e.g., "推广" meaning *promote*), and two one-character words (e.g., "勇抓" meaning *bravely seize*) comprising two independent one-character words. They found a reduced preview benefit (i.e., less facilitation from seeing a valid preview of the upcoming character before directly fixating it) in the latter case. In addition, Zang et al. (2016) directly manipulated the statistical likelihood of a character functioning as a one-character word. They found that when this likelihood was high, parafoveal processing was reduced; when the likelihood was low, readers engaged in greater parafoveal processing. Together, these findings suggest that when a character indicates a likely word boundary immediately after it, readers tend to treat it as a single word and engage less in parafoveal processing; in contrast, when a character signals the need for another to form a word, readers are more likely to focus on parafoveal information to anticipate the upcoming character. However, the gaze-contingent boundary paradigm relies

on artificial preview changes, which may induce unnatural reading behavior. It remains unclear whether this information affects word segmentation in natural, unaltered reading.

In the present eye-tracking experiment, we employed a natural reading paradigm without character changes to investigate whether the ability of character to function independently in sentences affects word segmentation during Chinese reading. Importantly, the present study considered not only the lexical properties of individual characters, but also how these properties interact with sentence context. We selected pairs of OASs where character A differed in its ability to function as a one-character word in sentences, while characters B and C remained the same. These OAS pairs were embedded in the identical sentence frame, where the correct segmentation was disambiguated as AB-C structure by the post-OAS context, while the prior context was identical and neutral with respect to word segmentation across conditions.

If the information about whether character can be used alone in sentences affects word segmentation, then readers should segment the OAS differently depending on whether character A can function independently in sentences. Specifically, when character A cannot stand alone in real contexts, readers combine it with subsequent character B to form a word, leading to the ultimately correct AB-C segmentation structure. When character A can stand alone, the neutral prior context allows for either the AB-C or A-BC segmentation structure to be contextually plausible. Therefore, regardless of whether character A can stand alone, readers adopt a contextually supported segmentation based on the combination of prior context and OAS, leading to no observable differences in first-pass eye-movement measures. However, when readers encounter post-OAS disambiguating information, those who initially segmented the OAS incorrectly as A-BC segmentation are expected to revise their segmentation. This error detection and correction process should result in longer second-pass reading times on the OAS region when character A can stand alone, compared to when character A cannot stand alone. By contrast, if the information about whether the character can stand alone in real contexts does not affect word segmentation, then there should be no such difference between these two situations.

Method

Participants Forty-eight college students (19 male and 29 female; age range 18–30 years, $M = 24.02$ years, $SD = 2.57$) participated in the eye-tracking experiment. All participants were native Chinese speakers with normal or corrected-to-normal vision and no history of neurological or language disorder. Two participants were excluded due to excessive noisy blinks, leaving a final sample of 46 participants for data analysis.

Power analysis To determine the number of participants required for the eye-tracking experiment, we ran a power analysis using the *powerSim* and *powerCurve* functions from the *simr* package (Green & MacLeod, 2016) in R (R Core Team, 2022). First, we collected pilot data from 10 participants and analyzed it using a *linear mixed-effects model* (LMM). The condition (i.e., whether character A can stand alone in real contexts) was entered as a fixed factor, with participants and items specified as crossed random effects, including intercepts and slopes (Baayen et al., 2008). Based on the prediction, second-pass reading time on the OAS region was used as the dependent variable. Then, based on this LMM, we simulated how the power varies as a function of sample sizes. The results showed that a sample of 40 participants would provide power of 91.60% (95% CI [88.82–93.88%]). To ensure that each condition was properly counterbalanced and there were enough participants after exclusion, we ultimately recruited 48 participants.

Stimuli Based on the column "char1_isword" of the OAS database, we selected 46 OASs where the first character (i.e., character A) could not function as a one-character word (i.e., *char1_isword* = 0; "cannot stand alone" condition). Each item was paired with a counterpart in which the character A could function as a one-character word (i.e., *char1_isword* = 1; "can stand alone" condition), while the word BC remained unchanged. To validate the manipulation of character A, a total of 18 participants who did not participate in the main experiment were presented with these characters A and asked to decide whether each could be used alone in the sentence (*0 = cannot stand alone*; *1 = can stand alone*). As shown in Table 2, the norming results aligned with the OAS database, with characters A that cannot function as standalone words (*char1_isword* = 0) being rated significantly lower in the probability of being perceived as able to stand alone.

