



University of
Massachusetts
Amherst

Stored Multiword Representations and their Usage during Chinese Reading

Item Type	Dissertation (Open Access)
Authors	Huang, Kuan-Jung
DOI	10.7275/55192
Rights	Attribution-ShareAlike 4.0 International
Download date	2026-03-07 00:44:55
Item License	http://creativecommons.org/licenses/by-sa/4.0/
Link to Item	https://hdl.handle.net/20.500.14394/55192

**Stored Multiword Representations and Their Usage
During Chinese Reading**

A Dissertation Presented

by

KUAN-JUNG HUANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2024

Psychological and Brain Sciences

© Copyright by Kuan-Jung Huang 2024
All Rights Reserved

Stored Multiword Representations and Their Usage During Chinese Reading

A Dissertation Presented

by

Kuan-Jung Huang

Approved as to style and content by:

Adrian Staub, Chair

Charles Clifton, Member

Lori Astheimer, Member

Brian Dillon, Member

Simon P. Liversedge, Member

Maureen Perry-Jenkins, Department Chair

Department of Psychological and Brain Sciences

To my grandpa

ACKNOWLEDGMENTS

I owe lots of thanks to my family, friends, colleagues, and mentors whose support enables me to finish my PhD.

I thank Adrian Staub for his wonderful mentorship. He has never seemed to lose confidence in me. He is always there to provide extremely insightful suggestions but at the same time gives me so much freedom to explore and grow as an independent scholar.

I am also particularly grateful for being able to express my exhaustion from work whenever I feel so, as he constantly checks in. The six years of running experiments, interpreting data, and evaluating theories with him have been a fun and enlightening ride.

I thank Brian Dillon too for his wonderful mentorship. I am so grateful that he *led me down the garden path*. His everlasting positivity and excitement make working with him barely stressful at all. His comments and thoughts on my/our work are so direct, in a very helpful way. Working along with him on computational psycholinguistics is so inspiring.

I thank Charles Clifton, Lori Astheimer, and Simon Liversedge for serving on my committee. I was introduced to Chuck late in my first year, and ever since then he has been so encouraging to me, always asking me to send him my data and manuscripts. He is always the first one to finish reading them and he sees important things I do not see.

Although I only have known Lori and Simon for a year, I cannot thank them enough for reading carefully through my very long manuscript and giving detailed, helpful feedback.

I thank Tal Linzen for hosting me at NYU for a year. The opportunity to work with him is invaluable. I have learned so much from the weekly lab meetings, every of our SAP meetings, and his writing and thinking about computational psycholinguistics.

Then I want to thank all the faculty at UMass CogPsych and the psycholinguistics faculty at UMass Ling for their teaching. Special thanks to Andrew Cohen and Jeffrey Starns who equipped me with the quantitative skills that I am now so proud of. Many times I just dropped by their offices always agreed to help on such short notice. I also thank John Kingston, Shota Momma, and Lyn Frazier for their input on many of my works.

I thank my co-authors who I have learned from about various aspects of psycholinguistics research through collaborations: Xingshan Li and Qiwei Zhang who invited me to their cool work on Chinese compound word processing; Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, and Will Timkey on syntactic ambiguity processing.

Then I thank my friends in the program: Jon and Sandarsh who shared the office with me. They friendly welcomed me to the department. Jon was also an approachable, helpful, and amazing labmate who I feel lucky to have had. And I had so much fun hiking and biking with Sandarsh. Brooke always made sure we hung out and caught up on our lives. I also had many great moments with Andrea, Merika, Patrick, Junha, Sean, John, Jerome, Mar, Anna, Yun, Chung-Yu, Melisa, Trina, Tori, Clara, Fran, Stelios.

I thank friends from Linguistics too for all the fun HSP trips and thought-provoking psycholinguistics workshops: Anissa, Erika, Özge, Suet-Ying, Jelly, Jed, and Maayan.

I whole-heartedly thank my only division-cohort/last-year housemate, Tejas. We went through the incredibly stressful job search process and we dissertated together. His moral support plays a very important role. He is super generous in offering help; I am grateful to have such a great cohort and to have him as a friend.

My three previous housemates, Monique, Can, and Larri, helped me so much outside of school. I specifically thank Monique for being so helpful when I first entered the States.

Three Taiwanese friends at UMass also have helped me a lot since I moved to Amherst. I thank Yueh-Chun, Hsin-Fei, and James for helping me accommodate to the life in the US.

My one-year visit at NYU was brightened with numerous amazing people. I especially thank Jai, Naomi, Anna, Alicia, and Marco for their extraordinary hospitality.

Next, I thank people back from Taiwan: Aleck Shi-Wei Chen dragged me into the psycholinguistics world when I was still a sophomore in college. His wonderful mentorship is what brought me here today, which I forever appreciate. Yun-Jia, Jing-Kai, Yi-Hsin and Hui-Tian listened to my complaint about any sort of things. Yung-Wei made sure we hung out whenever I went back.

I thank my dad and his wife, my brother, my grandparents and my cousins for their unconditional love and support. Janice and Louis visited me at Amherst and hosted me at Leesburg several times, which I so appreciate.

Second-to-last shout out to Grusha, Jane, and Suhas who not only have helped me intellectually but also emotionally. I feel spoiled for all they have done for me. Knowing them is one of the best things in my PhD life. I am very grateful for this friendship.

Lastly, I thank Stephen Bissonnette for supporting me through this tough road. I thank him for being my biggest fan. I thank him for happening to have a very linguistic-y mind and appreciating my research but making sure I also have a life. He is simply the best.

ABSTRACT

Stored Multiword Representations and Their Usage During Chinese Reading

SEPTEMBER 2024

KUAN-JUNG HUANG

B.A., NATIONAL CHENG KUNG UNIVERSITY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Adrian Staub

What are the building blocks of language stored in memory and how are they utilized in linguistic tasks? It has been proposed that meaningful strings of all lengths—morphemes, words, and sequences of multiple words—can be stored, with the last kind playing a crucial role in language processing. This dissertation investigates the existence of stored multiword representations and their usage in Chinese reading. Stored multiword representations are operationalized by using two words that frequently co-occur and comparing them with those that do not. Morphosyntactic structure is also manipulated, with the main comparison between noun-noun and verb-object sequences.

Two tasks are used to study visual recognition of multiword sequences: (1) a rapid masked visual presentation without sentence context probes how many words in a string can be simultaneously recognized and whether this limit is modulated by the co-occurrence frequency of the two words in the string; (2) a naturalistic sentence reading task with a gaze-contingent boundary change paradigm probes how far/deep Chinese

readers process downstream text not yet directly fixated (i.e., parafoveal processing) and whether this limit is modulated by the co-occurrence frequency of the two words in the downstream string.

The results show that co-occurrence frequency facilitates rapid visual recognition without sentence context, making parallel recognition of the two embedded words possible. This is the case for both noun-noun and verb-object sequences. In sentence reading, however, co-occurrence frequency influences online processing differently for strings of different structures. It facilitates processing extremely early on for noun-noun sequences: readers process Characters $n+3$ and $n+4$ in the parafovea beyond the visuo-orthographical level, while no such evidence is found for verb-object sequences.

However, relatively late foveal processing does appear to be facilitated by co-occurrence frequency, for both kinds. Based on the findings, I argue that while language users are highly sensitive to statistical regularities of word sequences of various structures, this possibly yields only familiarity with the surface multiword forms and ease of on-the-fly composition of the two embedded words for verb-object sequences. Compound nouns on the other hand may be lexically stored to have direct form-meaning mapping via frequent exposure, hence the additional early facilitation observed.

Future models of Chinese reading must incorporate mechanisms to explain the current data: (1) extremely fast access to frequently co-occurring strings' orthography, which suggests that a single decomposition route alone is likely insufficient; (2) distinctive processing patterns throughout the time course between noun-noun and verb-object sequences, which points toward structural/semantic composition in addition to utilization of word statistics and contextual probability.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS.....	v
ABSTRACT.....	viii
LIST OF TABLES.....	xiii
LIST OF FIGURES.....	xvi
CHAPTER	
I. INTRODUCTION.....	1
1.1. Background.....	1
1.1.1. Frameworks of visual word recognition in alphabetic systems.....	6
1.1.2. Chinese visual word recognition: data and theories.....	11
1.1.3. More on decomposition, segmentation, and interactive activation.....	21
1.1.4. Multicharacter string as one single word versus multiple words.....	25
1.1.5. The role of statistical regularities in stored, unified linguistic representations.....	30
1.1.6. Summing up and looking ahead.....	34
II. Can more than one word be recognized at a time?.....	37
2.1. Serial versus parallel word recognition.....	37
2.2. Experiment 1: recognition of 2-character noun-noun compounds and verb- object sequences.....	40
2.2.1. Participants.....	44
2.2.2. Materials.....	45
2.2.3. Procedures.....	47
2.2.4. Analysis.....	51
2.2.5. Results.....	55
2.2.6. Discussion.....	59
2.3. Reading strings of three/four Chinese characters.....	63
2.4. Experiment 2: recognition of 4-character noun-noun compounds, verb- object sequences, and subject-verb sequences.....	66
2.4.1. Participants.....	66
2.4.2. Materials.....	67
2.4.3. Procedures.....	71
2.4.4. Analysis.....	74
2.4.5. Results.....	76
2.4.5.1. Duration of pre-cues had little qualitative effect.....	76
2.4.5.2. Main results.....	78
2.4.6. Discussion.....	82
2.5. General Discussion for Experiment 1 and Experiment 2.....	85

2.5.1.1.	Recognition of (extremely) infrequent strings.....	85
2.5.1.2.	Recognition of meaningless/anomalous strings.....	88
2.5.1.3.	Stored representations or on-the-fly composition?.....	89
2.5.1.4.	A few notes on making inferences from the data.....	92
III.	Processing 4-character strings during sentence reading.....	95
3.1.	Segmenting/recognizing meaningful units from unspaced characters in a sentence.....	95
3.1.1.	Chinese readers do not read character by character.....	95
3.1.2.	Word-based reading with Chinese sentences.....	97
3.1.3.	The boundary-change paradigm and parafoveal processing.....	100
3.2.	Experiment 3: Parafoveal processing of 4-character strings of high/low frequency and of different morphosyntactic structures.....	107
3.2.1.	Participants.....	108
3.2.2.	Materials.....	108
3.2.3.	Procedures.....	115
3.2.3.1.	Online norming experiment.....	115
3.2.3.2.	Main eye-tracking experiment.....	116
3.2.4.	Analysis.....	118
3.2.4.1.	Norming.....	118
3.2.4.2.	Eye-movement measures.....	119
3.2.4.3.	Power analysis.....	120
3.2.5.	Results.....	124
3.2.5.1.	Norming results.....	124
3.2.5.2.	Eye-movement results.....	125
3.2.5.2.1.	Pretarget region.....	128
3.2.5.2.2.	First constituent region (C1C2).....	130
3.2.5.2.3.	Second constituent region (C3C4).....	133
3.2.5.2.4.	Whole-string region (C1C2C3C4).....	136
3.2.5.2.5.	Summaries of results of preview validity on eye movement measures.....	138
3.2.5.3.	Supplementary analysis (data with valid preview only).....	139
3.2.6.	Discussion.....	143
3.2.6.1.	Caveats on statistical inferences.....	143
3.2.6.2.	The word n+2 preview effect as a function of 4-character co-occurrence frequency and string structure.....	145
3.2.6.3.	Online sensitivity to morphosyntactic structure of the incoming strings.....	147
3.2.6.4.	Effect of foveal load on parafoveal processing.....	149
3.2.6.5.	String structure and frequency effects during normal sentence reading.....	151
3.2.6.6.	Stored representations or on-the-fly composition? II.....	153
IV.	General Discussion.....	156
4.1.	Findings in a nutshell.....	156
4.2.	The mental dictionary.....	159

4.3.	Chinese reading models.....	166
V.	Conclusion.....	170
5.1.	Limitations, future directions, and broader implications.....	170
5.2.	Concluding remarks.....	174

LIST OF TABLES

1	Summary of frequency and stroke measures for each item group in Experiment 1.....	46
2	Summary of frequency and stroke measures for each item subgroup in NN and VO groups.....	47
3	Summary of Gaussian parameters used to generate simulated data points in Experiment 1.....	54
4	Bayesian logistic regression models of accuracy for NN and VO conditions.....	58
5	Bayesian logistic regression models of accuracy for NN-FCO, NN-BCO, VO-FCO, and VO-BCO conditions.....	59
6	Summary of frequency and stroke measures for each item group in Experiment 2.....	68
7	Summary of frequency and stroke measures for each item subgroup in NN, VO, and SV groups.....	71
8	Bayesian linear regression model of RTs in Experiment 2.....	77
9	Bayesian logistic regression models of accuracy for NN, VO, and SV conditions.....	80
10	Bayesian logistic regression models of accuracy for NN-FCO, NN-BCO, VO-FCO, and VO-BCO conditions.....	82
11	Descriptive statistics for the relevant variables across the six conditions in Experiment 3.....	112
12	Summary of Zang et al. (2024)'s significant preview validity effects.....	120
13	Summary of Gaussian parameters used to generate simulated data points in	

Experiment 3.....	123
14 Descriptive statistics of reading measures on different regions.....	127
15 Bayesian linear regression model of gaze (log-transformed) duration on the pretarget region.....	129
16 Bayesian linear regression models of four reading measures on the C1C2 region.....	133
17 Bayesian linear regression models of four reading measures on the C3C4 region.....	135
18 Bayesian linear regression models of three duration measures on the whole target string region.....	138
19 Standard deviations of the random effects used for power simulations for Experiment 1 and Experiment 2.....	178
20 Bayesian logistic regression models of accuracy for SV conditions in Experiment 2 including both FCO and BCO strings.....	181
21 Comparisons for frequency measures from CCL corpus (2024) and from my own corpus for Experiment 2’s material.....	182
22 Comparisons for frequency measures from CCL corpus (2024) and from my own corpus for Experiment 3’s material.....	182
23 Estimates of fixed effects from the Bayesian model on Zang et al. (2024)’s gaze duration on C1C2 region.....	183
24 Standard deviations of the random effects used for power simulation for Experiment 3.....	185
25 Mean cloze probability of the first two characters and the last two characters,	

	given the context until the pretarget region.....	187
26	Mean conditional probability of C1C2, C3C4, and the whole string given the preceding context of each condition, estimated by the Chinese gpt2-xl (Zhao et al., 2023).....	188
27	Mean number of trials remaining after the trial-exclusion criteria in the six conditions in Experiment 3.....	189
28	Bayesian linear regression models of gaze duration on the whole-string region for the NN and AN groups.....	190

LIST OF FIGURES

1	A visual illustration of three possible scenarios of processing two simultaneously presented stimuli when there is no pre-cue.....	42
2	An AOC plot showing three predictions from three models for accuracy in the NPC condition, based on the accuracy in the PC condition.....	44
3	Illustration of the procedure for a trial in Experiment 1.....	49
4	Distributions of ISI duration for the first run in Experiment 1.....	56
5	AOC plots in Experiment 1.....	57
6	Illustration of the procedure for a trial in Experiment 2.....	72
7	AOC plots in Experiment 2A and 2B.....	77
8	Distributions of ISI duration for the first run in Experiment 2.....	78
9	Accuracy in Experiment 2 based on cueing condition, side, and string structure plotted on an AOC plot.....	79
10	AOC plots in Experiment 2.....	81
11	Posterior distributions of preview benefit in each of the six experimental conditions on the four eye movement measures on the C1C2 region.....	131
12	Posterior distributions of preview benefit in each of the six experimental conditions on the four eye movement measures on the C3C4 region.....	134
13	Posterior distributions of preview benefit in each of the six experimental conditions on the three duration measures on the whole target string region.....	136
14	Posterior distributions of landing position and number of fixations in each of the six experimental conditions, at C1C2, C3C4, and C1C2C3C4.....	142

15 Distributions of naturalness of items (on a scale of 1-7) in the six conditions in
Experiment 3.....186

16 Mean number of trials remaining after the trial-exclusion criteria in Experiment
3.....189

CHAPTER 1

INTRODUCTION

1.1. Background

Since the beginning of psychology as a scientific discipline, one of the central topics of psychology of language has been to understand the mechanisms underlying skilled reading (Cattell, 1886; Morton, 1969). More recently, modern neuroimaging techniques show that reading evokes activation of a specialized area in the brain, the left anterior fusiform gyrus, referred to as the visual word form area (Carreiras et al., 2014; Cohen et al., 2000). This functional specialization in response to an evolutionally recent cultural invention (i.e., writing systems) that allows rapid (within two hundred, or even a hundred, milliseconds) sensitivity to highly abstract visual stimuli in only literate people has continued fascinating cognitive psychologists.

Visual word recognition differs from auditory word recognition in several ways. While listeners passively hear speech unfold incrementally and continually without explicit demarcation between words, readers actively deploy their attention to a small part of the texts that are visually available to them *all at once*, and in most writing systems, words are clearly segmented with blank spaces. Despite these facts, which would suggest that all written letters within a word are processed in parallel to recognize the whole word (Nazir, Jacobs, & O'Regan, 1998; Reicher, 1969), it remains a debate whether a direct mapping between the meaning and the full form is how a word is recognized, especially for morphologically complex words (words that can be broken down into sub-constituents that have internal structures, e.g., *mousetrap*, Andrews,

Miller, & Rayner, 2004; Kuperman et al., 2009; Davis, Libben, & Segalowitz, 2019). That is, it is debated whether the word *mouse* is ever activated or even fully recognized prior to the recognition of *mousetrap*, and in what manner it is. Similarly, when encountering a verb like *cooked*, does one first decompose it into *cook* and *ed* and subsequently combine them together to compute a cooking event in the past (Rastle, Davis, & New, 2004; Solomyak & Marantz, 2010; Taft, 1994)? On the other hand, is it possible that units at an even higher-level (i.e., beyond a single word, e.g., *cook dinner*) are mentally represented and accessible such that more than one word can be recognized together as one unit without first being decomposed or segmented (Cutter, Drieghe, & Liversedge, 2014; Siyanova-Chanturia, Conklin, & van Heuven, 2011; Zang et al., 2024)?

These questions are further complicated when variation in language properties and writing systems is considered. Languages can differ in morphology, ranging from very synthetic (affixing multiple dependent morphemes to a root morpheme to form a word) to very analytic (nearly a one-to-one ratio of morphemes to words). In addition, while most writing systems adopt spacing between words, a few others do not; while most writing systems are alphabetic, some are logographic, and still others are syllabic or featural. These fundamental differences could potentially lead to differences in visual word processing for different language users (Frost, 2012; Li et al., 2022; Öksüz et al., 2024) and therefore warrant special investigation.

In this dissertation, I will study Chinese visual word recognition in three experiments. Chinese is a very analytic language that lacks inflectional morphology but its coinage of words is primarily through compounding (Zhou & Marslen-Wilson, 1995). The Chinese

writing system uses logograms, the visually most salient unit being a *character*, a fixed-size box of a variable number of strokes. The majority of words are two characters long (e.g., 70% in Sun et al.'s large-scale lexicon database¹, 2018). Almost every character has at least one meaning. That is, a character is a morpheme, most of the time. In some other rare cases a character has a meaning only when occurring together with another character (these are called monomorphic two-character words). Furthermore, Chinese characters are written adjacently without explicit spaces. For instance, 我-I 買-BUY 晚-NIGHT 餐-MEAL will be written as 我買晚餐 (I buy dinner). As the example demonstrates, (a) there are no inflectional morphemes (買-BUY is 買 regardless of tense or person or number), (b) one can visually recognize that there are 4 characters even if one has no knowledge of the language itself, (c) each character is small in width such that about four to five characters can easily fit within a reader's perceptual span (Inhoff & Liu, 1998; Yan et al., 2015), and (d) *I buy* (我買) and *dinner* (晚餐) are both visually represented as two adjacent characters but the former, presumably, is conceptually two separate "words" while the latter is one single "word".

This very last property of the Chinese writing system has motivated a great deal of research in the Chinese reading literature (Bai et al., 2008; Inhoff & Wu, 2005, Li, Rayner, & Cave, 2009; Perfetti & Tan, 1999; Xiong et al., 2023; Zang et al., 2016). That is, without visual cues to mark word boundaries, how do Chinese readers process visual

¹ Note that their categorization of words vs nonwords (phrases), while using a rather objective algorithm (Zhang et al., 2003), is still very arbitrary. For instance, *sing-song* is parsed as a word in the corpus but *listen-song* is not a word. Surprisingly *pick-song* is parsed as a word, despite the facts that both *listen* and *choose* unambiguously have the same part-of-speech and that co-occurrence of *pick-song* is actually lower than *listen-song* in a large untagged corpus (861 vs. 5968, PKU CCL corpus http://ccl.pku.edu.cn:8080/ccl_corpus/).

linguistic stimuli? Do Chinese readers process texts in a character-by-character manner? Or can boundaries of meaningful linguistic units somehow be extracted on the fly, and if so, what information is utilized when and how? Here, I will further address these questions with new experimental data, focusing on the role of morphosyntactic structure and word statistics. I will probe the limit of Chinese visual word recognition: can more than one word be recognized at a time as they are written closely without a space in between (cf., potentially only one word can in English, Brothers, Hoversten & Traxler, 2017; White, Palmer, & Boynton, 2018; 2020; White et al., 2019)? I will also probe the limit of parafoveal processing in Chinese sentence reading. Most importantly, how are these two limits modulated by string structure and frequency. The results will have important implications for models and theories of Chinese reading and eye movement control and language processing and learning in general.

The dissertation will unfold as follows. First, I will briefly review frameworks of visual word recognition for alphabetic scripts. Then I will review the literature of Chinese word recognition in isolation and in sentences and how they have shaped approaches to word recognition for Chinese scripts. Then, some elucidation on decomposition, segmentation, and interactive activation, specifically in the context of Chinese sentence reading, will follow. The potential need for segmentation in Chinese sentence reading further raises the question of how a lexical unit is defined in Chinese, which in another two sections I suggest can be addressed from (a) a morphosyntactic viewpoint and (b) a statistics-based viewpoint. These motivate the systematic comparisons of noun-noun and verb-object sequences and of frequent and infrequent sequences throughout the three experiments in the dissertation.

In Chapter 2, the empirical phenomenon of interest in Experiments 1 and 2 is whether two Chinese stimuli can be recognized simultaneously within a very short period of time, using a part-of-speech decision task with rapid parallel presentation. The research questions are whether parallel recognition depends on the morphosyntactic categories of the sequence and on the co-occurrence frequency of the two stimuli in the sequence and whether the two effects interact. The results can be used to infer the existence of stored, unified linguistic representations (that is, one single representation that encompasses both stimuli within the sequence) and the roles of the morphosyntax and frequency in the emergence of the representations.

Experiment 2 is extremely similar to Experiment 1 except for using 4-character-long strings (cf. two-character-long in Experiment 1). When containing a verb and a subject/object, 4-character strings are even more unambiguously two separate words, under certain linguistic theories. The inclusion of such stimuli can provide stronger evidence of stored, unified linguistic representations of multiword strings. To anticipate, it was found that even four-character verb-object or subject-verb sequences could be recognized in parallel, as noun-noun sequences were, but only robustly so when the co-occurrence frequency was high. This suggests that with enough exposure Chinese readers do store multiword representations in their memory.

Given the positive results in Experiments 1 and 2, Experiment 3 (Chapter 3) addresses whether such representations can be accessed very early on during naturalistic sentence reading. I adopted a gaze-contingent boundary-change paradigm to examine to what extent Chinese readers process the next four characters that they have not yet directly fixated. Contrary to the results in Experiments 1 and 2, differences emerged

between noun-noun and verb-object sequences: little lexical information of the last two characters was processed parafoveally for verb-object sequences, but much lexical information of the last two characters were processed parafoveally for noun/adjective-noun sequences. This pattern highlights the role of the linguistic task, whereby sentence reading heavily relies on incremental structure/meaning building.

Chapter 4 summarizes the findings in this dissertation and discusses them all together with respect to language processing (and learning) in general and Chinese sentence reading specifically. Chapter 5 concludes and points out the limitations of the dissertation and future directions.

1.1.1. Frameworks of visual word recognition in alphabetic systems

The earliest models of visual word recognition aimed at explaining the word frequency effect (Broadbent, 1967), the word-superiority effect (Reicher, 1969; Wheeler, 1970), the neighborhood effect (Havens & Foote, 1963), and the letter-transposition effect (Andrews, 1996). Here I will focus on the interactive activation (IA) model by McClelland and Rumelhart (1981) due to its undeniable prominence and its inspiration for a model of Chinese word recognition introduced later (see Norris, 2013, for a review).

The IA model is a localist connectionist model. Three levels of representation in the brain are assumed as three levels of nodes in the model. These are letter features, letters, and words. Each word, letter, or letter feature is represented by one node, and levels of nodes are intra- and inter-connected in a network. Parallel activation of nodes begins from the lowest level—letter features—by external visual stimuli of multiple letters. The activation of letter features feeds forward to the activation of the nodes on the next level—letters—in a cascaded manner: partial activation of letter features can activate

letter nodes that share those features. The activation of letters similarly feeds forward to activation of the nodes on the next level, i.e., word nodes, in a cascaded manner. Different from the letter feature level, the nodes within the letter level connect among one another via inhibitory links such that activation of one letter node will decrease activation of the others. Furthermore, activation of a letter node can also be excited by the upper-level word nodes via top-down feedback. Finally, word nodes are activated by activation of letter nodes and word nodes compete among one another via lateral inhibition. The dynamics among the three levels therefore are interactive and cascaded.

Assuming this simple architecture, the IA model was able to account for several basic findings, including the observation that letter identification is better when presented together with other letters that would make the string a word (or a pronounceable pseudoword) than when presented alone (the word-superiority effect, Reicher, 1969) and the observation that low-frequency words with many neighbors (words that only differ in one letter) are particularly poorly recognized (Broadbent & Gregory, 1968). Despite its great success, the IA model was limited in that it assumed letter-position-specific coding which was challenged by later findings of the transposed-letter effect (e.g., *jugde* is a more facilitative prime than *jupfe* for the target word *judge*, Perea & Lupker, 2003) and position-insensitive orthographic priming effects (e.g., *pivot* can prime *vote* despite *vot* being at different positions, Jordan, 1986). Furthermore, the original IA model's input is restricted to four-letter words. One obvious problem with this restriction is the representativeness of four-letter words for the structure of English vocabulary and its generalizability in word recognition mechanisms: words with more than four letters not only are longer in orthography but also tend to be more complex in morphology (e.g.,

review can be construed as *re-view*).

Online internal morphemic processing has been evidenced in many studies across languages (Bertram, Schreuder, & Baayen, 2000; Duñabeitia, Perea, & Carreiras, 2007; Rastle et al., 2004; Stanners et al., 1979; Taft & Forster, 1975; Taft & Zhu, 1995), and as such has motivated models that directly incorporate representation at the morphological level. For instance, Taft (1994) based his IA model on McClelland and Rumelhart (1981) but further assumed a body-, a morpheme-, and a concept-level of representation. These additional layers help explain that (a) it takes longer to decide that a string that is frequently the stem of affixed words is a word (e.g., **vent** in *advent*, *invent*, *prevent*, etc) or a nonword (e.g., **vive** in *revive* and *survive*) than strings that have similar surface frequency but do not usually serve as a stem (e.g., **coin** as a word or **nace** as a nonword), (b) lexical decision times are shorter for affixed words (e.g., **reproach**) whose stem (*proach*) is also in a more common word (*approach*) than that for their surface-frequency-matched counterparts (e.g., **dissuade**) whose stem is in a less common word (*persuade*), and (c) words that have a common affix but are not in fact affixed words (i.e., pseudo-affixed words such as *regatta* which does not connote any meaning of “again”) take longer to process than their surface-frequency-matched counterparts like *graffiti*. The crucial idea is that different morphemes have different strengths and activation of morpheme nodes can further activate word and concept nodes which provide facilitation sometimes and inhibition other times. For example, the presentation of *vent* does not only activate the visuo-orthographic nodes *V-E-N-T* and the word node of VENT but also VENT as a morpheme which will further activate other word nodes like INVENT and PREVENT; these activated word nodes compete with the word node VENT and therefore

increase decision times. Similarly, the presentation of *regatta* easily activates the high-frequency prefix *re-* which in turn activates the concept node of “again” which competes with the concept nodes of “boat” and “race”.

Note that while Taft’s interactive model (1994) assumed a morpheme layer, it did not treat decomposition (i.e., stripping *pre* from *prevent* to separately process *pre*) as a discrete stage but as an integral part of the access process for the whole word (see also Stevens & Plaut, 2022, for discussion). Stronger versions that assume an obligatory process of decomposing/segmenting morphemes have also been proposed (Crepaldi et al., 2010; Rastle et al., 2004; Taft, 2004). Notably, these frameworks posit that morpho-orthographic decomposition is extremely early such that it is blind to semantic information. For instance, pseudo-affixed words like *corner* could subliminally prime *corn* as effectively as *cleaner* primed *clean*. This suggests a purely structural decomposition: even though *corn* is unrelated to *corner* semantically, the *corner* prime will still be segmented first into *corn* and *er*. In addition, for both *corner-corn* and *cleaner-clean*, the priming effect was larger than the priming effect of *brothel* on *broth* (whereby *el* is not an affix), which suggests priming is not just visuo-orthographic but also morphemic.

Early morphological decomposition stands in contrast to the augmented addressed morphology model (Caramazza, Laudanna, & Romani, 1988) or the supralexical model (Giraud & Grainger, 2000) that posit that the very initial stage of processing a polymorphemic word is activation of the whole word at the form level. Access to relevant morphemic representation is only available after activation of the full form representation. This approach was motivated by the findings of a robust whole-

word/surface frequency effect on reading times of polymorphemic words (Burani & Caramazza, 1987; Schreuder & Baayen, 1995). A more eclectic approach is a parallel dual-route model (Baayen, Dijkstra, & Schreuder, 1997; Bertram et al., 2000; Schreuder & Baayen, 1995) whereby both the constituents (affixes and stems) and the full form are accessed simultaneously as soon as the word is encountered for all words. If the activation process via the full-form route finishes earlier than the activation process via the constituents, recognition of the whole word might not necessitate recognition of the parts. This view was supported by the findings that lexical decision times were influenced by the surface frequency for some polymorphemic words but were influenced by the base frequency of the stem for other polymorphemic words (Bertram et al., 2000). To date, the decomposition approach appears to have been the dominant view in the literature of visual word recognition (Gaston et al., 2021; Stevens & Plaut, 2022) given the overwhelming evidence of subliminal morpheme-based priming (Ciaccio, Kgolo, & Clahsen, 2020; Rastle et al., 2004) and neural activity to isolated word in magnetoencephalography (MEG) studies (Hsu, Pylkkänen & Lee, 2019; Lewis, Solomyak, & Marantz, 2011; Wray et al., 2022). However, an auxiliary parallel full-form route (e.g., Kuperman et al., 2009; Schmidtke & Kuperman, 2019) can easily help explain data involving extremely high-frequency complex words (Baayen, Wurm, & Aycok, 2007) or semantically opaque words (Jared, Jouravlev, & Joanisse, 2017) without compromising the assumption that all words go through early morphological decomposition. This more flexible framework also appears to be more efficient in that it assumes that all information (about constituent morphemes, whole word, morphological families) available for word recognition is utilized (Libben, 2006; Kuperman et al.,

2009).

1.1.2. Chinese visual word recognition: data and theories

To recap, the prominent view on visual word recognition for alphabetic scripts is that decomposition of words into stems and affixes (real or pseudo) occurs early and automatically for all words, but meanwhile words can also be recognized via a full form if the mapping between the form and lemma is strong enough to outrun recombination of decomposed constituents or if the meaning of the whole word is not transparently a combination of its constituents. However, in Chinese visual word recognition, it is reasonable to suspect that the extent to which each route is utilized might be different. For instance, it is possible that Chinese readers predominantly rely on the decomposition route because of the following reasons. First, a character is visually salient and almost every character can be associated with at least one meaning whether it is bound or free; hence a morpheme is visually salient. Furthermore, characters are written without spaces between words, and 70% of words are one-character in terms of token frequency² (Chinese Linguistic Data Consortium, 2003); decomposing a string into single characters/morphemes and processing each of them separately by default might be a smart heuristic strategy to avoid the need for more sophisticated word segmentation.

Several studies have revealed evidence of decomposition during Chinese visual word recognition. Zhang and Peng (1992), in one lexical decision task, found that when whole word frequency, family size (how many different characters can a character form a word together with), and number of strokes were experimentally controlled, the higher the frequency of the first/second character of a two-character coordinative compound was,

² *Token* frequency weighs each word by how frequently it appears, contrasting with *type* frequency which simply counts every unique word as 1.

the quicker it was for it to be categorized as a real word.

Zhou et al. (1999), using a series of primed lexical decision tasks, showed clear morphemic priming effects above and beyond orthographic and phonological priming effects. In each trial, a two-character target word was immediately preceded by a two-character prime (57 ms) in turn immediately preceded by a pattern mask (300 ms, masked priming). When the prime had, either at the first- or second-character position, the same morpheme (the same character denoting the same meaning, e.g., 華貴 splendid-expensive = LUXURIOUS as the target and 華麗 splendid-beautiful = MAGNIFICENT as the prime), response times to the target word were faster than when the prime had the same character denoting a different meaning (華僑 Chinese-emigrant) or when the prime had a different character with the same pronunciation (滑翔 slide-fly = GLIDE, both 華 and 滑 are pronounced as *hua*).

More recently, an MEG study of a lexical decision task (Hsu, et al., 2019) also suggests decomposition during visual word recognition. Four types of two-character words were adopted: (a) monomorphemic words (螳螂, MANTIS, each of the characters has no meaning by itself), (b) noun-noun compounds (花草, flower-grass = PLANT), (c) modifier-head compounds (汽車, gas-car = AUTOMOBILE), and (d) verb-object compounds (開車, drive-car). Compounds evoked larger brain activity in left anterior temporal cortex than monomorphemic words (i.e., (a) vs. (b)(c)(d)) as early as 200 ms (also Wei et al., 2023), and lexical decision was faster for compounds than monomorphemic words. The two findings suggest that morphological processing (decomposition) may be prelexical (as evidenced in early brain activity) and non-costly

(as evidenced in faster response time). Furthermore, brain activity in left posterior middle temporal gyrus differed during the 250-350 ms window between (b) and (c)(d), indicating the internal relation between morphemes within a compound word is assessed, which necessitates initial decomposition. In short, morphological decomposition has been demonstrated in Chinese studies by the constituent frequency effect, the morphemic-priming effect, and the morphological complexity effect in lexical decision tasks.

As in the literature on alphabetic visual word processing, early whole-word processing has also been evidenced, however, on top of the evidence of decomposition introduced above. First, in another experiment by Zhang and Peng (1992), they found that when words were controlled for first- and second-character frequency, family size, and number of strokes but differed in their whole word frequency, high-frequency two-character words were responded to faster than low-frequency counterparts. Secondly, in Mok (2009), the word-superiority effect was found to be larger in high-surface-frequency bimorphemic compound words than in low-surface-frequency ones and larger in semantically opaque compound words (horse-up = IMMEDIATELY) than in transparent ones (hand-gun = PISTOL). This suggests whole-word representation does influence online word recognition, and that high-surface-frequency compound words and semantically opaque compound words have stronger whole-word representation.

Note that these two pieces of evidence, as discussed in Section 1.1.1., are still compatible with a model with decomposition (e.g., Kuperman, et al., 2009). Arguably stronger evidence for exclusive whole-word processing came from some eye-tracking-during-Chinese-sentence-reading studies (Li et al., 2014; Ma, Li, & Rayner, 2015; Yang et al., 2012). In Li et al., (2014), eye-tracking data from reading non-manipulated corpus

texts were explained by multiple variables which included whole-word frequency. However, entering character frequency in the regression model did not significantly increase the variance explained for word reading times (in fact, it decreased the model fit). This finding paralleled an experimental eye-tracking study where first-character frequency and whole-word frequency were orthogonally manipulated (Ma et al., 2015). There, Ma et al. found that while high-whole-word-frequency two-character words were read significantly faster than their low-whole-word-frequency counterparts, a word with a high-frequency first character was *not* read significantly faster than a word with a low-frequency first character (when the second character was held the same and the whole-word-frequency was matched). Findings from these two studies suggested that character representation (similar to morphemic representation for Chinese) might not play a functional role in word recognition during sentence reading.

In addition to Ma et al. (2015) and Li et al. (2014), Yang et al. (2012) tested whether the embedded character is ever semantically activated or recognized when processing the whole word by looking at the effect of contextual plausibility. Sentences like (1a-1d) were used.

(1a) 围观的人看着他踢打门卫却无动于衷 (KICK DOORKEEPER)

(1b) 围观的人看着他哀求门卫却无动于衷 (ENTREAT DOORKEEPER)

(1c) 围观的人看着他踢打门却无动于衷 (KICK DOOR)

(1d) 围观的人看着他哀求门却无动于衷 (*ENTREAT DOOR)

(People watched him kick/entreat the doorkeeper/door but did nothing about it.)

In sentences (1a) and (1c) there is no implausibility either locally or globally. On the other hand, in sentence (1b) there is local implausibility if the first character DOOR is processed incrementally. Finally in sentence (1d) there is an outright implausibility. Based on previous findings, semantic implausibility can have its influence on the very first fixation, causing reading time to be longer (Rayner et al., 2004; Staub et al., 2007). Yang et al. found that the implausibility effect occurred as early as on the skipping rate, first fixation duration, and gaze duration, but only for condition (1d)—the one-character outright implausibility condition. For (1b) where the implausibility is only local, there was no effect on any measures on the target region (first-character, second-character, or both characters together). These were also true for later processing (go-past reading times and regress-in rates). Based on these findings, it was concluded that Chinese sentences are processed on a word-by-word, not character-by-character, basis.

In short, for isolated word recognition, evidence for both decomposition (morphological processing, Hsu et al., 2019; Zhang & Peng, 1992; Zhou et al., 1999) and whole-word access (Mok, 2009; Zhang & Peng, 1992) has been found. This pattern is by and large similar to that of isolated visual word recognition in alphabetic scripts and can be explained by interactive activation models (Taft, 1994; Li et al., 2009) or parallel multiple-route models (Kuperman et al., 2009). For sentence processing, earlier studies failed to find semantic or frequency effects from the first character embedded in a two-character word (Li et al., 2014; Ma et al., 2015; Yang et al., 2012). The next paragraphs, however, will introduce more recent studies that have provided evidence of character effects that suggests morphological processing even during sentence reading in Chinese.

One aspect overlooked in the earlier literature of Chinese sentence processing is the

interaction between whole-word frequency and embedded-character frequency. Recall that a very high-frequency polymorphemic word might form one single, directly-accessible lexical unit after having left a memory trace multiple times (Baayen et al., 2007; Kuperman et al., 2009), while infrequent compound words possibly can only be accessed by first being decomposed. Cui et al. (2021) explicitly tested this hypothesis by examining the embedded-character-frequency effect separately with high-surface-frequency compounds and with low-surface-frequency compounds. Their results showed exactly this pattern: there was no effect of embedded character frequency (neither first- or second-character) obtained for the former group but an effect of the first embedded character frequency for the latter group on the whole-word reading times as early as the first-fixation and gaze durations.

Intriguingly, the first-embedded-character-frequency effect observed in the low-frequency compounds in Cui et al. (2021) was an inhibitory one (i.e., the higher the character frequency, the *slower* the reading times). This inverse character frequency effect is opposite to that in isolated word recognition (Zhang & Peng, 1992 for Chinese; Taft, 1994 for English; Kuperman et al., 2009 for Finnish). Cui et al. (2021) interpreted this inverse character frequency effect to reflect, in essence, a morphological family size effect, which in turn was explained under the constraint hypothesis (Hyönä, Bertram, & Pollatsek, 2004). That is, characters that have higher character frequency tend to appear within many various multicharacter words (e.g., WATER in WATER-GUN, WATER-DISASTER, WATER-POWER, WATER-TUBE and so on). A high-character-frequency character thus is less informative in determining the identity of the whole word, while a low-character-frequency character—since it tends to be able to form only a few words—

is highly constraining about the upcoming character and thus informative about the identity of the whole word. Indeed, Cui et al.'s post-hoc analysis showed that the variable of morphological family size of the first character trumped the variable of the first character frequency. The inhibitory first-character frequency effect has also been reported in two other recent studies (Xiong et al., 2023; Yu, Liu, & Reichle., 2021). Xiong et al. (2023) also found the interaction between whole word frequency and first-character frequency, as in Cui et al. (2021), such that only for low-frequency words did the first-character frequency matter in processing time of the whole string.

Zhang, Huang, and Li (2024) provided an alternative view to the constraint hypothesis (Cui et al. 2021) in explaining the inhibitory character frequency effect. Instead of attributing it to morphological family size, they proposed that there are two types of frequency associated with the embedded constituents of a multicharacter words: one frequency is the surface frequency of the character (i.e., how many times the character appears in a corpus, whether it is embedded in another word or it is by itself a word), and the other frequency is the word frequency of the character as a single-character word. This approach aligned with the Chinese interactive activation model (Li et al., 2009; also Li & Pollatsek, 2020) which assumes three layers of nodes/representations: visual features, characters, and words, with the only excitatory feedback links between the word and character layer, and inhibitory intra-layer links among characters at the same position and among word nodes across overlapping positions, and three facilitatory feedforward links across all layers. Activations of nodes have different baselines based on a reader's prior experience (i.e., the frequency with which they are encountered), and recognition of a word is based upon the activation of a

word node reaching a certain threshold. Under this model, an intuitive prediction is that the higher frequency a character is (surface frequency), the faster a word containing it will be recognized due to the excitatory link between the character and the word. A less obvious prediction is that the higher the frequency of a character appearing as one independent word, the *slower* a word containing it will be recognized due to the intra-layer inhibitory connections among word nodes. That is, the single-character word nodes will compete more strongly with the presented multicharacter word node. Because these two variables (surface character frequency and word frequency of the character) tend to correlate strongly but have opposite effects, they might very often cancel each other out if the experimenters have not considered both variables, which can explain the elusiveness of the findings of a character frequency effect. Also note that this model predicts a facilitatory whole-word-frequency effect as well, and it simulates that for two-character words, the two-character word node almost always will win the competition over the embedded-one-character word nodes to eventually be recognized, thanks to its bottom-up feedforward support from *both* character nodes.

To demonstrate this, Zhang et al. (2024) conducted two studies of isolated word recognition. The first study was a reanalysis of two mega-lexical-decision-studies (10,022 two-character words read by 504 participants, Tsang, et al., 2018, and 25,286 two-character words read by 594 participants, Tse, et al., 2017). In both analyses, when both types of frequency associated with the first and second embedded characters were entered into the regression model along with the whole-word frequency and visual factors (the number of strokes for each character), both types of frequency had a significant effect on the response times. The character frequency was facilitatory while the word frequency of

a character was inhibitory. This was true for the first character in both studies; the effects from the second character were less clear in Tsang et al. (2018). The second study was a direct investigation of the effect of word frequency of the first embedded character on lexical decision times of two-character words. The experiment was a 2 (word frequency of the first embedded character, high vs. low) \times 2 (whole-word frequency, high vs. low) design, while character frequency, number of strokes, and morphological family size were carefully controlled for. Here, it was found that whole-word frequency facilitates response times while word frequency of the first character *inhibits* response times, with no evidence of interaction between the two. This offers a potential alternative interpretation to the morphological family size in explaining an inhibitory effect of character frequency in sentence reading (Cui et al., 2021); that is, it is a *word* frequency effect in disguise. Finally, it is worth noting that while an inhibitory word frequency effect can be experimentally shown, it was smaller compared to the whole-word frequency effect, in both their corpus study (Study 1) and experimental study (Study 2). This pattern is consistent with, as mentioned in the last paragraph, Li and Pollatsek (2020)'s model which simulates that the two-character word node almost always wins the recognition competition over the embedded single-character word nodes. Consistent with this simulation, in another very recent eye-tracking-during-sentence-reading study, Hyönä et al. (2024) examined the first-character *word* frequency effect with infrequent two-character compounds and failed to find a first-character word frequency effect on gaze durations on the whole-word region. While they found the effect on gaze durations on the first-character region, this only occurred inconsistently in one experiment but not another. The authors concluded that first-characters only played a fleeting role in Chinese

compound word recognition.

