Predictive processing during simultaneous interpreting: Insights from the visual-world paradigm, interpreting performance, and retrospective self-reports

Mingqing Xie

Predictive processing during simultaneous interpreting: Insights from the visual-world paradigm, interpreting performance, and retrospective self-reports

Mingqing Xie

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

School of Modern Languages and Cultures University of Durham

July 2025

Declaration

I hereby declare that this thesis is an original report of my research, which has been done after registration for the degree of PhD at Durham University, and materials contained in the thesis has not been submitted for a degree in this or any other institution. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Mingqing Xie

Signature:

Table of Contents

Abstract

Despite growing interests in predictive processing during simultaneous interpreting (SI), the real-time processing mechanisms supporting it have only received limited investigation, leaving a gap in our understanding of how interpreters dynamically utilise predictive processing in practice. This study systematically investigates predictive processing during SI through the triangulation of the visual-world paradigm, interpreting performance, and retrospective self-reports. A total of twenty-two professional interpreters and forty-four interpreting students were recruited to perform SI tasks involving multi-sentence paragraphs while viewing visual displays containing a target object, two semantic competitor objects, and one distractor objects. The collected data were analysed to uncover the presence, mechanism, and effects of predictive processing during SI.

First, this study examined whether professional and student interpreters predict semantic information about upcoming content in their second language during SI. Both groups demonstrated predictive eye movements toward the target objects before hearing the corresponding words in contextually constraining sentences, indicating the presence of predictive processing under demanding conditions. Second, by integrating eye-tracking, performance, and retrospective data, this study explored the mechanisms underpinning predictive processing during SI. Two distinct mechanisms were identified: prediction-by-production, characterised by top-down simulation of upcoming content, and prediction-by-association, driven by automatic semantic-thematic and lexical activation. These mechanisms appeared to operate in parallel, with interpreters switching between them depending on contextual and cognitive demands.

Thirdly, this study explored the relationship between predictive processing and interpreting performance. While no conclusive relationship was identified, high-performing interpreters, whether defined by short production latencies, high output quality, or self-reported anticipatory strategies, tended to exhibit earlier and more robust predictive fixations. The convergence across eye movements patterns, SI performance, and subjective attitudes points to a critical role of metacognitive control in shaping predictive processing during SI. Finally, this study investigated expertise-related differences in the cognitive rhythms and strategic control of predictive processing during SI. The professionals displayed more pronounced and temporally dynamic predictive fixations, more consistent SI performance across conditions, and greater metacognitive awareness, suggesting deeper engagement in top-down, production-based

prediction. In contrast, students demonstrated more static gaze patterns, prioritised input-output synchrony, and often adopted a streamlined processing strategy, reflecting reduced cognitive flexibility and heavier reliance on reactive, association-based prediction.

Together, these findings highlight predictive processing as a dynamic, individualised, and strategically modulated component of SI, shaped by both automatic cognitive mechanisms and metacognitive control. Implications are discussed for interpreter training and for future research into cognitive adaptation in high-demand language tasks.

Keywords: prediction, prediction-by-production, prediction-by-association, visual-world paradigm, EVS, interpreting quality, retrospection

Acknowledgements

The journey through a PhD programme is never easy, but I have been truly fortunate to be supported by my supervisor, senior researchers, family, friends, and colleagues throughout.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Binghan Zheng, for his inspiration, guidance, and encouragement, both during my master's and throughout the entire doctoral programme. My heartfelt thanks also go to the China Scholarship Council for funding my studies, and to St John's College and the School of Modern Languages and Cultures at Durham University for their generous support, especially during the pandemic.

I am deeply grateful to Professor Ricardo Muñoz Martín, Professor Yanping Dong, Professor Junying Liang, Professor Yanjing Wu, and Dr Yiguang Liu for their insightful comments and constructive suggestions on my research. I also thank the reviewers of my annual reviews for their valuable advice and support. Special thanks to Dr Patrick Sturt for his help with eye-tracking data analysis, and to Professor Chao Han and Professor Xia Xiang for their kind assistance during my data collection in Xiamen and Ningbo. I am also thankful to all the participants who took part in my experiments for their time and contributions.

I would also like to thank the current and former members of the TPR Lab at Durham University for their stimulating discussions, support, and encouragement. Thank you for sharing your experiences, piloting my experiments, and helping me refine my ideas. In particular, I would like to acknowledge Dr Hao Zhou, Dr Yu Weng, Ms Xueni Zhang, and Mr Haoshen He.

I am profoundly thankful for the unconditional love and unwavering support of my dearest parents and my partner, Dr Yifan Liang. I am also grateful to my wonderful friends, Miss Yijuan Jiang, Dr Ziyi Wang, Dr Hailin Yi, Dr Wei Li, Miss Lin Lin, Miss Shuyi Chen, and Miss Jing Li (Valen), for being by my side through all the highs and lows. My sincere thanks also go to my English grandparents, Mr Andrew West and Ms Jennifer Wilkinson, for making me feel at home in England. Finally, special thanks to my favourite dogs, Hana and Darcy, whose joy and comfort have sustained me more than words can express. The completion of this thesis would not have been possible without any of you.

List of Tables

List of Figures

List of Abbreviations

AoA	Age of acquisition
ANOVA	Analysis of variance
AOI	Area of interest
CC	Contextual constraint
CI	Consecutive interpreting
Cloze	Cloze probability
CPA	Cluster-based permutation analysis
CT-span	Critical verb to target word span
CV	Critical verb
df	Degree of freedom
Distra	Distractor
FluDel	Fluency of delivery
GCA	Growth-curve analysis
Elog	Empirical logit transformed
ERP	Event-related potentials
EVS	Ear-voice span
H	hypotheses
ICC	Intraclass correlation coefficient
ICC2k	ICC for average measures
Implau	Implausible competitor
InfoCom	Information completeness
L	Listening component
LME	Linear mixed effect
L1	First language
L2	Second language
M	Memory component
MSE	Mean square error
MTI	Master in translation and interpreting
NP	Noun phrases
NP1-nom	Nominative noun phrases
NP2-dat	Dative noun phrases
NP3-acc	Accusative noun phrases
P	Production component

PACS	Prediction-, association-, combinatorial-, and simulation-based
PC	Processing capacity
Plau	Plausible competitor
Pre	Predictable condition
PSE	Predictable sentence ending
RQ	Research question
SD	Standard deviation
SE	Standard error
SI	Simultaneous interpreting
SL	Source language
SOA	Stimulus onset asynchronies
SOV	Subject-object-verb
ST	Source text
SVO	Subject-verb-object
TA	Thematic analysis
Tar	Target
TL	Target language
TLQual	Target language quality
TP	Transitional probability
TW	Target word
Unpre	Unpredictable condition
VIF	Variance inflation factor
VN	Verb-noun association
VWP	Visual-world (eye-tracking) paradigm
WM	Working memory

Chapter 1 Introduction

1.1 Research background

Recent developments in neuroscience have proposed that the brain is fundamentally predictive in nature. According to Anokhin (1978), prediction is a core organising principle of brain activity, and later theories have expanded this idea into the concept of the "predictive brain" (Clark, 2013). Within this framework, the brain constantly generates forward models of sensory input, aiming to reduce uncertainty by pre-activating likely upcoming events based on prior experience and contextual cues. Prediction is therefore not a specialised or peripheral function, but a core computational feature of cognition, involved in perception, action, and language.

In the field of psycholinguistics, prediction has been defined more specifically as the pre-activation of linguistic features before they are encountered in the linguistic input (Pickering & Gambi, 1998). A wide range of empirical studies using behavioural and neurophysiological methods has demonstrated that language comprehenders can predict upcoming words at multiple levels, including semantic, syntactic, and phonological (Altmann & Kamide, 1999; Kutas & Hillyard, 1980; Ito, Pickering, & Corley 2018; van Berkum et al., 2005). This pre-activation can facilitate faster integration and smoother comprehension. For example, anticipatory eye movements in the visual-world paradigm (VWP) and reduced N400 amplitudes in event-related potentials (ERP) studies suggest that comprehenders can use contextual information to prepare for likely upcoming linguistic items even before they are fully spoken or read. This predictive ability enhances comprehension efficiency and contributes to the fluency of real-time language processing.

In the domain of simultaneous interpreting (SI), prediction also plays a critical role. However, within this context, it is often referred to as anticipation, which broadly captures an interpreter's ability to pre-emptively process or produce segments of speech before they become fully available in the source language (Wilss, 1978; Gile, 1992; Chernov, 1994). Anticipation in SI serves several key purposes: it helps reduce the cognitive load associated with working memory, thereby freeing up resources that interpreters can devote to production (de Groot, 2011) and self-monitoring (Chernov, 1994), while maintaining a manageable time lag between input and output (Gile, 2009). Beyond the potential benefits of successful anticipation, research has found that erroneous predictions do not incur additional cognitive costs compared to instances where no prediction is made during language comprehension (Frisson, Harvey, & Staub, 2017). Moreover, prediction has been shown to facilitate the processing of words semantically related to the predicted term (Staub, Grant, Astheimer, & Cohen, 2015). Over time, frequent incorrect

predictions may also lead individuals to learn and adapt from their errors, ultimately reducing the likelihood of mistakes (Dell & Chang, 2013). In this study, these two processes are collectively referred to as predictive processing, encompassing both comprehension-based prediction and production-based anticipation during SI (Hodzik & Williams, 2017).

Despite growing recognition of the importance of predictive processing in SI, the underlying mechanisms by which predictive processing operates remain unclear. Early studies that examined prediction in SI mainly focused on anticipation where interpreters produce translated equivalents before the corresponding source utterances are delivered (Jörg, 1995; Seeber, 2001; Wilss, 1978), and measured prediction effects through offline metrices such as production latency (Chmiel, 2021; Hodzik & Williams, 2017). While such measures have provided valuable insights, they may be susceptible to extraneous variables including environmental factors (e.g., room temperature) and may fail to isolate prediction-specific processes from other confounding cognitive processes, such as integration (Pickering & Gambi, 2018). More recently, online prediction has been examined using VWP (Amos, Seeber, & Pickering, 2022, 2023; Liu, Hintz, Liang, & Huettig, 2022). These studies demonstrated successful semantic prediction during SI, as evidenced by anticipatory eye movements towards objects before they were named. Nonetheless, they employed isolated sentences as stimuli, which do not fully replicate the realistic SI task, where source texts typically consist of multiple, contextually connected sentences. As Huettig and Mani (2016) argue, studying prediction in ecologically valid, naturalistic contexts is essential for advancing our understanding of how it functions in everyday language use.

Addressing these gaps, the present study systematically investigates predictive processing during simultaneous interpreting of coherent multi-sentence discourse. It adopts a refined framework that distinguishes between prediction-by-association, a bottom-up mechanism driven by lexical co-occurrence or semantic proximity, and prediction-by-production, a more effortful top-down simulation involving the interpreter's production system (Pickering & Gambi, 2018; Amos & Pickering, 2020). By integrating multiple analytic approaches, including anticipatory eye movements, latency measures, interpreting quality, and self-reported interpreting strategies, this study aims to offer a comprehensive account of how and when predictive processing operates during SI. Furthermore, it explores whether these mechanisms are used strategically, depending on task demands, interpreter expertise, and individual cognitive styles, thereby bridging cognitive theory and interpreter practice.

1.2 Methodology

The present study employed a mixed-methods design combining eye-tracking, behavioural performance measures, and retrospective self-reports to investigate predictive processing during SI of coherent discourse. The core objective was to examine whether interpreters anticipate upcoming content using prediction-by-association and/or prediction-by-production mechanisms, and how such predictive behaviours relate to interpreting expertise and performance.

Participants included professional interpreters and interpreting students with a minimum of one-year formal interpreting training. All participants completed an SI task in which they interpreted coherent English paragraphs into their native language (Mandarin Chinese). The source texts (STs) were semi-formal in register and literary in style, selected and edited to elicit semantic prediction through verb-mediated structures. Each paragraph contained both predictable and unpredictable sentences containing critical verbs (CVs), manipulated by varying cloze probability of target words (TWs). The experiment adopted a visual-world eye-tracking paradigm. During SI, participants viewed visual displays with four line-drawing images: one target object, one plausible competitor, one implausible competitor, and one unrelated distractor. Eye movements were recorded to identify anticipatory fixations prior to the TW onset. The experiment also measured ear-voice span (EVS) at the lexical (CV and TW) and sentence level (onset and offset), as well as duration measures, to assess production timing. To complement the online data, interpreting quality was assessed through both rubric-referenced scoring and item-based accuracy coded at both the word and sentence level. After each SI task, participants were asked to reflect on their SI process, engagement with visual inputs, and anticipatory strategies during SI.

Eye-tracking data were analysed using linear mixed effect modelling to examine prediction effects in the pre-target and post-target temporal windows, and growth curve analysis and cluster-based permutation analysis to examine more fine-grained temporal dynamics and prediction effects. Student t-tests, ANOVA, linear mixed effect modelling, and correlation analyses were used to relate predictive measures to interpreting accuracy and EVS control.

This multi-level approach allowed the study to triangulate findings across objective gaze behaviour, production timing, output quality, and subjective strategy use, offering a comprehensive view of predictive processing during SI in an ecologically valid context.

1.3 Aims, research questions, and hypotheses

This study investigates predictive processing during SI using an integrated approach that combines eye-tracking, behavioural performance data, and retrospective self-reports. In doing so, it addresses limitations in ecological validity and methodological scope in previous research. The investigation is structured around the following three core aims:

Aim 1: To examine the presence of predictive processing during SI of coherent discourse and identify its underlying mechanisms, with a particular focus on distinguishing between prediction-by-production and prediction-by-association.

Aim 2: To explore the relationship between predictive processing and SI performance, specifically investigating how prediction relates to the coordination of comprehension and production, as reflected in EVS and interpreting quality.

Aim 3: To compare professional and student interpreters in terms of their engagement in predictive processing, including behavioural indicators and self-reported strategies, to assess how expertise shapes the strategic and cognitive use of prediction.

Building on these aims, four research questions (RQs) and hypotheses (Hs) are proposed. To determine whether predictive processing is engaged during SI when interpreters work with coherent, multi-sentence discourse, the first research question asks:

RQ1: Do interpreters engage in predictive processing during SI of coherent discourse?

H1 (Main prediction): Professional interpreters exhibit significantly stronger evidence of predictive processing compared to interpreting students.

This question assesses the extent to which prediction occurs during realistic SI tasks and whether it is a universal feature of interpreting or one that emerges with training and expertise. The second research question focuses on the cognitive mechanisms underlying prediction. It asks whether prediction during SI operates through top-down processes such as simulation (prediction-by-production), bottom-up activation (prediction-by-association), or both:

RQ2: What are the underlying mechanisms supporting predictive processing during SI—prediction-by-production and/or prediction-by-association?

H2 (Parallel mechanisms): Both mechanisms operate in parallel during SI.

Exploratory Analysis**:** Should the data not fully support the parallel mechanisms hypothesis, an exploratory analysis will be conducted to examine whether one mechanism is predominantly engaged.

This question is central to understanding how predictive processing functions in the cognitively demanding context of SI and which mechanisms interpreters rely on under time pressure.

Given the potential role of prediction in supporting interpreting performance, by reducing memory demands and improving synchrony, the third research question explores this relationship:

RQ3: What is the relationship between predictive processing and SI performance?

H3 (Positive association): Stronger evidence of predictive processing will be positively correlated with better SI performance (e.g., higher accuracy, more adaptive EVS).

This question investigates whether interpreters who engage more in predictive processing deliver higher-quality renditions or better manage cognitive timing during SI.

Finally, the fourth research question investigates whether professional interpreters differ from students in how they engage prediction. This includes differences in the timing, mechanisms, and strategic use of prediction:

RQ4: What are the expertise-related differences in the cognitive rhythms and strategic engagement of predictive processing during SI?

H4 (Expertise Modulation): Professionals earlier and more refined predictive processing than students, indicating that expertise enhances metacognitive control during SI.

By comparing these groups, the study aims to determine whether prediction is an innate skill or one honed through experience and training.

1.4 Structure of the thesis

The remainder of this thesis is organised into nine chapters, each building upon the last to develop a comprehensive investigation of predictive processing in SI.

Chapter 2 presents a critical review of prediction in psycholinguistic research. Section 2.1 examines prediction in first language (L1) comprehension, Section 2.2 reviews findings from second language (L2) comprehension, and Section 2.3 discusses various linguistic and cognitive factors that modulate prediction during language comprehension.

Chapter 3 reviews anticipation in SI studies. Section 3.1 outlines key theoretical perspectives regarding the nature, presence, and potential effects of anticipation on SI performance. Section 3.2 discusses early empirical and corpus-based studies on anticipation, while Section 3.3 covers recent developments in the use of VWP to investigate the online dynamics of prediction during SI.

Chapter 4 introduces the theoretical framework underpinning the present study. Section 4.1 reviews neurophysiological and computational models of predictive language processing. Section 4.2 compares the prediction-by-production and prediction-by-association accounts in psycholinguistics, drawing on relevant empirical evidence. Section 4.3 discusses the application of the prediction-by-production framework to SI. Section 4.4 concludes with an overview of the current study's design.

Chapter 5 outlines the study's methodology. Section 5.1 details the manipulation of contextual predictability. Section 5.2 describes the visual-world eye-tracking paradigm and its associated data analysis techniques. Section 5.3 explains how SI performance was evaluated using EVS, a rubric-referenced scoring system, and item-based accuracy measures. Section 5.4 describes the thematic analysis of participants' retrospective self-reports.

Chapter 6 provides details on the experimental materials and procedures. Section 6.1 describes the preparation of auditory and visual stimuli, while Section 6.2 explains the technical setup and procedure of the experiment.

Chapter 7 presents the analysis and results. Section 7.1 reports findings from the eye-tracking data, followed by EVS and interpreting performance results in Sections 7.2 and 7.3, respectively. Section 7.4 examines the relationships between EVS and interpreting performance data. Section 7.5 presents findings from the retrospective self-reports. Sections 7.6 and 7.7 examine the relationship between eye-tracking data, performance metrics and retrospective data. Each subsection includes interim discussion to contextualise the findings.

Chapter 8 offers a general discussion that integrates and interprets findings across data types. Section 8.1 discusses the presence of predictive processing in SI with coherent discourse. Section 8.2 evaluates the mechanisms underlying prediction. Section 8.3 examines the relationship between predictive processing and SI performance. Section 8.4 explores expertise-related differences in predictive processing, including cognitive timing and strategic engagement.

Chapter 9 concludes the thesis. Section 9.1 summarises the key findings; Section 9.2 outlines implications for interpreting practice and pedagogy; Section 9.3 discusses the study's contributions and limitations; and Section 9.4 suggests avenues for future research.

Chapter 2 Prediction in language comprehension

2.1 Overview

Early evidence for prediction during language comprehension emerged from studies exploring the incrementality and contextual sensitivity of linguistic processing. Specifically, comprehenders demonstrate high efficiency in interpreting linguistic input as they encounter it incrementally. In a speech shadowing experiment, Marslen-Wilson (1973) found that listeners could repeat spoken sentences with a delay as short as 250 ms, suggesting that comprehension occurred almost simultaneously with perception. Similarly, in Swinney (1979), participants listened to sentences containing ambiguous words while simultaneously performing lexical decision tasks. The results demonstrated a facilitative effect of biasing context in selecting the appropriate meaning of lexically ambiguous words within a few hundred milliseconds. Likewise, Schwanenflugel and Shoben (1985) found faster lexical decisions for highly predictable words than for less predictable ones. Contextual sensitivity was also observed in reading studies. In an eye-tracking experiment, Ehrlich and Rayner (1981) observed that target words were read more quickly in highly predictable context. Frazier and Rayner (1982) reported fewer regressions when sentence structure aligned with predicted syntactic patterns, suggesting a role for syntactic prediction in guiding parsing decision. Trueswell, Tanenhaus, and Garnsey (1994) also found that readers used thematic role information to disambiguate syntactic structures, indicating predictive use of semantic cues in syntactic interpretation.

In a different line of research, Kutas and Hillyard (1980) used ERPs to examine brain responses to semantically congruent and incongruent words in written sentences. They found that a distinct negative ERP component occurring around 400 ms after the stimuli onset, later termed the N400, was elicited when sentences ended with an incongruent word. This suggests the brain's sensitivity to semantic violations and implies the involvement of semantic prediction during comprehension. Building on this study, Kutas and Hillyard (1984) explored the influence of word expectancy and semantic associations on sentence processing. They found an inverse relationship between N400 and the predictability of words within the context: specifically, words of higher predictability elicited smaller N400 amplitudes. These findings suggest that comprehenders may not only reactively process linguistic input as they encounter it, but also actively predict upcoming input.

However, although these studies support a prediction account of language comprehension, they do not directly demonstrate prediction. Instead, their findings are also compatible with an integration account. Integration refers to the process whereby, after processing a linguistic

input, comprehenders combine the activated linguistic representations with a mental model established by the preceding context (Pickering & Gambi, 2018). The primary difference between prediction and integration lies in the timing: prediction occurs before the linguistic input is encountered, whereas integration occurs afterwards, as bottom-up, reactive process. For instance, under a predictive interpretation, the shorter fixations associated with predictable words (Ehrlich & Rayner, 1981) are attributed to the facilitation provided by pre-activated linguistic representations. Under an integrative interpretation, however, those representations are not activated until the predictable word is encountered. Because predictable words are easier to be combined into the existing context, their processing is thus facilitated, leading to shorter fixation durations.

The integration account is also compatible with findings from ERP studies, in which reduced N400 amplitudes reflect lower integration difficulty. For example, in Bentin, McCarthy, and Wood (1985), participants were presented with word pairs in a lexical decision task and had to decide whether the second item was a real word or nonword. ERP recordings revealed reduced N400 amplitudes for semantically related word pairs (e.g., *nurse* following *doctor*). Using a similar paradigm, Praamstra and Stegeman (1993) found a reduction in N400 amplitudes when the second word was phonologically related to the first (e.g., *thing* following *king*). Under an integration account, these reduced N400 effects may result from associative priming, where semantically or phonologically related words are easier to process because they share features in memory. This shared activation facilitates the integration of the second word into the ongoing representation once it is encountered, without requiring prior prediction (Dosher & Rosedale, 1989; Meyer & Schvaneveldt, 1971; Ratcliff & McKoon, 1988).

Given difficulty to distinguish prediction from integration, the presence of prediction during language comprehension should be demonstrated more directly, by showing activation of relevant linguistic representations prior to the occurrence of the linguistic input. Alternatively, as Pickering and Gambi (2018) argued, there is a time lag between encountering the linguistic input and the responsive activation of relevant representations. If such activation occurs within this time lag, it cannot be the result of bottom-up processing alone; rather, it implies that pre-activation must have occurred. The studies reviewed below are discussed based on this criterion.

2.2 Prediction in native language comprehension

2.2.1 Prediction at the semantic level

Altmann and Kamide (1999) provide the earliest direct evidence for semantic prediction in language processing. They employed VWP to investigate whether verb information could provide cues for anticipating subsequent referents. Participants viewed a visual scene featuring one agent and four objects while listening to sentences containing mono-transitive verbs that were either semantically selective (e.g., *eat*) or semantically neutral (e.g., *move*). For example, the visual scene accompanying the sentence "*the boy will eat/move the cake*" depicted a boy sitting among a balloon, a cake, a toy car, and a toy train. They found that upon hearing the verb *eat*, participants began to fixate on the cake, the only edible item in the visual scene, before the object noun cake was explicitly mentioned. Such predictive eye movements were not observed when the verb was *move*, which matched all four items. Although this result did not illustrate prediction of the exact utterance, it provides strong evidence that participants used verb semantics information extracted from the verb (e.g., *eat*) to predict the semantic features (e.g., *edibility*) of upcoming referents. In a follow-up study, Kamide, Altmann, and Haywood (2003) duplicated this design but substituted the mono-transitive verbs with three-place verbs (e.g., "*spread the butter on the bread*"). Their results showed that the predictive inferences based on the verb's semantic information could extend beyond immediate direct objects to later-occurring objects in the sentences, further supporting the role of verb semantics in generating predictive representations.

Evidence for pre-activation of semantic features has also come from ERP studies. Federmeier and Kutas (1999) examined how contextual constraints and long-term memory influence sentence processing. They created sentential contexts that predicted a specific target word (e.g., "*They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of …*"), followed by one of the three critical words: 1) an expected word (e.g., *palms*), 2) an unexpected within-category word (e.g., *pines*), or 3) an unexpected between-category word (e.g., *tulips*). The results revealed that expected words elicited a late positive component in the N400 time window, while both types of unexpected words elicited typical N400 effects, suggesting an effect of prediction. Crucially, N400 amplitudes were smaller for within-category violations than for between-category ones, indicating an independent effect of semantic category. Similarly, Metusalem et al. (2012) investigated how real-world event knowledge shapes semantic processing during sentence comprehension. Participants read short

discourses ending with either an expected word (both syntactically and semantically appropriate with high predictability), an event-related word (semantically associated but not highly predictable), or an unrelated word (syntactically plausible but semantically irrelevant). They found a graded N400 response: smallest for expected words, larger for event-related words, and largest for unrelated words. Comparable graded N400 effects were also reported in other ERP studies manipulating contextual constraint, such as Boudewyn, Long, and Swaab (2015) and Paczynski and Kuperberg (2012), supporting predictive semantic processing driven by varying degrees of contextual constraint.

However, these graded N400 effects could also be explained by associative priming, as the preceding context in these studies often contained potential prime words (e.g., *tropical*) that are semantically associated with expected targets (e.g., *palms*). Although the lexical associative priming effects are typically short-lived and tend to dissipate within a single intervening word (Camblin et al., 2007; Masson, 1991; McKoon & Ratcliff, 1989), it remains possible that such primes partially facilitate sentence processing, leading to smaller N400 amplitudes. To more definitively isolate prediction from priming, Otten and van Berkum (2008) designed a study that compared participants' ERP responses to anomalous words presented in predictive contexts versus in priming control contexts. Both contexts included the same potential prime words, ensuring that any priming effects would be constant. The results showed that the anomalous nouns in the predictive context elicited a larger positivity than those in the priming control context. Since the anomalous nouns were equally implausible and difficult to integrate across conditions, the difference could not be attributed to integration difficulty. The most plausible explanation is that participants made predictions about upcoming words in the predictive context but failed to do so in the priming control context. The resulting ERP effect thus reflect a mismatch between predictions and actual utterances, rather than mere semantic association.

2.2.2 Prediction at the syntactic level

Prediction has also been observed at the syntactic level. Kamide et al. (2003) used VWP to investigate whether syntactic case marking in Japanese could trigger predictive processing in verb-final sentences. In three-noun phrase (3-NP) verb-final constructions, three types of case markers, i.e., nominative, dative, and accusative, are assigned to the verb's arguments to indicate the agent (NP1-nom), the goal (NP2-dat), and the theme (NP3-acc). In contrast, two-noun phrase (2-NP) verb-final sentences contain only NP1-nom and NP2-acc. Kamide and

colleagues designed two conditions: a 3-NP sentence in the dative condition and a 2-NP sentence in the accusative condition. The two conditions shared the same first two NPs. For example,

1a) Dative condition: ウェイトレスが客に楽し気にハンバーガーを運ぶ。
waitress-nom customer-dat merrily hamburger-acc bring.
(The waitress will merrily bring the hamburger to the customer.)

1b) Accusative condition: ウェイトレスが客を楽し気にからかう。
waitress-nom customer-acc merrily tease.
(The waitress will merrily tease the customer.)

The accompanying visual scene depicted a waitress, a customer, a hamburger, and a bin. They found that, up to the adverb (*merrily*), participants in the dative condition were significantly more likely to make anticipatory eye movements to the hamburger, the only plausible theme, than in the accusative condition. This indicates that participants integrated case markers incrementally and used them to predict the thematic structure of the upcoming verb phrase: specifically, anticipating *hamburger* (not *customer*) as the theme in the dative condition, and that *customer* as the only theme in the accusative condition.

Wicha, Moreno, and Kutas (2004) used ERPs to examine gender expectancy during sentence comprehension in Spanish. They manipulated both semantic congruity between a critical noun and its preceding context, and gender agreement between the critical noun and its preceding article. ERP responses were measure at the article and the critical noun. They found an N400 effect elicited by semantic incongruity, which was amplified when the gender agreement was also violated, suggesting a contribution of grammatical gender to predictive processing. Using a similar approach, van Berkum (2005) manipulated the syntactic gender of an adjective preceding a highly predictable noun. When the adjective mismatched the syntactic gender of the expected noun, a larger positive ERP amplitude was observed than when gender agreement was intact. Importantly, this effect disappeared in the absence of a constraining context, confirming that the response depended on predictive processing. Further support for grammatical gender-based prediction comes from ERP findings by Otten, Nieuwland, and van Berkum (2007), who observed a negative deflection between 300 ms and 600 ms following adjectives with unexpected gender in a predictive context. Similarly, Otten and van Berkum (2008) found a short-lived late negativity when the suffix of an adjective mismatched the syntactic gender of a highly predictable noun in a supportive discourse.

Collectively, these findings demonstrate that comprehenders can exploit syntactic features, such as case marking and grammatical gender, to form anticipatory representations during real-time sentence comprehension.

2.2.3 Prediction of phonological word forms

Pre-activation of semantic or syntactic information does not necessarily imply prediction of a specific word. However, when a specific word is predicted, other lexical representations, such as phonological word forms, should also be pre-activated. This hypothesis is supported by findings from an ERP study by DeLong, Urbach, and Kutas (2005). They asked participants to read sentences varying in contextual constraint, which ended with either an expected noun beginning with a consonant sound or a less likely noun beginning with a vowel sound. Participants' ERP responses were measured for both target nouns and their preceding indefinite articles (i.e., *a* or *an*). The study replicated the N400 effect for both nouns and articles. Moreover, N400 amplitudes demonstrated a negative correlation with the predictability of both target nouns and preceding articles. Since both types of articles (*a* and *an*) were syntactically and semantically congruous within the context, the N400 effect observed on the articles could not be attributed to integration difficulty. Instead, these results suggested that in highly constraining contexts, comprehenders generated specific lexical predictions and pre-activated the phonological form of the expected noun accordingly. Nonetheless, while some subsequent studies successfully replicated the article-elicited N400 effect (e.g., Martin et al., 2013), others failed to do so (Ito, Martin, & Nieuwland, 2017a; Nieuwland et al., 2018).

Using a different ERP paradigm, Ito et al. (2016) provided further evidence for phonological prediction. Participants were presented with high-constraint sentences (e.g., "*The student is going to the library to borrow a …*") that finished with one of the four word types: a predictable word (e.g., *book*), a phonologically distant semantic neighbour (e.g., *page*), a semantically unrelated phonological neighbour (e.g., *hook*), or a totally unrelated word (e.g., *sofa*). Sentences were presented word by word, with stimulus onset asynchronies (SOAs) of either 500 ms or 700 ms. At the shorter SOA (500ms), only semantic neighbours elicited reduced N400s relative to unrelated words. However, at longer SOA (700ms), both semantic and phonological neighbours elicited reduced N400s relative to unrelated words. These results suggest that phonological prediction is time-sensitive and may occur only when comprehenders have sufficient processing time. This supports the view that phonological prediction may not be a robust or automatic phenomenon in language comprehension (Ito et al., 2017a).

In a follow-up study (Ito, Pickering, et al., 2018), they employed VWP and had native English speakers listened to highly constraining sentences (e.g., "*The tourists expected rain when the sun wen behind the…*") while viewing visual displays that included the highly predictable word (e.g., *cloud*), an English phonological competitor (e.g., *clown*), a Japanese phonological competitor, or an unrelated object. In addition to significantly higher fixation proportions on the predictable object prior to its onset, English native speakers were also more likely to predictively fixate on the English phonological competitor, indicating anticipatory activation of word form. This effect was not observed for the Japanese phonological competitor, suggesting that phonological prediction during compression may be constrained by the comprehenders' native phonological system or familiarity with the phonological forms in question.

2.2.4 Summary of prediction in L1

To summarise, in L1 comprehending, prediction can occur at the semantic, syntactic and phonological levels. Prediction influences language comprehension independently of priming, although priming may serve to facilitate predictive processing (Lau, Holcomb, & Kuperberg, 2013; McRae, Hare, Elman, & Ferret, 2005). Prediction may also interact with integration. When certain linguistic features of an upcoming word are predicted, integration of that word into the context becomes easier. Moreover, both probability and precision of prediction correlate positively with the level of contextual constraint: the more constraining the context, the more likely and more specific the prediction. As demonstrated in studies of syntactic (van Berkum et al., 2005; Wicha et al., 2004) and phonological prediction (DeLong et al., 2005; Ito et al., 2016; Ito, Pickering, et al., 2018), highly constraining contexts can lead to the pre-activation of a specific lexical item, including its syntactic gender and/or phonological form—that is, a highly precise prediction. In contrast, under less constraining context or when processing time is limited, phonological prediction appears to be the first to diminish or disappear.

All the studies reviewed thus far have focused on L1 comprehension. However, in SI practice, interpreters often interpret from their L2 to L1, i.e., comprehending in an L2 and producing in their L1. The following section reviews studies examining prediction during bilinguals' L2 comprehension.

2.3 Prediction in second language comprehension

2.3.1 Prediction at the semantic level

As previously discussed, Ito, Pickering, et al. (2018) used VWP to investigate the time course of phonological prediction in both L1 and L2 English speakers (L1 Japanese). Their findings showed that both groups made predictive eye movements to target items well before word onset, although L2 speakers were slower than native speakers. Using a similar approach, Ito, Corley, and Pickering (2018) examined to what extent cognitive load influences predictive processing. Participants listened to sentences with verbs that either did or did not semantically constrain the identity of one of four visual items (i.e., the target item). In the load condition, participants performed an additional memory task. Both L1 and L2 speakers made more predictive fixations on the target objects in the predictable than the unpredictable sentences. However, this difference was only statistically significant in the no-load condition, suggesting that predictive processing in L2 may be more sensitive to cognitive resource limitations.

Hopp (2015) used VWP to test whether L1 and L2 German speakers (L1 English) integrate syntactic cues (i.e., case marking) and semantic cues (i.e., verb meaning) to guide prediction. In German, flexible word order is disambiguated by case-marked determiners: *der* (nominative) marks the subject, and *den* (accusative) marks the object. The results illustrated that L1 German speakers were able to use case marking to determine the grammatical role of the first noun phrase (subject or object) and made anticipatory fixations accordingly (i.e., to agent or patient referents). This pattern suggested they used both semantic and syntactic cues to make predictions. By contrast, L2 German learners did not modulate their fixations based on case marking. Instead, they consistently made predictive eye movements to the potential patient referent of the first noun, irrespective of the prenominal determiner. This suggests they predicted that the first noun was the agent, aligning with an L1 English subject-first processing bias (Hopp, 2006). This suggests that L2 comprehenders relied more on semantic than syntactic information in prediction.

Grüter, Lau, and Ling (2018) examined whether L2 Mandarin speakers utilise the semantic content of classifiers to generate predictions. In Mandarin, classifiers appear prior to nouns and encode both syntactic and semantic constraints (Allan, 1977; Li & Thompson, 1989). Previous research has shown that classifiers can trigger semantic prediction in L1 Mandarin speakers (Kwon, Sturt, & Liu, 2017). In this study, when participants heard a classifier that semantically matched both the target and a competitor, but syntactically matched only the target, L2 speakers

fixated more on the competitor until the noun was heard. This indicates that L2 Mandarin speakers relied more on semantic than syntactic information in generating predictions, and that syntactic constraints from classifiers were underutilised.

Using a modified version of the DeLong et al. (2005) design, Martin et al. (2013) conducted an ERP study involving both L1 and L2 English speakers (L1 Spanish). Participants read sentences ending with either a predictable or less predictable noun, each beginning with either a vowel or a consonant sound. Both groups exhibited N400 effects for less predictable nouns, indicating that semantic prediction occurred regardless of language background. These results also suggest that L1 and L2 English speakers may engage the same underlying mechanism for semantic prediction (Ito, Martin, & Nieuwland, 2017b; Kaan, 2014), even though processing efficiency may differ.

2.3.2 Prediction at the syntactic level

The studies reviewed above (Grüter et al., 2018; Hopp, 2015; Ito, Corley, et al., 2018; Ito, Pickering, et al., 2018) suggest that prediction in L2 comprehension may be limited to the semantic level, with little to no evidence for syntactic prediction. Further support for this conclusion comes from several other studies. For example, Mitsugi and MacWinney (2016) replicated the design of Kamide et al. (2003) in a study involving L1 and L2 Japanese speakers (L1 English). They found L2 learners of Japanese consistently made predictive eye movements towards representations of potential themes in both mono-transitive and ditransitive sentence structures. This pattern indicated that L2 learners did not integrate case-marking information, relying instead on semantic cues to guid prediction. In a follow-up study, Mitsugi (2017) further investigated whether L2 Japanese learners could predict the grammatical voice of the final verb using case markers. Again, L2 learners showed minimal use of case-marking cues, suggesting a limited ability to make syntactic predictions in L2. These findings are in line with Kaan, Kirkham, and Wijnen (2014), who used ERPs to test L2 English speakers' ability to distinguish elliptical from non-elliptical structures. Unlike native speakers, L2 participants failed to use syntactic cues predictively, highlighting further limitations in syntactic prediction during L3 comprehension.

However, there is evidence that advanced or near-native L2 speakers may be capable of syntactic prediction, particularly under favourable conditions. Hopp and Lemmerth (2018) used VWP to examine whether L2 German speakers (L1 Russian) could make gender-based predictions, despite the differing syntactic gender systems of the two languages. Results

showed that advanced L2 speakers made native-like use of gender agreement, largely unaffected by cross-linguistic incongruence. In contrast, intermediate L2 speakers did not exhibit predictive gender processing, suggesting that syntactic prediction in L2 may depend on proficiency level. Similarly, Foucart, Martin, Moreno, and Costa, (2014) conducted an ERP study involving L2 Spanish speakers (L1 French). Participants read sentences containing either a predictable or less predictable noun, with prenominal articles differing in syntactic gender. Since gender assignment is consistent between Spanish and French, the study aimed to test whether L2 speakers could anticipate the gender of upcoming nouns. N400 effects were observed on both articles and critical nouns, indicating that L2 speakers engaged in syntactic prediction under conditions of cross-linguistic similarity. Grüter, Lew-Williams, and Fernald (2012) also found that highly proficient L2 Spanish speakers used gender-marked determiners to anticipate upcoming nouns, much like native speakers. However, their predictive use of syntactic cues was less consistent than that observed in native speakers.

2.3.3 Prediction of phonological word forms

To date, there is no clear evidence that L2 speakers routinely engage in phonological prediction during language comprehension. As discussed earlier, in the VWP study by Ito, Pickering, et al. (2018), phonological prediction was observed in L1 speakers but not in L2 speakers (L1 Japanese). Similarly, in the ERP study by Martin et al. (2013), L2 speakers exhibited N400 effects at the semantic level, indicating semantic prediction, but no effects were observed at the phonological level. Furthermore, Ito et al. (2017b) replicated the ERP study by Ito et al. (2016), this time with L2 English speakers (L1 Spanish), and again failed to find evidence of phonological prediction.

However, a recent visual-world eye-tracking study by Lozano-Argüelles, Sagarra, and Casillas (2020) offers some evidence for phonological pre-activation among L2 speakers. They investigated whether L2 Spanish speakers can anticipate upcoming morphological information based on phonological cues, specifically suprasegmental (stress pattern) and segmental (syllabic structure) features. Participants listened to Spanish sentences in which the stress pattern (oxytone vs paroxytone) and syllabic structure of the preceding input provided cues about the likely morphological form of the upcoming word. Participants viewed two written options, and their anticipatory fixations were measured before the target word onset. The study included Spanish monolinguals, advanced L2 Spanish learners without interpreting experience, and advanced L2 learners with interpreter training. Results showed that while all groups used

phonological cues to guide anticipatory fixations, L2 speakers made slower and less consistent anticipatory fixations than native speakers, particularly in conditions where phonological competition was high. Notably, L2 learners with interpreter training outperformed their non-interpreting peers, demonstrating faster and more accurate prediction. These findings indicate that phonological prediction in L2 is possible, particularly at an abstract level (e.g., stress and syllable structure), but may require high proficiency or domain-specific training to emerge robustly.

2.3.4 Summary of prediction in L2

In summary, L2 speakers are more likely to engage in prediction at the semantic level, while syntactic and phonological prediction appear to be less frequent or absent. The likelihood of predictive processing in L2 seems to be closely tied to proficiency level. Highly proficient L2 speakers are more likely to predict at both semantic and, to some extent, syntactic levels, although such predictions tend to be slower or less consistent compared to those of native speakers. Intermediate L2 speakers generally show evidence of semantic prediction, but there is no current evidence supporting their ability to predict based on syntactic cues. This difficulty may reflect a general underuse of syntactic information in L2 comprehension compared to L1 processing (Marinis, Roberts, Felser, & Clahsen, 2005; Clahsen & Felser, 2006).

Several factors may explain the inconsistent findings regarding syntactic prediction in the reviewed L2 studies (Foucart et al., 2014; Grüter et al., 2012; Grüter et al., 2018; Hopp, 2015; Hopp & Lemmerth, 2018; Ito, Pickering, et al., 2018; Mitsugi, 2017; Mitsugi & MacWhinney, 2016). First, the varying difficulty levels of the experimental tasks could have affected results. As argued by Ito, Corley, et al. (2018), prediction requires cognitive resources during sentence processing. More demanding comprehension tasks may consume more of these resources, leaving insufficient capacity for predictive processing. Second, the degree of syntactic overlap between the L1 and L2 may influence syntactic prediction by L2 speakers. Some studies involved syntactic features shared across languages (e.g., Hopp & Lemmerth, 2018; Foucart et al., 2014), while others tested structures that differed between L1 and L2 (e.g., Mitsugi & MacWhinney, 2016). Syntactic similarity between L1 and L2 may facilitate predictive processing in L2 (Amos & Pickering, 2020).

Third, L1 regulation and production abilities may mediate L2 prediction. According to Zirnstein, van Hell, and Kroll (2018), the ability to manage L1 activation and engage in L2 production affects predictive processing. This aligns with findings by Martin, Branzi, and Bar

(2018), who showed that taxing the speech production system impairs prediction during sentence comprehension. Given that participants across studies vary in their L1 production capabilities, this factor may contribute to the mixed results observed in L2 prediction research.

2.4 Factors modulating prediction during language comprehension

Previous studies have shown that prediction can be modulated by several factors, such as working memory (WM) capacity (Huettig & Janse, 2016; Ito, Corley, et al., 2018; Lozano-Argüelles, Sagarra, & Casillas, 2023), production ability (Mani & Huettig, 2012; Zirnstein et al., 2018), age (Federmeier & Kutas, 2005; Huang, Meyer, & Federmeier, 2012; Wlotko & Federmeier, 2012), language proficiency (Hopp & Lemmerth, 2018; Lozano-Argüelles & Sagarra, 2021), and literacy (Huettig & Brouwer, 2015; Huettig & Pickering, 2019). The current section discusses mediating factors that are most relevant to the current study, including working memory, language proficiency, and domain-specific experience, i.e., interpreting experience.

2.4.1 Working memory

Working memory is a cognitive system that temporarily stores and manipulates information for ongoing processing. As a control mechanism, WM underpins one's ability to engage in complex and coherent thought, such as reasoning and decision-making (Baddeley, 2000, 2007). Previous research has demonstrated that WM capacity can modulate predictive processing during language comprehension. Huettig and Janse (2016) investigated individual differences in WM and processing speed in relation to anticipatory eye movements during spoken sentence comprehension. A group of Dutch speakers aged between 32 and 77 were recruited to complete three WM tasks, including a nonword repetition task, a backward digit span task, and a Corsi block-tapping task, to assess spatial WM capacity. In a VWP experiment, participants listened to sentences containing gender-marked articles while viewing four objects, only one of which matched the article's grammatical gender. Results revealed a positive correlation between WM capacity and predictive eye movements to the target object, suggesting that individuals with higher WM capacity were better able to make predictions during comprehension. Using the same paradigm, Ito, Corley, et al. (2018) introduced an additional WM load by requiring participants to remember words while listening to the target sentences. They found that predictive eye movements were significantly delayed under load. This suggests that prediction

draws on WM resources, and that occupying WM with an unrelated memory task reduces the availability of those resources for predictive processing.

Evidence from ERP studies further supports the role of WM in prediction. Federmeier and Kutas (2005) compared younger adults (Mean age = 20) and older adults (Mean age = 67) in their N400 responses to target nouns in either strongly or weakly constraining contexts. Both groups showed smaller N400 amplitudes in strongly constraining contexts, suggesting facilitated semantic processing. However, this effect was weaker and delayed in the older group, particularly among those with lower reading spans. These findings align with the age-related decline in WM capacity (Charlton et al., 2010; Hertzog, Dixon, Hultsch, & MacDonald, 2003), suggesting that individuals with reduced WM may be less able to retain and use the rich contextual information to support predictive processing.