For OASs in both conditions, words AB and BC were medium-frequency words, defined as occurring between 1 and 50 times per million. Because words BC were identical across conditions, their frequency and stroke count were the same. In the "cannot stand alone" condition, words AB (word frequency: $M = 1.86$, $SE = 0.12$; stroke count: $M = 15.30$, $SE = 0.58$) and BC (word frequency: $M = 1.82$, $SE = 0.12$; stroke count: $M = 15.20$, $SE = 0.63$) had comparable log-transformed word frequencies, $t(90) = 0.23$, $p = .817$, and stroke counts, $t(90) = 0.13$, $p = .900$. Similarly, in the "can stand alone" condition, words AB (word frequency: $M = 1.77$, $SE = 0.12$; stroke count: $M = 15.78$, $SE = 0.63$) and BC (the same as in the "cannot stand alone" condition) had comparable log-transformed word frequencies, $t(90) =$

Table 2 Properties of stimuli in experiment

	Cannot stand alone	Can stand alone	*t*	*p*
Probability of A being used alone	0.097 (0.01) Range: 0–0.28	0.912 (0.01) Range: 0.72–1.00	−43.46	< .001
Character frequency of character A	296 (55)	391 (48)	−1.31	.195
Stroke count of character A	8.17 (0.39)	8.65 (0.36)	−0.90	.372
Word frequency of word AB	8.03 (1.29)	8.03(1.60)	< 0.001	1.000
Stroke count of word AB	15.30 (0.58)	15.78 (0.63)	−0.56	.578
Bias toward OAS in isolation	0.60 (0.05)	0.67 (0.05)	−1.05	.297
Bias toward OAS with prior context	0.96 (0.02)	0.75 (0.04)	4.82	< .001
Bias toward OAS in the sentence	0.99 (0.01)	0.98 (0.01)	0.71	.482
Plausibility: Prior context + AB	6.02 (0.08) Range: 4.56–6.89	6.17 (0.06) Range: 4.61–6.72	−1.49	.141
Plausibility: The whole sentence	5.61 (0.09) Range: 4.00–6.47	5.59 (0.11) Range: 4.00–6.81	0.14	.887
Predictability of A	0.001 (0.001)	0	1.00	.320
Predictability of B	0.01 (0.01)	0.01 (0.01)	< .001	1.00
Predictability of C	0.01 (0.01)	0.004 (0.004)	0.34	.732
Sentence length	16.48 (0.24)	16.54 (0.24)	−0.20	.846

Note. The bias toward OAS reflects the probability of AB-C segmentation. Standard errors are given in parentheses. Predictability of word AB, BC, and ABC was 0

−0.27, $p = .785$), and stroke counts, $t(90) = 0.66$, $p = .513$. In addition, character A was carefully matched for frequency and stroke count across conditions; the frequency and stroke count of the two-character word AB were also controlled to minimize potential confounding effects (see Table 2 for details). The only difference between the two conditions was whether character A could stand alone in real contexts.

Each OAS pair was embedded in the same sentence frame, with a neutral prior context that was identical across conditions and post-OAS disambiguating information that consistently supported an AB-C segmentation. For example (1a and 1b), in the "cannot stand alone" condition, the OAS "奋力学" includes word AB "奋力" (meaning *to strive*) and word BC "力学" (meaning *mechanics*), whose character A "奋" could not be used alone in real contexts. In contrast, in the "can stand alone" condition, the OAS "卖力学" shares the same word BC, but with the word AB "卖力" (meaning *to exert effort*) whose character A "卖" (meaning *sell*) could be used alone. Importantly, the prior context was designed to be neutral with respect to segmentation. As a result, in the "can stand alone" condition, both AB-C and A-BC segmentations were contextually plausible based on the combination of prior context and the OAS.