To sum up, in the literature on visual processing (or reading) of Chinese texts, evidence for processing of both the whole word and the embedded characters has been reported. However, effects of whole-word properties appear to be more robust and/or much larger, with effects of subparts' properties mostly found only in paradigms using electrophysiological measures without sentence contexts (Hsu et al., 2019; Tsang & Zou, 2022; Wei et al., 2023). For behavioral measures, the latter effect is capturable but rather small compared to the former effect (in lexical decision, Zhang et al., 2024) or nonexistent or short-lived (in sentence reading, Hyönä et al., 2024), and it is easily trumped by whole-word frequency (i.e., only exists for infrequent words, Cui et al., 2021; Xiong et al., 2023). As such, the dominant view of Chinese word recognition is the interactive-activation approach (Li & Pollatsek, 2020; Reichle & Yu, 2018, but see more recently Wei et al., 2023; Yu, Tian, & Lau, 2024, for a stronger view of decomposition in Chinese word recognition with electrophysiological evidence).

Three points must be made about the work reviewed so far. First, and simply to reiterate, tasks undoubtedly play a role in how much the parts and whole matter. Second, the morphosyntactic structure of a word very likely is relevant, too. Most ERP studies that found character-level effects included, or specifically compared, noun-noun and verb-object sequences (Hsu et al., 2019; Tsang & Zou, 2022; Wei et al., 2023; Yu et al., 2024), while target words in most eye-tracking-during-reading studies are homogeneously nouns (Hyönä et al., 2024; Xiong et al., 2023; Yang et al., 2012; Yu et al., 2021). For example, Hyönä et al. (2024), who failed to find a robust effect of first-character word frequency, used exclusively nouns as their target words.

Finally, the work reviewed so far all used two-character strings, with the assumption that both characters are well within the perceptual span to be fully processed. In Chinese sentence reading, where numerous characters are presented without white spaces, an additional segmentation process might be needed for word recognition due to the constraint of visual attention and acuity (Bertram & Hyönä, 2003; Inhoff & Liu, 1998; Li et al., 2009). Understanding whether segmentation is distinct from decomposition or interactive activation and when segmentation happens is crucial for formalizing Chinese sentence reading (Li & Pollatsek, 2020; Yu et al., 2021).

Yet an even more fundamental question is: What constitutes a word in Chinese? In other words, what are Chinese readers segmenting for or recognizing, if not each single character? The next section (1.1.3) elaborates different conceptualizations of decomposition, segmentation, and interactive activation. Then Section 1.1.4. introduces a way to distinguish single-words vs. multiple-words in Chinese. Section 1.1.5. introduces yet a different framing for the Chinese lexicon, whereby single-words and multiple-words differ only quantitatively, such that a multiple-word can become a single-word-like (i.e., both can have stored, unified linguistic representations).

1.1.3. More on decomposition, segmentation, and interactive activation

The terms *decomposition* and *segmentation* have been introduced somewhat interchangeably so far. For instance, Taft (2004) used the term *morphological decomposition* while Rastle et al. (2004) used the term *morpho-orthographic segmentation* to refer to the same early process during word recognition for English. In both cases, the decomposed/segmented morphemes are still simultaneously processed en

route to recognition of the whole word³. For alphabetic scripts that adopt between-word spacing, as there is no need to segment words from a string, conceptualizing decomposition and segmentation as two different processes will be unnecessary, as they both work on morphemes within a clearly defined word. On the other hand, in word recognition for Chinese scripts, decomposition and segmentation can refer to fundamentally different processes. During recognition of a short character string (e.g., two characters) without sentential context, where there is little or no need to segment the characters⁴, decomposition might still happen for the presented string to be processed as two morphemes simultaneously.

In Chinese sentence reading, however, segmentation can be conceptualized as a necessary process to occur fully prelexically (even pre-decomposition) that is distinct from decomposition. For instance, a Chinese reader might always first segment the next two unrecognized characters from the rest of the string and lexically process⁵ only these two characters as a word. If the segmentation leads to errors (e.g., the two characters do not make one but two separate words, or the current first character should in fact be grouped together with the previous second character), reprocessing costs would emerge

³ Also recall that interactive action models such as that in Taft (1994) assumed no independent process of decomposition but assumed early simultaneous processing of the embedded morphemes. Still other models such as the dual-/multiple-route models (Hyönä et al. 2004; Kuperman et al., 2009) assumed early decomposition but suggest that morphemes are sequentially processed from left to right after being decomposed, especially for long compound words.

⁴ Note, however, a caveat that a lexical decision task where both words and nonwords are shown to the participants might unnaturally introduce a need to evaluate the combinability of the two characters as separate words, which may encourage segmentation (Xiong et al., 2023). Here one can simply consider the processes involved for a rapidly presented masked prime of a two-character word.

⁵ Here I define lexical processing as processes beyond visuo-orthographic processing, including morphemic/character, phonological, semantic, and syntactic processing, etc.

(Perfetti & Tan, 1999). Such a two-character assembly strategy as a heuristic will make segmentation a separate process independent from and preceding decomposition. Another prelexical approach to segmentation is that of Yu et al. (2021). Note that Yu and colleagues' model is a model of eye movements during Chinese reading but not a model of Chinese word recognition; the word identification system was thus not specified in detail. Every four unrecognized characters in a sentence are first segmented into either a one-, two-, three-, or four-character word by four simultaneous familiarity checks gauging how long it will take to eventually recognize each string if segmented that way. A second stage of lexical processing, which leads to full recognition of a word, only starts after the segmentation decision. Once segmented, the characters to the right will stop being lexically processed until the next cycle. As illustrated in these two approaches, then, an easy distinction between decomposition and segmentation in Chinese word processing is that both the left and right separated components will be lexically processed after decomposition while only the left component will be lexically processed after segmentation, with the right component being lexically (re)processed after an attention shift.

Yet another way to conceptualize segmentation in Chinese is the interactive activation framework by Li et al. (2009) and Li & Pollatsek (2020). In both models, there was no explicit segmentation process; instead, segmentation is through recognition of a single word⁶. All characters within the perceptual span (1 character to the left and 3 characters to the right of the current fixation) that are yet unrecognized go through the activation-

⁶ Note again that Li and colleagues' models, like Taft (1994), did not explicitly implement decomposition as a separate process either, but character/morphemic influences could be realized through interactive nodes between the character and word layers and within the word layer (Zhang et al., 2024).

for-recognition process at each time step. The word node whose activation reaches the recognition threshold first will be recognized and segmented from the rest of the character string. The component not belonging to the segmented (recognized) word will be lexically (re)processed after an attention shift.

A final note on the distinctions among decomposition, segmentation, and interactive activation is in reference to Yang et al. (2012). This study focused on the semantic processing of the first character of a two-character word and their interpretation of the findings is that the first character was not *segmented* as a single-character word and as such there was no semantic effect from this character as a word. With respect to processing of morphemes, there are multiple interpretations. First, the data are compatible with a strong holistic-word-processing hypothesis with neither decomposition nor any processing on the morpheme level for the segmented word (DOORKEEPER). Second, it could be that after being first segmented as one two-character word, the two characters still go through decomposition (DOOR-KEEPER) but either semantic processing from the morphemes is rather weak or only a fully recognized *word*'s meaning is eventually integrated with the preceding context (i.e., late semantic integration). Finally, the third possible interpretation is interactive activation. In both the example sentences above (1a & 1b), the whole word won the recognition competition over the single-character words and thus the semantic incompatibility of the first character in (1b) did not leave any trace of processing difficulty. Under interactive activation, if anything, one should expect (1a)'s processing time to be longer, as in this case *door* might be more predicted and its activation might be more boosted given the context *kick* and lead to more competition than given the context *entreat*. This indeed was numerically so for first-fixation duration

and gaze duration in Yang et al. (2012).

One issue for both Li and Pollatsek (2020) and Yu et al. (2021)'s models is knowing which character string has a corresponding word node for segmentation/recognition. Unlike languages that utilize white spaces in their writing system, Chinese users have difficulty reaching consensus on how many words a string consists of (He et al., 2021; Hoosain, 1992; Liu et al., 2013; Peng & Chen, 2004). On the implementation level, both Li and Pollatsek (2020) and Yu et al. (2021) used pre-parsed corpora as reference for existence of word nodes and their frequency, which is somewhat arbitrary (see Footnote 1). On the theory level, it is unclear what strings can be stored and represented as a word and whether word nodes are even the highest level of stored, unified linguistic representations. Must a multicharacter string be stored as multiple separate individual words and composed online or can it be stored as a single unit?

1.1.4. Multicharacter string as one single word versus multiple words

Categorizing a string containing multiple characters as one single word or multiple words can be objectively difficult (He et al., 2021; Hoosain, 1992; Liu et al., 2013; Peng & Chen, 2004). In an offline task (Liu et al., 2013), native Chinese readers were asked to explicitly segment sequences of characters into words by putting vertical lines as boundaries. The results showed substantial divergence among Chinese readers. For instance, noun-noun sequences and adjective-noun sequences are sometimes segmented as two separate words (45% of the time for noun-noun and 32% for adjective-noun) but other times as one single word unit.

Do individual differences in offline segmentation preference influence online processing? In a study (He et al., 2021), participants were first asked to naturally read

several experimental sentences in one task. Before the natural reading task, the same participants underwent a constituent decision task (whether each of 190 four-character strings should be explicitly segmented as one or two words). The results of this offline constituent decision task were used to group readers into one-word segmenters who almost always grouped four characters as one word and two-word segmenters who almost always grouped four characters into two separate words. In the main experiment, critical words were embedded in the experimental sentences and eye movement measures on the critical words were compared across one-word segmenters and two-word segmenters. The critical words were sorted into three groups based on an offline segmentation task by yet a different group of participants: one-word-unit (4-character strings that were almost always grouped as one word), ambiguous (approximately half of the time grouped as one word), and two-word-unit (almost always grouped as two words). Among the two factors and their interactions (one vs. two segmenters \times one- vs. ambiguous vs. two-unit), the only significant effect was the main effect of word unit such that readers' fixations were shorter and readers made fewer fixations on one-word-unit words than the other words. Given these patterns, the authors concluded that online reading is robust to individual differences in offline segmentation preference. Instead, certain strings of characters are likely to be represented as one meaningful linguistic unit such that those strings unanimously normed as one unit offline are processed in real-time as one as well.

One potential factor in such storage was proposed to be frequency of co-occurrence (Zang, 2019; Zang et al., 2024, see Section 1.1.5), as it too distinguished the one-word-unit group from the other two groups. Co-occurrence of characters, however, might not be the only variable influencing offline decision and online processing. Another kind of

information that might be utilized online is the morphosyntactic category of a character. For instance, in Liu et al. (2013), syntactic categories influenced offline segmentation decisions: a verb-noun sequence is more likely to be segmented into two individual words (78%) than a noun-noun (45%) or an adjective-noun (32%) sequence. That is, a verb-noun sequence is more often considered to be of multiple words than of one single word in Chinese.

Indeed, there are debates in Chinese syntax and pedagogical studies regarding whether a verb-object sequence should be defined as multiple words (Paul, 1988, Sybesma, 1999) or a compound word (Li & Thompson, 1989; Wang, 2009, also see Hsu, 2015, for a review). Here I present two theoretical reasons why a verb-object sequence is more likely than a noun-noun or adjective-noun sequence to be represented as two separate words rather than a compound word.

First, Chinese morphology of compound words is rich. Coordinative compounds are those with two same-category constituents (e.g., flower-grass = PLANT; teach-foster = EDUCATE), subject-predicate compounds are those with a noun as the first constituent and either a verb or an adjective as the second constituent (earth-shake = EARTHQUAKE; face-red = BLUSH), modifier-head compounds are those with either an adjective or an adverb as the first constituent and a noun or verb as the second constituent (big-door = GATE; lightly-look = DESPISE), verb-complement compounds are those with a verb as the first constituent and an adverb as the second constituent (push-wide = PROMOTE; change-better = IMPROVE), and verb-object compounds are those with a verb as the first constituent and a noun as the second constituent which is a direct object of the verb (draw-lot; ride-horse). One way to categorize these various types of

compound words into two groups is in terms of argument structures. Only for subject-predicate and verb-object compound words are there both one argument introducer and one argument (face-red⁷, ride-horse). In other cases, there is either only the argument introducer (teach-foster; change-better; lightly-look) or only the argument (big-door; flower-grass). As verb-argument realization has been shown to involve special online computation (e.g., Liao, Lau, & Chow, 2022), a sequence that contains both a verb and an argument may be more likely to be represented and function as two separate words, e.g., as evidenced by offline segment decision, Liu et al. (2013).

The second argument for verb-object (and subject-predicate) compounds as two words is their separability by other words. One can insert indefinitely many other characters/words between a verb-object sequence. For example, 我-I 常-often 爬-climb 山-mountain can be extended to 我-I 常-often 爬-climb 比較-relatively 少-few 人-people 爬-climb 的-complementizer 山-mountain (translation: I often climb mountains that fewer people climb.). Similarly, 你-your 臉-face 紅-red can be extended to 你-your 臉-face 可別-better not 因為-because 緊張-nervous 而-accordingly 紅-red (translation: Don't you blush just because you are nervous). This separability of the two characters that still maintains fully the meaning of the two characters is almost impossible for other compound types. Therefore, these two types of sequences should be more multiple-words-like than one-word-like than the other sequences.

In contrast, some other linguistic theories suggest that Chinese verb-object sequences are sometimes compound words, despite having both the verb and the syntactic

⁷ Adjectives in Chinese can function as verb-like predicates (Dixon, 2004).

complement (Luo, 2022). These verb-object sequences are known as bare noun incorporation constructions—contrasting cases where the noun following the verb is full-fledged (e.g., proper names, definite descriptions, etc.)—where the bare noun complement does not serve as a semantic argument to the verb but forms part of the event kinds because it denotes a prototypical theme. Such sequences have reduced discourse capacity. That is, the object nouns are non-specific and lack referential force. Note that under this theory whether the two words in the construction form a single compound word or two separate words is context dependent: only when used in a habitual context is it a compound word. Examples are 抽烟 SMOKE-CIGARETTE (to smoke in *Do you smoke?* 你抽烟吗?, cf., *Earlier I was smoking downstairs.* 我刚刚在楼下抽烟)⁸. These constructions tend to depict well-established, stereotyped activities, and thus usually have a *word* entry in a dictionary or a corpus.

Still others (Yeh, 2020) suggest a 4-level distinction among verb-object sequences (verb, word-like verb, phrase-like verb, and verb phrase⁹), depending on whether the two characters are separable by the perfective aspect marker 了 and whether the noun is referential. For instance, 投资 (CAST-RESOURCE, to invest) is a word because the two characters cannot be separated by 了; 打球 (HIT-BALL, to play ball) is phrase-like because they are separable by 了 but the object BALL is not any particular ball; 洗手 (WASH-HAND) is a verb phrase because they are separable and HAND always (implicitly) refer to someone's hands.

Note that even under these last two views (Luo, 2022; Yeh, 2020), only some verb-

⁸ Other examples include SING-SONG, READ-BOOK, CATCH-FISH, etc.

⁹ Their use of “phrase” corresponds to my term of “multiple-words”.

object sequences are single words, with the other verb-object sequences being either multiple-words or multiple-word-like. As we will see in the next section, yet other theories suggest abandoning the qualitative distinction between single-words and multiple-words and propose another psychological realistic unit—multi-constituent/multi-word units—whose strength of representation can vary quantitatively.

1.1.5. The role of statistical regularities in stored, unified linguistic representations

In the previous sections, I have discussed the difficulties of categorizing a string of characters into *one single word* or *multiple words* in Chinese by assuming a qualitative difference between single-words and multiple-words (or words vs. phrases, Huang, 1984). The proposed morphosyntactic approach to determine the string's status is also more or less a qualitative index. Others have proposed a less categorical distinction between single-words and multiple-words (Zang, 2019). This view—the Multi-Constituent Unit (MCU) Hypothesis—suggests that linguistic units such as spaced compound words, binomial word pairs, idioms, and common phrases may be stored and used as single linguistic representations, the strength of which varies depending on how frequent the exposure to the co-occurrence of the multiple constituents is (He et al., 2021; Zang et al., 2024) or how meaningful the multi-constituent unit is (Jolsvai, McCauley, & Christiansen, 2020; Yu et al., 2016; Zang et al., 2021; I will reserve an introduction of these Chinese empirical studies to Section 3.1.1. due to their high relevance to Experiment 3).

This view is consistent with the usage-based theory of language (Bybee, 2006). Indeed, outside of the Chinese literature, similar concepts of language users' storage of multiword expressions (MWEs) have been proposed (Jackendoff, 1997, see Contreras

Kallens & Christiansen, 2022 for a review). This line of research began with the findings that processing of idiomatic expressions that span multiple words (*kick the bucket*) is faster than that of literal ones (e.g., Swinney & Cutler, 1979). Recent evidence indicates that such a processing advantage is more associated with the formulaicity (rather than idiomaticity) of the expressions. For instance, formulaic (but literal and compositional) expressions are also processed faster than their counterparts (*in the middle of* vs. *in the front of*) in a self-paced reading task (Tremblay et al., 2011). In eye-tracking, Siyanova-Chanturia et al. (2011) found shorter and fewer fixations on high-frequency binomials than controls (*bride and groom* vs. *groom and bride*). That formulaic expressions exhibit such processing advantages suggests existence of MWEs that straddles the distinction between words and multiple-words.

Another way to explain the processing of formulaic expressions is familiarity, which can be quantitatively operationalized by the statistical regularities of their constituents. That is, the embedded constituents/words of a formulaic sequence tend to have a higher co-occurrence frequency (hence language users have more exposure to it), compared to their control with one word being replaced. Arnon and Snider (2010) adopted pairs of 4-word English sequences ranging in co-occurrence frequency: for *Where do you live* vs. *Where do you work* and *It was really funny* vs. *It was really big*, while in both pairs the first sequence is more frequent than their second counterpart, the difference is much bigger in the first pair than that in the second pair. It was found that reaction times in a phrase-decision task (*Is this phrase a possible sequence?*) were a function of the log co-occurrence frequency of the four words. Critically, that reaction times were shown to be sensitive to a wide range of co-occurrence frequency of four words indicates a continuum

of the familiarity effect. Taken together, the abovementioned studies and other similar findings in production tasks (Bannard & Matthews, 2008), recall tasks (Jacobs et al., 2016) or error-detection tasks (Huang & Staub, 2023) suggest that MWEs might be learned and represented.

It is worth noting that the multiword frequency effect can alternatively be explained by prediction and retrodiction, a view that emphasizes on-the-fly composition rather than accessible stored representation (Onnis & Huettig, 2021). For instance, *Where do you live* is easy to process not necessarily because the four words together form a familiar chunk but because of the highly predictable last word itself (the probability of *live* given *Where do you*) or because of the great ease of integrating this word into the previous three words. Onnis and Huettig illustrated this possibility by reanalyzing the datasets from Arnon and Snider (2010) and Bannard and Matthews (2008). In addition to having log co-occurrence frequency of the whole target string, they added into the regression model the conditional probability of the last word given the first three words as well as the conditional probability of the first three words given the last word. The results showed that a model that included all three predictors explained the data better than a model including only the log co-occurrence frequency variable. In fact, in the former model, the effect of log co-occurrence frequency became not significant (see, however, Goodkind and Bicknell, 2021, for co-occurrence frequency not reducible to predictability). Based on the findings, the authors argued for a strong limit on MWEs being stored as a single linguistic unit.

One way to further Onnis and Huettig's arguments, which were based on post-hoc regression analysis, is to experimentally tease apart co-occurrence frequency and

conditional probability and demonstrate an effect of the latter but no effect of the former. Yet such an approach will be extremely difficult, if not impossible, if one also aims to match the frequency of the embedded constituents and words. Yet another possibility then is to devise experiments that allow more inference about how the effect unfolds during the time course. Onnis and Huettig's prediction-and-retrodictio hypothesis predicts an effect to emerge late, since the facilitation is driven solely by the last word, while an MCU/MWE hypothesis might predict an earlier top-down effect from the whole string. The three experiments in this dissertation aim at shedding light on this issue.

Finally, as with the earlier section, it remains unclear whether morphosyntactic structure of the string plays a role in the formation of MWE representations. While some studies have used strings of various structures (as in Arnon and Snider, 2010), eye-tracking studies that reported a relatively early multiword frequency effect (on first-pass reading of the whole string) are exclusively binomials (e.g., *bridge and groom*, Siyanova-Chanturia et al., 2011) or idioms (Carrol & Conklin, 2020). In Chinese sentence reading, MCUs such as literal noun-noun compounds were shown to be processed as a whole with the semantic plausibility manipulation as introduced earlier (Yang et al., 2012; see also Wang et al., 2023 for similar findings with four-character strings), which suggests very early holistic processing. Literal verb-object sequences on the other hand have been examined by Jiang, Jiang, and Siyanova-Chanturia (2020), with a frequency manipulation. In an eye-tracking experiment, they adopted strings with either high or low co-occurrence frequency (e.g., 参加会议 vs. 参加游戏 *attend a meeting* vs. *attend a game*). Areas of interest were the final two characters as a word and the four characters as a whole. It was found that the multiword co-occurrence frequency had a sustained

facilitatory effect on processing time of the final two characters (first-fixation duration, gaze duration, rereading, and fixation counts), and there was a similar effect on processing time of the whole four-character region, although the effect on first-fixation duration was only trending toward significant. A similar MCU frequency effect was replicated in Jiang and Siyanova-Chanturia (2023), but again the effect on earlier fixation measures (first-fixation duration and gaze duration) was not significant¹⁰. It is thus yet unclear how early a verb-object MCU frequency effect can be. More importantly, no studies have investigated within a single experiment noun-noun and verb-object sequences and their interaction with co-occurrence frequency.

1.1.6. Summing up and looking ahead

To summarize the literature, from studies focusing on two-character processing, effects associated with both the embedded character (representing a morpheme in Chinese) and the whole word on response times, eye movements, and electrophysiological responses have all been documented. However, it is quite clear that effects associated with the whole word are stronger and more robust across paradigms, a pattern that can be explained by interactive-activation models (Li & Pollatsek, 2020; Zhang et al., 2024). This model assumes a high-level layer of representations (word nodes) that provide top-down facilitation for character recognition and eventually word recognition itself, and it does not assume a separate segmentation mechanism to deal with long strings of characters. However, implementing such a model requires a deeper understanding of Chinese readers' lexicon, namely what constitutes a word, and raises a question whether words are even the highest-level visual objects Chinese readers are

¹⁰ These two experiments, however, had rather low statistical power, each with 26 participants and 20 items in each condition.

trying to recognize among multiple unspaced characters during reading. I have considered two views of categorizing multicharacter strings, one where a qualitative distinction between single-words and multiple-words exists (Huang, 1984; Luo, 2022; Yeh, 2020) and the other where multi-constituent units (MCUs) straddle the distinction between single-words and multiple-words (Zang, 2019; Contreras Kallens & Christiansen, 2022). Under the former view, verb-object sequences, if the noun has clear reference and if the sequence is separable by perfective aspect marker and does not denote a stereotyped activity, are multiple-words, not single words. Under the latter view, even such a verb-object sequence can form a stored, unified linguistic representation as long as it is encountered frequently enough by the reader, although such a hypothesis has not been widely tested. On the other hand, noun-noun sequences are clearly single words or MCUs and by no means multiple-words, under both views.

Here I empirically test whether Chinese readers do form MCU representations, and whether they do so ubiquitously for multicharacter strings of all kinds of structures, by contrasting noun-noun and verb-object sequences. In Chapter 2, the empirical phenomenon of interest in Experiments 1 and 2 is whether two Chinese stimuli can be recognized in parallel within a very short period of time, using a part-of-speech decision task with rapid parallel presentation. The general rationale behind examining this phenomenon is as follows. A string whose embedded constituents must be processed serially and sequentially is a string not stored together, because if it is, the existing representation should provide top-down facilitation for it to be processed as one unit. If parallel recognition is possible only for noun-noun, not verb-object, sequences, this will provide support for a qualitative distinction between single-words and multiple-words in

Chinese, namely verb-object sequences do not get stored together as noun-noun sequences do. If parallel recognition is possible for frequent, not infrequent, sequences, regardless of the morphosyntax, this will provide support for existence of MCUs stored and represented in Chinese readers.

Experiment 1 uses two-character stimuli, while Experiment 2 uses 4-character strings, and includes also subject-verb sequences, which I suggested in Section 1.1.4 should also be considered multiple-words given their morphosyntax. To anticipate, it was found that even four-character verb-object and four-character subject-verb sequences could be recognized in parallel, as noun-noun sequences were, but only robustly so when the co-occurrence frequency was high. This suggests Chinese readers do form MCU representations, given enough exposure and regardless of their morphosyntactic structures. However, morphosyntax also appears to play a role, *when the multicharacter strings are very unfamiliar* to the readers: Readers' recognition of the embedded constituents of a string is worst (and most serial-like) for subject-verb sequences, compared to verb-object and noun-noun sequences.

CHAPTER 2

CAN MORE THAN ONE WORD BE RECOGNIZED

AT A TIME?

2.1. Serial versus parallel word recognition

Whether or not readers can process multiple words or one word at a time has been a debate (Reichle, et al., 2009; Snell & Grainger, 2019; Inhoff, Starr, & Shindler, 2000). This issue had been addressed mostly in the domain of sentence reading by examining whether words following a target word can influence processing time on the target word even before the eyes move beyond the target word (i.e., a parafoveal-on-foveal effect). Kennedy and Pynte (2005) argued for parallel word processing based on the analysis of eye movements during reading of English corpus data (non-experimentally manipulated, natural sentences). There, it was found that the frequency of Word $n+1$ had an facilitatory effect on Word n , a pattern not expected under a serial view where lexical processing of a word only starts after lexical processing of the previous word has been finished (Reichle et al., 2003). In an experimental study with four separate experiments (Brothers et al., 2017), however, no evidence of either frequency or predictability parafoveal-on-foveal effects was found at all (see also Veldre & Andrews, 2018a).

A different piece of evidence of seriality came from decision tasks with two simultaneously presented words with pre- and post-masks (White et al., 2018). Two words were rapidly flashed (42 ms) following and preceded by a mask (42 ms). The main manipulation was whether at the beginning of a trial participants saw a pre-cue to know which one of the words they would need to make a semantic decision on later. When

there was such a pre-cue, participants only needed to pay focused attention to one single word at a time; when there was no pre-cue they needed to split their attention to both words, to respond to the side that would be probed later. The results showed that, when plotted through “attention operating characteristic” curves (AOC, see Section 2.2 for details), participants’ performance fit the predictions of a serial processing model in which only one word could be processed at a time and was much worse than predicted by a limited-capacity parallel model. Similar results were found in lexical-decision tasks (White et al., 2019; White et al., 2020). On the contrary, performance in a color judgment task fell between the predictions of a limited-capacity parallel model and an unlimited-capacity parallel model. These findings suggest that linguistic processing can only take place for one word at a time in English. Further support came from a functional MRI study (White et al., 2019) which showed that the BOLD signal at the left anterior visual word form area was influenced by the frequency of only the attended word, with this locus argued to be the neurophysiological bottleneck of serial word recognition.

Despite compelling evidence of serial word recognition in English, word recognition in Chinese might in theory be more parallel, given that semantic information is more densely represented in a character (Hoosain, 1991; Hyönä et al., 2024; Li et al., 2022). Furthermore, Chinese characters are horizontally narrower than alphabetic words, and there are no explicit spaces demarcating word boundaries, making two words visually closer¹¹.

Some evidence of parallel word processing included Yan et al. (2009), who found a

¹¹ Indeed, using the same paradigm as White et al. (2018; 2019), White (2023) more recently found that when two short English words were juxtaposed close together, parallel recognition of two words is possible, at least for some participants (see also simulation by Reichle and Schotter, 2020, and Section 2.2.6. for discussion).

semantic parafoveal-on-foveal effect on gaze durations, arguing that much linguistic processing of Word $n+1$ has occurred early enough to influence concurrent processing of Word n (see also Yan & Sommer, 2015; Yang et al., 2009). When presented without sentence context, however, simultaneous recognition of two words appeared to be difficult. Li et al. (2009) investigated how Chinese readers recognize word(s) without sentence contexts. The task was to report as many characters as possible (or to search for a particular character) among four characters that were briefly presented. Conditions varied in terms of the numbers of words (e.g., two 2-character words or one 4-character word) within the string and the semantic relatedness between words. The conclusions were that characters within a word can be processed in parallel and processing efficiency decreases from left to right. That is, if Characters 1 and 2 formed a word, both characters' accuracy would be high. If the string contained 4 characters that do not form any words, character accuracy decreases sharply from left to right. Yet word recognition is serial. That is, when the string contained two 2-character words, character accuracy was adequately high only for the first two positions, but when the string contained only one single word, character accuracy was very high for all four positions. Finally, there was an effect of semantic relatedness: character accuracy in the third and fourth position was higher if the two 2-character words were semantically related, although still not close to ceiling as in the one-4-character-word condition. The authors interpreted the last finding as a context effect, such that Characters 1 and 2 were first recognized as a word and then the word recognition process of Characters 3 and 4 became facilitated due to this context. However, it has to be noted that the individually-thresholded presentation time in the study was extremely short (11-40 ms), so that in theory it did not allow even convert

attention to shift. Therefore, the last finding might require re-interpreting.

One limitation in Li et al. (2009) is that they did not implement masking for the critical experiments (Exps 3 & 5). This allows a possibility that visual information was encoded in parallel for all characters and retained for a short period, and the higher accuracy for the semantically-related condition was due to this retained visual memory enhanced by the semantic information from the first word so that the second word was correctly guessed more frequently. It needs to be verified whether such results hold even when masking is applied. Another limitation is the lack of use of AOC to characterize participants' performance as all-or-none, limited-capacity parallel, or unlimited parallel. In their Experiment 4—where on each trial participants were first given a character and then had to report whether it was present in a briefly flashed string—while performance at the last two character positions dropped significantly in the two-word condition, accuracy at both positions was around 85%, which is higher than chance level. More formal analysis is thus necessary to quantify the extent of parallelism. Finally, the 4-character strings in the study were intentionally positioned such that the participants' fixation was on the first character, with the rest of the three characters falling to the right of the fixation, for the purpose of mimicking natural sentence reading where readers read from left to right. This setting, however, could have given rise to the observation of the drop in recognition accuracy for the right two characters simply due to lower visual acuity. Thus, it is yet unknown, when given an optimal viewing position, whether Chinese readers can recognize two separate words in parallel at a given time point.

2.2.Experiment 1: recognition of 2-character noun-noun compounds and verb-object sequences

The goal of Experiment 1 is to probe the existence of noun-noun and verb-object MCUs. The rationale is that stored representation of an MCU will make it more likely for the two embedded words to be recognized in parallel due to the high-level, top-down influence. On the other hand, a string whose embedded constituents must be recognized serially in turn provides strong evidence of no stored representation of the whole string.

Here I adopted the task from White and colleagues (2018; 2019; 2020) with Chinese material, and with a part-of-speech (POS) judgment instruction (*Is this character a noun?* see Section 2.2.3. for details). Critically, as I have mentioned earlier, the primary manipulation in this paradigm is whether, *before the target string is shown*, participants get a visual cue showing which side of the string they should respond to (in my case, which character's POS they should judge). By comparing the accuracy in the precue (PC) and no-precue (NPC) conditions, it is possible to determine whether the two characters from both sides are processed in parallel, based on an attention operating characteristic plot (Scharf, Palmer, & Moore, 2011; Sperling & Melchner, 1978). Figure 1 visualizes three hypothetical scenarios in the NPC condition. If lexical processing of each side of the stimuli is fully parallel, the amount of the information processed for each side should be the same as that of the pre-cued, attended stimulus in the PC condition (Fig. 1a). Therefore, even without knowing in advance which side to respond to, the accuracy should be as high as that in the PC condition. On the contrary, if lexical processing for the two sides of the stimuli *must* occur serially and sequentially, at the point of the very brief presentation, the participant must choose one side of the stimuli to process first. In the example of Fig. 1c, the left side of the stimulus gets chosen to be processed first, which results in the same amount of information processed as that of the pre-cued, attended

stimulus in the PC condition. However, if on that trial, it is the stimulus on the right side that the participant needs to later respond to, they will have no information at all such that they must make a pure guessed judgment, hence the label “all-or-none serial processing”. Finally, yet another possibility is limited-capacity parallel processing (Fig. 1b) where concurrent lexical processing of each side of the stimuli is possible, but the information sampled has a fixed limit within an attended visual display per unit time (Shaw, 1980). Given that the region paid attention to in the NPC trial is twice as big as in the PC trial, the information processed for each side, despite gathered simultaneously, will be twice smaller in the NPC condition than that for the pre-cued, attended stimulus in the PC condition.

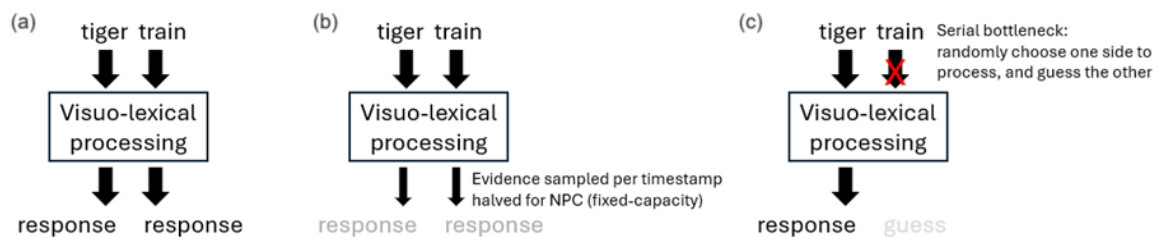


Figure 1. A visual illustration of three possible scenarios of processing two simultaneously presented stimuli when there is no pre-cue. Adapted from unpublished presentation slides of Dr. Dina V. Popovkina. The different shades of the blackness of *response/guess* reflect the different accuracy levels across the scenarios due to different amounts of information processed.

When incorporated with signal detection theory, the predicted accuracy under these three possible scenarios in the NPC conditions can be quantified, based on the accuracy in the PC conditions. This can be visualized in an AOC plot (Figure 2). The filled circle plotted on the x-axis reflects the accuracy for the stimulus on the right side in the PC condition, and the filled circle plotted on the y-axis reflects the accuracy for the stimulus on the left side in the PC condition. The axes start with 0.5 under the assumption that, in

the PC condition, the side not pre-cued is not paid attention to and therefore judgment accuracy on that side should be at chance. For the unlimited (fully) parallel scenario, accuracy on each side in the NPC condition should be comparable to that in the PC condition, and therefore it is expected to fall in the AOC plot at the intersection of the dashed lines. For the all-or-none serial scenario, accuracy on the left side in the NPC condition should be the same as that on the left side in the PC condition, if for all trials the participant chooses to attend the left stimulus, vice versa for the right side. Depending on the proportion of trials participants choose to attend the left side (from 0% to 100%), the accuracy is expected to fall on the diagonal line between the two filled circles. Finally, for the limited-capacity parallel scenario, accuracy in the NPC condition should fall on the curve in the AOC plot (see White et al. 2018 for how to calculate the curve with the conceptualization of limited sampling per display area per unit time described above).

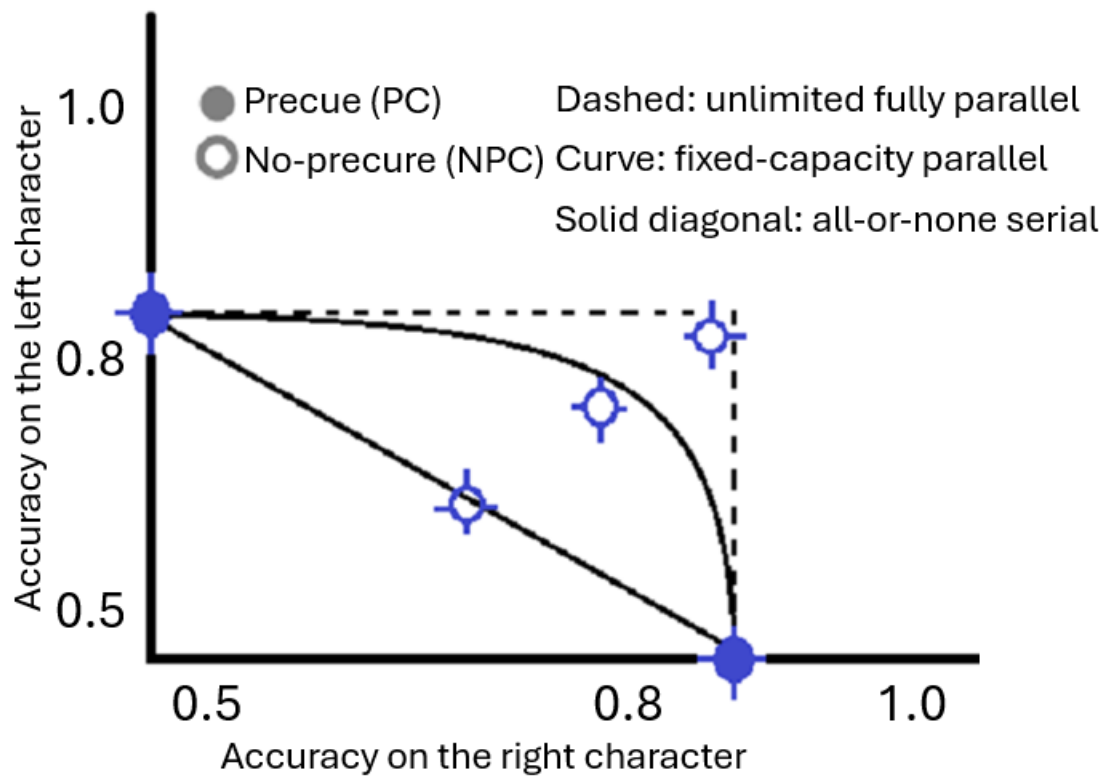


Figure 2. An AOC plot showing three predictions from three models for accuracy in the NPC condition, based on the accuracy in the PC condition. The unfilled circles are three potential outcomes, each consistent with one model’s prediction. Adapted from White et al. (2019).

2.2.1. Participants

Sixty-four participants in the Hampshire County in Western Massachusetts who have received at least 12 years of education in China, were recruited. They received \$24 US dollars for 90-min of participation. Twenty-four participants’ data were discarded due to the following criteria. Six later reported having attended high school in the United States, not fitting the criterion above. Seventeen had difficulty fixating centrally throughout the trials (some due to the lenses of their glasses being too reflective for calibration; some blinked too much) and felt too tired to continue. One participant’s accuracy in the PC condition was only 61.4%. In the end, 40 participants’ data were analyzed (mean age = 23 yr; 8 males, 29 females, and 3 refused to report). See Supplementary Material for

demographic and language-background summary. Among these 40 participants, 6 of them stopped at 75% of the experiment because they ran out of the time. Two participants' experiment shut down (after Block 11 and Block 15 respectively). Given that these participants still had a great number of trials run, their data were kept.

2.2.2. Materials

Five-hundred-forty-four two-character strings were created, as shown in Table 1. Critical stimuli were noun-noun (NN) and verb-object (VO) sequences. Included as control stimuli were verb-verb (VV) and subject-verb (SV) sequences, which are syntactically legitimate sequences in Chinese but were made meaningless or impossible¹². I expect readers to have no representations for these strings and hence processing should be less parallel, compared to the meaningful critical NN and VO items. They were thus included as controls. Inclusion of these items also ensured that participants could not make strategic guesses (e.g., recognizing the left character being a noun means the right character definitely would be a noun, if only NN and VO strings were used). Finally, 32 adjective-adjective sequences were also included as fillers; this is so that participants would not monotonously see only noun or verb characters throughout the experiment. Some characters appeared more than once in the 544 strings (maximal appearance times is 3), but no full stimulus string occurred twice. Mean frequency and number of strokes of each character were matched across groups. Frequency measurement was based on Sun et al. (2018) who combined the SUBTLEX-CH Corpus (33.5 million colloquial words, Cai & Brysbaert, 2010) and a modified Leiden Weibo

¹² For SV sequences, several items were in fact plausible, although very unnatural and uncommon (e.g., MOTH-COME, DRAGON-THROW, ONION-INCLUDE). Most items were semantically anomalous (LUNG-DELETE, TREASURE-SUBSTITUTE).

Corpus (101.4 million words of internet text, van Esch, 2012)¹³.

Efforts were made to include only characters that are POS-unambiguous, first based on the author’s intuition. Each character was then normed separately from the main experiment (see the end of Section 2.2.3 for the procedure), and items with a character whose POS is not unambiguous among Chinese readers were discarded post-hoc and not included in the analysis (see Section 2.2.4).

	Number of items (after ambiguity exclusion)	Whole-word frequency	First character frequency	Second character frequency	First character stroke	Second character stroke
Noun-Noun (NN)	145	4.45	10.2	10.5	9.12	8.62
Verb-Object (VO)	157	4.56	10.3	10.7	8.99	8.63
Verb-Verb (VV)	126	0	10.4	10.8	8.76	8.74
Subject-Verb (SV)	61	0	10.4	9.55	8.69	8.97
Adj-Adj (AA)	30	0	9.74	10.2	8.67	8.37

Table 1. Summary of frequency and stroke measures for each item group in Experiment 1. Frequency measures are natural-log transformation of raw counts from the combined corpus from Sun et al. (2018). The term whole-word frequency is conveniently used to be consistent with the literature (i.e., verb-object sequences that have an entry in Sun et al.’s corpus are defined as words).

Among all items in NN and VO, whole-word frequency ranges widely from very common NN sequences (VEGETABLE-KNIFE, COW-MEAT) to very uncommon ones (GOOSE-LEG, HORSE-ROPE), and similarly for VO sequences (KILL-PERSON, SAVE-LIFE being very common and WIPE-BOWL, BORROW-VIOLIN being very

¹³ About one-fourth of the NN and VO sequences has no word entry in this combined corpus due to the fact that this corpus was pre-parsed. I checked against the CCL corpus (2014 version), a raw, unparsed corpus of 581 million characters, for the co-occurrence frequency of the two characters. The frequency was then divided by 10 to approximately match the frequency measurement from Sun et al.’s combined corpus.

uncommon). This difference in frequency should lead to different strengths of representation of the string, which I further test by subgrouping the NN and VO conditions into strings of characters that frequently co-occur (FCO) and strings of characters that barely co-occur (BCO). The subgrouping criterion is based on the first and last quartiles of whole-word frequency of NN and VO, using the original 160 items. Table 2 shows the frequency and stroke measures of each subgroup.

	Number of items (after ambiguity exclusion)	Whole-word frequency	First character frequency	Second character frequency	First character stroke	Second character stroke
NN-FCO	37	7.04	10.9	10.9	8.65	8.62
NN-BCO	36	1.35	9.77	10.2	9.31	8.19
VO-FCO	36	6.94	10.3	11.4	8.94	7.58
VO-BCO	41	2.37	10.3	10.3	9.46	9.51

Table 2. Summary of frequency and stroke measures for each item subgroup in NN and VO groups. Frequency measures are natural-log transformation of raw counts from the combined corpus from Sun et al. (2018).

2.2.3. Procedures

The experiment was conducted in Mandarin. Participants first completed a language background questionnaire, and then were seated in front of the eye-tracker. The monitor was Dell U2412Mb 24 inches, with a refresh rate of 60 Hz. Resolution was set to 1920 × 1080. The distance between the screen and the participants' eyes was 750 cm. At this distance, each character (48 pixels) spans about 1 degree of visual angle. An instruction was shown on the screen with the experimenter giving a verbal walk-through simultaneously, as follows.

“The task is to judge Chinese characters' parts-of-speech. For each trial you will see a two-character string very briefly shown at the center of the screen, after which two squares will be displayed at the center of the screen. One of the squares will be colored

yellow. This yellow square indicates the character whose POS you need to judge. At the beginning of each block, a part-of-speech prompt will be shown—in some blocks “*is it a noun?*” and in others “*is it a verb?*”. (I used two different POS prompts to avoid potential positive/negative biases to a certain POS; the numbers of noun/verb blocks were balanced, and accuracy was averaged across POS prompts). Your judgements thus will always be “yes-or-no”. Press the left button for “yes” and the right button for “no”. Furthermore, there will also be two squares in the beginning of each trial; in some blocks, these squares will tell you in advance, before the string is shown, which side of the characters you need to judge; in other blocks, both squares will be gray such that you will not know which side to judge until after the character string disappears.”

Then a diagram of the trial procedure was shown, similar to Figure 3, except without the label of each stage and timing information. Participants were informed when to press the button. They were also instructed to fixate centrally on the screen, as marked by the fixation cross. If they did not, “please fixate centrally” would be shown on the screen and the trial would end, and they would not be able to press the button. They were also told not to directly look at the yellow square despite it being a cue. They could only use their peripheral vision to sense the cue.