In contrast, Otten and van Berkum (2009) found that syntactic prediction may not depend on WM capacity to the same extent. Building on their earlier study (Otten & van Berkum, 2008), they measured participants' ERP responses to prenominal determiners whose grammatical gender either matched or mismatched the expected noun. Participants were categorised by WM capacity, and early ERP effects (300–600 ms) of gender mismatch were found in both high and low WM groups. This suggests that some types of syntactic prediction, particularly those involving grammatical agreement, is not necessarily resource-demanding and may thus occur independently of WM capacity. However, only the low WM group exhibited a late negativity (900–1500 ms) to gender-inconsistent determiners. This later effect may reflect additional cognitive effort required to revise or suppress an initial prediction that was inconsistent with the input, or to maintain parallel interpretations, one with the expected noun and the other without it. Such processes are known to be WM-intensive (Ruchkin, Johnson, Canoune, & Ritter, 1990; Fiebach, Schlesewsky, & Friederici, 2001; King & Kutas, 1995; Gunter, Jackson, & Mulder, 1995).

In sum, the findings across these studies suggest that while WM may not determine whether prediction occurs, it does influence the timing and efficiency of prediction. Specifically, WM appears to support the integration of prediction-consistent input and the suppression or revision of inconsistent input. Therefore, WM is not a prerequisite for prediction itself, but a moderating factor in how effectively prediction is used during comprehension.

2.4.2 Language proficiency

Independent of its relationship with WM capacity, language proficiency has also been shown to influence predictive processing during language comprehension. As discussed earlier, L2 speakers generally demonstrate the ability to make predictions at the semantic level, although their processing tends to be slower and less consistent than that of native speakers (Grüter, et al, 2018; Hopp, 2015; Ito et al., 2017b; Ito, Pickering, et al., 2018; Martin et al., 2013). In contrast, syntactic and phonological predictions appear to be more variable across L2 speakers of different proficiency levels. Intermediate L2 speakers show little to no evidence of prediction at these levels (Kaan et al., 2014; Mitsugi, 2017; Mitsugi & MacWhinney, 2016), whereas highly proficient L2 speakers have demonstrated the ability to engage in syntactic (Foucart et al., 2014; Grüter et al., 2012; Hopp & Lemmerth, 2018) and even phonological predictions (Lozano-Argüelles et al., 2020). These findings suggest that L2 prediction is limited and highly dependent on language proficiency, particularly at the syntactic and phonological levels.

Further support comes from Peters, Grüter, and Borovsky (2015), who examined how anticipatory lexical activation during sentence comprehension varies across L2 proficiency levels. Using a visual-world eye-tracking paradigm, they compared the eye movements of high- and low-proficiency English speakers as they listened to sentences that semantically constrained the upcoming noun. The sentences were designed to promote either anticipatory activation (i.e., prediction of a likely upcoming word) or local coherence (i.e., facilitation based on immediate lexical associations). Their findings revealed that higher proficiency speakers made significantly more anticipatory fixations on the target object before it was mentioned. By contrast, lower proficiency speakers exhibited fixation patterns consistent with local coherence, suggesting that they relied more on bottom-up associations than on top-down prediction. This study highlights the critical role of proficiency in modulating prediction during L2 comprehension, with more proficient speakers being better able to exploit contextual constraints for prediction.

The improved predictive processing associated with higher L2 proficiency likely stems from more automatic lexical access, more efficient integration of syntactic and semantic information, and greater sensitivity to subtle probabilistic cues in the input (Clahsen & Felser, 2006; McDonald, 2006; Mitsugi, 2017; Ito & Pickering, 2021). These cognitive and linguistic advantages enable proficient L2 speakers to engage more effectively in top-down prediction.

By contrast, less proficient language users, such as young children and late L2 learners, tend to exhibit slower lexical access and limited grammatical knowledge, making it less likely for them to engage in top-down analysis but depend more on bottom-up, reactive processing strategies (Lew-Williams & Fernald, 2010). Additionally, because non-native speakers are often more dominant in their L1, cross-linguistic interference may influence the processing of lexical and grammatical features in L2 (Dussias, Kroff, Guzzardo, & Gerfen, 2013; Karaca, Brouwer, Unsworth, & Huettig, 2021; Spivey & Marian, 1999), while language-specific features of the L2 may not be reliably used for prediction (Foucart & Frenck-Mestre, 2011; Hopp, 2013; Lew-Williams & Fernald, 2010).

2.4.3 Interpreting experience

Interpreting experience, which is assumed to expand lexical and syntactic representations (Huettig & Pickering, 2019; Özkan, Hodzik, & Diriker, 2023), has been found to facilitate prediction during language comprehension. As discussed earlier, Lozano-Argüelles et al. (2020) reported the modulating effect of interpreting experience on prediction in non-interpreting L2 tasks. Specifically, L2 learners with interpreting experience made earlier and faster predictions than monolinguals and non-interpreter L2 learners under certain conditions. In a follow-up study, Lozano-Argüelles et al. (2023) examined whether this advantage was related to higher WM capacity and found an interaction between WM capacity and interpreting experience. Both interpreters and monolinguals benefited from higher WM capacity when using lexical stress and syllabic structure to make prediction, whereas non-interpreter L2 learners were hindered by higher WM capacity in L2 processing. They further explained that higher WM capacity allowed more possible continuations to be activated based on certain lexical stress and syllabic patterns. While interpreters enhanced their coordination ability through interpreting experience and deployed WM efficiently like monolinguals, non-interpreter L2 learners may have struggled with decision-making, resulting in slower prediction generation.

Similarly, Özkan et al. (2023) compared professional interpreters and trainees in a non-verbal listening task and found more efficient use of case-markers for prediction in individuals with more interpreting experience. They also observed that higher WM capacity facilitated prediction in professionals, but not in trainees. Collectively, these findings provide evidence for an interpreter advantage in prediction during L2 comprehension. This advantage is likely driven by a combination of factors: higher language proficiency facilitating lexical and syntactic activation; more efficient allocation of cognitive resources enabling flexible

prediction updates; and greater resilience to task demands, including time pressure and the cognitive load associated with concurrent production.

Using an ERP paradigm, Fan, Collart, and Chan (2022) investigated how interpreter expertise modulates sentence processing, particularly under conditions of language switching. Expert interpreters, intermediate interpreters, and interpreting students read Mandarin sentences that ended with either a congruent or incongruent final word. In some trials, the final word was in English, creating a code-switching condition. They found that all groups exhibited N400 effects to semantic incongruence and language switching, indicating early sensitivity to both factors. However, in the 500-700 ms window, differences emerged across groups. Expert interpreters with many years of interpreting experience showed sustained congruency effect in both monolingual and language switching conditions, while the other two less experienced groups only showed such effects in the language switching condition. Additionally, interpreting students exhibited a unique frontal negativity for incongruency in the language switching condition, suggesting increased cognitive demands. These findings highlight the role of expertise in supporting stable semantic processing and suggest that expert interpreters are better able to manage language switching without compromising predictive processing.

2.4.4 Summary of modulating factors

In summary, predictive processing during language comprehension is modulated by a variety of interacting factors. WM capacity supports the storage, maintenance, and updating of predictive representations, especially under time pressure or when multiple continuations must be evaluated. Language proficiency influences the accuracy, speed, and depth of prediction, with more proficient speakers showing more flexible and more extensive predictive processing, especially at the syntactic and the phonological levels. Interpreting experience, as a form of domain-specific expertise, enhances prediction by promoting automatic lexical access, efficient WM deployment, and adaptive processing strategies. Together, these factors suggest that predictive processing is not solely language-specific but also highly sensitive to broader cognitive and experiential profiles.

Chapter 3 Predictive processing in simultaneous interpreting

3.1 Theoretical discussion on predictive processing in SI studies

In SI studies, the process where interpreters predict the upcoming content is typically termed *anticipation*. Unlike *prediction* as defined in psycholinguistics, *anticipation* in SI studies refers more broadly to an interpreter's strategic use of context and cues to pre-emptively process or produce elements of speech before they are fully available. While prediction usually occurs without explicit intention, anticipation is often goal-directed and strategic. The present study aims to systematically investigate predictive processing during SI, encompassing both prediction in the psycholinguistic sense and anticipation as defined in SI studies.

The importance of predictive processing in SI has been long recognised in interpreting theory (Wilss, 1978; Lederer, 1981; Chernov, 1994; Gile, 1992, 2002, 2009; Setton, 1999; Amos & Pickering, 2020). Two main theoretical traditions have shaped early debates on the role of anticipation across language pairs. The "universalists" (Setton, 1999), or the "liberal arts community" (Moser-Mercer, 1994), rooted in Interpretive Theory and represented by scholars such as Lederer (1981) and Seleskovitch (1984), propose that language processing in SI relies on deverbalisation of the input, during which meaning is separated from linguistic form. From this perspective, anticipation is considered a universal phenomenon grounded in the language-independent message, and thus relatively unaffected by language pairs or cross-linguistic syntactic asymmetries. In contrast, the "bilateralists" (Setton, 1999) or the "natural science community" (Moser-Mercer, 1994) adopt an Information Processing framework and argue that language-specific features, especially syntactic and form differences, affect comprehension and reformulation during SI. According to this view, in language pairs with divergent syntactic structures, such as German and English, anticipation becomes not only beneficial but essential to compensate for delayed syntactic information and to generate a coherent and fluent output in the target language (Setton, 1999, 2005; Gile, 1992, 2002, 2009; Wilss, 1978).

Building on these foundational views, later researchers have identified the conditions for interpreters to anticipate upcoming speech. Unlike psycholinguistics studies, which typically examine prediction at specific linguistic levels, SI research considers a broader range of anticipation cues. Wilss (1978) and Lederer (1981) categorised anticipation cues into three types: 1) co-textual intralingual, such as semantic meaning, syntactic structure, lexical collocations (i.e., frequent occurrences of lexical items), and grammatical features (e.g., agreement markers and determiners); 2) extralinguistic situational cues, including world knowledge, prosody (e.g., intonation, stress), visual context (e.g., gestures, presentation slides),

and speaker identity and style; 3) standardised communication cues, like routine or formulaic expressions typical in institutional discourse (e.g., openings of speeches or diplomatic routine). This typology of cues aligns with Chernov's (1994; Chernov, Setton, & Hild, 2004) Probability Prediction (Anticipation) Model, which posits that successful anticipation depends on the objective redundancy of the message occurring at multiple levels, from syllables, words, to discourse structure. It is worth noting that what SI researchers label "anticipation via collocations" may overlap with lexical priming in psycholinguistics. For example, a noun-noun collocation (e.g., *doctor* and *nurse*) may reflect associate priming more than conscious prediction (Meyer & Schvaneveldt, 1971; Bentin et al., 1985). However, some researchers have argued that priming-triggered pre-activation should be considered a form of prediction (Lau et al., 2013; McRae et al., 2005). Overall, these factors jointly enable interpreters to predict and plan upcoming output, dynamically adjusting their processing in response to contextual and linguistic constraints.

The Effort Models, proposed by Gile (1992, 2002, 2009), offer a cognitive framework that complements these theoretical perspectives by explicitly modelling the resource constraints involved in SI. The Effort Models break the SI down into three simultaneous "Efforts" components, namely, listening (L), memory (M), and production (P), each requiring a share of the interpreter's processing capacity (PC). The model predicts that when the combined demand of L, M, and P approaches or exceeds available PC, interpreters may experience disfluency or errors. This model also accounts for the time-sensitive nature of processing. Delays in L reduce the time available for M and P, thereby increasing their processing demands. In his corpus-based study on Predictable Sentence Endings (PSEs) in Japanese, Gile (1992) found that predictability allowed interpreters to allocate less PC to L and more to M and P. Because no new input was being introduced during the PSEs, interpreters could also begin offloading information from memory, thereby reducing memory load and enhancing fluency in production. This highlights the strategic function of anticipation: by reducing cognitive load on L and M, anticipation releases resources for production and self-monitoring.

Building on this cognitive perspective, Amos and Pickering (2020) proposed a production-based model of prediction and argued for broader benefits of prediction during SI. Apart from earlier claims (Chernov, 1994; Gile, 2009; de Groot, 2011), they emphasised the importance of prediction when interpreting between languages with different word orders, where early prediction can compensate for structural delays. Drawing on psycholinguistic research, they noted that erroneous predictions are not costlier than no prediction (Frisson et al., 2017), and

that even semantically related but incorrect predictions can facilitate processing (Staub et al., 2015). In the long run, prediction errors may lead to improvements through error-based learning (Dell & Chang, 2013). These findings support the view that predictive processing is not only beneficial but adaptive, and that it can be strategically trained and reinforced in interpreters.

Taken together, these theoretical perspectives converge on the view that predictive processing in SI is a multi-faceted mechanism, shaped by intra- and extra-linguistic contexts and cognitive load. Predictive processing is not only beneficial but also adaptive, and can be strategically trained and reinforced in interpreters, to enable more efficient allocation of cognitive resources and facilitate output planning to ensure fluency and coherence even in the face of delayed or incomplete input.

3.2 Corpus-based analysis and early empirical studies of prediction during SI

Early SI studies mainly examined anticipation through corpus-based analysis and observation. One of the earliest studies, Wilss (1978), focused on syntactic anticipation in German-English interpreting. By comparing the timing of verbs rendered in English relative to the speaker's articulation, Wilss found that interpreters often produced English verbs before corresponding German verb had been fully uttered, indicating robust syntactic anticipation. This anticipatory behaviour was particularly frequent in German-English language pairs, where structural asymmetry requires interpreters to compensate for delayed grammatical information. By anticipating and front-loading verb translation, interpreters reduce ear-voice span and manage the cognitive load more efficiently.

With a similar focus, Jörg (1995) conducted an observational study with both German-English professional interpreters and interpreting students, with native language of either German or English. The results showed that successful anticipation occurred in approximately 50% of all the anticipation-likely sentences, incorrect anticipation in 2% and no anticipation in about 48%. Professional interpreters outperformed students, showing greater consistency and accuracy in anticipation. Moreover, L1 German interpreters had a better average performance than L1 English interpreters, indicating that prediction might be developed in one's L1 and thus occur more often in L1 than in L2. This is consistent with studies reviewed above that while prediction in L1 comprehension could occur at all linguistic levels, prediction in L2

comprehension mostly occurs at the semantic level but less at the syntactic or phonological levels (Mitsugi & MacWinney, 2016; Ito, Pickering, et al., 2018). As argued by Jörg, the proportion of successful anticipation might have been higher if the interpreted speeches had been spontaneous rather than prepared, as spontaneous speech tends to exhibit more redundancy, allowing for greater reliance on context. According to Gile's Effort Model (1992), prepared speeches with denser information place higher demands on listening and memory efforts, leaving less processing capacity for production and thereby limiting opportunities for anticipation.

Similarly, Kurz and Färber (2003) analysed a corpus of German-to-English SI performed by student interpreters from two native-language groups (German and English). Participants interpreted live or recorded German source speeches into English, and instances of anticipatory translation, particularly in sentence-final segments, were identified. They found that interpreters produced an anticipated translation in about one-third of all sentences, indicating that anticipation is a common strategy in German-to-English SI. Consistent with the findings of Jörg (1995), German native speakers demonstrated more frequent and accurate anticipation, highlighting the importance of source language proficiency and structural familiarity.

Through an experimental study, Seeber (2001) examined the effect of intonation in anticipation during SI. Professional L1 English L2 German interpreters were asked to interpret two German speeches: one delivered with monotonous intonation and the other with a lively prosody. While the number of anticipations was similar across both conditions, interpreters made earlier and more accurate anticipation in the monotonous condition, with fewer errors and more use of placeholders. Contrary to the original hypothesis, monotonous intonation did not hinder verb anticipation. Nonetheless, retrospective interviews revealed that interpreters subjectively believed lively intonation helped their performance. Seeber interpreted this discrepancy between participants' subjective impression and objective performance as the result of increased cognitive effort summoned to compensate for sub-optimal working conditions, aligning with findings from Moser-Mercer, Künzli, and Korac (1998).

In a more recent experimental study, Hodzik and William (2017) compared prediction in shadowing and SI tasks with L1 English L2 German non-interpreter bilinguals and interpreters. They manipulated contextual constraint (CC) level and the transitional probability (TP, the possibility that two words appearing together) and measured latencies between the sentence-final verbs in the input and output (the latencies). Negative latencies, suggesting output

preceding input, were treated as evidence of anticipation. The results showed significantly shorter latencies in high-CC contexts than in low-CC contexts in both tasks, indicating that the context semantic was used to generate expectation about upcoming content. However, TP only exhibited effects in shadowing tasks. Specifically, TP showed significant negative association with the latencies in the shadowing, but not in SI task. These results indicate that prediction during SI may rely more on high-level information (contextual cues) than on low-level information (statistical collocation). Also, anticipations were observed only in SI, not shadowing, implying that anticipation is a task-specific strategy.

Building on this approach, Chmiel (2021) conducted a longitudinal study comparing professional interpreters with interpreter trainees tested before and after a two-year training programme. All participants were late L2 learners. Using a word-translation task, participants translated target words embedded in high-constraint, low-constraint, or isolated sentences in both L1-to-L2 and L2-to-L1 directions. Word translation latencies served as the primary measure of anticipation. Across all groups, shorter latencies were observed in high-constraint sentences compared to low-constraint and isolation conditions, indicating anticipatory processing. Meanwhile, greater anticipations were observed in the L1-to-L2 direction than the other. Interpreter trainees demonstrated significant shorter translation latencies at the end of training, which was not observed among the professionals. However, the professionals demonstrated more efficient inhibitory control, suggesting that experience enhances executive control rather than increasing the extent of prediction.

3.3 Visual-world eye-tracking studies

Despite evidence for anticipation during SI, off-line measures do not directly capture the pre-activation of linguistic representations during SI tasks. The input-output latencies can be modulated by factors beyond prediction, such as integration difficulty and strategy preference (Timarová, Dragsted, & Hansen, 2011). Accordingly, the shorter latencies in highly constraining contexts could not be simply explained as the effect of prediction. In contrast, VWP provides a straightforward window for the time course of the predictive process, thereby capturing pre-activation through direct, observational means. Liu et al., (2022) replicated the design of Hintz, Meyer, and Huettig (2017) and explored whether bilinguals could predict upcoming semantically related words in their L1 during simultaneous and consecutive interpreting (CI) tasks. They found that most participants demonstrated predictive eye

movements during both SI and CI, suggesting that even under high processing demands, semantic prediction in the source language is generally preserved. However, approximately 25% of participants showed no predictive fixations in the SI condition, indicating individual variation in the manifestation of prediction under cognitive load.

Amos et al. (2022) replicated the design of Ito, Pickering, et al. (2018) and further investigated the predictive abilities of professional interpreters and translators with dual focuses on both semantic and phonological prediction. Participants, all L1 French professionals, were required to interpret English sentences (e.g., "*The man will open his…*"), while viewing a display containing the semantically predictable target (e.g., *mouth*), an English phonological competitor (e.g., *mouse*), a French phonological competitor (e.g., *bouchon* for *cork)*, or an unrelated object, with three other unrelated distractors. Semantic prediction was evidenced by early fixations on the target image prior to word onset. The results showed that both interpreters and translators made anticipatory fixations on the target object, indicating successful semantic prediction. However, there was no evidence of phonological prediction in either group. Furthermore, no significant differences were observed between interpreters and translators, suggesting that interpreting experience did not enhance predictive behaviour in this context.

Expanding on this line of enquiry, Amos et al. (2023) conducted a longitudinal study to examine whether SI training enhances prediction. Using the same visual-world design, they tested student interpreters both before and after two semesters of formal training. The primary aim was to determine whether training led to measurable changes in predictive behaviour. Results showed that students exhibited semantic anticipation even before training began, as indicated by anticipatory fixations on the target object around 350 milliseconds prior to the target word onset. Unexpectedly, no significant changes were observed after training in either the timing or extent of prediction. Moreover, prediction performance did not correlate with final interpreting achievement. These findings suggest that the ability to semantically anticipate during SI may be an inherent skill tied to general language comprehension rather than one that is substantially shaped by interpreting instruction—at least within the time frame and task constraints used in this study.

Together, these three studies converge on the finding that semantic prediction is reliably present during SI, even among novice interpreters or in cognitively demanding settings. However, they all used single sentences as auditory stimuli, whereas in realistic interpreting practices, typical utterances involve several coherent sentences. On one hand, the involvement

of multiple coherent sentences can foster the overall representation of the meaning of the text beyond the immediate, language-dependent representation, thereby building a "mental model" (Johnson-Laird, 1983) or "situation model" (Kintsch, 1998). This is supported by the evidence that higher accuracy was achieved in a cumulative cloze test that uses paragraphs involving multiple sentences than in single-sentence cloze test (Hoffman, 1980) and that comprehension difficulty may arise when there is incongruence between the ongoing sentence and the contextual information (O'Brien & Albrecht, 1992; Pickering & Traxler, 1998). On the other hand, retrieving contextual information may undermine prediction during SI as it requires the storage and activation of the inter-sentential information, that is, the "memory" effort, as termed in the Effort model for SI by Gile (2002, 2009), or the functioning of WM (Dong & Cai, 2015). It therefore remains unknown when interpreting full paragraphs, whether prediction is impeded due to higher demands for WM, or, on the contrary, facilitated by the richer information in the internal representation of context established by previous sentences.

Chapter 4 Theoretical framework

4.1 Neurophysiological and computational accounts of prediction during language comprehension

4.1.1 One-system account

Drawing on a simple recurrent network (Elman, 1990), Altmann and Mirković (2009) proposed a one-system account of linguistic prediction, in which sentence comprehension and event understanding are handled by a single, integrated cognitive system. This model is grounded in four core principles: mapping across domains, prediction, context sensitivity, and representation across time. First, the system continuously maps incoming linguistic input onto a representation of an event or situation, allowing linguistic and non-linguistic domains (such as visual or experiential knowledge) to be processed in parallel. Second, prediction is not treated as a separate process but as an emergent property of these dynamic mappings, with each new word updating the current mental state, which in turn adjusts expectations for upcoming input. Third, the model emphasises context: not just immediate linguistic cues but also visual scenes, prior discourse, and world knowledge are encoded and used to shape predictions. Lastly, comprehension unfolds over time via a memory of prior input states, so that the system's current state inherently reflects prior linguistic and conceptual context, guiding both interpretation and anticipation. Together, these principles explain how language users continuously and efficiently anticipate forthcoming input by aligning linguistic structures with representations of real-world events.

Empirical evidence for this one-system account comes from VWP studies. One key example is Altmann and Kamide (1999), who showed that upon hearing semantically constraining verbs like "*eat*" in the sentence "*The boy will eat the...*", participants made anticipatory eye movements to edible objects (e.g., *cake*) before the noun was spoken. This suggests that listeners integrate linguistic input with visual context in real time to anticipate plausible upcoming content. In a follow-up study (Altmann & Kamide, 2007), they found that even partial linguistic input could activate detailed event knowledge. After hearing "*The cat has killed...*", participants looked towards semantically associated objects like feathers, indicating prediction driven by event-based inference rather than simple lexical priming. These studies support the idea that prediction emerges from the system's ongoing mapping of language onto situation models that integrate linguistic and non-linguistic cues. Rather than invoking a separate prediction module, the one-system account sees anticipation as a natural consequence of incremental comprehension embedded in a unified, time-sensitive cognitive architecture.

However, not all findings are consistent with the one-system account. A study by Rommers, Meyer, and Huettig (2015) offers a different perspective by suggesting that prediction during language comprehension may involve separate verbal and non-verbal systems. In their visual-world eye-tracking experiment, participants listened to highly constraining sentences (e.g., "*In 1969, Neil Armstrong was the first man to set foot on the…*") while viewing a visual display with four images: the correct target (e.g., *moon*), a shape-related competitor (e.g., *tomato*), and two unrelated distractors. The results showed that listeners not only predictively fixated on the target object but also made anticipatory looks toward the shape-related competitor, suggesting that participants predicted upcoming content using both semantic and visual features. Meanwhile, anticipatory fixations on the correct target were associated with verbal ability (vocabulary and verbal fluency), while fixations on the shape competitor were linked to non-verbal attentional skill (measured through a spatial cueing task). This dissociation suggests that different types of prediction, semantic and visual, may rely on distinct cognitive systems. These findings challenge the one-system view and provide evidence that separate mechanisms, language-based and attentional- or perceptual-based, may work in parallel to support prediction during comprehension.

4.1.2 Two-system accounts

Huettig (2015) adopted the two-system proposed by Kahneman (2011) to language research and introduced a dual-system account of prediction in language processing. According to this framework, prediction arises from two qualitatively different cognitive systems. System 1 is fast, automatic, and associative—it uses simple mechanisms such as spreading activation based on frequently co-occurring words. In contrast, System 2 is slower, controlled, and resource-intensive—it relies on deliberate reasoning, working memory, and deeper syntactic or semantic analysis. Huettig (2015) also refers to these systems as the "dumb" route (System 1), which operates effortlessly through surface-level associations, and the "smart" route (System 2), which draws on structured linguistic knowledge and higher-level context. These two systems may work together or separately depending on the processing demands and the comprehender's cognitive resources. The dual-system view challenges models that assume prediction is always automatic and instead proposes that prediction can be either shallow and habitual or deep and reasoning-based, depending on the situation.

In addition to the behavioural evidence from studies like Rommers et al. (2015), neuroimaging studies offer further support for the dual-system account. Kuperberg (2007) has linked two

critical ERP components, i.e., N400 and P600, to the different types of predictive processing. The N400 is typically associated with automatic, memory-based semantic activation. Its amplitude is reduced when a word is predictable or semantically related to prior context, reflecting the operation of System 1, which relies on learned associations and lexical co-occurrence (Kutas & Hillyard, 1980; Federmeier & Kutas, 2005). In contrast, the P600, a positive deflection that occurs later, is often observed when the input violates syntactic or semantic expectations, requiring reanalysis or integration. This component is thought to reflect System 2 activity, as it involves controlled, effortful processing, especially when comprehenders need to revise or update their interpretation (Kuperberg et al., 2003).

Further support for the dual-system account comes from fMRI studies, which reveal that different brain regions are engaged depending on the type of predictive processing involved. For System 1, which supports fast, automatic, and associative prediction, neuroimaging studies have shown that temporal and inferior prefrontal regions are consistently activated during tasks involving semantic priming (Copland et al., 2003; Kotz, Cappa, von Cramon, & Friederici, 2002; Matsumoto et al., 2005; Rossell, Price, & Nobre, 2003). These same regions have also been found to respond to semantic incongruities in sentence processing, mirroring the neural sources of the N400 observed in ERP studies (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Kiehl, Laurens, & Liddle, 2002; Kuperberg et al., 2003). In contrast, System 2, associated with controlled and effortful processing, has been linked to activity in a broader network involving the posterior inferior frontal cortices, motor and parietal cortices, and middle and superior prefrontal cortices (Kuperberg, 2007; Kuperberg et al., 2003). These brain regions are typically involved in tasks that require syntactic parsing and executive control, particularly when predictions based on simple associations are insufficient and deeper combinatorial processing is needed. For example, when comprehenders are faced with structurally complex or ambiguous input, these frontal and parietal regions show increased activation, reflecting the additional cognitive resources required to generate or revise predictions based on higher-order linguistic and contextual analysis. These neuroimaging findings align with the dual-system account, supporting that prediction engages multiple cognitive mechanisms rather than a single system, with the specific neural network recruited depending on the complexity of the input and the comprehender's cognitive capacities.

4.2 Prediction-by-production and prediction-by-association

While the one-system and the two-system accounts offer valuable insights into the neurophysiological and computational underpinnings of prediction, they do not fully capture the cognitive mechanisms by which prediction is generated during real-time language comprehension. In contrast, the prediction-by-production and the prediction-by-association accounts (Pickering & Garrod, 2013; Pickering & Gambi, 2018) shift the focus from brain architecture to cognitive mechanisms, emphasising how predictions are generated and used during comprehension. Briefly, prediction-by-production suggests that listeners simulate upcoming speech by covertly engaging their own language production system, particularly when the context is highly constraining and sufficient processing time is available. Prediction-by-association, on the other hand, explains prediction as the automatic activation of related linguistic representations, based on established associative links in semantic memory (Huettig, 2015). Together, these two accounts provide a psychologically grounded framework for understanding prediction in both monolingual and bilingual contexts.

4.2.1 Conceptual foundations and models

Prediction-by-production accounts for linguistic prediction through the covert use of the language production system during comprehension. According to Pickering and Gambi (2018), comprehenders simulate the speaker's utterance by activating production mechanisms internally to anticipate upcoming linguistic material. This simulation begins with the inference of the speaker's communicative intention, derived from prior context, and leads to the generation of prediction at multiple levels, including semantic, syntactic, and phonological. Crucially, this process mirrors the stage of overt language production: comprehenders formulate a conceptual message, encode its grammatical structure, retrieve appropriate lexical items, and activate the corresponding phonological forms. Prediction-by-production is resource-dependent and optional, as it requires cognitive resources such as working memory; it is therefore less likely to occur under conditions of high task demand or reduced processing capacity. Nonetheless, when engaged, this mechanism allows for highly specific and structured predictions, which facilitate faster and more efficient language processing.

In contrast, prediction-by-association offers a less resource-intensive explanation for predictive processing. Instead of generating prediction through simulating production, comprehenders automatically activate related words or concepts based on previously learned statistical regularities and co-occurrence patterns in language (Huettig, 2015; Pickering & Gambi, 2018).

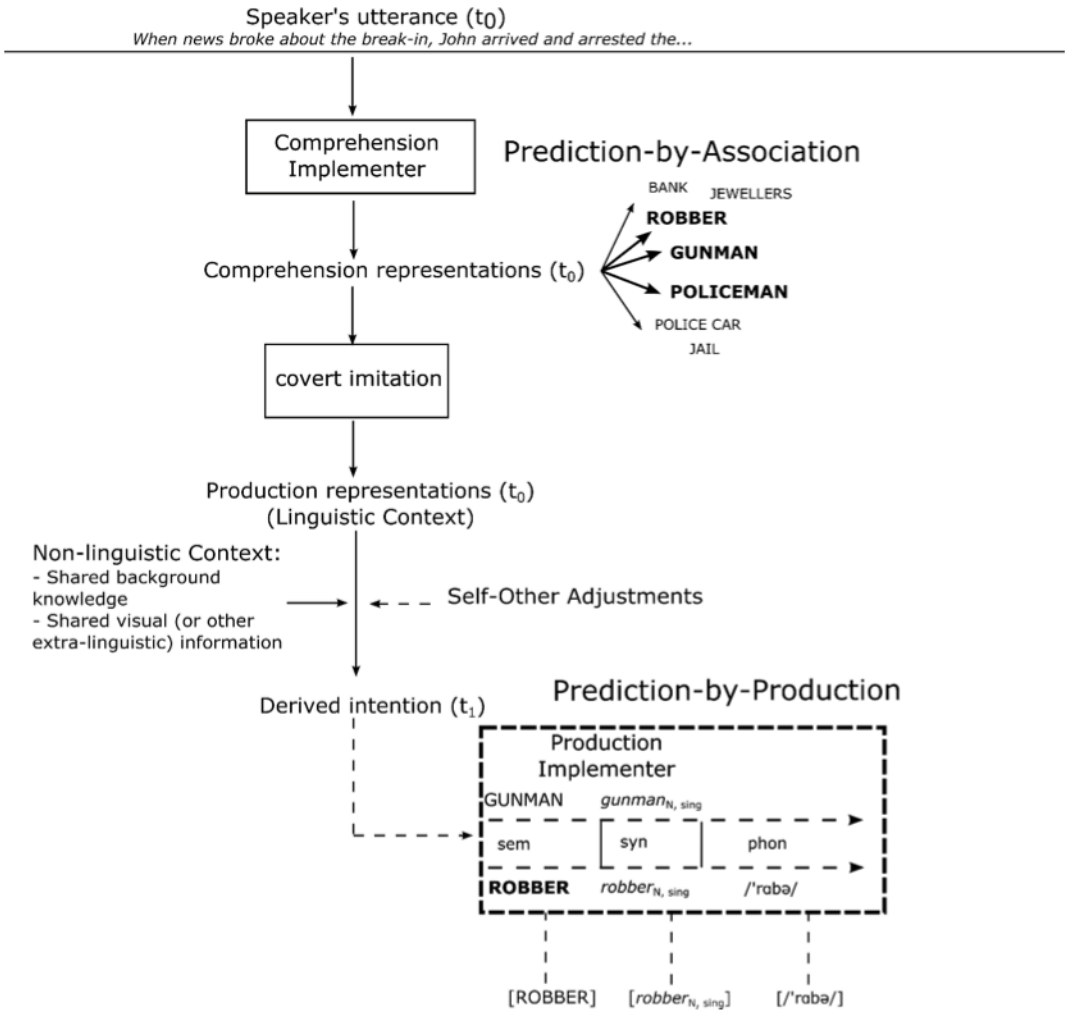
For instance, upon hearing the word “doctor,” related concepts such as “nurse” or “hospital” may become activated without deliberate effort. This form of prediction is fast, automatic, and relatively superficial, typically supporting low-precision expectations. It does not require deep syntactic or phonological processing, making it especially useful when the context is weakly constraining or when cognitive resources are limited. Prediction-by-association aligns with System 1 in the dual-system account, as it is automatic, experience-driven, and relatively inflexible. It also helps explain predictive effects observed in young children, ageing adults, or second-language speakers, who may lack the cognitive or linguistic capacity to engage in prediction-by-production but still demonstrate sensitivity to probabilistic cues in language.

Although prediction-by-production and prediction-by-association rely on distinct mechanisms, they are not mutually exclusive but often operate in a coordinated and temporally structured manner during language comprehension (Figure 4-1). When the listener hears the speaker’s utterance (t_0), an initial comprehension representation is formed. This representation rapidly activates a broad set of semantically or contextually related lexical items through prediction-by-association. At the same time, the listener may begin covert imitation, mapping comprehension representations onto the production system. This process incorporates both linguistic cues and non-linguistic context (such as background knowledge or visual information), leading to a derived interpretation of the speaker’s intention (t_1). At this point, prediction-by-production generates more detailed and structured expectations, activating specific semantic, syntactic, and phonological features of anticipated words. In this way, prediction-by-association provides a quick, low-effort foundation for predictive processing, while prediction-by-production adds precision when cognitive resources and contextual support permit. Together, they support a flexible and layered prediction, adapting dynamically to varying task demands and communicative contexts.

The prediction-by-production and prediction-by-association accounts are largely compatible with the multi-mechanism framework proposed by Huettig (2015), known as the PACS model. Huettig argues that predictive language processing relies on at least four partially distinct mechanisms: production-, association-, combinatorial-, and simulation-based processes. While the production-based mechanism in PACS share similarities with Pickering and Gambi’s (2018) prediction-by-production account, a key distinction lies in representational specificity. Pickering and Gambi suggest that prediction involves a forward model that draws on impoverished representations generated by the production system. In contrast, Huettig proposes that fully specified production representations, i.e., at the semantic, the syntactic, and

the phonological levels, are engaged during prediction. The association-based mechanism in PACS closely aligns with the prediction-by-association account, as both posit that prediction can arise from simple automatic spreading activation among stored linguistic representations.

Figure 4-1. The theoretical model of prediction-by-production and prediction-by-association from Pickering and Gambi (2018).



Although Huettig distinguishes between combinatorial and simulation-based mechanisms, these are not incompatible with Pickering and Gambi's framework. The combinatorial-based mechanism, which integrates multiple linguistic constraints to build structured meaning representations, arguably operates within the production system described in prediction-by-production. Likewise, the simulation-based mechanism, which draws on embodied experience and world knowledge to support prediction, can be understood as contributing to early stages of production-based prediction, particularly in deriving communicative intentions. Thus, while the PACS model offers a broader taxonomy of predictive mechanisms, its core components are

conceptually integrated into the processes described by Pickering and Gambi, supporting a unified view of predictive processing as both flexible and multi-layered.

4.2.2 Empirical evidence

4.2.2.1 Evidence for prediction-by-production

A series of visual-world eye-tracking studies by Knoeferle and colleagues provide support for the prediction-by-production account by demonstrating how listeners use linguistic input, visual context, and world knowledge to generate structured predictions during sentence comprehension. In Knoeferle, Crocker, Scheepers, and Pickering (2005), participants viewed static scenes involving two human characters and an action, such as a man and a girl engaged in a tickling event. They then listened to temporarily ambiguous German sentences, such as *Der Mann, den das Mädchen kitzelt...* ("The man that the girl is tickling…"), where the verb-final structure delayed disambiguation until the sentence's end. Importantly, thematic roles in German can only be resolved using case markings and the final verb. The eye-tracking data revealed that listeners made anticipatory fixations to the character most likely to serve as the agent or patient, based on the interaction of visual cues and early syntactic information. Such anticipatory fixations occurred before the verb was encountered, suggesting that listeners used linguistic and visual context to simulate the unfolding utterance. Rather than relying solely on passive association, participants internally generated a structured representation of the sentence by engaging their own production mechanisms.

Knoeferle and Crocker (2006, 2007) extended this work by showing that world knowledge and recent visual experience also contribute to predictive processing. In the 2006 study, listeners viewed event scenes involving plausible actions (e.g., a boxer punching a politician) while hearing verb-final German sentences. They found that participants used both real-world plausibility and linguistic input to anticipate upcoming sentence elements, again before the critical verb appeared. In the 2007 study, the researchers tested how recently viewed dynamic scenes influenced sentence interpretation. Participants watched short animations of events, followed by a static scene and a spoken sentence. When the sentence aligned with the recently observed event, listeners showed anticipatory fixations to relevant referents; when it conflicted, processing slowed. These results further support prediction-by-production by showing that listeners use recent perceptual input and real-world expectations to internally model likely sentence continuations.

Evidence from ERP studies also support prediction-by-production. As discussed earlier in Section 2.1.1., ERP studies have showed that semantic prediction can lead to the activation of syntax (Otten et al., 2007; Otten & van Berkum, 2008; van Berkum et al., 2005; Wicha et al, 2004) and even to activation of form (DeLong et al., 2005; Ito et al., 2017b). These findings support prediction-by-production as they show a pattern of comprehenders' prediction that aligns with the directional flow of language production, proceeding from semantics, to syntax, and then to form. Additional support comes from speech perception and dialogue tasks showing that listeners adjust their phonemic perception based on predicted words. In phoneme restoration studies, for instance, listeners perceive missing or degraded speech sounds more clearly when the surrounding context strongly suggests a particular word (Gagnepain, Henson, & Davis, 2012; Sohoglu, Peelle, Carlyon, & Davis, 2012). This effect is enhanced when the predicted word is specific and semantically plausible, suggesting that the comprehender has pre-activated the word's phonological form. These findings suggest that comprehenders use production mechanisms to predict not just meaning but the sound structure of upcoming speech, further supporting the claim that prediction-by-production engages deep, form-level representations in real time.

4.2.2.2 Evidence for prediction-by-association

Metusalem et al. (2012), on the other hand, provide compelling evidence for prediction-by-association. As discussed earlier, they observed a graded N400 response, with the smallest amplitude for expected word, larger for event-related words, and the largest for unrelated words. This graded N400 effects suggest that event-related words were activated not through top-down analysis of the preceding discourse, but via associative links, aligning with the prediction-by-association account.

Two studies by Kukona and colleagues (2011, 2014) provide converging evidence for the prediction-by-association account by demonstrating that predictive processing during language comprehension often involves the automatic activation of semantically related words rather than the structured simulation of a specific upcoming utterance. In Kukona et al. (2011), participants listened to sentences of varying constraint (e.g., "*Toby arrests/notices the...*") while viewing visual displays containing the target (e.g., *crook*), semantically related items (*policeman*), and unrelated distractors. Anticipatory fixations emerged not only to the most predictable referent but also to semantically related alternatives, especially in early time windows, suggesting that listeners activated a network of plausible lexical candidates based on

prior co-occurrence patterns. Similarly, Kukona, Cho, Magnuson, and Tabor, (2014) found that while processing simple predictive sentences like "*The boy will eat the cake*", listeners shifted attention to both semantically and phonologically related competitors (e.g., *cookie*, *cape*) before the critical noun was heard. These patterns of lexical competition were further supported by computational modelling, which showed that associative overlap could account for the observed interference effects. Together, these studies indicate that listeners draw on automatic spreading activation to pre-activate a range of plausible continuations based on learned associations. Unlike prediction-by-production, which involves syntactic and phonological simulation via the production system, the findings here point to fast, low-effort, probabilistic prediction that arises from associative memory.

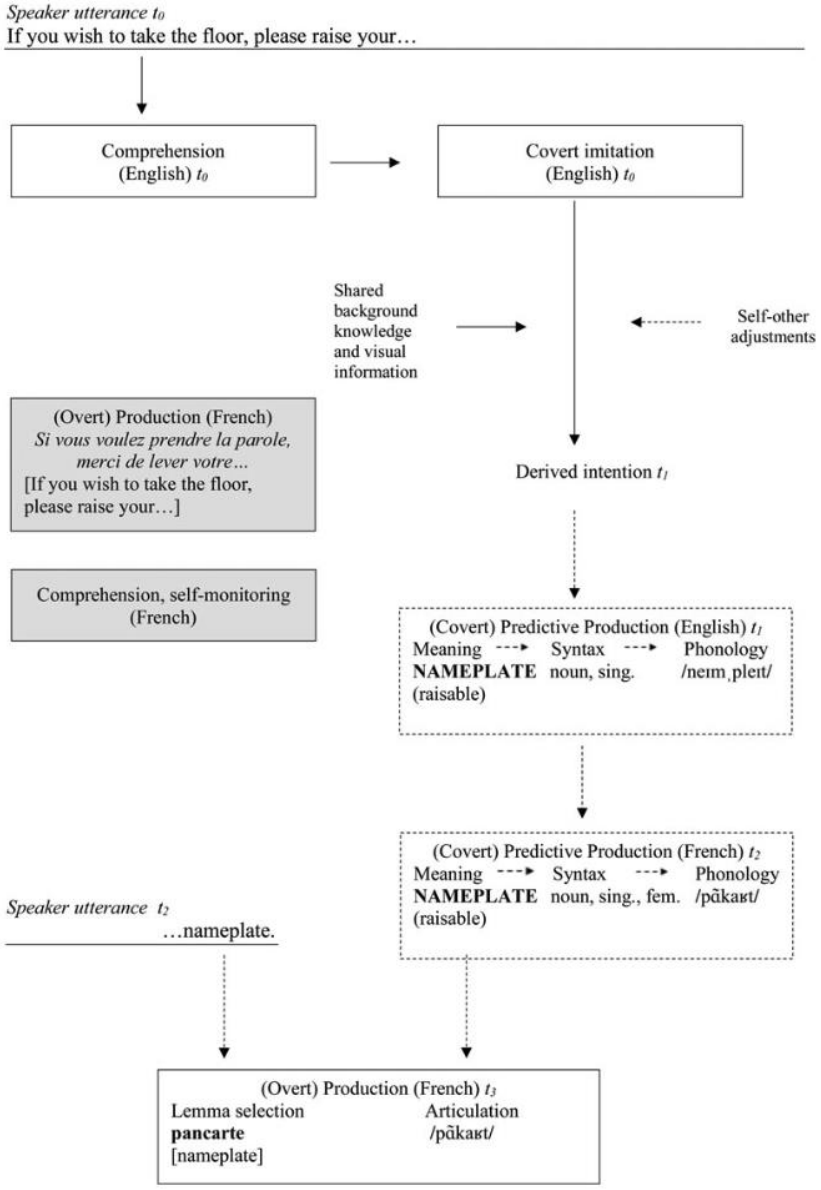
4.3 Prediction-by-production in SI

Amos and Pickering (2020) extended the prediction-by-production account to accommodate the unique cognitive demands of SI by proposing a tailored model of how interpreters generate predictions in real time. Unlike typical comprehension tasks, SI requires the interpreter to comprehend and produce language concurrently, often with incomplete input and under intense time pressure. Interpreters must not only understand the speaker's message as it unfolds but also reformulate and articulate it in the target language, frequently before the source utterance is complete. To account for these unique conditions, Amos and Pickering argued that prediction-by-production in SI serves two critical functions: supporting comprehension of the source language and facilitating early preparation for production in the target language. This dual function entails a dual use of the production system: both to predict upcoming elements in the source speech and to plan and produce the interpreter's own speech. Likewise, the comprehension system is used dually: for understanding the original speech and for self-monitoring the interpreter's output.

Given the cross-linguistic nature of SI, they proposed separate models of prediction-by-production in SI based on the degree of syntactic alignment between the source and target languages. In SI between syntactically matched language pairs, prediction-by-production operates in a manner broadly similar to language comprehension but is shaped by the unique demands of bilingual reformulation. Upon hearing the speaker's utterance (t_0), the interpreter begins to comprehend and simultaneously engage in covert imitation of the source-language structure. This process is informed by both linguistic input and contextual cues and leads to the

derivation of speaker's intended message (t_1). What distinguishes SI from typical comprehension is that interpreters must not only predict the next word in the source language but also prepare its equivalent in the target language. This involves a cascade of internal processes that mirror overt production, including semantic analysis, syntactic encoding, and phonological retrieval, first in the source language (e.g., *nameplate*) and then reformulated in the target language (e.g., *pancarte* in French). The close alignment in syntactic structures allows a relatively direct transition from prediction to production. Unlike comprehension, where prediction primarily facilitates understanding, prediction-by-production in SI also supports immediate and fluent speech output.