(1a). Cannot stand alone
刘涛一直在奋力学课本上的知识
Liu Tao has been **striving to learn** the knowledge from the textbook

(1b). Can stand alone
刘涛一直在卖力学课本上的知识
Liu Tao has been **exerting effort to learn** the knowledge from the textbook

As shown in Table 2, the sentence-level factors were well controlled. To assess sentence plausibility, we recruited two separate groups of 18 participants (who did not take part in the main eye-tracking experiment) for each condition. One group rated the plausibility of the whole sentences, while the other evaluated the plausibility of the sentence segment consisting of the prior context and word AB. No significant differences in plausibility ratings were found between the "cannot stand alone" and "can stand alone" conditions. To assess predictability, another group of 18 participants per condition (also independent of the main experiment) were presented with the sentence segments preceding the OAS region and were asked to write down three characters they predicted would come up next. Predictability was found to be very low across both conditions. Finally, sentence length was comparable between conditions.

Offline word segmentation task To assess the effectiveness of the manipulation regarding whether character A could stand alone, and to explore its potential impact on word segmentation in an offline setting, we conducted three offline word segmentation tasks. Unlike the main eye-tracking experiment, which captured real-time reading behavior, these offline tasks involved explicit judgments made without time pressure. A total of 30 participants, who did not participant in the eye-tracking experiment, took part in each

task. All OASs were divided into two counterbalanced lists. In the first task, participants were presented with isolated OASs and asked to indicate their preferred word segmentation by inserting a slash "/." If they believed the first two characters formed a word, they placed the slash between the second and third characters; if they believed the last two characters formed a word, they placed the slash between the first and second characters. In the second task, participants saw a sentence fragment ending with an OAS (i.e., prior context and an OAS). They were informed that the sentence was incomplete and asked to insert a slash "/" to mark the word boundary according to their preferred interpretation. In the third task, participants were presented with full sentences and instructed to segment them into individual words using slashes "/" based on their understanding of the sentence. Results are shown in Table 2. For the isolated OASs, the probability of AB-C segmentation was comparable between the two conditions. However, when prior context was provided, the probability of AB-C segmentation was significantly higher when character A could not be used alone compared to when it could. When the full sentence was presented, the probability of AB-C segmentation was comparable between the two conditions.

These findings from norming studies provide evidence that the information about whether characters can stand alone in sentences affects offline segmentation outcome in real contexts. Specifically, when presented in isolation, neither the AB-C nor A-BC structure was proven implausible without context, so participants might segment OAS as either AB-C or A-BC structures on the basis of their individual language experience, regardless of whether character A could function as a one-character word. Due to the left-to-right reading direction in Chinese reading, the word on the left side (e.g., AB) has advantages over the word on the right side (e.g., BC). Thus, in both the "cannot stand alone" and "can stand alone" conditions, participants tended to make an AB-C segmentation with a probability generally ≥ 0.6, which is consistent with previous studies (Huang et al., 2021). However, when the OAS was presented with prior context, characters A that could not stand alone needed to be combined with character B to function in the sentence. Thus, compared to when character A could stand alone, participants were more likely to make an AB-C segmentation when character A could not stand alone. This confirms that the effect of character A's standalone status on segmentation occurs only when sentence context is available, supporting the validity of our manipulation. With the full sentence available, participants could correctly segment the OAS as the AB-C structure in both situations. The change in probability of AB-C structure from isolation to having prior context was significant when A could not stand alone (from 0.60 to 0.96; $t(90) = -7.04$, $p < .001$), but not when A could stand alone (from 0.67 to 0.75; $t(90) = -1.36$, $p = .174$).

Apparatus Participants' eye movements were recorded using an EyeLink 1000 eye tracker with a sampling rate of 1,000 Hz. The materials were presented on a 21-inch Sony Multiscan G520 CRT monitor (resolution, 1024 × 768 pixels; refresh rate, 150 Hz) connected to a Dell PC. Each sentence was displayed on a single line in Song 20-point font, and the characters were shown in black (RGB: 0, 0, 0) on a gray background (RGB: 128, 128, 128). A chin rest and forehead rest were employed to minimize head movement throughout the experiment. Participants were seated 58 cm from the computer screen; at this distance, one character subtended a visual angle of approximately 1°. For each participant, the viewing was binocular, but only the right eye was monitored.