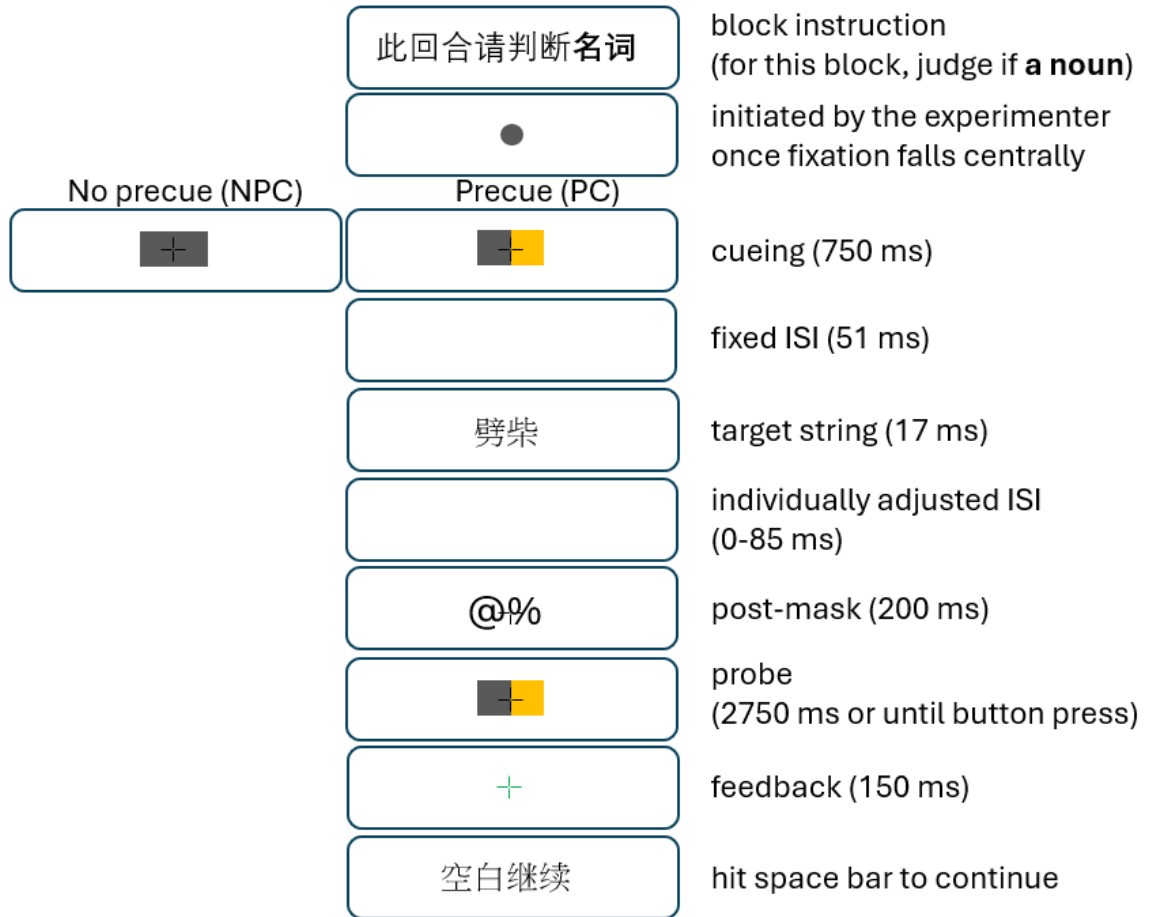


Figure 3. Illustration of the procedure for a trial in Experiment 1. The block instruction on the top of the figure only occurs once at the beginning of each block. ISI: inter-stimulus-interval.

Then a few examples of characters with their POSs labeled were shown. Participants were told that all characters should in theory be unambiguous, but in cases they found the probed character ambiguous, they should go with their instinct (what POS most aligns with that character). Participants were also explicitly told there might be adjectives in the string, despite the question prompt always asking about either a noun or a verb. In cases where an adjective was probed, they should always just press “no”, regardless of which block they were in, since an adjective is neither a noun nor a verb. Finally, they were explicitly told to make the POS judgment solely for the probed character and ignore any

context.

A ten-trial guided-practice block followed. This very first practice block was guided (with the experimenter interfering) because the task was extremely difficult for first-timers, with almost all participants reporting not seeing any characters for the first few trials. They were told to simply familiarize themselves with the procedures and to learn when to press the button, at this stage. If they could not yet perceive any characters, they could just randomly press either button. They were also assured that there would be enough practice runs for them to become capable of seeing the characters before the official blocks.

After the guided-practice block, 2 to 10 practice blocks followed (each with 10 PC trials), depending on their performance in the blocks. The ratio of NN:VO:VV:SV:AA trials in these practice blocks was made similar to that in the official blocks. A staircase inter-stimulus-interval (ISI) adjustment was applied during these practice blocks. The ISI between the target string and the post-mask was initially set to 34 ms. After every two blocks (a verb block and a noun block combined), if the average accuracy across the twenty trials was below 65%, the ISI increased 17 ms for the next two blocks. If the average accuracy across the twenty trials was above 75%, the ISI decreased 17 ms for the next two blocks. The practice blocks continued until either (a) the accuracy for the most recent two blocks fell between 65%-75%, (b) the ISI could not be further decreased (already 0 ms), or (c) all 10 practice blocks were run. This thresholding procedure ensured that the PC conditions could serve as a baseline for comparison with the NPC conditions (i.e., to avoid a ceiling or floor effect).

Finally, after the PC practice blocks, one NPC practice block of 14 trials was

administered. After that, the participant was asked if they had further questions. They were reminded again that all judgments should be solely about the probed character without context being considered. They were also told to do their best and not to be stressed about getting every single trial correct.

There were 16 official blocks, each with 34 trials. The manipulation of PC/NPC was blocked, and so was the manipulation of noun/verb prompt. Every four-block set constituted a run (PC-noun, PC-verb, NPC-noun, and NPC-verb); within each run the order of the blocks was randomized. In each block, there were 10 NN, 10 VO, 8 VV, 4 SV, and 2 AA trials. Within a block, half of the trials were with a left probe and half the right probe, with the order randomized. The ISI was updated every run, if accuracy averaged across the two PC blocks was not between 70% to 80%.

For each participant, each item only appeared once (either with a left or right probe, a noun or verb prompt, PC or NPC) throughout the experiment. In which block an item appeared was also balanced (early or late in the experiment). In total, there were 16 Latin Square counterbalance lists administered.

The whole experiment took on average an hour and a half. After the experiment was over, participants completed an offline POS judgment task on a spreadsheet where each row contained one single character. They were asked to type out, as quickly as possible, whether the character is a noun, a verb, or an adjective on the next column and not overthink. If they found it hard to decide the POS of the character, they should leave the cell blank.

2.2.4. Analysis

Trials with the participant not fixating centrally (0.5° to the left or right and 1° above

or below the center of the screen) between the offset of the cueing and the onset of the probe were aborted, unless the non-central fixation occurred before the target string was even shown, in which case the trial would immediately be repeated. About 5.7% of the trials were aborted in the PC conditions and 4.3% of the trials were aborted in the NPC conditions.

Trials with time-out (no response within 2750 ms) were automatically coded as incorrect trials. Filler trials (AA) were not analyzed due to the low number of observations and the fact that, in Mandarin, an adjective can sometimes serve as a predicate, making it verb-like and not unambiguous. About 5% of the items containing a character that was not highly unambiguous regarding their POS, based on the offline judgment task, were further discarded before analysis: if the agreement on POS was below 75%, the character was considered ambiguous; if the agreement on POS was at or above 80%, the character was not considered ambiguous; if the agreement was between 75%-80% but fewer than 10% of the responses were the counterpart POS (e.g., if the character was supposed to be a verb and only 8% of the people judged it to be a noun), the character would not be considered ambiguous¹⁴.

As mentioned, trials were collapsed across noun-prompt and verb-prompt trials for accuracy calculation, despite this factor being experimentally manipulated. Therefore, for each string structure (NN/VO/VV/SV) per participant, there were four conditions (PC-left, PC-right, NPC-left, and NPC-right) whose mean accuracy was calculated. I then further calculated the grand mean accuracy for each of these four conditions across

¹⁴ This last, more liberal criterion was adopted because, for example, even if a supposedly-noun character tends to be judged as an adjective, participant still could get the trial correct by pressing “no” to the verb-prompt, whether or not they deemed the character a noun or an adjective.

participants, and the by-subject standard deviation. These four conditions' accuracy will be plotted as 3 dots ([PC-right, 0.5], [0.5, PC-left], [PC-right, PC-left]) on an AOC plot, respectively for each of the four string structures.

In addition to AOC plotting, statistical tests were performed to eventually determine whether there was a difference in accuracy between the PC and NPC conditions, as well as whether string structure and co-occurrence frequency modulate the effect of pre-cueing. This statistical analysis combined the noun- and verb-prompt trials (as in AOC plotting) but also the left- and right-probe trials. Two power analyses, using simulation, were conducted for this statistical analysis, one without considering the factor of co-occurrence frequency (a model involving only pre-cueing \times structure) and the other with it (pre-cueing \times structure \times frequency). The former assumes that noun-noun sequences are MCUs that can be recognized as single units and thus their recognition accuracy is not subject to pre-cueing while verb-object sequences are multiple-words (Huang, 1984; Sybesma, 1999) that must be serially recognized. The latter assumes that *only high-frequency noun-noun sequences* are MCUs whose recognition accuracy is not subject to pre-cueing: an interaction between pre-cueing and frequency for noun-noun items and a further interaction among pre-cueing, frequency and structure. For each simulated trial, a linear combination was performed for samples drawn independently from multiple Gaussian distributions that respectively reflected fixed intercepts and slopes, on probability scale; random subject and item effects were also applied in the linear combination. This summed probability then was used to generate a binomial response (correct or incorrect) for each trial. The resulting simulated dataset went through a frequentist generalized linear mixed-effect model to determine if the critical z-values are

larger than 1.96. Table 3 shows the assumed fixed parameters for generating the simulated responses. Details for the simulation are reported in Appendix A.

Simulation 1 (all NN and VO items)	Mean	SD	Note
Intercept (NN with pre-cues)	0.80	0.03	The ISI thresholding keeps accuracy around 80%
No-precue effect (for NN)	-0.02	0.02	Barely an accuracy drop from not getting a pre-cue
VOvsNN (with pre-cues)	0	0.015	No structure effect when there is a pre-cue
No-precue effect × VOvsNN	-0.07	0.03	A big accuracy drop from not getting a pre-cue specifically for VO strings
Simulation 2 (only most and least frequent NN and VO items)	Mean	SD	Note
Intercept (frequent NN with pre-cues)	0.80	0.03	The ISI thresholding keeps accuracy around 80%
Infrequency effect (for NN with pre-cues)	-0.03	0.02	A very small frequency effect even when there is a pre-cue
No-precue effect (for frequent NN)	-0.02	0.02	When strings are frequent NN, barely an accuracy drop from not getting a pre-cue
VOvsNN (for frequent strings with pre-cues)	0	0.015	No structure effect when there is a pre-cue
Infrequency × No-precue (for NN)	-0.07	0.03	A big accuracy drop from not getting a pre-cue for infrequent NN strings
Infrequency × VOvsNN (with pre-cues)	0	0.015	No structure effect when there is a pre-cue
No-precue effect × VOvsNN (for frequent strings)	-0.07	0.03	A big accuracy drop from not getting a pre-cue for frequent VO strings
Infrequency × No-precue × VOvsNN	0.07	0.03	Frequency modulates the pre-cueing effect for NN but not VO strings

Table 3. Summary of Gaussian parameters used to generate simulated data points in Experiment 1. For random effects, see Table A2 in the Appendix.

The first simulation showed a 95% chance of obtaining a significant two-way interaction when all items were included. The second simulation showed a 20% chance of obtaining a significant three-way interaction, a 21% chance of obtaining a significant two-way interaction between pre-cueing and string structure, and a 33% chance of

obtaining a significant two-way interaction between pre-cueing and frequency for NN items. This very low power for the second analysis is likely due to the subgrouping based on first and last quartile co-occurrence frequency.

In analyzing the experiment, due to the high power to detect a two-way interaction when including all the NN and VO items, I first ran a regression model with the full dataset. Then, due to the low power with the subsets of items, I ran multiple simple regression models that included only the main effect of pre-cueing, each time at a particular level of frequency and structure (e.g., with only trials of infrequent NN strings), instead of running one model that included all three factors (frequency \times structure \times pre-cueing). The limitations of this statistical practice are noted in Sections 2.5.4 and 5.1.

2.2.5. Results

I first report the grand means (collapsing across string structure and probe position) in the PC and NPC conditions. The mean accuracy for the former was 84% and for the latter was 80%. That accuracy in the PC condition was around 80% indicates the staircase thresholding method worked adequately. Figure 4 shows the distributions of ISI between the target string and the post-mask for the first run and last run of the official blocks, as well as the distribution of ISI throughout Experiment 1 across all runs. It was found that for the majority of Chinese readers to correctly judge one of the two characters' POS, it only requires 17 ms of presentation of the two-character string and 0-34 ms of blank frame before the 200 ms post-mask.

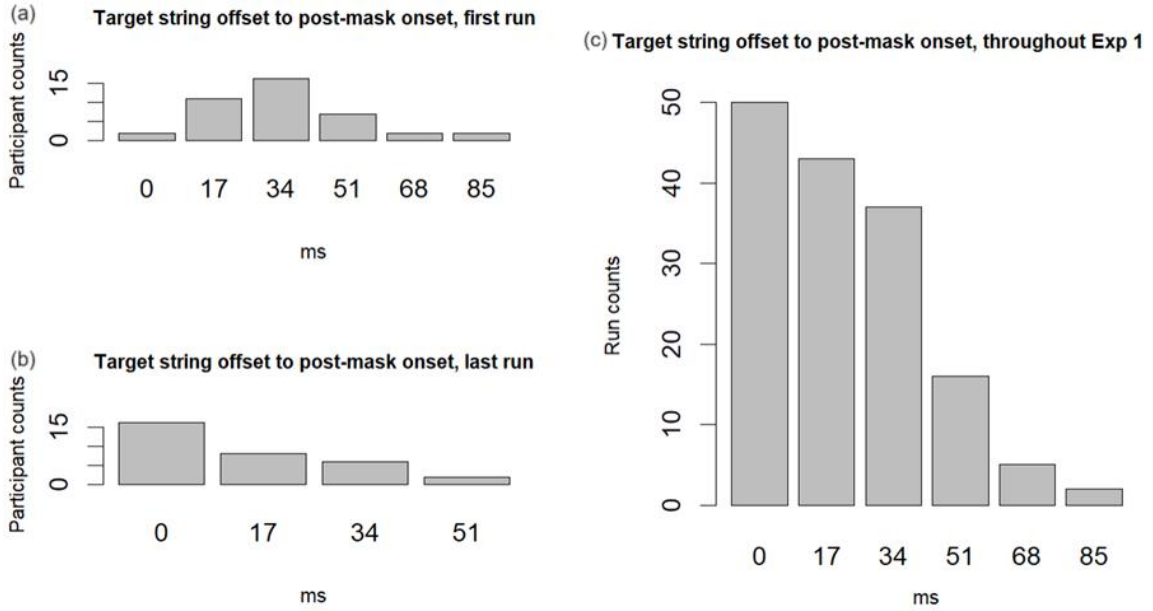


Figure 4. Distributions of ISI duration for the first run in Experiment 1. (4a); for the last run in Experiment 1 (4b); for all the runs in Experiment 1 (3c). Y-axes differ between (4c) and (4a)(4b) because each participant had 4 runs in total.

Fig. 5a shows the results by cueing condition, side and string structure in an AOC plot. The predictions (the diagonal line and the curve) are plotted individually for each string structure, based on the accuracy in the PC condition. The dotted lines (the predictions for unlimited fully parallel processing) in Figure 2 are omitted here for ease of exposition; I will evaluate against full parallelism with statistical tests (see below).

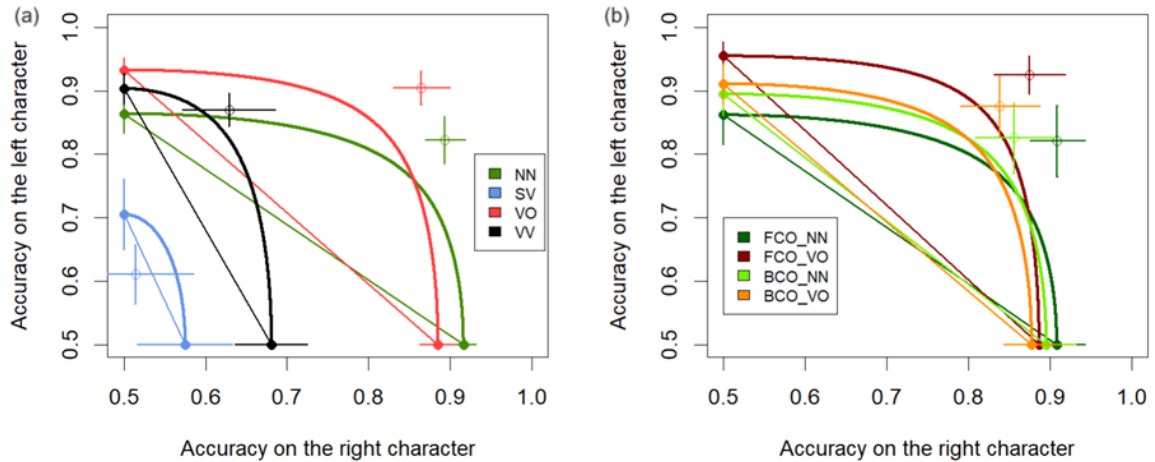


Figure 5. AOC plots in Experiment 1. (a) Accuracy in Experiment 1 based on cueing condition, side, and string structure plotted on an AOC plot. (b) Accuracy from only NN and VO types further separated based on whole-word frequency (FCO: strings of frequently co-occurring characters; BCO: strings of barely co-occurring characters), based on cueing condition and side. Error bars reflect 1.96 * by-subject standard errors.

As is shown in Fig. 5a, NPC accuracy in both NN (green) and VO (red) sequences is clearly above the prediction of the limited-capacity parallel model. In contrast, nonword controls (sequences with two characters not able to be composed semantically), VV and SV, have NPC accuracy reliably lower than their PC baseline. In the VV case, NPC accuracy falls within the prediction by the limited-capacity parallel model. In the SV case, NPC accuracy falls within the prediction by the all-or-none serial model. One notable pattern is that the four string structures appear to already differ in their accuracy in the PC conditions, with NN and VO having the highest, somewhat comparable accuracy, and VV and SV having accuracy much lower than NN and VO.

Although it is clear that NPC accuracy in NN and VO sequences falls above the limited-capacity parallel curve, to more closely address whether full parallel recognition is possible for NN and VO, I ran a Bayesian logistic regression model (Bürkner, 2017) in R (R Core Team, 2021) on the accuracy (aggregating across left and right sides) to see if

accuracy differs across PC and NPC conditions and whether string structure interacts with pre-cueing. Factors were sum-coded, with PC condition coded as -0.5, NPC condition as 0.5, VO condition as -0.5, and NN condition as 0.5. Table 4 shows the results of the Bayesian logistic regression model. It is shown that, while visually on the AOC plot the NPC points fall close to the intersection of PC accuracy, statistically, accuracy in NPC is slightly but reliably lower than that in PC (87.2% vs. 90%). Importantly, there is very little evidence showing that this pre-cueing effect differs between NN and VO items (85.8% vs. 89% for NN and 88.5% vs. 90.9% for VO).

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
Intercept	2.31	0.12	2.07	2.55
NPC_PC	-0.23	0.09	-0.39	-0.05
NN_VO	-0.26	0.10	-0.46	-0.06
Interaction	-0.06	0.15	-0.35	0.22

Table 4. Bayesian logistic regression models of accuracy for NN and VO conditions. Only fixed effects estimates are shown here. R formula: Correctness ~ NPC_PC * NN_VO + (1+ NPC_PC * NN_VO|subj) + (1+NPC_PC|item); PC are coded as -0.5 and NPC coded as 0.5, VO as -0.5, and NN as 0.5. Chains =4, each with 8000 iteration, warmup = 4000. A logit link function was used and the default priors in the brms package were used. All Rhats = 1.00.

To examine the role of co-occurrence frequency in parallel recognition, subsequent analyses were performed separately including trials of items at a particular frequency and structure level (high/low and NN/VO). Fig. 5b shows the results by cueing condition, side, string structure, and frequency in an AOC plot. As with Fig. 5a, NPC accuracy in all subgroups unambiguously falls above the limited-capacity parallel curves. To determine where reliable difference in recognition accuracy exists between the PC and NPC conditions, four Bayesian logistic regression models were run. Table 5 shows the results of the Bayesian models. For FCO subgroups, whether NN or VO, there is no evidence

that accuracy differs between PC and NPC conditions. For BCO subgroups, it is NN sequences whose accuracy in NPC condition is statistically lower than in PC condition (84.3% vs. 89.7%). VO-BCO sequences have statistically comparable accuracy between PC and NPC conditions (86.1% vs. 89.6%), although there is a numerical trend.

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
NN-FCO				
Intercept	2.30	0.21	1.92	2.76
NPC-PC	-0.11	0.23	-0.55	0.37
NN-BCO				
Intercept	2.49	0.23	2.07	2.96
NPC-PC	-0.54	0.21	-0.96	-0.12
VO-FCO				
Intercept	2.59	0.19	2.24	3.00
NPC-PC	-0.22	0.25	-0.69	0.30
VO-BCO				
Intercept	2.56	0.24	2.12	3.05
NPC-PC	-0.25	0.24	-0.71	0.23

Table 5. Bayesian logistic regression models of accuracy for NN-FCO, NN-BCO, VO-FCO, and VO-BCO conditions. Only fixed effects estimates are shown here. R formula: Correctness ~ NPCvsPC + (1+NPCvsPC|subj) + (1+NPCvsPC |item); PC are coded as 0 and NPC coded as 1. Chains =4, each with 8000 iteration, warmup = 4000. A logit link function was used and the default priors in the brms package were used. All Rhats = 1.00.

2.2.6. Discussion

Experiment 1 adopted White and colleagues' method to investigate whether two stimuli presented simultaneously can be recognized in parallel, using Chinese material. Physically, the two stimuli are unambiguously two Chinese characters, yet psychologically, it is less clear whether they represent one Chinese word or two Chinese words, which is the theoretical question asked here. The main empirical question is whether the extent of parallel processing differs across string structure and frequency. That different string structures reveal different results suggests that our item manipulation is successful and sensitive to the task manipulation (pre-cueing). By including strings that definitely constitute two words as the control groups, I was able to demonstrate the limit

of Chinese readers' processing. In the case of VV where both verbs are not semantically related and together meaningless (e.g., DYE-ROW (as in *row a boat*), READ-INCREASE), processing is by no means fully parallel but mostly consistent with limited-capacity parallel. This suggests that even in Chinese, which lacks white spaces, putting visual stimuli adjacent to each other, parallel recognition of two words is not possible. In the case of SV where two words are mostly semantically incompatible (see Footnote 12), processing even becomes strictly serial. Taken at face value, these two results from the control stimuli might suggest that Chinese word processing is strictly serial, but the participants' accuracy potentially benefited from using a guessing mechanism in the VV condition. That is, for trials where the participant was probed on the side that they did not process, they could still get the answer correctly if they simply responded based on what they recognized on the other side, because both sides have the same POS (White et al., 2020).

Another possibility for the two results above is that the very low accuracy for SV in the NPC condition (or in fact both the PC and NPC conditions) arose from the anomaly/impossibility after composing the two words (e.g., SAUCE-CHAT, LUNG-DELETE). This is different from VV, where even though the two words could not be composed, they could simply be perceived as a linear list of two unrelated words, whereas the SV sequences that gave compositionally anomalous meanings might have yielded further costs to response accuracy. Such anomalous/impossible meanings only emerged from the SV condition, which accounted for only about 12% (64/544) of all the trials. This might also have led to some response biases for the participants (i.e., denying a perceived character and inferring that it should be another character that would make

the whole string meaningful or a list of unrelated words). Indeed, the overall accuracy of the SV items is far below the thresholding 80%, which suggests that some other factors in decision-making (rather than in lexical processing, e.g., response bias) might have been involved. I adopt this interpretation, given the results from Experiment 2, where better controls and instructions are used.

My target conditions (NN and VO), in contrast to the control groups, showed performance well above that predicted by limited-capacity parallel processing. For NN that are very frequent (FCO), there was no reliable difference, indicating parallel recognition. Within the experiment, FCO NN sequences are surely psychologically single words, consistent with our empirical observation. On the other hand, FCO VO sequences, which can be either multiple words or MCUs under different theoretical views, were empirically recognized as one unit. This then supports the view that Chinese readers do store VO sequences as unified linguistic units, or MCUs.

For BCO strings, neither NN nor VO sequences should be MCUs, given their infrequent occurrence. However, NN sequences should be understood as one unit, as the two embedded nouns together refer to one entity. It is possible that with unfamiliar compound nouns (for which readers do not yet have stored representations, due to little exposure), processing must start bottom-up. Therefore, no evidence for fully parallel processing was observed. For VO sequences, it is unexpected that even VO-BCO sequences were processed fully in parallel. Note that as seen in Table 2, the least frequent 25% of all the VOs used still appear to contain some not-uncommon strings. This non-zero frequency might be the reason why they still are MCUs. This suggests that forming representations for 2-character Chinese VO sequences requires only little experience with

them.

Finally, one also seemingly puzzling pattern is that the degree of parallelism appears to be a function of how high the accuracy is in the PC condition. In Experiment 2 it will be revealed that this is not always the case.

Experiment 1 contrasted with White and colleagues' earlier series of work (White et al., 2018; 2020; White et al., 2019), where they showed strictly serial processing of two English words. The words in each sequence they adopted were all unrelated, which can be seen as similar to my VV, SV, NN-BCO, VO-BCO conditions. Only the SV sequences showed the same pattern as theirs, which I suggested above might not truly reflect serial processing. It appears, then, that Chinese word processing is more parallel than English word processing (Yan et al., 2010; Yan & Sommer, 2015; Yang, 2009; Yu, Wu, & Gu, 2023). This could be because Chinese words are written adjacently, and two words span a very narrow region of space. That these nonlinguistic factors might be critical is supported by simulation showing that two short, not long, English words can be simultaneously identified (Reichle & Schotter, 2020). Empirically, a recent experiment by White (2023) showed that when two short English words are juxtaposed closer together, accuracy in the PC and NPC conditions will not be different, for many participants.

Furthermore White (2023) also showed that two English words are more likely to be processed simultaneously, even with wider distance in between, when the two words form a compound word (e.g., water fall). This is consistent with our findings and interpretation that a stored representation that encompasses the presented stimuli provides top-down facilitation for the two stimuli (in my case two characters) to be processed in parallel. Thus, we can infer from Experiment 1 that the VO-FCO, VO-BCO, and NN-

FCO strings used in this experiment all are MCUs that have stored representations.

I postpone a deeper discussion of the MCU hypothesis and a general conclusion about how many separate unrelated words Chinese readers can process, and how fast, until after Experiment 2 is presented. Experiment 2 will use 4-character strings as targets and attempt to address the puzzling pattern of SV sequences by adopting better control groups. Why 4-character strings? While it has been shown that in Chinese sentence reading information from up to 4 characters to the right of the current fixation is utilized (Yan et al., 2015), whether all characters within this perceptual span are recognized fully in parallel is yet unclear. Li et al. (2009) suggested that only 4-character strings that form single words (i.e., idioms) go through fully parallel recognition, but not strings of two two-character words, yet their experimental design did not provide readers optimal viewing position. Furthermore, as will be seen, morphosyntax in Chinese allows generation of infinite noun-noun, verb-object, and even subject-verb sequences, many of which frequently occur in daily conversation. It thus stands to reason that, like 2-character strings, these frequently co-occurring 4-character strings are eligible as stored MCUs and would be recognized fully in parallel as idioms would. Uncovering the existence of such units, which are in fact very common and nontrivial, is an important step toward a full picture of Chinese sentence reading.

2.3. Reading strings of three/four Chinese characters

While Section 1.1.2. reviewed studies on processing of 2-character strings (with or without sentence context), which is what the literature on Chinese reading mostly is about, many studies also have investigated reading of words or multiple-words consisting of three or four characters (Inhoff & Wu, 2005; Li et al., 2009; Wang et al., 2023; Yen et

al., 2012; Zhou & Li, 2021, to name a few). As introduced in Section 1.1.4., one type of two-character compound words is coordinate: two synonymous morphemes together mean one thing (e.g., 号码, DIGIT-CODE = number); another type is modifier head (e.g., 电话, ELECTRICITY-SPEECH = telephone). Using the modifier-head rule recursively, one can get 电话号码 (ELECTRICITY-SPEECH-DIGIT-CODE, telephone number), which is a four-character noun-noun MCU. Interestingly, despite its extremely frequent usage, this four-character string is not listed as a word in Sun et al.'s Chinese lexical database; instead, among the four-character words (3355 out of 48644 word entries) almost all are non-compositional/metaphorical idioms or proper nouns (e.g., 社会主义, SOCIETY-DOCTRINES = socialism). Whether or not Chinese readers do store 4-character compositional noun-noun compounds as MUCs can be addressed empirically.

Wang et al. (2023) extended the semantic implausibility paradigm (Yang et al., 2012) to three-character noun-noun compounds. Recall that Yang et al. (2012) had two-character compound nouns such as 守门员, DOOR-KEEPER preceded by verbs that either semantically fit or does not fit the first character of the compound noun (e.g., *kick* or *entreat*, but both verbs fit with the whole compound noun). Yang et al. found that this local semantic implausibility did not influence any stages of eye-movements on the compound words. This pattern was replicated using 4-character noun-noun compounds (e.g., 海洋生物, SEA-OCEAN-LIVING-THING = *sea creature*, Wang et al., 2023), where no local implausibility effect was observable at any stage of reading on the 4-character strings (e.g., the two characters SEA-OCEAN were read similarly fast whether they were preceded by PROTECT or by EAT). Moreover, the study additionally manipulated the novelty of the four-character strings, including novel noun-noun

compounds such as (海洋植物, SEA-OCEAN-PLANT-THING = *sea plant*). It was found that, with novel 4-character noun-noun compounds, a local implausibility effect emerged as early as first-fixation duration on the first two-characters (e.g., *protect sea plant* vs. *eat sea plant*; the latter case was read slower), where mean first fixation duration on *sea* was longer in the latter than that in the former. These findings are consistent with my Experiment 1's results reported above, where fully parallel recognition was observed for NN-FCO compounds but not for NN-BCO (novel) compounds. Based on Wang et al. (2023, see also Zhou and Li 2021 for the same paradigm and findings with 3-character noun-noun compound nouns), it is predicted that fully parallel recognition would be observed for 4-character-NN-FCO compounds in Experiment 2 here. Other studies, using the boundary-change paradigm, also have shown three- or four-character strings are processed together as one single unit (Yu et al., 2016; Zang et al., 2021; Zang et al., 2023; Zang et al., 2024), but I will introduce them later in the next chapter due to their closer relevance to Experiment 3.

Whether fully parallel recognition is generalizable to 4-character *verb-object* sequences is slightly less clear. In Experiment 5 of Li et al., (2009), unmasked 4-character strings were presented (23.5 ms) to elicit whole-string reports. Three types of stimuli were used: idiomatic one word, two related 2-character words, and two unrelated 2-character words. Items in the semantically related two-word condition were adjective-noun or adverb-verb sequences (e.g., 美满婚姻, HAPPY-MARRIAGE = happy marriage; 突然来临, SUDDENLY-HAPPEN = suddenly happen). There, it was found that recognition accuracy for the first two characters was around 90% for all three types of stimuli. For the last two characters, accuracy was respectively about 85%, 60%, and 20%,

for idioms, related words, and unrelated words. This was suggested by the authors to indicate that adjective-noun and adverb-verb sequences must be serially processed. This early piece of evidence suggests that fully parallel recognition of 4-character strings might be limited to noun-noun compounds (Wang et al., 2024 above) or idioms (Li et al., 2009).

Recall, however, strings in this study were placed to the right of the participants' fixation, rather than centered around the participant's fixation. The accuracy drop thus could have been due to the worse visual acuity at the edge. Experiment 2 addresses whether all characters in 4-character strings can be recognized simultaneously, comparing noun-noun, verb-object, and subject-verb sequences, the latter two of which are under some theories distinctly two words because they both consist of an argument and an argument introducer and also because they are separable by indefinite number of characters/words (see Section 1.1.4). In Experiment 1, we saw results suggesting that subject-verb sequences were processed serially with a confounding factor of semantic implausibility. Experiment 2 includes subject-verb sequences as one of the target conditions in which all items are semantically plausible.

2.4. Experiment 2: recognition of 4-character noun-noun compounds, verb-object sequences, and subject-verb sequences.

2.4.1. Participants

Participants were recruited from the same participant pool as in Experiment 1. In total, 57 participants were recruited. They received \$30 US dollars for 105-min of participation. 17 participants had difficulty fixating centrally throughout the trials (some due to the lenses of their glasses being too reflective for calibration; some blinked too

much) and felt too tired to continue. In the end, 40 participants' data were analyzed (mean age= 23 yr; 6 males, 1 non-binary, and 33 females). See Supplementary Material for demographic and language-background summary. Among these 40 participants, 3 of them stopped at six-sevenths of the experiment because they ran out of time. Given that these participants still had a great number of trials run, their data were kept. Among the 40 participants analyzed, 14 had participated in Experiment 1, but the two experiments were conducted at least one month apart. These participants were not debriefed immediately after Experiment 1.

2.4.2. Materials

Five-hundred-sixty 4-character strings were created, as shown in Table 6. Critical stimuli were noun-noun (NN, 112 items), verb-object (VO, 112 items), and subject-verb (SV, 112 items) sequences. In general, each noun/object/subject consists of two noun characters (e.g., 底线, BOTTOM-LINE = bottom line) that together refer to one entity and each verb consists of two verb characters that together refer to one action/event (e.g., 查看, CHECK-LOOK = examine). Thus the POS structures for the three sequences *generally* are NNNN, VVNN, and NNVV. In practice, since I will use only the leftmost and rightmost character as my loci of interest¹⁵, the second character and third character's POSs were not strictly selected to follow these patterns. For example, the third character in NN and VO may be an adjective (e.g., 新年, NEW-YEAR = new year; 公园, PUBLIC-GARDEN = park), and the second character in VO may be an adverb/particle (e.g., RUSH-TO = rush to). These are acceptable as long as the two characters together

¹⁵ This is because visual acuity is lowest on the edges. The pre-cueing effect, if any, should be greatest, for the outer two characters. I will therefore only look at the leftmost and rightmost characters as my loci of interest.

unambiguously form a noun or a verb and globally the whole sequences are clearly NN, VO, or SV.

	Number of items (after ambiguity exclusion)	WF	F1	F4	F12	F34	S1	S4	S1S2	S3S4
Noun-Noun (NN)	95	2.93	10.5	10.5	7.91	7.79	7.72	8.58	15.3	16.2
Verb-Object (VO)	101	2.92	10.5	10.6	7.86	7.73	8.38	8.19	15.7	15.8
Subject-Verb (SV)	104	2.91	10.6	10.5	8.05	7.75	8.1	8.02	15.8	16.7
VNVV	51	0	9.49	10.4	0.69	6.3	9.8	9.49	18.3	18.8
VVNV	52	0	9.12	9.6	6.35	0.63	9.35	9.87	19.3	18.3

Table 6. Summary of frequency and stroke measures for each item group in Experiment 2. WF: frequency of the whole string; F1: frequency of the first character; F4: frequency of the fourth character; F12: frequency of C1C2; F34: frequency of C3C4; S1: number of strokes of the first character; S4: number of strokes of the fourth character; S1S2: number of strokes of C1C2; S3S4: number of strokes of C3C4. Frequency measures are natural-log transformation of raw counts and were based on the combined HuanQiuRenWu and Weibo corpus I created (105 million characters).

Included as control stimuli were VVNV (56 items) and VNVV (56 items). Inclusion of these items ensured that participants could not make strategic guesses (e.g., recognizing the first two characters being a verb means the last character definitely is a noun, if only NN, VO, and SV were used). These two types of sequences were nonwords and non-MCUs, and were generated by replacing one character from a plausible co-occurring 4-character string. For instance, a VNVV is 杀程佩戴 (KILL-PROCEDURE-CARRY-WEAR), which was derived from a real phrase 全程佩戴 (FULL-PROCEDURE-CARRY-WEAR), which means to wear (something) throughout the process. By doing so, the strings were globally nonsensical but maintained certain local

associations among the embedded characters. The whole sequences thus had clearly more than one unit but were not easily distinguishable from the true MCU targets, preventing participants from further using heuristic strategies in the experiment (i.e., simply focusing on whether there is some relationship between the second and third characters).

Another 56 items were included with a POS structure of ANNA, to avoid participants monotonously seeing only nouns and verbs throughout the experiment. As in Experiment 1, these items were not analyzed due to the ambiguity of adjectives in Mandarin as adjectives and predicating verbs. Finally, 56 items (with NNNN, VNVV, and VVNV structures) were included as fillers. For these items, the character probed was either the second or the third character, contrary to the other 504 items with which either the first or the fourth character was probed. These items were simply added to make the reading task more naturalistic, and therefore not analyzed in the end.

Some characters appeared more than once in the 560 strings (maximum of 5 times) and some two-character strings appeared more than one (maximum of 3 times), but no full stimulus string occurred twice. Mean frequency and number of strokes of each character were matched across groups. For frequency measurement, I created my own *unparsed*, combined corpus of web text scraped from HuanQiuRenWu (Global People Magazine) and from Weibo (see Appendix B for detail about the corpus). The corpus totals 105 million characters, whose unigrams' (one-character strings), bigrams' (two-character strings), and 4-grams' (four-character strings) frequency was extracted. I further checked the frequency against the PKU CCL unparsed corpus¹⁶ (Zhan et al., 2019, which

¹⁶ I did not directly use the PKU CCL corpus since the search engine only provided the number of documents containing the queried character strings, rather than the actual counts of appearance. See Appendix D for CCL frequency measures.

was updated in 2024, with 4.75 billion characters, after Experiment 2's material had been created). The two corpora showed very consistent frequency measures.

As in Experiment 1, among all items in NN, VO, and SV, whole-string frequency ranges widely from FCO to BCO strings. Moreover, in this experiment, rather than using first/last quartiles to subgroup FCO and BCO strings in a post-hoc fashion, I intentionally included 32 (one-fourth of all the items) *extremely uncommon* but plausible strings in each string structure (e.g., 月饼拼盘, MOONCAKE-PLATTER for NN; 等待特权, AWAIT-PRVILIGE for VO; 男性颤抖, MALE-SHIVER for SV), to contrast them with the 32 (one-fourth, first quartile) top frequent items in the material.

Table 7 shows the frequency and stroke measures of each subgroup. Note that the properties of the sub-constituents unfortunately were not as well controlled as in Experiment 1. Specifically, the FCO groups appear to also have systematically, but only slightly, higher frequency of the embedded characters (see Section 2.4.6. on how this might not be a concern).

	Number of items (after ambiguity exclusion)	WF	F1	F4	F12	F34	S1	S4	S12	S34
NN-FCO	24	4.82	10.7	11.4	8.51	8.53	8	7.62	15.6	15.8
NN-BCO	27	0	10.2	9.68	7.08	7.15	7.3	9.15	15.6	17.6
VO-FCO	30	4.69	10.5	10.8	8.04	8.10	8.43	7.83	15.2	15.3
VO-BCO	28	0	10.6	10.2	7.83	6.98	8.5	8.64	16.4	16.6
SV-FCO	33	4.63	11.0	10.4	8.75	8.19	8.97	8	16	16.9
SV-BCO	29	0	10.1	10.4	7.34	7.30	8.07	6.97	15.1	16.3

Table 7. Summary of frequency and stroke measures for each item subgroup in NN, VO, and SV groups. WF: frequency of the whole string; F1: frequency of the first character; F4: frequency of the fourth character; F12: frequency of C1C2; F34: frequency of C3C4; S1: number of strokes of the first character; S4: number of strokes of the fourth character; S1S2: number of strokes of C1C2; S3S4: number of strokes of C3C4. Frequency measures are natural-log transformation of raw counts.

Efforts were made to include only characters that are POS-unambiguous, first based on the author’s intuition. Each character was then normed separately from the main experiment (see the end of Section 2.2.3 for the procedure), and items with a character whose POS is not unambiguous among Chinese readers were discarded post-hoc and not included in the analysis (see Section 2.4.4).

2.4.3. Procedures

The physical setting was exactly the same as in Experiment 1. The procedures (Figure 6) were by and large similar to Experiment 1 except for a few minor differences, noted below.

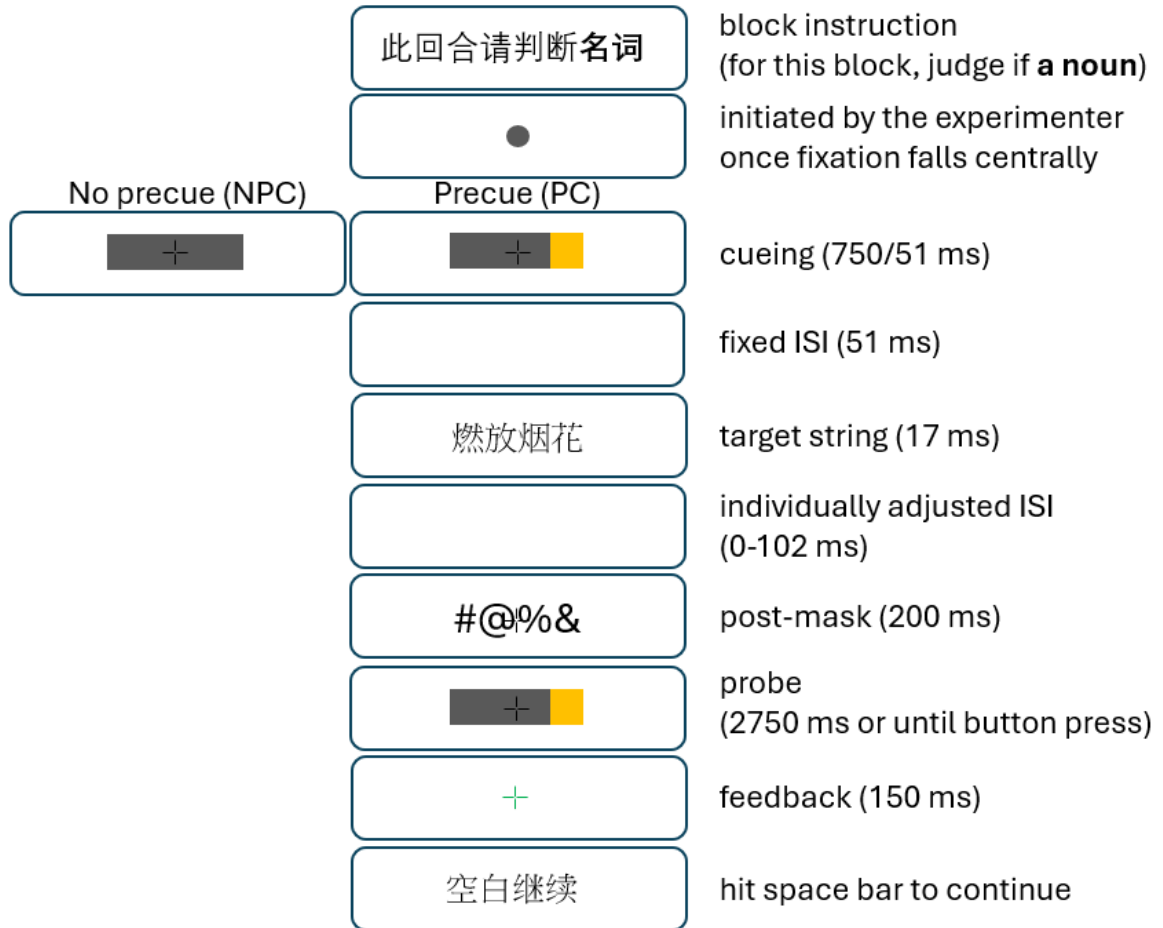


Figure 6. Illustration of the procedure for a trial in Experiment 2. The block instruction on the top of the figure only occurs once at the beginning of each block. Note the ISI between cueing squares and the target string was initially set to 750 ms and changed to 51 ms after 15 participants had been run (see explanations in the main text).

First, when participants were told to make the POS judgment solely for the probed character and ignore any context, they were additionally told that a certain proportion of the trials will contain nonsensical strings. This was not in the instruction in Experiment 1 and was changed as such in the hope of reducing any potential response bias. Second, the staircase thresholding stage before the official experiment was reduced to 8 blocks, rather than 10 blocks, to avoid the experiment being too long. Third, immediate abortion of trials due to non-central fixation occurred only if the fixation was 0.625° to the left or

right and 1.25° above or below the center of the screen (instead of the $0.5^\circ/1^\circ$ criterion used in Experiment 1). However, if the fixation was 0.5° - 0.625° to the left/right and 1° - 1.25° above/below the center of the screen, the trial continued but was marked and later discarded prior to analysis.