Figure 4-2. The prediction-by-production model for syntactically matched language pairs in SI from Amos and Pickering (2020).

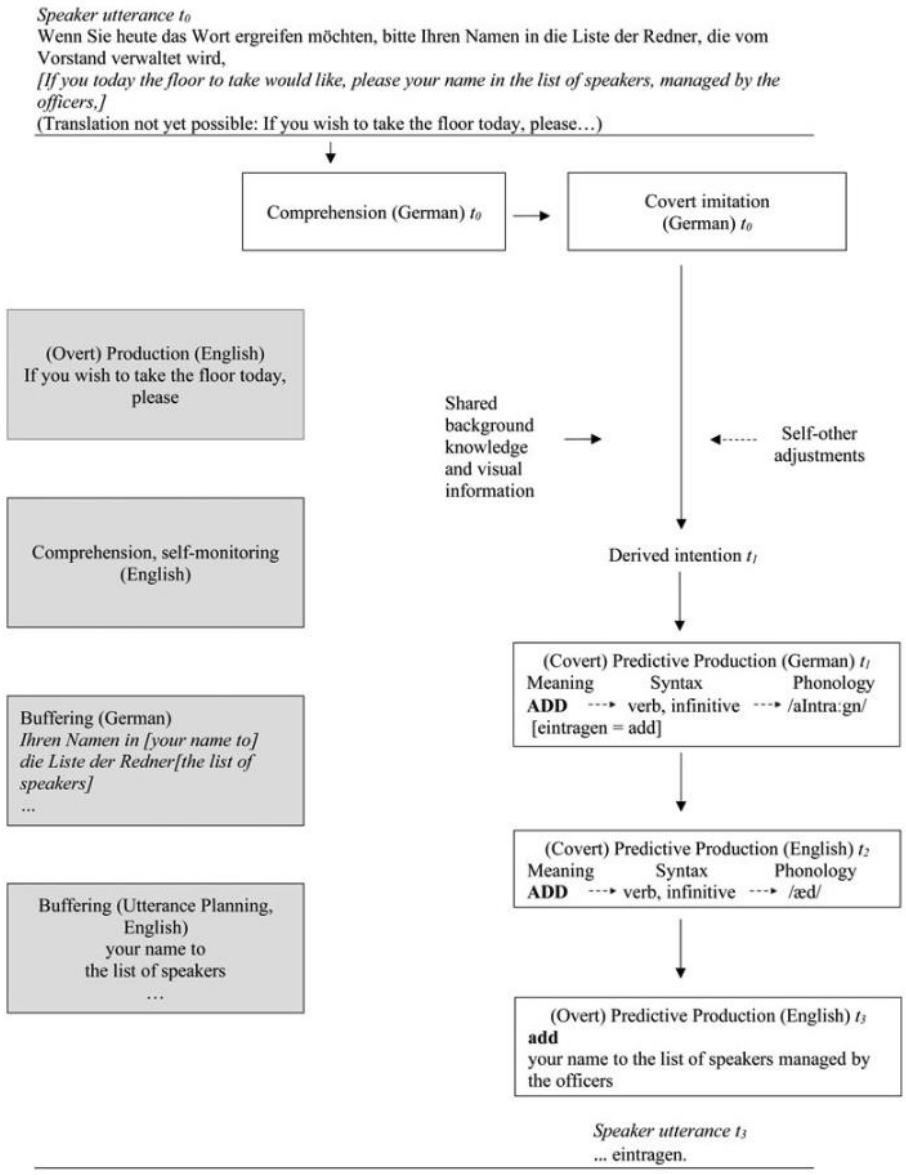


In syntactically mismatched language pairs (e.g., German and English), where key sentence elements like verbs are delayed in the source language, prediction becomes even more critical. While prediction-by-production follows a similar overarching workflow, additional cognitive strategies are required to manage cross-linguistic differences in word order. As shown in Figure 4-2, the source language (German) verb appears sentence-finally, delaying access to crucial propositional information. To maintain fluency in the target language (English), interpreters must engage in early prediction and buffering. After initial comprehension and covert imitation at t_0, the interpreter derives the speaker's intended meaning and initiates covert predictive production in the source language. Critically, the interpreter pre-activates the upcoming verb (e.g., *eintragen* "to add") using semantic, syntactic, and contextual cues before it is encountered in the input. This verb is then reformulated in the target language at t_2 and prepared for early production at t_3. Unlike syntactically matched pairs, where production proceeds linearly with the input, mismatched pairs require buffering: for example, the phrase "*your name to the list of speakers*" must be held in WM until the predicted verb can be overtly expressed. This increased complexity highlights the interpreter's need to restructure and re-order content in advance, reinforcing the strategic importance of prediction-by-production in managing syntactic asymmetry during SI.

In sum, while prediction-by-production in language comprehension involves covert simulation primarily to aid understanding, its role in SI is broader and more dynamic. Interpreters simulate upcoming source language input not only to support comprehension but also to prepare coherent and timely output in the target language. This dual application reflects the adaptive nature of the production system in demanding bilingual contexts and illustrates how prediction facilitates interpreters to stay ahead of the speech stream, mitigate processing delays, and resolve ambiguities more efficiently.

However, Amos and Pickering's account (2020) focuses exclusively on prediction-by-production, leaving a cavity in understanding how a bottom-up, automatic mechanism, such as prediction-by-association, might also operate during SI. While they argue for a central role for prediction-by-production during SI, this process is cognitively demanding and may not always be available under the high processing loads and time pressure that characterise SI practices. In contrast, prediction-by-association is relatively automatic, low-effort, and therefore makes it well-suited to operate in SI especially when cognitive resources are limited.

Figure 4-3. The prediction-by-production model for syntactically mismatched language pairs in SI from Amos and Pickering (2020).



This perspective also aligns with findings from Liu et al. (2022). They used the same stimuli as Hintz et al. (2017), in which the target objects were more strongly associated with critical verbs than distractors. The evidence of prediction in Liu et al. (2022) is therefore compatible with both prediction-by-production and prediction-by association. From the prediction-by-production perspective, upon hearing a critical verb in the sentence, interpreters covertly simulate its production in conjunction with the preceding linguistic context, forming an intention that aligns with the speaker's likely communicative goal. This derived intention then selectively activates the representation of the target object, as it constitutes the most contextually appropriate continuation, given the verb's semantic constraints and the sentence

structure. Consequently, the interpreters focus their attention on the target object, filtering out distractors that conflict with the anticipated message. By contrast, the prediction-by-association account explains prediction as the result of activation spreading through associative memory networks. Upon hearing the verb, its conceptual representation activates related features, with the target receiving increased activation due to its stronger associative link compared to the distractor. This activation bias leads to more fixations on the target, even in the absence of a deliberately derived communicative intention. In other words, it is possible that interpreters may have relied on both mechanisms during SI.

4.4 Present study

The present study systematically investigates predictive processing during SI through the triangulation of VWP, interpreting performance, and retrospective self-report. The present study primarily examines prediction, defined as pre-activation of linguistic elements in psycholinguistic research. Given the interpreting-specific context of the current research, the present study also extends to incorporate the broader notion of anticipation commonly used in interpreting studies, namely the strategic use of linguistic and extralinguistic cues to pre-emptively process or produce upcoming content (Chernov, 1994; Moser-Mercer, 1994; Seeber, 2001).

Using VWP, the present study examined real-time predictive processing during SI of extended discourse comprising full paragraphs. By examining this demanding SI task, the study aims to explore the cognitive mechanisms underpinning prediction during SI, with a focus on testing two accounts: *prediction-by-production* (i.e., top-down contextual analysis) and *prediction-by-association* (i.e., bottom-up activation based on general linguistic associations) (Pickering & Gambi, 2018; Amos & Pickering, 2020). Offline measures, including latency measures and interpreting quality, were incorporated to examine how predictive processing relates to or interact with performance. These findings were further triangulated with retrospective self-report, to provide a comprehensive perspective on predictive processing during SI. By including both professional and student interpreters, the study also investigates expertise-related differences in the cognitive rhythm and metacognitive control of predictive processing during SI.

In the visual-world eye-tracking experiment, participants listened to and simultaneously interpreted four English paragraphs into Chinese while viewing visual displays. Each paragraph included several verb-mediated experimental sentences, each containing a critical verb (CV) that created either a predictable (e.g., *eat*) or unpredictable (e.g., *buy*) context for a target word (TW) (e.g., *In the station store, commuters are eating/buying freshly made bread.*). Each visual display contained four objects: the target (e.g., *bread*), a distractor (e.g., *bone*), and two semantic competitors. The two competitors were designed to compete with the target based on their association strength with the CVs and their plausibility within the context. The implausible competitor (e.g., *turkey*) was highly associated and semantically compatible with the predictive verb but was contextually implausible in the predictable condition (commuters are unlikely to eat a whole turkey at a train station in the morning). The plausible competitor (e.g., *juice*) was strongly associated with the unpredictive verb and contextually plausible in the unpredictable condition. The distractor (bone) was unrelated to either CVs or implausible in both conditions.

This design independently manipulated contextual plausibility and verb-noun associative strength to differentiate between top-down analysis of the context (i.e., prediction-by-production), and bottom-up activation through lexical association (i.e., prediction-by-association). Accordingly, hypotheses H4-H6 are revised as follow:

> H4: Under a combined account, participants would predictively fixate more on the target in the predictable condition and on the plausible competitor in the unpredictable condition, reflecting associative processing in the absence of strong contextual cues.
>
> H5: Under a prediction-by-production only account, participants would predictively fixate on the target in the predictable condition and on both the target and plausible competitor in the unpredictable condition, due to their higher contextual plausibility in the given contexts.
>
> H6: Under a prediction-by-association only account, participants would predictively fixate on the implausible competitor in the predictable condition and on the plausible competitor in the unpredictable condition based on their strong associations with the CVs.

Chapter 5 Research methodology

5.1 Manipulation of predictability

5.1.1 Cloze probability as a proxy for prediction-by-production

The cloze test is commonly used to measure the predictability of a word in a given context. Participants are typically asked to complete sentence fragments by filling in gaps, usually under no time pressure. The proportion of participants who supply a particular word is then calculated as the word's cloze probability (Bloom & Fischler, 1980). Cloze probability has been found to correlate with reading time in eye-tracking (Frisson, Rayner, & Pickering, 2005; Rayner, Slattery, Drieghe, & Liversedge, 2011) and self-paced reading studies (Roland, Yun, Koenig, & Mauner, 2012; Smith & Levy, 2013), as well as the N400 component in ERP studies (Kutas & Hillyard, 1984). However, the instructions given to participants in cloze tasks vary across studies, possibly leading to different strategies being used. Some studies provide minimal instructions, asking participants simply to complete the sentence or context (Otten et al., 2007; Altarriba, Kroll, Sholl, & Rayner, 1996), without explicitly encouraging prediction (s). Others use more specific instructions, such as asking participants to supply the first word or noun that comes to mind (Frisson et al., 2017; Martin et al., 2013; Thornhill & van Petten, 2012; Chou, Huang, Lee, & Lee, 2014), the most likely word (Ashby, Rayner, & Clifton, 2005; Dambacher et al., 2012; Staub, 2011), or the most suitable continuation (DeLong et al., 2005; Kleinman, Runnqvist, & Ferreira , 2015). These differences, along with individual participant preferences, may result in reliance on different language systems, e.g., production versus comprehension, when completing the task. Therefore, the cognitive mechanisms underlying the cloze test remain uncertain, and it is unclear what cloze probability truly reflects.

To address this issue, Staub et al., (2015) conducted an experiment in which participants produced spoken continuations to sentence fragments presented via rapid serial visual presentation. The latency of the spoken response was measured. They found a strong correlation between spoken responses and offline cloze norms when the target words had high cloze probability (≥ 90%). Moreover, response latency was negatively correlated with both cloze probability and contextual constraint: that is, participants responded more quickly when the target word was more predictable and when the context was more constraining. These results suggest that cloze tests, even when administered offline and without time constraints, reliably reflect language users' predictive tendencies. Staub et al. proposed an activation-based race model to explain their findings: all potential sentence completions are activated in parallel and independently race toward a response threshold. In this model, high-cloze words are more

strongly activated and therefore reach the threshold more quickly than low-cloze words. However, this model challenges the interpretation of cloze probability as a direct measure of predictability, because a word may be strongly activated without being explicitly predicted. In contrast, other researchers argue that the pre-activation of features associated with a likely upcoming word should still be considered part of predictive processing (Kutas, DeLong, & Smith, 2011).

Further insight into the relationship between cloze probability and prediction comes from an ERP study by Chou et al. (2014). They examined the effects of cloze probability and semantic constraint on ERP components associated with prediction (P200 and N400). Instead of using sentences as stimuli like typical ERP studies, they created a set of seventy-two Chinese phrases in a fixed "numeral + classifier + noun" format and manipulated classifier constraint (strong vs. weak) and noun cloze probability (high, low, implausible). Classifier constraint was operationalised as the inverse of the number of nouns that could plausibly follow a given classifier. Classifiers compatible with many nouns were considered weakly constraining, while those compatible with only a few were considered strongly constraining. The results showed an interaction between classifier constraint and noun cloze probability on the N400. For weakly constraining classifiers, the N400 amplitude decreased as noun cloze probability increased, with high-cloze nouns eliciting the smallest N400. A similar pattern was observed for strongly constraining classifiers, but the difference between low-cloze and implausible nouns was not statistically significant. These findings suggest that the N400 is modulated by both cloze probability and contextual constraint. However, the use of classifier-noun combinations limits its generalisability to full sentence comprehension, where multiple contextual cues typically contribute to prediction.

In summary, although the underlying cognitive mechanisms of the cloze test are debated, a broad body of ERP and eye-tracking studies supports its utility as an index of predictive processing. Despite differing theoretical interpretations, the current study adopts the conventional use of cloze probability as a proxy for predictability. To best approximate the conditions of SI and drawing on the prediction-by-production framework (Pickering & Garrod, 2007, 2013; Pickering & Gambi, 2018), the cloze test used here was designed to encourage participants to rely on their language production system. Moreover, in alignment with the VWP used in the main SI task, the cloze test deviates slightly from the standard format. Instead of freely completing sentence fragments, participants are presented with several line-drawing objects for each gap and asked to name the one they believe is most likely to come next. This

design better simulates the visual and lexical constraints present in the main experimental task and allows for a more targeted measure of predictive processing.

5.1.2 General word association as a proxy for prediction-by-association

General word association refers to the tendency for one word (the cue) to automatically evoke another (the response) based on mental associations formed through experience (de Deyne, Navarro, & Storms, 2013; Hintz et al., 2017; Nelson, McEvoy, & Schreiber, 2004). These associations reflect how words are linked in the mental lexicon and are typically measured using free association tasks, in which participants respond to a cue word with the first word that comes to mind. De Deyne et al. (2013) introduced continuous association tasks, where participants generate multiple associations per cue, resulting in denser semantic networks and more accurate modelling of lexical relationships. Importantly, general word associations span multiple representational domains, including semantic (e.g., "doctor–nurse"), phonological (e.g., "king–ring"), orthographic (e.g., "table–cable"), and experiential or perceptual associations (e.g., "sun–hot") (Shelton & Martin, 1992). These associations reflect overlapping layers of memory, including statistical co-occurrence, shared conceptual features, and surface-level form similarities, making general word association a rich source for studying prediction in language processing.

Previous VWP studies have revealed significant effects of general word association on predictive eye movements. For example, Yee and Sedivy (2006) demonstrated that upon hearing a spoken word such as "candle," participants quickly shifted their gaze not only to the named item but also to a semantically related item in the visual display (e.g., "lamp"). This suggests that hearing a word automatically activates associated concepts, which influence gaze pattern even before the target is explicitly articulated. Using the same paradigm, Duñabeitia et al. (2009) investigated how abstract and concrete words differ in their associative representations. They found that abstract-associative pairs (e.g., "smell–nose") elicited stronger gaze patterns than concrete pairs (e.g., "crib–baby"). These findings suggest abstract words are more strongly represented through association-based network, whereas concrete words rely more on semantic or conceptual similarity. In a related study, Iordanescu, Graboweck, and Suzuki (2011) found that both spoken words and corresponding characteristic sounds (e.g., "dog" or a barking sound) facilitated visual search for matching images. Participants located associated targets more quickly, even when those targets were rare,

showing that auditory cues can activate related visual representations and guide anticipatory attention.

However, not all studies find a robust effect of general word association. Hintz et al. (2017), using a similar VWP design, presented participants with four-object displays and sentences containing verbs that were either functionally or associatively related to one of the objects. The results showed that only functional associations (e.g., “cut” and “carrot”) reliably predicted anticipatory fixations on the target object, whereas general associative strength did not. This suggests that although general word association can influence visual attention, its predictive utility may be limited when more goal-relevant functional information is available.

In summary, despite mixed findings, general word association has been shown to influence predictive eye movements under certain conditions. The present study followed Hintz et al. (2017) and used verb-mediated sentences as stimuli. Although Hintz et al. observed predictive effects only for functional associations, the current study used general word association as a proxy for prediction-by-production. This decision is grounded in the theoretical distinction between mechanisms: functional associations likely reflect structural relationships mediated by covert production, aligning with the prediction-by-production account (Pickering & Gambi, 2018). In contrast, general word association, measured through a continuous free association task (de Deyne et al., 2013; Hintz et al., 2017) where participants were asked to supply the first three nouns that came to mind, reflects more automatic spreading activation. Given the high cognitive demands of SI, prediction-by-association may operate automatically without specific orientation or constraint. Therefore, the continuous free association task is adopted as a practical approximation of this underlying mechanism.

5.2 The visual-world eye tracking paradigm

The visual-world eye tracking paradigm has been widely used to study real-time language comprehension. In a typical setup, participants listen to spoken language while viewing a visual scene displaying several objects. Their eye movements are recorded and analysed to examine how linguistic input guides visual attention over time. This paradigm is grounded in the tendency for people to rapidly direct their gaze to visual items that are either mentioned in, or related to, the spoken language. In a foundational study, Cooper (1974) demonstrated that participants were more likely to fixate on objects that were either explicitly named or

conceptually related to the concurrent language they were listening to. For example, upon hearing the word "dog", participants were more likely to look at a picture of a dog, and upon hearing "Africa" (the probe word), they tended to look at pictures of a lion, zebra, or snake. These language-driven fixation responses could occur closely time locked to the speech: about 55% of such fixation responses were initiated upon the probe words being pronounced (in some cases even upon the initial syllable, e.g., '*ze*' in zebra), and about 40% of the fixation responses occurred within 200ms after the word termination. Given the high degree of linguistic sensitivity and tight temporal couplings between auditory inputs and fixation responses, Cooper proposed the paradigm as a powerful tool for investigating speech perception, memory, and language processing in real time.

However, partly due to the limited development of eye tracking technique at the time, this paradigm received relatively little attention until Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) published a study in *Science* investigating the joint effects of linguistic and non-linguistic inputs on language processing. They used a similar method to examine how visual context affects the interpretation of temporarily ambiguous instructions, such as "Put the apple on the towel in the box." Participants viewed scenes with either one or two referents. In the one-referent visual context, the visual scene contained an apple on a towel, another towel without an apple, a pencil, and a box; in the two-referent visual context, the pencil was replaced with a second apple on a napkin. Distinctive eye movement patterns emerged across the two visual contexts. When presented with one-referent visual context and ambiguous instructions, participants tended to first look at the only apple, then at the towel without an apple, and finally at the box. When presented with two-referent visual context, participants shifted their attention between the two apples, but upon hearing "the towel", quickly fixated on the apple on the towel, and then they directly fixated on the box—rarely fixating on the empty towel. These patterns suggest that participants interpreted "on the towel" as a destination when only one apple was present, but as a modifier when two apples were available. The findings provide strong evidence for the interaction effect between non-linguistic inputs (i.e., visual contexts) and linguistic inputs on real-time sentence processing, and suggest that eye movements can serve as a window into the mental mechanisms underlying spoken language comprehension.

Building on Cooper's and Tanenhaus et al.'s work, subsequent research has extended VWP to explore the fine-grained dynamics of semantic activation during language comprehension. These studies have demonstrated that a listener's eye movements are influenced not only by directly mentioned words but also by conceptually or associatively related items, reflecting the

richness and immediacy of semantic processing in real time. For example, Huettig and Altmann (2005) presented participants with spoken sentences while displaying visual scenes containing a named object (the target, e.g., "*piano*"), a word from the same conceptual category[1] as the target (the semantic competitor, e.g., "*trumpet*"), or both. Even when the named object was absent, participants fixated more on the semantic competitor, indicating that conceptually related representations were activated and guided attention even in the absence of lexical overlap. Further evidence of this phenomenon comes from Yee and Sedivy (2006), who adopted a target-present design and presented participants with visual displays containing both a target object (e.g., "key") and an associatively related competitor (e.g., "lock"), along with unrelated distractors. They found that participant directed more fixations to the associated competitors, indicating a role of associative links based on frequent co-occurrence in modulating visual attention. Taken together, these studies demonstrate that semantic activation during comprehension is rapid and flexible, extending beyond lexical identity to encompass broader conceptual categories or associations, and that these different types of semantic relationships can dynamically shape the mapping between linguistic and visual referents.

In addition to relatively fast and automatic forms of semantic relatedness, more complex semantic links can also guide visual attention. One example is affordance, which involves top-down analysis of object-function relationships in context. Chambers, Tanenhaus, and Magnuson (2004) examined whether the functional properties of visual objects influence how listeners resolve syntactic ambiguities. Participants heard instructions containing ambiguous phrases such as "*Pour the egg in the bowl over the flour*" while viewing displays that included two eggs, a bowl and flour. The affordance of the eggs was manipulated: in some trials, both eggs were liquid (i.e., pourable), while in others, only one was. When both eggs were pourable, participants initially looked at the bowl (interpreting "*in the bowl*" as a destination). However, when only one egg was pourable, participants quickly fixated on that egg early on (interpreting "*in the bowl*" as a modifier to identify which egg). These patterns suggest that affordance information embedded in the visual display interacts with syntactic processing, illustrating how conceptual knowledge of object function is utilised in real time to resolve linguistic ambiguity.

[1] Altmann (2005) made a clear methodological distinction between semantically related and semantically associated words. Semantically related words were defined as words that share categorical or feature-based relationships (e.g., "dog" and "cat", which both belong to the category "animal" and share some common features). Semantically associated words, on the other hands, were defined based on co-occurrence and associative norms (e.g., "dog" and "bone", which are not from the same semantic category but frequently appear together and are linked through real-world associations)

Beyond conceptual affordance, perceptual features such as colour and shape have also been shown to interact with semantic processing in guiding eye movements. In a series of experiments, Huettig and Altmann (2011) investigated whether stored colour knowledge (e.g., frogs are typically green), perceived surface colour, or categorical relationships determine fixations during spoken word comprehension. Their results showed that surface colour exerted a strong and immediate effect on visual attention, while stored colour knowledge exerted a weak and delayed effect only when conflicting with the surface colour. This suggests that perceptual salience can override conceptual expectations in directing gaze. Similarly, Dahan and Tanenhaus (2005) demonstrated that shaped-based similarity can act as a visual-semantic cue. In their study, participants heard the word "snake" while viewing a display containing a snake, a rope (which shares visual features with a snake), and unrelated items. Results showed increased fixations on the rope, supporting the idea that visual features can trigger semantic activation and influence gaze even in the absence of lexical overlap.

Phonological information also plays a crucial role in guiding visual attention during language comprehension. In Allopenna, Magnuson, and Tanenhaus (1998), participants heard spoken instructions like "pick up the beaker" while viewing a display containing four objects: the target (e.g., "beaker"), an onset competitor (e.g., "beetle"), an offset competitor (e.g., "speaker"), and an unrelated distractor. Eye-tracking data revealed that listeners rapidly fixated on both the onset and the offset competitors shortly after the target word onset. However, onset competitors attracted fixations earlier and more strongly than offset competitors, indicating a fine-grained temporal sensitivity to the phonological unfolding of the spoken word. These findings suggest that listeners use phonological information incrementally and predictively to narrow down lexical candidates during speech comprehension.

Apart from comprehension tasks, similar effects have been observed during language production. Huettig and Hartsuiker (2008) employed an object-naming task where participants named a visually presented target object (e.g., "pizza") while viewing a display that included semantic competitors (e.g., "bread") and visually related competitors (e.g., "coin", similar in shape). They found that participants were more likely to fixate on these related objects, indicating that both semantic and visual properties become activated during lexical selection and utterance planning. In a follow-up study (Huettig & Hartsuiker, 2010), they explored phonological effects during production. In a similar picture-naming task, they found that participants fixated more often on phonological competitors (e.g., "beetle" when naming "beaker") than on unrelated distractors. These findings demonstrate that during language

production, just like in comprehension, speakers activate and integrate multiple types of lexical information, including semantic, visual, and phonological features, which in turn modulate their visual attention.

Collectively, this body of evidence indicates that VWP captures a complex, temporally sensitive interplay of semantic, perceptual, and phonological processes during language comprehension and production. These effects occur without any explicit instruction on visual search (Huettig, Rommers, & Meyer, 2011) and closely time-locked to the concurrent speech, sometimes even within 100 ms (Altmann, 2011). The timing of activation further varies depending on the nature of the relation between the spoken word and the visual referents. For instance, Huettig and McQueen (2007) showed that when the visual scene appeared at the sentence onset, phonologically related items were fixated before semantically or perceptually related items. However, when the visual scene appeared shortly before the target word onset (200 ms), participants first fixated on perceptually related items, followed by semantically related ones, with no significant fixations to phonological competitors. These results suggest a cascading model of lexical access in the speech recognition system, where different levels of representation are activated in parallel but with varying speeds and temporal dynamics, from phonological, to visual, and then semantic. The visual recognition system, on the other hand, follows a reversal of that order. The rapidity of this activation and mapping process is also mediated by word frequency (Dahan, Magnuson, & Tanenhaus, 2001; Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Objects whose names are of higher frequencies are more likely to be fixated at the early stage of the eye movement. These findings affirm VWP as a rich methodological tool for investigating the temporal dynamics of multi-level language processing under naturalistic, time-constrained conditions.

5.2.1 Typical properties of the visual-world paradigm

5.2.1.1 Target-present and target-absent design

The studies reviewed above primarily employed two types of visual stimulus design: the target-present and the target-absent designs. In a target-present design, the visual display includes a target explicitly mentioned in the spoken input, presented alongside competitors that are semantically, phonologically, and/or visually related (e.g., Chambers et al., 2004; Dahan & Tanenhaus, 2005; Yee & Sedivy, 2006). This design is especially suited for examining the time course and competition between the target and its competitors, enabling researchers to track how quickly specific linguistic features are activated or suppressed. In contrast, a target-absent

design distributes the target and the competitor objects across different trials or displays (Huettig & Altmann, 2005, 2011; Rommers, Meyer, Praamstra, & Huettig 2013; Ito, Pickering, et al., 2018). This approach isolates the bottom-up activation triggered by linguistic input alone and is well-suited for investigating whether language can activate related mental representations in the absence of a direct referent.

The present study employed a target-present design for two main reasons. First, the study aimed to investigate the underlying mechanisms of predictive processing during SI, with a specific focus on the prediction-by-production and the prediction-by-association accounts. Presenting the target and the competitors within the same visual display allowed for direct observation of how participants allocated their visual attention, thereby revealing the competition and temporal dynamics between top-down and bottom-up predictive mechanisms. By contrast, presenting the targets and the competitors in separate displays, as in the target-absent design, might still elicit predictive fixations, indicating the presence of prediction, but would not allow researchers to determine whether these fixations were driven by top-down contextual analysis or by bottom-up associations. Second, the target-absent design inherently requires more trials and a larger participant pool to achieve sufficient statistical power (i.e., separate trials for the target, the competitors, and the distractor as a baseline). In contrast, the target-present design allows all these object types to be included within a single trial. Unlike previous SI studies that used independent sentences as stimuli (Amos et al., 2022; Liu et al., 2022) and thus had greater flexibility in the number of trials, the present study involved coherent discourse, making it impractical to embed a large number of trials. Besides, given the risk of interpreter fatigue, an excessively long task may negatively impact both SI performance and predictive processing.

5.2.1.2 Visual stimuli type

A major factor that can influence language-driven eye movement in VWP is the types of visual stimuli. In addition to simple line drawings, pictures, and real objects, complex photographic scenes (Andersson, Ferreira, & Henderson, 2011) and printed words (McQueen & Viebahn, 2007; Huettig & McQueen, 2007) were also used as visual stimuli. Andersson et al. (2011) found that despite complex photography involving multiple objects distributed across the scenes, participants still tended to fixate on target objects in even the most demanding conditions (speech delivered in high speed and objects of high density). McQueen and Viebahn (2007) replicated the design of Allopenna et al. (1998) but used printed words as visual stimuli. They observed that competitors sharing onset or offset phonemes with the target words

attracted more fixations than unrelated distractors. Compared to pictures, printed words can represent more abstract concepts and are particularly advantageous in studies focused on orthographic or phonological processing. When competitor items share phonological or orthographical similarity with the target item, they are more likely to attract fixations in printed word than in picture (Salverda & Tanenhaus, 2010). As shown in Huettig and McQueen (2007), phonological competitor effects were stronger with printed words, and in contrast, semantic or shape-based competitor effects were stronger with pictures. Given that the present study investigates semantic prediction, and that printed words may interfere with the processing of the auditory input, simple line drawings were used to reduce potential confounds from perceptual overlap or linguistic interference.

5.2.1.3 Preview time and speech rate

Another critical factor influencing predictive eye movements is the preview period, defined as the time allowed to visually inspect the scene before auditory stimuli begin. Longer preview times give participants more opportunity to extract and encode visual information and generate expectations about the upcoming sentence (Ferreira, Foucart, & Engelhardt, 2013). In Huettig and Guerra (2019), when participants were granted four seconds of preview, they made predictive fixations to the targets under normal or fast speech rates. In contrast, when granted one second of preview, they could only made predictive fixations under slower speech rates. Additionally, short or no preview periods significantly reduce the influence of visual context on prediction, especially for subtle cues such as phonological similarity (Huettig & McQueen, 2007). To minimise the impact of visual context while ensuring sufficient time to recognise all presented objects under time-pressure, the present study adopted a 1500 ms preview period. This approach aimed to balance ecological plausibility with methodological control, allowing for the detection of predictive processing while limiting confounds from visual salience or scene semantics.

5.2.1.4 Limits and comparisons to the ERP paradigm

A major limitation of this paradigm is its sensitivity to multiple stimulus-related factors, such as competitor's shape (Dahan &Tanenhaus 2005; Huettig & Hartsuiker, 2008), colour (Huettig & Altmann, 2011), word frequency (Dahan et al., 2001) and types (Huettig & McQueen, 2007), all of which can interfere with language-driven eye movements. This sensitivity necessitates careful control of the visual stimuli. By comparison, the ERP paradigm is less demanding in

terms of visual design, as it is primarily concerned with linguistic features and is less influenced by perceptual attributes.

Nonetheless, VWP is more suitable than the ERP paradigm for investigating predictive processing during SI. This is because ERP responses are highly susceptible to muscular activity, particularly near the head area, such as blinking or mouth movements (Luck & Kappenman, 2012). During SI tasks, interpreters must speak continuously, and their articulatory movements inevitably generate artefacts that severely interfere with ERP recordings. One study by Kurz (1995) attempted to examine predictive processing in SI using ERPs. However, interpreters were instructed to produce the output silently in their minds rather than aloud, which significantly compromised the ecological validity of the task. Furthermore, the ERP paradigm typically requires a large number of trials to achieve reliable results, making it impractical for SI studies, as interpreters are prone to fatigue after multiple interpreting sessions.

To sum up, considering its nature of reflecting the information activation in a closely time-locked manner and its advantage over the ERP paradigm, VWP provides an effective way to investigate the predictive processing during SI. It is worth noting that perceptual features (such as colour and shape) and linguistic features (such as word frequency and age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) associated with the visual stimuli needs to be manipulated carefully to avoid interference on the language-driven eye movements. The preview time of the visual scene should also be controlled so that participants at least recognise the perceptual features of all visual objects in the scene (Rommers et al., 2013).

5.2.2 Data analysis

There is no explicit homogeneity in pre-processing eye-tracking data in psycholinguistics studies using VWP (Ito & Knoeferle, 2023). A major reason for this is the variability of visual stimuli and tasks across VWP studies, which significantly affects participants' viewing patterns. For example, while many studies use objects in the form of colour pictures, photographs, or monochromatic line drawings embedded in natural scenes or arranged in arrays, some studies utilise printed words as visual stimuli (McQueen & Viebahn, 2007; Salverda & Tanenhaus, 2010; Ito, 2019). These different forms of visual stimuli can elicit variances in eye movement measures. The typical mean fixation duration for scene perception is around 330 ms, whereas for reading it is 225 ms (Rayner, 1998). Additionally, the arrangement of objects influences viewing behaviours. Identifying an object in a natural scene containing many objects is much

slower compared to identifying an object arranged in an array of four to five objects (Henderson & Ferreira, 2004).

Tasks used in VWP studies range from language comprehension (Snedeker & Trueswell, 2004; Chambers et al., 2004; Huang & Snedeker, 2009; Hvelplund, 2014), word recognition (Canseco-Gonzalez et al., 2010; Weber & Cutler, 2004), to production (Bock, Irwin, Davidson, & Levelt, 2003; Griffin & Bock, 2000; Huettig & Hartsuiker, 2010) and WM tasks (Ito, Corley, et al., 2018). These tasks involve diverse underlying cognitive processing, leading to differences in eye movement patterns and measures. For instance, in studies involving a production task, eye movements can be modulated by speech planning as participants tend to fixate on an object until a corresponding phonological representation is retrieved, with fixation duration positively correlating with the length of the object name (Meyer, Roelofs, & Levelt, 2003; Meyer, van der Meulen, & Brooks, 2004). Due to these reasons, the selection of appropriate measures and the determination of threshold parameters remain flexible and at the discretion of the researchers. Researchers tend to establish their own measurements and rules and often do not report how the eye-tracking data quality was assessed and how eye-tracking data were cleaned.

The specialty of the current study lies in the simultaneous involvement of language comprehension and production tasks, adding to the complexity of underlying cognitive processing. Previous SI studies using VWP (Amos et al., 2022, 2023; Liu et al. 2022) did not report their data cleaning processes. Both studies relied on Eye-Link software to identify fixations and compute relevant metrics. It remains unknown to what extent and how interpreting tasks affect eye movement patterns and measures of viewing objects, particularly in the setting of VWP. Previous interpreting studies using eye-tracking (e.g., Chen, Kruger, & Doherty, 2020) and guidance on processing eye-tracking data in translation studies (e.g., Hvelplund, 2014) are not suitable references for the current study either, as these studies and guidance mostly address text reading, which differs from scene perception in the current study. They often rely on measures such as mean fixation duration, gaze sample to fixation percentage, which, as discussed above, could vary due to the change of tasks. Thus, the eye-tracking data were processed following a more general criterion based on the physiology of eye movements (Komogortsev et al., 2010; Holmqvist et al., 2023).

5.2.2.1 T-test, ANOVA, and linear-mixed effect model

T-test and ANOVA have been used to analyse eye-tracking data in early psycholinguistic studies using VWP (Allopenna et al., 1998; Altmann & Kamide, 1999). To conduct *t*-tests or ANOVAs, the data need to meet several assumptions: independence of observations, normality, and homogeneity of variances. First, the independence of observations requires that observations be independent within each group and between groups. The current study adopted a multi-level sampling scheme, where each participant was assigned ten trials in each of the two conditions (twenty trials in total for each participant), and their eye movements were sampled repeatedly at fixed time intervals (1.6 ms) during these trials. Thus, there were three levels in the sampling hierarchy: level one consists of individual observations in each trial by each participant; level two consists of all trials aggregated by conditions or by participants; and level three are the participants and conditions. These levels formed various clusters of individual observations, within which eye-tracking data are more strongly correlated than those from different clusters. For example, eye-tracking data from the same trial are more strongly correlated than those from different trials. Such within-cluster correlation also applies to other sampling levels and their combinations, e.g., the combination of a participant in a single condition (Barr, 2008).

Second, the normality assumption for *t*-test requires the data to be normally distributed, while for ANOVA, it requires the residuals (differences between the observed and predicted values) to be normally distributed. In the current study, the dependent variable was a categorical variable in the setting of VWP as participants' fixations were categorised by the four regions of objects (i.e., four AOIs) presented in the display. Such a categorical variable follows a multinominal distribution rather than a normal distribution (Barr, 2008). An intuitive approach to address this issue is to transform the categorical variable into a continuous variable by computing fixation proportions. However, the bounded nature of fixation proportions, which always range between 0 and 1, poses challenges for parametric tests like *t*-test and ANOVA. For ANOVAs, the confidence intervals (CIs) around the estimates can extend beyond the interpretable range between 0 and 1 (Jaeger, 2008)). Also, there is an inherent correlation between group means and variances of bounded data; means around 0.5 has smaller variances than those close to 0 and 1 (Jaeger, 2008). In other words, fixation proportions ranging between 0 and 1 naturally violates the assumption of homogeneity of variances, which requires the variances of the data in each group should be approximately equal.

A method to transform bounded data to unbounded data is to perform a logit transformation using a formula $log\big(Proportion/(1\text{-}Proportion)\big)$ (Hintz et al., 2017). However, a logit transformation is undefined when the proportion equals 0, i.e., when participants completely ignore an interested object in a pre-defined time span, or when the proportion equals 1, i.e., when the interested object is attended throughout the pre-defined time span. Both situations are not uncommon in the analysis of eye-tracking data, especially when fixations are aggregated into small time bins (e.g., 50 ms). Instead, an empirical logit transformation (Elog) is recommended to resolve the undefined boundary problem. Essentially, this involves adding a small value to both the numerator and determinator of the logit transformation equation (Barr, 2008). The present study used 0.5 and computed empirical logit transformations of fixation proportions with a formula *log ((Proportion+0.5)/(1-Proportion+0.5))* (Barr, 2008).

Due to the failure to meet the assumptions for *t*-test and ANOVA, linear mixed effect (LME) models (or multi-level regression models) were fitted to compare participants' fixations during the prediction and post-target time windows. The term "mixed effect" emphasises the inclusion of both fixed and random effects. Fixed effects are the effects of interest that will be estimated in the analysis, usually corresponding to the experiment manipulations. Random effects capture the variability at different levels in the hierarchical structure of the data, e.g., by-participant variability, by-trial variability. Therefore, LME models address the issue of non-independence arising from the multi-level sampling scheme by incorporating the diverse clusters of the sampling levels to model non-independence. This approach also eliminates the need for data aggregation by sampling levels, which is typically required by parametric tests such as *t*-tests or ANOVAs. Thus, more information is incorporated into the data analysis, thereby improving statistical power (Barr, 2008). Furthermore, LME models do not require the data to meet the assumptions of normality and homogeneity of variances, thereby are suitable for analysing the eye-tracking data from the current study.

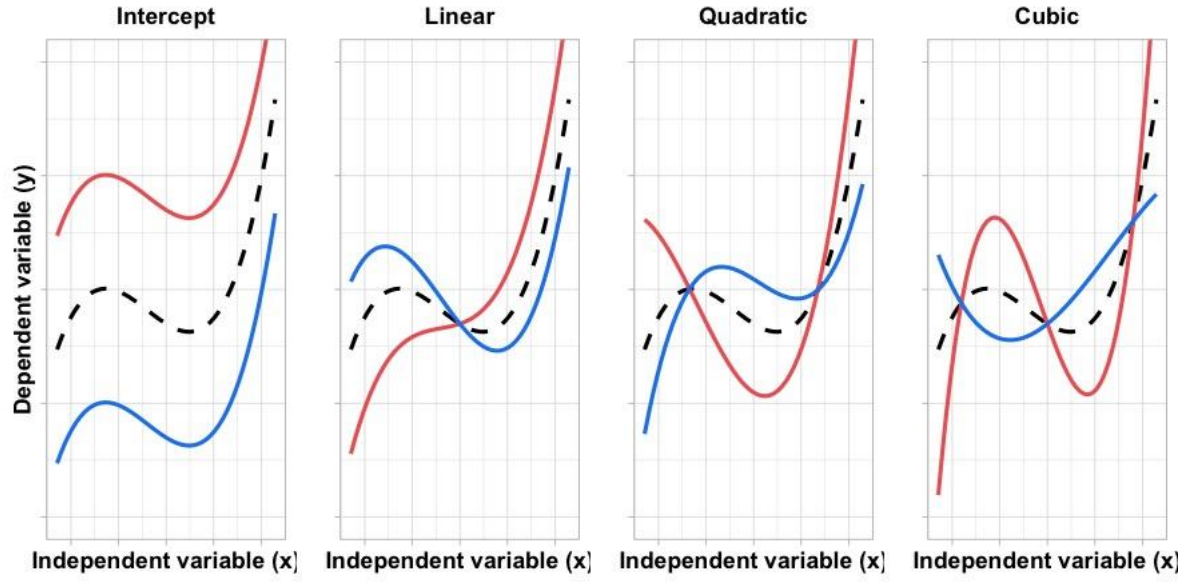
5.2.2.2 Growth-curve analysis

The present study used GCA to examine the non-linear dynamic changes of the condition and the AOI effects over time. The GCA was originally developed for developmental psychology studies using a longitudinal design, where data are collected over several months or even years (Singer & Willett, 2003). In psycholinguistic studies where measures (e.g., eye movements, reaction times) are monitored throughout a given time span, the data structure resembles that of longitudinal studies, except with much smaller measurement intervals (e.g., second,

millisecond). This method was thus adapted to apply in psycholinguistic studies (Magnuson et al., 2007; Mirman, Dixon, & Magnuson, 2008; Kukona et al., 2011; Ito, Pickering, et al., 2018) to model the temporal trajectory of an effect.

The essence of the GCA lies in its ability to capture the curvilinear relationship between measures and time. In both developmental and psycholinguistic studies, changes in measures over time do not always follow a straight linear trajectory with a fixed slope, as assumed by linear models such as ANOVAs and LME models. Instead, they often exhibit a more complex curvilinear form, with (multiple) increase and decrease occurring consecutively as the time unfolds. Therefore, time is incorporated in general linear (mixed effect) models as power polynomials to represent such a curvilinear relationship between dependent variables and time. For instance, a second order (or quadratic) polynomial term ($time^2$) can represents a curve with a single inflection, and a third order (or cubic) polynomial term $time^3$ can represents a curve in a sigmoidal form with two inflections. However, these natural power polynomials are highly collinear, making it difficult to estimate the unique contribution of each polynomial. To address this issue, orthogonal power polynomials – linear transformations of natural power polynomials – are introduced to avoid the collinearity. Terms of different orthogonal polynomial orders can affect the form of the curve independently (Magnuson et al., 2007; Mirman et al., 2008).

Figure 5-1. Effects of manipulating individual polynomial terms on the curvilinear form of a data set using a cubic polynomial.



#Notes: The dashed lines represent the original data set. The red lines represent data sets with increased coefficients of each term. The blue lines represent data sets with decreased coefficients of each term.

To demonstrate the effects of manipulating individual polynomial terms on the shapes of GCA curves, we adapted the Figure 6 from Mirman et al. 2008 and generated a data set using a cubic polynomial. Figure 5-1 presents the curve of the original data set (in dashed line) and how the curve changes when each polynomial term is changed. The intercept term reflects the overall vertical height of the curve; the linear (or first order) term reflects the overall angle (slope) of the curve; the quadratic term reflects the rate of change on two symmetric sides around a central inflection point; and the cubic term reflects the changes in that rate of change (Mirman et al., 2008). Based on the interpretation of each polynomial terms, it is not hard to find that a GCA can capture the overall average, angle or slope, and rate of changes of an effect but cannot estimate its exact time span. Another disadvantage of the GCA is that it does not control for the autocorrelation inherent in eye-movement data. Due to the physiological nature of eye movements, eye gazes must be executed in a chronological and spatial sequence (Rayner, 1998), leading to natural autocorrelation in successive fixations. Stone, Lago, and Schad (2021) found that different lengths of analysis time bins resulted in different levels of correlation between two consecutive bins. When fixations were aggregated by 50 ms, 100 ms, and 150 ms bins, the correlation coefficient between two consecutive bins were over 0.9, 0.8, and 0.75, respectively. The autocorrelation can lead to inflated Type I error[2] rates (Huang & Snedeker, 2020).

5.2.2.3 Cluster-based permutation test

The CPA can determine the time span that presents a significant effect and inherently address the autocorrelation issue of eye-tracking data through its methodological framework. The CPA was initially proposed by Maris and Oostenveld (2007) in the context of psychophysiological studies to analyse electroencephalogram and magnetoencephalogram data and was later applied to psycholinguistic studies using VWP (Barr, Jackson, & Phillips, 2014; Hahn, Snedeker, & Rabagliati, 2015; Amos et al., 2022). The general procedures of a CPA are as below: 1) conduct a series of statistical tests (e.g., *t*-test, ANOVA, LME models) at each time bin or spatial location, and detect clusters of temporally or spatially contiguous points whose test statistics exceed the predefined significance threshold (e.g., *t*-value for *t*-test); 2) calculate the cluster-mass statistics by summing the test statistics of each point in each detected cluster; 3) permute the data a large number of times (at least 1000) following the exchangeability

[2] Type I error is also known as false positive. Specifically, it refers to the incorrect rejection of a null hypothesis that the results are not significant, while the null hypothesis is actually true.

assumption for the CPA[3] (Maris & Oostenveld, 2007); 4) repeat the statistical testing, cluster detection, and cluster-mass statistic calculation steps on each of the permuted data sets and store the largest cluster mass statistic for each permuted data set, and these stored maximum cluster mass statistics form a null distribution of cluster masses; 5) assess the significance of each cluster detected in the original data set through calculating the proportion of cluster-mass statistics from the null distribution that are larger than that from the original data set, and this proportion is called Monte Carlo *p*-value.