Procedure Upon entering the lab, participants were given the experimental instructions and a brief description of the apparatus. The eye tracker was calibrated at the beginning of the experiment and again during the experiment as needed. A three-point calibration and validation procedure was followed, and the maximal error of validation was below 0.5° for the visual angle. Next, each participant read five sentences for practice, followed by 46 experimental sentences and 46 filler sentences in random order. Each sentence appeared after participants fixated on a character-sized box at the location of the first character of each sentence. Participants were asked to read the sentences silently and answer the questions following approximately one third of the sentences. To avoid strategic reading that might interfere with the results while keeping participants engaged in reading for comprehension, comprehension questions were presented only after filler sentences. The filler sentences consisted of common, unambiguous sentences, such as "那座大桥的修建便利了两地的交通" (meaning *The construction of the bridge facilitated transportation between the two locations*), and the length of filler sentences ($M = 15.72$, $SE = 0.35$) was similar to that of sentences containing OASs. Comprehension questions were directly related to the content of the filler sentences and required yes/no responses. After reading each sentence, participants pressed a response button to start the next trial.

Results and discussion

The mean accuracy on the questions following filler sentences was 93.91%, indicating that the participants understood the sentences well. Noisy blinks resulted in the exclusion of 6.29% of the trials. Fixations with durations longer than 1,000 ms or shorter than 80 ms (approximately 1.97%) were also excluded from the analysis. The following eye-movement measures were analyzed:

(1). *first-pass reading time* (the summed duration of all first-pass fixations on the OAS region before moving to another region),
(2). *second-pass reading time* (the summed duration of all fixations on the OAS region following the first-pass reading, including zero durations when the OAS region was not regressed to; see Birch & Rayner, 2010; Clifton et al., 2007; Kim et al., 2018),
(3). *number of fixations* (the number of fixations the OAS region received during first-pass reading, or including any regressions made back to it), and
(4). *total reading time* (the sum of all fixations on the OAS region).

Data were analyzed using LMMs. The condition (i.e., whether character A can stand alone in real contexts) was entered as fixed effects, specifying the participants and items as crossed random effects, including intercepts and slopes (Baayen et al., 2008). Following Barr et al. (2013), we used the maximal model that could converge. We first constructed a model with a maximal random factor structure. When the maximal model failed to converge, we used a zero-correlation parameter model and dropped the random components that generated the smallest variances. The *lmer* function from the *lme4* package (version 1.1–35.5.1.5; Bates et al., 2015) was used. We report the regression coefficients (*b*s, which estimate the effect size), standard errors (*SE*s), *t* values, and corresponding *p* values. We estimated and reported the *p* values for the effects using the *summary* function of the *lmerTest* package (version 3.1-3.1; Kuznetsova et al., 2017). Fixation duration measures were log-transformed, except for second-pass reading time, which included zero time. Detailed eye movement measures and fixed-effects estimates from the LMMs for all measures are shown in Table 3.

The effect of character A's ability to function as a one-character word was not significant for first-pass reading measures, including first-pass reading times and number of first-pass fixations. However, it was significant for the second-pass reading times, total reading times, and total number of fixations. Specifically, when character A could stand alone, readers had longer reading times on the OAS region after reading the post-OAS disambiguating information, compared to when character A could not stand alone.

These results suggest that the information about whether character A can stand alone as an independent word in sentences can be used by Chinese readers in word segmentation. When character A could not stand alone, readers were more likely to combine it with the next character (i.e., B) to form a word in the sentence, leading to an AB-C segmentation of the OAS. In contrast, when character A could stand alone, readers might segment the OASs as either AB-C or A-BC. These findings are consistent with previous studies using the gaze-contingent boundary paradigm, which has shown that when a character requires another character to form a word, it can influence word segmentation, as reflected in increased parafoveal processing (Cui et al., 2013; Xie et al., 2025; Zang et al., 2016). However, the present study extends this literature in two important ways. First, it used a more natural reading paradigm without artificially manipulating parafoveal preview, thereby preserving the ecological validity of the reading process. Second, it examined the word segmentation issue not only in terms of initial segmentation decisions but also in how readers revise them when conflicting information arises later in the text.