Finally, the duration of the pre-cueing squares was initially set to the same as in Experiment 1 (750 ms). After 15 participants were run, it was found that participants had a hard time fixating centrally, more so than participants in Experiment 1. Trials aborted due to non-central fixation accounted for 5.7% (PC) and 4.3% (NPC) in Experiment 1, yet trials aborted due to non-central fixation accounted for 11.2% (PC) and 8.7% (NPC) for these 15 participants. Given this high rate, a concern was that greater effort was needed to suppress saccade programming when the stimuli (both the cueing squares and the characters) were long. This could have impaired the perceptual processing during the presentation of the character string and the judgment, specifically in the PC condition: the participant might have ended up totally ignoring the yellow pre-cue (or paying attention exclusively to the center), as even paying covert attention to the edge frequently led to unwanted saccade and trial termination. I therefore shortened the duration of the cueing squares from 750 ms to 51 ms. At this rate, participants still would be able to sense the yellow square and utilize the information (Talgar, Pelli, & Carrasco, 2004) but arguably would barely have enough time to program a saccade (Kowler et al., 1995). Indeed, non-fixation rates dropped to 7.1% (PC) and 6.6% (NPC) for the rest of the 25 participants with 51 ms. Since the two sub-datasets do not differ qualitatively in their results (see Section 2.4.5), I will analyze Experiment 2 combining all 40 participants.

There were 28 official blocks, each with 20 trials. As in Experiment 1, every four-

block set constituted a run (PC-noun, PC-verb, NPC-noun, and NPC-verb); within each run the order of the blocks was randomized. In each block, there were 4 NN, 4 VO, 4 SV, 2 VNVV, 2 VVNV, 4 fillers. Within a block, 9 trials were with a leftmost probe (first character), 9 with a rightmost probe (fourth character), 1 with a second-character probe, and 1 with a third-character probe, with the order randomized. The same staircase thresholding method was adopted in the official blocks as in Experiment 1 after every run (4 blocks).

As in Experiment 1, for each participant, each item only appeared once (either with a left/right/leftmost/rightmost probe, a noun or verb prompt, PC or NPC) throughout the experiment. In which block an item appeared was also balanced (early or late in the experiment). In total, there were 16 Latin Square counterbalance lists administered.

The whole experiment took on average 105 minutes. After the experiment was over, participants completed an offline POS judgment task on a spreadsheet where each row contained one single character. They were asked to type out, as quickly as possible, whether the character is a noun, a verb, or an adjective on the next column and not overthink. If they found it hard to decide the POS of the character, they should leave the cell blank.

2.4.4. Analysis

All participants' accuracy in the PC condition was above 70%, which is expected under the staircase thresholding method. Analysis was almost identical to that in Experiment 1, except that here I analyzed data on the first and fourth characters.

Across the string structures, about 7-15% of the items containing a character that was not highly unambiguous regarding their POS, based on the offline judgment task, were

further discarded before analysis, using the same criteria described in Experiment 1 (Section 2.2.4). VNVV and VVNV's data were aggregated together as controls.

As in Experiment 1, in addition to AOC plotting, statistical tests were performed to eventually determine whether there was a difference in accuracy between the PC and NPC conditions, as well as whether string structure and co-occurrence frequency modulate the effect of pre-cueing. Two power analyses, using simulation, were conducted for this statistical analysis, one without considering the factor of co-occurrence frequency (a model involving only pre-cueing \times structure) and the other with it (pre-cueing \times structure \times frequency). The former assumes that noun-noun sequences are MCUs that can be recognized as single units and thus their recognition accuracy is not subject to pre-cueing while verb-object/subject-verb sequences are multiple-words (Huang, 1984; Sybesma, 1999) that must be serially recognized. The latter assumes that *only high-frequency noun-noun sequences* are MCUs whose recognition accuracy is not subject to pre-cueing: an interaction between pre-cueing and frequency for noun-noun items and a further interaction among pre-cueing, frequency and structure. Simulation parameters were the same as in Experiment 1 except that the number of trials in each condition was based on the design of Experiment 2. Details for the simulation are reported in Appendix A.

The first simulation showed a 92% chance of obtaining a significant two-way interaction when all items were included. The second simulation showed a 21% chance of obtaining a significant three-way interaction, a 22% chance of obtaining a significant two-way interaction between pre-cueing and string structure, and a 29% chances of obtaining a significant two-way interaction between pre-cueing and frequency for NN

items. The very low power for the second analysis is likely due to the subgrouping based on first and last quartile co-occurrence frequency. As in Experiment 1, in analyzing the results, due to the high power to detect a two-way interaction when including all the NN and VO items, I first ran a regression model with the full dataset. Then, due to the low power with the subsets of items, I ran multiple simple regression models that included only the main effect of pre-cueing, each time at a particular level of frequency and structure (e.g., with only trials of infrequent NN strings), instead of running one model that included all three factors (frequency \times structure \times pre-cueing). The limitations of this statistical practice are noted in Sections 2.5.4 and 5.1.

2.4.5. Results

2.4.5.1. Duration of pre-cues had little qualitative effect

As noted in Section 2.4.2, the duration of the pre-cue was changed from 750 ms to 51 ms after 15 participants had been collected. Here I call the experiment with 750 ms of cueing squares 2A, and the experiment with 51 ms of cueing squares 2B. I first show that there are no detectable qualitative differences between 2A and 2B in terms of accuracy or RTs.

The grand mean of accuracy (across string structures) in the PC and NPC conditions in 2A were 81.5% vs. 77.7% and in 2B were 81.8% vs. 76.9%. As in Experiment 1, the accuracy in the PC condition fell close to 80%, suggesting that the staircase thresholding worked as expected. Figure 7 shows the results by cueing condition, side, and string structure. The only notable difference between 2A and 2B is in the controls (VNVV-VVNV). Overall, it is clear that all three target conditions (NN, VO, and SV) had NPC accuracy well above the limited-capacity parallel curves.

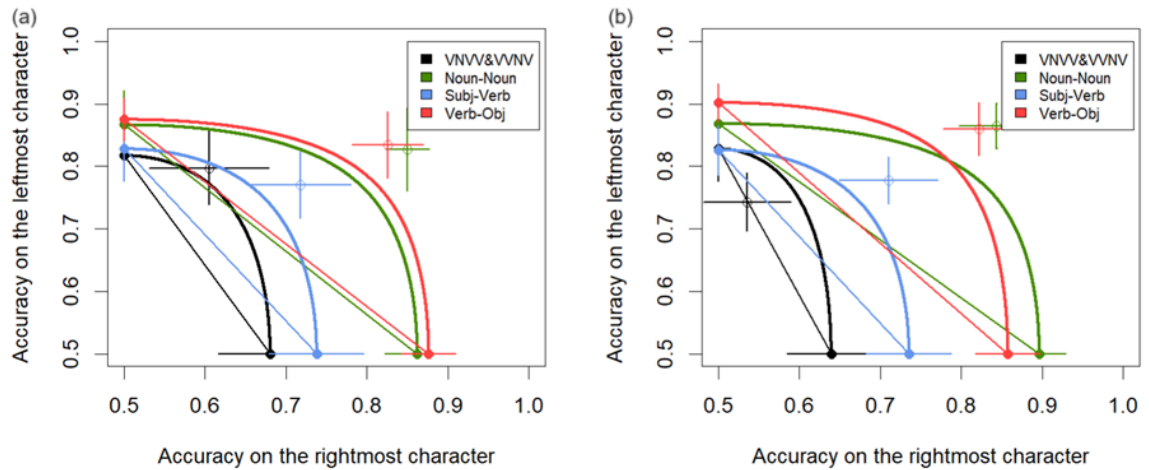


Figure 7. AOC plots in Experiment 2A and 2B. (a) AOC plot with the first 15 participants (750 ms of cueing squares). (b) AOC plot with the last 25 participants (51 ms of cueing squares). Error bars reflect 1.96 * by-subject standard errors.

For RTs, I ran a Bayesian linear regression model, with log-transformed RTs as the dependent variable and cueing condition and sub-experiment, as well as their interaction, as fixed effects. Table 8 showed that, while cueing had a reliable effect (972 ms vs. 1200 ms) and participants in 2B tended to have faster RTs ([-0.22, 0.02]), the interaction between cueing and sub-experiment had a coefficient estimate largely centered around zero. This suggests that the efficiency of the cues was comparable between 2A and 2B.

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
Intercept	6.88	0.03	6.82	6.94
A_B	-0.10	0.06	-0.22	0.02
PC_NPC	0.20	0.02	0.17	0.23
AB × PC_NPC	0.02	0.03	-0.04	0.07

Table 8. Bayesian linear regression model of RTs in Experiment 2. Only fixed effects estimates are shown here. R formula: $\log(\text{RT}) \sim A_B * PC_NPC + (1+PC_NPC|subj) + (1+A_B * PC_NPC|item)$; Sub-experiment A was coded as -0.5 and sub-experiment B was coded as 0.5; PC was coded as -0.5 and NPC coded as 0.5. Chains =4, each with 8000 iterations, warmup = 4000. The default priors in the brms package were used. All Rhats = 1.00.

Given these overall qualitative similarities between the two sub-experiments, I will report the results combining data from the two sub-experiments (total N= 40) in the rest of this section.

2.4.5.2. Main results

Figure 8 shows the distributions of ISI between the target string and the post-mask for the first run and last run of the official blocks, as well as the distribution of ISI throughout Experiment 2 across all runs. Similar to Experiment 1, it was found that for the majority of Chinese readers to correctly judge one of the two characters' POS, it only requires 17 ms of presentation of the two-character string and 0-34 ms of blank frame before the 200 ms post-mask.

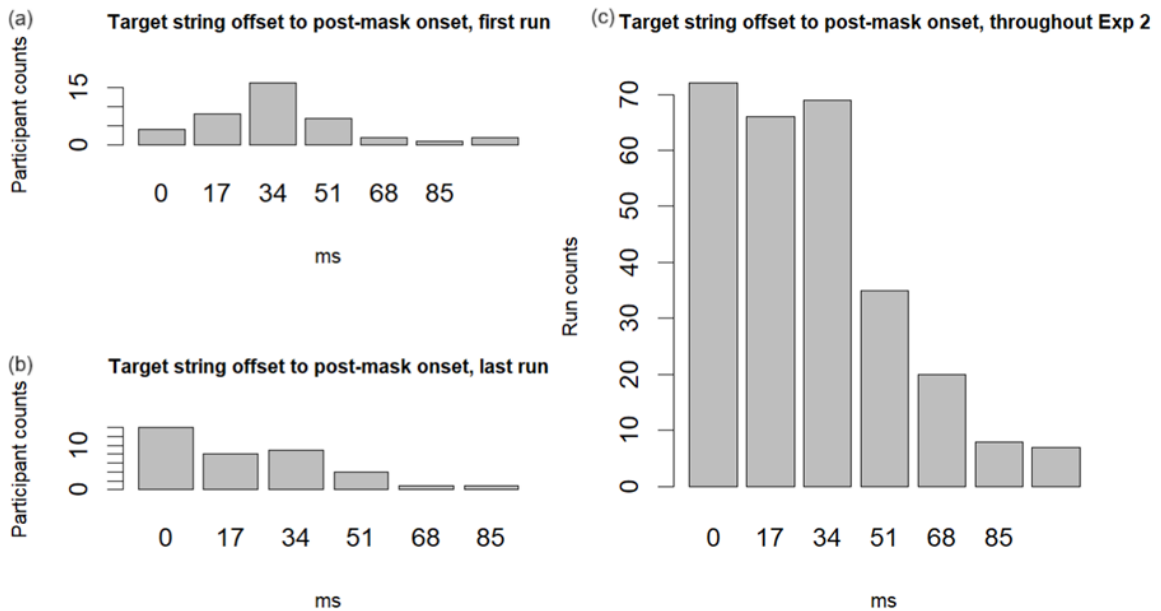


Figure 8. Distributions of ISI duration for the first run in Experiment 2. (a); for the last run in Experiment 2 (b); for all the runs in Experiment 2 (8c). Y-axis differs between (c) and (a)(b) because each participant had 7 runs in total.

Figure 9 shows the results by cueing condition, side and string structure in an AOC plot. The predictions (the diagonal line and the curve) are plotted individually for each

string structure, based on the accuracy in the PC condition.

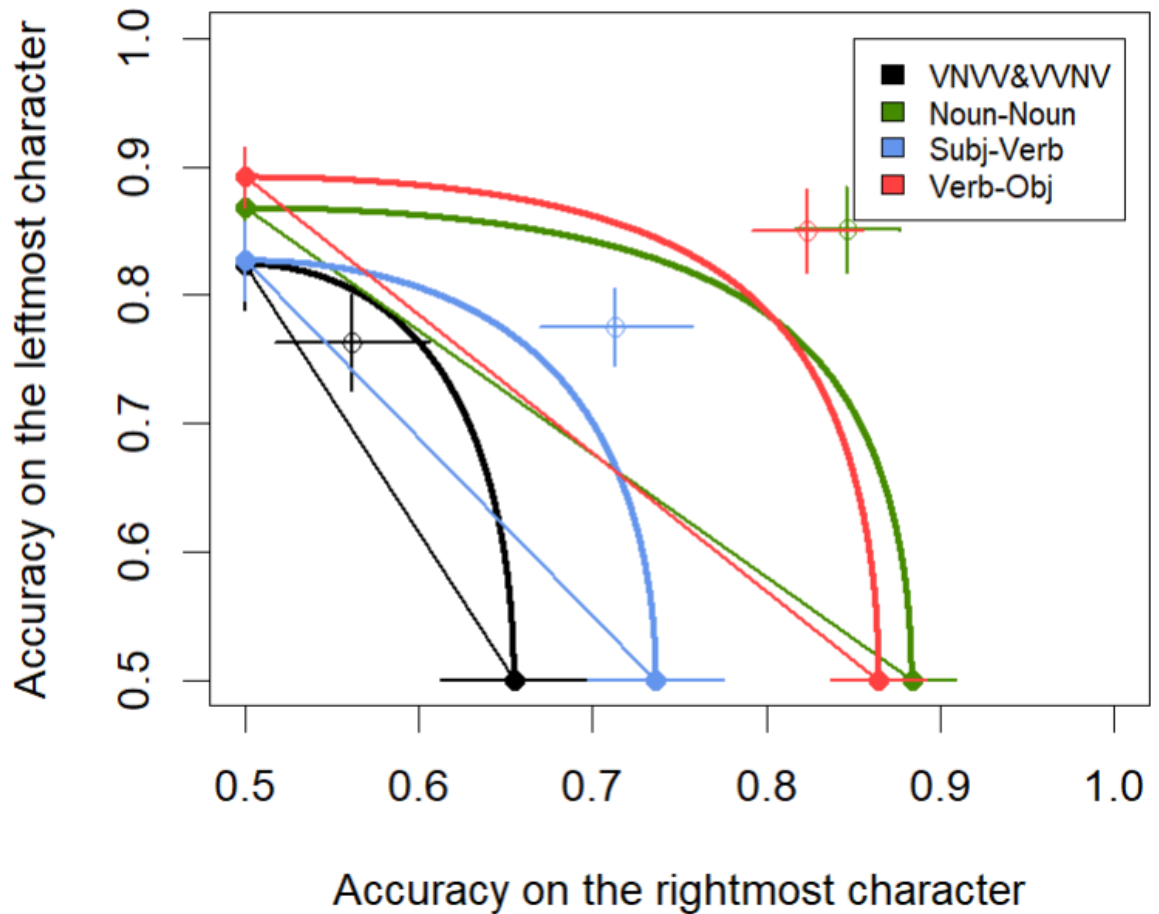


Figure 9. Accuracy in Experiment 2 based on cueing condition, side, and string structure plotted on an AOC plot. Error bars reflect $1.96 \times$ by-subject standard errors.

As is shown in Figure 9, NPC accuracy in NN (green), VO (red), and SV (blue) sequences is clearly above the prediction of the limited-capacity parallel model. In contrast, nonword controls (combining VNVV-VVNV) have NPC accuracy much lower than their PC baseline, falling closer to the all-or-none serial prediction.

Contrary to Experiment 1, parallel processing does not hinge on the accuracy in the PC conditions. That is, while SV sequences had lower accuracy overall in the PC condition compared to NN and VO sequences in the PC conditions, their NPC accuracy

still fell well above the limited-capacity parallel curve.

As in Experiment 1, a Bayesian logistic regression model on accuracy (aggregating across left and right sides) was run to determine whether there was a reliable difference in accuracy between the PC and NPC conditions and whether the effect was modulated by string structure. Table 9 shows the results of the Bayesian logistic regression model. It is shown that, while visually on the AOC plot the NPC points fall close to the intersection of PC accuracy, statistically, accuracy in NPC is reliably lower than that in PC (81% vs. 84.3%). Importantly, there is very little evidence showing that this pre-cueing effect differs between NN and VO items (84.9% vs. 87.5% for NN, 83.7% vs. 87.8% for VO, and 74.4% vs. 78.1% for SV).

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
Intercept	1.78	0.09	1.60	1.96
NPC_PC	-0.26	0.08	-0.42	-0.11
NN_VO	0.03	0.11	-0.19	0.25
SV_VO	-0.67	0.11	-0.89	-0.44
Interaction (NN_VO)	0.10	0.15	-0.20	0.41
Interaction (SV_VO)	0.12	0.14	-0.15	0.39

Table 9. Bayesian logistic regression models of accuracy for NN, VO, and SV conditions. Only fixed effects estimates are shown here. R formula: Correctness ~ NPC_PC * (NN_VO+SV_VO) + (1+NPC_PC * (NN_VO+SV_VO)|subj) + (1+NPC_PC|item); PC are coded as -0.5, NPC coded as 0.5, VO as (-1/3, -1/3), NN as (2/3, -1/3), and SV as (2/3, -1/3). Chains =4, each with 8000 iterations, warmup = 4000. A logit link function was used and the default priors in the brms package were used.

Recall that among all items in NN, VO, and SV, whole-string frequency ranges widely from FCO to BCO strings. Figure 10 shows the results by cueing condition, side and string structure in two AOC plots (separately for FCO and BCO strings). Figure 10a shows that accuracy in NPC was far above the limited-capacity parallel curve for all

structures of FCO strings. In contrast, Figure 10b shows that for BCO strings, only NN sequences' accuracy in NPC fell far above the limited-capacity parallel curve. For VO sequences, accuracy was near the curve, and for SV sequences, accuracy was between the curve and the all-or-none serial diagonal line.

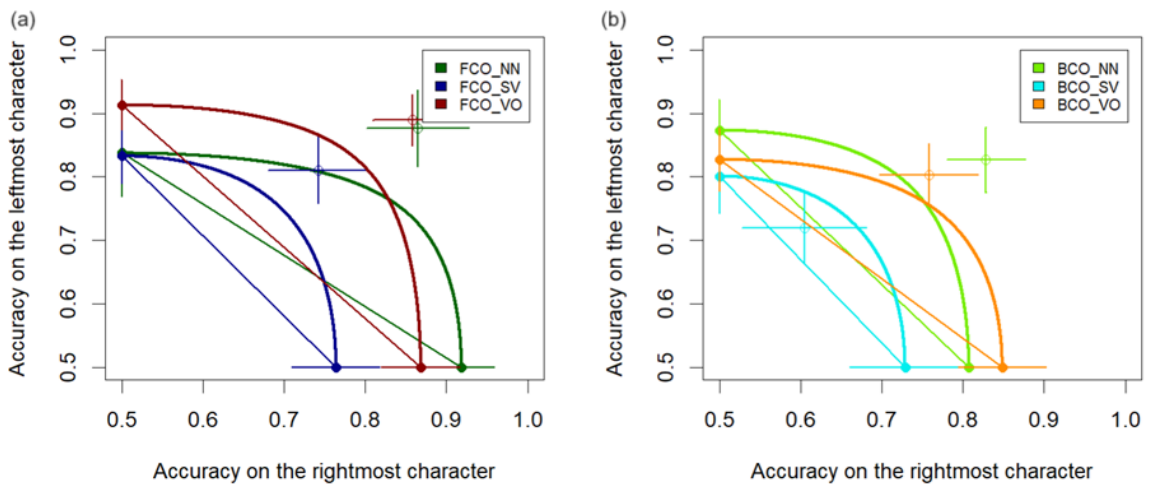


Figure 10. AOC plots in Experiment 2. (a) Accuracy of FCO strings in Experiment 2 based on cueing condition, side, and string structure plotted on an AOC plot. (b) Accuracy of BCO strings in Experiment 2 based on cueing condition, side, and string structure plotted on an AOC plot. FCO: strings of frequently co-occurring characters; BCO: strings of barely co-occurring characters. Error bars reflect 1.96 * by-subject standard errors.

Bayesian logistic regression models were run to determine how fully in parallel the characters in each subgroup are processed (Table 10). The models showed that for all FCO strings, there was no evidence of a difference between the PC and NPC conditions ($[-0.57, 0.47]$ for NN, $[-0.54, 0.62]$ for VO, and $[-0.44, 0.26]$ for SV), indicating fully parallel processing of all FCO strings. BCO sequences, on the other hand, showed differences among the three structures. For NN, there was no evidence of a difference between the PC and NPC condition ($[-0.46, 0.46]$). For VO, the Bayesian model suggests a tendency of accuracy lower in the NPC than in the PC condition, with the 95% credible

interval mostly lying at the negative values ([-0.67, 0.10]), despite overlapping zero.

Finally, for SV, it was clear that accuracy in the NPC condition was statistically lower than that in the PC condition ([-0.81, -0.15]).

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
NN-FCO				
Intercept	2.19	0.23	1.78	2.70
NPC-PC	-0.04	0.27	-0.57	0.47
NN-BCO				
Intercept	1.74	0.19	1.39	2.13
NPC-PC	-0.01	0.23	-0.46	0.46
VO-FCO				
Intercept	2.32	0.22	1.94	2.79
NPC-PC	0.03	0.29	-0.54	0.62
VO-BCO				
Intercept	1.66	0.16	1.37	2.00
NPC-PC	-0.29	0.19	-0.67	0.10
SV-FCO				
Intercept	1.52	0.18	1.18	1.88
NPC-PC	-0.09	0.18	-0.44	0.26
SV-BCO				
Intercept	1.28	0.18	0.93	1.66
NPC-PC	-0.47	0.17	-0.81	-0.15

Table 10. Bayesian logistic regression models of accuracy for NN-FCO, NN-BCO, VO-FCO, and VO-BCO conditions. Only fixed effects estimates are shown here. R formula: Correctness ~ PC_NPC + (1+PC_NPC|subj) + (1+PC_NPC|item); PC are coded as 0 and NPC coded as 1. Chains =4, each with 8000 iterations, warmup = 4000. A logit link function was used and the default priors in the brms package were used. All Rhats = 1.00.

Since SV strings are the ones whose two separate (FCO & BCO) models' results differed, I ran a model combining SV-FCO and SV-BCO strings that directly tested the interaction. The model showed a hint of an interaction ([-0.11, 0.84], see Appendix C for the full model results), supporting that SV-FCO strings were processed more in parallel than SV-BCO strings.

2.4.6. Discussion

Experiment 2 aimed at addressing the same questions as Experiment 1 with material consisting of 4-character strings. It also further investigated recognition of subject-verb sequences that are semantically plausible. The experiment design showed sensitivity to the item manipulation in the expected direction: four-character controls that form neither single words nor MCUs were processed in quite a serial fashion, while four-character strings that are semantically plausible were clearly processed in parallel, although not fully parallel. Further subgrouping based on whole-string frequency showed that all FCO strings were processed fully in parallel regardless of their morphosyntactic structures. On the other hand, for BCO strings, a gradient pattern of parallelism was revealed, with NN strings showing full parallelism, VO strings showing only little evidence of full parallelism, and SV strings showing strong evidence *against* full parallelism (Figure 10b and Table 10). These findings extend the results from Experiment 1, indicating the existence of stored MCUs not only for 2-character frequent NN and VO strings but also for 4-character frequent NN, VO, and even SV strings.

Before further discussing/interpreting the results, I first address one noticeable potential confound in Experiment 2: the differences in properties of the embedded constituents among structures and among frequency subgroups. Specifically, it was clear that within each of the three structures, the frequent subgroups not only differed in the whole-string frequency but also in almost all other frequency measures¹⁷ (note that while whole-string frequency very likely naturally correlates with their embedded constituents' frequency to some degree, the correlations seen in my items were mostly due to my own

¹⁷ Stroke measures had this similar issue, although to a lesser extent (they are on raw scale, not log scale) and for some in an opposite direction to frequency measures (i.e., high frequent groups tend to have *more* strokes for SV). I therefore do not consider it to be a confound.

oversight. It is in theory possible to dissociate the two frequencies).

I argue that the differences in embedded constituents' frequency should not have contributed to the processing differences between FCO and BCO groups (Figure 10a vs. Figure 10b) nor to the processing differences among the three structures of BCO strings (Figure 10b), for the following reasons. First, among the three structures, the NN sequences are the group whose C1, C4, C1C2, and C3C4 frequency differs between the FCO and BCO subgroups (for C1, $p = .1$; for the other three, $ps < 0.0005$). However, NN sequences are the group that showed the most similar processing patterns between the FCO and BCO subgroups. Second, when the three structures were compared within the FCO strings and within the BCO strings, it was within FCO strings that embedded constituents' frequency differed (for C1, $p = 0.09$, for C4, $p < .005$, for C1C2, $p = 0.08$, for C3C4, $p > .1$), not within BCO strings. However, the processing patterns were very similar among the three structures within FCO strings. Therefore, processing patterns and embedded frequencies appear to be dissociated, and it is more probable that the observed differences in processing patterns were due to the differences in whole-string structure and frequency.

Unlike Experiment 1, it was found that parallel processing in the NPC condition did not rely on accuracy in the PC condition. Within the PC condition, accuracy differed quite substantially between SV sequences and NN/VO sequences. Nevertheless, SV sequences still showed more than limited-capacity parallel processing (all items combined) and fully parallel processing (FCO strings only), as NN and VO sequences did. This confirms the validity of the current task to address serial versus parallel processing.

The results of FCO strings contrasted with Li et al. (2009), where 4-character strings whose first two characters and last two characters are related, and in fact co-occur frequently (e.g., 美满婚姻, HAPPY-MARRIAGE = happy marriage; 突然来临, SUDDENLY-HAPPEN = suddenly happen) were suggested to be processed serially. Here I examined strings (VO & SV) that were even more likely to be categorized as multiple-words given that they contain both a verb and an argument (Huang, 1984; Sybesma, 1999). That FCO strings of various structures had accuracy compatible with the unlimited-capacity parallel prediction can be easily explained by all of them being stored as MCUs: The stored MCUs have nodes at the word level which can be partially activated and interactively provide feedback to the character level activation. When frequent enough, it competes more strongly and can win over the embedded 2-character word nodes for recognition. That stored MCUs can be of various string structures provides strong support for usage-based language learning (Bybee, 2006) and against a morphosyntax-based, categorical distinction between single-words and multiple-words (Zang, 2019; Contreras Kallens & Christiansen, 2022).

The results for BCO strings were more complicated and diverged a little from Experiment 1's BCO results. I therefore discuss them together in the next section.

2.5. General Discussion for Experiment 1 and Experiment 2

2.5.1. Recognition of (extremely) infrequent strings

A discrepant finding about BCO strings between Experiment 1 and Experiment 2 is that in Experiment 1 NN-BCO strings were not processed fully in parallel while in Experiment 2 they were. Not only were NN-BCO strings in Experiment 2 longer but they were also *extremely* infrequent strings. Since it is not possible that the participants have

stored representation for these strings, other mechanisms that might have given rise to this seemingly full parallelism should be considered. One possibility is the POS congruency effect, or the flanker effect (Eriksen & Eriksen, 1974). That is, when task-relevant features are processed in parallel, they can facilitate each other if they are of the same category (but see White et al., 2020 for *selection errors* or *response biases* as an explanation for a congruency effect that does not rely on parallel processing). This effect was recently reported with Dutch material (an alphabetic writing system with white space) in a POS judgment task with rapid parallel presentation, similar to the current paradigm (Snell, 2024), which was argued to reflect parallel word processing. In Experiment 2 then, it is possible that all characters are processed in parallel to a certain degree, and with evidence of presence of a noun accumulating from each of the multiple characters, the probability of reporting any of the characters being a noun increases and boosts the performance to fully parallel-like. This congruency effect might be especially strong in Experiment 2, since there are *four* characters providing bottom-up input. In theory, however, the two-character NN-BCO sequences in Experiment 1 should also have benefited from the congruency to show a fully parallel processing pattern. It remains unclear why it was not the case in Experiment 1 (see Section 2.5.4 for statistical concerns).

For VO-BCO sequences, in both experiments, evidence is mostly consistent with fully parallel processing, although in Experiment 2 the credible interval for NPCvsPC difference had most density at negative values [-0.67, 0.10]. This different pattern—hint of limited parallelism in E2 and hint of full parallelism in E1—is expected since strings were shorter and were not extremely uncommon in Experiment 1.

Now consider just the BCOs within Experiment 2, which were well matched across the three structures. Here, structure had a gradient effect, with the degree of parallelism being $NN > VO > SV$. The former difference could have been a POS congruency effect, as explained above. Putting that speculation aside, what is clear is that SV-BCO sequences showed a qualitative difference from NN- and VO-BCO sequences. One possible driving factor is semantic relatedness, or meaningfulness of the string (Jolsvai et al., 2020). I consider semantics unlikely to be a critical factor given that all strings are *extremely* infrequent and all can be composed into meaningful entities/events, but this possibility can be evaluated in the future by quantifying the relatedness between the two words within each item and examining whether the group means in semantic relatedness follows the empirical pattern ($NN > VO > SV$). Another more plausible explanation is the structure effect itself. Since there are no stored high-level (4-character) representations for these strings, the readers must first deal with the two 2-character words separately (arguably in parallel to a certain extent). The differences in accuracy among the three structures in the NPC condition suggest that, in addition to purely recognizing each part of the whole string, participants must have also been engaged in composing the two parts into a bigger part. This composition can be semantic, relational/thematic, or morphosyntactic/structural, or all three. One could interpret the difference as reflecting that understanding who the agent in a verb phrase is harder than understanding who the patient in a verb phrase is (Cohn & Paczynski, 2013). Another interpretation is that integrating a noun as a subject of a verb requires more processing than integrating a noun as an object to a verb, consistent with the theory of external versus internal arguments (Bowers, 1993, Önem, 2024). Still another possibility that does not rely on an assumption

of active real-time composition is to assume that abstract constructions, too, can be stored. While in these BCO cases where the exact combination of the lexical items has not been experienced, readers might have more abstract storage of constructions/schemas whose strengths vary: [NP[NP NP]], [VP[V NP]], and [TP[NP VP]] (Culicover, Jackendoff, & Audring, 2017). The easier the partially recognized lexical items fit into these stored schemas the more top-down feedback it provides back to the character recognition or the POS determination process. Future studies are required to disentangle these possibilities (see Section 4.2 for more discussion).

2.5.2. Recognition of meaningless/anomalous strings

The experimental paradigm was first invented to address the question of whether a (English) reader can process two words at a time (White et al., 2018). Early results showed strong evidence for a behavioral and neuronal serial bottleneck for word recognition in English (White et al., 2018; 2020; White et al., 2019). A more recent study, however, suggested that the serial bottleneck observed might have been attributable to the physical distance between the two words (White, 2023). Furthermore, semantic-relatedness induced more parallel processing (White, 2023), which could be attributed to stored representations of the whole sequences.

The controls used in the current study provided venues to address the theoretical question of how many words a reader can process at a time. The unrelated/meaningless VV sequences showed evidence of limited-capacity parallel processing. The globally anomalous four-character VNVV and VVNV sequences showed evidence of processing between seriality and limited-capacity parallelism, although note that these sequences contained three, not two, separate units. While even more suitable stimuli—unrelated,

incongruent, syntactically-non-composable, two-unit strings—could have been used to address this question more definitely, overall, the current findings suggest that in Chinese reading, two words can be processed with a certain—not full—degree of parallelism (Yan et al., 2009; Yan et al., 2010; Yan & Sommer, 2015; Yang et al., 2009; Yu et al., 2023).

2.5.3. Stored representations or on-the-fly composition?

Although accuracy in the NPC condition for the FCO strings was pretty clearly comparable to that in the PC condition, full recognition enabled by stored representations of the whole strings might not be the only explanation for the high accuracy. One possibility is that it is easier to compose two frequently co-occurring words, either due to higher forward conditional probability, higher backward conditional probability (Onnis & Huettig, 2021), or closer semantic relatedness. Assuming that the default processing mode of Chinese words is limited-capacity parallel for two words, the two partially-recognized words might be able to trigger the composition process perhaps through activated semantic features (Coltheart et al., 2001) or a rapidly extracted syntactic frame (Koriat & Greenberg, 1994; Wen, Snell, & Grainger, 2019). The easier composition of the two words into a phrase whose representation is made on the fly then provides top-down facilitation to determining POSs of the embedded characters/words.

I note that even if such a possibility exists within a cascaded interactive model, it must be emphasized how effective and fast this composition needs to happen. In the experiment, the target only appeared for 17 ms and was followed by a mask within 34 ms, for over half of the participants. At a similar rate, it has been shown, based on single neuron firing in macaques (Rolls et al., 1999), that only about one third of the information (compared to no masking) is available to be processed, and essentially no

information is gained 250 ms post-stimulus. This suggests that any cascaded interactive composition processes that could have given rise to decent accuracy comparable to the PC condition must finish within this period. Indeed, analysis of RTs¹⁸ in my experiments showed that the difference between the PC and NPC conditions was across the board about 200 ms (which is no more than the duration of the mask itself), providing a converging suggestion about how fast the processes need to be (i.e., within 250 ms after the onset of the target string).

Whether the underlying processes in the NPC condition are unfolded as such can be informed by EEG/MEG studies. In a recent EEG study with a priming design (Wang et al., 2024), a target word (2000 ms) was immediately preceded by a prime presented for 50 ms. There were five types of primes: W+M+, W-M+, W-M-, semantically related, and totally unrelated. W means whole word semantics, M means morpheme meaning, + means congruent, and – means incongruent. For example, with 面颊(cheek) as the target, a W+M+ prime would be 面庞(face), a W-M+ prime would be 面试(interview), a W-M- prime would be 面粉(flour), a semantically related prime would be 容貌(appearance), and an unrelated prime would be 环境(environment). Behaviorally, it was found that only the W+M+ and semantically related primes provided facilitation in *both* RTs and accuracy in lexical decision. The W-M+ and W-M- primes provided the same facilitation in RTs *but no facilitation in accuracy*. Electrophysiologically, it was shown that comparisons between the W+M+ condition and the unrelated condition and between

¹⁸ I refrain from adopting RTs as my main variable of interest given that it was not registered in the prospectus. It also should be noted that RT analysis that included only the correct trials would further remove 10-30% of the data, which would lead to even lower statistical power. Furthermore, RTs as a variable appear to have little sensitivity to composition processes that are able to be captured by concurrent ERP measures (Hsu et al., 2019; Wang et al., 2024).

the semantically related condition and the unrelated condition revealed sustained differences in brainwaves as early as in the 100-150 ms window post-target. Comparisons between the W-M+ condition and the unrelated condition and between the W-M- condition and the unrelated condition also revealed differences in brainwaves, albeit not as early or as sustained. More critically, direct comparison between the W-M+ condition and the W-M- condition revealed no significant difference throughout the 100-500 ms. These last two findings suggest that morphosemantic information (the only property that differed between W-M+ and W-M- primes) is not available with 50 ms masked presentation. In fact, even when presented for 200 ms, qualitatively similar results were observed. On the other hand, W+M+ and semantically related primes appeared to provide extremely early (100-150 ms post-target, or 150-200 ms post-prime) facilitation in processing (in RTs, judgment accuracy, and brainwaves), compared to the unrelated primes, which most likely could be attributed to the meaning activation from the full form.

Wang et al. (2024) concluded that direct full-form-meaning mapping most likely underlies Chinese compound recognition, which would align more with the interpretation of stored representations for my current experiments, if we assume that frequent 4-character strings are also on the same level as 2-character compounds words. However, given that the full meaning of a 2-character string can be accessed so quickly (150 ms the latest), cascaded interactive composition into a 4-character phrase is not impossible either. I thus deem both stored representations and on-the-fly interpretations likely at this point. A different paradigm will be used in Experiment 3 which might provide a different angle in answering this question.

2.5.4. A few notes on making inferences from the data

Before closing Chapter 2, I note a few caveats regarding data interpretation. I argue that while future studies should improve on these issues to replicate the findings for better science, none of the issues substantially compromise the conclusions made above, in my opinion.

First, the experiments adopt a POS judgement task, rather than a recognition or lexical decision task to investigate word/MCU recognition¹⁹. Intuitively, if one assumes that POS judgment is only possible after full recognition of the character itself, the accuracy of POS judgment could easily be used as a proxy for recognition accuracy, since POS ambiguity was normed and also PC/NPC was a within-item manipulation.

Meanwhile, I have mentioned a possibility of cascaded processing where some linguistic information might be accessible prior to full recognition of the character²⁰. However, even under such an assumption, the different patterns seen between the FCO strings and BCO strings still must have arisen from the properties of the whole string (either co-

¹⁹ The choice of POS judgment as the linguistic task, rather than lexical decision or semantic judgment, is due to noun-noun compounds and verb-object sequences being my target stimuli. Because of this, it would be extremely hard, if not impossible, to come up with a question prompt whose judging criterion is comparable across the noun and verb constituent (e.g., Is the character animate? would not be applicable to a verb). A lexical decision task (Is the character a real character?) would not be ideal either, as all my target strings consist of real characters. This means that accuracy measure will be based solely on one question whose correct answer is always yes. This violates the assumption of (signal-detection-based) attention operating curves (White et al., 2018), which requires measurement of both hits and correct rejections. This assumption is maintained in my task by alternating the noun and verb question prompts.

²⁰ Still another interpretation of the POS judgment data relies on inference. It could be the case that when participants only recognized a part of the string, they used that information to infer the POS of the unrecognized part. This could have made meaningless/anomalous strings less accurate and plausible strings more accurate, as participants might guess toward the string being plausible. This could also have made strings of certain syntactic structure more accurate than strings of others. However, this mechanism is less likely to explain the difference within plausible strings of the same syntactic structure (e.g., FCO VO vs. BCO VO). Still, future studies with direct examination on character identification accuracy should be done to corroborate the current findings.

occurrence frequency, conditional probability, or semantic relatedness), not partial morphosyntactic information from any of the embedded characters, to the extent that the embedded frequency of C1, C4, C1C2, and C3C4 has been adequately controlled. Usage of POS judgment therefore should not be a concern, although in the future, direct examination on recognition performance can further corroborate the current findings.

Second, it must be acknowledged that the statistical power in the two experiments might not be high, with 40 participants each, and the critical constructions each having 95-157 items that were further broken down into PC/NPC. When further divided into FCO and BCO subgroups, item numbers were even down to 24-41 for each subgroup and further broken down to PC/NPC. As such, I did not directly run models with interactions that involved co-occurrence frequency. A statistical interaction, however, *must* be demonstrated to provide *evidence of qualitatively different patterns* for different conditions (Nieuwenhuis et al., 2011). Future studies with many more participants and items thus are needed to provide more definite evidence for parallel recognition enabled by co-occurrence frequency, and hence stronger evidence for stored multiword units.

Relatedly, the usage of AOC plots is most helpful only when standard errors are short enough to distinguish data and predictions. The AOC plots were also created by aggregating accuracy across participants for each cuing condition and each side. Bimodal distributions of processing modes across multiple individuals might have given rise to a limited-capacity processing pattern when averaged. This issue applies to item-wise variation: within the FCO groups, variance in co-occurrence frequency exists across items, but they were analyzed in an aggregated fashion. Future studies should devise experiments statistically powerful enough to allow meaningful participant/item-wise

analysis (see Section 4.2 for theoretical motivations, also Section 5.1).

With all this being said, what is crucial throughout the two experiments is that by manipulating the global properties (frequency, structure, or plausibility) of the strings I was able to provide preliminary evidence of different levels of processing modes, as suggested by different degrees of PC-NPC difference: limited-capacity parallel processing was found for nonword/non-MCU conditions, which was coupled with robust fully parallel processing for *all* FCO strings. Also note that a post-hoc model with only SV strings in Experiment 2 did show marginal interaction between pre-cueing and frequency (Table C1). The overall pattern in the two experiments thus provides some evidence for stored multiword units.

This set of new data informs candidates of models of visual word recognition or reading, which will be further discussed in Chapter 4. Chapter 3 reports the third experiment in the dissertation, investigating processing of 4-character strings within a sentence.

CHAPTER 3

PROCESSING 4-CHARACTER STRINGS DURING SENTENCE READING

3.1. Segmenting/recognizing linguistic units from unspaced characters in a sentence

3.1.1. Chinese readers do not read character by character

Do Chinese readers process a sentence incrementally on a character-by-character basis? Or are they able to somehow identify words out of the string even if there are no explicit visual cues? Readers of language systems that deploy spaces between words clearly show difficulty when they read texts of words that are glued altogether (Rayner, Fischer, & Pollatsek, 1998). Compared to reading normal texts, the overall fixation durations are longer, saccadic lengths are shorter, and hard words are read even longer than easy words (i.e., the word frequency effect is larger). Therefore, spaces that mark word boundaries are important for word recognition processes during sentence reading, for languages whose default writing system is spaced. Is this true also for Chinese readers?

Several studies (e.g., Bai et al., 2008; Chen et al., 2021; Zang et al. 2013) address this question by having Chinese speakers read Chinese sentences in different modes of presentation: (1) normal unspaced texts, (2) texts that are spaced between every character, (3) texts that are spaced between each word, and (4) texts that are spaced but result in sequences of nonwords. Some other studies, instead of using spacing, used color-alternation to correctly or incorrectly mark word boundaries (Bai et al., 2008; Zhou et al., 2018). Among these findings, it was robustly found that adding a boundary between

every character disrupts reading processes: Chopping a sentence into single characters led to significantly more regressive saccades, more total numbers of fixations and longer overall reading time, compared to normal presentation (Bai et al., 2008; Chen et al., 2021). It is also apparent that adding boundaries that result in nonword grouping causes disruption to reading measures across the board (Bai et al., 2008; Chen et al., 2021; Zhou et al., 2018). In contrast, reading is facilitated with word-based spacing, including fixation durations (sometimes as early as on first-fixation durations, Zang, et al., 2013), landing position (closer to the word center), and numbers of fixations (more skipping and more single-fixations). Overall sentence reading rates, however, were comparable between the two conditions in both Bai et al. (2008) and Zang et al. (2013), which was argued to reflect an overall familiarity effect for normal, unspaced reading (Chen et al., 2021). When participants read texts printed from right to left (an unfamiliar format in Chinese), the overall reading time was longer for the normal condition than the word-based spacing condition. However, given enough practice (familiarization with 30 minutes of right-to-left reading for ten days), overall reading times became comparable between the two conditions, as in the left-to-right presentation studies. But this practice did not make the difference in mean fixation durations between the two conditions disappear, suggesting that there really is facilitation from word-based spacing, at least for local eye movement measures.

To sum up, studies that artificially added boundaries strongly suggested that Chinese readers do not read sentences on a character-by-character basis, as evidenced from the robust differences between the normal and character-based marking conditions. Yet, readers cannot perfectly segment strings into their correct word groupings online, as

evidenced from the facilitation from word-based spacing conditions in mean fixation durations and numbers of fixations, even though spacing is an unfamiliar format. This is not at all surprising, as in the unspaced, normal texts there are no visual cues for word boundaries at all, and thus perfect online segmentation is basically impossible. Despite this, normal unspaced reading does not seem too disruptive and proves to be more efficient than a character-based strategy.

3.1.2. Word-based reading with Chinese sentences

While Chinese readers might not perfectly extract word boundaries on the fly, several pieces of evidence suggest that Chinese sentence reading is mostly word-based (e.g., Hyönä et al., 2024; Ma et al., 2015; Yang et al., 2012). Some therefore have argued that Chinese words are the psychological real units during Chinese sentence reading (Li et al., 2015).

These studies have been introduced in Section 1.1.2. Here I will briefly reiterate them. Ma et al. (2015), in a factorial design, manipulated the frequency of the first character and the frequency of the whole 2-character word. These 2-character target strings were embedded in a sentence with the preceding and following contexts held the same. It was found that first-character frequency did not influence either first-fixation or gaze durations of the whole word region. On the other hand, whole-word frequency influenced both the first-fixation and gaze durations. This was taken to reflect word-based processing in Chinese sentence reading. More recently, Hyönä et al. (2024) operationalized the first-character frequency using the first-character *word* frequency measure and found a significant effect only in gaze durations on the first character region and only in one experiment but not another. Finally, Cui et al. (2021) and Xiong et al.