This methodological framework ensures the CPA account for the autocorrelation present in eye-tracking data by respecting the temporal structure of the data in the clustering and permutation procedures. Specifically, clusters of significance are formed by grouping temporally adjacent data points, and the permuted data are created by shuffling the data points within time bins. Through this approach, the arbitrariness in choosing the length of analysis time bins can also be avoided (Barr et al., 2014), as time bins are grouped into clusters regardless the length of each time bin, and significance is assessed on the cluster level. The multiple comparison problem, which is common when using parametric tests across multiple time bins, is also solved. The significance of a cluster is estimated not through multiple comparisons between the test statistic of each data point in the cluster and the critical alpha level (either corrected or not), but through a one-time comparison between the cluster-mass statistics of the whole cluster with the null distribution, thereby avoiding the inflated rate of Type I errors and reducing the family-wise error rate[4] (Maris & Oostenveld, 2007; Huang & Snedeker, 2020).

It is worth noting that significant clusters identified by the CPA do not indicate the onset of an effect (Sassenhagen & Draschkow, 2019). In other words, the onset of a significant cluster is not necessarily the exact onset of an effect. This is due to the mathematical methods involved in the formation of clusters and the assessment of cluster significance. While significance on the cluster level is assessed reliably by reducing multiple comparisons to a one-time

[3] The exchangeability assumption for the CPA requires that the labels of observations are exchangeable in the null distribution. For example, in a study that makes within-subject comparison of observations under two conditions, the null hypothesis is that the observation should be invariant across the two conditions, and therefore exchanging the condition labels of observations within each subject should make no differences. It remains unknown if there are any differences between subjects. Therefore, the condition labels should be exchanged within subjects. Following the same logic, in between-subject comparisons, the condition labels should be shuffled between subjects.

[4] Family-wise error rate refers to the probability of making at least one Type I error across a set of multiple hypothesis tests.

comparison, there is no such control in identifying significant individual time points. Furthermore, since the null distribution consists of the maximum cluster-level statistic from each permuted data set, when there is a cluster that extends through a large part of the time course, the CPA may be less sensitive to the smaller cluster (Maris & Oostenveld, 2007). An approach to address this issue is to introduce a vector-valued test statistic. Simply speaking, when an effect is expected to appear in several time spans of different lengths, the largest cluster identified in the original data set should be compared with a null distribution that consists of the maximum cluster-statistic from each permuted dataset, and the second largest cluster with a null distribution that consists of the second largest cluster-statistics, and so on. The current study adopted this approach.

5.3 Ear-voice span

Ear–voice span (EVS) refers to the interval between hearing a SL word or segment and producing its equivalent in a TL (Timarová et al., 2011). It is a key indicator of processing delay and simultaneity in SI, offering insights into the cognitive load and coordination required for comprehension, memory, and production to operate in parallel (Defrancq, 2015; Gile, 2008; Timarová et al., 2011). Early studies measured EVS in terms of word lag (i.e., the number of words by which the interpreter trails behind the speaker) (Gerver, 1969; Treisman, 1965). This approach was gradually replaced by measurements in temporal units, which provide better cross-study comparability. Empirical findings suggest that the average EVS for professional interpreters typically ranges from two to five seconds (Defrancq, 2015; Lee, 2002).

5.3.1 Mediating factors

However, EVS is highly variable (Ono, Tohyama, & Matsubara, 2008) and influenced by multiple factors. One frequently studied factor is speaker delivery rate, although findings remain mixed. Some studies report a positive correlation between delivery rate and EVS (Barik, 1973; Gerver, 1969; Collard & Defrancq, 2019), while others report a negative correlation (Lee, 2002). The speech type, e.g., spontaneous and prepared speech, may also influence EVS. Adamowicz (1989) found that the EVS was shorter in prepared texts than in spontaneous ones, possibly due to greater predictability of the former. Similar results were also observed in Podhajská (2008). Kim (2005) and Ono et al. (2008) investigated the effect of language pairs or interpreting directionality on the EVS. The former observed minimal variation in EVS across

Korean-to-Japanese and Korean-to-Chinese interpreting, while the latter found significant differences between English-to-Japanese and Japanese-to-English interpreting. These inconsistencies may reflect the varying degrees of syntactic alignment across language pairs.

Beyond linguistic and task-related variables, individual preferences and interpreter characteristics have been argued to play a central role in modulating EVS (Anderson, 1994; Timarová et al., 2011). For instance, EVS has been shown to negatively correlate with interpreting experience (Timarová et al., 2014. 2015). Similarly, Doi, Sudoh, and Nakamura (2021) found that interpreters with fifteen years of interpreting experience showed significantly shorter EVS than those found in interpreters with four years and one year of experience. Díaz-Galaz, Padilla, and Bajo, (2015), however, found no significant differences between professional interpreters and interpreting students. Other factors, such as gender (Collard & Defrancq, 2019), age (Timarová et al., 2015), and WM capacity (Timarová et al., 2014) have been explored, but their effects remain inconclusive.

Despite mixing results observed across studies, some consistent patterns have emerged. Longer EVS tends to consistently associate with more complex text segments, longer sentences, and the presence of visual inputs (Barik, 1973; Gerver, 1969; Ruiz Rosendo & Galván, 2019). Similarly, faster speech rates and scripted (as opposed to spontaneous) speech typically elicit longer EVS. Interpreting experience remains one of the more robust modulating factors, with more experienced interpreters generally displaying shorter and more consistent EVS (Doi et al., 2021).

Janikowsky and Chmiel (2025) extended this line of enquiry by analysing EVS in SI between Polish and English using data from the Polish Interpreting Corpus. They employed LME models to examine both text-specific variables (e.g., interpreting direction, speech type, delivery rate, word position, lexical class) and interpreter-specific variables (experience level, memory capacity, individual variation). Their findings revealed that EVS was modulated by interpreting direction, with longer EVS observed in less typical direction. Consistent with previous studies, they found that EVS was influenced by speech rate and speech types, with longer EVS observed in faster speeches and scripted speeches (Barik, 1973; Collard & Defrancq, 2019; Gerver, 1969). Additionally, they found opposite effects of word position within source texts and target texts: EVS increased as the source text progressed but decreased as the target text unfolded. Contrary to earlier studies that reported shorter EVS in more experienced interpreters (Doi et al., 2021; Timarová et al., 2014), Janikowsky and Chmiel

(2025) found that that more experienced interpreters demonstrated longer EVS. This suggests that experienced interpreters may strategically delay production to ensure greater accuracy or coherence. Interestingly, WM capacity did not significantly affect EVS, supporting the view that EVS may reflect more specialised interpreting skills rather than general cognitive capacity.

5.3.2 Methodological considerations

A critical methodological consideration in EVS research lies in the choice of measurement point. Given syntactic mismatch between SL and TL and strategic variation in information restructuring, EVS can vary substantially depending on where and how it is measured. Different studies have used various approaches: Treisman (1965) and Barik (1973) measured EVS at fixed intervals of five seconds; Lee (2002) and Kim (2005) did so at sentence onsets; and Lamberger-Felber (2001) at the beginning of segments that were difficult for interpreters. Timarová et al. (2011) measured the EVS for items in an English-to-Czech interpreting rendition using three different methods adopted from Treisman (1965), Barik (1973), and Lee (2002), and found comparable means and medians. However, they found considerable variation in minimum and maximum values, differing up to two seconds, indicating that the choice of measurement point affects granularity rather than overall trends. Consequently, EVS measurement strategies should be tailored to specific research aims.

In addition to its role as a measure of processing delay, EVS has recently been adopted as an indicator of predictive processing during SI. Hodzik and Williams (2017) measured the EVS for sentence-final verbs in SI from German to English, whereas Chmiel (2021) examined the EVS for sentence-final nouns in a word-translation task between English and Polish. In both studies, EVS was negatively correlated with contextual constraints: interpreters exhibited shorter EVS in highly constraining contexts, suggesting that prediction facilitated earlier output initiation. Adopting this framework, the present study used EVS to investigate the relationship between predictive processing and SI performance, as well as the influence of trial-specific features (e.g., word frequency, sentence position). The EVS was measured at both the lexical level (i.e., for the CV and the TW) and at the sentence level (sentence onset and offset). The present study explored how trial-specific features affect EVS, how EVS interacts with predictive processing, and whether these relationships differ between professional interpreters and interpreting students.

5.4 Interpreting quality assessment

A variety of scoring methods have been employed in interpreting assessment, each reflecting different priorities in terms of validity, reliability, diagnostic usefulness, and practicality (Han, 2022). Traditional approaches include atomistic (error-analytic) scoring, which involves identifying and counting specific errors such as omissions or mistranslations. While this method provides highly detailed feedback, it is labour-intensive and often criticised for its low inter-rater reliability. Another common approach, checklist-based scoring, uses a pre-defined inventory of quality criteria to rate aspects such as delivery and accuracy. Although the assessment criteria are often conceptually related, they are presented individually, which can increase cognitive load for raters (Fowler, 2007; Perez, Hartley, Mason, & Peng, 2003). In response to these limitations, rubric-referenced scoring methods have gained popularity (Lee, 2008; Liu, 2013; Setton & Dawrant, 2016; Tiselius, 2009). Multi-dimensional rubrics with detailed performance descriptors strikes a useful balance between diagnostic depth and reliability, making them increasingly common in formal assessment contexts (Han, 2022).

In addition to these human-rater approaches, more recent methods include comparative judgement and automated scoring. Comparative judgement asks raters to compare performance pairs rather than assigning absolute scores, which often improves reliability and reduces rater bias, although it provides limited diagnostic feedback (Han & Lu, 2021). Automated scoring systems, by contrast, evaluate interpreting using quantifiable features such as fluency, lexical diversity, or alignment with reference texts. These systems are efficient and scalable, but they continue to face challenges in capturing the nuanced aspects of human performance and ensuring fairness. As Han (2022) suggests, the future of interpreting assessment may lie in hybrid approaches that integrate the strengths of both human judgement and automated systems, depending on the goals and context of the evaluation.

Recognising that a mixed-methods approach may yield more reliable and informative results than either system alone (Amini, 2018; Waddington, 2001), the present study adopted a multi-method scoring, incorporating both rubric-referenced assessment and item-based analysis.

5.4.1 Rubric-referenced assessment

The rubric-referenced assessment employed a rating scale adapted from Han (2016) and Chen, Yang, and Han (2022) (see Figure 5-2). The rating scale comprised three eight-point subscales: information completeness (the extent to which source-text propositional content is rendered),

fluency of delivery (the extent to which disfluencies, such as un/filled pauses, long silence, and fillers, are present in target-language rendition), and target-language quality (the extent to which expressions are natural to a native speaker of the target language). Each subscale was designed with flexibility, allowing scores to be reduced to a four-band scale by collapsing two adjacent points into a single score band.

To evaluate the reliability of scoring, Han (2016) applied Generalisation Theory to analyse the impact of various error sources (e.g., raters, tasks, and dimensions) on overall score dependability. He found that information completeness yielded the most reliable scores. Decision studies further indicated that increasing the number of tasks most effectively improved score reliability for information completeness, while increasing the number of raters was more beneficial for enhancing fluency and target-language quality dimensions. Additionally, composite scores remained stable across different weighting schemes, although information completeness contributed most to total score variance. These findings suggest that tailoring the number of tasks and raters to specific scoring dimensions can significantly improve the reliability of interpreting assessments.

For the current study, two professional interpreters, each with a minimum of three years' interpreting experience, were recruited to perform the rubric-referenced assessment. A detailed account of their rating procedures and the evaluation of inter-rater reliability is provided in Section 7.2.1 of Chapter 7.

Figure 5-2. The rubric-referenced rating scale adopted from Han (2016) and Chen et al. (2022).

Band/ Scoring criteria	**Information Completeness** (InfoCom)	**Fluency of Delivery** (FluDel)	**Target Language Quality** (TLQual)
Score			
Band IV (Score range: 7-8)	A substantial amount of original messages delivered (i.e., > 80%), with a few number of deviations, inaccuracies, and minor/major omissions.	Delivery on the whole fluent, containing a few disfluencies such as (un)filled pauses, long silence, fillers and/or excessive repairs.	Target language idiomatic and on the whole correct, with only a few instances of unnatural expressions and grammatical errors.
Band III (Score range: 5-6)	Majority of original messages delivered (i.e., 60-70%), with a small number of deviations, inaccuracies, and minor/major omissions.	Delivery on the whole generally fluent, containing a small number of disfluencies.	Target language generally idiomatic and on the whole mostly correct, with a small amount of instances of unnatural expressions and grammatical errors.
Band II (Score range: 3-4)	About half of original messages delivered (i.e., 40-50%), with many instances of deviations, inaccuracies, and minor/major omissions.	Delivery rather fluent. Acceptable, but with regular disfluencies.	Target language to a certain degree both idiomatic and correct. Acceptable, but contains many instances of unnatural expressions and grammatical errors.
Band I (Score range: 1-2)	A small portion of original messages delivered (i.e., < 30%), with frequent occurrences of deviations, inaccuracies, and minor/major omissions, to such a degree that listeners may doubt the integrity of renditions.	Delivery lacks fluency. It is frequently hampered by disfluencies, to such a degree that they may impede comprehension.	Target language stilted, lacking in idiomaticity, and containing frequent grammatical errors, to such a degree that it may impede comprehension.

5.4.2 Item-based assessment

The item-based assessment in this study was adapted from the semi-automatic, graded semantic scoring method for evaluating SI quality proposed by Zhang (2016). This method is based on a frame-based analysis, which evaluates semantic equivalence between the source and target texts using FrameNet. In this framework, lexical items (typically verbs or nouns) evoke semantic frames, i.e., structured conceptual scenarios. Each frame is composed of multiple frame elements (FEs), which represent the participants, roles, or properties relevant to that scenario. In the analysis, both the original utterance and the interpreter's rendition are manually annotated for frames and their associated FEs. The evaluation focuses not on direct lexical

matching, but on whether the interpreter successfully conveys the key FEs of each frame. Scoring is based on the extent of FE overlap between the source and target: full matches receive full credit, partial matches receive reduced scores, and missing or incorrect FEs are penalised. This method allows for a graded assessment of meaning transfer that captures both accuracy and completeness of interpretation.

In the present study, a FrameNet was constructed for each experimental sentence by manually annotating the FEs in both the source text and the participants' interpreting renditions (see Section 7.3.2 of Chapter 7). However, unlike Zhang (2016), who employed a graded rubric assigning partial credit based on how accurately FEs were matched between interpreter's rendition and the source, the present study adopted an all-or-none scoring strategy. This decision was made for the reliability of assessment and simplicity of data interpretation. Each participant completed twenty experimental trials, corresponding to twenty FrameNets. Because the graded method requires raters to apply subjective judgements (e.g., evaluating whether an FE was appropriately interpreted), repeating this judgement within each participant could lead to inconsistency and rater bias across trials and reduce the reliability of the assessment. By contrast, the all-or-none method required all FEs to be accurately conveyed, except those classified as modifiers. Otherwise, the interpretation was marked as ineffective or null. This binary scoring approach enabled straightforward calculation of by-participant metrics and facilitated direct comparison with results from the rubric-referenced assessment, where each participant received a single score per dimension.

5.5 Thematic analysis

Thematic Analysis (TA) is a flexible and widely used method for identifying, analysing, and reporting patterns or themes within qualitative data (Braun & Clarke, 2006). It offers a structured yet adaptable approach to organising and interpreting rich, detailed data, making it particularly suitable for exploring participants' experiences, perceptions, and meaning-making processes. Compared to other qualitative approaches, such as Discourse Analysis (Gee, 2014) or Interpretative Phenomenological Analysis (Smith & Osborn, 2003), TA is less constrained by specific epistemological frameworks or methodological assumptions, making it more accessible to researchers across disciplines. Instead, TA offers a balanced approach that supports both data-driven (inductive) and theory-informed (deductive) analysis, particularly

suited to studies that seek to uncover implicit cognitive processes through retrospective accounts, where themes may reflect both conscious reflections and latent patterns.

The present study used TA for its capacity to bridge data and theory: inductive coding captured interpreters' self-reported challenges and strategies, while deductive analysis contextualized these findings within predictive processing models. This hybrid approach ensured themes were grounded in participants' voices while addressing the study's theoretical aims. The analytic process follows Braun and Clarke's (2006) six-phase framework: (1) familiarization with data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining/naming themes, and (6) producing the report. TA's rigor was further enhanced through reflexivity and iterative coding. Two coding cycles were conducted at a six-month interval to assess intra-coder consistency, minimizing bias and ensuring thematic stability (Nowell, Norris, White, & Moules, 2017).

Chapter 6 Experimental setup and procedure

6.1 Pre-tests

Several pre-tests were conducted to prepare the experimental materials, including the source text and visual stimuli. A free association test was conducted to measure association strength between CVs and TWs and to select competitor words. A visual similarity test and a naming test were conducted to ensure that created visual objects represent their corresponding target, competitor, and distractor words and that visual objects representing competitor words are unrelated to TWs. A cloze test was conducted to measure the predictability of TWs.

6.1.1 The source text preparation

The ST was adapted from a travel story depicting a journey to Norway (see Appendix 1). Complicated and less frequent words (e.g., mesmerised) were eliminated or replaced by relatively simpler and more frequent synonyms. Phrases and sentences that were deemed to be difficult to be translated into Chinese were either removed or paraphrased. The simplification of the ST was intended to reduce the cognitive load devoted to comprehension and interpreting tasks and increase the possibility of prediction embodied by predictive eye movements. The ST was divided into four paragraphs, comparable in length and readability (see Table 6-1). The adapted ST was also evaluated for naturalness by two native English speakers.

Table 6-1. Profiles and readability of the ST

	Paragraph 1	**Paragraph 2**	**Paragraph 3**	**Paragraph 4**
Number of words	314	333	305	334
Number of sentences	26	24	24	23
Number of long words	21	18	22	18
Percent of longwords	6.69%	5.41%	7.21%	5.39%
Average words per sentence	12.08	13.88	12.71	14.52
Average syllables per word	1.39	1.34	1.38	1.36
Flesch Kincaid Reading Ease	77.1	79.7	76.9	77.4
SMOG Index	5.5	5.3	5.8	5.4

#Notes: Long words are defined as words with more than three syllables. Flesch Kincaid Reading Ease scores is based on a scale from 0 to 100, and a higher score means easier to read. A value between 60 and 80 should be easy for a 12- to 15- year-old (native speakers) to understand. SMOG Index estimates the years of education a person needs to comprehend writing.

Each stimulus paragraph included five verb-mediated experimental sentences (e.g., *In the station store, commuters are eating/buying freshly made bread.*), three in the predictable

condition with predictive verbs (e.g., *eating*) and two in the unpredictable condition with unpredictive verbs (e.g., *buying*). All twenty experimental sentences had the same subject-verb-object (SVO) structure, with adjective phrases (e.g., *freshly made*) separating verbs and objects. These adjective phrases are three-to-six syllables long, thereby creating a time window allowing for potential predictive eye movements upon hearing the CVs. At least one filter sentence was inserted between two experimental sentences. Two counterbalanced versions of the ST were created, with the predictive verbs in one version replaced with unpredictive counterparts in the other version, and vice versa.

6.1.2 Word length and frequency

Spoken word recognition has been shown to be jointly mediated by word length, frequency, and age of acquisition (Grosjean, 1980; Hudson & Bergman, 1985; Metsala, 1997; Garlock, Walley, & Metsala, 2001; Magnuson et al., 2007). In the current study, the CV were controlled for word length and lexical frequency to ensure comparable difficulties in recognising and processing spoken CVs across conditions, thereby leaving equivalent amounts of time for participants to make predictive eye movements. Instead of controlling word length by the number of letters (Hintz et al., 2017), the present study measured word lengths by the number of syllables since it provides a more direct indicator of the word duration in the audio recording (Marslen-Wilson, 1984). The mean numbers of syllables of predictive and unpredictive CVs were 1.85 (SD = 0.86) and 1.85 (SD = 0.81), respectively, with no significant difference between the two groups ($p > .1$). It can be assumed that the durations of CVs were almost the same across the two conditions, giving the same amount of time for participants to process and react to these verbs.

Word frequencies were measured through use frequency per million words retrieved from the British National Corpus. Corpus-based word frequencies were log-transformed. The mean log-transformed word frequency for predictive CVs was 4.37 (SD = 0.58), significantly lower than the mean for unpredictive CVs (Mean = 4.83, SD = 0.58) ($t(38) = –2.536$, $p = .015$). Consistent with the pattern observed in Hintz et al. (2017), this might be because predictive CVs tended to describe more specific actions or movements, while unpredictive CVs had fewer restrictions and thus were more often used. Yet, this difference does not undermine the conclusion that prediction is facilitated by predictive verbs during interpreting if predictive eye movements are observed in the predictable condition but not in the unpredictable condition. On the contrary, if participants succeed in making predictive eye movements upon hearing predictive CVs

which were of lower frequencies but fail to do so upon hearing unpredictive verbs which were of higher frequencies, it implies that the easier and quicker integration of unpredictive verbs into the context does not necessarily facilitate prediction for the target, and that top-down assessment of the context may play a greater role in prediction.

6.1.3 Cloze test

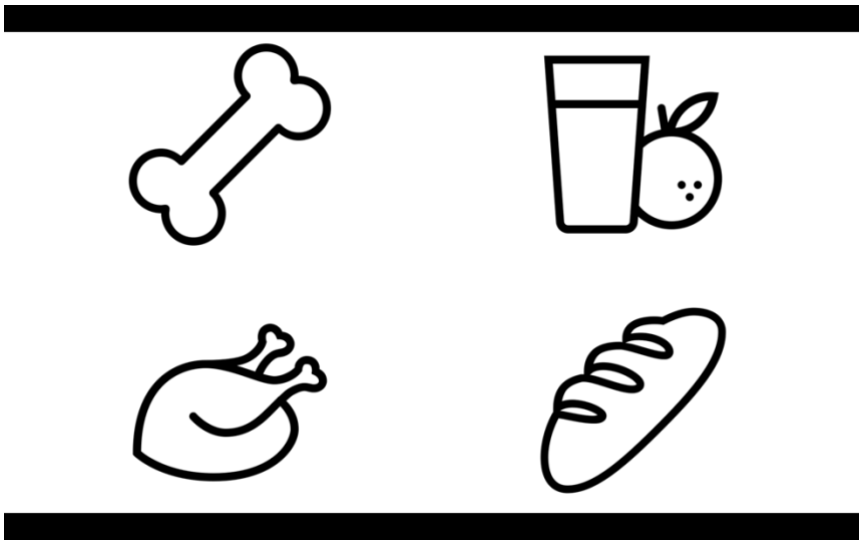
To measure the predictability of the TWs, a cloze test was administered to 48 students (Mean age = 23.73 years, *SD* = 1.12) enrolled in a master's programme in Translation and Interpreting (MTI), all of whom were native speakers of Mandarin Chinese and second-language speakers of English. They were randomly assigned to one of the two versions of the ST, in which the TWs from the experimental sentences were omitted (e.g., *In the station store, commuters are eating/buying freshly made* ___.). For each sentence, four object options were provided. Participants were instructed to read the ST sequentially and complete sentences by selecting one of the four matched objects that was most likely to appear in the context and supplying a label for the chosen object. They were required to fill in the blanks before reading the following sentences, and the answers could not be changed once they started reading the following sentences. This instruction was intended to simulate the interpreting scenario where prediction for the forthcoming content can only be based on the previous context.

The predictability of a TW was measured by dividing the number of participants who provided the correct answer (i.e., the TW or its synonym) by the total number of participants who were assigned the same version of the ST. After excluding invalid answers (0.5%), the average predictability score for TWs in the predictable condition was .88 (*SD* = .11), significantly higher than in the unpredictable condition (Mean = .63, *SD* = .24; $t(27) = 2.84$, $p = .008$). The predictability of the TWs in the unpredictable condition was relatively higher compared with the results in previous studies adopting the same paradigm (Hintz et al., 2017; Ito, Corley, et al., 2018; Liu et al., 2022). This might be because these studies all used single sentences as stimuli, while the stimuli in the current research were paragraphs comprised of multiple sentences, which might create stronger contextual constraints even in the unpredictable condition. The average predictability of the plausible competitors in the unpredictable condition was .32 (*SD* = .23).

6.2 Visual stimuli preparation

Each visual display consisted of four objects representing the target, the distractor, and the two competitor words, matched to their corresponding experimental sentence (see example in Figure 6-1). Objects were depicted with monochrome line drawings, and their positions were randomised using a Latin-square design within a (virtual) 2×2 grid. The four images in each display were unrelated perceptually or linguistically. A series of validation procedures were taken to prepare and evaluate the visual stimuli.

Figure 6-1. Example display for an experimental sentence: *In the station store, commuters are eating/buying freshly made bread*.



#Notes: The target, *bread,* on the lower right; the plausible competitor, *juice,* on the upper right; the implausible competitor, *turkey,* on the lower left, and the distractor, *bone,* on the upper left.

6.2.1 Free association test

The free association test was adapted from the free verb-noun association test by Hintz et al. (2017). An independent group of MTI students (N = 48; Mean age = 22.83, *SD* = 1.26) were recruited. None of them participated in any of the other pre-tests or the formal experiment. These participants were assumed to share similar general and language-specific knowledge with the student participants in the formal experiments, as they were mostly of the same age group and were all first- or second-year MTI students whose L1 is Mandarin Chinese and L2 is English. Their free associations of given critical words were thus postulated to be similar.

Twenty predictive and twenty unpredictive CVs, were listed in eight different randomised orders. The participants were randomly assigned one of the eight lists and were instructed to read the critical words one by one and write down the first three nouns that came into their

minds. This instruction restricted only the word class of associated words (i.e., noun) but not any other linguistic aspects. In other words, participants were allowed to provide words that were semantically, contextually, phonologically, and/or orthographically associated with the CVs. This was to simulate the simultaneous interpreting procedure in the formal experiment, during which interpreters, without any external restriction or interference, could think of any associated words upon hearing the critical words.

The results of one participant were eventually excluded as s/he failed to fulfil the task due to a misunderstanding of the instruction. The association strength was computed by dividing the count of occurrences of a particular noun by the total number of the participants. There were 105 missing values (1.3% of the data) where participants failed to provide any answers, or where the provided answers were not nouns. Table 6-2 illustrates the average association strength between TWs and CVs in different conditions. The average association strength of the TWs with the predictive CVs (Mean = .14) is slightly higher than those with unpredictive CVs (Mean = .07), but the difference is not significant ($t(34) = 1.652$; $p = .108$). Competitor and distractor words were selected based on the association test results. The plausible competitors exhibited similar or higher association strengths with corresponding unpredictive verbs than the TWs and unpredictive verbs did. The implausible competitors had similar or higher association strengths with predictive verbs than the TWs and predictive verbs. The distractors were incompatible in the contexts and were of almost no association with either predictive or unpredictive CVs. However, in a single visual display, the distractor was in the same semantic category (van Overschelde, Rawson, & Dunlosky, 2004) with one of the other three objects presented simultaneously. This design was intended to avoid eye movements directed to distractor objects due to their unique semantic categorical features among all the objects in the same visual display.

Table 6-2. Association strengths between the competitor and the distractor words and the CVs

	Plausible competitor		**Implausible competitor**		**Distractor**	
	Predictive	**Unpredictive**	**Predictive**	**Unpredictive**	**Predictive**	**Unpredictive**
Max	0.21	0.94	0.72	0.38	0	0.02
Min	0	0	0	0	0	0
Mean	0.02	0.16	0.33	0.04	0	0.00
SD	0.06	0.29	0.25	0.10	0	0.00

6.2.2 Word length, frequency, and age of acquisition

The mean log-transformed word frequency for the TWs was 4.12 (*SD* = 0.47), significantly lower than that of the plausible competitor words (Mean = 4.53; *SD* = 0.39; $t(36) = -2.937$; $p = .006$) and the implausible competitor words (Mean = 4.82; *SD* = 0.56; $t(37) = -4.192$; $p < .001$), but not significantly different from that of distractor words (Mean = 4.33; *SD* = 0.47; $t(37) = -1.383$; $p = .175$). Again, such a difference does not undermine the conclusion but is based on a slightly different rationale. If participants made predictive eye movements to the target objects but not to the two competitor objects, it to some extent rebuts the integration account, which suggests that predictive eye movements to the targets are made because they are more easily integrated into the context but implies that predictions are made based on the contextual information.

A subjective frequency rating adapted from Chen and Dong (2019) was employed as a complement measure since the validity of the corpus frequency has long been questioned (Gernsbacher, 1984; Kuperman & van Dyke, 2013), due to its sensitivity to sampling biases and limited reliability for low-frequency words (Brysbaert et al., 2011; Brysbaert & New, 2009). Furthermore, the predictive power of corpus frequency norms on L2 lexical processing is even more limited as corpus frequency norms were mostly derived from corpora built on L1 materials and may not truly reflect the input received by L2 speakers (Chen & Dong, 2019). Previous studies have shown a strong correlation between the two measures of word frequencies (Brysbaert & Cortese, 2011; Ernestus & Cutler, 2015), even including subjective frequencies rated by L2 speakers (Chen & Dong, 2019). The current study, therefore, used subjective frequency ratings as a valid measure of word frequencies among L2 speakers. Target, competitor, and distractor words were evenly distributed across two lists. Forty-eight MTI students (Mean age = 23.12; *SD* = 1.33) were randomly assigned one of the two lists and instructed to assess word frequencies on a 7-point Likert scale[5]. Twenty-four pseudowords were inserted into the two lists to examine the validity of participants' results: the results were deemed invalid when more than six (25%) pseudo words were rated higher than 3. The data of one participant was excluded based on this criterion. The results show that the subjective ratings were generally similar across the four types of words. There were some competitor and distractor words whose subjective frequency ratings were exceptionally low relative to other

[5] 1 = no exposure to the word at all; 2 = very infrequent exposure to the word; 3 = infrequent exposure to the word; 4 = moderate exposure to the word; 5 = frequent exposure to the word; 6 = very frequent exposure to the word; 7 = extremely frequent exposure to the word.

words appearing in the same visual scenes. These words were replaced by words of higher frequencies and lower AoA to ensure none of the words was new to L2 speakers. After replacing low-frequency words, target, competitor, and distractor words were compared in terms of their AoA (Kuperman et al., 2012). The results are presented in Table 6-3.

Table 6-3. Subjective frequency ratings for target, competitor, and distractor words

	Mean	***SD***	**ANOVA**
Target	5.42	1.01	
Plausible competitor	5.38	1.24	$F(3, 76) = 1.55$
Implausible competitor	5.67	0.84	$p = .21$
Distractor	4.98	0.98	

6.2.3 Visual similarity test and naming test

A visual similarity test, adapted from Hintz et al. (2017), and a naming test were conducted to evaluate (1) the extent to which the typical visual representations of the target, the competitor, and the distractor words resembled the intended referents, and (2) whether these visual objects could invoke the intended concepts of the words. Notably, the cognitive process underpinning these two tests occurred in reversed order. The visual similarity test required participants to recognise the printed words first, generate mental representations of semantic concepts of these words, and then compare these mental representations with the physical shapes of displayed objects. Conversely, the naming test required participants to recognise the physical shapes of displayed objects first and then generate lexical names to denote the semantic concepts invoked by the displayed objects. The generated names were then compared with the printed words (by the researcher).

Eleven participants (Mean age = 25.64, *SD* = 4.15, L1 = Mandarin Chinese, L2 = English) who were students studying at British universities were paid to participate in the visual similarity test. Thirty-six sets of printed words (including target, competitor, and distractor words) were created and matched with visual objects that were intended to depict these words. Within each set, the order of visual objects and printed words was randomised. Participants were instructed to rate to what extent each printed word was represented by a corresponding visual object on a rating scale from 0 (No similarity) to 10 (Identical). The average visual similarity rating across all objects was 9.65 (*SD* = 0.42), indicating the created visual objects accurately depicted their corresponding printed words.

Another group of thirteen Chinese students from British universities (Mean age = 25.70, *SD* = 3.35, L1 = Mandarin Chinese, L2 = English) was recruited to perform the naming test. 144 visual objects were presented in a randomised order, and participants were instructed to provide a name for each object. The naming consistency was calculated by the number of participants who produced the intended name for each object by the total number of participants. A set of four objects appearing in the same visual display were eventually removed due to relatively low naming consistency. Objects that were mistakenly named by more than six participants were replaced. The overall mean naming consistency was .96 (*SD* = .08). It is worth noting that the semantic concepts invoked by the given objects could vary across cultures. However, the participants in the naming test shared the same cultural background with those in the formal experiment. Therefore, it is reasonable to assume that the visual stimuli would invoke similar semantic concepts in both groups.

6.3 Formal experiment

6.3.1 Participants

Twenty-two professional interpreters and forty-four MTI students participated in the formal experiment. All participants were based in mainland China at the time of testing. They were native speakers of Mandarin Chinese who had acquired English (L2) through formal education and received at least one year of professional interpreting training. Table 6-4 presents a detailed profile of the participants' language background. The professional group were significantly older than the student group and had a later age of L2 acquisition and greater exposure to the L2. The student group reported no professional interpreting experience at the time of testing. There were no significant differences between the two groups in their L2 proficiencies. All participants reported normal or corrected-to-normal vision and no history of language disorders or hearing impairments. Prior to participation, they were informed about the purpose and procedures of the experiment and provided written informed consent. Ethical approval for the study was obtained from the Ethics Committee of Durham University.

Table 6-4. Background information of the two groups and t-test comparison results

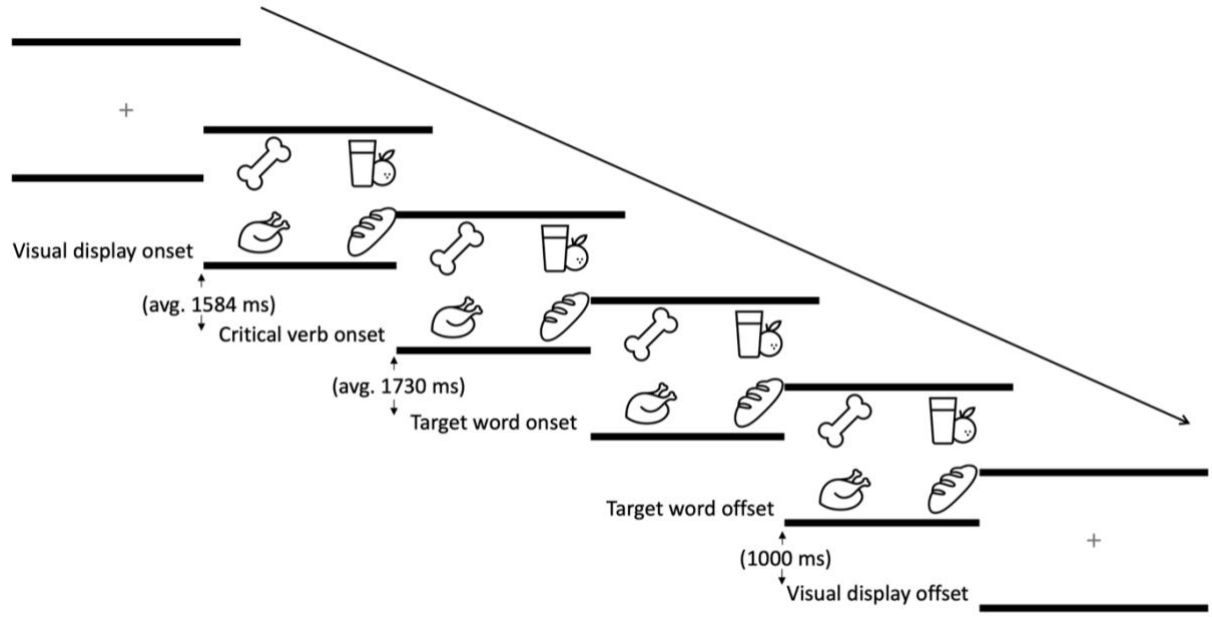
	Professional (N = 22)	**Student (N = 44)**	**Comparison**
Age (yrs)	32.27 (*SD* = 5.15)	23.71 (*SD* = 2.56)	***
Age (yrs) of acquisition of L2	10.00 (*SD* = 1.66)	7.69 (*SD* = 2.59)	***
Time (yrs) of exposure to L2	22.26 (*SD* = 4.25)	15.97 (*SD* = 2.88)	***
Time (yrs) in profession	7.77 (*SD* = 4.30)	-	-
TEM-8	77.88	78.74	
Self-rated English proficiency			
Listening	6.00 (*SD* = 0.76)	5.69 (*SD* = 0.87)	
Reading	6.23 (*SD* = 0.75)	6.00 (*SD* = 0.73)	
Speaking	5.86 (*SD* = 0.84)	5.54 (*SD* = 0.78)	
Writing	5.50 (*SD* = 0.80)	5.46 (*SD* = 0.78)	

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.TEM-8 stands for Test for English Major—Grade Eight, an examination for university students majoring in English. The self-rated English proficiency was assessed on a seven-point Likert-type scale (from 1 = "very low" to 7 = "very high").

6.3.2 Stimuli

The auditory stimuli were recorded by a male native American English speaker who spoke in the standard American accent. As the standard American accent is more generally accepted in China, the ST was read in this accent to reduce listening difficulty. The speaker read the ST at a consistent rate of approximately 2.5 syllables per second, or 110 words per minute. A five-second pause was inserted after every two-to-four sentences. The relatively slow speed of delivery and the five-second intervals between sentence groups were intended to reduce interpreting task difficulty and create optimal conditions for predictive eye movements. Two versions of auditory stimuli were created. Each visual display appeared on the screen at least 1.5 s prior to the CV onset, allowing participants sufficient time to preview and identify all displayed objects (Huettig et al., 2011), and disappeared 1s after the TW offset (see Figure 6-2). Between each visual stimulus, a small fixation cross was presented at the screen centre to guarantee participants' predictive eye movements always start from the central point of the screen, which is equidistant from each of the four objects.

Figure 6-2. Timeline of a visual display for a single experiment sentence.



#Notes: The CV onset was at about 1600 ms (Mean = 1584 ms, *SD* = 231 ms) after the visual display onset, and the TW onset was at about 1750 ms (Mean = 1730 ms, *SD* = 337 ms) after the CV onset. The visual display disappeared 1000 ms after the TW offset.

Due to an unintentional programming oversight, the number of trials per condition was not perfectly balanced across all participants. Specifically, half of the participants (Group 1) were presented with twelve trials in the predictable condition and eight in the unpredictable condition, while the remaining participants (Group 2) received the inverse ratio of eight predictable and twelve unpredictable trials. While this imbalance represents a technical limitation, the potential influence on the findings was mitigated through two complementary approaches. First, the primary analyses employed LME models and CPA, both of which are inherently robust to unequal trial distributions and help control for variance across participants and conditions (Barr et al., 2013; Maris & Oostenveld, 2007). Second, a robustness check was conducted using subsampled datasets that reduced the degree of imbalance while preserving maximal statistical power (Sherman, 1998). For Group 1 participants, ten predictable trials were randomly selected from the original twelve while all eight unpredictable trials were retained. Conversely, for Group 2 participants, all eight predictable and twelve unpredictable trials, all eight predictable trials were retained, and ten unpredictable trials were randomly selected from twelve. Although this approach did not create perfectly balanced datasets, it substantially reduced the degree of imbalance (from a 12:8 to a 10:8 ratio) while avoiding excessive data loss that could compromise analytical sensitivity. The main findings remained

largely consistent across the original and the subsampled datasets, providing converging evidence for the robustness of the observed effects. In cases where discrepancies emerged, such as shifts in significance level or direction of effect, these were clearly noted in the results section and further addressed in the discussion. Full details of the robustness check and comparative results are provided in Appendix X.

6.3.3 Apparatus

The visual stimuli were presented on a 23.8-inch EIZO FlexScan EV2451 monitor at a resolution of 1920 × 1080 pixels (refresher rates 55–76 Hz). The ST recordings were played though a Sennheiser PC8USB headset. Eye movements were registered using a Tobii Pro Spectrum eye tracker with a sampling rate of 600 Hz. The viewing distance was around 65 cm.

6.3.4 Procedure

One day before the formal experiment, participants received a glossary of potentially challenging words and phrases, which they were instructed to review in advance. None of the words and phrases included in the glossary appeared in the experimental sentences. Participants were instructed to read the glossary at least once before attending the experiment. The formal experiment started with a brief orientation session outlining the procedure, and the experiment consisted of a warm-up session followed by four study sessions. The warm-up paragraph, approximately one minute in duration, provided experiment instructions and introduced the narrative context of the travel story to be interpreted. Some visual displays were presented during the warm-up session to simulate the formal experiment. The warm-up material did not contain any of the experimental sentences. Participants were informed that they could withdraw from the experiment any time they did not want to continue.

Each session started with a five-point calibration of the eye tracker. Instructions regarding calibration were provided. Participants were asked to keep their heads still following calibration and throughout each session, and to focus on the central fixation cross whenever no visual stimulus was presented. Participants were informed that objects presented in the visual displays might or might not be relevant to the spoken sentences, and were encouraged, but not required, to attend to the displays. Their visual search behaviours therefore remained autonomous. Participants were allowed to take notes during interpreting tasks as long as they kept looking at the screen. Participants were assigned to one of two versions of the auditory stimuli.

A retrospective interview was conducted after each interpreting task. Participants were asked to assess the interpreting task difficulty on a scale from 1 to 7 (1 = extremely easy, 7 = extremely difficult) and the speech rate on a scale from 1 to 5 (1 = extremely slow, 5 = extremely fast). They were asked to answer three questions: Q1. Please recall and evaluate the interpreting process, including but not limited to difficulties and problems encountered and strategies or skills used to resolve these problems; Q2. Did the visual display influence the interpreting performance? Q3. Did you try to anticipate upcoming contents? Extended questions were asked based on participants' responses. They were also asked to recall the visual objects they saw during interpreting tasks and whether these objects were seen before they heard the actual words representing these objects. A five-minute break followed each interview, when participants were permitted to review the glossary. The total duration of the experiment ranged from 45 to 60 minutes.

Chapter 7 Data analysis and results

7.1 Eye-tracking data

7.1.1 Data preparation and cleaning

The raw eye-tracking data were processed using Tobii Pro Lab 1.162. Fixations were identified through the Tobii I-VT filter. Adjacent fixations were merged based on two threshold parameters: a maximum time interval of 75 ms and a maximum visual angle of 0.5 degrees, between separate fixations to be merged. The time interval threshold was grounded in the physiological characteristics of eye movements that computing a saccade requires a certain latency, and thus making a saccade to fixate elsewhere and then another to re-fixate on the previous area takes longer than 75 ms. The visual angle threshold was selected because eye movements during fixation are usually less than 0.25 degrees of visual angle with some micro saccades smaller than 0.5 degrees (Komogortsev et al., 2010). Fixations shorter than 60 ms were discarded, as the lower bound of fixation duration is commonly reported to range between 60 and 100 ms (Komogortsev et al., 2010).