The present findings can be accounted for by CRM. Based on CRM, all possible word candidates composed by characters within perceptual span are activated in parallel. Thus, when reading the OAS, characters A, B, and C activate word candidates A, B, C, AB, and BC. Since each character belongs to only one word, word candidates that share the same character compete with each other for segmentation. Thus, the word candidate AB is inhibited by competitors such as candidates BC and A. When character A cannot function as a one-character word in sentences, its corresponding word-level unit (i.e., word A) is unlikely to be activated, reducing the number of competitors in the segmentation process. In this case, the candidate AB only competes with BC, allowing AB to win the competition and

Table 3 Eye movement measures and results of the linear mixed-effects models

Measures	Cannot stand alone	Can stand alone	*b*	*SE*	*t*	*p*
First-pass reading time (ms)	446 (20)	450 (22)	−0.001	0.03	−0.03	.973
Second-pass reading time (ms)[a]	487 (49)	562 (56)	**80.19**	**26.29**	**3.05**	**.002**
Number of fixations (first pass)	1.63 (0.07)	1.63 (0.07)	−0.01	0.05	−0.12	.905
Number of fixations (total)	3.66 (0.23)	3.94 (0.26)	**0.29**	**0.10**	**2.95**	**.003**
Total reading time (ms)	962 (60)	1041 (66)	**0.08**	**0.04**	**2.12**	**.040**

Note. Standard errors are given in parentheses. Significant effects are indicated in bold

[a]To ensure the robustness of our findings, we also analyzed second-pass reading times after excluding zero durations, and the observed effects remained significant (*cannot stand alone* condition: $M = 675$ ms, $SE = 43$; *can stand alone* condition: $M = 765$ ms, $SE = 52$; $b = 0.10$, $SE = 0.04$, $t = 2.42$, $p = .016$)

hence be segmented more easily. By contrast, when character A can function independently in sentences, its word-level unit is also activated. In this situation, the candidate AB has to compete with both candidates BC and A, making it less likely to be segmented. Therefore, compared to when character A can function as a one-character word in sentences, word AB is more likely to win the competition and be segmented when character A cannot. Since the correct segmentation in all stimuli was the AB-C structure, readers were more likely to make a correct segmentation when character A could not be used alone in sentences, leading to fewer revisions and hence shorter second-pass reading times.

The absence of effects during first-pass reading can be attributed to the design of the present study, in which the prior context was designed to be neutral with respect to word segmentation, and disambiguating information only appeared after the OAS. During first-pass reading, the combination of neutral prior context and the OAS allows both AB-C and A-BC segmentations to be contextually plausible. When character A could stand alone, readers might initially segment the OAS as AB-C or A-BC, both of which were supported by the context at that point and hence did not impose processing difficulty. When character A could not stand alone, readers likely used this information to make the correct AB-C segmentation, which was also contextually plausible. Therefore, no significant difference was observed for first-pass reading measures across conditions. It should be noted that this is just speculation regarding the reason for the null effect, and future studies should be conducted to test this possibility.

However, differences emerged during second-pass reading, reflecting readers' error detection and correction processes. Specifically, in the "can stand alone" condition, some readers initially adopted the incorrect A-BC segmentation. Upon encountering the post-OAS disambiguating information, they revised their interpretation, leading to increased second-pass reading times on the OAS region. Conversely, in the "cannot stand alone" condition, readers relied on the lexical property that character A could not function independently, allowing an initial AB-C segmentation that matched later context, and thus eliminating the need for reanalysis. These findings align with the two-stage word segmentation strategy proposed by Ma et al. (2014), which suggests that readers first rely on local lexical cues (e.g., word frequency or character properties) to make an initial segmentation, and then integrate contextual information to confirm or revise it as needed.

Overall, these findings deepen our understanding of the types of information Chinese readers rely on to segment words. In the absence of explicit visual cues like inter-word spaces, Chinese readers actively utilize multiple sources, including character-level properties such as positional probability and the ability to be used alone in sentences (as demonstrated in the present study), word-level features such as frequency and morphological type, and higher-level information like emotional valence and sentence context. These sources of information collectively support efficient and accurate word segmentation during Chinese reading. As demonstrated in the present experiment, the OAS database reported in this study made it easier to find sufficient stimuli for experimental design, providing a valuable tool for future research on Chinese reading.