(2023) observed that the first-character frequency effect only was observed in infrequent compound words, not in frequent ones. That first-character frequency effects are not observed ubiquitously is consistent with a view that Chinese sentence reading is (largely) word-based.

Yang et al. (2012) manipulated the semantic fit of a verb and its following 2-character compound noun, creating conditions with and without local implausibility (e.g., ENTREAT DOOR-KEEPER vs. KICK DOOR-KEEPER). This local implausibility between the verb and the embedded first character did not yield a longer reading time on the 2-character compound noun throughout the time course. This suggested that Chinese readers process and recognize the whole compound as one and do not evaluate the semantic fit between the first embedded noun and the preceding verb. Recent studies have extended the findings to 3-character compound nouns (Zhou & Li, 2021) and 4-character compound nouns (Wang et al., 2023). Moreover, Wang et al. (2023) found an interaction between local implausibility and co-occurrence frequency of the whole string. That is, local implausibility had an effect for novel/infrequent compounds (e.g., EAT SEA-PLANT vs. PROTECT SEA-PLANT) but had a null effect for familiar/frequent compounds (e.g., EAT SEA-CREATURE vs. PROTECT SEA-CREATURE).

Given the abundant empirical evidence, a theoretical question is *how* Chinese readers engage in word-based processing without explicit word boundaries. The interactive-activation framework (Li et al., 2009; Li & Pollatsek, 2020) relies simply on activation and competition processes for word recognition without positing a specific segmentation mechanism. Under this view, the most frequent string within the perceptual span—supported by bottom-up evidence and constrained by visual acuity (greatest for the

leftmost characters) as well as top-down preceding sentential context with more predictable/plausible words receiving a greater boost (Huang, Staub, & Li, 2021; Yao, Staub, & Li, 2022)—wins the competition and gets recognized. This activation-competition-based approach was further supported by reading of ambiguous character strings (Ma, Li, & Rayner, 2014; Ma, et al., 2017). Yu et al. (2021) proposed a similar, familiarity-based approach to Li and Pollatsek (2020), although their model did not emphasize competition among word nodes/candidates. As noted in Section 1.1.3., a limitation of these models is word nodes being somewhat arbitrarily defined. Furthermore, it is unclear whether word nodes are the highest level of representations that can provide top-down feedback and are the units that readers attempt to recognize or segment for.

Wang et al. (2023) with the local semantic implausibility manipulation showed that frequent 4-character noun-noun compounds were likely to be processed as one unit, yet these strings normally are not listed in a dictionary or in a corpus, due to their literal compositionality (SEA-OCEAN-LIVING-THING). Furthermore, in Experiment 1 and Experiment 2, I have shown that, besides noun-noun compounds, even strings that typically would not be categorized as single words (i.e., literal verb-object or subject-verb sequences) might in fact have corresponding stored representations that can be utilized in real-time processing. Given the suggestive evidence that 4-character NN- and VO-FCO strings can be processed fully in parallel as if they are single units when presented alone, a subsequent question is to ask whether they are also processed together as single units during naturalistic sentence reading. To address this question, I will adopt a boundary-change paradigm to examine how much information of the string can be processed before

the string is even directly fixated (known as parafoveal processing). The next section introduces the paradigm and several studies that have informed our understanding of MCU processing during sentence reading.

3.1.3. The boundary-change paradigm and parafoveal processing

Chinese readers appear to engage in more/deeper parafoveal processing than readers of alphabetic languages. That is, they are more sensitive to linguistic information of the later stimuli even when the eyes have not been fixated on them, arguably due to the visual density of characters and the lack of spaces (Hoosain, 1991; Hyönä et al., 2024; Li et al., 2022). This has been demonstrated by the frequency or semantic parafoveal-on-foveal effects (frequency/semantics of the word following a target word can influence processing time on the target word even before the eyes move beyond the target word, Yan et al., 2009; Yan et al., 2010; Yan & Sommer, 2015; Yang et al., 2009), which is generally not found in the English literature (e.g., Brothers et al., 2017; Veldre & Andrews, 2018a).

Other studies report a preview validity effect using a gaze-contingent boundary-change paradigm (Rayner, 1975). In such an experiment, participants are asked to read naturally one sentence at a time, with a target region embedded within (e.g., *This is a **cute** dog*). Unknown to the participants, an invisible boundary is set before the target region by the experiment program. There will be at least two conditions (the valid preview condition and the invalid preview condition). In the valid preview condition, the sentence will be presented in a normal way, with the invisible boundary bearing no relevance. In the invalid preview condition, the sentence will initially be presented with the target region containing a word/string that is different from the word/string in the sentence in

the valid preview condition (e.g., *This is a **good** dog*). Once the participants' eye fixation is registered as being to the right of the boundary, the word/string in the target region (*good*) will change to the word in the sentence in the valid preview condition (*cute*). That is, the sentence will be changed back to a normal sentence (*This is a cute dog*).

Importantly, the change will happen extremely rapidly such that participants will rarely be aware of a change of stimulus occurring. The difference in eye movement measures on the target word can be taken to infer how much processing has started parafoveally before the target word is fixated: if having been shown a different string has no impact on any eye movement measures on the target region, participants most likely do not process the word in the preview at all while their eyes are still on the region prior to the target region. In reality, despite not consciously noticing the change, readers' reading times on the target word are robustly affected by the manipulation, suggesting readers do not wait until they directly fixate a word to process it (see Schotter, Angele, and Rayner, 2012, for a review).

In the Chinese literature, the preview effect can be obtained even with the changed target word not immediately following the invisible boundary (an analogous English example is *They can| see **faces/words** on the screen.*, where | represents the invisible boundary and the bolded words represent the words before and after the change; note the word *see* remains the same pre- and post-boundary-crossing). This is known as the Word $n+2$ preview effect and is elusive in the English literature but less elusive and stronger in size in the Chinese literature (Vasilev & Angele, 2017, for a review and meta-analysis). The Word $n+2$ preview effect in Chinese, however, had been suggested to occur only when Word $n+1$ was short (one-character) and high-frequency (Yan et al., 2010; Yang et

al., 2009; Yang et al., 2012).

More recently, Yu et al. (2016) showed that the Word $n+2$ preview effect might in fact be more common. They compared a preview effect for 3-character *idioms* with that for 3-character *literal strings*. An analogous English example is *kick the bucket* as an idiom and *carry the bucket* as a literal string. The structure of the 3-character string was VNN (verb-object), and all the verbs were of rather low frequency. The boundary was in front of the verb ($n+1$), but only the noun ($n+2$) went through a change. The invalid preview of the noun was two characters that could not form a word, as in (2a-2d)

(2a) 周丽丽觉得|揭疮疤打小报告是小孩们的行为 (idiom, identical preview)

(2b) 周丽丽觉得|揭俘郡打小报告是小孩们的行为 (idiom, invalid preview)

(2c) 周丽丽觉得|留疮疤打小报告是小孩们的行为 (phrase, identical preview)

(2d) 周丽丽觉得|留俘郡打小报告是小孩们的行为 (phrase, invalid preview)

(揭-疮疤, PICK SCAB = *reveal a bad secret* (figuratively); 留-疮疤, LEAVE SCAB = *leave scabs*; 俘郡 = nonword shown as the invalid preview; | indicates preview

boundary that was unknown and invisible to the participants; characters in red were the characters going through the change; red fonts were used for illustration purpose only, not in the actual presentation)

Their critical finding is that, both on the $n+2$ region and the whole 3-character region, there was a main effect of preview on first-fixation, single-fixation, and gaze durations such that reading times were longer with an invalid preview. However, there was no interaction with structure (idiom vs. literal string). That is, the extent to which the post-

boundary reading benefited from the correct preview is similar, whether the 3 characters constituted an idiom or a literal string. Moreover, that there was an invalid preview effect for 3-character literal strings suggests that word $n+2$ can be processed via parafoveal view, even when word $n+1$ is rather low-frequency.

Zang et al. (2021) examined the preview effect using idiomatic/literal three-character strings with a modifier-noun structure. The design was essentially identical to Yu et al. (2016) except that it involved a MMN (MM means a 2-character modifier) rather than a VNN structure. The first constituent is always a modifier for its following noun, for both the idiom and literal string, but the idiom's meaning is more figurative than the literal one's. An example is as (3a-3d).

(3a) 小辉充分利用各种|垫脚石当上了区域经理。 (idiom, identical preview)

(3b) 小辉充分利用各种|垫脚平当上了区域经理。 (idiom, invalid preview)

(3c) 小辉充分利用各种|垫脚布让家里保持乾淨。 (literal, identical preview)

(3d) 小辉充分利用各种|垫脚平让家里保持乾淨。 (literal, invalid preview)

(垫脚-石, FOOTSTEPPING STONE = *pawn* (figuratively); 垫脚-布,

FOOTSTEPPING PAD = *bathroom rugs*; 垫脚-平, FOOTSTEPPING FLAT = nonword;

| indicates preview boundary that was unknown and invisible to the participants;

characters in red were the characters going through the change; red fonts were used for illustration purpose only, not in the actual presentation)

The rationale here is that if figurative idioms are more lexicalized as a MCU than literal ones, then their constituents should be more likely to be processed together and as a

whole and processed to a greater extent even from the parafoveal view. The results showed that both the idiomatic and literal conditions showed a preview effect, such that having a valid preview facilitated foveal reading time of the first two characters, the last character, and whole region across all the fixation duration measures. This preview validity effect interacted with the string type such that a bigger benefit from a valid preview was present for idioms than for literal strings. There was also a main effect of string type such that idioms were read faster than literal strings across all fixation duration measures, regardless of the preview condition. Overall, both findings on foveal and parafoveal processing supported Zang et al.'s (2021; Zang, 2019) hypothesis that idioms are more lexicalized than their literal counterparts, even when their structures are exactly the same (modifier-noun). Similar findings using strings of a MNN structure were reported in Zang et al. (2023).

To summarize this section so far, given that the logographic script is denser, and characters are short in width, Chinese readers can process quite a bit more than just the character/word currently fixated. The extent of this processing of the right stimuli can be addressed by demonstrating a preview validity effect and further its interaction with the target string's meaning (idiomatic vs. literal). The three experiments respectively used VNN (Yu et al., 2016), MMN (Zang et al., 2021), and MNN strings (Zang et al., 2023), with the latter two experiments finding comparable effects and interactions between the experiments (i.e., both literal and idiomatic strings showed a preview validity effect, but idioms had a stronger one). These differed from the VNN experiment where only a main effect of preview validity was shown.

The first thing to point out is that while none of the studies reported whole-string

frequency, it seems clear that the literal strings adopted are of low frequency. That all 3-character strings in these studies (even the literal, infrequent verb-object strings) showed a preview validity effect is consistent with Experiment 1 and Experiment 2 of mine, suggesting that two words can be processed in parallel to a great degree. Second, since in all three experiments the idioms can be construed as MCUs (due to their figurative meanings by convention) and the literal ones are not (due to their low frequency), the interactions observed in Zang et al (2021) and Zang et al. (2023) support the MCU hypothesis (Zang, 2019). Finally, it is less clear why there was no interaction found in Yu et al. (2016). One possibility is that the VNN idioms used were not frequent enough. Given the constraints (three characters, with the verb being only one character and low frequency), there might not be many eligible frequent VNN idioms. Therefore, despite the meanings being fixed and non-literal, the strengths of the stored representations might be fairly weak.

Finally, rather than using idiomatic and literal strings, Zang et al. (2024) created 4-character strings, with each item having three variants that can be categorized into three conditions: one-word condition, ambiguous condition, two-words condition. The three variants only differed in the last two characters. This categorization was normed and confirmed in a word-boundary-placing task with 55 participants: Participants were asked to decide where the word boundary should be within each 4-character string. Almost no one placed a word boundary in the strings in the one-word condition, about half of the participants place a boundary in the middle of the strings and about half placed no word boundaries in the ambiguous condition, and almost everyone placed a word boundary in the middle of the strings in the two-words condition. The $n+2$ boundary change paradigm

was adopted, as shown in (4a-4c).

(4a) 张雨桐认为|安全系数达到百分之二十才合格。(one-word, identical preview)

(4b) 张雨桐认为|安全新率达到百分之二十才合格。(one-word, invalid preview)

(4c) 张雨桐认为|安全需求得到满足才最重要。(ambiguous, identical preview)

(4d) 张雨桐认为|安全甯坐得到满足才最重要。(ambiguous, invalid preview)

(4e) 张雨桐认为|安全就是要始终高于一切。(two-word, identical preview)

(4f) 张雨桐认为|安全荆芟要始终高于一切。(two-word, invalid preview)

(安全系数, SAFETY FACTOR; 安全需求, SAFETY NEED; 安全就是, SAFETY

IS; | indicates preview boundary that was unknown and invisible to the participants; characters in red were pseudo-characters that were presented invalidly and went through a change; red fonts were used for illustration purpose only, not in the actual presentation)

In the invalid preview condition, the two characters are initially pseudo-characters, and only turn into the correct characters after participants' fixation crosses the invisible boundary. The mean frequencies and strokes of the last two characters across the three conditions (one-word, ambiguous, and two-words) were matched. However, the whole-string frequency and the syntactic structure were not matched across the three conditions: The one-word condition had the highest mean of co-occurrence frequency, followed by the ambiguous condition, and the two-words condition the lowest. The strings in the one-word condition and the ambiguous condition also tend to be of NNNN or MMNN structure, while the last two characters in the two-word condition are less homogenous in terms of their parts-of-speech.

Consistent with the MCU hypothesis, the one-word condition, which was normed offline to be one-unit and was of high co-occurrence frequency, showed the greatest preview validity effect. The ambiguous condition also showed a reliable, but smaller, preview validity effect, while the two-words condition showed no preview validity effect. Thus Zang et al. (2024) extended Zang et al. (2021) and Zang et al. (2023) with 4-character strings and further showed that offline word-unit judgment or co-occurrence frequency is an indicator for an MCU. They also extended He et al. (2021) by showing the MCU processing advantages start as early as in the parafovea.

3.2.Experiment 3: Parafoveal processing of 4-character strings of high/low frequency and of different morphosyntactic structures

While Zang et al. (2024) showed that a parafoveal validity effect can be shown with strings as long as 4 characters, inference can only be made about noun strings (either modifier-noun or noun-noun); it is still unknown what is the role of morphosyntactic structure in the usage of MCUs during naturalistic sentence reading.

While Experiment 1 and Experiment 2 found minimal differences in processing between noun-noun and verb-object sequences (almost identical patterns for FCO strings) that were presented very briefly in isolation, it is possible that differences in processing patterns for the two structures emerge in naturalistic sentence reading (Xiong et al., 2023). One possibility is that sentence comprehension relies heavily on structure building/composition (Stanojevic et al., 2023; Mollica et al., 2020) in order to construct sentence-level meanings, and composing the relation between a verb and a noun may be harder than composing the relation between two nouns (e.g., processing a verb involves thematic-role assignment, MacDonald, Pearlmutter, & Seidenberg, 1994). Compared to a

noun-noun compound, building a VO structure may impose extra processing cost, and this could inhibit a preview effect. Experiment 3 thus bridges the gap to further understand the role of task in processing of different structures of MCUs.

3.2.1. Participants

Sixty-four participants from the same pool as in Experiment 1 and Experiment 2 were recruited. Three participants did not finish the experiment due to poor calibration of eye-tracking; one participant's experiment crashed early in the study. In the end, 60 participants' data were analyzed (mean age = 24 yr; 17 males and 43 females), which included one participant who went through five-sixths of the experiment and was too tired to continue (but their data were kept). Twenty-one of the 60 participants had participated in Experiment 1 and/or Experiment 2, with at least one month intervening between their participation in E1 and/or E2 and their participation in E3. See Supplementary Material for demographic and language-background summary.

3.2.2. Materials

A target 4-character string in each sentence. In the invalid preview condition, the last two characters in the string were first presented invalidly as two pseudo-characters and turned into the last two characters in the actual target string once the program detected the eyes crossing the invisible boundary. While the boundary was placed in front of the whole 4-character string, the first two characters did not go through any changes pre- and post-boundary (see 5a-5f below for example).

One hundred eighty critical 4-character items were created, divided into 6 groups (30 items each): frequent-NN, infrequent-NN, frequent-VO, infrequent-VO, frequent-AN (adjective-noun), and infrequent-AN. Here, rather than using barely co-occurring

sequences. I used infrequent strings of non-zero frequency with highly plausible meanings (see 5d-5f below). Since no explicit POS judgment was involved in the experiment, the POS of each character did not necessarily fit a NNNN, VVNN, or AANN structure, as long as the whole string was unambiguously NN, VO, or AN. Each target string was embedded in a sentence, with the preceding contexts differing across items. This design differed from Zang et al. (2024) where all the texts up until the first two characters of the target strings were the same across the three triplets within an item (see 4a-4f above). This choice was because here I directly manipulate the syntactic structure as a factor. While it is possible to make the preceding context the same across three string types, pros and cons exist with such a design (see the last paragraph in this section). Here, to match the preceding context, the pre-target region (the two characters prior to the target strings) always contained a frequent two-character string of similar kinds across all items: for NN and AN, it was always a generic determiner or quantifier (e.g., my, the, one, this, few); for VO, it was always an auxiliary verb or temporal adverb (e.g., can, will, never, immediately). By using these specific strings at the pre-target region, the syntactic predictability of the target words can also be tightly controlled across string types. I confirmed this with a cloze norming task in an online experiment with 43 participants (see Section 3.2.3.1). The cloze norming results also showed that all 4-character strings had zero predictability. Naturalness of the whole sentences was also normed in the online experiment (see Section 3.2.3.5.1 for norming results).

Examples are shown in (5a-5f, in the actual experiment there was no bold font, underlines, or colored texts; they are only for illustrative purpose to the readers of this manuscript). | indicates the invisible boundary, the pretarget strings are underlined. In

bold fonts are the last two characters of the target string. / indicates alternation: Colored in red are the characters shown invalidly in the preview which turned in to the correct two characters (black, bold font) after eyes passed the invisible boundary. These characters in the invalid preview are pseudocharacters, matched on orthography and number of strokes with their valid counterparts.

(5a) 我怀疑你的|高中**匪寨**/同学不过是想骗你的钱。 Frequent-NN

I suspect your HIGH-SCHOOL-CLASSMATE is simply trying to rip you off.

(5b) 周琦的姊姊从此|失去**砵刃**/信心不愿意接手任何案件。 Frequent-VO

Chou-Chi's sister since then LOST-CONFIDENCE, not willing to overtake any cases.

(5c) 大家都希望这种|最高**焯尖**/权力不要落入邪恶之人的手中。 Frequent-AN

Everyone hopes this HIGHEST-POWER does not fall into some evil person's hands.

(5d) 我想修改我的|商品**伙莽**/订单却一直找不到操作说明。 Infrequent-NN

I want to revise my MERCHANDISE-ORDER but keep not finding instructions.

(5e) 黄冠霖正在|制作**獾当**/帽子给他的女儿当礼物。 Infrequent-VO

Kuan-Lin Huang right now (is) MAKING-HAT for his daughter as a gift.

(5f) 根据规定你的|私人**乍峴**/宠物若要上火车是需要额外买票的。 Infrequent-AN

According to the rules your PERSONAL-PET, if taken onto the train, requires additionally buying a ticket.

(In the examples here, both the invalid (red) and valid (black) characters are shown in one sentence, with / indicating pre- and post-boundary presentation)

Table 11 shows the descriptive statistics for each group's relevant variables. Frequency

measurements are log-transformation of raw counts that are based on the Weibo-HuanQiuRenWu corpus (105 million characters) I made myself but later corroborated with the 2024's PKU CCL corpus (see Appendix D).

String types	WP	WF	PreF	F12	F34	F1	F2	F3	F4	S12	S34	WF/ F12	WF/ F34	Natur.	SynPred ²¹
Freq-NN	7.1 (1.3)	4.46 (0.4)	12.7 (1.1)	8.81 (0.5)	8.77 (0.8)	11.3 (0.9)	11.3 (1.2)	11.6 (1.0)	11.1 (0.9)	15.5 (3.8)	14.0 (3.8)	.017 (.01)	.018 (.02)	6.11 (0.42)	0.89 (0.13)
Infreq-NN	6.8 (1.4)	0.83 (0.3)	12.9 (1.5)	8.77 (0.5)	8.70 (0.8)	11.2 (0.8)	11.1 (0.9)	11.2 (1.1)	10.9 (1.0)	15.9 (3.8)	14.9 (4.6)	.000 (.00)	.000 (.00)	5.83 (0.49)	0.83 (0.18)
Freq-VO	6.3 (1.4)	4.47 (0.4)	12.2 (1.6)	8.70 (0.7)	8.66 (1.0)	11.1 (1.0)	10.9 (1.0)	11.2 (0.9)	11.1 (1.2)	14.7 (3.6)	14.9 (3.8)	.021 (.03)	.026 (.03)	6.03 (0.27)	0.81 (0.20)
Infreq-VO	5.1 (1.1)	0.83 (0.4)	11.9 (1.9)	8.74 (0.6)	8.66 (1.0)	11.1 (1.1)	11.0 (1.1)	11.0 (1.2)	11.0 (1.5)	14.7 (4.0)	15.1 (5.0)	.000 (.00)	.000 (.00)	5.93 (0.37)	0.88 (0.13)
Freq-AN	6.7 (0.9)	4.46 (0.4)	13.4 (0.7)	8.56 (0.8)	8.83 (0.9)	11.8 (1.1)	11.5 (1.1)	11.2 (1.0)	11.0 (0.7)	14.7 (4.4)	15.7 (4.2)	.024 (.03)	.019 (.02)	6.11 (0.38)	0.12 (0.1) 0.79 (0.2)
Infreq-AN	6.6 (1.4)	0.78 (0.2)	13.2 (1.1)	8.68 (0.7)	8.77 (0.9)	11.4 (1.0)	11.3 (1.0)	10.9 (1.0)	11.0 (1.0)	14.7 (4.3)	15.3 (4.1)	.000 (.00)	.000 (.00)	5.99 (0.50)	0.12 (0.2) 0.81 (0.2)

Table 11. Descriptive statistics for the relevant variables across the six conditions in Experiment 3. WP: word position in the sentence; WF: target strings' frequency; PreF:

²¹ The numbers in the second row for AN reflect the syntactic predictability of a noun.

pretarget strings' frequency; F12: C1C2 of the target string's frequency; S12: C1C2 of the target string's number of strokes; WF/F12: conditional probability of the whole string given C1C2; WF/F34: conditional probability of the whole string given C3C4. Natur: normed naturalness of the whole sentences; SynPred: syntactic predictability (see 3.2.4.1).

There were in total 3 (string structures, NN/VO/AN) \times 2 (frequency, high/low) \times 2 (preview, valid/invalid) conditions. The valid/invalid variable was a within-item manipulation; each participant only saw one version of each item. There were therefore 2 Latin Square counterbalance lists. Each participant saw 15 items for each of the conditions. Along with the 180 critical items, participants also saw 45 filler sentences of similar lengths of random sentence structures. All sentences were presented in a totally random order. These 225 sentences were preceded by 4 practice sentences. One third of the sentences were followed by a comprehension question for attention check.

The choice of including AN strings, in addition to the comparison between NN and VO strings, was to further address a potential concern for—or an empirical question about the role of—incremental processing. (5a) and (5d) illustrated potential misinterpretations due to incremental processing. In (5a) the subject of the embedded sentential clause should be CLASSMATE, not HIGHSCHOOL. However, it is in theory possible that readers will initially adopt an interpretation of YOUR HIGHSCHOOL only to find out the ultimate reading should be YOUR HIGHSCHOOL CLASSMATE, which can cause processing difficulty (Xiang, 2013). This scenario of initially adopting an incorrect interpretation would be more likely to happen in the invalid preview condition: Given that the last two pseudocharacters do not form a 4-character unit with HIGHSCHOOL, readers might recognize/segment HIGHSCHOOL and immediately integrate it with YOUR to form a subject role, prior to directly fixating on the four-

character string. If so, once fixated, HIGHSCHOOL-CLASSMATE would yield an even longer reading time. The difference in reading time between the valid and invalid conditions for NN strings might thus be a composite of a regular validity effect *and* a misinterpretation effect. For VO strings there would not be such a complication: even if readers recognize/segment the verb in the invalid condition and immediately integrate the verb with the preceding auxiliary verb, prior to directly fixating on the four-character string, they do not need to revise their interpretation once fixating directly the four-character string.

Whether this is the case can be known by adding another type of string (AN). This can be easily seen in (5c) and (5f) where initially integrating HIGHEST with THIS or integrating PERSONAL with YOUR would not induce processing cost when the noun is later integrated to it: THIS HIGHEST POWER has no conflict with THIS HIGHEST, and YOUR PERSONAL PET has no conflict with YOUR PERSONAL. Therefore, comparing three types of strings will provide a bigger picture about parafoveal processing.

Finally, the choice of not having a triplet of an item (NN/VO/AN) that shared the same preceding context is due to the difficulty of matching syntactic predictability of words in the target region. That is, given one sentence context, *I suspect your*, it is much more likely that the incoming text will be a noun than a verb²². It is possible that the higher syntactic predictability of a noun coming next will facilitate recognizing a noun in the parafovea (Staub & Clifton, 2006). The preview validity effect might thus be stronger for NN strings than for VO strings, in this example. This difference, however, will reflect

²² in Mandarin a relative clause can be formed after a determiner, but this will be an infrequent sentence structure

the role of the context in MCU usage but tell us little about the role of the structure of the MCU itself during sentence processing. On the other hand, by using different preceding contexts I can match the syntactic predictability of the target strings across the three string types.

3.2.3. Procedures

3.2.3.1. Online norming experiment

An online experiment was conducted on IbexFarm to ensure that the naturalness of the sentences and the predictability of the target strings were comparable across the 6 string types. The experiment consisted of 2 parts. It started with a cloze task (Taylor, 1953) where on each trial participants were provided a sentence context and were asked to continue the sentence. Participants were told that they do not have to finish the sentence, and their response can be of any length. An example question prompt was given, and four sample answers to this example prompt were given; they varied from containing two characters, containing three characters, containing four characters, to containing fourteen-characters-ending-with-a period. Participants were told that every sentence is independent and a response on one trial should not influence or be influenced by a response from a different trial. Three practice trials were included. For each item in the main eye-tracking experiment, all the characters preceding the target 4-character string were presented as the prompt on each trial in this norming task, e.g., 我怀疑你的 _____ (I suspect your _____), from (5a). Each participant underwent 90 trials (half of the 180 critical items in the main experiment).

The second part is a naturalness rating task. Participants read one full sentence at a time, and were asked to rate each sentence, on a scale of 1 to 7, based on how natural the

sentence sounds, 6 or 7 being extremely natural, and 1 or 2 being awkward-sounding. The instruction specified that the rating should not be based on semantics, as some sentences might be hard to understand due to lack of a wider context. Participants should only decide whether the sentence sounds like a sentence that some person would produce under some specific context. If they think it is a sentence that would occur in their daily life, in literature, or in films or television, they should give it a higher score. If they think it is a sentence that would not occur in any kind of situation, they should give a lower score. They were told that every sentence is independent and a response on one trial should not influence or be influenced by a response from a different trial. Three practice trials were included. Each participant read 90 sentences (the other half of the 180 critical items in the main experiment that they did not give a cloze response to). Twenty filler sentences were also included in addition to the 90 sentences. These sentences were all nonsensical, with the degree varying widely.

Forty-three participants were recruited via group chats of Chinese international students (in New York, California, Maryland, or Michigan) via mobile application WeChat. They were paid 7 dollars for their 30 minutes of participation.

3.2.3.2. Main eye-tracking experiment

The experiment was conducted in Mandarin. Participants came to the lab and filled out the consent form and the demographic form. They were then seated at the eye-tracker with a head and chin rest. The monitor was 19 inches Dell UltraScan P991, with a refresh rate of 120 Hz. Resolution was set to 1024 x 768. The distance between the screen and the participants' eyes was 56.5 cm. At this distance, each character (27 pixels) spans about 1 degree of visual angle. Eye movements were recorded with an SR Research

EyeLink1000 eye tracker with a sampling rate of 1,000 Hz. The eye tracker was calibrated using a three-point horizontal calibration procedure with an average error below 0.30 degrees of visual angle. Each trial started with a drift check, and re-calibration was run whenever needed.

Participants were asked to read one sentence each time, not out loud but silently. After finishing reading the sentence, they should press the right button to proceed. They were told that for a proportion of sentences, there will be a comprehension question immediately following. The questions were all two-alternative forced-choice questions. They used the left or right button to select from the two choices. On each trial, they fixated centrally first and then saw a black square on the very left edge of the screen. Once the black square is fixated, it will trigger the presence of a sentence. Emphasis was made about reading at a natural pace, not too fast or too slowly. They were told to avoid fast scanning through the sentence, but to read for comprehension. They were told to press the right button once they think they understand the sentence. The experiment took on average 50 minutes to complete. After the experiment, participants were asked if they noticed anything abnormal in the experiment. If they said they notice the boundary change, a further question probed on what proportion of the trials they noticed a change. They were then debriefed regarding the actual percentage of the trials that involved a change, and the purpose of the experiment.

After the experiment was over, participants completed an offline POS judgment task on a spreadsheet where each row contained one single character (for Experiment 1 and 2's items). They were asked to type out, as quickly as possible, whether the character is a noun, a verb, or an adjective on the next column and not overthink. If they found it hard

to decide the POS of the character, they should leave the cell blank. They were paid \$25 for 75 minutes of their participation.

3.2.4. Analysis

3.2.4.1. Norming

The naturalness of each sentence was calculated across all participants. A glimpse at the data was taken halfway through the study, and items that were rated with a mean score below 4 were edited. Two items unexpectedly ended up not being rated due to errors in executing the script during this revision process. Two other items unexpectedly ended up not having cloze responses due to errors in executing the script during this revision process.

The cloze responses were analyzed in the following way. First, lexical predictability only considered exact four-character matches, which no participant ever produced for any item. Two other lexical predictability measures were calculated, one considering the first two-character matches and the other considering the last two-character of the target string produced by the participants as the first two-character of their responses. Second, the cloze responses were re-coded into different syntactic categories by the author to calculate syntactic predictability. Regardless of the length of the response, I only considered the very first element of the response. For NN items, the syntactic predictability is counted as the number of noun responses divided by the total number of responses, for each item. For VO items, the syntactic predictability is counted as the number of verb responses divided by the total number of responses, for each item. For AN items, the syntactic predictability is counted as the number of adjective responses divided by the total number of responses, for each item. However, since adjectives are

optional modifiers, they occurred much less often. As such for AN items, I also reported the syntactic predictability based on noun responses (see Footnote 21).

3.2.4.2. Eye-movement measures

Four eye movement measures were analyzed. First-fixation duration is the very first fixation on the region defined. Gaze duration is the sum of all fixations prior to a fixation that exits the region defined, either exiting forward to the right or backward to the left. Go-past time is the sum of all fixations until a fixation goes to the right of the region defined; the fixations include those that are not on the region defined. Skipping probability is reported. Skipping is only defined when it happens during first-pass reading.

First-fixation duration, as its name suggests, is the very first fixation on a region—the moment participants would see the newly-changed string, and therefore a difference on this measure between the valid and invalid preview condition might mostly reflect processing induced by the visual/perceptual change (Dimigen et al., 2012; McConkie & Zola, 1984), although the very first fixation duration also can reflect deeper processing such as semantics (Rayner et al., 2004), in non-boundary-change paradigms. Gaze-duration includes the first fixation, and second, and so on, until the eyes move out of the region for the first time, therefore definitely encompassing full processing of the target string. If parafoveal processing of the 4 characters goes through processing beyond visuo-orthographic processing, then a difference can be expected in this measure. Go-past time subsumes gaze duration but in addition includes all the fixations in the regressive-path. This includes a scenario where participants only fixate one time on the target region and immediately regress back to earlier points. This scenario can be triggered by low-level

perceptual discrepancy pre- and post- invisible boundary.

Indeed, in Zang et al. (2024), the most diagnostic measure is first-fixation and gaze durations on C1C2 and on the whole string. Table 12 summarizes where Zang et al. (2024) found and not found preview validity effects for each of their three experimental groups: Measures that were non-diagnostic (all groups showing significant or all groups not showing significant preview validity effects) are marked with a minus sign. Measures that were diagnostic to differentiate the three string types and provided evidence consistent with the MCU hypothesis are marked with a plus sign. The measure that differentiated the three string types but provided evidence not directly consistent with the MCU hypothesis is skipping probability on C3C4. To reiterate, based on Zang et al. (2024), *skipping probability* of C1C2 and *first-fixation* and *gaze* durations of C1C2 and of the whole string are the loci where I anticipate to find evidence that differentiates different types of strings.

	One-word	Ambiguous	Two-word
C1C2 – Skipping probability +	V	X	X
C1C2 – First-fixation duration +	V	V	X
C1C2 – Gaze duration +	V	V	X
C1C2 – Go-past time -	V	V	△
C3C4 – Skipping probability *	X	X	V
C3C4 – First-fixation duration -	X	X	X
C3C4 – Gaze duration -	X	X	X
C3C4 – Go-past time -	X	X	X
WholeString – First-fixation duration +	V	V	X
WholeString – Gaze duration +	V	V	X
WholeString – Go-past time	V	V	V

Table 12. Summary of Zang et al. (2024)’s significant preview validity effects. Included here are 11 measures that I will report in the current experiment (out of the 18 measures originally reported in Zang et al., 2024). + indicates measures that differentiated the experimental groups; - indicates measures yielding same results among the groups; * indicates the measure that differentiated the groups but were not easily interpretable regarding the MCU hypothesis. △ indicates a marginal preview validity effect.

3.2.4.3. Statistical Power

To estimate the power of the current study, I referred to Zang et al. (2024)'s existing dataset, which involved a 2 (valid/invalid preview) \times 3 (one-word/ambiguous/two-word) design. The interaction between validity and one-word-vs.-two-word contrast is analogous to the interaction between validity and string structure (NN-vs.-VO) of the high-frequency subset in the current study: their one-word stimuli and my frequent noun-noun stimuli were hypothesized to be unified single processing units while their two-word stimuli and my verb-object stimuli were not, under the hypothesis of structure building and thematic processing during sentence reading. For simulation, effect sizes and variances were based on their empirical findings: to do so, I first ran a Bayesian linear mixed-effects model of their gaze duration data on the C1C2 region with full random-effect structures to obtain the fixed and random effect estimates. Fixed effects were directly taken as the expected fixed effect sizes and as means of the distributions where the simulated data were sampled from. The standard deviations of these distributions were chosen in a less systematic way, by trial and error, yet eventually each simulated dataset ended up having comparable estimated variances as the original estimates of Zang et al.'s. Each simulated observation was a sum of samples taken independently from multiple Gaussian distributions that respectively reflect fixed intercept, effects and interactions, and random subject and item intercepts and slopes. No correlations among random effects were assumed since the Bayesian model of the original dataset did not yield any reliably strong indications of non-zero correlations. Data loss (e.g., blinking or skipping) was assumed to be at the same rate as in Zang et al: for each participant a loss rate between 0.1 to 0.6 (mean = 0.35) was drawn and randomly applied to the total trials. A hundred experiments were simulated. Each stimulated dataset

was entered into a frequentist linear mixed effect model to estimate whether the critical interaction had a t-value higher than 1.96.

The current study, however, differed from Zang et al. (2024) in that it involved a 2 (valid/invalid preview) \times 3 (NN/VO/AN) \times 2 (frequent/infrequent) manipulation. For the infrequent stimuli, I assumed very little to no effect of preview validity and reliable frequency effects for all string structures based on Jiang and Siyanova-Chanturia (2023) and Wang et al. (2023). Frequency effects were assumed to be slightly smaller for the invalid preview conditions. The study also differed from Zang et al. in having only 60 participants but 180 items (with only 2, rather than 6, counterbalance lists). The simulated experiments were based on these specifications for the power estimate of the current study. For the regression model in each simulation, the effect of frequency and the effect of validity was sum-coded (frequent: 0.5; infrequent: -0.5; invalid: -0.5; valid: 0.5) and the 3-level effect of structure was be dummy-coded (VO: 0,0; NN: 1,0; AN: 0,1, see Section 3.2.5.2 for reasons for this contrast coding scheme). Table 13 summarizes the fixed intercept and slopes that generated the simulated observations (log RTs, for parameters of random effects see Appendix E). Note that the coding scheme used to generate the samples was different from that of the regression model, out of convenience, as linear combinations are more intuitive when factors were all dummy-coded.

	Mean	sd	Note
Intercept (frequent, invalid NN)	5.48	0.15	
Infrequency effect (for invalid NN)	0.03	0.1	very small frequency effect with invalid preview
Validity effect (for frequent NN)	-0.09	0.15	robust preview effect
VOvsNN (for frequent, invalid NN)	0.02	0.15	Based on Zang et al. (2024)
ANvsNN (for frequent, invalid NN)	0.05	0.1	Based on Zang et al. (2024)
Infrequency \times validity (for NN)	0.06	0.15	barely any preview effect for infrequent NN
Infrequency \times VOvsNN (for invalid)	-0.01	0.15	very small frequency effect with invalid preview regardless of structure
Infrequency \times VOvsAN (for invalid)	-0.005	0.15	very small frequency effect with invalid preview regardless of structure
Validity \times VOvsNN (for frequent)	0.10	0.1	no preview effect for VO even for frequent strings
Validity \times ANvsNN (for frequent)	0.03	0.1	smaller preview effect for frequent AN
Infrequency \times Validity \times VOvsNN	-0.03	0.2	even smaller preview effect for infrequent VO
Infrequency \times Validity \times ANvsNN	-0.01	0.2	similar preview effects for infrequent AN and infrequent NN

Table 13. Summary of parameters used to generate simulated data points in Experiment 3. Gray shades indicate fixed effect size taken directly from a Bayesian model with Zang et al. (2024)'s data. For random effects, see Table E2 in the Appendix.

The simulations showed that only 1 out of 100 simulations resulted in a significant three-way interaction (validity \times frequency \times NNvsVO). However, for the two-way interaction (validity \times NNvsVO, aggregating across the two frequency levels), 76 out of 100 simulations revealed a significant result. For the two-way interaction (validity \times frequency at the VO level), power was only 11%. Since noun-noun sequences, rather than verb-object sequences, are the ones hypothesized to be most likely to benefit from co-occurrence frequency to become MCUs, for each simulated dataset, another reduced frequentist model was run including only the noun-noun trials to estimate the power for the two-way interaction (validity \times frequency at the NN level). The power to detect such

an interaction was also very low, 22%.

As the power analysis indicated extremely low power to detect a three-way interaction, I would not interpret any null three-way interaction. To probe if preview validity effects are qualitatively different across string structures and frequency without interpreting the three-way interaction, I will derive from the full model the simple effect of preview validity for each of the six string groups (NN/VO/AN \times frequent/infrequent) and visualize the posterior distributions of the simple effects. For the sub-conditions with visually different effect sizes of preview validity, a post-hoc reduced regression model including only trials of the sub-conditions will be run to provide quantitative evidence of a two-way interaction. The limitations of this statistical practice are noted in Section 5.1.

3.2.5. Results

3.2.5.1. Norming results

All items were rated for naturalness by at least 19 people (except 3 items by 9 people and 2 items by 0 people due to iterative item editing to improve naturalness). All 178 rated items had mean naturalness higher than 4.5 on a 1-7 scale. Mean naturalness of filler implausible sentences was 3.02. Mean naturalness of the 6 critical conditions was across the board around 6 (Table 11; see also Appendix F for visualization of the distribution of item naturalness), with items with frequent strings rated slightly more natural than items with infrequent strings. However, given the means were all close to 6 (out of the maximum 7), I do not consider this difference practically meaningful.

All items had cloze responses from at least 21 people (except 3 items by 8 people and 2 items by 0 people due to iterative item editing to improve naturalness). Overall predictability of subparts of the target strings given the context was very close to zero

(see Appendix F for the mean cloze probability aggregated across the items for each condition). The extremely low word predictability is further corroborated by a large language model (Appendix G, Chinese gpt2-xl, Zhao, et al., 2023).

3.2.5.2. Eye-movement results

As a reminder, this experiment involves a 2 (valid/invalid preview) \times 3 (NN/VO/AN) \times 2 (frequent/infrequent) manipulation. The critical investigation is whether the preview validity effect differs across the six types of strings.

All participants' accuracy was above 85%, indicating decent engagement in the reading task (mean: 95.6%; min: 85.7%; max: 100%). Twenty-two people reported either zero or one trial of detecting a change. Sixteen reported detecting a change in 2-4 trials. Fifteen reported detecting a change in 5-10 trials. Five reported detecting a change in 10% of all trials and two reported 25% and 33% of all trials had a word change. Overall, 53 out of 60 participants noticed no more than 10% of the actual invalid preview trials.

Raw reading duration measures (first-fixation duration, gaze duration, and go-past time) were first log-transformed before entering Bayesian linear regression models to alleviate the issue with skewed residual distributions.

Trials with blinking or track loss during the very first fixation on the target 4-character region were discarded. Trials with the boundary change occurring early (before participants' eyes cross the boundary) or too slow (9-ms or more after eyes already were on the post-boundary region) were discarded. These resulted in 12% trial loss. Two additional trials with the total number of fixations in the sentence fewer than 3 were discarded. Trials with first-fixation duration shorter than 80 ms on the target 4-character region were discarded, leading to an additional 0.9% of trials removed. Finally, trials with

no fixation on any region prior to the target region were discarded (2%). On average, 20% out of the 15 trials in each condition (NN/VO/AN \times frequent/infrequent \times valid/invalid) were discarded (see Appendix H).

For all the Bayesian models reported in Sections 3.2.5.2., the effect of frequency and the effect of validity will be sum-coded (frequent: 0.5; infrequent: -0.5; invalid: -0.5; valid: 0.5) and the 3-level effect of structure will be dummy-coded (VO: 0,0; NN: 1,0; AN: 0,1). This means the intercept corresponds to the mean of VO strings and the simple effects are simple effects at the level of VO. I chose this contrast scheme because uncovering simple effects in a model containing a 3-way interaction is extremely difficult. Setting a particular level of a factor as baseline/reference level by using dummy coding will make interpretation of the intercept and the coefficients directly interpretable. However, this comes with a cost of increasing uncertainty of estimates for simple effects at the other non-baseline/reference levels, due to more interaction terms involved in the structural equation to derive the simple effects. Given this asymmetry in estimate uncertainty, I chose to use VO as the baseline/reference level to maximize the possibility of finding a simple preview validity effect at this level, at the cost of decreasing the possibility of finding a simple preview validity effect at the NN or AN level. Given that preview benefit has been demonstrated with NN and AN strings (Zang et al., 2024), I am more certain a priori that a preview validity effect could be replicated and therefore less concerned about not having enough statistical power for these two string structures.

Table 14 first shows the descriptive statistics of all reading measures on the pretarget, C1C2, C3C4, and C1C2C3C4 regions.