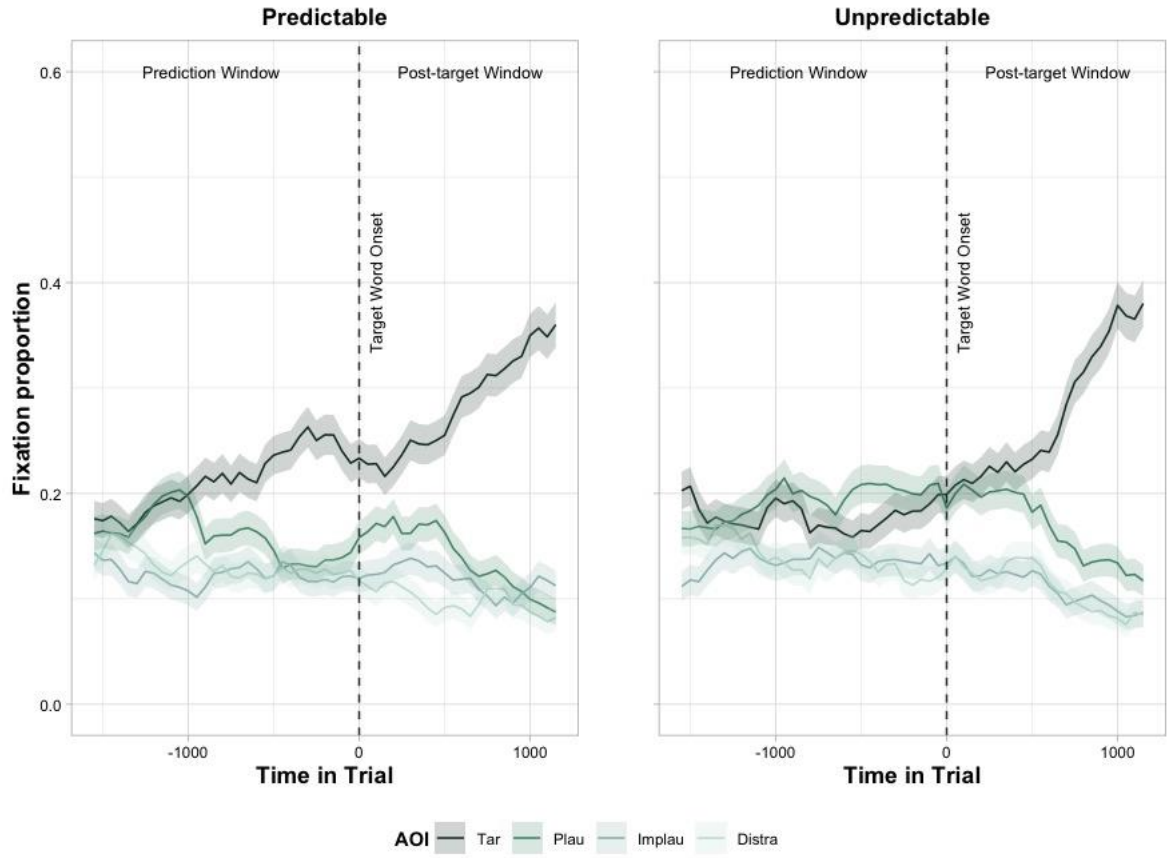
Four areas of interest (AOIs) were created, corresponding to the four objects: the targets (Tar), the plausible competitors (Plau), the implausible competitors (Implau), and the distractors (Distra). All AOIs were identical in shape (square) and size, and their positions were symmetrical relative to the centre point of the display, either vertically or horizontally. Each trial was annotated with two temporal markers: the critical verb onset (CV onset), the target word onset (TW onset). Two time-windows of interests were defined: the prediction window, spanning from the CV onset to the TW onset, and the post-target window, extending from the TW onset to 1000 ms thereafter.

Subsequently, the eye-tracking data were processed and cleaned using the *eyetrackingR* package (version 0.2.0, Dink & Ferguson, 2015) in R Studio (version 2024.04.2+764, RStudio Team, 2024). A two-step exclusion procedure was applied based on each participant's average track-loss rate, defined as the proportion of samples in which the eye-tracker failed to record the participant's eye position. This was calculated by dividing the number of track-loss samples by the total number of samples across all trials. There is no universally accepted threshold for track-loss rate. For instance, Dink and Ferguson (2015) excluded trials with more than 25% track-loss, while de Kloe et al. (2022) used a 75% threshold. Such variation is justifiable, given the heterogeneity of research aims, tasks, and participant characteristics across visual-world paradigm studies. Accordingly, data exclusion criteria should be tailored to the specific behavioural patterns of participants in a given experiment.

As a first step, trials with a track-loss rate higher than one standard deviation above the mean track-loss rate across all participants were excluded, following the method used by Cui and Zheng (2022), whose study involved visual search, a process comparable to the visual perception task involved in the current study. As a second step, data from participants who contributed fewer than 25% of all trials (i.e., fewer than five trials) were excluded, in line with Liu et al. (2022). The rational for the second step is that participants who contributed very few trials may have failed to attend to the visual stimuli regularly and consistently across trials. Consequently, the limited data from their remaining trials were considered unreliable for estimating individual eye movement patterns. This two-step procedure allowed the exclusion of participants who did not engage meaningfully with the visual displays, while retaining valid trials and increasing data variance. Using this method, eye-tracking data from nine student participants were excluded. In comparison, Experiment 2 of Liu et al. (2022) excluded eleven of forty-one participants (26.8%), whereas the current study excluded nine of sixty-six participants (13.6%), a proportion that should be considered acceptable. The remaining fifty-seven participants contributed an average of 18.46 trials (*SD* = 3.25).

Fixation proportion for each AOI was calculated by dividing the number of gaze samples in each AOI by the total number of samples in each 50 ms bin, spanning from the CV onset to 1000 ms after the TW onset. Following the approach of Ito, Pickering, et al. (2018), both blinks and fixations outside the four defined AOIs were included in the denominator when computing fixation proportions. A 180 ms forward shift was added to account for saccade programming latency. This latency is defined as the delay between the reception of a visual signal and the initiation of a corresponding eye movement, typically ranging from 150 ms to 200 ms (Findlay, 1997; Salverda, Kleinschmidt, & Tanenhaus, 2014). Altmann (2011) reported an even shorter estimate of approximately 100 ms based on a re-analysis of two previously published visual-world paradigm studies. Given the complexity of the current experiment task, which was likely to prolong saccade programming latencies, and the need to retain a sufficient amount of data to guarantee statistical power, a 180 ms forward shift was deemed appropriate. Consequently, the two time-windows were re-defined: the prediction window extended from –1550 ms to 180 ms, and the post-target window from 180ms onwards. Figure 7-1 present an overview of the time course of eye movements across all participants in the prediction and the post-target time windows.

Figure 7-1. Time course of the fixation proportions of the four objects in the two conditions.



#Notes: The shaded areas represent ±1 SE. Time 0 ms represents the TW onset.

Figure 7-1 illustrates that, in the predictable condition, the four objects attracted similar fixations at the CV onset until approximately –700 ms, when the targets began to receive more fixations than the other three objects. An obvious downward inflection was observed for the targets at around –250 ms, whereas an opposite trend emerged for the plausible competitors. Around 200 ms after the TW onset, fixations on the targets increased again, while fixations on the other three objects decreased gradually. In the unpredictable condition, the clear divergence between the targets and the other three objects was not observed. Instead, both the targets and the plausible competitors appeared to attract more fixations than the other two objects from –1000 ms onwards, but the differences were much smaller than those in the predictable condition. Again, fixations on the targets exceeded those on the other three after the TW onset; however, this occurred around 500 ms, later than in the predictable condition.

7.1.2 By-group analysis for the prediction and the post-target windows

As an initial step in analysing the eye-tracking data, each time window of interest was examined to investigate the effects of condition and AOI. Fixation proportions were aggregated

by trial within participants for both the prediction and the post-target windows. LME models were constructed for each group under each condition using the *lmer()* function from the *lme4* package (version 1.1-31.1, Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was the empirical logit transformation of fixation proportions (Elog). The fixed effects included an interaction between AOI and condition. The condition effect was centred, with the two conditions being sum-coded using the *contr.sum()* function from the *stats* package. Therefore, the main effect of condition (unpredictable) reflects the difference between the unpredictable condition and the grand mean of both the predictable and the unpredictable conditions. The AOI effect was examined using a treatment contrast with the *contr.treatment()* function from the *stats* package, where each object was compared against the targets. In other words, the main effect for AOI (e.g., the Plau AOI) reflects the difference between that AOI and the Tar AOI.

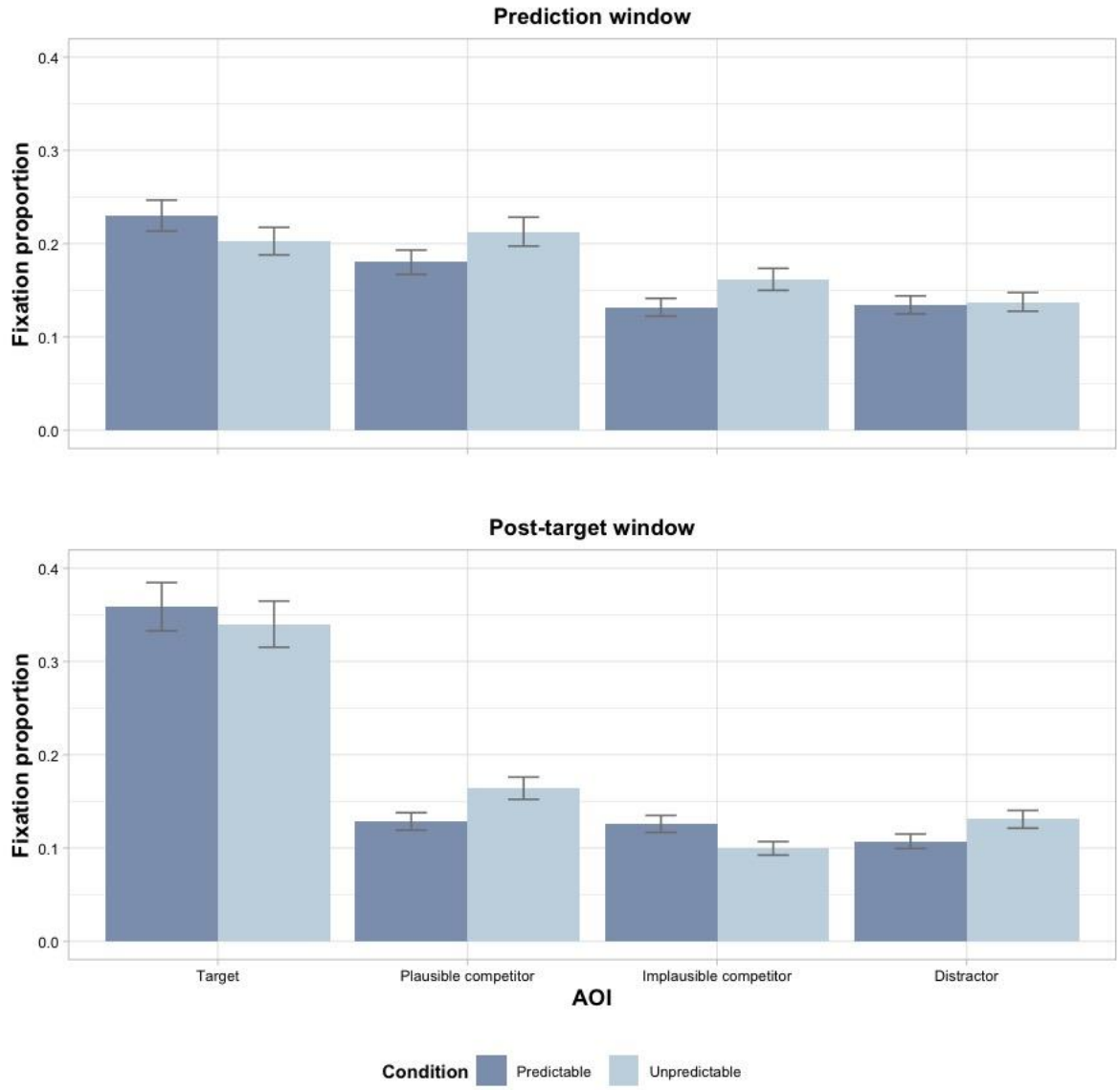
LME models were initially fitted using a maximal random-effects structure, specifying both participants and trials as random effects with intercepts and slopes. The *"bobyqa"* optimiser was employed to aid model convergence. As none of the full models exhibited convergence warning, no reduction procedures were necessary, and the final models for both groups in both time windows retained the maximal random-effects structure. To ensure the reliability of these findings, the same LME models were rerun on a subsampled dataset following the subsampling procedure described in Chapter 6 Section 6.3.2. The results remained consistent in both effect directions and magnitude, confirming the robustness of the main findings (see Appendix 2).

7.1.2.1 Results for the professional interpreters

Figure 7-2 illustrates the average fixation proportions of each object under the two conditions within the professional group. Consistent with the observations of Figure 7-1, during the prediction window, there was an interaction of AOI and condition for the targets and the plausible competitors. Specifically, in the predictable condition, the targets received more fixations than the plausible competitors, whereas in the unpredictable condition, the plausible competitors received slightly more fixations than the targets. Both the targets and the plausible competitors were fixated on more frequently than the plausible competitors and the distractors in both conditions. Both the implausible competitors and the distractors attracted more fixations in the unpredictable condition compared to the predictable condition. In the post-target window, fixation proportions for the targets clearly outnumbered those of the other three objects across both conditions. The targets and the implausible competitors showed a slight decrease in the

unpredictable condition, whereas the plausible competitors and the distractors exhibited an opposite trend.

Figure 7-2. The professional interpreters' average fixation proportions of the four objects in the prediction and the post-target windows.



#Notes: The error bars represent ±1 *SE*.

LME models further confirmed the significance and the magnitude of effects observed in Figure 7-2 (see Table 7-1). In the prediction window, no significant main effect of condition was found. The targets showed marginally significant advantages over the implausible competitors and the distractors in the predictable condition, as indicated by the marginally significant negative estimates. Consistent with the previous observations, the professionals demonstrated a greater fixation preference for the targets over the plausible competitors in the

predictable condition, as evidenced by a significant interaction between the Plau AOI and the condition. Specifically, while the plausible competitors did not significantly differ from the targets in the predictable condition, they attracted significantly more fixations relative to the targets in the unpredictable condition. This suggests that in the absence of predictive CVs, the professionals were more likely to consider the plausible competitors than the targets prior to the onset of the TW. There was also a marginally significant interaction effect for the distractors, indicating a tendency for increased fixations on the distractors in the unpredictable condition compared to the predictable condition, although this did not reach conventional levels of statistical significance.

Table 7-1. LME model for the professional group in the prediction window

Fixed effect	**Estimate**	**Std. Error**	***df***	***t***	***p***
(Intercept)	−3.063	0.490	29	−6.251	< .001 ***
Plausible competitor	−0.697	0.634	27	−1.099	0.282
Implausible competitor	−1.467	0.769	27	−1.908	0.067 †
Distractor	−1.396	0.723	26	−1.932	0.064 †
Unpredictable condition	−0.455	0.312	308	−1.457	0.146
Plau × Unpre	1.029	0.425	1384	2.420	0.016 *
Implau × Unpre	0.630	0.426	1381	1.480	0.139
Distra × Unpre	0.706	0.425	1382	1.659	0.097 †

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

In the post-target window, the targets attracted significantly more fixations than all other object types, as indicated by the significant negative estimates for the plausible competitors, the implausible competitors, and the distractors. This suggests that participants consistently directed their gaze toward the targets after the onset of the TW. The negative estimates for the unpredictable condition and the interaction between the Implau AOI and the condition, despite both being insignificant, align with the observation that the targets and the implausible competitors attracted slightly fewer fixations in the unpredictable condition in the post-target window. Unlike in the prediction window, no significant interaction between condition and object type was observed, indicating that the predictability of the CVs did not significantly modulate fixations in the post-target window. This pattern implies that once the target word was heard, participants reliably shifted their attention to the targets regardless of condition.

Table 7-2. LME model for the professional group in the post-target window

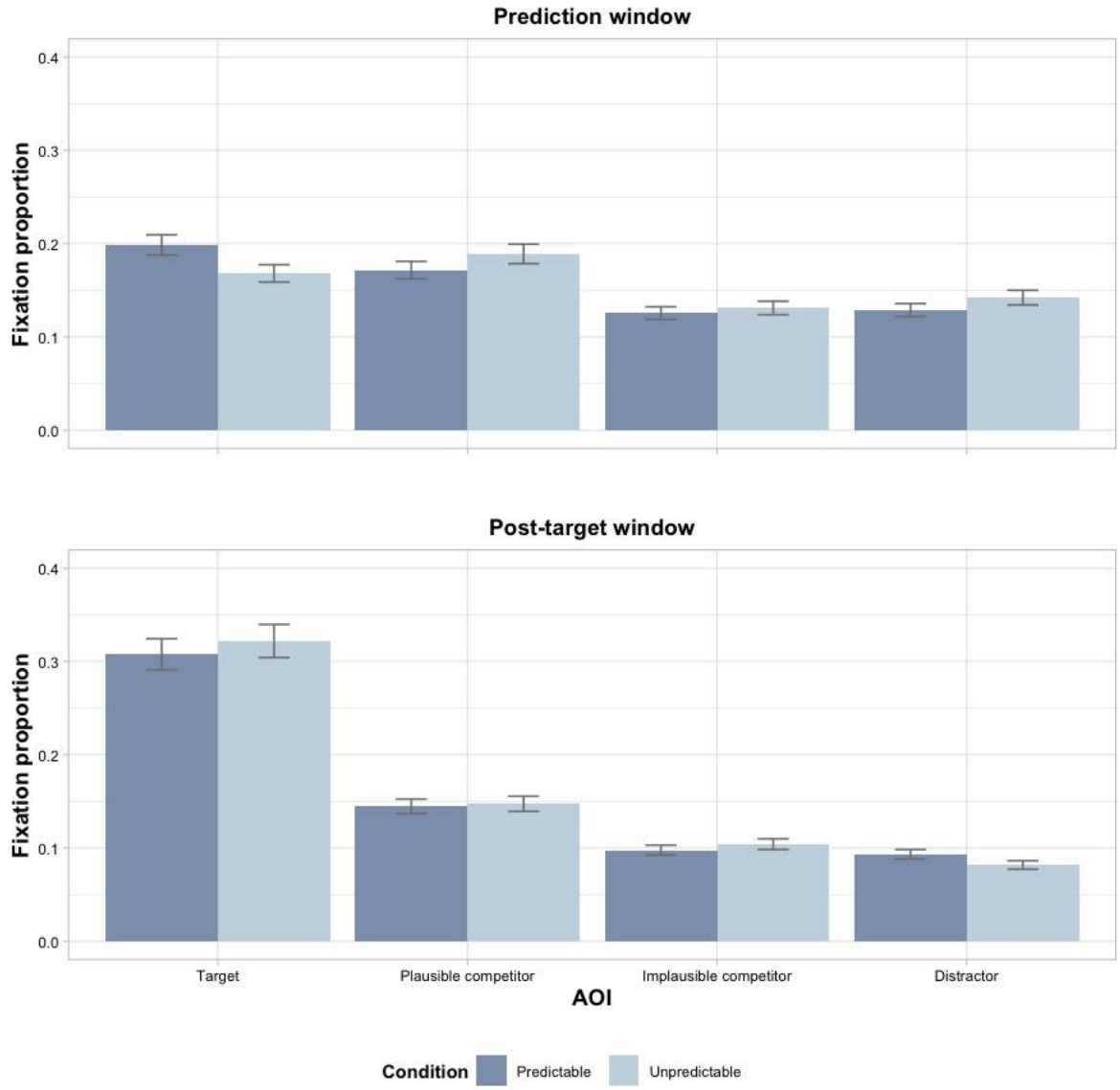
Fixed effect	**Estimate**	**Std. Error**	***df***	***t***	***p***
(Intercept)	−2.028	0.486	34	−4.173	< .001 ***
Plausible competitor	−2.621	0.656	38	−3.997	< .001 ***
Implausible competitor	−2.754	0.783	32	−3.519	0.001 **
Distractor	−2.931	0.696	36	−4.209	< .001 ***
Unpredictable condition	−0.183	0.314	298	−0.583	0.560
Plau × Unpre	0.619	0.427	1408	1.448	0.148
Implau × Unpre	−0.229	0.428	1404	−0.536	0.592
Distra × Unpre	0.594	0.427	1408	1.389	0.165

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

7.1.2.2 Results for the interpreting students

Figure 7-3 illustrates the average fixation proportions of each object under the two conditions in the student group. Similar to the professionals, students demonstrated an interaction between the Plau AOI and the condition in the prediction window, as indicated by the reduced fixation proportions for the targets and increased fixation proportions for the plausible competitors in the unpredictable condition relative to the predictable condition. Both the implausible competitors and the distractors attracted slightly more fixations in the unpredictable condition. In the post-target window, all objects except the distractors demonstrated slight increases of fixation proportions, while the distractors were attended even less in the unpredictable condition. Compared to the professionals, fixation proportions of the student group in each condition and their between-condition differences were smaller across all four objects in both time windows.

Figure 7-3. The interpreting students' average fixation proportions of the four objects in the prediction and the post-target windows.



#Notes: The error bars represent ±1 *SE*.

In the prediction window, LME models revealed no significant main effects or interactions for any of the AOIs or conditions in the student group. Although Figure 7-3 showed higher fixation proportions for the targets than those for the other three objects in the predictable condition, such differences did not reach statistical significance. Similarly, the observed higher fixation proportions for the plausible competitors than those for the other three objects in the unpredictable condition were not statistically significant. These non-significant effects suggest that students may not have made anticipatory fixations based on the predictability of the CV during the prediction window. It is also possible that the anticipatory fixations were not

pronounced throughout the whole prediction window or only occurred during a brief period within the prediction window.

Table 7-3. LME model for the student group in the prediction window

Fixed effect	Estimate	Std. Error	*df*	*t*	*p*
(Intercept)	−4.042	0.421	30	-9.607	< .001 ***
Plausible competitor	−0.176	0.570	24	−0.309	0.760
Implausible competitor	−1.057	0.657	23	−1.608	0.121
Distractor	−0.897	0.616	24	−1.455	0.159
Unpredictable condition	−0.043	0.244	538	−0.177	0.860
Plau × Unpre	0.068	0.337	2500	0.201	0.841
Implau × Unpre	0.136	0.337	2497	0.402	0.688
Distra × Unpre	0.253	0.337	2498	0.750	0.454

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

In the post-target window, students demonstrated a pattern consistent with that of the professionals. Specifically, the targets attracted significantly more fixations than all other object types, as indicated by the significantly negative estimates for the plausible competitors, the implausible competitors, and the distractors. This demonstrates that students consistently directed their gaze toward the targets after the TW onset. No significant main effect of condition or interaction between the condition and the AOI was observed. These findings suggest that students, like the professionals, reliably shifted their attention to the targets in the post-target window regardless of whether the condition was predictable or unpredictable.

Table 7-4. LME model for the student group in the post-target window

Fixed effect	Estimate	Std. Error	*df*	*t*	*p*
(Intercept)	−2.862	0.478	42	−5.983	< .001 ***
Plausible competitor	−1.724	0.526	31	−3.275	0.003 **
Implausible competitor	−2.749	0.700	28	−3.925	< .001 ***
Distractor	−2.605	0.576	32	−4.521	< .001 ***
Unpredictable condition	0.185	0.232	1193	0.797	0.426
Plau × Unpre	−0.323	0.326	2503	−0.991	0.322
Implau × Unpre	0.045	0.326	2493	0.137	0.891
Distra × Unpre	−0.432	0.326	2501	−1.327	0.185

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

To summarise, both groups exhibited an interaction between the Plau AOI and the condition in the prediction window. Specifically, fixation proportions for the targets were higher than those for the plausible competitors in the predictable condition, whereas the opposite trend was observed in the unpredictable condition. Meanwhile, the implausible competitors did not attract more fixations in the predictable condition, despite their strong lexical associations with the predictive CVs. These patterns support the simultaneous engagements of both the prediction-by-production and the prediction-by-association mechanisms during SI. The high predictability associated with the predictive CVs allowed interpreters to engage in a top-down analysis, utilising the production system to infer and imitate the speaker's intent and speech. As a result, the implausible competitors were excluded due to their limited contextual plausibility, while the targets were anticipated because they were compatible with both the predictive CVs and the broader context. When both the targets and the plausible competitors were compatible with the unpredictive CVs and the context, the latter attracted more fixations than the former due to the stronger lexical association between the plausible competitors and the unpredictive CVs. This suggests a role for prediction-by-association in the absence of strong predictive cues. However, this interaction reached significance only in the professional group, indicating a higher level of predictive processing in the professionals than in students.

In the prediction window, the implausible competitors and the distractors showed increased fixation proportions in the unpredictable relative to the predictable conditions. This suggests that fixations were more evenly distributed across the four AOIs in the predictable condition but more concentrated in the predictable condition. In the post-target window, fixation proportions for each object were highly consistent across the two conditions, with the targets attracting significantly more fixations than the other three objects in both conditions. This pattern reflects language-driven eye movements following the explicit mention of the targets in the linguistic input. Overall, the professionals showed higher fixation proportions than students across both conditions and all objects, suggesting greater cognitive flexibility of the professionals in coordinating both audio and visual inputs.

Unexpectedly, the student group showed no significant condition effect for any object, nor any significant AOI effect between the targets and the other three objects in the predictable condition during the prediction window. A potential explanation is that these differences were not significant for more than half of the prediction window, thereby counterbalancing any

potentially significant effects within a smaller portion of the time window. Aggregating fixation proportions across broad time windows may have obscured more fine-grained temporal patterns. Consequently, while such analyses provide a general overview of the eye movement patterns, they are limited in their ability to identify the timing and progressing of effects as the discourse unfolds. The ability to capture these temporal dynamics is a key advantage of the visual-world paradigm. The following sections report results from alternative approaches that examine the time course of these effects in greater details.

7.1.3 By-group analysis for temporal dynamics of the prediction effect

For a finer-grained time course of the effects, GCA were conducted to examine the non-linear dynamic changes of the condition and the AOI effects over time. GCA models were constructed for each group using the *lmer()* function from the *lme4* package (version 1.1-31.1, Bates et al., 2015). The model evaluated the Elog fixation proportions predicted by fixed effects of condition (predictable vs. unpredictable) and AOI (Tar vs. Plau, Implau, or Distra), and the interaction of the two on all time terms. The condition effect was sum-contrast coded, with 0.5 representing the predictable condition and –0.5 representing the unpredictable condition. The AOI effect was treatment-contrast coded, with the targets as the reference level.

The appropriate polynomial terms (quadratic, cubic, or quartic) were determined based on several considerations, balancing the goodness of model fit, rates of Type I error, and the interpretability of polynomial terms. There is no consensus on the criterion for selecting polynomial terms. The goodness of model fit provides a critical criterion for excluding higher order polynomial terms that do not significantly improve the model fit. However, solely relying on the goodness of model fit and including higher order polynomial terms (e.g., quartic terms and above) to enhance model fit may result in inflated rates of Type I error (Huang & Snedeker, 2020). Besides, higher order polynomial terms tend to be sensitive to the asymptotic tails of the curve and often lack clear cognitive interpretation in the context of visual-world paradigm studies (Mirman et al., 2008), and including higher order polynomial terms into the random structure require heavy computation. In this study, visual inspection of Figure 7-1 indicates that the changes in effects (i.e., inflection points) that are most relevant to the research interest mostly occurred around the middle part (approximately –700 ms to the TW onset) and the last third (around 250ms onwards) of the time course. In other words, among all the inflection points observed from Figure 7-1, those appearing in the time window around –700 ms to the TW onset and from 250 ms onwards were most relevant to the interests, as these inflection

points often corresponded to the orientation of fixations away from the other three objects to targets or vice versa. Including quadratic and cubic polynomial terms was deemed sufficient to capture the shift of attention in these time windows. Given the potential risk of inflated rates of Type I error and the complexity of interpretation, the quartic polynomial term was excluded from the GCA models despite significantly enhancing the model fit. Based on these criteria, cubic terms were retained in the final models.

GCA models were initially fitted using a maximal random-effects structure, specifying both participants and trials as random effects with intercepts and slopes. The *"bobyqa"* optimiser was employed to aid model convergence. In cases where the full model failed to converge, an iterative reduction strategy was applied followed Bates et al. (2015) to balance model fit and parsimony. This reduction procedure commenced with the removal of item-level correlations, followed by the stepwise elimination of random slopes, and ceased when no further convergence warning occurred, and goodness of fit, as indicated by the Akaike information criterion value, no longer improved significantly with further parameter trimming. Participant-level weights, calculated as the inverse of the number of contributed trials with a formular $1/number\ of\ contributed\ trials$, were included to account for variability in trial contributions across participants. P-values for fixed effects were derived using Satterthwaite's approximation method. The final models for both groups in both time windows all included maximal random-effects structure.

The final GCA model for the professional group retained uncorrelated random intercepts and slopes for the linear term (ot1) across trials, specified as $(1 + Predictable + ot1 \,||\, Trial)$, capturing trial-specific variations in the temporal dynamics of the fixation proportions. Random intercepts for participants were modeled as $(1 + Predictable \mid Participant)$, accounting for individual difference in baseline values. The final GCA model for the student group included uncorrelated random intercepts and slopes for the linear term across both trials and participants, specified as $(1 + Predictable + ot1 \,||\, Trial)$ and $(1 + Predictable + ot1 \,||\, Participant)$, respectively.

For the simplicity of interpretation, the AOI and the condition effects were further examined separately. To compare fixation proportions on the four AOIs in greater detail, CPAs were conducted for the AOI effect (Maris & Oostenveld, 2007; Ito & Knoeferle, 2023). The CPA utilised one-sided t-tests to identify time bins during which (Elog) fixation proportions on the targets were significantly higher than those on each of the other objects, under each condition.

T-tests were used due to the convergence issues with the LME models and the lighter computation required by *t*-tests. Three contrast labels were created: TP for Tar vs. Plau, TI for Tar vs. Implau, and TD for Tar vs. Distra. A one-sided *t*-test was conducted for each contrast on each time bin using the *t.test()* function from the *stats* package, to examine if the means of fixation proportions for the targets aggregated by participants were significantly higher than those for the other objects in the contrast. Statistical significance was determined using a dual criterion: t-value exceeding 2 and p-values below .05. Neighbouring time bins meeting these criteria were grouped into clusters, and cluster-mass statistics were computed as the sum of t-values within each cluster. Permutations (n = 1000) were performed by shuffling data within participants for each time bin from –1550 ms to 1180 ms. For each permutation, the cluster detection and cluster-mass statistic calculations were repeated, and the maximum cluster-mass statistics were stored to create a null distribution. Statistical significance of clusters was evaluated using Monte Carlo p-values derived from this null distribution. Following Amos et al. (2022), we only clusters longer than 150 ms will be reported here.

To examine time course of the condition effect, a CPA was performed based on LME models, including a fixed effect of the condition and by-participant and by-trial random intercepts. Considering the large number of observations required to assess random effects, random slopes were excluded from the LME models to avoid singular fit. The LME models were run on each 50 ms time bin from –1550 ms to 1180 ms, and the data were permuted 1000 times, using the *clusterperm.lmer()* function from the *permutes* package (version 2.8, Voeten, 2022). The outputs of the *clusterperm.lmer()* function specified all detected clusters and their cluster-mass statistics and Monte Carlo *p*-values, and significant clusters were identified based on Monte Carlo p-values smaller than 0.05. Considering the typical mean fixation duration for scene perception is around 330 ms (Rayner, 1998), between-condition difference shorter than 150 ms were not considered stable, and thus only clusters equal to or longer than 150 ms were reported here.

To ensure the reliability of the findings despite the slight imbalance in trial numbers between conditions, a robustness check was conducted by rerunning the same GCA models on a subsampled dataset that reduced the imbalance. The results of the robustness check are reported alongside the main results for each group in the respective results sections. CPAs were not rerun as they are computationally intensive and generally robust to moderate trial imbalance.

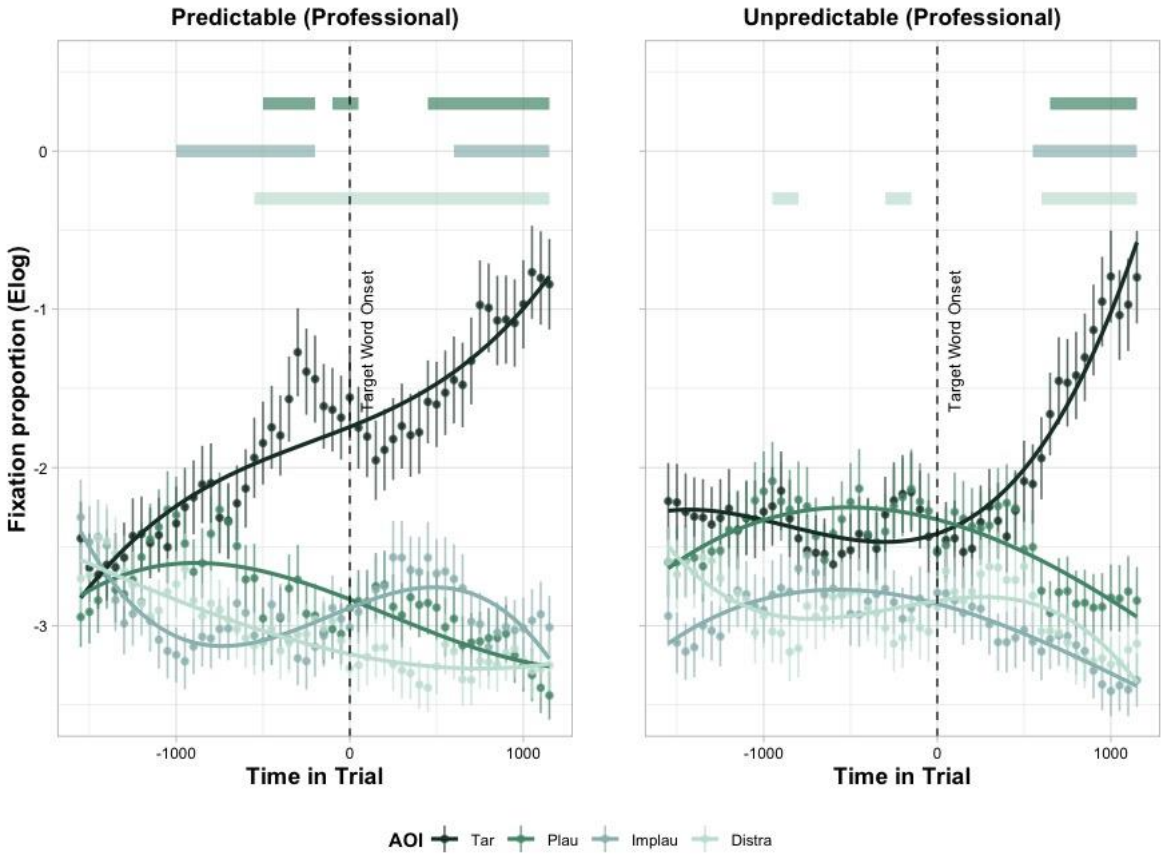
7.1.3.1 Results for the professional interpreters

The GCA for the professional interpreter group revealed significant interactions between the condition and the AOI, reflecting differential temporal dynamics in fixation proportions across the four presented objects as a function of predictability. Specifically, a significant interaction was observed between the predictable condition and the plausible competitors ($\beta = -0.677$, $SE = 0.059$, $t = -11.557$, $p < .001$), with the negative estimate indicating reduced fixation proportions on the plausible competitors in the predictable condition compared to the unpredictable condition. Similarly, significant interactions were found for the predictable condition × the implausible competitors ($\beta = -0.242$, $SE = 0.059$, $t = -4.127$, $p < .001$) and the predictable condition × the distractors ($\beta = -0.441$, $SE = 0.059$, $t = -7.520$, $p < .001$), suggesting lower fixation proportions on these two objects under the predictable condition. These findings indicate that the professionals were more likely to attend to non-target objects in the unpredictable condition. This pattern could be attributed to the influence of unpredictive CVs, which restricted TW predictions and led to increased fixation drift across all four objects. The plausible competitors, in particular, were more likely to attract participants' attention as they were compatible with both the unpredictive CVs and the global contexts.

In terms of temporal dynamics, significant interactions emerged between the predictable condition and the linear ($\beta = 1.124$, $SE = 0.308$, $t = 3.646$, $p < .001$), suggesting a shaper overall increase in the predictable condition. A significant interaction was also found for the quadratic terms ($\beta = -2.338$, $SE = 0.308$, $t = -7.583$, $p < .001$), indicating a significant between-condition differences for the targets in around the midpoint of the timeline. This aligns with the peak in the predictable condition and the U-shaped curve in the unpredictable condition around −500 ms relative to the TW onset. For the plausible competitors, significant interactions were observed for the linear ($\beta = -1.741$, $SE = 0.436$, $t = -3.993$, $p < .001$) and the quadratic terms ($\beta = 2.863$, $SE = 0.436$, $t = 6.567$, $p < .001$). For the implausible competitors, significant interactions were observed for the quadratic ($\beta = 3.727$, $SE = 0.436$, $t = 8.548$, $p < .001$) and the cubic terms ($\beta = -1.076$, $SE = 0.459$, $t = -2.469$, $p = .014$). The distractors also exhibited significant interactions with the predictable condition across all time terms: the linear ($\beta = -1.943$, $SE = 0.436$, $t = -4.456$, $p < .001$), the quadratic ($\beta = 2.942$, $SE = 0.436$, $t = 6.747$, $p < .001$), and the cubic terms ($\beta = 1.185$, $SE = 0.459$, $t = 2.719$, $p = .007$). These significant quadratic interactions align with the observed pattern where the fixation differences between the targets and each of the other three objects peaked around the middle of the time course under the predictable condition, whereas the between-AOI differences in the unpredictable

condition remained relatively stable until approximately 500 ms after the TW onset (Figure 7-4). Results from the robustness check confirmed these main findings for the professional group, with effect directions preserved and significance levels remaining unchanged or varying only slightly (e.g., from $p < .001$ to $p < .01$; see Appendix 3).

Figure 7-4. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the professional interpreter group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

The CPA for the AOI effect identified significant clusters where fixation proportions on the targets exceeded those on the other three objects before the TW onset in the predictable condition but not in the unpredictable condition. In the predictable condition, three significant clusters were detected for the TP contrast: from –500 ms to –200 ms (cluster mass statistic = 21.933, $p < .001$), from –100 ms to 50 ms (cluster mass statistic = 10.005, $p < .001$), from 450 ms to 1150 ms (cluster mass statistic = 64.600, $p < .001$). The TI contrast showed two significant clusters: from –1000 ms to –200 ms (cluster mass statistic = 55.029, $p < .001$) and from 600 ms to 1150 ms (cluster mass statistic = 52.909, $p < .001$). For the TD contrast, a

single significant cluster was identified, spanning from –550 ms to 1150 ms (cluster mass statistic = 143.384, $p < .001$). The significant clusters before the TW onset demonstrate that the professionals preferentially fixated on the targets before hearing the TW, suggesting successful TW prediction under the predictable condition. This is consistent with the GCA results that estimates for interactions between the predictable condition and each of non-target objects were significantly positive for the quadratic term. The fact that significant clusters for the TI and TD contrast appeared earlier than those for the TP contrast suggests that the professionals excluded the implausible competitor and distractor objects before the plausible competitors. After the TW onset, the significant clusters reflect confirmatory fixations after hearing the TW and integrated the TW into their comprehension and/or production process.

In the unpredictable condition, significant clusters appeared only after the TW onset for the TP and TI contrasts: from 650 ms to 1150 ms for the TP contrast (cluster mass statistic = 47.122, $p < .001$) and from 550 ms to 1150 ms for the TI contrast (cluster mass statistic = 64.812, $p < .001$). For the TD contrast, there are three significant clusters: from –950 ms to –800 ms (cluster mass statistic = 11.352, $p < .001$), from –300 ms to –150 ms (cluster mass statistic = 11.264, $p < .001$), and from 600 ms to 1150 ms (cluster mass statistic = 57.912, $p < .001$). The absence of significant clusters for the TP and TI contrasts before TW onset indicates a delay in fixating on the targets, which suggests that participants relied on bottom-up processing of the auditory input rather than pre-activating target-related information. This also supports that fixations were more evenly distributed among the AOIs in the unpredictable condition. The two significant clusters for the TD contrast before TW onset are likely driven by random fluctuations or subtle biases, as they do not consistently favour the targets.

The CPA for the condition effect identified a significant positive cluster for the targets, extending from −600 ms to 250 ms (cluster mass statistic = 152.401, $p < .001$), indicating higher fixation proportions for the targets in the predictable condition than in the unpredictable condition. This positive cluster is in line with the GCA result, which showed that the predictable condition had a significant negative effect on the quadratic term. For the plausible competitors, three significant negative clusters were observed: from −250 ms to −50 ms (cluster mass statistic = 68.525, $p < .001$), from 250 ms to 450 ms (cluster mass statistic = 38.418, $p = .003$), and from 1000 ms to 1150 ms (cluster mass statistic = 38.658, $p = .003$). These negative clusters indicate higher fixation proportions for the plausible competitors in the unpredictable condition during these time spans. The implausible competitors also exhibited three significant clusters, but in different direction, extending from −1550 ms to −1400 ms (cluster mass statistic

= 39.951, $p < .001$), from 300 ms to 650 ms (cluster mass statistic = 52.256, $p < .001$), and from 900 ms to 1100 ms (cluster mass statistic = 31.690, $p < .001$). In contrast, a negative cluster was identified for the distractors extending from 250 ms to 500 ms (cluster mass statistic = 76.721, $p < .001$). The presence of significant clusters after the TW onset suggests that the professionals may not have fixated exclusively on the targets despite their explicit mention. Instead, they continued to shift their attention across the other three objects.

7.1.3.2 Results for the interpreting students

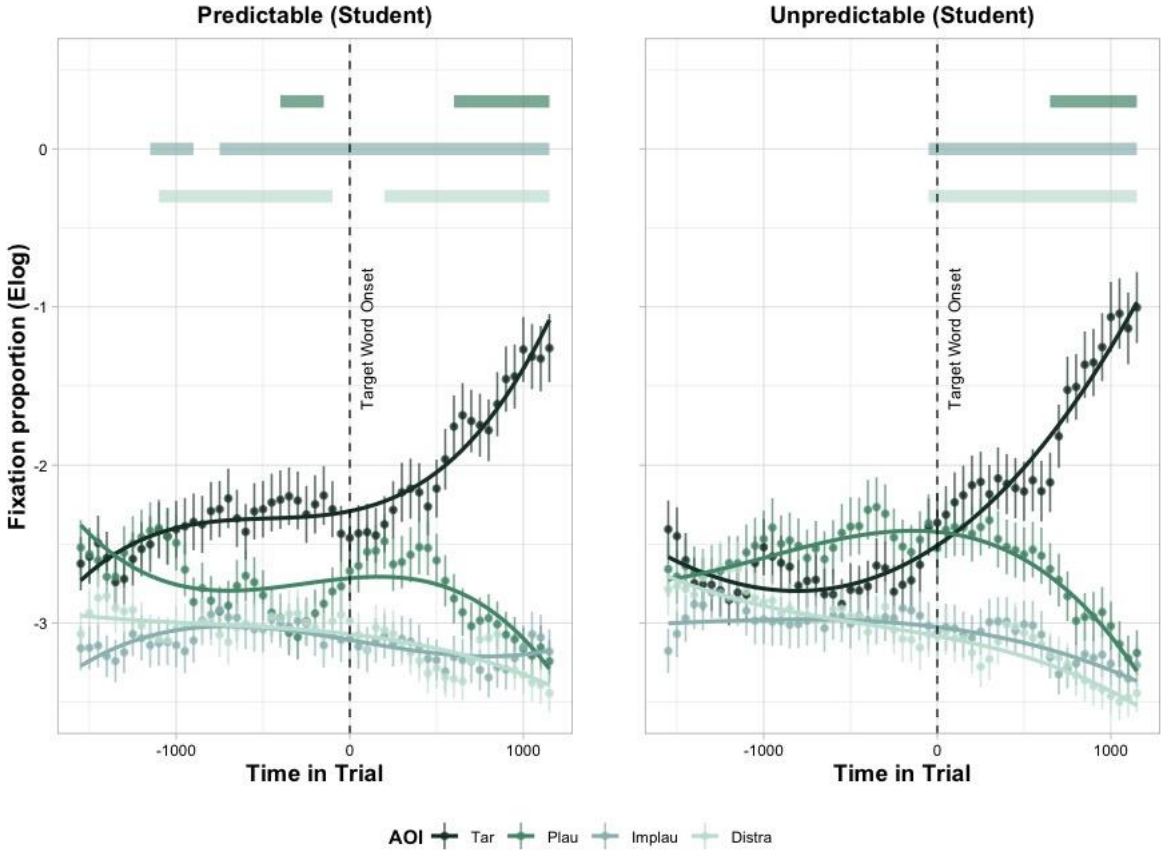
The GCA results for the interpreting student group illustrated patterns similar to those observed for professional interpreters, with significant interactions between condition and AOI (Figure 7-5). In particular, the predictable condition was associated with reduced fixation proportions on the non-target objects. Significant interactions were observed between the predictable condition and the plausible competitors ($\beta = -0.310$, $SE = 0.043$, $t = -7.258$, $p < .001$), the implausible competitors ($\beta = -0.218$, $SE = 0.043$, $t = -5.092$, $p < .001$), and the distractors ($\beta = -0.183$, $SE = 0.043$, $t = -4.276$, $p < .001$). These results indicate that, like the professionals, the students showed a greater tendency to fixate on non-target objects in the unpredictable condition. However, the effect sizes are smaller for the students than for the professionals, suggesting that compared to the professionals, the students had less pronounced attentional shifts to the targets in the predictable condition.