General discussion

Accurate word segmentation is crucial for reading comprehension in Chinese, where the absence of inter-word spaces can lead to ambiguity in word boundaries. One example of this challenge is the *overlapping ambiguous string* (OAS), a three-character string (ABC) where character B can belong to word AB or BC. Research using OASs, both in isolation and embedded in sentences, has provided valuable insights into the cognitive processes underlying Chinese reading (Chen et al., 2024; Huang et al., 2021; Huang et al., 2024; Huang & Li, 2020, 2024; Liu et al., 2023; Ma et al., 2014; Zhao et al., 2024), highlighting the need for a comprehensive and systematically annotated OAS database. In this study, we introduce an OAS database comprising 952,497 OASs, each with 43 types of linguistic information, including properties of constituent characters, constituent words, and the OAS as a whole. To demonstrate its utility, we conducted an eye-tracking reading experiment. Specifically, we manipulated whether character A in the OAS could be used alone in sentences, and embedded the OAS into a controlled sentence frame, where prior context was neutral regarding segmentation and post-OAS context provided disambiguating information favoring an AB-C structure. Results showed that when character A could stand alone, readers had longer second reading times on the OAS region after reading the post-OAS disambiguating information, compared to when it could not. These findings suggest that Chinese readers use knowledge about a character's lexical status for word segmentation: when a character cannot stand alone, readers are more likely to group it with the next character to form a word; when a character can stand alone, readers may either segment it independently or combine it with the next character. This experiment is suggestive for understanding Chinese word segmentation and also demonstrates how the OAS database is useful in conducting research. Manually creating enough OAS stimuli is challenging. The OAS database provides a comprehensive and systematically organized collection of OASs, allowing for economic, precise, and flexible manipulation of lexical properties. Researchers can use it to address specific research questions. For instance, to investigate how character frequency influences word segmentation,

researchers can use the database to select OASs within targeted character frequency ranges while controlling for other variables. In sum, the OAS database aids researchers in preparing stimuli to test theoretical and computational hypotheses about Chinese word segmentation, thereby deepening our understanding of Chinese reading.

Beyond its applications to Chinese reading, the OAS database can also contribute important knowledge to fields such as artificial intelligence, education, and writing system reform. In artificial intelligence, the rapid development of large language models (LLMs) has prompted studies on whether LLMs perform tasks similarly to humans (e.g., Aher et al., 2023). The OAS database offers a tool for investigating how LLMs handle semantic and structural ambiguities compared to humans (see Liao et al., 2024, for related research using OASs). In education, the strings with ambiguous word boundaries, like OAS, pose challenges for children and L2 learners. The OAS database can help examine which linguistic cues learners rely on during word segmentation, thereby informing the design of textbooks and teaching strategies. In terms of writing system reform, studies have shown that adding inter-word spaces improves Chinese readers' comprehension of ambiguous materials (Hsu & Huang, 2000a, b), and that inserting spaces after words is more effective than before (Liu & Li, 2014). With LLMs enabling easy insertion of spaces, the OAS database offers a valuable resource for identifying optimal spacing strategies and assessing whether such modifications improve reading efficiency for the general public and early-stage learners (see Chen et al., 2020, for related research), thus providing an empirical basis for potential changes to the writing system.

To conclude, the present study introduces a reliable and valid database of Chinese OASs, enabling more targeted and flexible experimental designs for Chinese word segmentation research. The OAS database offers a powerful tool for advancing our understanding of the cognitive mechanisms underlying Chinese reading. Additionally, it has potential applications in fields such as artificial intelligence, education, and writing system reform.

Author note We would like to thank Jiayu Liu for database sorting.

Authors' contributions Linjieqiong Huang: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing–Original draft preparation, Writing–Reviewing and Editing

Chenxi Li: Writing–Reviewing and Editing

Xingshan Li: Conceptualization, Methodology, Formal analysis, Visualization, Supervision, Writing–Reviewing and Editing, Funding acquisition

Funding This research was supported by a grant from the National Natural Science Foundation of China (32371156).