Pretarget	Skipping Prob. (%)		First-fixation Duration		Gaze Duration		Go-past Time	
	valid	Invalid	Valid	Invalid	valid	invalid	valid	invalid
Freq NN	35 (3.2)	34 (3.4)	216 (6.64)	212 (4.87)	236 (9.95)	233 (8.15)	301 (13.1)	289 (14.6)
Infreq NN	31 (3.3)	35 (3.2)	222 (5.98)	218 (5.93)	247 (9.54)	234 (7.93)	306 (15.2)	296 (15.4)
Freq VO	28 (3.0)	27 (2.8)	230 (5.90)	221 (5.81)	265 (9.69)	253 (10.3)	347 (16.2)	325 (15.8)
Infreq VO	26 (2.6)	27 (2.7)	224 (5.97)	229 (7.10)	258 (10.4)	257 (10.3)	322 (18.6)	321 (16.2)
Freq AN	35 (3.1)	34 (3.4)	219 (5.77)	216 (6.24)	244 (9.90)	238 (10.4)	305 (15.8)	289 (14.2)
Infreq AN	34 (3.0)	34 (3.1)	211 (5.86)	211 (5.37)	236 (9.59)	240 (9.97)	284 (13.1)	283 (14.0)
C1C2								
Freq NN	29 (2.8)	28 (3.2)	231 (5.83)	243 (5.76)	250 (7.98)	270 (9.18)	302 (13.2)	349 (15.8)
Infreq NN	26 (2.8)	27 (3.1)	235 (5.84)	245 (6.23)	257 (8.19)	275 (9.2)	328 (15.0)	343 (14.0)
Freq VO	29 (2.9)	26 (3.5)	236 (6.15)	240 (6.36)	271 (10.4)	269 (10.4)	342 (16.3)	356 (18.5)
Infreq VO	30 (3.4)	28 (3.3)	241 (6.46)	239 (6.38)	275 (10.4)	274 (10.1)	333 (14.8)	356 (16.6)
Freq AN	28 (2.8)	26 (3.2)	235 (5.66)	249 (6.52)	258 (8.24)	271 (8.54)	340 (16.5)	344 (15.0)
Infreq AN	28 (3.0)	27 (3.3)	241 (6.11)	252 (6.95)	268 (8.8)	282 (10.1)	333 (14.2)	356 (18.7)
C3C4								
Freq NN	29 (2.4)	24 (2.3)	226 (5.33)	230 (5.48)	249 (8.38)	251 (7.02)	321 (12.7)	319 (12.9)
Infreq NN	28 (2.5)	28 (2.7)	238 (6.42)	235 (5.16)	265 (8.88)	259 (6.9)	339 (14.2)	349 (13.2)
Freq VO	26 (2.2)	25 (2.4)	232 (5.57)	222 (5.59)	259 (7.52)	245 (7.92)	340 (15.9)	335 (19.6)
Infreq VO	28 (2.7)	23 (2.4)	231 (6.37)	235 (6.12)	264 (9.58)	270 (9.53)	356 (14.4)	388 (19.5)
Freq AN	29 (2.5)	27 (2.5)	227 (5.72)	228 (6.33)	249 (7.64)	251 (8.63)	293 (10.9)	306 (12.6)
Infreq AN	27 (2.5)	27 (2.6)	230 (5.75)	232 (4.85)	255 (7.63)	255 (6.93)	334 (11.6)	339 (14.3)
C1C2C3C4								
Freq NN	3.9 (1)	3.4 (1)	231 (5.4)	242 (5.43)	368 (18.6)	402 (20.2)	468 (25.0)	516 (26.2)
Infreq NN	4.1 (1.1)	4.7 (1.4)	239 (5.71)	249 (5.98)	411 (21.2)	421 (20.9)	521 (28.3)	535 (25.8)
Freq VO	4.3 (1.1)	4.0 (1.2)	240 (5.91)	240 (6.11)	415 (22.4)	405 (21.5)	522 (27.8)	548 (30.6)
Infreq VO	7.2 (1.7)	4.9 (1.4)	241 (6.15)	243 (5.75)	439 (25.0)	442 (23.7)	535 (27.8)	593 (32.9)
Freq AN	4.0 (1)	4.1 (1.5)	235 (5.77)	247 (6.05)	381 (19.1)	403 (21.0)	482 (28.5)	505 (26.0)
Infreq AN	4.9 (1.3)	3.4 (1)	241 (5.51)	252 (5.71)	407 (19.4)	418 (18.8)	508 (23.3)	536 (28.3)

Table 14. Descriptive statistics of reading measures on different regions. By-subject standard errors in parentheses.

3.2.5.2.1. Pretarget region

Table 11 above showed that strings in the pretarget region slightly differed in frequency among the three string types (although all were rather frequent). The syntactic category of the strings in the pretarget region also differed across the three string types (with a determiner/possessive/quantifier for NN and AN, but an auxiliary verb or adverb for VO). To examine the influence of these properties of the pretarget strings, a Bayesian linear regression model on gaze duration of this region was run. Another question is whether there is a parafoveal-on-foveal effect of preview validity. Note that the pretarget region was 3-4 characters in front of the characters whose validity was manipulated. At this distance, the parafoveal-on-foveal effect typically had not been observed (Zang et al., 2021; Zang et al., 2023; Zang et al., 2024, but see Yu et al., 2016).

Table 15 shows the Bayesian linear regression model on gaze duration on the pretarget (2-character) region. There were effects of target string structure, with VO's pretarget region read longer (261 ms) than NN (242 ms) and AN (240 ms). There was also an interaction between NNvsVO contrast and frequency, meaning that the structural difference was stronger when the target strings were frequent (236 ms vs. 265 ms). Furthermore, there was an effect of pretarget strings' frequency, which was an expected, canonical frequency effect. This effect was smaller when the target strings were frequent. This might be because, as seen in Table 11, in the frequent target groups there was less variation in pretarget string's frequency, by design error.

Regarding parafoveal-on-foveal validity effects, no reliable effects associated with preview validity were observed, consistent with (Zang et al., 2024).

Gaze duration	Estimate	SE	95-CI (low)	95-CI (high)
Intercept	5.43	0.03	5.36	5.49
NNvsVO	-0.04	0.02	-0.08	-0.01
ANvsVO	-0.03	0.02	-0.07	0.01
Validity	0.01	0.02	-0.02	0.05
Frequency	0.03	0.03	-0.02	0.08
NNvsVO×Validity	0.02	0.03	-0.03	0.07
ANvsVO×Validity	0.00	0.03	-0.05	0.06
NNvsVO×Freq	-0.06	0.04	-0.14	0.01
ANvsVO×Freq	-0.02	0.04	-0.10	0.06
Validity×Freq	0.04	0.04	-0.03	0.10
NNvsVO 3-way	-0.06	0.05	-0.16	0.03
ANvsVO 3-way	0.00	0.05	-0.10	0.11
Pretarget Freq	-0.03	0.01	-0.05	-0.01
PreFreq×Validity	0.00	0.01	-0.02	0.03
NNvsVO×PreFreq	0.01	0.02	-0.03	0.04
ANvsVO×PreFreq	-0.03	0.02	-0.08	0.02
PreFreq×Freq	0.03	0.02	-0.00	0.06

Table 15. Bayesian linear regression model of gaze (log-transformed) duration on the pretarget region. VO is the reference level. NNvsVO: contrast subtracting VO from NN; ANvsVO: contrast subtracting VO from AN. PreFreq: frequency of the string on the pretarget region, log-transformed and centered. NNvsVO 3-way: three-way interaction among validity, frequency, and structure (NN-VO). ANvsVO 3-way: three-way interaction among validity, frequency, and structure (AN-VO). All Rhats =1.00. The default priors in the brms package were used.

Since an effect of pretarget strings' frequency and an interaction with the effect of structure were seen already at the pretarget region, I added the pretarget strings' frequency and its interaction with preview validity as covariates for the analysis on the target region. The concern is that foveal load (frequency) might affect parafoveal processing: that is, if the currently fixated string is requiring the reader more effort to process, the reader might have less resource to preprocess the string in the parafovea (Henderson & Ferreira, 1990; Kennison & Clifton, 1995, but see Zhang et al., 2019 and Section 3.6.3 for discussion). Note that the experiment was not designed to address this specific question, and efforts had been made to equalize the frequency of the pretarget strings. Still, the potential frequency difference in the pretarget region across the six

groups (Table 11) motivated this statistical decision. As this effect and its interaction would simply serve as covariates, I will report them only in the tables, not in the text, in the results section. In Section 3.6.3, I will discuss it in relation to the current main research questions.

For each region, I will first visualize the posterior distributions of preview benefit (using the invalid preview conditions as the baselines for condition subtraction) in hopes of easier grasp of the full picture. Each figure is based on the four Bayesian models whose actual statistics will be reported in text and in a table following the figure.

To plot the figure of posterior distributions of preview benefit, each posterior sample was first transformed back from log-RT to raw RTs through exponentiation. Each posterior distribution contains all samples from the 4 Monte Carlo Markov Chains (5000 samples each) from a Bayesian model. Pretarget strings' frequency (and its interaction with string structure and frequency) as covariates, though entered in the models, were omitted for visualization.

3.2.5.2.2. First constituent region (C1C2)

Figure 11 shows the posterior distributions of preview benefit for different measures at the C1C2 region. As mentioned in Section 3.2.4.2., it is expected that skipping probability, first-fixation duration, and gaze duration are most likely to differentiate different types of strings. Here I found that except for skipping probability, all three other measures distinguish VO strings from NN and AN strings. There seems no strong evidence for frequency modulating the preview benefits.

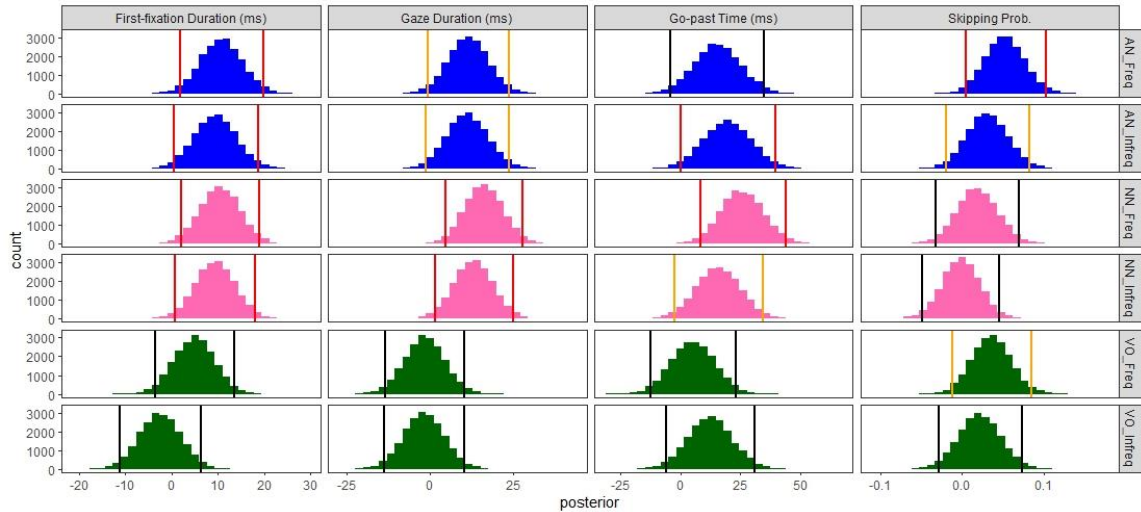


Figure 11. Posterior distributions of preview benefit in each of the six experimental conditions on the four eye movement measures on the C1C2 region. Vertical bars reflect 95% Credible Intervals; red indicating not overlapping with zero; orange indicating 89% Credible Intervals not overlapping with zero.

Table 16 shows the Bayesian models that Figure 11 was based on. Simple effects are in reference to VO strings (aggregated across frequent and infrequent strings). For skipping probability, there was only a hint of a validity effect such that valid preview induced more skipping of C1C2. For first-fixation duration, there was an effect of structure (ANvsVO) and a strong hint of its interaction with validity, such that AN's first-fixation was longer but only when it was with an invalid preview (252 ms vs. 239 ms²³). There were also interactions between structure (NNvsVO) and validity and between structure (ANvsVO) and validity. This suggests preview benefits were larger for NN strings (11 ms) and AN strings (12 ms) than for VO strings (1 ms), as Figure 11 also suggests. There was also a weak hint of an interaction between validity and frequency ([-0.09, 0.02]), such that preview benefit was larger when the target strings had higher

²³ All numbers in the parentheses are based on Table 14's group means. Arithmetic averaging is applied to further average out group means whenever needed.

frequency (4 ms vs. -2 ms), for VO specifically.

For gaze duration, there was an effect of structure (NNvsVO) and a reliable interaction with validity. This means that NN's preview benefit was reliably bigger (19 ms) than VO's (-1 ms). This difference in preview benefit size was also true for ANvsVO, given another interaction observed (13 ms vs. -1 ms).

For go-past time, there was a reliable effect of preview validity. However, there was a hint of its interaction with NNvsVO, suggesting that while preview validity effects were present for both (in fact all three) structures, it was larger for NN than for VO (31 ms vs. 18 ms).

Skipping Probability					First-fixation Duration			
	Est.	SE	95-CI (low)	95-CI (high)	Est.	SE	95-CI (low)	95-CI (high)
Intercept	-1.30	0.20	-1.69	-0.91	5.42	0.02	5.37	5.46
NNvsVO	-0.06	0.09	-0.24	0.12	-0.00	0.01	-0.03	0.03
ANvsVO	-0.10	0.10	-0.29	0.10	0.03	0.01	-0.00	0.05
Validity	0.17	0.11	-0.03	0.39	-0.01	0.01	-0.03	0.02
Frequency	-0.07	0.13	-0.31	0.18	-0.02	0.02	-0.05	0.02
NNvsVO×Validity	-0.12	0.14	-0.40	0.16	-0.04	0.02	-0.08	-0.00
ANvsVO×Validity	0.09	0.15	-0.20	0.40	-0.04	0.02	-0.08	0.00
NNvsVO×Frequency	0.24	0.18	-0.11	0.60	-0.00	0.03	-0.05	0.05
ANvsVO×Frequency	-0.02	0.19	-0.38	0.35	0.00	0.03	-0.05	0.05
Validity×Frequency	0.09	0.21	-0.31	0.51	-0.03	0.03	-0.09	0.02
NNvsVO 3-way	0.03	0.28	-0.53	0.58	0.03	0.04	-0.05	0.10
ANvsVO 3-way	0.06	0.28	-0.49	0.61	0.03	0.04	-0.05	0.10
Pretarget Frequency	0.01	0.04	-0.07	0.09	-0.01	0.01	-0.02	0.00
Pretarget Freq×Validity	-0.13	0.06	-0.25	-0.01	-0.01	0.01	-0.03	0.01
Gaze Duration					Go-past Time			
Intercept	5.51	0.03	5.45	5.57	5.65	0.04	5.58	5.72
NNvsVO	-0.03	0.02	-0.06	0.01	-0.02	0.03	-0.07	0.03
ANvsVO	0.00	0.02	-0.03	0.04	0.02	0.03	-0.03	0.07
Validity	0.01	0.02	-0.03	0.04	-0.03	0.02	-0.08	0.01
Frequency	-0.02	0.02	-0.03	0.04	-0.01	0.04	-0.08	0.06
NNvsVO×Validity	-0.07	0.02	-0.12	-0.02	-0.04	0.03	-0.11	0.02
ANvsVO×Validity	-0.05	0.03	-0.10	-0.00	-0.03	0.03	-0.10	0.04
NNvsVO×Frequency	-0.01	0.03	-0.08	0.05	-0.04	0.05	-0.14	0.06
ANvsVO×Frequency	-0.01	0.03	-0.08	0.05	-0.00	0.05	-0.10	0.10
Validity×Frequency	-0.00	0.03	-0.07	0.06	0.02	0.04	-0.06	0.11
NNvsVO 3-way	-0.01	0.05	-0.11	0.08	-0.06	0.06	-0.19	0.06
ANvsVO 3-way	-0.00	0.05	-0.10	0.09	-0.01	0.07	-0.14	0.12
Pretarget Frequency	-0.01	0.01	-0.03	0.00	-0.04	0.01	-0.06	-0.02
Pretarget Freq×Validity	-0.01	0.01	-0.03	0.01	0.02	0.02	-0.01	0.05

Table 16. Bayesian linear regression models of four reading measures on the C1C2 region. VO is the reference level. Pretarget frequency log-transformed and centered. All Rhats = 1.00. The default priors in the brms package were used.

3.2.5.2.3. Second constituent region (C3C4)

Figure 12 shows the posterior distributions of preview benefit for different measures at the C3C4 region. As mentioned in Section 3.2.4.2., No effects are expected here, based on Zang et al. (2024). Indeed, the only preview benefits were found on skipping probabilities. Since no effects were expected at this region, I will only report the effects observed in the section but refrain from further discussing them.

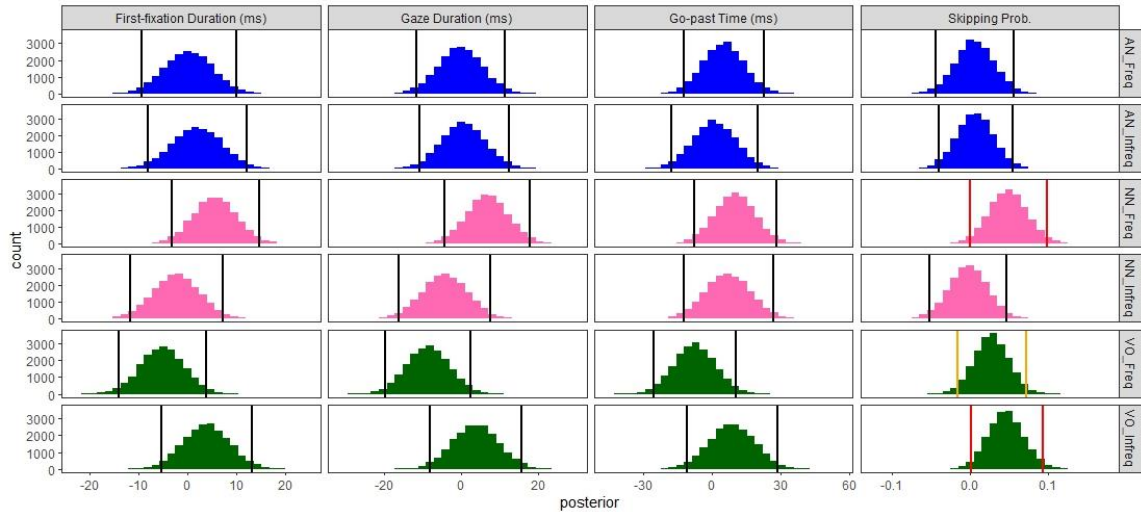


Figure 12. Posterior distributions of preview benefit in each of the six experimental conditions on the four eye movement measures on the C3C4 region. Vertical bars reflect 95% Credible Intervals; red indicating not overlapping with zero; orange indicating 89% Credible Intervals not overlapping with zero.

Table 17 shows the Bayesian models that Figure 12 was based on. Simple effects are in reference to VO strings (aggregated across frequent and infrequent strings). For skipping probability, there were effects of structures, both NNvsVO and ANvsVO, such that the object in a VO was skipped (25.5%) less than the head noun in an NN (27.3%) or AN (27.5%), on average. There was also an effect of preview validity: the object in an VO was skipped more when the preview was valid (27% vs 24%). Finally, there was a hint of a three-way interaction (NNvsVO \times Frequency \times Validity). As can be seen in Figure 12 above, frequency modulated the validity effects in opposite ways for NN and VO.

For first-fixation duration, there was a hint of an interaction between validity and frequency in an unexpected direction such that the frequent VO group showed a *benefit from an invalid preview* (-10 ms). There was also a strong hint of a 3-way interaction with NNvsVO: while the frequent VO group showed a reverse preview benefit, the frequent

NN group showed a trend for a canonical preview benefit (4 ms). For gaze duration, the pattern was very similar (-14 ms vs. 2 ms). Additionally, there was a hint of an effect of frequency, specifically for VO (15 ms, although as seen from Table 14, it was mostly due to frequent VO strings *with invalid preview* being read especially fast).

For go-past time, there was an effect of structure: AN strings elicited shorter go-past time on average (318 ms) than VO strings (355 ms). Second, infrequent VO strings elicited longer go-past time than frequent VO strings (34 ms), regardless of validity.

Skipping Probability					First-fixation Duration			
	Est.	SE	95-CI (low)	95-CI (high)	Est.	SE	95-CI (low)	95-CI (high)
Intercept	-1.36	0.15	-1.67	-1.07	5.37	0.02	5.33	5.42
NNvsVO	0.14	0.08	-0.01	0.29	0.01	0.02	-0.02	0.04
ANvsVO	0.17	0.08	0.01	0.32	-0.00	0.02	-0.04	0.03
Validity	0.23	0.10	0.03	0.43	0.00	0.02	-0.03	0.03
Frequency	0.00	0.11	-0.21	0.21	-0.03	0.02	-0.07	0.02
NNvsVO×Validity	-0.09	0.14	-0.37	0.19	-0.01	0.02	-0.05	0.03
ANvsVO×Validity	-0.19	0.14	-0.47	0.09	-0.01	0.02	-0.05	0.04
NNvsVO×Frequency	-0.07	0.15	-0.37	0.23	-0.01	0.03	-0.07	0.05
ANvsVO×Frequency	0.10	0.15	-0.20	0.39	0.01	0.03	-0.05	0.07
Validity×Frequency	-0.11	0.19	-0.50	0.27	0.04	0.03	-0.02	0.10
NNvsVO 3-way	0.41	0.27	-0.11	0.95	-0.08	0.04	-0.16	0.01
ANvsVO 3-way	0.11	0.27	-0.41	0.63	-0.03	0.04	-0.12	0.05
Pretarget Frequency	-0.02	0.04	-0.10	0.06	0.01	0.01	0.00	0.03
Pretarget Freq×Validity	0.10	0.06	-0.02	0.22	-0.01	0.01	-0.03	0.01
Gaze Duration					Go-past Time			
Intercept	5.46	0.03	5.41	5.51	5.64	0.03	5.58	5.70
NNvsVO	-0.01	0.02	-0.05	0.03	-0.02	0.03	-0.08	0.03
ANvsVO	-0.02	0.02	-0.07	0.02	-0.05	0.03	-0.11	0.01
Validity	0.01	0.02	-0.02	0.05	-0.00	0.02	-0.05	0.05
Frequency	-0.05	0.03	-0.11	0.01	-0.09	0.04	-0.17	-0.01
NNvsVO×Validity	-0.02	0.02	-0.07	0.03	-0.03	0.03	-0.10	0.04
ANvsVO×Validity	-0.01	0.03	0.06	0.04	-0.01	0.03	-0.08	0.06
NNvsVO×Frequency	-0.01	0.04	-0.09	0.07	0.01	0.06	-0.11	0.12
ANvsVO×Frequency	0.03	0.04	-0.05	0.11	0.01	0.06	-0.10	0.12
Validity×Frequency	0.05	0.03	-0.02	0.12	0.06	0.05	-0.04	0.15
NNvsVO 3-way	-0.10	0.05	-0.20	-0.01	-0.07	0.07	-0.20	0.06
ANvsVO 3-way	-0.05	0.05	-0.15	0.05	-0.07	0.07	-0.20	0.06
Pretarget Frequency	0.02	0.01	-0.00	0.03	-0.01	0.01	-0.04	0.01
Pretarget Freq×Validity	-0.01	0.01	-0.03	0.01	-0.01	0.02	-0.04	0.02

Table 17. Bayesian linear regression models of four reading measures on the C3C4 region. VO is the reference level. Pretarget frequency log-transformed and centered. All

Rhats = 1.00. The default priors in the brms package were used.

3.2.5.2.4. Whole-string region (C1C2C3C4)

Figure 13 shows the posterior distributions of preview benefit for different measures at the C1C2 region. Since Chinese readers barely skip 4-target strings, no model was run for the skipping measure. As mentioned in Section 3.2.4.2., it is expected that first-fixation duration and gaze duration are most likely to differentiate different types of strings. This appears to be the case: gaze durations distinguish VO strings from NN and AN strings. Moreover, frequency seems to modulate the preview benefits for NN strings in particular.

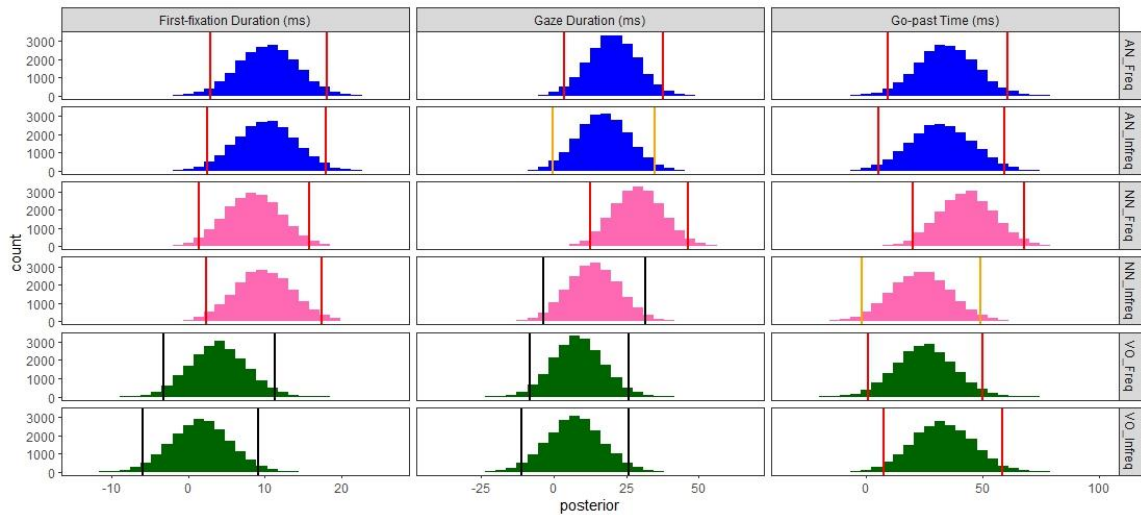


Figure 13. Posterior distributions of preview benefit in each of the six experimental conditions on the three duration measures on the whole target string region. Vertical bars reflect 95% Credible Intervals; red indicating not overlapping with zero; orange indicating 89% Credible Intervals not overlapping with zero.

Table 18 shows the Bayesian models that Figure 13 was based on. Simple effects are in reference to VO strings (aggregated across frequent and infrequent strings). For first-fixation duration, AN strings were read slower than VO strings (244 ms vs 241 ms).

There was a hint of a small validity effect for VO strings (1 ms²⁴). Structure interacted with validity such that the validity effects were stronger for NN and AN strings than for VO strings (10 ms for NN, 1 ms for VO ,and 11 ms for AN). For gaze duration, there were effects of structure and strong indications of interactions: preview benefits were bigger for NN (22 ms) and AN strings (17 ms) than for VO strings (-3 ms). Within VO strings, there was a strong effect of frequency (30 ms), and an indication of small preview benefit. However, when plotted out separately for frequent and infrequent VO strings (Figure 13 above), the preview benefit for each did not seem reliable.

For go-past time, VO strings were again on average longer (550 ms) than NN (510 ms) or AN strings (508 ms). Frequent VO strings also had shorter go-past time than infrequent VO strings (29 ms), regardless of preview validity.

²⁴ Note as per the last footnote, the numbers in the parentheses are based on arithmetic group means. They might differ from the model-based mean differences that take into accounts random item and subject effects.

Skipping Probability					First-fixation Duration			
	Est.	SE	95-CI (low)	95-CI (high)	Est.	SE	95-CI (low)	95-CI (high)
Intercept					5.42	0.02	5.38	5.47
NNvsVO					0.00	0.01	-0.02	0.02
ANvsVO					0.02	0.01	-0.00	0.04
Validity					-0.01	0.01	-0.04	0.01
Frequency					-0.01	0.02	-0.04	0.02
NNvsVO×Validity					-0.03	0.02	-0.06	0.00
ANvsVO×Validity					-0.03	0.02	-0.07	0.00
NNvsVO×Frequency					-0.02	0.02	-0.06	0.03
ANvsVO×Frequency					-0.00	0.02	-0.05	0.04
Validity×Frequency					-0.01	0.02	-0.05	0.03
NNvsVO 3-way					0.02	0.03	-0.05	0.08
ANvsVO 3-way					0.01	0.03	-0.06	0.07
Pretarget Frequency					-0.01	0.00	-0.02	0.00
Pretarget Freq×Validity					-0.01	0.01	-0.03	0.00
Gaze Duration					Go-past Time			
Intercept	5.87	0.05	5.77	5.97	6.06	0.05	5.95	6.17
NNvsVO	-0.05	0.02	-0.09	-0.00	-0.04	0.03	-0.10	0.01
ANvsVO	-0.04	0.02	-0.08	0.01	-0.03	0.03	-0.09	0.03
Validity	-0.02	0.02	-0.06	0.01	-0.07	0.02	-0.11	-0.03
Frequency	-0.05	0.03	-0.11	0.00	-0.05	0.04	-0.13	0.02
NNvsVO×Validity	-0.04	0.03	-0.09	0.01	-0.01	0.03	-0.07	0.04
ANvsVO×Validity	-0.03	0.03	-0.08	0.02	-0.01	0.03	-0.07	0.05
NNvsVO×Frequency	-0.01	0.04	-0.09	0.08	-0.03	0.05	-0.13	0.08
ANvsVO×Frequency	0.00	0.04	-0.08	0.08	-0.01	0.05	-0.11	0.10
Validity×Frequency	-0.00	0.04	-0.07	0.07	0.01	0.04	-0.07	0.09
NNvsVO 3-way	-0.04	0.05	-0.14	0.05	-0.07	0.06	-0.18	0.05
ANvsVO 3-way	-0.01	0.05	-0.11	0.09	-0.02	0.06	-0.14	0.09
Pretarget Frequency	0.01	0.01	-0.01	0.03	-0.03	0.01	-0.05	-0.01
Pretarget Freq×Validity	-0.03	0.01	-0.05	-0.00	0.01	0.01	-0.02	0.03

Table 18. Bayesian linear regression models of three duration measures on the whole target string region. VO is the reference level. Pretarget frequency log-transformed and centered. All Rhats=1.00. The default priors in the brms package were used.

3.2.5.2.5. Summaries of results of preview validity on eye movement measures

On the region of the first embedded constituent (C1C2, Figure 11), for skipping, there was a preview benefit for frequent and infrequent AN strings, and an indication of preview benefit for frequent VO strings. For first-fixation duration, preview benefit was robustly found for NN and VN but not for VO strings. For gaze duration, preview benefit was found for NN and an indication of it for AN strings, but nothing for VO strings.

Similarly, only NN and AN strings showed some evidence of preview benefit on go-past time.

On the region of the second embedded constituent (C3C4, Figure 12), for skipping probability, preview benefit was found in frequent NN, frequent VO and infrequent VO strings. For duration measures, preview validity had almost no effects for all string types and frequency. However, an odd trend of a reverse preview validity (i.e., *preview benefit* for the invalid preview) was seen for frequent VO strings on gaze duration. Overall, it is safe to say preview validity has minimal effects on this region.

On the region of the whole target string (C1C2C3C4, Figure 13), the results overall are qualitatively the same as those on the first embedded constituent (C1C2) region. Most interestingly, there was an indication of preview benefit being larger for frequent NN strings than for infrequent NN strings in gaze duration. This was consistent with a post-hoc model that directly examined the preview validity between frequent NN and infrequent NN strings (34 ms vs. 10 ms, Validity \times Frequency: [-0.12, 0.02], see Appendix I). AN strings—although throughout the time course showing patterns similar to NN strings— on the other hand, did not show any indication of a preview validity \times frequency interaction in a post-hoc model (22 ms vs. 11 ms, interaction: [-0.08, 0.05], see Appendix I).

3.2.5.3. Supplementary analysis (data with valid preview only)

In addition to the main analysis (the preview effect and its interaction with frequency/string type), a supplementary analysis was also conducted. This supplementary analysis included only trials of the valid preview condition. This analysis directly compared frequent and infrequent strings and directly compared the three string types, as

well as the interaction between frequency and string type. This is a supplementary analysis, rather than the main analysis, since the embedded first two characters were not identical across the frequent and infrequent conditions. Recall for example, Jiang et al. (2020) investigated the effect of co-occurrence frequency by comparing pairs (e.g., 参加会议 vs. 参加游戏 *attend a meeting* vs. *attend a game*) where the first two characters were the same and only the whole-frequency of the string differed. The lexical mismatch in my experiment therefore was not ideal for such an investigation. However, to the extent that frequency measures were carefully matched across the 6 string types, comparison among the six might still be informative.

For exploratory purposes, I only reported two eye movement measures: landing position (which character is first fixated, within the region defined) and number of fixations during first-pass reading (how many fixations are made before eyes first leave either to the left or right from the region defined). These two coarse measures (cf., millisecond-based duration measures) respectively reflect the earliest process (landing position is similar to the skipping measure; the saccade decision is made before the string is directly fixated) and a relatively late process (an aggregate measure similar to gaze duration). For landing position, rather than treating a skip as *NA*, I re-coded a skip as 2 (for a 2-character region), or 4 (for a 4-character region), meaning that for this trial the landing position of is 2/4 characters ahead. These models also included pretarget strings' frequency and its interaction with target strings' frequency/structure as covariates, but they will not be reported in text or in the figure.

All eye movement measures are conducted with the regions defined in three different ways (the first embedded constituent, C1C2; the second embedded constituent, C3C4; the

whole string C1C2C3C4). Note unlike the main analysis, the 3-level effect of structure will be dummy-coded (VO: 0,0; NN: 1,0; AN: 0,1); the effect of frequency will also be dummy-coded (Frequent: 1; Infrequent:0). Simple effects thus are in reference to infrequent VO items. This is because the analysis is only on trials with valid preview and thus does not involve a three-way interaction. Using dummy-coding makes it easy to directly interpret the intercepts, simple effects, and interactions.

As in the main results, I first report a visualization of the posterior distributions from the supplementary analysis (Figure 14). Note that what is plotted here is not any effect but raw measurements (landing position in units of characters and number of fixations). The takeaway from the plot is that landing position is not quite a sensitive measure: the only difference observed is the eyes landed closer for infrequent NN strings than infrequent VO strings (whole-string region). For number of fixations, it is quite clear that NN strings benefit from frequency at all three regions. AN strings showed the same pattern, although to a lesser degree at C3C4. On the other hand, frequency effects were found for VO strings at C3C4 and the whole region, *but not as early as at C1C2*. The next few paragraphs describe the actual statistics.

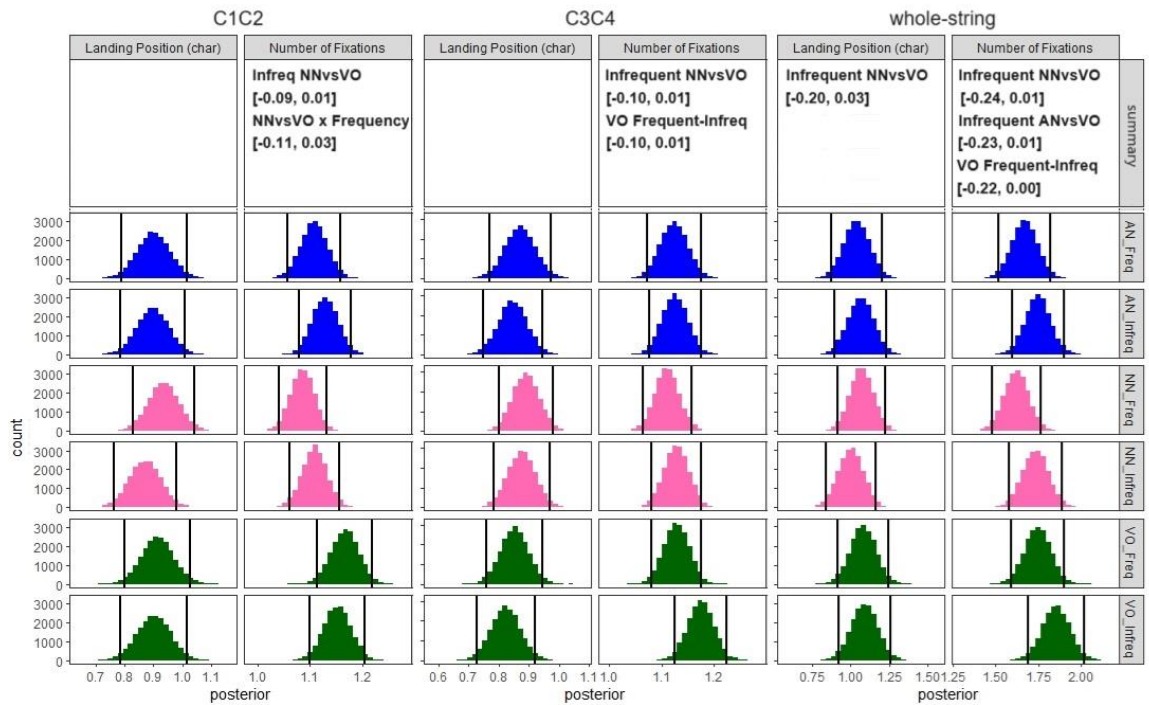


Figure 14. Posterior distributions of landing position and number of fixations in each of the six experimental conditions, at C1C2, C3C4, and C1C2C3C4. Vertical bars reflect 95% Credible Intervals, not colored since these are not distributions of any effects. Landing position is in character unit.

On C1C2 region, for landing position, there were no reliable effects. While the frequency effect did not strongly interact with NNvsVO, the trend was pointing to a direction where frequent NN strings had farther mean landing position than infrequent NN strings ([-0.07, 0.17]).

For number of fixations, there was a strong hint of NNvsVO difference ([-0.09, 0.01]). There was a weak hint of an interaction between the two ([-0.11, 0.03]). AN strings showed a similar pattern ([-0.08, 0.03] for ANvsVO, and [-0.11, 0.04] for the interaction). Altogether, these numerical trends suggest that verbs require more processing. Verb processing did not benefit from co-occurrence frequency of the whole sequence, while the first noun of NN strings and the adjective of AN strings numerically

appeared so.

On C3C4 region, for landing position, there were no reliable effects or even weak hints observed. For number of fixations, there were again strong hints of effects of structures, with VO strings receiving higher number of fixations ($[-0.10, 0.01]$ for NNvsVO; $[-0.11, 0.01]$ for ANvsVO). There was also a strong hint of frequency effect ($[-0.10, 0.01]$). Thus, strings of all 3 structures had a benefit from the embedded two components being frequently co-occurring at this later stage.

On the whole string region (C1C2C3C4), for landing, there was a hint of a difference between infrequent VO and infrequent NN strings ($[-0.20, 0.03]$), with landing being *shorter for the latter*, which is unexpected if one assumes NN strings are more one-unit like than VO strings. Note while not close to reliable ($[-0.08, 0.24]$), the interaction between frequency and string structure was positive, meaning that the structure effect in this unexpected direction was just with infrequent strings. No difference between AN and VO strings was found.

For number of fixations, there was an effect of NNvsVO ($[-0.24, -0.01]$), a hint of an effect of ANvsVO ($[-0.23, 0.01]$), and a target strings' frequency effect ($[-0.22, 0.00]$). No other effects were found or hinted. This suggests co-occurrence frequency effects were across the board facilitatory for processing of the whole string, yet VO strings are harder to process than NN/AN, regardless of frequency.

3.2.6. Discussion

3.2.6.1. Caveats on statistical inferences

Bayesian (logistic) linear models were used for analysis in Experiment 3. Compared to Experiments 1 and 2, a much higher number of models were run on different regions

and for different measures. While unlike frequentist approaches, a large number of observations is not required and no planned tests are needed for credible interval construction, discretion is still needed when using heuristic decision rules to make conclusions, as is the case in Experiment 3 whose result interpretations' were based on 89% and 95% credible intervals (Kruschke, 2021). It should thus be acknowledged that false positive discoveries might exist in the results reported.

In light of this issue, I would first note how this current experiment can be seen as a conceptual replication of Zang et al.'s (2024) Experiment 3. Both the one-word group and ambiguous group in Zang et al. used predominantly noun and adjective as the first two characters and another noun as the last two characters, thus paralleling my noun-noun and adjective-noun groups. The frequency was not controlled for in Zang et al. and varied widely within each group, but both their one-word group and ambiguous group had mean co-occurrence frequency much higher than my infrequent groups and slightly lower than my frequent groups. Their two-word group had mean co-occurrence frequency comparable to my infrequent groups.

As mentioned in Section 3.2.4.2., since no significant preview validity effects were found on C3C4 in Zang et al. (2024) except for skipping probability for the two-word group (which I in fact replicated with my most two-word-like group, verb-object), I consider all the other effects found on C3C4 here potentially false positive discoveries that await replications in the future, and therefore would not interpret them in the discussion.

On the other hand, effects for the most diagnostic measures—first fixation and gaze durations on C1C2 and on the whole string—were successfully replicated, differentiating

the verb-object strings from the other two string structures. This main finding should not be considered as a potential false positive discovery. Regarding skipping of C1C2, the preview benefit was not replicated for noun-noun compounds, the reason of which remains unclear. Note however in Figure 11 there clearly was a trend of preview validity for skipping of C1C2 for frequent noun-noun strings. Finally, the preview benefit for go-past time at C1C2 and the whole string was also largely replicated.

3.2.6.2. The word $n+2$ preview effect as a function of 4-character co-occurrence frequency and string structure

Experiment 3 factorially manipulated preview validity, string structure, and string frequency with 4-character material. The preview manipulation was on the third and fourth characters (hence on word $n+2$ under theories that categorically distinguish single-words and multiple-words). It has been demonstrated that Chinese readers can be sensitive to stimuli that are this far from their current fixation, but this sensitivity is not ubiquitous for all kinds of strings in the parafovea (Zang et al., 2024). Both string frequency and string structure might modulate the word $n+2$ preview effect: a frequently co-occurring string has been hypothesized to be distinctly stored (Contreras Kallens & Christiansen, 2022; Divjak & Caldwell-Harris, 2015; Zang, 2019), making word $n+2$ in essence part of $n+1$. These representations may not be limited to idioms or compound nouns (Arnon & Snider, 2010; Jiang et al., 2020). Indeed, in the present Experiment 1 and Experiment 2 we saw that even literal, compositional verb-object (and even subject-verb) sequences, when sufficiently entrenched, appear to have corresponding representations to provide extremely early top-down facilitation. On the other hand, some evidence suggested that noun-noun or modifier-noun sequences are even more likely to

be represented or are represented more strongly (e.g., by comparing results from Yu et al., 2016 and Zang et al., 2023). Experiment 3 thus orthogonally manipulated string structure and string frequency and examined how the preview effects are similar/different across the six string types (cf., Zang et al., 2024).

The results are clear. Across all fixation measures frequent NN strings are the group that showed most consistent preview benefit, followed by frequent AN strings. Infrequent NN strings and infrequent AN strings also showed certain degrees of preview benefit, although numerically smaller. Frequent NN strings showed a larger preview benefit than infrequent NN strings on the whole string region. This last finding differentiates NN strings from AN strings. On the other hand, no preview benefit was found in almost all measures for VO strings, even the frequent ones, except for go-past time²⁵. The difference between NN/AN strings and VO strings in preview benefit sizes was supported by a two-way interaction for several measures.

Overall, Experiment 3 suggests that during sentence reading, Chinese readers do orthographically process more than two characters in the parafovea, as evidenced in go-past time across all conditions. However, when the first two characters constitute a verb, readers do not process the third and fourth characters in the parafovea beyond visuo-orthography. The degree of processing of a string in the parafovea differs between NN/AN strings and VO strings even when they are all frequent strings. This contrasts with Experiment 1 and Experiment 2 where NN and VO strings were processed in very similar fashions when both were frequent. This highlights the role of linguistic task in

²⁵ Go-past time is most sensitive when containing regressive eye movements, which could be triggered even by trans-saccadic visuo-orthographic mismatches. This might be why preview validity effects are commonly found for this measure (see Table 12).

online usage of MCUs. I will discuss more about the differences between Experiment 3 and the previous two experiments, as well as broader implications for MCUs and the mental lexicon in Chapter 4. The remainder of this chapter focuses on the mechanisms during sentence reading that give rise to the observations.

3.2.6.3. Online sensitivity to morphosyntactic structure of the incoming strings

One interpretation of the results is that Chinese readers are sensitive to the morphosyntactic structure of the incoming strings not yet fixated and will allocate attentional resources differently depending on the structure. Drieghe et al. (2018)'s experiment using the boundary-change paradigm with one-character strings shed light on this issue. Their design was different from Experiment 3 since the invisible boundary was placed immediately before the one character that went through the change (i.e., a word $n+1$ preview manipulation), and the invisible boundary lay within a two-character compound string (cf., mine lying in the beginning of the target string). Two types of compound strings were used (NN and AN). Their results showed that on the second character, while preview validity effects were found in both single-fixation duration and gaze duration measures for both NN and AN strings, the effects were bigger for AN strings than for NN strings. Based on the findings, the authors concluded that the extent to which readers process stimuli in the parafovea is influenced by the morphosyntactic structure of the currently fixated word.

A similar finding of such quick adjustment of where to focus parafoveally was recently reported by Luo, Tan, and Yan (2023). The critical comparison was between MNN and MMN strings. Using a natural sentence reading paradigm (i.e., no boundary change), it was found that forward saccade was farther for MNN than for MMN. It was

concluded that readers can evaluate the morphological structure of the string in the parafovea and attempt to target the head/stem of the string as the saccade goal.

One difference between Experiment 3 and Drieghe et al. (2018) and Luo et al. (2023) is that for my NN/AN and VO conditions, the strings differed already at the pretarget region. For NN/AN, my pretarget region always contained a determiner/quantifier/possessor that is very predictive of a noun coming next, while for VO, my pretarget region always contained an auxiliary verb or a temporal adverb that is very predictive of a verb.