In terms of temporal interactions, the practicable condition showed significant effects on all temporal terms. For the linear term ($\beta = -0.944$, $SE = 0.226$, $t = -4.176$, $p < .001$), the negative estimate indicates that the overall increase of fixation proportions for the targets was higher in the unpredictable condition than in the predictable condition. For the quadratic term ($\beta = -1.021$, $SE = 0.225$, $t = -4.535$, $p < .001$), the significant effect reflects the divergent curvatures of the fixation trajectories across the two conditions. Finally, the cubic term was also significant ($\beta = 0.718$, $SE = 0.225$, $t = 3.190$, $p = .001$), suggesting a more complicated pattern of fluctuations at the two ends of the time course. Significant effects were also observed for condition × AOI interactions on each time term. For the linear term, there were significant interactions between the predictable condition and the plausible competitors ($\beta = 0.726$, $SE = 0.318$, $t = 2.279$, $p = .023$), the implausible competitors ($\beta = 1.392$, $SE = 0.318$, $t = 4.372$, $p < .001$), and the distractors ($\beta = 1.627$, $SE = 0.318$, $t = 5.109$, $p < .001$). Similarly, significant quadratic term interactions were observed for the plausible competitors ($\beta = 2.171$, $SE = 0.318$, $t = 6.819$, $p < .001$), the implausible competitors ($\beta = 1.084$, $SE = 0.318$, $t = 3.404$, $p < .001$),

and the distractors ($\beta = 0.892$, $SE = 0.318$, $t = 2.800$, $p = .005$). For the cubic term, significant interactions emerged for the plausible competitors ($\beta = -1.094$, $SE = 0.318$, $t = -3.437$, $p < .001$) and the distractors ($\beta = -0.640$, $SE = 0.318$, $t = -2.012$, $p = .044$). These temporal dynamics are similar to the results observed for professional interpreters, where fixation differences between the targets and the non-target objects peaked around the middle of the time course in the predictable condition. However, the magnitudes of the temporal interaction effects were smaller for the students than for the professionals.

The robustness check generally confirmed the main predictive effects in the student group (see Appendix 3): the targets consistently attracted more fixations than the other three objects in the predictable condition, and significant between-AOI differences remained on the quadratic terms, reflecting a prediction effect around the midpoint of the time course. However, several higher-order interactions became less significant or non-significant in the subsampled dataset. These reductions are likely due to the decreased numbers of trials and greater between-subject variability within the student group, which increased the sensitivity of the results to changes in trial structure. The implications of this finding are further addressed in Section 7.1.6.

Figure 7-5. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the interpreting student group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

The CPA for the AOI effect in the interpreting student group also resembled the patterns observed in the professional interpreter group. In the predictable condition, two significant clusters emerged for each contrast: from –400 ms to –150 ms (cluster mass statistic = 15.465, $p < .001$) and from 600 ms to 1150 ms (cluster mass statistic = 53.807, $p < .001$) for the TP contrast; from –1150 ms to –900 ms (cluster mass statistic = 13.888, $p < .001$) and from –750 ms to 1150 ms (cluster mass statistic = 147.260, $p < .001$) for the TI contrast; from –1100 ms to –100 ms (cluster mass statistic = 54.268, $p < .001$) and from 200 ms to 1150 ms (cluster mass statistic = 93.724, $p < .001$) for the TD contrast. These early clusters, appearing before the TW onset, align closely with the GCA findings, which demonstrated significant quadratic interactions for all non-target objects in the predictable condition. In the unpredictable condition, only one significant cluster was identified for each contrast: from 650 ms to 1150 ms (cluster mass statistic = 48.913, $p < .001$) for the TP contrast; from –50 ms to 1150 ms (cluster mass statistic = 110.131, $p < .001$) for the TI contrast; from –50 ms to 1150 ms (cluster mass statistic = 116.305, $p < .001$) for the TD contrast. These results indicate that in the unpredictable condition, fixation differences across four objects emerged much later and persisted uniformly through the latter part of the time course.

Similar to the professionals, the CPA for the condition effect revealed a significant positive cluster for the targets in the student group, extending from −800 ms to −150 ms (cluster mass statistic = 106.194 $p < .001$), consistent with the negative effect of the predictable condition on the quadratic term in the GCA model. For the plausible competitors, a single significant negative cluster was identified from −500 ms to −50 ms (cluster mass statistic = 103.273, $p < .001$), indicating higher fixation proportions for the plausible competitors in the unpredictable condition. No significant clusters were found for either the implausible competitors or the distractors, suggesting relatively minor variations in the eye movements for these two objects across conditions.

To summarise, the GCAs and CPAs of the condition and the AOI effects captured the temporal evolution of the effects and identified the specific time spans in which significant effects occurred—details not accessible through the by-window LME analysis. Consistent with the speculations outlined in Section 1.2, these effects were not sustained uniformly across the

entire time course or the prediction window but varied dynamically as the audio stimuli unfolded. In both groups, the CPAs for the AOI effect revealed significant clusters for all between-AOI contrasts before the TW onset in the predictable condition. This indicates that both the professionals and students had already fixated predominantly on the targets prior to the TW onset, reducing their attentions to the other three object. Among the non-target objects, the implausible competitors were excluded earliest, followed by the distractors, with the plausible competitors excluded latest, around –500 ms for the professionals and –400 ms for students. The exclusion of the non-target objects occurred much later in the unpredictable condition. In particular, for the plausible competitors, no significant cluster for the TP contrast was observed until 650 ms after the TW onset, suggesting that participants hesitated between the targets and the plausible competitors until the TW was explicitly mentioned in the audio input.

In both groups, the targets were more likely to be predicted prior to the TW onset in the predictable condition than in the unpredictable condition, as evidenced by the significant condition effect on the quadratic term in the GCA models and the significant cluster before the TW onset in the CPAs. Conversely, the plausible competitors attracted more attention before the TW onset in the unpredictable condition, as indicated by the significant interactions between the Plau AOI and the condition on the linear and the quadratic terms in the GCA models, and by the significant negative clusters before the TW onset identified in the CPAs. These results for the targets and the plausible competitors also align with the interaction effect found in the by-window analysis. Among the professionals, both the implausible competitors and the distractors demonstrated significant interaction with the condition on the cubic terms, as well as significant clusters after the TW onset, suggesting that the professionals continued to shift their attention towards these two objects even after hearing the TW. In contrast, the absence of significant clusters for these two objects in students suggests comparable eye movement patterns across conditions.

7.1.4 By-trial analysis for effects of predictability and lexical association

A by-trial CPA was conducted to estimate the contribution of cloze probability and verb-noun association strengths to eye movements in the prediction window from the CV onset to the TW onset. The CPA was based on linear regression models with trial-aggregated (Elog) fixation proportions on each object in every 50 ms time bin as the dependent variable predicted by cloze probability (Cloze) and verb-noun association (VN). Given the relatively small sample size

(i.e., forty data points per time bin), which might limit the statistical power of linear regression, the model was fitted on the aggregated data from every two consecutive time bins. Clusters were identified for each predictor as groups of consecutive time bins where the regression estimates were statistically significant and had the same direction (either positive or negative). The cluster mass was calculated by summing the t-statistics within each identified cluster. The data were permuted 1000 times for each predictor by shuffling that predictor across trials, and the regression analysis and cluster detection procedures were repeated for each permuted dataset. Statistical significance of clusters was evaluated following the same method as described in 1.3.

For the professional group, the by-trial CPA revealed significant effects of cloze probability on all objects except the distractors (Table 7-5). Cloze probability was positively associated with fixation proportions on the targets but negatively associated with the two competitors, highlighting its critical role in guiding predictive processing—enhancing fixation on the anticipated targets while suppressing competitors. Both the targets and the plausible competitors identified three significant clusters, indicating a dynamic and flexible use of cloze probability in prediction strategies. The negative cluster for the implausible competitors, spanning from –1050 ms to –750 ms, suggests that implausible competitors were deprioritised during this period. This finding is consistent with the significant cluster identified in the CPA for the AOI effect in the predictable condition (–1000 ms to –200 ms), where the targets received more fixations than the implausible competitors. Contrary to our expectation that verb-noun association would be positively associated with fixation proportion, the effect of verb-noun association was significantly negative on the targets in the early stage following the CV onset (from –1450 ms to –1250 ms).

Table 7-5. By-trial CPA for the effect of cloze probability and verb-noun association for the professional group

AOI	Predictor	Cluster	Estimate sign	*p*	Cluster mass
Tar	Cloze	–1550 to –1250	Positive	< .001***	8.639
		–950 to –650	Positive	< .001***	11.665
		–350 to – 150	Positive	< .001***	15.057
	VN	–1450 to –1250	Negative	< .001***	7.356
Plau	Cloze	–1550 to –1250	Negative	<.001***	11.726
		–950 to –650	Negative	<.001***	10.517
		–250 to 50	Negative	<.001***	13.404

Implau	Cloze	–1050 to –750	Negative	<.001***	13.746

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. Acronyms: 1) Tar: Target; 2) Plau: Plausible competitor; 3) Implau: implausible competitor; 4) Cloze: cloze probability; 5) VN: verb-noun association

For the student group (Table 7-6), only one significant cluster of the cloze probability effect was identified for each object. Compared to the professionals, these clusters tended to have a longer duration, suggesting that cloze probability triggered less frequent eye movements in the students. Interestingly, a positive cluster for the distractors was observed near the TW onset (from –50 ms to 150), indicating an increase in fixation proportion on the distractors during this period. No effect of verb-noun association was observed for any of the four objects.

Table 7-6. By-trial CPA for the effect of cloze probability and verb-noun association for the student group

AOI	Predictor	Cluster	Estimate sign	*p*	Cluster mass
Tar	Cloze	–1550 to –150	Positive	< .001***	41.961
Plau	Cloze	–1050 to 150	Negative	< .001***	48.064
Implau	Cloze	–1550 to –1050	Negative	< .001***	21.554
Distra	Cloze	–50 to 150	Positive	< .001***	7.082

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. Acronyms: 1) Tar: Target; 2) Plau: Plausible competitor; 3) Implau: implausible competitor; 4) Distra: Distractor; 5) Cloze: cloze probability; 6) VN: verb-noun association

7.1.5 Between-group comparisons

To compare the results from the professional and student groups, a GCA model was run on both groups and specified an interaction by the expertise group. The expertise effect was sum-contrast coded, with 1 representing the predictable condition and –1 representing the unpredictable condition. The interactions between expertise and condition were significant on the linear ($\beta = 1.020$, $SE = 0.190$, $t = 5.381$, $p < .001$) and quadratic terms ($\beta = -0.664$, SE = 0.189, $t = -3.509$, $p < .001$), reflecting more pronounced fixation shifts toward the targets for the professionals compared to students, especially under the predictable condition. This interaction was also significant on the cubic terms, ($\beta = -0.508$, SE = 0.189, $t = -2.685$, $p = .007$), suggesting more frequent attention shifts in the professionals. The expertise effect for the plausible competitors was evident in a significant interaction with the predictable condition on the linear time term ($\beta = -1.233$, $SE = 0.267$, $t = -4.611$, $p < .001$), indicating that the

professionals demonstrated a steeper decline in fixations over time compared to students. The expertise effect for the plausible competitors was also significant in the predictable condition on the cubic terms ($\beta = 0.800$, $SE = 0.267$, $t = 2.989$, $p = .003$), revealing that the professionals made dynamic adjustments in their attention shifts during predictable trials, whereas students exhibited more static attention patterns.

For the implausible competitors, expertise effects appeared primarily in interaction with the predictable condition on the linear ($\beta = -0.893$, $SE = 0.267$, $t = -3.339$, $p < .001$) and the quadratic terms ($\beta = 1.322$, $SE = 0.267$, $t = 4.942$, $p < .001$). Specifically, when predictability was high, the professionals showed a sharper reduction in fixations on the implausible competitors over time. The most pronounced expertise effects were observed in the distractors. A significant three-way interaction between the expertise, the predictable condition and the distractors was significant the linear term ($\beta = -1.785$, $SE = 0.267$, $t = -6.674$, $p < .001$), demonstrating that the professionals rapidly disengaged from distractors, particularly under predictable conditions. Furthermore, significant interactions with the quadratic ($\beta = 1.025$, $SE = 0.267$, $t = 3.834$, $p < .001$) and cubic terms ($\beta = 0.913$, $SE = 0.267$, $t = 3.414$, $p < .001$) revealed that students displayed a slower and more inconsistent pattern of disengagement, underscoring differences in suppression efficiency between the groups.

The robustness check on the between-group GCA comparisons showed that most effects, particularly linear and quadratic interactions involving predictability, expertise, and AOI, remained consistent across the original and the subsampled datasets (see Appendix 4). These results confirmed the stability of core expertise-related differences in fixation dynamics despite change in trial structure. A few interactions showed reduced significance in the subsampled data set (e.g., Pre × Pro × Implau × ot1 shifted from $p < .001$ to $p = .114$), likely due to reduced statistical power resulting from trial subsampling. In contrast, some higher-order interactions became more significant, especially on the cubic term (e.g., Pre × Pro × Plau × ot3), potentially reflecting random variability introduced by the subsampling process. Overall, the main expertise-related effects were preserved, confirming the robustness of the key findings while highlighting that higher-order interactions are more sensitive to trial count and sampling variability.

In sum, the professionals exhibited faster disengagement from non-target objects than students in the predictable condition, as evidenced by significant interactions between expertise and the predictable condition on linear terms for all four objects. The targets, the implausible

competitors, and the distractors also showed significant interactions involving expertise and the predictable condition on the quadratic terms, suggesting more pronounced between-condition differences for these three objects among the professionals, particularly around the middle of the timeline. Additionally, the significant interactions on the cubic terms indicate that the professionals shifted their visual attention more frequently over time.

7.1.6 Interim discussion

The findings provide compelling evidence for successful semantic prediction during SI in both professional and student interpreters. In the predictable condition, both groups demonstrated significantly higher fixation proportions on the target objects than the non-target ones before the TW onset. This effect is supported by significant quadratic interactions between the predictable condition and non-target objects in the GCA results, and also by the significant clusters appearing before the TW onset in the CPA results for the AOI effect in both groups. In contrast, in the unpredictable condition, this anticipatory fixation pattern was not observed. Instead, both groups showed delayed shifts of visual attention from the non-target to the target objects, emerging approximately 600 ms after the TW onset. This temporal lag suggests that participants were unable to predict the targets in the unpredictable condition, likely due to the absence of strong contextual constraints.

7.1.6.1 A dual route of prediction-by-production and prediction-by-association

The above findings provide insights into the cognitive mechanisms underlying prediction during SI. The observed eye movement patterns support a dual-route model involving both prediction-by-production and prediction-by-association. In the predictable condition, both groups fixated more on the targets than on the implausible competitors before the TW onset, despite both objects being semantically compatible with the predictive CV, and the predictive CV was more strongly associated with the implausible competitors than with the targets. This pattern suggests that the interference from the stronger association between the implausible competitors and the CV was mitigated by the low plausibility of the implausible competitors in the global context. This pattern supports the top-down processing of contextual information, i.e., a covert imitation of the speaker to derive the intended message, aligning with accounts of prediction that emphasise the role of the production system in guiding semantic prediction (Federmeier, 2007; Huettig, 2015; Pickering & Gambi, 2018; Pickering & Garrod, 2013). In contrast, in the unpredictable condition, while both the targets and the plausible competitors were compatible with the unpredictive CV and plausible in the context, the plausible

competitors attracted more fixations than the targets, particularly among student interpreters. The increased fixation to the plausible competitors may reflect their stronger association with the unpredictive CV, highlighting a role of prediction-by-association in guiding predictive eye movements.

Another possible explanation of increased fixations on the plausible competitors in the unpredictable condition is integration. Specifically, due to their higher corpus-based frequencies relative to the targets, the plausible competitors might have been more readily accessible and thus more easily integrated into the context, leading to increased fixations (Dahan et al., 2001; Magnuson et al., 2007). However, if word frequency were the primary driver of this effect, we would also expect the implausible competitors—whose corpus-based frequencies were higher than those of the targets—to attract more fixations in the predictable condition. This pattern was not observed. Furthermore, subjective frequency ratings (Table 6-3) were similar across all four word-types, suggesting that differences in word frequency likely did not significantly influence L2 speakers' access to the displayed objects. This suggests that word frequency alone is unlikely to account for the observed fixation patterns. Instead, the stronger association between the plausible competitors and the unpredictive CV may have made them more salient in the visual display, reinforcing a prediction-by-association mechanism.

It is noteworthy that the by-trial CPA revealed a limited contribution of general verb-noun association to predictive eye movements. Specifically, a negative effect on the target fixations in the early stage following CV onset in the professional group. One possible explanation of the negative effect of verb-noun association relates to the extended preview time: participants had approximately 1.5 seconds of preview time before the CV onset. During this period, they may have already attended to the targets and the plausible competitors based on their higher relevance and plausibility in the context. However, professionals, who are assumed to have greater cognitive flexibility and more efficient L2 processing abilities than students (Kaan & Grüter, 2021; Ito & Pickering, 2021), may have adopted a "watchful waiting" strategy, a form of anticipatory readiness for unexpected developments of events (Özkan et al., 2023). Therefore, verb-noun association might have initially redirected their visual attention away from the targets, as they remained strategically open to less expected outcomes, leading to a short-lived negative effect.

The higher relevance and plausibility of the targets and the plausible competitors may also explain the significant effect of cloze probability on these two objects in the early time window following the CV onset (from –1550 ms to –1250 ms) and the absence of significant cluster on the distractors in the professional group. Upon hearing the CV, the cloze probability appeared to influence fixations on these two objects in a faster rate than the other two objects, which were less relevant to the context. The distractors, in particular, were neither plausible in the context nor compatible with the CV, resulting in its low relevance to the context. According to the utility view of prediction proposed by Kuperberg and Jaeger (2016), a rational comprehender aims to maximises the utility of prediction in pursuit of communicative goals. Therefore, unlikely objects, such as the distractors, may be suppressed as quickly as possible to minimise interference and conserve processing resources. This pattern further supports the involvement of a prediction-by-production mechanism.

Overall, the fixation patterns did not align exclusively with either mechanism. Instead, both groups appeared to integrate prediction-by-production and prediction-by-association mechanisms, showing a tendency to fixate on the targets in the predictable condition and plausible competitors in the unpredictable condition. This pattern aligns with the view that top-down analysis and bottom-up activation is interacting to optimise cognitive efficiency.

7.1.6.2 Expertise-related differences in the predictive processing during SI

Overall, the group level analysis revealed similarities in prediction between the interpreting students and professional interpreters. Both groups showed robust anticipatory effects in the predictable condition, as evidenced by pronounced fixation shifts toward the targets relative to the other three objects. However, in the unpredictable condition, increases of fixations on the targets were delayed until the TW onset, with fixations more evenly distributed across the four objects. Consistent with H11, the professionals were more likely to predict the targets than the students, as evidenced by the professionals' higher fixation proportion on the targets in the predictable condition and the significant quadratic interaction between the expertise and the condition in the between-group GCA. The professionals also exhibited faster suppression of unrelated objects, especially in the predictable condition, supported by the significant interactions between expertise and conditions on the linear term across all three non-target objects.

Meanwhile, the professionals demonstrated greater flexibility and dynamism in visual attention patterns and strategic use of prediction cues, whereas the students exhibited more static and

less differentiated eye movement patterns. Specifically, the professionals demonstrated faster and more frequent attention shifts across the four objects in both conditions, as evidenced by the significant estimates on the cubic terms in the between-group GCA and the presence of multiple shorter clusters in the cloze probability effect observed in the by-trial CPA. Even after forming a prediction in the predictable condition, they made more pronounced gaze shifts away from the targets (see Figure 7-4). These results are consistent with the finding of Özkan et al. (2023) that professional interpreters returned their gaze to the baseline following initial predictive fixations. Furthermore, the professionals also exhibited more pronounced disengagement from the implausible competitors and the distractors, particularly under the predictable condition, as indicated by significant linear and quadratic interactions in the between-group GCA. These dynamic patterns of visual attention shift likely reflect that, as the ST unfolds, professional interpreters continuously update their prediction by integrating prior knowledge and new bottom-up inputs (Kuperberg & Jaeger, 2016). Such rapid and efficient adjustment, including the ability to flexibly shift toward relevant targets and suppress distractors, highlights enhanced cognitive flexibility and superior inhibitory control, which is likely attributable to their extensive interpreting experience (Özkan et al., 2023, Lozano-Argüelles et al., 2020, 2021).

In contrast, the students made less pronounced and less frequent visual attention shifts, as indicated by the smaller absolute values of estimates on the quadratic terms in the GCA and the presence of single, longer cluster of the cloze probability effect identified in the by-trial CPA. This aligns with Liu et al. (2022), who found that some participants did not move their eyes at all. This suggests that the extreme cognitive demands of SI may have hindered the students' ability to actively update prediction by integrating prior contextual knowledge and new incoming bottom-up inputs, an ability more readily observed in professionals. Instead, when encountering a potential continuation for the speaker's utterances, especially one that strongly aligned with their internal representation of context, the students appeared more likely to settle for a "good enough" interpretation (Ferreira, 2003; Kuperberg, 2007; Kuperberg & Jaeger, 2016) to conserve cognitive and metabolic resources for cognitive sub-processes. As a result, the students relied more on a shallow processing of broader contextual cues leading to less precise target identification.

The robustness check further reinforced these group differences. For the professionals, only minor changes in significance were observed in the subsampled dataset, suggesting greater consistency in eye movement patterns across trials under each condition and smaller between-

subject variability. In contrast, the student group exhibited a more noticeable reduction in significance in the subsampled dataset, particularly on higher-order time terms. This may indicate greater between-subject variability and less consistent fixation behaviour within the student group. Another possible explanation is a training effect. Specifically, over the course of the experiment, some student participants may have gradually learned to search for the potential continuation, regardless of the sentence condition. While this training effect does not invalidate the observed prediction effects, it may suggest that the students relied more heavily on global contextual cues without fully processing the predictive potential of the CV, again reflecting a more shallow, less dynamic use of predictive strategies.

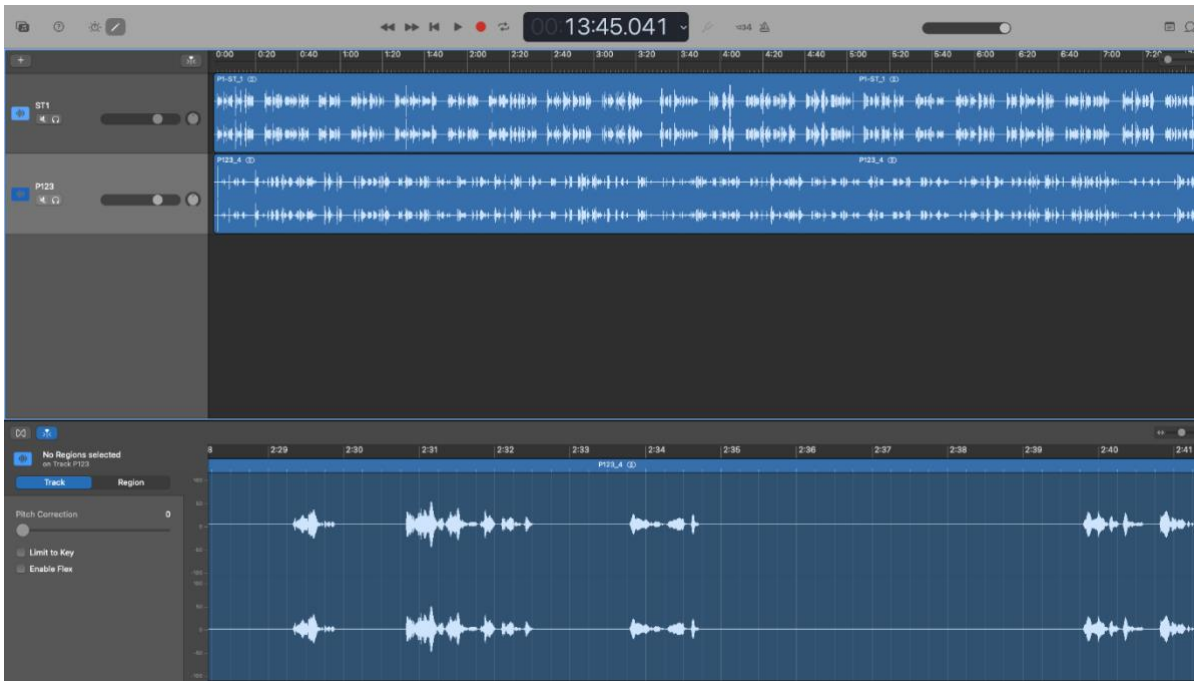
7.2 Ear-voice span data

7.2.1 Data preparation and cleaning

To prepare the EVS data, each participant's interpreting rendition for all four paragraphs was merged into a single audio file, which included both the experimental and filler sentences in their entirety. Data from all sixty-six participants were initially examined, including those excluded from the eye-tracking analysis due to low data quality. Data from one student participant were later excluded from the EVS analysis due to an audio recording failure[6]. The recordings were processed using GarageBand on MacOS, an audio editing software selected for its ability to support multi-track editing and its suitability for fine-grained temporal analysis. Specifically, it allows for precise visual inspection of sound waveforms through adjustable zoom functions, facilitating the accurate manual annotation of onset and offset points at a millisecond level. In this study, a temporal resolution of approximately 40 ms (0.04 s) was adopted during annotation. This level of precision was deemed sufficient given that EVS is typically analysed using a basic time unit of 0.1 seconds (Timarová et al., 2011), enabling reliable detection of subtle timing differences relevant to interpreting latency. A representative screen shot with varying zoom units is provided in Figure 7-6.

[6] This participant's eye-tracking data were also excluded from analysis.

Figure 7-6. The interface of GarageBand



For each of the twenty experimental sentences, four critical time points were manually annotated in both the interpreting and source recordings: sentence onset, CV onset, TW onset, and sentence offset.

1) Sentence onset was defined as the time point when the participant started producing the interpreted equivalent of the current experimental sentence in the target language, following the cessation of output related to the preceding sentence. Crucially, this point did not necessarily correspond to the interpreted equivalent of the first word in the source sentence, as interpreters often restructure or reorder content due to individual strategies, cognitive load, or syntactic differences between the source and target languages (Timarová et al., 2011). The judgement was based on the semantic boundary between sentences, with the first semantically relevant word marked as the onset (Barik, 1973). Filler words were included in this judgement, but disfluencies or non-lexical vocalisations (such as “em”) were excluded.

2) CV onset referred to the first appearance of an acceptable interpreted equivalent of the CV. As interpreters may repeat, revise, or reformulate their outputs to aid fluency, particularly under cognitive strain (Gile, 2009; Pöchhacker, 2004), the CV equivalent could appear multiple times. Only the first occurrence was annotated, as it most closely reflects the moment at which comprehension, lexical retrieval, and overt production of

the CV equivalent were first completed. Subsequent repetitions were disregarded, as they are likely to result from later-stage production processes such as self-monitoring, discourse planning, or hesitation (Levelt, 1989; Mazza, Turatto, & Caramazza, 2009).

3) TW onset referred to the first appearance of the interpreted equivalent of the TW. This was identified using the same criteria applied to the CV onset.

4) Sentence offset was defined as the time point at which the final interpreted segment of the current experimental sentence concluded, prior to the onset of interpretation of the following sentence. Similar to the onset, the offset did not necessarily coincide with the interpreted equivalent of the last word in the source sentence, as the interpreted sentence did not always map directly onto the source sentence structure. It was determined based on the conclusion of semantically relevant interpreted content before the participant transitioned to the next sentence. In cases where output from the next sentence was integrated into the same utterance, only the segment corresponding to the current experimental sentence was considered in the annotation. This approach ensured consistency in delimiting the interpreted segments across participants.

For any time point where no clear equivalent was produced, a missing value was recorded for that specific time point within the trial. Annotation was conducted twice by the researcher with a three-month interval. The results of the two annotations were highly consistent and thus should be considered valid and accurate. The second annotation was used for final analysis.

The same four time points were also annotated in the ST recordings. EVS values were then computed by subtracting the corresponding time stamps in the source recording from those in the participants' interpreting renditions, yielding four measures per trial: sentence onset EVS, CV-EVS, TW-EVS, and sentence offset EVS. In addition to these EVS, two further temporal measures were derived. First, the time interval between the onset of the interpreted equivalent of the CV and that of the TW, termed the CT-span, was calculated by subtracting the CV onset from the TW onset. This measure captures the temporal gap between the processing and production of the verb and the subsequent target word, potentially reflecting the influence of predictive processing. Given that the interpreted equivalent of the TW occasionally appeared before that of the CV, the CT-span was defined as the absolute value of this difference. Second, the duration of the interpreted equivalent of each experimental sentence, referred to as the sentence-span, was computed by subtracting the sentence onset from the sentence offset, providing an estimate of overall production length for each trial.

7.2.2 By-group analysis for effects of predictability and trial-specific features

A series of statistical tests was conducted to examine each expertise group independently. To assess the effect of condition, paired-sample t-tests were first performed to compare the six temporal measures between the predictable and the unpredictable conditions. A further by-trial analysis was conducted using LME models to investigate how trial-specific linguistic features contributed to the variations of the temporal measures. These features were categorised into three levels: the lexical level, including CV frequency and TW frequency; the sentence level, including cloze probability, the association strength between the CV and the TW (verb-noun association), and sentence length; the discourse level, including preceding sentence length and within-paragraph position. Although verb-noun association reflects lexical co-occurrence, it was excluded from the lexical level due to its significant negative correlation with CV frequency ($r = -.350$, 95% CI [−.597, −.043], $t(38) = -2.305$, $p = .027$), in order to control for shared variance. Instead, it was classified at the sentence level, as it reflects relational semantics operating at the phrasal or the sentential level, and engages prediction mechanism potentially utilised during SI.

The discourse-level features captured the influence of the surrounding text (co-text), especially the preceding sentence, on the processing of the experimental sentence, given that experimental sentences were embedded within coherent discourse. It is worth noting that, within each paragraph, every two to four sentences were grouped into a segment, with a five-second interval inserted between segments. When an experimental sentence appeared as the first sentence of a segment, i.e., following the five-second interval, it was assumed that the interpreting of the final sentence in the preceding segment was typically completed during that interval. This interval presumably allowed participants to offload some memory and processing demands and cognitively reset before continuing. As the influence of that preceding sentence on the experimental sentence was likely minimal in such case, its length was coded as zero to approximate this presumed minimal effect. Additionally, previous research suggests that EVS may correlate with the location of the measurement point in the text (Timarová et al., 2011), with shorter EVSs observed earlier in the discourse. To account for this, the position of each experimental sentence within its paragraph was included as a predictor. Within-paragraph position was indexed by the word count of the text preceding each experimental sentence in that paragraph.

LME models were fitted using the *lmer()* function from the *lme4* package in R. For each of the six temporal measures, three LME models were constructed, one for each linguistic level, with the temporal measure as the dependent variable and a by-participant random intercept. The lexical-level model included CV frequency and TW frequency as the fixed effects; the sentence-level model included cloze probability, verb-noun association, and sentence length; and the discourse-level model included preceding sentence length and within-paragraph position. Despite a significant positive correlation between cloze probability and verb-noun association ($r = 0.354$, $t(38) = 2.335$, $p = .025$), the two features were included simultaneously in the sentence-level model to examine the combined influence of distinct prediction mechanisms. To address potential multicollinearity that could inflate standard errors, variance inflation factor (VIF) was computed using the *vif()* function from the *car* package in R for each sentence-level model. All VIF values were below 1.5, well within the commonly accepted threshold of 2.5, indicating limited multicollinearity.

Unlike previous studies using independent sentences as stimuli, which allowed flexible sentence order to counterbalance the potential effect of stimuli order, the present study used four paragraphs that were arranged in a fixed sequence following a coherent narrative thread. Because paragraph order represents a higher-level grouping factor that nests other trial features, including it as a fixed effect in the LME models could absorb shared variance and obscure the effects of other meaningful predictors. Therefore, the effect of paragraph order was examined separately using one-way ANOVA.

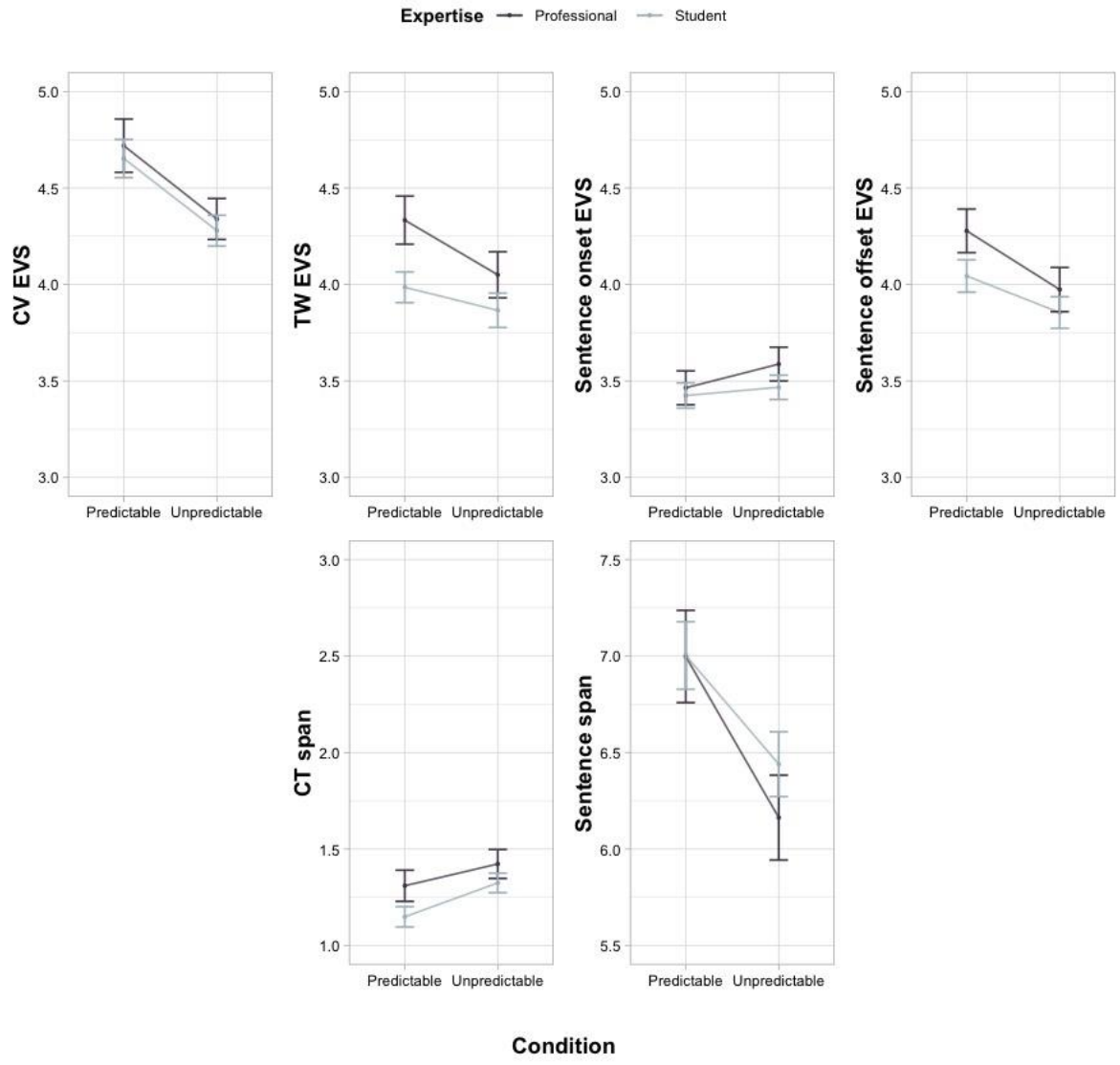
A robustness check was conducted by rerunning these tests on a subsampled dataset following the procedure described in Section 6.3.2. The results of the robustness check are reported alongside the main results for each group in the respective results sections.

7.2.2.1 Results for the professional interpreters

In the professional group, paired-sample *t*-tests revealed a significant effect of condition on CV-EVS, with longer delays observed in the predictable condition (Mean = 4.720 s, $SD = 1.723$ s) than in the unpredictable condition (Mean = 4.340 s, $SD = 1.421$ s), $t(301) = 2.183$, $p = .030$ (see Figure 7-7). TW-EVS was slightly higher in the predictable condition (Mean = 4.333 s, $SD = 1.664$ s) than in the unpredictable condition (Mean = 4.050 s, $SD = 1.607$ s), but the difference did not reach statistical significance, $t(356) = 1.640$, $p = .102$. At the sentence level, sentence onset EVS did not differ significantly between conditions (predictable: Mean = 3.465 s, $SD = 1.246$ s; unpredictable: Mean = 3.588 s, $SD = 1.277$ s), $t(412) = -0.991$, $p = .322$.

Sentence offset EVS showed a marginal trend towards higher values in the predictable condition, but the difference was not statistically significant (predictable: Mean = 4.278 s, *SD* = 1.556 s; unpredictable: Mean = 3.974 s, *SD* = 1.644 s), $t(394) = 1.893$, $p = .059$. No significant difference was found for CT-span (predictable: Mean = 1.310 s, *SD* = 0.967 s; unpredictable: Mean = 1.423 s, *SD* = 0.935 s), $t(290) = -1.024$, $p = .307$. Finally, sentence-span differed significantly between conditions, with the predictable condition yielding longer durations (Mean = 6.998 s, *SD* = 2.683 s) than the unpredictable condition (Mean = 6.163 s, *SD* = 2.584 s), $t(258) = 2.571$, $p = .011$. The robustness check showed consistent results in the subsampled dataset, except a slight increase of significance for sentence-span (from $p = .011$ to $p = .006$, see Appendix 5).

Figure 7-7. By-group means of EVS measures in each condition.



#Notes: The error bars represent ±1 SE.

LME models revealed several significant effects on the EVS measures. For CV-EVS, at the lexical level, CV frequency showed negative contributions, indicating that more frequent CVs facilitated faster production. At the sentence level, cloze probability exhibited positive effects, indicating that more constraining contexts were linked to longer delays in the production of CV equivalents. Neither verb-noun association nor the experimental sentence length made a significant contribution to CV-EVS. At the discourse level, the within-paragraph position of the experimental sentence had a significant positive effect, with CV-EVS increasing as the sentence appeared later in the paragraph. The robustness check revealed a largely consistent pattern, with one exception: the effect of CV frequency became marginally significant (from $p = .038$ to $p = .052$, see Appendix 6). This reduction in significance suggests that the facilitation effect of higher CV frequency on production timing may have been limited or less stable among the professionals.

In contrast, TW-EVS had no significant associations with any one of the frequency measures or with cloze probability. It was negatively associated with verb-noun associations. Additionally, TW-EVS was positively associated with the lengths of the preceding sentence, suggesting that the production of TW equivalents was delayed when the experimental sentence was preceded by a longer sentence. These findings were confirmed by the robustness check (see Appendix 6).

Table 7-7. LME models for CV-EVS and TW-EVS in the professional group

CV-EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	4.891	0.979	325	4.994	< .001 ***
CV frequency	−0.274	0.131	318	−2.088	.038 *
TW frequency	0.219	0.186	316	1.183	.238
Sentence					
(Intercept)	3.416	0.371	281	9.214	< .001 ***
Cloze probability	1.467	0.379	315	3.876	< .001 ***
Verb-noun association	−0.582	0.564	323	−1.032	.303
Sentence length	0.006	0.021	314	0.279	.781
Discourse					
(Intercept)	3.692	0.216	91	17.109	< .001 ***

Preceding sentence length	0.002	0.011	313	0.191	.849
Within-paragraph position	0.005	0.001	316	5.513	< .001 ***
TW-EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	3.482	1.009	348	3.452	< .001 ***
CV frequency	−0.040	0.141	339	−0.286	.775
TW frequency	0.221	0.184	337	1.201	.231
Sentence					
(Intercept)	3.727	0.390	311	9.568	< .001 ***
Cloze probability	0.412	0.438	337	0.941	.348
Verb-noun association	−2.014	0.569	346	−3.540	< .001 ***
Sentence length	0.029	0.021	337	1.402	.162
Discourse					
(Intercept)	3.337	0.220	97	15.172	< .001 ***
Preceding sentence length	0.027	0.011	338	2.520	.012 *
Within-paragraph position	0.004	0.001	338	4.147	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Sentence onset EVS were significantly associated with all trial-specific features, except for cloze probability. Specifically, CV frequency, verb-noun association, and sentence length exhibited significant negative effects, suggesting that more frequent CV, stronger association links between the CV and the TW, and a longer sentence were associated with shorter delay in the initiation of sentence production. In contrast, TW frequency, preceding sentence length, and within-paragraph position had significant positive effects, indicating that sentence initiation may be delayed by more frequent TW, longer preceding sentence, or the later occurrence of the experimental sentence within the paragraph. The robustness check revealed two key changes: the effects of CV frequency and verb-noun association became non-significant in the subsampled dataset (see Appendix 6). These reductions were particularly related to the CV, possibly reflecting greater variability in verb-related features across trials, which made these effects more sensitive to subsampling.

Sentence offset EVS appeared less sensitive to lexical-level features, with no significant effect of either frequency measure. Instead, sentence-level features had more pronounced effects on

sentence offset EVS, with a positive effect of cloze probability and a negative effect of verb-noun association. Sentence offset EVS was less affected by preceding sentence, as indicated by non-significant effect of preceding sentence length but increased as the experimental sentence occurred later within the paragraph. These results were confirmed by the robustness check (see Appendix 6).

Table 7-8. LME models for sentence onset EVS and sentence offset EVS in the professional group

Sentence onset EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	3.059	0.709	407	4.313	< .001 ***
CV frequency	−0.202	0.095	338	−2.132	.034*
TW frequency	0.344	0.129	336	2.663	.008**
Sentence					
(Intercept)	4.047	0.273	291	14.822	< .001 ***
Cloze probability	0.351	0.276	392	1.272	.204
Verb-noun association	−0.974	0.414	398	−2.351	.019 *
Sentence length	−0.052	0.014	390	−3.633	< .001 ***
Discourse					
(Intercept)	2.841	0.166	79	17.165	< .001 ***
Preceding sentence length	0.039	0.008	390	5.079	< .001 ***
Within-paragraph position	0.002	0.001	391	3.258	.001 **
Sentence offset EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	4.063	0.908	391	4.476	< .001 ***
CV frequency	0.008	0.120	374	0.069	.945
TW frequency	0.006	0.169	373	0.035	.972
Sentence					
(Intercept)	3.689	0.359	244	10.266	< .001 ***
Cloze probability	0.842	0.349	373	2.409	.016 *
Verb-noun association	−1.935	0.518	379	−3.733	< .001 ***
Sentence length	0.001	0.018	372	0.037	.971
Discourse					
(Intercept)	3.325	0.226	63	14.738	< .001 ***

Preceding sentence length	0.004	0.010	373	0.362	.718
Within-paragraph position	0.005	0.001	373	5.664	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

For the duration measures, the CT-span was sensitive to lexical-level features, with a positive association with CV frequency and a negative association with TW frequency. Verb-noun association also demonstrated significant negative effect on CT-span, indicating that higher-frequency TW and/or stronger lexical associations between the CV and the TW facilitated more rapid transitions between them, whereas higher-frequency CV contributed to extended transitions. Similar to the EVS measures, CT-span was positively associated with within-paragraph position, indicating that the later-occurrence of the experimental sentence within the paragraph impeded the transition between the CV and the TW.

Sentence-span demonstrated significant association with all linguistic features. Similar to CT-span, sentence-span also exhibited a positive association with CV frequency and a negative association with TW frequency. At the sentence level, cloze probability and sentence length had positive effects, whereas verb-noun association had negative effects. At the discourse level, sentence-span decreased with a longer preceding sentence but increased with within-paragraph position. The robustness check confirmed these findings and additionally revealed a significant negative effect of cloze probability on CT-span observed in the subsampled dataset (see Appendix 6).

Table 7-9. LME models for CT-span and sentence-span in the professional group

CT-span					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	1.737	0.654	280	2.659	.008 **
CV frequency	0.213	0.092	283	2.307	.022 *
TW frequency	−0.329	0.122	279	−2.706	.007 **
Sentence					
(Intercept)	1.726	0.241	287	7.176	< .001 ***
Cloze probability	−0.467	0.289	277	−1.618	.107
Verb-noun association	−1.267	0.366	291	−3.465	< .001 ***
Sentence length	0.011	0.014	279	0.781	.435
Discourse					

(Intercept)	1.163	0.127	158	9.182	< .001 ***
Preceding sentence length	−0.011	0.007	278	−1.508	.133
Within-paragraph position	0.002	0.001	282	2.923	.004 **
Sentence-span					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	8.419	1.849	261	4.552	< .001 ***
CV frequency	1.287	0.262	261	4.909	< .001 ***
TW frequency	−1.882	0.344	261	−5.471	< .001 ***
Sentence					
(Intercept)	−0.745	0.442	255	−1.686	.093
Cloze probability	1.119	0.531	246	2.109	.036 *
Verb-noun association	−1.782	0.669	258	−2.664	.008 **
Sentence length	0.531	0.026	245	20.251	< .001 ***
Discourse					
(Intercept)	6.444	0.312	173	20.657	< .001 ***
Preceding sentence length	−0.175	0.019	247	−9.328	< .001 ***
Within-paragraph position	0.011	0.002	253	5.978	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

One-way ANOVAs showed significant effects of paragraph order on all four EVS measures as well as on sentence-span (see Table 7-10). Post hoc comparisons using Tukey's HSD tests indicated consistent patterns for CV-EVS, TW-EVS, sentence offset EVS, and sentence-span, with significantly higher values in Paragraph 4 compared to Paragraph 1, 2, and 3 ($ps < .001$), and no significant differences among the first three paragraphs. Sentence onset EVS showed a slightly different pattern, with Paragraph 4 significantly higher than the first two paragraphs ($ps < .01$), while Paragraph 3 did not differ significantly from the others. In contrast, paragraph order had no significant effect on CT-span, suggesting stable transition intervals between the CV and the TW across task stages. The robustness check also confirmed these findings (see Appendix 7).