Availability of data and materials The original data and materials are publicly available from https://osf.io/tegwd/?view_only=f92d59ea18254967bb69971dc7e6af7e.

Code availability The R code is publicly available from https://osf.io/tegwd/?view_only=f92d59ea18254967bb69971dc7e6af7e.

Declarations

Conflicts of interest/competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This project was approved by the Institutional Review Board of the Institute of Psychology of Chinese Academy of Sciences.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication The participant has consented to the submission of their data to the journal.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (vol. 202, pp. 337–371). PMLR. https://proceedings.mlr.press/v202/aher23a.html

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Birch, S., & Rayner, K. (2010). Effects of syntactic prominence on eye movements during reading. *Memory & Cognition, 38*(6), 740–752. https://doi.org/10.3758/MC.38.6.740

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS one, 5*(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729

Chen, K., Gu, L., Zuo, H., Bai, Q., & Zhu, Y. (2020, December 4). *The effect of word-word space reading and resolutions on ambiguities of Chinese segmentation by advanced L2 learners of Chinese*. https://doi.org/10.31124/advance.13271798.v1

Chen, R., Huang, L., Perea, M., & Li, X. (2024). The role of semantic information in Chinese word segmentation. *Language, Cognition and Neuroscience, 40*(1), 41–55. https://doi.org/10.1080/23273798.2024.2390003

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.

Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. *Eye Movements*, 341–371. https://doi.org/10.1016/B978-008044980-7/50017-3

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*(2), 403–416. https://doi.org/10.1080/17470218.2012.720265

Gan, K. W., Palmer, M., & Lua, K. T. (1996). A statistically emergent approach for language processing: Application to modeling context effects in ambiguous Chinese word boundary perception. *Computational Linguistics, 22*(4), 531–553. https://doi.org/10.5555/256329.256337

Green, P., & MacLeod, C. J. (2016). Simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hsu, S.-H., & Huang, K.-C. (2000). Effects of word spacing on reading Chinese text from a video display terminal. *Perceptual and Motor Skills, 90*(1), 81–92. https://doi.org/10.2466/pms.2000.90.1.81

Hsu, S.-H., & Huang, K.-C. (2000). Interword spacing in Chinese text layout. *Perceptual and Motor Skills, 91*(2), 355–365. https://doi.org/10.2466/pms.2000.91.2.355

Huang, L., & Li, X. (2020). Early, but not overwhelming: The effect of prior context on segmenting overlapping ambiguous strings when reading Chinese. *The Quarterly Journal of Experimental Psychology, 73*(9), 1382–1395. https://doi.org/10.1177/1747021820926

Huang, L., & Li, X. (2024). The effects of lexical-and sentence-level contextual cues on Chinese word segmentation. *Psychonomic Bulletin & Review, 31*(1), 293–302. https://doi.org/10.3758/s13423-023-02336-9

Huang, L., Staub, A., & Li, X. (2021). Prior context influences lexical competition when segmenting Chinese overlapping ambiguous strings. *Journal of Memory and Language, 118*, Article 104218. https://doi.org/10.1016/j.jml.2021.104218

Huang, L., Zhang, X., & Li, X. (2024). Chinese readers utilize emotion information for word segmentation. *Psychonomic Bulletin & Review, 31*(4), 1548–1557. https://doi.org/10.3758/s13423-023-02436-6

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. *Psychology of Learning and Motivation, 67*, 239–283. https://doi.org/10.1016/bs.plm.2017.03.008

Kim, Y. S. G., Vorstius, C., & Radach, R. (2018). Does online comprehension monitoring make a unique contribution to reading comprehension in beginning readers? Evidence from eye movements. *Scientific Studies of Reading, 22*(5), 367–383. https://doi.org/10.1080/10888438.2018.1457680

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lexicon of Common Words in Contemporary Chinese Research Team. (2008). *Lexicon of common words in contemporary Chinese*. Commercial Press.

Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review, 127*(6), 1139–1162. https://doi.org/10.1037/rev0000248

Li, M., Gao, J., Huang, C., & Li, J. (2003). *Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation*. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* (pp. 1–7). Association for Computational Linguistics. https://doi.org/10.3115/1119250.1119251

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology, 1*(3), 133–144. https://doi.org/10.1038/s44159-022-00022-6

Liang, F., Gao, Q., Li, X., Wang, Y., Bai, X., & 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.

Liao, W., Wang, Z., Shum, K., Chan, A. B, & Hsiao, J. (2024). Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading. *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1961–1967). https://escholarship.org/uc/item/2sm8g139

Liu, P., & Li, X. (2014). Inserting spaces before and after words affects word processing differently in Chinese: Evidence from eye movements. *British Journal of Psychology, 105*(1), 57–68. https://doi.org/10.1111/bjop.12013

Liu, J., Gu, J., Feng, C., Shi, W., Biemann, C., & Li, X. (2023). Cross-modal impact of recent word encountering experience. *Scientific Studies of Reading, 28*(2), 101–119. https://doi.org/10.1080/10888438.2023.2234518

Ma, G., Li, X., & Rayner, K. (2014). Word segmentation of overlapping ambiguous strings during Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1046–1059. https://doi.org/10.1037/a0035389

Perfetti, C. A., & Harris, L. N. (2013). Universal reading processes are modulated by language and writing system. *Language Learning and Development, 9*(4), 296–316. https://doi.org/10.1080/15475441.2013.813828

R Core Team. (2022). *R: A language and environment for statistical computing (Version 4.2.2) [Computer software]*. R Foundation for Statistical Computing. Retrieved March 23, 2021, from https://www.R-project.org

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*, 65–81. https://doi.org/10.1016/0010-0285(75)90005-5

Sun, M., & Zou, J. [孙茂松, & 邹嘉彦]. (2001). Review of Chinese automatic word segmentation research [汉语自动分词研究评述]. *Contemporary Linguistics* [当代语言学], *3*(1), 22–32.

Sun, M., Chen, X., Zhang, K., Guo, Z., & Liu, Z. (2016). *THULAC: An efficient lexical analyzer for Chinese* [Computer software]. Retrieved July 2, 2025, from https://github.com/thunlp/THULAC

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Xie, F., Chen, W., Zhang, L., Cao, X., & Warrington, K. L. (2025). Exploring the role of word segmentation on parafoveal processing during Chinese reading. *Journal of Cognitive Psychology, 37*(1), 1–14. https://doi.org/10.1080/20445911.2024.2429176

Xun, E., Rao, G., Xie, J., & Huang, Z. [荀恩东, 饶高琦, 谢佳莉, & 黄志娥]. (2015). Diachronic retrieval for modern Chinese word: System construction and its application [现代汉语词汇历时检索系统的建设与应用]. *Journal of Chinese Information Processing* [中文信息学报], *29*(3), 169–176.

Xun, E., Rao, G., Xiao, X., & Zang, J. [荀恩东, 饶高琦, 肖晓悦, & 臧娇娇]. (2016). The construction of the BCC Corpus in the age of Big Data [大数据背景下BCC语料库的研制]. *Corpus Linguistics*[语料库语言学], *3*(1), 93–118.

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*(5), 1007–1029. https://doi.org/10.1007/s11145-011-9321-z

Zang, C., Wang, Y., Bai, X., Yan, G., Drieghe, D., & Liversedge, S. P. (2016). The use of probabilistic lexicality cues for word segmentation in Chinese reading. *Quarterly Journal of Experimental Psychology, 69*(3), 548–560. https://doi.org/10.1080/17470218.2015.1061030

Zang, C., Fu, Y., Du, H., Bai, X., Yan, G., & Liversedge, S. P. (2024). Processing multiconstituent units: Preview effects during reading of Chinese words, idioms, and phrases. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 50*(1), 169–188. https://doi.org/10.1037/xlm0001234

Zhao, H., Huang, L., & Li, X. (2024). Readers may not integrate words strictly in the order in which they appear in Chinese reading. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02614-0

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open practices statement The OAS database, materials, data, and analysis code are publicly available at https://osf.io/tegwd/?view_only=f92d59ea18254967bb69971dc7e6af7e. None of the reported studies were preregistered.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.