The use of preceding context in segmenting the upcoming strings has been demonstrated (Huang, et al., 2021; Huang & Li, 2024). For AN strings, it is possible that after processing the pretarget word, the readers looked for a noun and thus focused farther to the right, having sensed that the first two characters in the parafovea were not a noun. This would be consistent with Drieghe et al. and Luo et al. where the head/stem of the string in the parafovea was selected for the next word to be processed. Presumably, this strategy should apply whether the AN string is frequent or infrequent, as long as the first two characters clearly indicate an adjective, and thus might be why there was no reliably big difference in preview benefit between frequent AN and infrequent AN groups.

For NN strings, it is theoretically possible that the readers adopted the same strategy—since the head/stem for a NN compound also lies in the last two, not the first two characters—although it might be less likely. This is because the first two characters also were a noun, and in many cases—although not for all the items—the first noun was compatible with its previous context (e.g., in example sentences (5a) and (5d), I SUSPECT (that) YOUR HIGHSCHOOL is fine, and I WANT TO REVISE MY

MERCHANDISE is also rather fine). Thus, for the frequent NN case it is most likely that the greater processing to the far right in the parafovea reflects that the whole 4-character-string was stored and processed as one unit.

On the other hand, this syntactic-prediction hypothesis will predict that readers only will exclusively focus on the first two characters in the VO condition. This is because the first two characters are the most predicted POS (verb) and the head. Moreover, pure verbs (that is, if the string contains no nouns at all) are exclusively of 1 to 2 characters in Chinese. Therefore, readers will not parafoveally process up to 4 characters, unless the whole VO strings are what they are directly predicting. The fact that I did not observe a preview validity effect suggests that the verb stems per se, but not the whole VO strings, were what the readers predicted.

3.2.6.4. Effect of foveal load on parafoveal processing

Still another possibility regarding the role of preceding context is the foveal-load-on-parafoveal-processing hypothesis (Henderson & Ferreira, 1990; Kennison & Clifton, 2002; Veldre & Andrews, 2018b). We saw in Table 11 that strings in the pretarget region slightly differed in frequency among NN/VO/AN conditions (although all rather frequent): the VO groups (whether frequent or infrequent) both had slightly lower mean pretarget strings' frequency compared to their NN and AN counterparts. In the analysis of pretarget reading time, it was also found that, even with pretarget strings' frequency as a covariate, there still was a processing time difference between VO and NN/AN conditions in the pretarget region. Possibly then, the difficulty in processing an auxiliary verb (or a temporal adverb), compared to processing a determiner, made it harder to process farther parafoveally. This difference in processing difficulty between an auxiliary

verb and a determiner can be of semantic or syntactic nature, given that frequency already has been controlled for in the statistical model.

Recently, Zhang et al. (2019) addressed the question of foveal load on parafoveal processing in Chinese sentence reading. They manipulated the frequency of the two-character pretarget word in a one-character boundary change paradigm. There, a preview benefit was observed whether or not the pretarget word was frequent or infrequent, and there was no observable difference in the benefit sizes. One possibility for the findings is a ceiling effect. That is, using only a one-character word as the target word and setting the boundary immediately preceding it, it may be possible to get a robust effect regardless of the frequency of the pretarget word. Note that Zhang et al. did find forward saccade length influenced by pretarget strings' frequency. It was simply that difference in fixation durations between the preview conditions was not found to be influenced by pretarget strings' frequency.

In Experiment 3, I included pretarget strings' frequency and its interaction with preview validity, simply out of potential concern. Overall, the interaction was neither ubiquitous nor uncommon²⁶. Except for one case (skipping of C1C2), the direction of the interactions found in all other cases was in the expected direction (i.e., the harder the foveal word the less the parafoveal processing). The interactions were observed most reliably on skipping of the C3C4 region and on first-fixation and gaze durations of the whole string region. Overall then, there were some indications of foveal load modulating

²⁶ A spillover effect was also found frequently (i.e., the ease of processing of the pretarget region persists to facilitate the next word). There was also frequently a main effect of pretarget strings' frequency in the go-past time measures which is expected since go-past time can include reading time on the pretarget region. The last type of pretarget strings' frequency effect, when interacting with preview validity (a *positive* interaction, which was observed once), is not a foveal-load on parafoveal processing effect.

the extent of parafoveal processing, through frequency of the pretarget strings. This is not entirely consistent with the recent study by Zhang et al. (2019), which future studies can shed light on.

More importantly, here, even when pretarget strings' frequency was controlled for, the manipulated structure and frequency effects as well as their interactions were still demonstrated to affect parafoveal processing. However, target strings of different structures also already differed in the pretarget region in word classes. This difference could have modulated the preview benefit, either caused by processing load or by (syntactic) prediction, as discussed in the previous section. Regardless of the cause, a highly dynamic attention allocation strategy for processing texts in the parafovea is suggested in this experiment (Drieghe et al., 2018; Henderson & Ferreira, 1990; Liu, Reichle, & Li, 2015; 2016; Luo et al., 2023).

3.2.6.5. String structure and frequency effects during normal sentence reading

A supplementary analysis that only included trials with valid preview was done to directly examine the effects of string structure and frequency on reading 4-character strings. In the English literature, multiword frequency effects in eye-tracking studies were only reported for binomial strings or were confounded with idiomaticity (Carrol & Conklin, 2020; Siyanova-Chanturia et al., 2011), and on whole-string region and on gaze duration, rather than as early as first-fixation duration. In the Chinese literature, Wang et al. (2023) used noun-noun compounds and did report first-fixation durations of different regions, finding that whole-string frequency effect only occurred as early as on first-fixation duration on the second constituent (C3C4). There was also a frequency effect on gaze duration on the whole string (C1C2C3C4). In work by Jiang and colleagues (2020;

2023) who studied verb-object sequences, a frequency effect also occurred only on first-fixation duration on the second constituent. For the whole string region, Jiang et al. (2020) found an effect on gaze duration, and a trend on first-fixation duration; Jiang and Siyanova-Chanturia (2023) did not even find a significant effect on gaze duration, but only a trend of it. It thus remains a question how fast a multiword frequency effect can be.

Here, since I used a different design from the three abovementioned Chinese studies, my first two characters were not matched lexically across the six conditions (although frequency was carefully matched). Therefore, I only explored coarse eye movement measures: landing position and number of fixations. These two measures reflect the earliest process (landing position is similar to the skipping measure, the programming decision of which is made before the string is directly fixated) and a relatively late process (an aggregate measure similar to gaze duration).

For landing position, barely any differences were found, except for a trend of infrequent NN strings' landing being closer than VO strings on the C1C2C3C4 region²⁷.

For number of fixations, the frequency effect was across all three strings on both the C3C4 region and on the whole string, consistent with both Jiang and colleagues (2020; 2023) and Wang et al. (2023). This could reflect strength of representation or familiarity of the whole string but could also reflect ease of integration/composition (Onnis & Huettig, 2021), given that it is a relatively late measure. Furthermore, VO strings required more fixations on all three regions (C1C2, C3C4, and C1C2C3C4), and, unlike NN and

²⁷ . This only finding about infrequent NN strings could indicate that infrequent NN strings were initially segmented as two separate words, and the participants were trying to target the first noun as a saccade goal. This effect was not as reliable and did not interact strongly with frequency ([-0.08, 0.24]). Therefore, I refrain from interpreting it too strongly.

AN strings, did not benefit from whole-string frequency as early as on C1C2. This is strong evidence of serial processing of the two components and evidence of structural building. The difference likely reflects (a) difference in the verb vs. noun/adjective itself (that is, a single-word-level difference, Matchin et al., 2019; Perani et al., 1999), (b) relational processing of argument structure (Liao et al., 2022), or both.

Overall, the current results suggest a quite stark difference in how frequency effects unfold between NN/AN strings and VO strings, although future studies with better lexical control can be even more informative.

3.2.6.6. Stored representations or on-the-fly composition? II

Chapter 2 ended up with a possibility that the co-occurrence frequency effect on POS judgment was due to rapid cascaded composition, since C1C2 and C3C4 very likely have a tighter relationship in FCOs than in BCOs. Given that parafoveal validity effect presumably occurred during the last fixations (which lasted on average 250 ms) prior to first focal fixation on the target word, it is also possible that C1C2 and C3C4 had been separately recognized and composed within this time window.

Consider first an extreme case where two embedded words were recognized strictly serially from the parafovea²⁸. In the current case, having fully recognized C1C2 might facilitate recognition of C3C4 more when the whole string is frequent than when it is infrequent (Onnis & Huettig, 2021). However, one piece of evidence that strongly rules out this strictly sequential recognition possibility, is that, in the NN conditions, this scenario should have sometimes led to local implausibility or led to the need to re-

²⁸ Indeed, word $n+2$ preview effect was initially explained to be a result of sequential processing of words in the parafovea. Specifically, it could explain earlier findings that the word $n+2$ preview effect only was obtained when word $n+1$ was easy to recognize (Yang et al., 2012).

interpret the strings, which should have induced observable processing cost (Wang et al., 2023). This possible incremental issue does not exist for the AN sequences. I found, however, for the valid preview condition, all reading measures at all regions were shorter for frequent NN sequences than frequent AN sequences. This suggests frequent NN strings did not go through incremental recognition of the embedded two nouns that sequentially got composed/integrated.

Moving to parallel processing of two separate words, a few differences in the settings across the two paradigms require considering. First, instead of being presented centrally, the target strings in Experiment 3 were viewed from character $n-1$, $n-2$, or even $n-3$. This eccentricity of viewpoint very likely made perceptual evidence of the last two characters really weak. Note that in Li et al. (2009) where participants' fixation was always at the center of the first character of the 4-character string, some degree of asymmetric recognition already emerged. At the viewing distance in the current experiment, even less perceptual evidence about C3C4 may be accumulated. Hence, the possibility of partially activated syntactic/semantic/lexical features from the two separate embedded words bootstrapping for the two words to be composed prior to full recognition seems low.

On top of this difference in visual acuity, it also must be noted that there was additionally ongoing processing from the pretarget region, compared to the paradigm in Experiments 1 and 2. Therefore, this 250 ms window needs to encompass the concurrent processes for the currently foveated characters and the composition processes for the upcoming 4 characters. The two factors considered (i.e., highly asymmetric visual acuity and ongoing processing of the pretarget string), it does not seem likely that the processing ease for frequent NN strings in Experiment 3 was due to on-the-fly composition.

I argue that the most plausible explanation is that most frequent NN strings (and perhaps several, but not as many, frequent AN strings) are represented as one unit (MCUs), and the linguistic information from the whole string can be extremely quickly utilized, when activated by minimal perceptual evidence, to guide attention allocation for parafoveal lexical processing and target for a saccade goal. Whether or not VO strings are eligible to be stored/represented as MCUs remains a question. One possibility is that they can, at least in terms of pure visual perception, but sentence reading/comprehension requires more than visually recognizing stored units. I will discuss this further in Chapter 4.

CHAPTER 4

GENERAL DISCUSSION

4.1. Findings in a nutshell

Experiments 1 and 2 adopted a part-of-speech (POS) judgment task with rapid parallel presentation (17 ms) of character strings. On each trial participants judged one of the characters embedded in a two-character string (Experiment 1) or a four-character string (Experiment 2). The to-be-judged character was probed after a 200 ms mask that almost immediately followed the character string. In the baseline condition, each trial began with a brief cue of which character will be probed. This allows participants to pay covert attention to the to-be-probed character in advance. This pre-cued baseline condition was compared to the condition when no precue was available. The performance difference between the two conditions was evaluated against three predictions: serial processing, limited-capacity parallel processing, and fully parallel processing. Length, structure, and frequency of strings were also manipulated.

Across the two experiments, it was found that (a) strings that contain two words that together are meaningless or anomalous (VV, SV, VNVV, VVNV) had a performance difference below, or right at, the limited-capacity parallel prediction, (b) strings that contain two plausible, but not-frequently-co-occurring words had a performance difference above the limited-capacity parallel prediction but below the fully parallel prediction (NN-BCO, VVOO-BCO), and (c) strings that contain two meaningful and frequently-co-occurring words had a performance difference compatible with the fully parallel prediction (NN-FCO, VO-FCO, NNNN-FCO, VVOO-FCO, SSVV-FCO). Three

subgroups that did not fit these categorizations were VO-BCO (fully-parallel), NNNN-BCO (fully-parallel), and SSVV-BCO (below limited-capacity parallel). Despite the last few exceptions, it was clear that FCOs exhibited a great deal of ease of processing compatible with what a fully-parallel-word-recognition-model will predict.

Given this observation (specifically, the similarity between NNNN-FCO and VVOO-FCO strings), Experiment 3 examined processing of 4-character-strings during naturalistic sentence reading. In addition to NNNN (noun-noun) and VVOO (verb-object) comparisons, AANN (adjective-noun) strings were included. String frequency was again manipulated, factorially along with string structure. The experiment adopted a boundary-change paradigm to inspect the extent to which Chinese readers process strings to the right of the current fixation. In one condition, the last two characters (C3C4) of the target string were presented invalidly while participants still were fixating at any region prior to the target string. In the other condition, the target string was presented normally throughout the time participants read. By comparing these two conditions for different reading measures, it can be inferred how much processing has occurred for C3C4 before any characters in the string were directly fixated. Results showed that, among the six string groups, frequent noun-noun showed the greatest processing facilitation from a valid preview of the last two characters. There also were indications of processing facilitation from a valid preview for the infrequent noun-noun and frequent as well as infrequent adjective-noun strings, although less robust and reliable. Most importantly, there was little evidence of such facilitation for either frequent or infrequent verb-object strings. Taken together, the extent to which Chinese readers can process the third and fourth of the characters in the upcoming string in a sentence is a function of both the

structure and the frequency of the string. Finally, additional analysis on only trials with a valid preview again hinted a structure×frequency interaction at an early locus: for number of fixations at C1C2, frequency effects were found for noun-noun and adjective-noun strings, but not for verb-object strings. On the other hand, at later loci (C3C4 and whole string), frequent strings received lower numbers of fixations, regardless of string structure.

To summarize the three experiments, effects of co-occurrence frequency of character strings were robustly found, yet when and what extent they played a role in the two linguistic tasks depend on the structure of the string. Facilitation in rapid visual recognition, parafoveal processing, and late foveal processing was found in all three experiments for frequent noun-noun compounds. In contrast, for verb-object sequences, facilitated processing was only found in rapid visual recognition and late foveal processing, not in parafoveal processing during sentence reading.

The findings support the existence of multiword representations but highlight the role of linguistic tasks. The findings also impose constraints on how these stored units are utilized during reading. I argue that while language users are highly sensitive to statistical regularities of various word sequences, this possibly yields only familiarity with the surface multiword forms and ease of on-the-fly composition for verb-object sequences. Compound nouns on the other hand may be lexically stored to have direct form-meaning mapping via frequent exposure, hence the additional early facilitation observed.

Two main implications from the experiments will be discussed: (1) what types/levels of linguistic content are stored in memory and (2) how are they accessed/utilized during reading.

4.2. The mental dictionary

The dissertation began with discussion of how a visual word is recognized, and a focus on Chinese reading was motivated by the fact that its writing system does not utilize white space as a separator. This raised the question of how Chinese readers recognize meaningful units without explicit visual cues. However, even in languages where word units can be *relatively* easily defined, there have been, and still are, debates regarding what linguistic units are stored and how they are accessed. Some argue a word's meaning can be memorized as a single atomic unit along with its full form and syntactic properties, and they can be directly retrievable (Giraudo & Grainger, 2001; Schmidtke & Kuperman, 2019). Others argue that a word's meaning must be derived from morphosyntactic composition over even smaller linguistic units (Halle & Marantz, 1994; Haspelmath, 2017; Krauska & Lau, 2023; Taft & Frost, 1975).

Regarding the Chinese literature, I have argued that word-level units do have stored representations via which semantic information can be directly accessed, based on the following evidence. First, effect sizes of frequency of embedded characters (morphemes) are much smaller relative to those of whole-word frequency in lexical decision tasks (Xiong et al., 2023; Zhang et al., 2024). Second, effects of frequency of embedded characters (morphemes) are elusive and often trumped by whole-word frequency effects in natural reading tasks (Cui et al., 2021; Hyönä et al., 2024; Xiong et al., 2023). Third, effects of semantics from the whole word emerge earlier than, or can totally override, semantics from the parts (Yang et al., 2012; Wang et al., 2024). This does not happen when the whole words are infrequent (Wang et al., 2023).

Building upon these pieces of evidence, I further ask if units bigger than words can

also be stored. Recent theories (Contreras Kallens & Christiansen, 2022; McCauley & Christiansen, 2019; Zang, 2019) have suggested that, in addition to morphemes and words, meaningful sequences of all lengths (known as multiword expressions, MWEs, or in Chinese, multi-constituent units, MCUs) are eligible for storage. Indeed, Experiments 1 and 2 of mine lent support to the existence of such stored units. Note that the critical finding is not only that all constituents embedded in MCUs can be recognized in parallel, but that MCUs made possible a fully-parallel-like pattern that their counterparts did not show. This finding applied to three structurally different MCUs: noun-noun, verb-object, and subject-verb. This can be taken as evidence for concrete usage/exemplar-based language learning (Ambridge, 2020; McCauley & Christiansen, 2019) that disregards the role of abstract syntactic categories.

The non-MCU counterparts were well matched in terms of the subparts' frequencies, and notably, the fully-parallel-like pattern for MCUs was observed with *extremely* rapid presentation (17 ms + 0-34 ms of a blank frame + 200 ms mask). These two reasons make the MCUs being stored a highly likely inference. Still, one can argue for a rapid composition within 250 ms, which would not require the whole string being stored (Onnis & Huettig, 2021). Results from Experiment 3 speak against such a possibility, since the processing advantages obtained must have come while the eyes still had not fixated the target string. It is not plausible that rapid sequential word recognition and composition can be done in 250 ms for a string that is in the parafovea (see Section 3.2.6.6). Thus, the empirical data in this dissertation provide strong evidence for *stored* multiword representations.

This strong conclusion, however, only applies to noun-noun sequences, and possibly

to adjective-noun sequences (Zang et al., 2024), but not to verb-object sequences, given that there was barely evidence of a preview benefit for this string group. This difference in MCU advantages for verb-object sequences across the two paradigms raises the question whether frequent, literal verb-object sequences can be thought of as true stored MCUs. First note that the preview benefit found for frequent noun-noun sequences was in *gaze duration* of C1C2 and of the whole sequence; processing ease on these areas and stages is clear signature of facilitated lexical processing for C3C4 in the parafovea. Preview benefit of frequent verb-object sequences, on the other hand, was only displayed in go-past time of C1C2C3C4, which does not clearly index lexical processing of C3C4 (see Footnote 25). This suggests little lexical processing of C3C4 even when the verb-object sequences are frequent. Furthermore, when looking only at the trials with valid preview, co-occurrence frequency influenced number of fixations as early as at the first constituent for noun-noun strings. Verb-object strings on the other hand only benefited from co-occurrence frequency at the second constituent and at the whole region. These suggest that during sentence reading verb-object sequences do not function as one processing unit.

One possibility is that verb-object sequences *are* stored as meaningful linguistic units but cannot be accessed parafoveally as one²⁹ during sentence reading. It could be that processing difficulty at the foveated pretarget region prevents far and deep parafoveal processing. While this possibility in theory can be true, it must be noted that in the present Experiment 3, the pretarget region contains a rather frequent string and is highly predictive of a verb coming next. This setting is where one would expect to see

²⁹ Note that my data do not suggest merely a decomposition of the verb and the object but more a likely full segmentation of the two constituents.

substantial parafoveal processing. Granted, in a future experiment, one can even have a setup where the preceding context is highly predictive of the *exact 4-character string* to further encourage deep parafoveal processing of the whole 4-character string. Yet it has been shown that effects of word predictability and word frequency are additive (Goodkind & Bicknell, 2021; Shain, 2024; Staub, 2015). Thus, if multiword representations function as word-like representations, we should not expect multiword predictability (i.e., a context effect) to interact with multiword frequency, but this remains an empirical question.

Thus, another more plausible interpretation is that language users simply do not store verb-object sequences as single lexical units in the memory, and the processing advantage found in Experiment 2 for verb-object sequences may simply reflect perceptual learning. Different levels of processing of linguistic signals are supported by different brain networks (Fedorenko, Ivanova, & Regev, 2024). Familiar percepts can enjoy low-level visual/perceptual advantages (Xue and Poldrack, 2007). Visual advantages might be most relevant when the task was only to recognize the string itself and when the stimuli were flashed for only 17 ms.

Thus, when all three experiments were taken together, the results in fact cast doubt on the claim that learning of multiword sequences as stored units is blind to syntactic structure (McCauley & Christiansen, 2019). Note, however, noun-noun strings and verb-object sequences differ in many aspects: the absence/presence of an argument introducer (Matchin et al., 2019) and the separability can be thought of as syntactic/structural (Yeh, 2020, Section 1.1.4). Yet they also differ in *concepts*, with a noun-noun string denoting only the entity while a verb-object string denoting both an event and an entity. This

difference can be of semantic nature. Whether the difference in the processing patterns between the two types of sequences truly reflects a difference in syntax remains a question.

Knowing what type and size of a linguistic unit is eligible to be stored and utilized directly, the next important question is what gives rise to their emergence. While certainly *some* input must be needed for a unit to be stored (Erker & Guy, 2012), raw co-occurrence frequency is considered an unreliable and ineffective cue for learning MCUs/MWEs (McCauley & Christiansen, 2019): too many sequences would be stored simply because the two adjacent words themselves are highly frequent words. Indeed, computational modeling work has shown that a model that learns MWEs by utilizing backward transitional probability performs in a more human-like way in both child comprehension and production tasks than a model that learns by raw co-occurrence frequency (McCauley & Christiansen, 2019). Other attempts have been recently made by Houghton and Morgan (2023) with English noun-noun compounds in an experimental setting with adult readers, but they failed to find a difference in processing patterns whether familiar compounds were operationalized through co-occurrence frequency or transitional probability, contrary to McCauley and Christiansen (2019). However, the task adopted by Houghton and Morgan is a Maze task where participants were forced to make an active choice for every word presented serially. This rather unnatural task might have interrupted the holistic processing for compounds that are holistically stored.

The way of operationalization of stored multiword units via co-occurrence frequency in the current dissertation is an intuitive and practical implementation. Specifically, considering the competition account of visual word recognition (Li & Pollatsek, 2020;

Zhang et al., 2024), efforts were made to equalize the embedded words' frequency across the item groups within each experiment. Due to this frequency control, frequent co-occurrence item groups necessarily had a higher conditional probability (e.g., Table 11³⁰). Thus, the current study is not suitable to adjudicate this question, but this important question should be pursued in future studies³¹.

Finally, whether the familiarity effect (frequency or conditional probability) of multiword sequences is continuous or discrete, i.e., threshold-based, also bears a theoretical significance (Arnon & Snider, 2010; Contreras Kallens & Christiansen, 2022; Divjak & Caldwell-Harris, 2015). Under a generative account that emphasizes the role of morphosyntax (Chomsky, 1995) or a words-and-rules view (Pinker, 1999), which posit a qualitative difference between single-word-units and multiple-word-units (i.e., phrases), surface frequency can influence the processing of stored forms in the mental inventory but not on-the-fly compositional ones (Pinker & Ullman, 2002). The current study adopted a factorial design, dichotomizing items into frequent and infrequent groups with an extreme frequency manipulation. This was to maximize effect size and statistical power, given the relatively small sample size of participants. Building upon the positive findings in the current work, future studies should devise experiments statistically powerful enough to allow meaningful item-wise analysis. This can address how multiword frequency across a wide range continuously influences *online* sentence processing (cf., existing evidence for other offline/isolated-stimuli studies, Arnon &

³⁰ Conditional probability here is calculated as position-insensitive, hence slightly differs from transitional probability.

³¹ Still other probability/statistics-based cues for learning processing units have been proposed, including lexicality probability (Zang et al., 2016) or positional word probability (Liang et al., 2023; Yen et al., 2012). However, operationalizing these probabilities with 4-character strings is difficult due to the need for a pre-specified lexicon and from a pre-parsed corpus.

Snider, 2010; Caldwell-Harris, Berant, & Edelman, 2012; Jacobs, et al., 2016).

To conclude this section, the current study provided strong evidence of stored multiword representation for Chinese noun-noun compounds. The representations yield processing facilitation across the time course both in isolation or with sentence context: visual/perceptual processing, parafoveal processing beyond orthography, and foveal processing. In contrast, no evidence of such representations was found for verb-object sequences: while co-occurrence frequency does matter for sequences of this type, the processing facilitation was only seen for visual/perceptual processing and foveal processing. These findings indicate that linguistic units above word-level can be stored but not all types of multiword sequences are ready to be stored and utilized as one processing unit.

I have outlined three important future questions: (1) whether the nature of the difference between noun-noun compounds and verb-object sequences is semantic or structural, (2) whether co-occurrence frequency as opposed to other statistical regularities for multiple words is the driving factor for storage of units, and (3) whether this factor works on a continuous basis. In closing, I note that despite the qualitative differences between noun-noun and verb-object sequences, that the latter type still benefits from co-occurrence frequency at some processing stages provides strong support for the notion that language users are highly sensitive to statistical regularities of various word sequences (Erickson & Thiessen, 2015; Romberg & Saffran, 2013). However, it appears to be case that noun-noun compounds are stored in a way that allows their linguistic information to be directly utilized at various levels of sentence processing. The next section discusses how these stored units are utilized in these different processes.

4.3. Chinese reading models

The empirical results in the study have implications for models of Chinese reading. Given the robust and early effects of co-occurrence frequency that were found across the three experiments, it seems necessary to posit multiword representations. This can be either symbolic nodes or distributed connections. The corresponding orthography (for verb-object strings) and semantics (noun-noun strings) should be directly addressable without the mediation of their embedded parts (Wang et al., 2024). Parallel access to information from the parts can still be assumed (Schmidtke & Kuperman, 2019), but a model that only relies on a compulsory single route of full morphological parsing very likely will fail to explain the current data. Note also that positing stored and accessible representations does not entail that frequent noun-noun compounds do not go through morphological composition at some later point to, for example, understand the modifier-head relationship (Hsu et al., 2019).

Another implication is for the debate of serial versus parallel processing of words and how it relates to multiword sequences. Understanding of processing mode is specifically important in reading because, unlike speech, stimuli are available all at once to the comprehender and the comprehender must actively search for the next unit to process. The first two experiments suggest that orthographical processing mode goes from partly parallel for non-MCUs to fully parallel for MCUs. The third experiment suggests that even with visual eccentricity noun-noun MCUs are processed as one-unit-like starting from the parafovea while verb-object sequences show a great degree of seriality in processing. While the current versions of Li and Pollatsek's (2020) and Yu et al's (2021) models allow four characters to be recognized simultaneously, many frequent 4-character

noun-noun compounds simply are not specified to be in the inventory. This implementation issue thus needs to be addressed. A possibility is to use a large unparsed corpus and list every occurrence of a 4-character string and then apply some minimum-row-frequency filter as a gatekeeper (Erker & Guy, 2012). One can also follow McCauley and Christiansen (2019) using average backward transitional probability as the filter.

Yet even with the inclusion of more 4-character units, a more difficult problem is how to simulate the difference between noun-noun compounds and verb-object sequences. One simpler way is to further (manually) filter out all sequences that are not noun/adjective-noun and idioms from the model's vocabulary—leaving the question of how humans actually do so to the language learning/acquisition field. With this approach, a 4-character verb-object sequence will not have a corresponding node to be activated to provide feedback for the string to be processed together.

Next, what is required in a Chinese reading model is both rapid utilization of the foveated (pretarget) word and the rapid utilization of the next few characters to influence the degree of parafoveal processing. Liu et al. (2015) found that the foveated word influenced saccade targeting, and Liu et al. (2016) found that the parafoveal word also influenced saccade targeting. The model by Li and Pollatsek (2020) has implemented the modulation of foveal processing load on parafoveal processing to explain the findings by Liu et al. (2015) but there was no additional mechanism explicitly implemented to simulate Liu et al.'s (2016) results. More importantly, as discussed in Sections 2.4.6 and 3.2.6., the utilization of words in the fovea and parafovea might not solely be based on frequency and predictability (since both variables were to a great degree controlled for in Experiment 3): Even when statistically controlling for the pretarget strings' frequency, the

regression model still suggested that auxiliary verbs/temporal adverbs took longer to read than determiners/quantifiers/possessors (but see Staub, 2023, for no evidence of word class effects in English). The participants also appeared to use morphosyntactic information from the parafovea to influence their attention allocation for the stimuli in the parafovea (Drieghe et al., 2018; Luo et al., 2023). By sensing the upcoming $n+1$ and $n+2$ being an adjective, the readers might decide to look further ahead to search for the head, hence some preview benefit from valid C3C4, even for infrequent strings. By sensing the upcoming $n+1$ and $n+2$ being a verb, the readers might decide to pay exclusive attention to the first two characters, simply because (they know) verbs are hard to process. This biased focus to only the first two characters thus might reduce the activation of the object needed to activate the whole 4-character-string node; hence there is no preview benefit. Finally, the difference in landing positions among the three structures of infrequent strings also hints the utilization of morphosyntactic/semantic information in the parafovea (but see Footnote 27).

Morphosyntactic and/or thematic/semantic features thus must be utilized in a model (e.g., Rabe et al., 2024). The processing of these features might also need to be cascaded and interactive, due to how fast these effects unfold, as just illustrated regarding Experiment 3. In Experiments 1 and 2, too, we saw processing modes being influenced by structural differences among the three BCO strings as well as semantic plausibility. A model of this kind has been proposed by Snell et al. (2018; Wen, et al., 2019) for reading of languages that use white space. Such an architecture might be more pertinent for writing scripts without white space and with semantic information densely packed in characters. In the Chinese literature, models positing semantic and phonological features

also have been proposed, although they are models of character identification/word recognition, not of sentence reading (a localist model, Taft, Zhu, & Peng, 1999; a distributed model, Chang, Welbourne, & Lee, 2016).

This brings us to the final but perhaps most important point: a Chinese reading model should also incorporate a composition mechanism (Stanojevic et al., 2023; Mollica et al., 2020). Computational models of language processing or learning that were trained solely to learn word co-occurrence statistics have been shown to capture a certain degree of variation in reading times (Smith & Levy, 2013) or produce human-like performance mimicking developing children (McCauley & Christiansen, 2019). However, in the present study, the structural effects at different time courses observed across the current three experiments are highly unlikely to be attributable to frequency/predictability: foveal reading of a verb-object sequence requires more re-fixations than that of a noun-noun or adjective-noun sequence, even when frequency and predictability had been tightly controlled for (Table 11 and Table G1). Similarly, the difference between VO-BCO and SV-BCO strings—which cannot be attributed to predictability or parts of speech—may provide even more striking evidence of qualitative difference in structure building. Given these findings, a reading model with an algorithm that utilizes morphosyntactic and/or thematic/semantic features and links composition processes to reading time is clearly needed.

CHAPTER 5

CONCLUSION

5.1.Limitations, future directions, and broader implications

While the experiments in this dissertation yielded various exciting findings, several limitations must be noted. First, all three experiments involved relatively few participants or few observations per condition. Due to the sample sizes in the three experiments, I opted to not directly interpret null effects of three-way interactions in Experiment 3 and not directly run two-way interaction models for most of the analysis in Experiments 1 and 2. Instead, my inference mostly was based on the simple effects (invalid versus valid; PC versus NPC) separately tested within each string type at each frequency level. These alone cannot be taken as strong evidence of qualitative differences among different types/frequencies of strings (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). The results thus warrant replication in the future. Note that Experiment 3 here, with 60 participants and 180 items, may not be seen as a small-sized experiment, compared to most eye-tracking-in-reading experiments. However, the power analyses here indicate that it is still underpowered to reliably detect a two- or three-way interaction. Such an issue possibly applies to most other studies in the literature and requires more attention (Brysbaert & Stevens, 2018).

With this acknowledgment kept in mind, I, however, stress that all my key conclusions were at least supported by “marginal” two-way interactions. (1) Preview benefit for first-fixation and gaze durations can be shown for noun/adjective-noun

sequences but cannot be observed for verb-object sequences at C1C2 and at the whole string: all 95% CIs for NNvsVO×Validity (or ANvsVO×Validity) interactions were (almost) not overlapping with zero (Tables 14 & 16). (2) Preview benefit for gaze duration at the whole string region for noun-noun compounds was modulated by their frequency (i.e., the Validity×Frequency interaction [-0.12, 0.02], Appendix I). (3) Foveal benefit at C1C2 emerged for noun-noun strings, not for verb-object strings (i.e., the NNvsVO×Frequency interaction [-0.11, 0.03]). Finally, (4) a difference between PC and NPC condition can be found for SV-BCO strings but not for SV-FCO strings (i.e., the PCvsNPC×Frequency interaction [-0.11, 0.84], Appendix C). What is not yet strongly supported is the three-way interactions and I look forward to the revelation of them in future studies with much higher statistical power.

Aside from revealing the three-way interactions, an even much larger-scale study can also provide further insight into multiword representations, by demonstrating a continuous frequency effect across a wide range or a threshold-based frequency effect. This will require a large number of observations *per item*.

Experiment 1 and 2 also can benefit from a larger sample size. Specifically, with enough observations per item, one might even be able to plot an AOC curve for each item. It can then be seen if the NPC unfilled circle moves away from the limited-parallel curve to the fully-parallel intersection point as a particular function of co-occurrence frequency or of transitional probability. Other improvements for Experiments 1 and 2 include answering even more precisely how fast activation/composition is needed for recognition of the whole string. The response time measures in the current experiments are simply too long to be informative (about 1000 ms for PC and 1200 ms for NPC

condition), possibly due to the masking and the POS judgment task itself. Future studies can address this by, for example, allowing responses right from the onset of the masking by not positioning the mask and the probe at the same area.

Another limitation due to the relatively small sample size is the inability to directly test existing models such as that of Li and Pollatsek (2020). Based on Zhang et al. (2024), we might also expect word-level competition among the first embedded word (C1C2), the second embedded word (C3C4), and the whole multiword string (C1C2C3C4) in our current experiments. The examination of this will shed light on how similarly a stored multiword unit functions to a lexicalized word unit. This analysis also would benefit from better lexical control (e.g., creating co-occurrence frequency pairs, and in each pair exact same first two characters are shared).

This relates to another current limitation/future direction: a direct comparison of the time course of co-occurrence frequency effect between noun-noun compounds and strings of other structures. In the current Experiment 3, a difference in time course was hinted. A frequency effect emerged at C1C2 for noun/adjective-noun strings but not for verb-object strings. This, however, was based on a coarse reading measure: number of fixations during first-pass reading. Future studies can extend Jiang and colleagues (2020; 2023, who examined co-occurrence frequency effects on fixation durations of verb-object sequences) by examining *both* noun-noun compounds and verb-object sequences within the same experiment. Importantly, I argue that inspection of first-fixation and gaze durations at the C1C2 region should be critical in addressing how word-like a multiword representation is.

The trickiest part of comparing strings of different structures in sentence reading is to match the preceding context. Utilizing cutting-edge language models, I have checked that the differences in eye movements between different structures are unlikely to be due to target phrase/word predictability (Appendix G). It still leaves a possibility that other non-predictability contextual effects (e.g., semantics or syntax) have given rise to the differences in eye movements. This should be addressed empirically in the future by holding the preceding context the same across targets of different structures. The variance not attributable to the preceding contexts then requires further understanding as to whether it reflects differences in structure building or differences in meaning building of the target strings. While teasing apart syntactic and semantic composition is notoriously hard (Călinescu, Ramchand, & Baggio, 2023; Pykkänen, 2019), working toward this problem is important to understand why and how readers do not store verb-object strings as single units.

Beyond psycholinguistics, the current study has implications for pedagogy and artificial intelligence. For instructors of Chinese as a second language, it is important to emphasize word collocation and provide intensive exposure to frequently co-occurring sequences. Furthermore, extra emphasis can be made about noun-noun compounds being the dominant coinage usage in Chinese. This meta-morphosyntactic awareness might help learners not to incrementally recognize and integrate part of a long noun into the context (e.g., EAT **SEA-CREATURE**) to avoid local semantic misfit or reanalysis (Wang et al., 2023). On the other hand, for instructors of English (or any other language that does not have as many compound nouns as Chinese) as a second language, explicit corrections on overgeneralized and excessive use of noun-noun compounds can be given for L2

writing/speaking to prevent the comprehenders from local semantic misfit or reanalysis (Staub et al., 2007)³².

Finally, some insights from the current study may be shared with the field of natural language processing. A recent study (Si et al., 2023) has shown the effect of tokenization methods and vocabulary construction algorithms on downstream Chinese NLP task performance. Current common practices of tokenization and segmentation for Chinese texts is character-based with all vocabularies being single characters (e.g., BERT, Devin et al., 2018). Si et al. (2023) showed how different ways of encoding characters and constructing vocabularies can improve the model. For instance, having the vocabulary include combinations of characters (MCUs). One question is whether improving the tokenization and/or vocabulary construction algorithm to further mimic real Chinese users' mental dictionary (i.e., containing a great number of 4-character strings of noun-noun, adjective-noun, and idioms) can make language models more human-like. Other questions relate to decoding and model evaluation, such as how models with a character-based or even sub-word-based algorithm suffer specifically on reading tasks because there is little chance for them to process longer multiword chunks in parallel.

5.2. Concluding remarks

The dissertation aims at addressing two grand questions in psycholinguistics: what linguistic units can be stored in the memory and how they are utilized for linguistic tasks. I focus on multiword sequences which had not traditionally been considered stored units until two decades ago (Abbot-Smith & Tomasello, 2006). I investigated processing of

³² My apology if the readers of this dissertation have found excessive use of unusual English noun-noun compounds and got garden-pathed several times.

such units in Chinese, for its writing system and morphology can afford multiple morphemes/words to be uniformly packed within the perceptual span of a fixation (Yan et al., 2015). Frequency as well as structure of multiword sequences is manipulated. The empirical questions are how these two factors influence online processing of visual multiword sequences. Three experiments were conducted, using two paradigms.

The empirical results are clear. Results from Experiment 1 and Experiment 2 speak against Li et al. (2009) and White et al. (2018; 2020), showing fixed-capacity parallel processing of two words is possible for Chinese. The degree of parallelism increases if the two embedded words co-occur frequently, consistent with White (2023), regardless of string structure. Results from Experiment 3 largely replicate Zang et al. (2024). I extend their results by more clearly attributing the parallel lexical processing of Character $n+1$ to $n+4$ to high frequency and noun/adjective-noun structure, *not verb-object structure*. Experiment 3 and Experiments 1 and 2 therefore diverged in how frequency plays a role for verb-object sequences, which could be due to the linguistic tasks (Fedorenko et al., 2024). I have argued however that it most likely reflects storability constraints for different multiword string structures. Taken together, the results support the notion that language users are highly sensitive to statistical regularities of various word sequences (Erickson & Thiessen, 2015; Romberg & Saffran, 2013) but speak against a strong interpretation by McCauley and Christiansen (2019) who suggested multiword sequences of all structures can be stored as single linguistic units.

For future Chinese reading models, I propose that multiword representations must exist (either symbolic or distributed) and must be directly addressable (Wang et al., 2023; Wang et al., 2024). However, some mechanisms will be needed to filter out non-storable

sequences such as verb-object sequences (cf. Jiang and colleagues, 2020; 2023). The models also should be able to encode morphosyntactic or thematic/semantic features. These features should also be accessible extremely quickly in a cascaded and interactive way (Snell et al., 2018) to influence outgoing saccade targeting (Liu et al., 2015; 2016; Luo et al., 2023) and attention allocation for lexical processing of the text in the parafovea (Drieghe et al., 2018). Most importantly, the models should be able to compose structurally and/or semantically (Stanojevic et al., 2023; Mollica et al., 2020). These composition processes should be linkable to reading times. Finally, an important question is to understand why and how verb-object sequences are not stored the way noun-noun compounds are.

APPENDIX A

POWER ANALYSIS FOR EXPERIMENT 1 AND EXPERIMENT 2

Two power analyses were conducted for Experiment 1, one based on the dataset including all NN and VO items and the other based on the dataset including only NN and VO items whose co-occurrence frequency is in the first and last quartiles. The latter only involves co-occurrence frequency as an additional factor (high vs. low) and its interactions with pre-cueing and string structure. To simulate a trial response from a particular condition (e.g., infrequent NN without a precue), a linear combination was first performed as follows: $rnorm(1, mean_intercept, sd_intercept) + rnorm(1, mean_noprecue_{frequent_NN}, sd_noprecue_{frequent_NN}) + rnorm(1, mean_infrequency_{precue_NN}, sd_infrequency_{precue_NN}) + rnorm(1, mean_noprecue \times infrequency_{NN}, sd_noprecue \times infrequency_{NN}) + item_i_intercept + item_i_noprecue_{frequent_NN} + item_i_infrequency_{precue_NN} + item_i_noprecue \times infrequent_{NN} + sub_j_intercept + sub_j_noprecue_{frequent_NN} + sub_j_infrequency_{precue_NN} + sub_j_noprecue \times infrequent_{NN}$, where intercept refers to the mean correct probability of the pre-cued, frequent NN condition, $item_i$ refers to random effects associated with item i , sub_j refers to random effects associated with subject j . The output value is used as a probability to generate a binomial response (correct or incorrect) for that particular trial. Trial loss due to eyes not fixating at the center was assumed at the mean rate of 12.5%, with each participant's loss rate randomly drawn from a uniform distribution between 5% to 20%, and applied randomly to the total number of trials.

Table A1 reports standard deviations of the random effects used for simulations.

Simulation 1 (all NN and VO items)		
Random effects	Subject SD	Item SD
Intercept (NN with pre-cues)	0.03	0.06
No-precue effect (for NN)	0.01	0.03
VOvsNN (with pre-cues)	0.01	0.03
No-precue effect \times VOvsNN	0.01	0.03
Simulation 2 (only most and least frequent NN and VO items)		
Random effects	Subject SD	Item SD
Intercept (frequent NN with pre-cues)	0.03	0.01
Infrequency effect (for NN with pre-cues)	0.01	0.01
No-precue effect (for frequent NN)	0.01	0.01
VOvsNN (for frequent strings with pre-cues)	0.01	0.01
Infrequency \times No-precue (for NN)	0.01	0.01
Infrequency \times VOvsNN (with pre-cues)	0.01	0.01
No-precue effect \times VOvsNN (for frequent strings)	0.01	0.01
Infrequency \times No-precue \times VOvsNN	0.01	0.01

Table A1. Standard deviations of the random effects used for power simulations for

Experiment 1 and Experiment 2. The simulated dataset then was entered into a frequentist generalized mixed-effect model with a logit link function. All factors were sum-coded, with a full random-effect structure. A hundred experiments were simulated.

For the first simulation based on all NN and VO items, the effect of interest was the two-way interaction between pre-cueing and string structure. For the second simulation based on only the most and least frequent NN and VO items, the effect of interest was the three-way interaction among pre-cueing, string structure, and co-occurrence frequency, as well as the two-way interaction between pre-cueing and string structure. For the second simulation, in addition to a full model including the three-way interaction, another reduced model was run that only included NN items for the power of two-way interaction between pre-cueing and frequency.

Similar power analyses were conducted for Experiment 2. All parameters were the same as in Experiment 1 except that the number of trials in each condition was based on the design of Experiment 2.

APPENDIX B

THE WEIBO-HUANQIURENWU CHINESE CORPUS

HuanQiuRenWu (Global People Magazine) is a bi-monthly magazine (whose issues from June 2008 to February 2017 were available on the official website of the publisher people.com.cn). The magazine features politics, world news, finances, arts and history, and life and culture. These issues have int total 34.3 million characters.

Weibo is a Chinese microblogging website and is one of the biggest social media platforms in China, launched in late 2009 and (<https://en.wikipedia.org/wiki/Weibo>) To scrape internet text from Weibo, I adopted the Automated Search Engine Queries method by Sharoff (2006). Five hundred frequent Chinese words (1 or 2 characters) were selected from the top 8645 words from Sun et al. (2018). After sampling these 500 words, 6000 queries were made by further sampling 2 words as a pair each time from the 500 words for 6000 times. The 6000 queries (word pairs, further discarding 85 queries due to repetitions) were than used as key words on the search engine of Weibo. All microblogs from the top 50 pages of search results were collected at the time of September, 2023. However, since the search is by default sorted by years, meaning the original top 50 pages heavily contained microblogs in the most recent years, a filter was further applied to discard microblogs that were from 2021-2023. Thus, the corpus contains texts of Weibo microblogs from 2010-2020. This resulted in 70.7 million characters.

The HuanQiuRenWu and Weibo unparsed corpora were then combined into a 105-million-character contemporary internet corpus.