Table 7-10. One-way ANOVA of the paragraph effect for the professional group

	df	***F***	***MSE***	***p***
CV-EVS	(3, 331)	24.84	2.05	< .001 ***

TW-EVS	(3, 355)	25.46	2.23	< .001 ***
Sentence onset EVS	(3, 410)	7.93	1.52	< .001 ***
Sentence offset EVS	(3, 392)	12.94	2.37	< .001 ***
CT-span	(3, 294)	1.03	0.90	.381
Sentence-span	(3, 260)	9.24	6.47	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Overall, the professionals exhibited a tendency towards longer temporal measures in the predictable condition compared to the unpredictable condition. At the lexical level, CV frequency had negative effects on CV-EVS, and positive effects on CT-span and sentence-span, suggesting that more frequent CVs were associated with earlier production of CV equivalents, but also with extended CV-TW intervals and overall sentence durations. In contrast, TW frequency showed an opposite pattern, with higher-frequency TWs linked to longer delays in sentence initiation but shortened CV-TW intervals and sentence durations. At the sentence level, cloze probability was positively associated with CV-EVS, sentence offset EVS, and sentence-span, while verb-noun association was negatively associated with all temporal measures except CV-EVS and sentence onset EVS. These findings suggest that more constraining contexts contributed to delayed CV production and longer sentence durations, whereas stronger verb-noun association generally accelerated and shortened production. Sentence length was negatively associated with sentence onset EVS and positively associated with sentence-span, indicating that when interpreting longer experimental sentences, the professionals tended to initiate earlier and proportionally extend sentence production.

At the discourse level, preceding sentence length showed positive effects on TW-EVS and sentence onset EVS, and a negative effect on sentence-span. This suggests that when the experimental sentence was preceded by a longer sentence, the professionals tended to delay sentence initiation and the production of TW equivalents, but reduced the overall duration of sentence production. Within-paragraph position exhibited positive effects on all six temporal measures, indicating increasing lags between input and output as the discourse progressed. Similarly, the order of the paragraph within the overall SI task were positively associated with all temporal measures except CT-span, significantly longer lags were observed in Paragraph 4.

7.2.2.2 Results for the interpreting students

The student group demonstrated patterns largely consistent with those observed in the professional group across all four EVS measures and the sentence-span measure (see Figure 7-7). Specifically, the students also showed significantly longer CV-EVS in the predictable condition (Mean = 4.653, *SD* = 1.604 s) than in the unpredictable condition (Mean = 4.279 s, *SD* = 1.426 s), $t(529) = 2.944$, $p = .003$. A significant between-condition difference was also observed for sentence-span, with longer sentence durations in the predictable condition (Mean = 7.003 s, *SD* = 2.529 s) compared to the unpredictable condition (Mean = 6.440 s, *SD* = 2.597 s), $t(441) = 2.325$, $p = .021$. No significant effects of condition were found for TW-EVS (predictable: Mean = 3.986 s, *SD* = 1.439 s; unpredictable: Mean = 3.866 s, *SD* = 1.596 s), $t(637) = 1.001$, $p = .317$; sentence onset EVS (predictable: Mean = 3.425 s, *SD* = 1.289 s; unpredictable: Mean = 3.467 s, *SD* = 1.237 s), $t(769) = -0.470$, $p = .638$; or sentence offset EVS, (predictable: Mean = 4.044 s, *SD* = 1.559 s; unpredictable: Mean = 3.855 s, *SD* = 1.537 s), $t(700) = 1.621$, $p = .106$. The most distinct divergence from the professional group emerged in CT-span: while the professionals showed no significant difference across conditions, the students exhibited significantly faster transitions between the CV and the TW in the predictable condition (Mean = 1.149 s, *SD* = 0.813 s) than in the unpredictable condition (Mean = 1.325, *SD* = 0.853 s), $t(514) = -2.410$, $p = .016$. The robustness check showed completely consistent results for the student group in the subsampled dataset (see Appendix 5).

The LME models for the student group revealed broader effects of trial-specific features on both CV-EVS and TW-EVS. For CV-EVS, in addition to the significant predictors observed in the professional group, there was a significant negative effect of verb-noun association strength, suggesting that a stronger lexical link between the CV and the TW facilitated earlier production of CV equivalents. For TW-EVS, like the professionals, the students exhibited negative effect of verb-noun association and positive effects of preceding sentence length and within-paragraph position. Additionally, TW-EVS in the students was significantly modulated by TW frequency and experimental sentence length, with more frequent TWs and longer sentence length associated with longer delays in producing TW equivalents. The patterns observed for sentence onset EVS and sentence offset EVS in the student group were largely consistent with those of the professional group, with one exception: while no significant effect of preceding sentence length was found on sentence offset EVS in the professionals, the students showed a marginally significant positive effect ($p = .076$). This additional effect suggests that the students may have delayed concluding the experimental sentence that

followed longer preceding sentences. The robustness check confirmed these key findings and additionally revealed marginally significant positive effect of TW frequency on both CV-EVS and sentence offset EVS in the subsampled dataset, which were absent in the original dataset (see Appendix 6). The marginally significant effect of preceding sentence length on sentence offset EVS in the original dataset also became statistically significant in the subsampled dataset.

Table 7-11. LME models for the EVS measures in the student group

CV-EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	4.957	0.739	561	6.707	< .001 ***
CV frequency	−0.303	0.098	554	−3.088	.002 **
TW frequency	0.222	0.139	549	1.597	.111
Sentence					
(Intercept)	3.796	0.277	552	13.728	< .001 ***
Cloze probability	1.528	0.286	551	5.345	< .001 ***
Verb-noun association	−2.037	0.402	572	−5.070	< .001 ***
Sentence length	−0.016	0.016	550	−1.020	.308
Discourse					
(Intercept)	3.677	0.149	257	24.694	< .001 ***
Preceding sentence length	0.003	0.008	549	0.429	.668
Within-paragraph position	0.005	0.001	552	6.844	< .001 ***
TW-EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	2.003	0.693	628	2.889	.004 **
CV frequency	0.131	0.096	610	1.373	.170
TW frequency	0.329	0.129	611	2.544	.011 *
Sentence					
(Intercept)	3.388	0.274	608	12.377	< .001 ***
Cloze probability	0.219	0.308	608	0.711	.478
Verb-noun association	−2.016	0.369	631	−5.461	< .001 ***
Sentence length	0.049	0.015	610	3.363	.001 **
Discourse					
(Intercept)	3.000	0.147	190	20.441	< .001 ***

Preceding sentence length	0.025	0.007	609	3.337	.001 ***
Within-paragraph position	0.005	0.001	610	7.349	< .001 ***
Sentence onset EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	3.378	0.493	765	6.848	< .001 ***
CV frequency	−0.375	0.065	731	−5.750	< .001 ***
TW frequency	0.443	0.090	729	4.909	< .001 ***
Sentence					
(Intercept)	4.009	0.210	529	19.129	< .001 ***
Cloze probability	0.123	0.206	732	0.600	.549
Verb-noun association	−0.746	0.285	742	−2.618	.009 **
Sentence length	−0.043	0.010	728	−4.101	< .001 ***
Discourse					
(Intercept)	2.893	0.127	130	22.862	< .001 ***
Preceding sentence length	0.031	0.006	728	5.633	< .001 ***
Within-paragraph position	0.002	0.001	727	3.909	< .001 ***
Sentence offset EVS					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	3.277	0.673	688	4.871	< .001 ***
CV frequency	0.027	0.089	667	0.302	.763
TW frequency	0.136	0.128	664	1.065	.287
Sentence					
(Intercept)	3.655	0.265	592	13.806	< .001 ***
Cloze probability	0.678	0.267	666	2.539	.011 *
Verb-noun association	−2.456	0.360	680	−6.823	< .001 ***
Sentence length	0.006	0.014	663	0.450	.653
Discourse					
(Intercept)	2.945	0.149	165	19.756	< .001 ***
Preceding sentence length	0.013	0.007	662	1.777	.076 †
Within-paragraph position	0.006	0.001	662	9.247	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Regarding the CT-span, the students exhibited a pattern different from the professionals. While CV frequency exerted a significant positive effect on CT-span in the professionals, it only exerted a marginally significant positive effect in the students (p = .052). By contrast, there was no significant effect of cloze probability in the professionals, whereas a marginally significant negative effect was found in the students (p = .058). The patterns of sentence-span were consistent in both groups, with significant negative associations with TW frequency, verb-noun association, and preceding sentence length, and significant positive associations with CV frequency, cloze probability, sentence length and within-paragraph position. The robustness check showed largely consistent results, especially for sentence-span. However, the marginally significant effect of CV frequency on CT-span became non-significant in the subsampled dataset, indicating a more limited effect of CV frequency on CT-span in the students (see Appendix 6). Similar to the professionals, the subsampled dataset also revealed an additional significant effect of cloze probability on CT-span. This may be due to the subsampling procedure reducing the number of trials where verb-noun association was extremely strong that overshadowed or diminished the detectability of the cloze probability effect. With those trials reduced, previously non-significant or marginally significant effects of cloze probability could reach statistical significance.

Table 7-12. LME models for CT-span and sentence-span in the student group

CT-span					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	1.752	0.442	499	3.962	< .001 ***
CV frequency	0.120	0.062	493	1.946	.052 †
TW frequency	−0.261	0.082	494	−3.205	.001 **
Sentence					
(Intercept)	1.429	0.167	518	8.559	< .001 ***
Cloze probability	−0.379	0.200	489	−1.898	.058 †
Verb-noun association	−0.731	0.239	519	−3.064	.002 **
Sentence length	0.015	0.010	498	1.473	.141
Discourse					
(Intercept)	1.060	0.084	318	12.586	< .001 ***
Preceding sentence length	−0.007	0.005	491	−1.403	.161
Within-paragraph position	0.002	0.0005	501	3.364	< .001 ***

Sentence-span					
Predictor	**Estimate**	**Std. Error**	***df***	***t***	***p***
Lexical					
(Intercept)	5.402	1.419	446	3.806	< .001 ***
CV frequency	1.455	0.202	446	7.217	< .001 ***
TW frequency	−1.312	0.258	446	−5.085	< .001 ***
Sentence					
(Intercept)	−0.635	0.342	444	−1.854	.064 †
Cloze probability	1.400	0.402	414	3.485	< .001 ***
Verb-noun association	−1.900	0.507	443	−3.746	< .001 ***
Sentence length	0.508	0.020	423	25.430	< .001 ***
Discourse					
(Intercept)	6.144	0.227	295	27.058	< .001 ***
Preceding sentence length	−0.161	0.014	424	−11.783	< .001 ***
Within-paragraph position	0.012	0.001	435	9.262	< .001 ***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

As for the effect of paragraph order, the students' temporal measures (except for CT-span) generally increased as the task progressed, mirroring the pattern observed in the professionals. However, Tukey's HSD tests revealed a slightly more complicated pattern in the students. While Paragraph 4 elicited significantly longer EVSs and sentence durations than the three preceding paragraphs (*ps* < .001), Paragraph 3 also yielded significantly higher values than Paragraphs 1 and 2 across all four EVS measures. In contrast to the professionals, who demonstrated consistent CT-span across four paragraphs, the students showed a significantly shorter CT-span in Paragraph 3 than Paragraph 2 ($p = .022$). The robustness check also confirmed these findings (see Appendix 7).

Table 7-13. One-way ANOVA of the paragraph effect for the student group

	df	*F*	*MSE*	*p*
CV-EVS	(3, 580)	61.93	1.76	< .001***
TW-EVS	(3, 644)	84.28	1.66	< .001***
Sentence onset EVS	(3, 768)	24.08	1.46	< .001***
Sentence offset EVS	(3, 699)	50.60	1.98	< .001***
CT-span	(3, 519)	3.21	0.69	.023 *
Sentence-span	(3, 445)	16.92	6.01	< .001***

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Overall, the students exhibited a pattern largely consistent with that of the professionals, with a few notable differences. At the lexical levels, the significant positive effect of CV frequency on CT-span observed in the professionals was only marginally significant in the students. TW frequency, in contrast, showed an additional positive effect on TW-EVS, suggesting a delay in producing equivalents of more frequent TWs. At the sentence level, cloze probability showed a marginally significant negative effect on CT-span, whereas verb-noun association exhibited a significant negative effect on CV-EVS, indicating that the students produced CV equivalents earlier when the CV and the TW had a stronger lexical association. Sentence length had an additional positive effect on TW-EVS, suggesting delayed production of TW equivalents in longer sentences. At the discourse level, preceding sentence length showed a marginally significant positive effect on sentence offset EVS, indicating that the students tended to conclude sentence production later when it followed a longer sentence. Within-paragraph position and paragraph order showed similar patterns to those observed in the professionals, with two notable exceptions. First, paragraph order was also positively associated with CT-span, which was absent in the professionals. Second, unlike the professionals, who exhibited extended temporal measures only in Paragraph 4, the students showed these extensions as early as in Paragraph 3.

7.2.3 Between-group comparisons

To investigate differences between the professionals and the students, independent-sample t-tests were conducted on six temporal measures (see Table 7-14). Analyses first examined overall group differences, followed by condition-specific comparisons to assess how each group responded to varying levels of predictability in the ST. The results revealed a significant between-group difference in TW-EVS, with the professionals exhibiting longer lags in the

production of TW equivalent than the students. While the professionals also showed higher values in CT-span and sentence offset EVS, these differences were only marginally significant. Condition-specific analyses indicated that these between-group differences were largely attributable to divergent processing of predictive cues. Specifically, significant group differences emerged under the predictable condition, whereas no significant differences were observed in the unpredictable condition across any measure. The robustness check revealed mostly consistent results in the subsampled dataset, except that the marginally significant between-group differences observed in the original dataset (e.g., sentence offset EVS and CT-span) became non-significant in the subsampled dataset (see Appendix 8).

Table 7-14. Between-group comparisons for the EVS and the duration measures

Overall					
	Professionals	**Student**	***t***	***df***	***p***
CV-EVS	4.517 (*SD* = 1.578)	4.448 (*SD* = 1.519)	0.646	674	0.519
TW-EVS	4.191 (*SD* = 1.639)	3.926 (*SD* = 1.519)	2.515	693	0.012 *
Sentence onset EVS	3.528 (*SD* = 1.262)	3.446 (*SD* = 1.262)	1.061	845	0.289
Sentence offset EVS	4.120 (*SD* = 1.697)	3.948 (*SD* = 1.550)	1.718	794	0.086 †
CT-span	1.369 (*SD* = 0.950)	1.244 (*SD* = 0.839)	1.883	557	0.060 †
Sentence-span	6.562 (*SD* = 2.659)	6.702 (*SD* = 2.578)	–0.689	538	0.491
Predictable					
CV-EVS	4.720 (*SD* = 1.723)	4.653 (*SD* = 1.604)	0.391	307	0.696
TW-EVS	4.333 (*SD* = 1.664)	3.986 (*SD* = 1.439)	2.349	322	0.019 *
Sentence onset EVS	3.465 (*SD* = 1.246)	3.425 (*SD* = 1.289)	0.365	420	0.715
Sentence offset EVS	4.278 (*SD* = 1.556)	4.044 (*SD* = 1.559)	1.666	389	0.096 †
CT-span	1.310 (*SD* = 0.967)	1.149 (*SD* = 0.813)	1.670	260	0.096 †
Sentence-span	6.998 (*SD* = 2.683)	7.003 (*SD* = 2.529)	–0.017	251	0.986
Unpredictable					
CV-EVS	4.340 (*SD* = 1.421)	4.279 ((*SD* = 1.426)	0.455	369	0.650
TW-EVS	4.050 (*SD* = 1.607)	3.866 (*SD* = 1.596)	1.236	371	0.217
Sentence onset EVS	3.588 (*SD* = 1.277)	3.467 (*SD* = 1.237)	1.113	423	0.267
Sentence offset EVS	3.974 (*SD* = 1.644)	3.855 (*SD* = 1.537)	0.846	405	0.398
CT-span	1.423 (*SD* = 0.935)	1.325 (*SD* = 0.853)	1.082	293	0.280
Sentence-span	6.163 (*SD* = 2.584)	6.440 (*SD* = 2.597)	–1.001	287	0.318

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

7.2.4 Interim discussion

Overall, both the professionals and the students demonstrated distinct temporal patterns between the predictable and the unpredictable conditions, with the predictable condition generally associated with longer EVSs and extended sentence durations. There was no significant between-group difference in sentence onset EVS, CV-EVS, sentence offset EVS, or sentence-span, suggesting that both groups initiated and concluded the sentence production at similar timings, with comparable timelines for CV production. However, the professionals maintained a stable CT-span across both conditions, whereas the students transitioned between the CV and the TW more quickly in the predictable condition. Additionally, in the predictable condition, the professionals showed significantly longer TW-EVS. This pattern indicates that the between-group differences in the temporal dynamics of interpreting outputs emerged primarily in the later part of the sentence, particularly after the CV. The emergence of between-group difference specifically in the predictable condition and in the later sentence segment suggests that the two groups may have adopted distinct strategies for processing predictive cues. These findings challenge the conventional expectation that predictive processing facilitates interpreting performance by reducing lags between input and output. The observed variations in the temporal measures can be further understood through the contribution of trial-specific features at each level.

7.2.4.1 Lexical-level features

At the lexical level, both groups exhibited negative effects of CV frequency on CV-EVS. A potential explanation for these negative effects is the greater accessibility of more frequent CVs, which may enhance comprehension and facilitate integration into the mental representation of the context. Moreover, such CVs may be more readily selected during early planning due to their greater availability in interpreters' mental lexicons and higher flexibility to fit in a range of syntactic and semantic contexts (Kempen & Harbusch, 2019; Wen & van Heuven, 2017). Given the pivotal role of verbs in shaping syntactic structures (Antón-Méndez, 2020; Vigliocco et al., 2011), a CV equivalent may be identified early to anchor sentence structure.

Additional insights into how CV frequency influenced early production processes come from the robustness check using the subsampled dataset. In the original dataset, both groups showed significant a negative effect of CV frequency on sentence onset EVS. However, in the subsampled dataset, this effect remained significant for the students but became non-significant

for the professionals. This divergence suggests a possible expertise-related difference in the role of lexical accessibility at sentence initiation. Specifically, while more frequent CVs may consistently facilitate lexical access, the professional's decision on when to initiate sentence production appeared to rely more on a range of strategic or higher-level planning factors (e.g., discourse structure, anticipation of upcoming elements, or speaker pacing). This strategic flexibility in when to launch the utterance may introduce more noise in sentence onset timing and weaken the statistical detectability of lexical frequency effects when trial numbers are reduced. In contrast, student interpreters, who are more likely to rely on lexical accessibility during early sentence planning, showed a more stable effect of CV frequency on sentence onset EVS, even in the subsampled dataset. Their initiation of sentence production may be more related to how easily the CV can be accessed and retrieved, resulting in a clearer and more consistent association between CV frequency and sentence onset timing.

However, despite these advantages at the initial production stage, higher-frequency CVs were associated with both longer CV-TW intervals and longer sentence durations in the professionals. This may be attributed to the more possible semantic continuations or syntactic structures supported by more frequent CVs (Ellis, O'Donnell, & Römer, 2014; Kempen & Harbusch, 2019; van de Velde & Meyer, 2014), which can reduce predictive value and increase planning demands as the sentence unfolds. Instead of committing to a single interpretation early, the professionals may continue evaluating multiple structural possibilities enabled by higher-frequency CVs, thereby slowing down the overall production to prepare for reformulation, maintain coherence, or accommodate anticipated structural complexity. This interpretation aligns with the operational definition of CV onset as the first appearance of an appropriate CV equivalent, regardless of later correction or revisions. The positive effect of CV frequency on CT-span was only marginally significant in the students, reflecting the presence of post-CV planning in the students, but to a more limited extent than in the professionals. Although the students also showed a significant positive association between CV frequency and sentence-span, the extended sentence duration may not have resulted from active refinement and elaboration, as observed in the professionals, but rather from inefficient or reactive formulation strategies. Instead of proactively structuring complex sentences like the professionals, the students tended to produce longer, more linear sentences when lower-frequency CVs posed challenges in lexical retrieval. This interpretation could be further supported or refuted by examining the quality of their interpreting renditions. Collectively,

these findings suggest that CVs exert structural influence not only at the outset of planning but also throughout sentence production.

In contrast, TW frequency was positively associated with sentence onset EVS but negatively associated with CT-span and sentence-span in both groups. Although this combination of delayed EVS measures and shortened duration measures may appear contradictive, it likely reflects the differing strategic focuses and temporal dynamics in processing nouns versus verbs during SI. Unlike higher-frequency CVs, which prompt rapid initial selection due to a wide array of semantically and syntactically compatible translation options, high-frequency TWs often require greater lexical precision and contextual sensitivity. This can lead to increased cognitive effort during early-stage planning, delaying production onset. Nevertheless, once production begins, this preparatory planning may extend as far as to the segment between the CV and the TW, subsequently supporting faster transitions reducing overall sentence duration. This suggests that while high-frequency CVs enable early structural anchoring, high-frequency TWs demand more refined lexical retrieval before articulation, resulting in divergent impacts on timing. The additional positive association between TW frequency and TW-EVS observed in the student group further suggests that, when encountering more frequent TWs, the students may have experienced greater lexical competition in selecting a suitable TW equivalent in the target language.

The differential impacts of CV and TW frequency on the EVS measures further suggest that both groups tended to prioritise verbs over nouns during both sentence comprehension and production. Given the verb's prominent syntactic role, as well as its earlier occurrence in the experimental sentence, interpreters likely allocated earlier and more substantial cognitive efforts to integrate the CV into both their mental representations and overt outputs. Consequently, the production of CV equivalents remained relatively stable across co-textual variations. In contrast, the TW, which typically appeared later in the experimental sentence, was more susceptible to the cumulative processing demands both within and across sentence boundaries. This interpretation is supported by the significant positive effects of preceding sentence length on TW-EVS, and the absence of such effects on CV-EVS.

7.2.4.2 Sentence-level features

The pattern of sentence-level features likely reflects prediction mechanisms potentially involved during SI. On the one hand, the effects of cloze probability support the prediction-by-production account. In both groups, cloze probability exerted positive effects on CV-EVS,

sentence offset EVS, and sentence-span. These finding suggest that, in response to constraining contextual cues, the interpreters may have engaged in more extensive predictive processing, allocating cognitive resources towards the production system for utterance planning or reformulation, aligning with the longer CV-EVS in the predictable condition. These efforts likely resulted in more integrated or syntactically complex output, consistent with the longer sentence-span observed in the predictable condition compared to the unpredictable condition. This interpretation aligns with the view that prediction enables higher-level processing adjustments, such as restructuring or pre-activate subsequent content (Gile, 2009; Seeber, 2013).

At the same time, greater contextual predictability also appeared to facilitate more efficient local CV-TW sequencing in both groups. In the original dataset, cloze probability did not significantly affect CT-span, possibly due to the masking effects of strong verb-noun associations in certain trials. However, in the subsampled dataset, a significant negative effect of cloze probability on CT-span emerged in both groups. While this shared pattern suggests that strong contextual cues helped the interpreters plan or initiate the TW more rapidly, leading to a shorter interval between the CV and the TW, the underlying mechanisms may differ across groups. For the professionals, shorter CT-span likely reflect enhanced structural planning or more efficient lexical retrieval without sacrificing completeness. That is, greater predictability allowed them to formulate the upcoming structure earlier and more fluently, while still producing structurally faithful rendering. In contrasts, the students may have responded adaptively to contextual predictability, opting to streamline their output by omitting or compressing the intermediate elements to ease processing. This interpretation aligns with the significantly shorter CT-span in the predictable condition observed only in the student group. Thus, while both groups benefited from predictability, their strategies reflected distinct cognitive approaches: the professionals elaborated with fluency, while the students simplified with efficiency.

Another possible explanation of the positive association between cloze probability and sentence-span is that higher cloze probability tended to associate with less frequent CVs, as evidenced by significantly lower CV frequency in the predictable condition than in the unpredictable condition. With fewer translation options in the target language (Wen & van Heuven, 2017), less frequent CVs may cause greater difficulty in early lexical selection, thereby delaying the sentence initiation and the production of CV equivalents. Meanwhile, less frequent CVs tend carry more specific semantic meanings and provide stronger predictive cues,

thus inviting more robust predictive processing and leading to greater sentence elaboration. In other words, lower-frequency CVs may slow early production but result in more top-down elaboration, whereas higher-frequency CVs allow early sentence onset but require more active construction downstream. Both pathways may contribute to longer sentence spans, albeit through distinct underlying mechanisms.

This interpretation suggests a U-shaped relationship between CV frequency and sentence-span. To test this, additional LME models were constructed for each group with sentence-span as the dependent variable and CV frequency as the independent variable interacted with a second-order polynomial term, including a by-participant random intercept. The results showed a significant positive effect on the linear term (professionals: $\beta = 12.547$, $SE = 2.768$, $t = 4.533$, $p < .001$; student: $\beta = 19.516$, $SE = 2.753$, $t = 7.089$, $p < .001$), but no significant effect on the quadratic term (professionals: $\beta = -3.709$, $SE = 2.881$, $t = -1.288$, $p = .199$; students: $\beta = -4.051$, $SE = 2.854$, $t = -1.419$, $p = .157$) in either groups. This reinforces the idea that sentence-span extended with CV frequency. Taken together, these findings indicate that the positive association between cloze probability and sentence-span observed in both groups cannot be fully explained by the increased predictability associated with less frequent CVs. Rather, they suggest that the interpreters may have drawn upon broader predictive cues embedded in the discourse, such as contextual information and structural expectations, to support planning and formulation during SI. In this sense, predictive processing appears to have extended beyond the immediate lexical item (i.e., the CV), engaging higher-order mechanisms that facilitate discourse-level anticipation and production.

The observed pattern of verb-noun association strength, on the other hand, supports the co-existence of prediction-by-association account during SI. In both groups, verb-noun association exerted negative effects on TW-EVS, sentence onset EVS, sentence offset EVS, and the two duration measures. This pattern suggests that interpreters may have benefited from strong lexical associations in anticipating the TW, enabling them to rapidly generate a semantic representation of the verb-noun segment within their mental context. Unlike the predictive processing invited by increased cloze probability, which involved more top-down elaboration and resulted in expanded output, stronger lexical links appear to have triggered bottom-up activation that was relatively automatic and demanded minimal cognitive resources. This process likely accelerated the sentence initiation and enhanced fluency and cohesion in their outputs. The facilitating effect of stronger lexical associations between the CV and the TW on earlier sentence initiation also supports the notion that early-stage planning may span across

the segment linking the CV and the TW. Compared to the professionals, the students exhibited an additional significant negative effect of verb-noun association on CV-EVS. The broader significant effects of verb-noun association observed in the student group echoes the view that the students may have engaged less in top-down analysis and syntactic adjustment than the professionals. Instead, the students may have relied more heavily on bottom-up, surface-level lexical associations to initiate production.

Sentence length showed positive effects on sentence-span and negative effects on sentence onset EVS in both groups. The positive association indicates that longer sentences elicited proportionally longer production durations. In contrast, the negative effect on sentence onset EVS, however, likely reflects a discourse-level tendency for longer sentence to be followed by shorter ones, serving as a strategy to ease comprehension difficulty. Therefore, when preceded by a longer sentence, the experimental sentence tended to be shorter, thereby reducing its production duration. The students demonstrated an additional positive effect of sentence length on TW-EVS, which was not found in the professionals. This finding supports the interpretation that the processing and production of the verb were prioritised over those of the noun, resulting in relatively stable lags between the input and output of CVs across varying sentence lengths. Conversely, as the TW typically appeared later in the experimental sentence, the students may have delayed its processing and production to a greater extent in longer sentences, likely due to the cumulative processing demands of preceding text.

7.2.4.3 Discourse-level features

At the discourse level, both groups demonstrated largely consistent patterns across all six temporal measures. Preceding sentence length exerted positive effects on TW-EVS and sentence onset EVS, and a negative effect on sentence span in both groups. The positive effects suggest that longer preceding sentences delayed the initiation of the experimental sentence and the production of the TW equivalents. The student group exhibited an additional (marginally) significant positive association with sentence onset EVS, indicating a stronger cumulative effect of longer preceding sentence in the students compared to the professionals. The negative effect on sentence span further supports the interpretation that longer preceding sentences were likely followed by shorter experimental sentences, resulting in reduced production durations.

The position of the sentence within a paragraph exhibited positive effects on all six temporal measures, suggesting that lags between input and output gradually increased as the discourse progressed. Similarly, the order of the paragraph within the overall SI task were positively

associated with all temporal measures except CT-span. These patterns reflect a cumulative impact of discourse complexity and interpreter fatigue, both within individual paragraphs and across the broader task. As linguistic input accumulated, interpreters' cognitive resources may have been increasingly taxed, resulting in slower production. Notably, post hoc Tukey's HSD tests revealed different patterns of the professionals and the students. The professionals exhibited significantly higher values in Paragraph 4 than in the three preceding paragraphs across all four EVS measures. These findings indicate that the professionals maintained a relatively consistent interpreting rhythm during the first three paragraphs, which spanned approximately thirty minutes with intermittent breaks, after which signs of fatigue became more pronounced, as evidenced by longer EVSs. This pattern is consistent with previous studies suggesting a thirty-minute threshold for cognitive endurance in SI (Klonowicz, 1990; Kurz, 2003; Moser-Mercer, Künzli, & Korac, 1998).

Meanwhile, the students showed extended EVSs as early as in Paragraph 3, suggesting an earlier appearance of interpreting fatigue in the students. However, the significantly shorter CT-span in Paragraph 3 than in Paragraph 2 may reflect the students' growing adaptability to contextual predictability and the experimental setting in which the visual display consistently included the target objects among the four presented options. The students may have learnt to utilise both linguistic cues and visual information to anticipate the TW and streamline their interpreting output by transitioning more quickly from the CV to the TW. Yet, this effect of increasing adaptability was not seen in Paragraph 4, possibly due to mounting interpreter fatigue that became even more pronounced than in Paragraph 3 and counteracted the students' growing familiarity with the experimental setting.

In sum, both groups showed similar tendency to prioritise the production of verbs over nouns in their processing and planning, and both were subject to the cumulative effects of processing both with and across paragraphs. A primary difference between the two groups lay in their strategic engagement with predictive processing. The professionals appeared to exploit predictive cues more proactively and strategically, engaging in deeper processing and formulation to ensure semantically faithful and structurally refined outputs. In contrast, the students seemed to engage with predictive information more reactively and superficially, adopting a streamlining strategy to maintain synchrony with the source, which resulted in less adjusted and less fluent interpretations.

7.3 Interpreting quality

7.3.1 Rubric-referenced assessment

The rubric-referenced scores were rated by two professional interpreters with a minimum of three years' interpreting experience who did not participate in the experiment. Both raters were native speakers of Mandarin Chinese with English as their second language. Prior to the assessment, they were provided with the ST, the rating scale adopted from Han (2016), and a six-minute tutorial video briefly explaining the scale. The rating comprised three eight-point subscales: information completeness (InfoCom), fluency of delivery (FluDel), and target-language quality (TLQual). Where a significant discrepancy occurred between the two raters (i.e., a score difference greater than two on any subscale), the recording was re-assessed by both raters. A composite score of overall interpreting quality was computed following a weighted scheme: 50% InfoCom, 30% FluDel, and 20% TLQual.

To evaluate inter-rater reliability across the four dimensions of interpreting quality, intraclass correlation coefficients (ICCs) were computed using the *ICC()* function from the *psych* package in R. Following the guidelines of McGraw and Wong (1996), a two-way random-effects model with absolute agreement was applied, reporting both single and average measures. For InfoCom, the ICC for average measures (ICC2k) was .93, 95% CI [.89, .96], indicating excellent agreement between raters. FluDel yielded an ICC2k of .96, 95% CI [.93, .97], also reflecting a high degree of reliability. The ICC2k for target language quality was .92, 95% CI [.86, .95], and for overall quality, it was .97, 95% CI [.95, .98], both suggesting strong inter-rater consistency. These results confirm that the ratings used in the subsequent analysis were robust and reliable.

Independent samples t-tests were conducted to compare each subscale score between the professionals and the students. The results indicated that the professionals outperformed the students across all three subscales as well as in overall interpreting quality. The largest between-group difference was observed in InfoCom, while the smallest was found in TLQual. Pearson's correlations analyses were then conducted to examine the relationships among these three subscale cores within each group. In the professionals, InfoCom was strongly and positively correlated with FluDel, $r = .930$, 95% CI [.836, .971], $t(20) = 11.292$, $p < .001$, and with TLQual, $r = .915$, 95% CI [.804, .965], $t(20) = 10.171$, $p < .001$. A similarly strong correlation was observed between fluency of delivery and target language quality, $r = .918$, 95% CI [.810, .966], $t(20) = 10.366$, $p < .001$. Similar patterns were observed among the students:

InfoCom was positively associated with FluDel, $r = .876$, 95% CI [.783, .931], $t(42) = 11.786$, $p < .001$, and with TLQual, $r = .853$, 95% CI [.745, .918], $t(42) = 10.595$, $p < .001$. The correlation between FluDel and TLQual was also strong, $r = .872$, 95% CI [.777, .929], $t(42) = 11.572$, $p < .001$. These results indicate consistently robust interrelations among the quality dimensions across both professional and student interpreters.

Table 7-15. Between-group comparisons for the rubric-referenced interpreting scores

	Professionals	**Student**	***t***	***df***	***p***
InfoCom	5.909 (SD = 1.065)	4.886 (SD = 0.858)	3.621	39	0.001 **
FluDel	5.523 (SD = 1.096)	4.571 (SD = 0.841)	3.341	38	0.002 **
TLQual	5.636 (SD = 0.990)	4.986 (SD = 0.712)	2.551	36	0.015 *
Overall	5.739 (SD = 1.034)	4.811 (SD = 0.778)	3.436	37	0.001 **

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

7.3.2 Item-based assessment

The item-based accuracies were calculated as proportions of accurately interpretated CVs, TWs, and experimental sentences in participants' interpreting renditions. The criteria for correct interpretation of CVs and TWs did not require literal translations but was judged based on semantic and syntactic appropriateness, as well as fidelity to the ST. Sentence-level accuracy was evaluated in two steps. First, each sentence was annotated for frame elements (Zhang, 2016), including agent, action, patient, modifiers, locations, and relational components. Accurate interpretation required correct rendering of at least all elements except modifiers. Second, the overall meaning of the sentence had to remain faithful to the original intent. The judgement process was carried out twice by the researcher, with a three-month interval between assessments. Intra-rater consistency between the two assessments was determined by calculating the proportion of items judged identically in both instances, yielding consistency rates of 97.65% for CVs, 99.23% for TWs, and 96.42% for sentences. These results indicate a high level of intra-rater reliability. The results from the second round of judgement were used in subsequent analyses. Pearson's correlation analyses revealed significant positive associations between rubric-referenced scores and item-based accuracies ($ps < .001$, $rs > .550$), supporting the reliability of the assessment framework.

Paired-sample t-tests were conducted within each group to assess between-condition differences in accuracy of CV, TW, and sentence interpretations. For the professionals,

accuracy tended to be higher in the unpredictable condition compared to the predictable condition, although these differences were not statistically significant across all measures. This trend was also observed in the students, with much more pronounced distinctions. Specifically, the students showed a significant improvement in CV accuracy in the unpredictable condition compared to the predictable condition, while no significant differences were observed for TW accuracy. Sentence interpretation accuracy for the students was also significantly better in the unpredictable condition. These patterns suggest that unpredictability may enhance certain aspects of interpreting quality, especially for less experienced interpreters.

Table 7-16. Between-condition comparisons for item-based accuracies

Professional	**Predictable**	**Unpredictable**	***t***	***df***	***p***
CV accuracy	0.723 (SD = 0.198)	0.805 (SD = 0.166)	–1.480	41	0.147
TW accuracy	0.833 (SD = 0.136)	0.826 (SD = 0.137)	0.184	42	0.855
Sentence accuracy	0.591 (SD = 0.251)	0.634 (SD = 0.225)	–0.606	42	0.548
Student					
CV accuracy	0.611 (SD = 0.198)	0.766 (SD = 0.187)	–3.738	84	<0.001 ***
TW accuracy	0.758 (SD = 0.176)	0.779 (SD = 0.186)	–0.546	84	0.587
Sentence accuracy	0.490 (SD = 0.217)	0.587 (SD = 0.223)	–2.041	84	0.044 *

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Independent-sample t-tests were conducted to compare overall interpreting accuracies between the two groups. The professionals significantly outperformed the students across all three measures. For CVs, the difference was significant, $t(99) = 2.999$, $p = .003$; for TWs, the difference was also significant, $t(118) = 2.845$, $p = .005$; and for sentence interpretation, the difference remained statistically significant, $t(84) = 2.195$, $p = .031$. These results indicate a consistently higher level of interpretation accuracy among the professionals compared to the students. To further examine whether a condition effect contributed to these between-group differences, independent-sample t-tests were performed for each condition. In the predictable condition, the professionals demonstrated significantly higher accuracy than the students in interpreting CVs, $t(42) = 2.161$, $p = .036$. While the professionals also exhibited higher mean accuracy than the students in interpreting TWs and overall sentences, these differences were not statistically significant: $t(53) = 1.914$, $p = .061$ for TWs, and $t(37) = 1.600$, $p = .118$ for sentences. In the unpredictable condition, no significant differences were found between the professionals and the students across any of the measures. Accuracy for CVs was slightly

higher in the professionals than the students, but the difference was non-significant, $t(47) = 0.847$, $p = .401$. Similarly, TW recognition and sentence interpretation did not differ significantly between groups, with $t(55) = 1.146$, $p = .257$ and $t(42) = 0.803$, $p = .427$, respectively. These findings suggest that expertise played a more prominent role in modulating the processing CVs in the predictable condition than in the unpredictable condition.

7.3.3 Interim discussion

The rubric-referenced assessment revealed significant advantages for the professionals over the students across all quality dimension, with the largest advantage observed in information completeness. This aligns with the item-based assessment, which showed that the professional group achieved a higher overall level of interpreting accuracy than the student group. This expertise-related advantage primarily stemmed from the professionals' enhanced performance in processing predictable sentences, as evidenced by the significant and marginally significant between-group differences in CV and TW accuracies in the predictable condition. Within-group comparisons showed that the professionals maintained relatively stable accuracy across all three measures in both conditions, whereas the students demonstrated significantly lower CV and sentence interpretation accuracies in the predictable condition. These comparisons also revealed the most pronounced between-condition differences in CV accuracy for both groups, suggesting greater difficulty in processing and producing CV equivalents in the predictable condition than in the unpredictable condition.

These findings further support the previous interpretations of the EVS data. First, CVs played a critical role in determining syntactic structure and thus exerted impacts beyond the lexical level. Second, predictive CVs, which tend to be of lower frequency, offered fewer translation options and demanded more efforts in lexical retrieval, thereby increasing processing and production difficulties, especially in early utterance planning. Third, the two groups may have employed different strategies to cope with the increased processing demands associated with predictive CVs. The professionals may have taken advantage of the higher predictability associated with predictive CVs and engaged in extensive planning and formulation to preserve structural fidelity, resulting in renditions comparable to those in unpredictable condition, which involved more frequent, less predictive CVs. The students, by contrast, were more vulnerable to the increased processing difficulties posed by less frequent predictive CVs. The greater efforts required for lexical retrieval may have impeded their predictive processing, contributing to slightly lower TW accuracy than that of professional in the predictable condition. Notably,

no significant between-condition differences in TW accuracy were observed in either group, suggesting relatively consistent processing of TWs in either condition. This may be explained by the identical TWs used across conditions and the visual availability of target objects, which may have facilitated TW processing.

7.4 Relationships between EVS and interpreting quality

To explore the relationship between EVS and interpreting quality, Pearson's correlation tests were conducted between EVS measures and corresponding item-based accuracy scores at both lexical and sentence levels. For both the professionals and the students, correlations were computed between (1) CV-EVS and CV accuracy, (2) TW-EVS and TW accuracy, (3) sentence onset EVS and sentence accuracy, and (4) sentence offset EVS and sentence accuracy. To further examine condition-specific patterns, additional correlation analyses were conducted separately for the predictable and the unpredictable conditions. These analyses aimed to determine whether synchrony between comprehension and production were associated with interpreting accuracy, and whether this association differed by condition or expertise. To assess the stability of the EVS-accuracy associations, a robustness check was conducted by rerunning the correlation tests on the subsampled dataset.

7.4.1 By-group analysis

In the professionals, sentence accuracy showed a significant negative correlation with sentence onset EVS, $r = –.572$, 95% CI [–.742, –.331], $t(42) = –4.514$, $p < .001$, and a marginally significant negative correlation with sentence offset EVS, $r = –.256$, 95% CI [–.514, –.044], $t(42) = –1.719$, $p = .093$. No significant correlations were observed between lexical-level EVS and accuracy. When analysed by condition, sentence accuracy showed a marginally significant negative correlation with sentence onset EVS in the predictable condition and a significant negative correlation in the unpredictable condition. The robustness check exhibited increased significance in the overall correlations between sentence offset EVS and sentence accuracy (from $p = .093$ to $p = .034$) and the correlation between sentence onset EVS and sentence accuracy in the predictable condition (from $p = .088$ to $p = .048$) (see Appendix 9).

Table 7-17. Pearson's correlations between EVS measures and item-based accuracies for the professional group

Predictable						
EVS	**Item-based accuracy**	***r***	**95% CI**	***t***	***df***	***p***
CV-EVS	CV accuracy	.077	[–.356, .483]	0.344	20	.734
TW-EVS	TW accuracy	.041	[–.387, .455]	0.185	20	.855
Sentence onset EVS	Sentence accuracy	–.373	[–.686, .058]	–1.796	20	.088 †
Sentence offset EVS	Sentence accuracy	–.195	[–.570, .247]	–0.890	20	.384
Unpredictable						
EVS	**Item-based accuracy**	***r***	**95% CI**	***t***	***df***	***p***
CV-EVS	CV accuracy	.045	[–.384, .458]	0.203	20	.841
TW-EVS	TW accuracy	–.004	[–.425, .419]	–0.016	20	.987
Sentence onset EVS	Sentence accuracy	–.606	[–.819, –.248]	–3.409	20	.003 **
Sentence offset EVS	Sentence accuracy	–.208	[–.579, .234]	–0.951	20	.353

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

The students exhibited a more complicated pattern than the professionals. Overall, CV accuracy was significantly negatively correlated with CV-EVS, $r = –.262$, 95% CI [–.449, –.053], $t(84) = –2.487$, $p = .015$. TW accuracy also showed a negative correlation with TW-EVS, $r = –.230$, 95% CI [–.421, –.019], $t(84) = –2.163$, $p = .033$. Sentence accuracy was significantly negatively correlated with sentence onset EVS, $r = –.397$, 95% CI [–.562, –.203], $t(84) = –3.970$, $p < .001$, while no significant correlation was found for sentence offset EVS. In both conditions, sentence accuracy consistently showed significant negative correlations with sentence onset EVS. In the unpredictable condition, a marginally significant negative correlation was observed between CV accuracy and CV-EVS. The robustness check revealed reduced significance in the overall correlations between the lexical level measures, indicating limited stability in these effects across trials (see Appendix 9).

Table 7-18. Pearson's correlations between EVS measures and item-based accuracies for the student group

Predictable						
EVS	**Item-based accuracy**	***r***	**95% CI**	***t***	***df***	***p***
CV-EVS	CV accuracy	.061	[–.243, .355]	0.394	41	.695
TW-EVS	TW accuracy	.027	[–.276, .325]	0.173	41	.863
Sentence onset EVS	Sentence accuracy	–.435	[–.650, –.155]	–3.091	41	.004 **
Sentence offset EVS	Sentence accuracy	.066	[–.239, .359]	0.423	41	.674

Unpredictable						
EVS	**Item-based accuracy**	***r***	**95% CI**	***t***	***df***	***p***
CV-EVS	CV accuracy	–.281	[–.536, .021]	–1.878	41	.068 †
TW-EVS	TW accuracy	–.057	[–.352, .247]	–0.368	41	.715
Sentence onset EVS	Sentence accuracy	–.320	[–.566, –.022]	–2.166	41	.036 *
Sentence offset EVS	Sentence accuracy	–.226	[–.493, .080]	–1.485	41	.145

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Although the reported significant correlations are small to moderate in magnitude, they are considered meaningful given the complexity of human behaviour. Cohen, Cohen, West, and Aiken's (2003) conventional benchmarks classify Pearson's r values of 0.10, 0.30, and 0.50 as small, medium, and large effects, respectively, subsequent empirical analyses have indicated that these thresholds may be overly conservative in psychological research. For instance, Gignac and Szodorai (2016), based on 708 meta-analytic correlations, found that the 25th, 50th, and 75th percentiles corresponded to r = 0.11, r = 0.20, and r = 0.29, respectively. Based on these findings, they recommended considering correlations of 0.10, 0.20, and 0.30 as relatively small, typical, and relatively large, respectively, in individual differences research. Similarly, Weinerová et al. (2022) conducted a comprehensive analysis of over 12,000 correlations in social and developmental psychology, and reported median r values around 0.17 to 0.31, depending on the sample and context. These findings support the interpretation of *r* values in the range of 0.20–0.30 as meaningful within the current research context.