APPENDIX C

POST-HOC TWO-WAY INTERACTION MODEL IN

EXPERIMENT 2

	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound	R-hat	Bulk ESS	Tail ESS
SV							
Intercept	1.27	0.13	1.01	1.53	1.00	10734	11027
NPC-PC	-0.30	0.11	-0.53	-0.07	1.00	27299	13182
HF-LF	0.43	0.18	0.09	0.78	1.00	13665	12928
NPC-PC	0.36	0.24	-0.11	0.83	1.00	24729	12853

Table C1. Bayesian logistic regression models of accuracy for SV conditions in Experiment 2 including both FCO and BCO strings. Only fixed effects estimates are shown here. R formula: Correctness ~ NPCvsPC * HFvsLF (1+ NPCvsPC * HFvsLF |subj) + (1+ NPCvsPC * HFvsLF|item); PC are coded as 0 and NPC coded as 1; FC are coded as 0 and HF as 1. Chains =4, each with 8000 iterations, warmup = 4000. A logit link function was used and the default priors in the brms package were used. All Rhats = 1.00.

APPENDIX D

COMPARISONS BETWEEN CCL and WEIBO- HUANQIURENWU CORPRA

	Number of items (after ambiguity exclusion)	WF CCL	WF	F12 CCL	F12	F34 CCL	F34
Noun-Noun	24	7.89	4.82	12.02	8.51	12.09	8.53
Noun-Noun	27	0.72	0	10.26	7.08	10.20	7.15
Verb-Object	30	7.75	4.69	11.51	8.04	11.51	8.10
Verb-Object	28	0.45	0	11.28	7.83	10.18	6.98
Subject-Verb	33	7.56	4.63	11.76	8.75	11.72	8.19
Subject-Verb	29	0.57	0	10.29	7.34	10.83	7.30

Table D1. Comparisons for frequency measures from CCL corpus (2024) and from my own corpus for Experiment 2's material.

	WF CCL	WF	F12 CCL	F12	F34 CCL	F34	WF/F12 CCL	WF/F12	WF/F34 CCL	WF/F34
Freq NN	8.42 (0.55)	4.46 (0.4)	12.49 (0.78)	8.81 (0.5)	12.60 (0.77)	8.77 (0.8)	0.024 (0.022)	.017 (.01)	0.026 (0.042)	.018 (.02)
Infreq NN	2.48 (1.14)	0.83 (0.3)	12.11 (1.16)	8.77 (0.5)	11.95 (1.12)	8.70 (0.8)	1.95e-04 (4.12e-05)	.000 (.00)	1.85e-04 (2.78e-04)	.000 (.00)
Freq VO	8.06 (1.01)	4.47 (0.4)	12.53 (1.02)	8.70 (0.7)	12.00 (0.80)	8.66 (1.0)	0.023 (0.037)	.021 (.03)	0.032 (0.033)	.026 (.03)
Infreq VO	2.71 (0.54)	0.83 (0.4)	12.31 (1.33)	8.74 (0.6)	11.67 (1.03)	8.66 (1.0)	9.6e-05 (8.48e-06)	.000 (.00)	2.33e-04 (2.58e-04)	.000 (.00)
Freq AN	8.58 (0.84)	4.46 (0.4)	12.37 (0.94)	8.56 (0.8)	12.55 (0.63)	8.83 (0.9)	0.032 (0.026)	.024 (.03)	0.029 (0.034)	.019 (.02)
Infreq AN	2.21 (0.86)	0.78 (0.2)	12.24 (1.19)	8.68 (0.7)	11.75 (0.99)	8.77 (0.9)	6.9e-05 (6.17e-05)	.000 (.00)	1.36e-04 (1.64e-04)	.000 (.00)

Table D2. Comparisons for frequency measures from CCL corpus (2024) and from my own corpus for Experiment 3's material.

APPENDIX E

POWER ANALYSIS FOR EXPERIMENT 3

Power analysis for Experiment 3 was conducted based on Zang et al. (2024). Table E1 reports a Bayesian model of gaze duration on the C1C2 region of their study. Dummy-coding was used for the contrasts of both factors: for wordhood, the one-word condition was coded as the baseline (0, 0) to contrast respectively the two-word condition and the ambiguous condition; for preview validity, invalid preview was coded as 0 and valid preview coded as 1. The Bayesian model assumes a full random-effect structure, but no correlations among random effects were strongly indicated to be non-zero. All effects thus were drawn from independent distributions.

Fixed effects	Estimate	SE	95% Credible Interval-lower bound	95% Credible Interval-upper bound
Intercept	5.53	0.03	5.48	5.59
valid preview	-0.09	0.03	-0.15	-0.04
Two-word vs. one-word	0.02	0.03	-0.03	0.07
Ambiguous vs. one-word	0.05	0.02	0.00	0.10
Valid preview × Two-word vs. one-word	0.10	0.04	0.02	0.17
Valid preview × Ambiguous vs. one-word	0.03	0.04	-0.04	0.10

Table E1. Estimates of fixed effects from the Bayesian model on Zang et al. (2024)'s gaze duration on C1C2 region. RT is log-transformed. Default priors in the brms package were used. All Rhats = 1.00. All Rhats = 1.00.

In addition to the effects observed in Zang et al. (2024), the other effects (and interactions) unique to the current study were assumed to have a certain effect size based on the hypotheses (see Table 13 in the main text). For each simulated response,

depending on the condition, a random sample is drawn from multiple Gaussian distributions, the sum of which is the log-transformed gaze duration. For instance, a gaze duration in the frequent, valid-preview verb-object condition is $\text{rnorm}(1, \text{mean_intercept}, \text{sd_intercept}) + \text{rnorm}(1, \text{mean_VOvsNN}_{\text{frequent_invalid}}, \text{sd_VOvsNN}_{\text{frequent_invalid}}) + \text{rnorm}(1, \text{mean_validity}_{\text{high_NN}}, \text{sd_validity}_{\text{high_NN}}) + \text{rnorm}(1, \text{mean_validity} \times \text{VOvsNN}_{\text{frequent}}, \text{sd_validity} \times \text{VOvsNN}_{\text{frequent}}) + \text{item}_i \text{_intercept} + \text{item}_i \text{_VOvsNN}_{\text{frequent_invalid}} + \text{item}_i \text{_validity}_{\text{frequent_NN}} + \text{item}_i \text{_validity} \times \text{VOvsNN}_{\text{frequent}} + \text{sub}_j \text{_intercept} + \text{sub}_j \text{_VOvsNN}_{\text{frequent_invalid}} + \text{sub}_j \text{_validity}_{\text{frequent_NN}} + \text{sub}_j \text{_validity} \times \text{VOvsNN}_{\text{frequent}}$, where intercept refers to the frequent, invalid-preview, NN condition, item_i refers to random effects associated with item i , sub_j refers to random effects associated with subject j .

To simulate a slightly asymmetrical distribution of log-transformed gaze duration, an additional term was stochastically drawn from an exponential distribution ($\lambda = 3/10$) and added to/subtracted from a simulated response by using the following function $\text{sample}(c(\text{rexp}(4, 3), -\text{rexp}(7, 10), \text{rep}(0, 4)), 1)$. Finally, any response with a duration ms shorter than 80 ms or longer than 1200 (raw-scale) was replaced, using the following functions.

```
trialdf[which(trialdf$RT<4.39),'RT'] <- runif(sum(trialdf$RT<4.39,na.rm=T),4.39,4.6)
```

```
trialdf[which(trialdf$RT>7.08),'RT'] <- 5+rexp(sum(trialdf$RT>7.08,na.rm=T),2.5)
```

```
trialdf[which(trialdf$RT>7.08),'RT'] <- runif(sum(trialdf$RT>7.08,na.rm=T),7.08,7.13)
```

Table E2 reports standard deviations of the random effects used for simulations.

Random effects	Subject SD	Item SD
Intercept (invalid, frequent NN)	0.16	0.09
Infrequency effect (for invalid NN)	0.03	0.03
Validity effect (for frequent NN)	0.045	0.045
VOvsNN (for frequent, invalid NN)	0.06	0.09
ANvsNN (for frequent, invalid NN)	0.03	0.06
Infrequency × validity (for NN)	0.06	0.06
Infrequency × VOvsNN (for invalid)	0.06	0.06
Infrequency × VOvsAN (for invalid)	0.06	0.06
Validity × VOvsNN (for frequent)	0.045	0.09
Validity × ANvsNN (for frequent)	0.06	0.06
Infrequency × Validity × VOvsNN	0.075	0.075
Infrequency × Validity × ANvsNN	0.075	0.075

Table E2. Standard deviations of the random effects used for power simulation for Experiment 3.

APPENDIX F

ONLINE NORMING RESULTS FOR MATERIAL IN EXPERIMENT 3

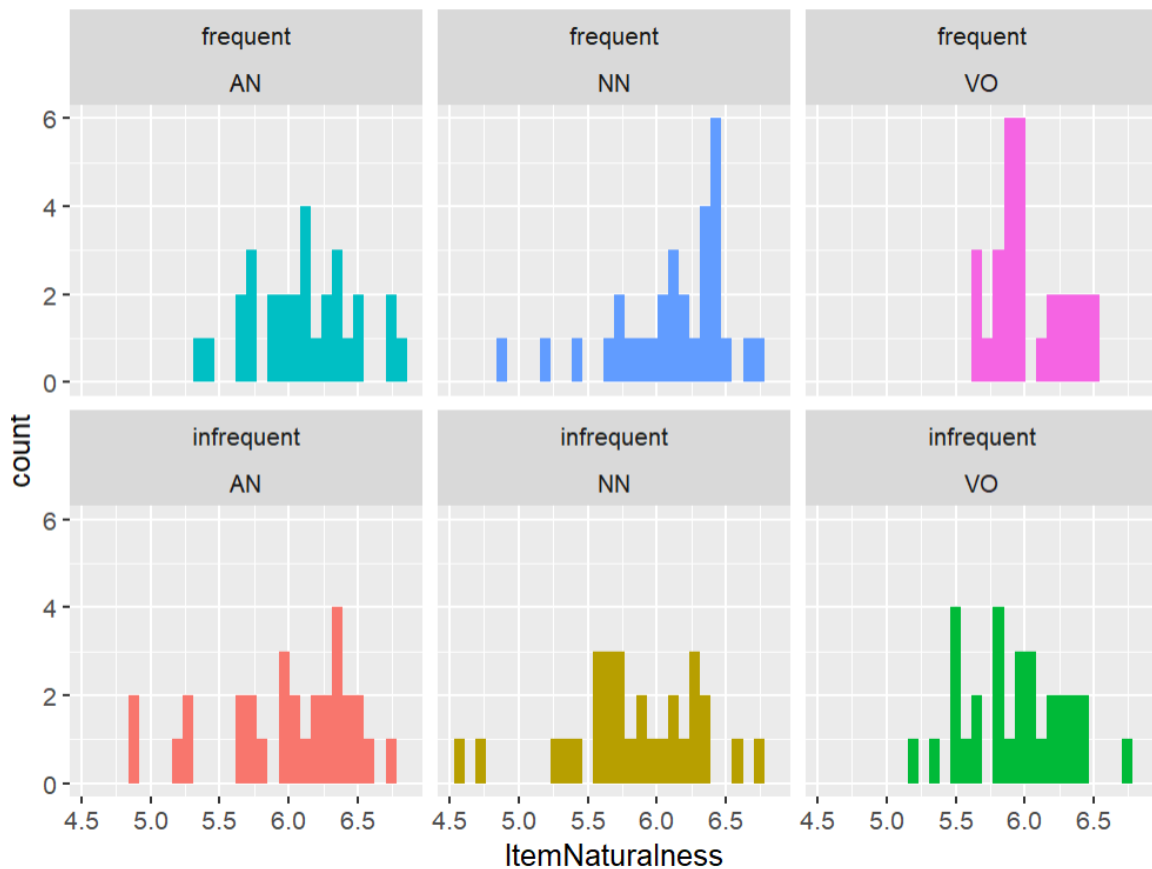


Figure F1. Distributions of naturalness of items (on a scale of 1-7) in the six conditions in Experiment 3.

Two items in the frequent NN, one item in the frequent VO, one item in the frequent AN, one item in the infrequent VO, and one item in the AN received a cloze response that was matched with the first and second character of the target string. One item in the frequent NN and one item in the frequent AN received a cloze response that was matched with the third and fourth character of the target string. As seen in Table F1, all groups essentially

had zero cloze probability.

	Frequent NN	Infrequent NN	Frequent VO	Infrequent VO	Frequent AN	Infrequent AN
C1C2 match	0.00476 (0.0146)	0 (0)	0.00238 (0.0106)	0.00238 (0.0106)	0.00264 (0.0112)	0.00238 (0.0106)
C3C4 match	0.00227 (0.0101)	0 (0)	0 (0)	0 (0)	0.00252 (0.0107)	0 (0)

Table F1. Mean cloze probability of the first two characters and the last two characters, given the context until the pretarget region.

APPENDIX G

LANGUAGE-MODEL BASED LEXICAL PREDICTABILITY

Freq	Structure	P(C1C2 pre-context)	P(Whole String pre-context)	P(C3C4 pre-context+C1C2)	P(C3C4 pre-context)
High	AN	0.00089 (0.0018)	0.000035 (0.00011)	0.136 (0.215)	0.00062 (0.00084)
High	NN	0.00077 (0.0015)	0.000017 (0.00003)	0.0530 (0.0828)	0.00202 (0.00437)
High	VO	0.00213 (0.0037)	0.000083 (0.00037)	0.0637 (0.151)	0.000036 (0.000108)
Low	AN	0.00041 (0.0008)	0.0000002 (0.0000001)	0.0022 (0.0074)	0.00245 (0.00420)
Low	NN	0.00025 (0.0005)	0.0000004 (0.0000002)	0.0012 (0.0026)	0.00189 (0.00527)
Low	VO	0.00105 (0.0019)	0.0000001 (0.0000002)	0.00049 (0.0007)	0.00001 (0.00002)

Table G1. Mean conditional probability of C1C2, C3C4, and the whole string given the preceding context of each condition, estimated by the Chinese gpt2-xl (Zhao et al., 2023).

APPENDIX H

TRIAL LOSS INFORMATION IN EXPERIMENT 3

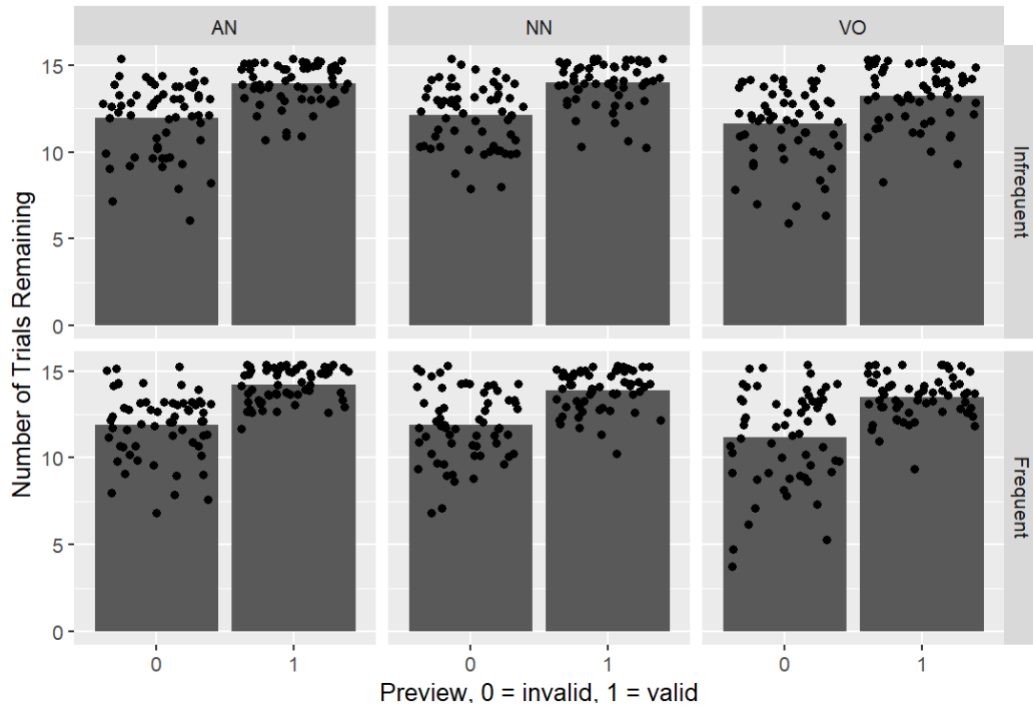


Figure H1. Mean number of trials remaining after the trial-exclusion criteria in Experiment 3. Each dot (jittered) in each bar represent a participant’s number of remaining items in that condition.

Preview	String Type	Frequency	Mean Number of trials remaining (by-subj SD)
Invalid	NN	Low	12.10 (1.81)
Invalid	NN	High	11.88 (1.97)
Invalid	VO	Low	11.61 (2.21)
Invalid	VO	High	11.20 (2.67)
Invalid	AN	Low	11.98 (2.06)
Invalid	AN	High	11.90 (1.82)
Valid	NN	Low	14.01 (1.21)
Valid	NN	High	13.90 (1.17)
Valid	VO	Low	13.23 (1.66)
Valid	VO	High	13.51 (1.20)
Valid	NN	Low	13.96 (1.19)
Valid	NN	High	14.23 (0.85)

Table H1. Mean number of trials remaining after the trial-exclusion criteria in the six conditions in Experiment 3.

APPENDIX I

POST-HOC TWO-WAY INTERACTION MODELS IN EXPERIMENT 3

Table I1 shows the Bayesian models on gaze duration (log-transformed) of the whole string region separately for just NN strings and just AN strings, with the infrequent-invalid condition as the baseline level. While for both models there was a strong hint of preview validity ([-0.09, 0.01]; [-0.10, -0.01]), only in the NN model was there a strong hint of an interaction between validity and frequency ([-0.12, 0.02]). This suggests that preview benefit was even stronger for frequent NN strings.

Gaze Duration (NN)					Gaze Duration (AN)			
	Est.	SE	95-CI (low)	95-CI (high)	Est.	SE	95-CI (low)	95-CI (high)
Intercept	5.88	0.05	5.78	5.98	5.89	0.05	5.79	5.99
Frequent	-0.04	0.03	-0.10	0.03	-0.04	0.03	-0.11	0.02
Valid	-0.04	0.02	-0.09	0.01	-0.06	0.02	-0.10	-0.01
Frequent × Valid	-0.05	0.04	-0.12	0.02	-0.02	0.03	-0.08	0.05
Pretarget Freq	0.01	0.02	-0.02	0.05	-0.00	0.02	-0.04	0.03
Pretarget Freq × Valid	-0.01	0.02	-0.05	0.02	-0.02	0.02	-0.06	0.01

Table I1. Bayesian linear regression models of gaze duration on the whole-string region for the NN and AN groups. Dummy coding was used for the effects of frequency and preview type, with infrequent condition and invalid preview as baseline. Pretarget Freq: frequency of the string on the pretarget region, log-transformed and centered.

BIBLIOGRAPHY

- Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review*, 23, 275–290.
- Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6), 509-559.
- Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, 35(6), 775-800.
- Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, 16(1-2), 285-311.
- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1), 67-82.
- Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of memory and language*, 37(1), 94-117.
- Baayen, R. H., Wurm, L. H., & Aycocock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The mental lexicon*, 2(3), 419-463.
- Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: evidence from eye movements. *Journal of experimental psychology: Human perception and performance*, 34(5), 1277.
- Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science*, 19(3), 241-248.
- Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of memory and language*, 42(3), 390-405.
- Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of memory and language*, 48(3), 615-634.
- Bowers, J. (1993). The syntax of predication. *Linguistic inquiry*, 24(4), 591-656.

- Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological review*, 74(1), 1.
- Broadbent, D. E., & Gregory, M. (1968). Visual perception of words differing in letter digram frequency. *Journal of Memory and Language*, 7(2), 569.
- Brothers, T., Hoversten, L. J., & Traxler, M. J. (2017). Looking back on reading ahead: No evidence for lexical parafoveal-on-foveal effects. *Journal of Memory and Language*, 96, 9-22.
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, 1(1).
- Burani, C., & Caramazza, A. (1987). Representation and processing of derived words. *Language and cognitive processes*, 2(3-4), 217-227.
- Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, 80, 1-28.
- Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711-733.
- Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6), e10729.
- Caldwell-Harris, C., Berant, J., & Edelman, S. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In *Frequency Effects in Language Representation*, Volume 2.
- Călinescu, L., Ramchand, G., & Baggio, G. (2023). How (not) to look for meaning composition in the brain: A reassessment of current experimental paradigms. *Frontiers in Language Sciences*, 2, 1096110.
- Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, 11(41), 63-65.
- Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, 28(3), 297-332.
- Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in cognitive sciences*, 18(2), 90-98.
- Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, 63(1), 95-122.

- Ciaccio, L. A., Kgoro, N., & Clahsen, H. (2020). Morphological decomposition in Bantu: A masked priming study on Setswana prefixation. *Language, Cognition and Neuroscience*, 35(10), 1257-1271.
- Chang, Y. N., Welbourne, S., & Lee, C. Y. (2016). Exploring orthographic neighborhood size effects in a computational model of Chinese character naming. *Cognitive psychology*, 91, 1-23.
- Chen, M., Wang, Y., Zhao, B., Li, X., & Bai, X. (2021). The trade-off between format familiarity and word-segmentation facilitation in Chinese reading. *Frontiers in Psychology*, 12, 602931.
- Chinese Linguistic Data Consortium, 2003 <https://doi.org/10.35111/n069-0642>
- Chomsky, N. (1995). Language and nature. *Mind*, 104(413), 1-61.
- Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in cognitive science*, 9(3), 542-551.
- Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M. A., & Michel, F. (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2), 291-307.
- Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of Agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, 67(3), 73-97.
- 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.
- Contreras Kallens, P., & Christiansen, M. H. (2022). Models of language and multiword expressions. *Frontiers in Artificial Intelligence*, 5, 781962.
- Crepaldi, D., Rastle, K., Coltheart, M., & Nickels, L. (2010). ‘Fell’ primes ‘fall’, but does ‘bell’ prime ‘ball’? Masked priming with irregularly-inflected primes. *Journal of memory and language*, 63(1), 83-99.
- Cui, L., Wang, J., Zhang, Y., Cong, F., Zhang, W., & Hyönä, J. (2021). Compound word frequency modifies the effect of character frequency in reading Chinese. *Quarterly Journal of Experimental Psychology*, 74(4), 610-633.
- Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science*, 9(3), 552-568.

- Cutter, M. G., Drieghe, D., & Liversedge, S. P. (2014). Preview benefit in English spaced compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), 1778.
- Davis, C. P., Libben, G., & Segalowitz, S. J. (2019). Compounding matters: Event-related potential evidence for early semantic access to compound words. *Cognition*, 184, 44-52.
- Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., ... & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, 330(6009), 1359-1364.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials. *Neuroimage*, 62(1), 381-393.
- Divjak, D., & Caldwell-Harris, C. L. (2015). Frequency and entrenchment. *Handbook of cognitive linguistics*, 39, 53-75.
- Dixon, Robert M. W. 2004. "Adjective classes in typological perspective". *Adjective Classes: a Cross-Linguistic Typology* ed. by Robert M. W. Dixon & Alexandra Y. Aikhenvald
- Drieghe, D., Cui, L., Yan, G., Bai, X., Chi, H., & Liversedge, S. P. (2018). The morphosyntactic structure of compound words influences parafoveal processing in Chinese reading. *Quarterly Journal of Experimental Psychology*, 71(1), 190-197.
- Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007). Do transposed-letter similarity effects occur at a morpheme level? Evidence for morpho-orthographic decomposition. *Cognition*, 105(3), 691-703.
- Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37, 66-108.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1), 143-149.
- van Esch, D. (2012). Leiden weibo corpus.
- Erker, D., & Guy, G. R. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, 526-557.

- Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 1-24.
- Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *The Behavioral and brain sciences*, 35(5), 310.
- Gaston, P., Stockall, L., VanWagenen, S., & Marantz, A. (2021). Memory for affixes in a long-lag priming paradigm. *GLOSSA-A JOURNAL OF GENERAL LINGUISTICS*.
- Girardo, H., & Grainger, J. (2000). Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and cognitive processes*, 15(4-5), 421-444.
- Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint arXiv:2103.04469*.
- Halle, M., & Marantz, A. (1994). Some key features of Distributed Morphology. *MIT working papers in linguistics*, 21(275), 88.
- Haspelmath, M. (2017). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 51(s1000):31–80.
- Havens, L. L., & Foote, W. E. (1963). The effect of competition on visual duration threshold and its independence of stimulus frequency. *Journal of Experimental Psychology*, 65(1), 6.
- He, L., Song, Z., Chang, M., Zang, C., Yan, G., & Liversedge, S. P. (2021). Contrasting off-line segmentation decisions with on-line word segmentation during reading. *British Journal of Psychology*, 112(3), 662-689.
- Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3), 417.
- Hoosain, R. (1992). Psychological reality of the word in Chinese. In *Advances in psychology* (Vol. 90, pp. 111-130). North-Holland.
- Houghton, Z. N., & Morgan, E. (2023). Does Predictability Drive the Holistic Storage of Compound Nouns?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, No. 45).
- Hsu, Y. W. (2015). Verb-object compound processing in Mandarin (unpublished master's thesis). National Taiwan Normal University, Taiwan.

- Hsu, C. H., Pylkkänen, L., & Lee, C. Y. (2019). Effects of morphological complexity in left temporal cortex: An MEG study of reading Chinese disyllabic words. *Journal of Neurolinguistics*, 49, 168-177.
- Huang, C.-T. J. (1984). Phrase structure, lexical integrity, and Chinese compounds. *Journal of Chinese Language Teachers Association*, 19(2), 53-78.
- Huang, K.- J., & Staub, A. (2023). The Effect of Phrase Frequency on Failure to Notice Word Transpositions. Poster at the Psychonomic Society's 64th Annual Meeting.
- 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.
- Huang, L., Staub, A., & Li, X. (2021). Prior context influences lexical competition when segmenting Chinese overlapping ambiguous strings. *Journal of Memory and Language*, 118, 104218.
- Hyönä, J., Bertram, R., & Pollatsek, A. (2004). Are long compound words identified serially via their constituents? Evidence from an eyemovement-contingent display change study. *Memory & cognition*, 32, 523-532.
- Hyönä, J., Cui, L., Heikkilä, T. T., Paranko, B., Gao, Y., & Su, X. (2024). Reading compound words in Finnish and Chinese: An eye-tracking study. *Journal of Memory and Language*, 134, 104474.
- Inhoff, A. W., & Liu, W. (1998). The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance*, 24(1), 20.
- Inhoff, A. W., Starr, M., & Shindler, K. L. (2000). Is the processing of words during eye fixations in reading strictly serial?. *Perception & Psychophysics*, 62(7), 1474-1484.
- Inhoff, A. W., & Wu, C. (2005). Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & cognition*, 33(8), 1345-1356.
- Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.
- Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, 87, 38-58.
- Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, 87, 38-58.

- Jared, D., Jouravlev, O., & Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 422.
- Ji, H., & Gagné, C. L. (2007). Lexical and relational influences on the processing of Chinese modifier-noun compounds. *The Mental Lexicon*, *2*(3), 387-417.
- Jiang, S., Jiang, X., & Siyanova-Chanturia, A. (2020). The processing of multiword expressions in children and adults: An eye-tracking study of Chinese. *Applied Psycholinguistics*, *41*(4), 901-931.
- Jiang, S., & Siyanova-Chanturia, A. (2023). The processing of multiword expressions in L1 and L2 Chinese: Evidence from reaction times and eye movements. *The Modern Language Journal*, *107*(2), 565-605.
- Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2020). Meaningfulness beats frequency in multiword chunk processing. *Cognitive Science*, *44*(10), e12885.
- Jordan, T. R. (1986). Testing the BOSS hypothesis: Evidence for position-insensitive orthographic priming in the lexical decision task. *Memory & Cognition*, *14*(6), 523-532.
- Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision research*, *45*(2), 153-168.
- Kennison, S. M., & Clifton, C. (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: implications for eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 68.
- Koriat, A., & Greenberg, S. N. (1994). The extraction of phrase structure during reading: Evidence from letter detection errors. *Psychonomic Bulletin & Review*, *1*(3), 345-356.
- Kowler, E., Anderson, E., Doshier, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision research*, *35*(13), 1897-1916.
- Krauska, A., & Lau, E. (2023). Moving away from lexicalism in psycho-and neuro-linguistics. *Frontiers in Language Sciences*, *2*, 1125127.
- Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature human behaviour*, *5*(10), 1282-1291.
- Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 876.

- Lewis, G., Solomyak, O., & Marantz, A. (2011). The neural basis of obligatory decomposition of suffixed words. *Brain and language*, *118*(3), 118-127.
- Li, X., Bicknell, K., Liu, P., Wei, W., & Rayner, K. (2014). Reading is fundamentally similar across disparate writing systems: a systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General*, *143*(2), 895.
- 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.
- Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review*, *127*(6), 1139.
- Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive psychology*, *58*(4), 525-552.
- Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.
- Li, X., & Wang, T. (2008). Lexicon of common words in contemporary Chinese.
- Libben, G. (2006). Why study compound processing? An overview of the issues. *The representation and processing of compound words*, 1-22.
- 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.
- Liao, C. H., Lau, E., & Chow, W. Y. (2022). Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language*, *126*, 104350.
- Liu, P. P., Li, W. J., Lin, N., & Li, X. S. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading?. *PloS one*, *8*(2), e55440.
- Liu, Y., Reichle, E. D., & Li, X. (2015). Parafoveal processing affects outgoing saccade length during the reading of Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1229.
- Liu, Y., Reichle, E. D., & Li, X. (2016). The effect of word frequency and parafoveal preview on saccade length during the reading of Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 1008.

- Luo, Q. (2022). Bare nouns, incorporation, and event kinds in Mandarin Chinese. *Journal of East Asian Linguistics*, 31(2), 221-263.
- Luo, Y., Tan, D., & Yan, M. (2023). Morphological structure influences saccade generation in Chinese reading. *Reading and Writing*, 36(5), 1339-1355.
- 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.
- Ma, G., Li, X., & Rayner, K. (2015). Readers extract character frequency information from nonfixated-target word at long pretarget fixations during Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, 41(5), 1409.
- Ma, G., Pollatsek, A., Li, Y., & Li, X. (2017). Chinese readers can perceive a word even when it's composed of noncontiguous characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(1), 158.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4), 676.
- Matchin, W., Liao, C. H., Gaston, P., & Lau, E. (2019). Same words, different structures: An fMRI investigation of argument relations and the angular gyrus. *Neuropsychologia*, 125, 116-128.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, 88(5), 375.
- McConkie, G. W., & Zola, D. (1984). Eye movement control during reading: The effect of word units. In *Cognition and motor processes* (pp. 63-74). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Miozzo, A., Soardi M., Cappa, SF. (1994). Pure anomia with spared action naming due to a left temporal lesion. *Neuropsychologia*, 32: 1101-9.
- Mok, L. W. (2009). Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and cognitive processes*, 24(7-8), 1039-1081.
- Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., ... & Fedorenko, E. (2020). Composition is the core driver of the language-selective network. *Neurobiology of Language*, 1(1), 104-134.
- Morton, J. (1969). Interaction of information in word recognition. *Psychological review*, 76(2), 165.

- Nazir, T. A., Jacobs, A. M., & O'Regan, J. K. (1998). Letter legibility and visual word recognition. *Memory & cognition*, 26(4), 810-821.
- Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, 14(9), 1105-1107.
- Norris, D. (2013). Models of visual word recognition. *Trends in cognitive sciences*, 17(10), 517-524.
- Öksüz, D., Brezina, V., Monaghan, P., & Rebuschat, P. (2024). Individual word and phrase frequency effects in collocational processing: Evidence from typologically different languages, English and Turkish. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication.
- ÖNEM, E. E. (2024). Processing of Internal and External Arguments in Focus in Simple Declarative Sentences in Turkish. *Dil ve Edebiyat Dergisi*, 20(1), 1-16.
- Onnis, L., & Huettig, F. (2021). Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly?. *Brain Research*, 1772, 147674.
- Paul, W. (1988). *The syntax of verb-object phrases in Chinese: Constraints and reanalysis*. Languages Croises.
- Peng, R., & Chen, J. (2004). Even words are right, odd ones are odd: Explaining word segmentation inconsistency among Chinese readers. *Chinese Journal of Psychology*, 46(1), 49-55.
- Perani, D., Cappa, S. F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., & Fazio, F. (1999). The neural correlates of verb and noun processing: A PET study. *Brain*, 122(12), 2337-2344.
- Perea, M., & Lupker, S. J. (2003). Does judge activate COURT? Transposed-letter similarity effects in masked associative priming. *Memory & Cognition*, 31(6), 829-841.
- Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In *Reading chinese script* (pp. 127-146). Psychology Press.
- Pinker, S. (1999). Out of the minds of babes. *Science*, 283(5398), 40-41.
- Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in cognitive sciences*, 6(11), 456-463.
- Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, 366(6461), 62-66.

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2024). SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language*, *135*, 104496.
- Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic bulletin & review*, *11*, 1090-1098.
- Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive psychology*, *7*(1), 65-81.
- Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision research*, *38*(8), 1129-1144.
- Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1290.
- Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of experimental psychology*, *81*(2), 275.
- Reichle, E. D., Liversedge, S. P., Pollatsek, A., & Rayner, K. (2009). Encoding multiple words simultaneously in reading is implausible. *Trends in cognitive sciences*, *13*(3), 115-119.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, *26*(4), 445-476.
- Reichle, E. D., & Schotter, E. (2020). A Computational Analysis of the Constraints on Parallel Word Identification. In *CogSci*.
- Reichle, E. D., & Yu, L. (2018). Models of Chinese reading: Review and analysis. *Cognitive Science*, *42*, 1154-1165.
- Rolls, E. T., Tovée, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *Journal of cognitive neuroscience*, *11*(3), 300-311.
- Romberg, A. R., & Saffran, J. R. (2013). Expectancy learning from probabilistic input by infants. *Frontiers in psychology*, *3*, 610.

- Scharff, A., Palmer, J., & Moore, C. M. (2011). Extending the simultaneous-sequential paradigm to measure perceptual capacity for features and words. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 813.
- Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, 116, 250-267.
- Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, 74, 5-35.
- Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. *Morphological aspects of language processing*, 2, 257-294.
- Shaw, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 277–296). New York: Routledge.
- Si, C., Zhang, Z., Chen, Y., Qi, F., Wang, X., Liu, Z., ... & Sun, M. (2023). Sub-character tokenization for Chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 11, 469-487.
- Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (2011). Seeing a phrase “time and again” matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3), 776.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302-319.
- Snell, J. (2024). The reading brain extracts syntactic information from multiple words within 50 milliseconds. *Cognition*, 242, 105664.
- Snell, J., & Grainger, J. (2019). Readers are parallel processors. *Trends in Cognitive Sciences*, 23(7), 537-546.
- Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6), 969.
- Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, 22(9), 2042-2057.
- Sperling, G., & Melchner, M. J. (1978). The attention operating characteristic: Examples from visual search. *Science*, 202(4365), 315-318.

- Stanners, R. F., Neiser, J. J., Herson, W. P., & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior*, 18(4), 399-412.
- Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., & Hale, J. T. (2023). Modeling Structure-Building in the Brain With CCG Parsing and Large Language Models. *Cognitive science*, 47(7), e13312.
- Staub, A. (2023). The function/content word distinction and eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: evidence from either... or. *Journal of experimental psychology: Learning, memory, and cognition*, 32(2), 425.
- Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 1162.
- Stevens, P., & Plaut, D. C. (2022). From decomposition to distributed theories of morphological processing in reading. *Psychonomic Bulletin & Review*, 29(5), 1673-1702.
- Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD) A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods*, 50, 2606-2629.
- Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, 18(5), 523-534.
- Sybesma, R. (1999). *The Mandarin VP*. Springer Science & Business Media.
- Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast?. *Memory & cognition*, 37, 529-540.
- Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and cognitive processes*, 9(3), 271-294.
- Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology Section A*, 57(4), 745-765.
- Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6), 638-647.

- Taft, M., & Zhu, X. (1995). The representation of bound morphemes in the lexicon: A Chinese study. *Morphological aspects of language processing*, 293-316.
- Taft, M., Zhu, X., & Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language*, 40, 498–519.
- Talgar, C. P., Pelli, D. G., & Carrasco, M. (2004). Covert attention enhances letter identification without affecting channel tuning. *Journal of vision*, 4(1).
- Taylor, W. L. (1953). Cloze Procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.
- Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2), 569-613.
- Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior research methods*, 50, 1763-1777.
- Tsang, Y. K., & Zou, Y. (2022). An ERP megastudy of Chinese word recognition. *Psychophysiology*, 59(11), e14111.
- Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, 49, 1503-1519.
- Vasilev, M. R., & Angele, B. (2017). Parafoveal preview effects from word N+ 1 and word N+ 2 during reading: A critical review and Bayesian meta-analysis. *Psychonomic Bulletin & Review*, 24, 666-689.
- Veldre, A., & Andrews, S. (2018a). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language*, 100, 1-17.
- Veldre, A., & Andrews, S. (2018b). How does foveal processing difficulty affect parafoveal processing during reading?. *Journal of Memory and Language*, 103, 74-90.
- Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, 1(1), 59-85.
- Wang, Y., Li, Z., Jiang, M., Long, F., Huang, Y., & Xu, X. (2024). Time course of Chinese compound word recognition as revealed by ERP data. *Language, Cognition and Neuroscience*, 39(1), 55-75.

- Wang, J., Yang, J., Biemann, C., & Li, X. (2023). Mechanism of semantic processing of lexicalized and novel compound words: An eye movement study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(11), 1812.
- Wei, Y., Niu, Y., Taft, M., & Carreiras, M. (2023). Morphological decomposition in Chinese compound word recognition: Electrophysiological evidence. *Brain and Language*, 241, 105267.
- Wen, Y., Snell, J., & Grainger, J. (2019). Parallel, cascaded, interactive processing of words during sentence reading. *Cognition*, 189, 221-226.
- White, A. L. (2023). Processing bottlenecks in visual word recognition. Poster at The 15th Annual Meeting of the Society for the Neurobiology of Language.
- White, A. L., Palmer, J., & Boynton, G. M. (2018). Evidence of serial processing in visual word recognition. *Psychological Science*, 29(7), 1062-1071.
- White, A. L., Palmer, J., & Boynton, G. M. (2020). Visual word recognition: Evidence for a serial bottleneck in lexical access. *Attention, Perception, & Psychophysics*, 82, 2000-2017.
- White, A. L., Palmer, J., Boynton, G. M., & Yeatman, J. D. (2019). Parallel spatial channels converge at a bottleneck in anterior word-selective cortex. *Proceedings of the National Academy of Sciences*, 116(20), 10087-10096.
- Wray, S., Stockall, L., & Marantz, A. (2022). Early Form-Based Morphological Decomposition in Tagalog: MEG Evidence from Reduplication, Infixation, and Circumfixation. *Neurobiology of Language*, 3(2), 235-255.
- Xiang, M. (2013). Resolving structural ambiguities in Chinese. In: Rint Sybesma, Wolfgang Behr, Zev Handel & C.T. James Huang (eds.), *Encyclopedia of Chinese Language and Linguistics*. Leiden: Brill Academic Publishers.
- Xiong, J., Yu, L., Veldre, A., Reichle, E. D., & Andrews, S. (2023). A multitask comparison of word-and character-frequency effects in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(5), 793.
- Xue, G., & Poldrack, R. A. (2007). The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. *Journal of cognitive neuroscience*, 19(10), 1643-1655.
- Yan, M., Kliegl, R., Shu, H., Pan, J., & Zhou, X. (2010). Parafoveal load of word N+ 1 modulates preprocessing effectiveness of word N+ 2 in Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6), 1669.

- Yan, M., Richter, E. M., Shu, H., & Kliegl, R. (2009). Readers of Chinese extract semantic information from parafoveal words. *Psychonomic bulletin & review*, *16*(3), 561-566.
- Yan, M., & Sommer, W. (2015). Parafoveal-on-foveal effects of emotional word semantics in reading Chinese sentences: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1237.
- Yan, M., Zhou, W., Shu, H., & Kliegl, R. (2015). Perceptual span depends on font size during the reading of Chinese sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 209.
- Yang, J., Rayner, K., Li, N., & Wang, S. (2012). Is preview benefit from word $n+2$ a common effect in reading Chinese? Evidence from eye movements. *Reading and Writing*, *25*, 1079-1091.
- Yang, J., Staub, A., Li, N., Wang, S., & Rayner, K. (2012). Plausibility effects when reading one-and two-character words in Chinese: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1801.
- Yang, J., Wang, S., Xu, Y., & Rayner, K. (2009). Do Chinese readers obtain preview benefit from word $n+2$? Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1192.
- Yao, P., Staub, A., & Li, X. (2022). Predictability eliminates neighborhood effects during Chinese sentence reading. *Psychonomic Bulletin & Review*, 1-10.
- Yeh, H. H. (2020). Chinese verb phrases: Continuum of patterns with different lexical statuses. Stanford University.
- Yen, M. H., Radach, R., Tzeng, O. J. L., & Tsai, J. L. (2012). Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and writing*, *25*, 1007-1029.
- Yin, H., Libben, G., & Derwing, B. L. (2022). How the Chinese writing system can reveal the fundamentals of hierarchical lexical structure. *Journal of Cultural Cognitive Science*, *6*(2), 199-218.
- Yu, L., Cutter, M. G., Yan, G., Bai, X., Fu, Y., Drieghe, D., & Liversedge, S. P. (2016). Word $n+2$ preview effects in three-character Chinese idioms and phrases. *Language, Cognition and Neuroscience*, *31*(9), 1130-1149.
- Yu, L., Liu, Y., & 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.

- Yu, X., Tian, X., & Lau, E. (2024). Electrophysiological responses to syntactic and “morphological” structures: evidence from Mandarin Chinese. *bioRxiv*, 2024-01.
- Yu, R., Wu, Y., & Gu, F. (2023). Parallel phonological processing of Chinese characters revealed by flankers tasks. *Frontiers in Psychology*, *14*, 1239256.
- Zang, C. (2019). New perspectives on serialism and parallelism in oculomotor control during reading: The multi-constituent unit hypothesis. *Vision*, *3*(4), 50.
- Zang, C., Fu, Y., Bai, X., Yan, G., & Liversedge, S. P. (2021). Foveal and parafoveal processing of Chinese three-character idioms in reading. *Journal of Memory and Language*, *119*, 104243.
- Zang, C., Fu, Y., Du, H., Bai, X., Yan, G., & Liversedge, S. P. (2023). 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..
- Zang, C., Liang, F., Bai, X., Yan, G., & Liversedge, S. P. (2013). Interword spacing and landing position effects during Chinese reading in children and adults. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 720.
- 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.
- Zang, C., Wang, S., Bai, X., Yan, G., & Liversedge, S. P. (2024). Parafoveal processing of Chinese four-character idioms and phrases in reading: Evidence for multi-constituent unit hypothesis. *Journal of Memory and Language*, *136*, 104508.
- Zhan, W., Guo, R., Chang, B., Chen, Y., & Chen, L. (2019). The building of the CCL corpus: Its design and implementation, *Corpus Linguistics*, 2019, Vol.6, No.1, pp.71-86
- Zhang, Q., Huang, K. J., & Li, X. (2024). Competition between parts and whole: A new approach to Chinese compound word processing. *Journal of Experimental Psychology: Human Perception and Performance*.
- Zhang, H. P., Liu, Q., Cheng, X., Zhang, H., & Yu, H. K. (2003, July). Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the second SIGHAN workshop on Chinese language processing* (pp. 63-70).
- Zhang, M., Liversedge, S. P., Bai, X., Yan, G., & Zang, C. (2019). The influence of foveal lexical processing load on parafoveal preview and saccadic targeting during Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 812.

- Zhang, B., & Peng, D. (1992). Decomposed storage in the Chinese lexicon. In *Advances in psychology* (Vol. 90, pp. 131-149). North-Holland.
- Zhao, Z., Li, Y., Hou, C., Zhao, J., Tian, R., Liu, W., ... & Yan, K. (2022). Tencent pretrain: A scalable and flexible toolkit for pre-training models of different modalities. *arXiv preprint arXiv:2212.06385*.
- Zhou, J., & Li, X. (2021). On the segmentation of Chinese incremental words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(8), 1353.
- Zhou, X., & Marslen-Wilson, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive processes*, 10(6), 545-600.
- Zhou, X., Marslen-Wilson, W., Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology reading Chinese compound words. *Language and cognitive processes*, 14(5-6), 525-565.
- Zhou, W., Wang, A., Shu, H., Kliegl, R., & Yan, M. (2018). Word segmentation by alternating colors facilitates eye guidance in Chinese reading. *Memory & cognition*, 46, 729-740.
- Wang, H. F. 王海峰. (2009). 基於大型語料庫的現代漢語離合詞定量研究. *華語文教學研究*, 6(1), 59-89.