7.4.2 Interim discussion

In both groups, sentence accuracy exhibited relatively consistent negative correlations with sentence onset EVS in both conditions, suggesting that earlier initiation of sentence production yielded more accurate sentence renditions. This is consistent with findings from previous studies which reported that prolonged EVS tends to impair the SI performance (Díaz-Galaz et al., 2015; Lee, 2002). The robustness check confirmed these associations, with increased significance observed in the professionals in the predictable condition. This strengthens the interpretation that early sentence initiation supports higher output quality, possibly by reducing memory load and allowing more fluent delivery.

At the lexical level, the association between EVS measures and interpreting accuracy was not significant in the professionals, and this pattern remained stable in the robustness check. This

suggests that the professionals were adept at adjusting word order and managing timing delays without compromising lexical translation accuracy. This flexibility possibly reflects greater WM capacity and more automatised lexical retrieval processes in the professionals, which allowed them to buffer planned output more effectively. In contrast, the students showed significant negative associations between the two lexical-level EVS measures and their respective accuracy scores in the original dataset. However, these effects were reduced or non-significant in the robustness check, indicating that these associations were less stable across trials. This suggests that the observed effects in the original dataset may have been driven by specific high-variance trials and not generalisable across the full sample. The reduction in significance reinforces the interpretation that student interpreters' performance was more variable, and their lexical processing was more susceptible to task demands and trial-level differences.

Taken together, these findings suggest that while sentence-level timing is a robust predictor of accuracy for both groups, lexical-level timing is less stable indicator with greater individual variances, especially among less experienced interpreters.

7.5 Self-reported retrospection data

The participants' retrospection data were analysed through a hybrid thematic analysis, combining inductive and deductive approaches. At the initial stage, the researcher reviewed the raw responses repeatedly to familiarise with the data. Inductive coding was applied to label participants' responses without theoretical preconceptions, allowing themes to emerge from the data. These preliminary themes were then refined through iterative review, focusing on patterns in interpreter's anticipation strategies, visual-input influence, and reported challenges. Subsequently, deductive coding was employed to re-examine the data through the lens of prediction-by-production and prediction-by-association, ensuring alignment with established frameworks. To enhance reliability, the coding process was conducted twice by the researcher at a six-month interval, with an intra-coder consistency of 95.68%. The final reported results derive from the second coding iteration, ensuring rigorous and reflective engagement with the data. The reports below start from codes of responses to the last question concerning anticipation, as they are most relevant to the present study.

7.5.1 Codes concerning anticipation

The analysis of participants' responses to the last question revealed three overarching themes: anticipation mode or cues, contents of anticipation, and the frequency of anticipation (see Appendix 10). Five anticipation modes were identified, differing mainly in the cues participants relied upon. The most frequently reported was world-based anticipation, where participants drew on their world knowledge or prior experiences to anticipate upcoming content (Quote 1-4). World knowledge is understood as the extra-linguistic information that supports comprehension and the establishment of mental models. The second most common was discourse-based anticipation, in which participants used linguistic cues from earlier segments of the discourse to infer likely continuations. This mode involved an active analysis of both lexical semantics and discourse coherence, such as anticipating a contrastive description of daytime after hearing "unlike the peace and quietness in the night…" (Quote 5-6). A third mode, glossary-based anticipation, reflected participants' use of pre-task materials to anticipate potential topics or scenarios (Quote 7). Some participants also reported visual-based anticipation, in which interpreters used information from the visual display to inform their anticipations (Quote 8). Finally, lexical-based anticipation referred to anticipations triggered by individual lexical items. This ranged from automatic lexical association (e.g., "sushi" leading to "fish") (Quote 9), to more inferential reasoning (e.g., "drive" prompting the prediction of "motorcycle") (Quote 10).

> Quote 1: It's easier to guess content what's coming when it's something we see in daily life. Like, when I heard "boarding the train" and "sitting on my seat", I could picture the train and imagine what might come next. (P110)

> Quote 2: When the porter said, "In case of emergency…", I already knew he'd say not to use the lift and take the stairs, just like those signs you always see. (P126)

> Quote 3: I imagined myself on the train and made some predictions based on my own travel experience. (P132)

> Quote 4: From "artist", "musician", and "central market", I figured "drum" would be more likely to show up than "piano", like something you'd see on the street. (117)

> Quote 5: When I heard "unlike… at night", I thought it might describe the city during the day as a comparison (P129)

Quote 6: The tourists recommended “sushi” to the narrator, so I guessed he would end up buying some. (P219)

Quote 7: The glossary had “make-believe game”, so I figured there might be some children playing games. (P123)

Quote 8: Seeing a stroller and a milk bottle made me think of a scene involving taking care of a baby. (P223)

Quote 9: When I heard the tourists were having “sushi”, I instantly thought of fish. (P218)

Quote 10: When I heard “drive”, I guessed some kind of vehicle would follow, possibly a motorcycle. (P213)

Regarding the anticipated contents, the most frequently anticipated elements were broad topics, events, or scenarios, rather than specific words. Participants reported constructing mental models of unfolding situations, anticipating agents, locations, actions, relations, or likely outcomes (Quote 11-12). In some cases, specific lexical items were anticipated, particularly when prompted by strong cues (Quote 13). Another pattern involved anticipation of semantic category or hypernym, where participants anticipated a more general category rather than a precise words (Quote 14). Notably, several participants also reported incorrect anticipations, which occurred at both the scenarios (four times by four participants) and the word items (four times by four participants) (Quote 15-16).

Quote 11: Sometimes I imagined what kind of scene it was. For example, when I heard “magic show” and “magician”, I guessed something might disappear. (P113)

Quote 12: It was lunch time, so I figured the speaker might go to the dining coach next. (P132)

Quote 13: The speaker said “musician” and “beat”, and I saw a drum on the display, so I expected he’d mention “drum”. (P217)

Quote 14: When the blonde lady boarded the train, I figured she probably had luggage, but I wasn’t sure if it was a backpack or a suitcase. (P217)

Quote 15: I thought the blonde lady would grab a coffee after getting on, but she just went to sleep right away. (P122)

> Quote 16: When the speaker talked about a butcher shop, I expected "steak" to come next. So, I was a bit surprised when it turned out to be "chicken". (P209)

With respect to anticipation frequency, participants' responses fell into four levels, closely tied to their attitudes toward anticipation. Approximately one-third of participants reported always or actively anticipating during the SI task, expressing a neutral to positive view of anticipation and providing multiple instances of anticipations in their retrospections. In contrast, about a quarter of participants expressed negative attitudes, stating that SI's cognitive demands left little room for anticipation (Quote 17-18). Some participants reported that they did not avoid anticipation but only occasionally did so, especially when they felt they had sufficient cognitive capacity. Finally, six participants (four students and two professionals) claimed that they never engaged in anticipation. A re-examination of their responses confirmed that none of these six participants provided any codes related to anticipation modes or content.

> Quote 17: I didn't really try to predict because if I got it wrong, it'd just take extra time and effort to fix. (P120)

> Quote 18: Predicting takes extra energy, so I mostly just followed along without trying to guess what was next. (P131)

Overall, a greater proportion of professional interpreters reported engaging in predictive processing during SI. Especially in terms of anticipation modes, the professionals more frequently reported drawing on world knowledge and discourse cues to support their anticipation. In contrast, student interpreters appeared more reliant on visual information, as reflected in the higher proportion of the students who reported using visual-based anticipation strategies. With regard to anticipated contents, about half of participants in both groups reported anticipating general topics or scenarios. However, the professionals were more likely than the students to report anticipating specific words. As for the frequency of anticipation, a larger number of the professionals indicated that they did not engage in anticipation at all during the SI tasks.

7.5.2 Codes concerning visual inputs

The analysis of participants' responses to the second question identified three major themes: viewing pattern, audio-visual input interaction, and effects of visual inputs (see Appendix 11). Regarding viewing pattern, some participants reported that they automatically looked at the visual display when objects appeared, although they sometimes found the visuals distracting

(Quote 19-20). Others describe consistently and actively engaging with the visual input, especially after recognising a link between the visuals and the source text (Quote 21-22). In contrast, some participants only looked at the images when they had spared cognitive resources (Quote 23), while others tried to avoid looking altogether, though they occasionally glanced at the display (Quote 24). A few participants reported completely ignoring the visuals and focusing solely on the SI task (Quote 25). It is worth noting that viewing behaviours varied throughout the SI task: some participants reported engaging with the visuals at the beginning but later withdrawing as the task became more demanding, and vice versa (Quote 26-27).

> Quote 19: The pictures sometimes distracted me, but I just couldn't help looking at them to find the target. (P104)
>
> Quote 20: I was distracted by some pictures in the last paragraph, so I decided not to look at them. But I could still see them from the corner of my eyes. (P118)
>
> Quote 21: I usually skimmed the four objects to see what might come up, but not in a focused way. I didn't want to get too distracted. (P132)
>
> Quote 22: After a few times, I realised one of the four objects would be mentioned, so I started checking for target objects every time when they appear on the screen. (P229)
>
> Quote 23: I only looked at the pictures when I had spare energy. Otherwise, I just stared at the cross in the centre of the screen. (P211)
>
> Quote 24: I tried not to look at the pictures too much. Not all of them were useful and some were quite distracting. (P224)
>
> Quote 25: I focused on what I was hearing. I didn't have the capacity to process pictures too. (P216)
>
> Quote 26: At first, I kept looking at the pictures, but I found they distracted me because I'd start thinking about what they meant. So, I gave up on them later. (P209)
>
> Quote 27: I didn't look at the pictures in the beginning, but I found them helpful in reminding me of what I'd just heard. So, I started using them more. (P218)

In terms of audio-visual input interaction, participants most frequently reported that they located the target object after hearing the corresponding TW. However, some identified the target before hearing it, guided by contextual cues (Quote 28). Others reported locating the target almost simultaneously with hearing the word (Quote 29). While some intentionally

searched for the target once they understood how the visual stimuli related to the audio input, others mistakenly focused on an object that ultimately proved irrelevant (Quote 30).

> Quote 28: I saw the grapes before hearing the word. The speaker mentioned wine, so I guessed the grapes were relevant. (P119)
>
> Quote 29: I happened to see the drum on the screen when I heard the word "drum". (P126)
>
> Quote 30: When I heard "toys", I looked at the teddy bear, but it turned out to be the doll. (P213)

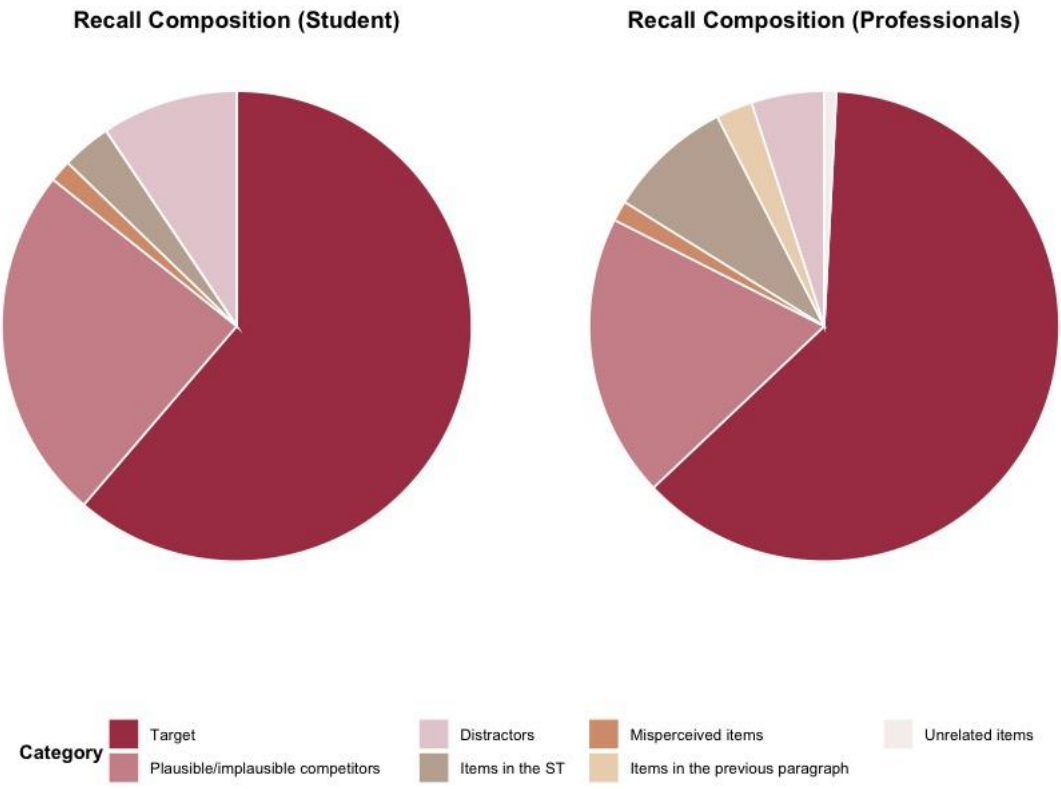
With regards to the influence of visual inputs, participants reported that the visuals distracted them from the SI tasks (Quote 31-32), while others said the visuals misled their comprehension or predictions (Quote 33). In contrast, some participants found the visuals helpful for recalling or confirming parts of the source text (Quote 34-35). Some participants expressed neutral views as they found themselves unaffected by visual inputs. A few participants noted that the visual input supported their interpretation when lexical access was limited. (Quote 36).

> Quote 31: Sometimes the pictures distracted me from producing outputs. (P214)
>
> Quote 32: I saw the suitcase and started wondering why it wasn't black like in the audio. That threw me off a bit. (P128)
>
> Quote 33: The watering can mislead me. I thought the man in the garden would water the flowers, but he was painting it. (P121)
>
> Quote 34: The pictures sometimes helped remind me of the source text. Like, if I missed a few words, the pictures filled in the gap. (P219)
>
> Quote 35: The visuals helped me double-check what I heard. They confirmed that I got it right. (P125)
>
> Quote 36: I couldn't think of a good Chinese equivalent for "fireplace" until I saw it on the screen. (P123)

Participants were also asked to recall objects from the visual display (see Figure 7-8). The target objects were the most frequently remembered, followed by the two types of competitors. Interestingly, participants occasionally recalled items mentioned only in the audio, or misperceived certain images. Only the professionals reported recalling items from previous

sessions, and one professional recalled some items that had appeared in neither the audio ST nor the visual displays.

Figure 7-8. The composition of recalled object types in each group



7.5.3 Codes concerning SI tasks

The analysis of participants' responses to the first question revealed four major themes: subjective perceptions of ST difficulty, challenges encountered during SI, interpreting strategies adopted, and meta-cognitive reflections (see Appendix 12). The most frequently reported source of difficulty was the information density. Several participants noted that the first paragraph contained a particularly dense flow of information, characterised by frequent scene shifts (Quote 37). However, many found the second paragraph less intensive, despite the subjective ratings showing no significant difference in difficulty between the two paragraphs. Two additional sources of difficulty were frequently mentioned: unfamiliar content and unfamiliar language style. Though related, these two challenges differed in nature. Unfamiliar content referred to content words, such as uncommon place names or culturally specific terms, which were hard to visualise or translate (Quote 38-39). Unfamiliar language style referred to

the ST's literary tone, which participants found uncommon in SI practices, particularly because of its elaborate use of modifiers (Quote 40-42). Unexpectedly, some participants, especially the professionals, found slow speech delivery to be an impediment. Some noted that slow pacing made it harder to form a coherent mental representation of the message, resulting in increased memory load (Quote 43-44). Others found that slow speech placed higher demands on their production.

> Quote 37: The information density was quite high for me, especially with the story switching scenes so often (P102).
>
> Quote 38: I'm not familiar with the natural scenes describe in the story. It's hard to imagine a train running by the sea. (P224)
>
> Quote 39: I wasn't sure what "porter" meant, so I didn't know how to interpret it. (P104)
>
> Quote 40: It's easy to understand literary descriptions, but hard to translate them well. (P111)
>
> Quoe 41: I've never interpreted a literary text before. It was beautiful, so I tried to polish my output more. (P120)
>
> Quote 42: The source text was quite literary, with lots of details and adjectives as modifiers, not something I usually interpret. (P230)
>
> Quote 43: The speech itself wasn't hard, but it was so slow I couldn't organise the info properly. I had to hold too much in memory. (P132)
>
> Quote 44: Because the speech was slow, I had to wait for the full meaning, which increased memory pressure. (P229)

In terms of challenges during SI, the two most commonly reported were code-switching difficulties (e.g., failing to find an equivalent in the target language) and listening comprehension failures (e.g., missing key details) (Quote 45-46). These two challenges reflect distinct stages of interpreting: the former is related to comprehension, and the latter to production. Some participants also reported intentional omissions due to time pressure or uncertainty, particularly with adjectives and modifiers (Quote 47). Others described difficulties maintaining the naturalness of their output due to time constraints or the ST's literary style (Quote 48-49). A few mentioned misunderstanding or inaccuracies in their renditions.

> Quote 45: I got stuck on “Trondheim”, even though it was in the glossary. I just couldn’t recall the Chinese name. (P114)

> Quote 46: The sentence about “sushi” was structure differently from Chinese, so it was a bit hard to interpret. (P106)

> Quote 47: I heard the adjectives but didn’t have time to interpret them. (P127)

> Quote 48: There were lots of modifiers, and I tried to use more elegant language, but that kind of distracted me from listening. (P107)

> Quote 49: I tried translating every modifier, but it took extra time to organise and find the right words. (P111)

Participants also shared interpreting strategies they employed. Some prioritised content words, omitting modifiers to manage cognitive load (Quote 50). Others reported filling in gaps using context or prior knowledge when comprehension failed (Quote 51). A few mentioned taking notes, even if they did not refer back to them. Some adopted a syntactic linearity approach, translating word-for-word to avoid processing syntactic reordering, and used fillers to maintain fluency.

> Quote 50: There were lots of adjectives, which I don’t usually deal with. I focused more on nouns as I figured they mattered more. (P127)

> Quote 51: I couldn’t remember exactly what the porter said about “emergency”. I realised I might have interpreted something wrong halfway through, so I just followed my own logic and reinterpreted the sentence. (P223)

In terms of meta-cognitive strategies, some participants reported adjusting to the task and the dual audio-visual input over time, especially after the first paragraph. Several participants mentioned intentionally controlling their EVS to manage output fluency or memory load (Quote 52-53). Only student participants reported fatigue, particularly in the later sections of the task (Quote 54).

> Quote 52: I extended the EVS as I usually do. It helps organising my output. (P113)

> Quote 53: When the speech was slow, I waited for all the info to come through, but then I’d forget the beginning. (P120)

7.5.4 Interim discussion

The analysis of the retrospection data provided further insight into how participants engaged in predicative processing during SI tasks. World-based anticipation was the most frequently reported mode, in which participants reported using world knowledge or prior experience to anticipate upcoming input. This is consistent with the simulation-based prediction (Huettig, 2015), which emphasises the use of mental imagery to simulate events and generate specific predictions based on sensorimotor experience. This anticipation mode also supports the prediction-by-production account. The simulation process suggests a perspective-taking process, based on which participants inferred the speaker's communicative intention and projected plausible continuations based on message-level coherence. This is distinct from automatic associative activation; rather, it reflects intentional, context-sensitive processing (Pickering & Gambi, 2018). Similarly, the frequently reported discourse-based anticipation involved top-down analysis of linguistic context, drawing on message-level understanding rather than lexical priming. In contrast, visual-based and lexical-based anticipations may reflect either production-based or association-based processes. For example, the prediction of "grapes" upon seeing its image could stem from a top-down evaluation of contextual relevance or from lexical association with "wine" mentioned earlier. Overall, participants' reflections suggest the operation of both prediction-by-production and prediction-by-association during SI.

Although the eye-tracking data revealed significant predictive fixations on the targets prior to the TW onset, only a few participants reported having identified the targets before beforehand. They more often indicated that they only occasionally anticipated upcoming content. This discrepancy between the eye-tracking data and self-reports likely reflects a gap between implicit cognitive processing and conscious awareness. The predictive eye movements provide strong evidence of the pre-activation of semantic element, yet this process typically occurs below the level of conscious awareness. In contrast, retrospective self-reports are subject to memory reconstruction and biases. Participants may not accurately recall the timing of their predictive fixations. Identifying the targets after hearing it involves conscious integration into the mental model, which is more memorable. This interpretation is also supported by the finding that participants most frequently recalled the target objects than any other types of objects. The memory biases can also be supported by instances where participants recalled items from previous sessions, items mentioned only in the audio input, or even items that were completely irrelevant.

Notably, professional interpreters provided more detailed and reflective retrospections than student interpreters, as evidenced by the higher proportions of coded responses from the professionals across nearly all categories. Especially in terms of challenges encountered during SI, the professionals reported more instances of inaccurate interpretations, despite having significantly higher information completeness and overall interpreting quality. This suggests that the professionals may exhibit greater cognitive flexibility to cope with multi-channel inputs while managing to self-monitor the output. There were, however, a few notable differences in student responses. First, the students more frequently reported visual-based anticipation, indicating a greater reliance on surface-level lexical associations rather than deep contextual analysis. This is consistent with the eye-tracking findings showing more static fixation patterns and fewer visual attention shifts, as well as increased fixations on the plausible competitors when contextual constraints were low. Second, the students more commonly reported omitting modifiers and focusing on content words in their outputs. This aligns with the EVS data, which showed the students had shorter CT-span in the predictable condition, suggesting that the students tended to adopt a streamlined production strategy facilitated by stronger contextual predictability. Third, only the students reported interpreting fatigue, in contrast to the professionals. This finding is consistent with the EVS patterns: the students began extending their EVSs as early as in Paragraph 3, whereas the professionals demonstrated significantly longer EVSs only in Paragraph 4, indicating higher resilience under sustained cognitive load.

7.6 Relationships between eye movements and SI performance

To examine the relationships between eye movements and SI performance, all participants, regardless of their expertise, were re-labelled based on their EVSs and interpreting quality. The EVS-based labels were determined using a new EVS measure, i.e., sentence onset-to-TW EVS, computed by subtracting the sentence onset in participants' interpreting outputs from the TW onset in the source text. This measure captured the onset of sentence production relative to the onset of the TW. The TW onset was used as the temporal anchor to maintain alignment with the previous eye-tracking analysis, where it served as a fixed reference point suitable for cross-participant and cross-trial comparisons. The quality-related labels were based on rubric-referenced scores of overall interpreting quality.

On average, participants began interpreting 1670 ms (*SD* = 2655 ms) before the TW onset (professionals: −1627 ms (*SD* = 2691 ms), students: −1693 ms (*SD* = 2636 ms)). Participants whose average sentence onsets occurred earlier than the average were labelled as "Early" interpreters, while those who began later were classified as "Late" interpreters. The average overall interpreting score across all participants was 5.145 (*SD* = 1.03). Participants with overall scores above this average were labelled as "High-quality" interpreters; and those with lower scores were labelled as "Low-quality" interpreters. Table 7-22 illustrates the distribution of participants following re-labelling. Notably, participants whose eye-tracking data were deemed invalid were also excluded from this analysis. About 40% of professional interpreters were classified as the early interpreters and 60% as the late interpreter. Around 77% of the professionals produced interpreting outputs rated above the average, while the remaining 23% contributed to below-average outputs. The students were divided roughly equally in terms of their EVSs, whereas fewer than a quarter produced high-quality interpreting, with the majority falling into the low-quality group. To investigate their eye movement patterns, the by-group GCAs and the CPAs of the condition and the AOI effect described in Section 7.1. were repeated for each re-labelled group.

Table 7-19. Distribution of the professional and the student groups in the re-labelled groups

	Professional			**Student**		
	Early	Late	Sum	Early	Late	Sum
High-quality	7	10	17	4	4	8
Low-quality	2	3	5	14	13	27
Sum	9	13	22	18	17	35

7.6.1 Relationships between eye movements and lag in SI

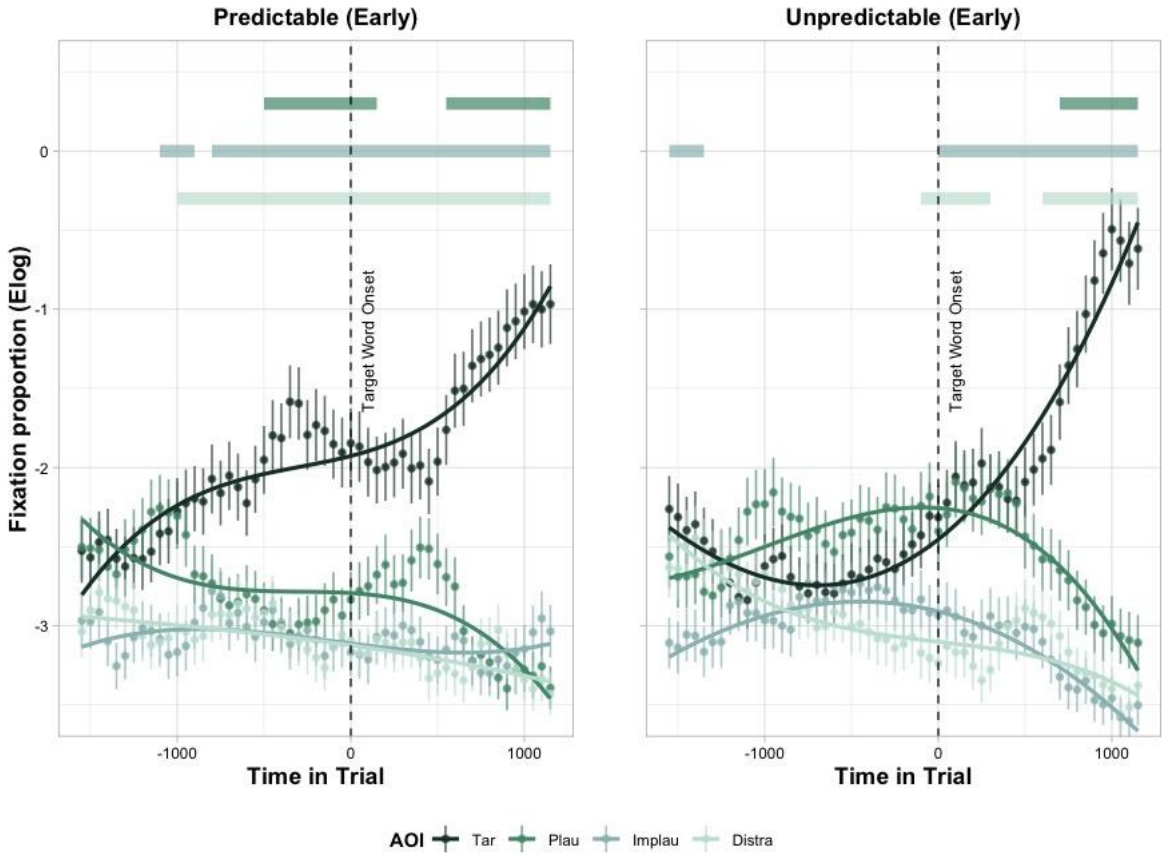
7.6.1.1 Results for the early interpreters

The GCA for the early interpreter group revealed a significant negative interaction between the condition and the Plau AOI (β = −0.578, *SE* = 0.050, *t* = −11.628, *p* < .001), indicating that fixation proportions on the plausible competitors were significantly lower in the predictable condition than in the unpredictable condition. Similarly, significant reductions in the unpredictable condition were observed for the implausible competitors (β = −0.301, *SE* = 0.050, *t* = −6.052, *p* < .001) and the distractors (β = −0.352, *SE* = 0.050, *t* = −7.089, *p* < .001),

suggesting that the early interpreters were more likely to consider non-target alternatives in the unpredictable condition.

With regard to the time terms, the predictable condition was significantly interacted with all time-terms. For the linear term (β = −0.824, SE = 0.260, t = −3.144, p = .002), the negative estimate indicates a flatter overall trajectory for fixations on the targets under the predictable condition. The interaction between predictability and the quadratic term was also significant (β = −2.382, SE = 0.261, t = −9.087, p < .001), reflecting the downward inflation in the predictable condition and the upward inflation in the unpredictable condition around the midpoint of the time course. The cubic term was also significant (β = 0.620, SE = 0.261, t = 2.366, p = .018), suggesting subtle differences in fixation fluctuations at the boundaries of the timeline. For the plausible competitors, interactions with the predictable condition were significant on both the quadratic (β = 3.816, SE = 0.371, t = 10.296, p < .001) and the cubic terms (β = −0.775, SE = 0.371, t = −2.092, p = .036). The significant positive estimate on the quadratic term was in the opposite direction to that for the targets, suggesting converse patterns for these two objects across conditions. The implausible competitors and distractors showed similar patterns, with significant interactions on the linear term (the implausible competitors: β = 1.497, SE = 0.370, t = 4.042, p < .001; the distractors: β = 1.634, SE = 0.370, t = 4.411, p < .001), and the quadratic terms (the implausible competitors: β = 3.581, SE = 0.371, t = 9.663, p < .001; the distractors: β = 1.931, SE = 0.371, t = 5.211, p < .001), indicating that the early interpreters gradually reduced attention to non-targets over time in the predictable condition. The robustness check confirmed the main findings, except that the effect on the cubic terms became less significant (see Appendix 13).

Figure 7-9. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the early interpreter group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

The CPA for the AOI effect revealed that two significant clusters for the TP contrast in the predictable condition, extending from −500 ms to 150 ms (cluster mass statistic = 50.864, $p < .001$), and from 550 ms to 1150 ms (cluster mass statistic = 63.308, $p < .001$). By contrast, only one significant cluster was observed in the unpredictable condition, occurring after the TW onset, from 700 ms to 1150 ms (cluster mass statistic = 45.168, $p < .001$). For the TI contrast, two significant clusters were found in the predictable condition: from −1100 ms to −900 ms (cluster mass statistic = 11.983, $p < .001$), and from −800 ms to 1150 ms (cluster mass statistic = 155.854, $p < .001$). In the unpredictable condition, two significant clusters emerged: from −1550 ms to −1350 ms (cluster mass statistic = 14.807, $p < .001$), and from the TW onset to 1150 ms (cluster mass statistic = 102.410, $p < .001$). The TD contrast showed a long, continuous cluster in the predictable condition, spanning from −1000 ms to 1150 ms (cluster mass statistic = 181.982, $p < .001$), suggesting an early exclusion of distractors in constraining

context. In the unpredictable condition, two significant clusters were identified: from −100 ms to 300 ms (cluster mass statistic = 26.047, $p < .001$), and from 600 ms to 1150 ms (cluster mass statistic = 63.176, $p < .001$). These findings support an early exclusion of non-target objects prior to the TW onset in the predictable condition, but not in the unpredictable condition.

The CPA for the condition effect revealed a positive cluster for the targets, extending from −850 ms to −100 ms (cluster mass statistic = 207.644, $p < .001$), indicating significantly higher fixation proportions for the targets in the predictable condition during this time span. Two negative clusters were identified for the plausible competitors, from −550 ms to 300 ms (cluster mass statistic = 169.716, $p < .001$) and from 650 ms to 900 ms (cluster mass statistic = 41.237, $p < .001$), suggesting that the early interpreters were more likely to fixate on the plausible competitors in the unpredictable condition than in the predictable condition. No significant cluster was identified for the implausible competitors, indicating consistent fixation patterns across conditions, whereas a negative cluster was observed for the distractors from 450 ms to 650 ms (cluster mass statistic = 26.256, $p < .001$).

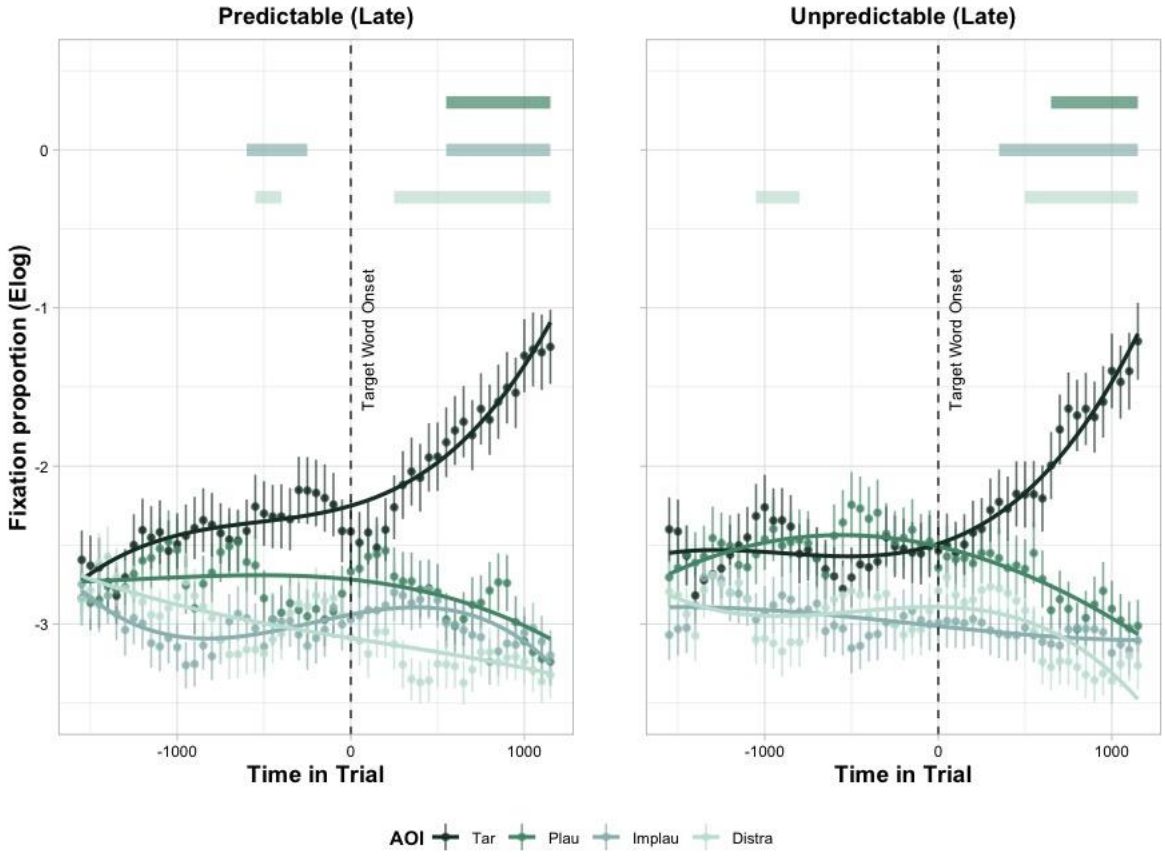
7.6.1.2 Results for the late interpreters

In the late interpreter group, the GCA also revealed significant interactions between the predictable condition and the three non-target AOIs: the plausible competitors ($\beta = -0.308$, $SE = 0.048$, $t = -6.412$, $p < .001$), the implausible competitors ($\beta = -0.147$, $SE = 0.048$, $t = -3.054$, $p = .002$), and the distractors ($\beta = -0.196$, $SE = 0.048$, $t = -4.068$, $p < .001$). These significant estimates were smaller than those in the early interpreter group ($\beta = -0.578$, −0.301, and −0.352, respectively), suggesting less pronounced between-condition differences in the late interpreter group.

In terms of temporal dynamics, the late interpreters only showed a significant condition effect on the quadratic term ($\beta = -0.669$, $SE = 0.253$, $t = -2.647$, $p = .008$), suggesting a limited modulation of predictability on a small portion of the time course. The three-way interactions among the condition, the AOI, and the time terms revealed fewer significant effects than those in the early interpreter group. The predictable condition × the plausible competitors was only significant on the quadratic term ($\beta = 1.115$, $SE = 0.358$, $t = 3.118$, $p = .002$); the predictable condition × the implausible competitors was marginally significant on the quadratic term ($\beta = 0.614$ $SE = 0.358$, $t = 1.716$, $p = .086$), and significant on the cubic term ($\beta = -0.842$, $SE = 0.358$, $t = -2.356$, $p = .018$). The robustness check showed largely consistent significance pattern, except reduced significance found in the interactions with the implausible competitors

on the quadratic term in the subsampled dataset (see Appendix 13). These results indicate that the late interpreters show a weaker and temporally delayed influence of predictability on their gaze patterns during sentence production.

Figure 7-10. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the late interpreter group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

The CPA for the AOI effect revealed fewer and later significant clusters in the late interpreter group. For the TP contrast, significant clusters appeared only after the TW onset in both conditions (predictable: from 550 ms to 1150 ms (cluster mass statistic = 51.134, $p < .001$); unpredictable: from 650 ms to 1150 ms (cluster mass statistic = 49.813, $p < .001$). For the TI contrast, two significant clusters were found in the predictable condition, from −600 ms to −250 ms (cluster mass statistic = 2.402, $p < .001$), and from 550 ms to 1150 ms (cluster mass statistic = 65.510, $p < .001$). In the unpredictable condition, only one significant cluster was identified, from 350 ms to 1150 ms (cluster mass statistic = 72.050, $p < .001$). For the TD contrast, two significant clusters were observed in each condition. In the predictable condition,

clusters extended from −550 ms to −400 ms (cluster mass statistic = 11.711, $p < .001$) and from 250 ms to 1150 ms (cluster mass statistic = 88.528, $p < .001$); in the unpredictable condition, from −1050 ms to −800 ms (cluster mass statistic = 18.781, $p < .001$) and from 500 ms to 1150 ms (cluster mass statistic = 73.445, $p < .001$).

The between-condition differences were also less pronounced in the late interpreter group, as shown by the CPA for condition effect. No significant clusters were observed for the targets or the implausible competitors, suggesting relatively consistent fixation patterns across conditions for these two objects. A negative cluster was identified for the plausible competitors, from −500 ms to −100 ms (cluster mass statistic = 81.761, $p < .001$), indicating higher fixation proportions in the unpredictable condition prior to the TW onset. Similarly, the distractor exhibited a negative cluster extending from 450 ms to 650 ms (cluster mass statistic = 26.256, $p < .001$).

7.6.1.3 Interim discussion

The present findings reveal clear differences in the predictive processing patterns of the early and the late interpreter groups, particularly in terms of the temporal dynamics and the magnitude of the predictive effects. Both groups showed evidence of modulating their eye movements in response to the contextual predictability; however, these effects were more pronounced and temporally earlier in the early interpreter group. Overall, the early interpreters showed more robust predictive fixations on the targets and stronger suppression of fixations on the other three objects in constraining contexts. This is supported by the larger estimates for the interaction between the condition and the AOIs on the quadratic term in the early interpreters, and by the presence of the significant cluster for the TP contrast prior to the TW onset, which was absent in the late interpreters. This stronger differentiation implies a greater sensitivity to the predictive cues associated with the CVs and a more refined ability to anticipate and constrain lexical competition during incremental sentence planning.

Regarding temporal dynamics, the early interpreter group showed widespread interactions between the condition and the time terms, with significant effects across all polynomial terms (linear, quadratic, and cubic). These interactions were present not only for the targets but also for the non-target AOIs, indicating that the early interpreters adjusted their gaze patterns in real-time to efficiently manage referential selection. By contrast, the late interpreter group exhibited fewer and weaker interactions with time, suggesting a delayed and more restricted integration of predictive cues, resulting in less dynamic modulation of visual attention over

time. These findings also align with the findings of Amos et al. (2022), who observed significantly higher proportion of fixations on the target versus the unrelated object in interpreters who began interpreting earlier.

There are two potential interpretations of this between-group difference. On one hand, deeper engagement of predictive processing may have enabled interpreters to commit to a referent selection earlier and to plan their outputs more efficiently, thereby facilitating the earlier initiation of sentence production. On the other hand, greater synchronicity between comprehension and production may have supported a more active involvement of the production system, which, according to the prediction-by-production account, facilitated greater and more effective prediction. These explanations are not mutually exclusive and may jointly contribute to the observed patterns. However, it remains uncertain whether increased prediction enhances interpreting quality.

7.6.2 Relationships between eye movements and SI quality

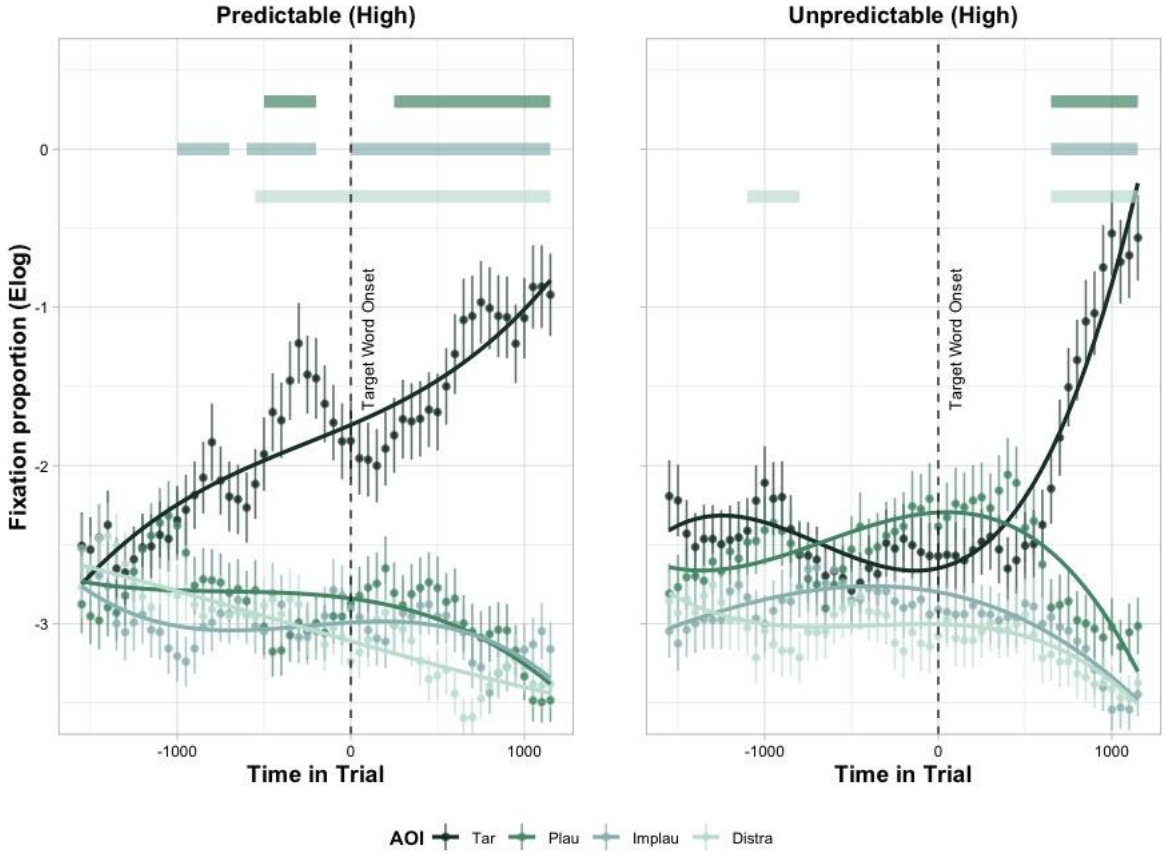
7.6.2.1 Results for the high-quality interpreters

The GCA for the high-quality interpreter group revealed a significant negative interaction between the condition and the Plau AOI ($\beta = -0.726$, $SE = 0.052$, $t = -13.882$, $p < .001$), indicating significantly lower fixation proportions on the plausible competitors in the predictable condition than in the unpredictable condition. Similar negative condition effects were observed for the implausible competitors ($\beta = -0.462$, $SE = 0.052$, $t = -8.836$, $p < .001$) and the distractors ($\beta = -0.362$, $SE = 0.052$, $t = -6.916$, $p < .001$), suggesting that high-quality interpreters reduced attention to non-target items when contextual constraint increased.

With regard to the time terms, the predictable condition significantly interacted with all-time terms: the linear ($\beta = 0.859$, $SE = 0.275$, $t = 3.118$, $p = .002$), the quadratic ($\beta = -2.990$, $SE = 0.276$, $t = -10.847$, $p < .001$), and the cubic terms ($\beta = -1.038$, $SE = 0.275$, $t = -3.771$, $p < .001$). The positive linear estimate reflects an overall increase in target fixations over time in the predictable condition, while the significant negative quadratic and cubic terms suggest a steeper peak and faster decline compared to the unpredictable condition. For the plausible competitors, significant interactions with the predictable condition emerged on the linear ($\beta = -1.537$, $SE = 0.390$, $t = -3.946$, $p < .001$), the quadratic ($\beta = 3.921$, $SE = 0.390$, $t = 10.064$, $p < .001$) and the cubic terms ($\beta = 1.551$, $SE = 0.389$, $t = 3.985$, $p < .001$), indicating divergent temporal dynamics from the targets. The implausible competitors showed a similar interaction on the quadratic term ($\beta = 3.998$, $SE = 0.390$, $t = 10.262$, $p < .001$), and a marginally significant

interaction on the cubic term (β = 0.670, SE = 0.389, t = 1.721, p = .085). The distractors also significantly interacted with all three-time terms: the linear (β = −1.771, SE = 0.390, t = −4.548, p < .001), the quadratic (β = 3.370, SE = 0.390, t = 8.649, p < .001), and cubic terms (β = 1.532, SE = 0.389, t = 3.937, p < .001). These significant three-way interactions suggest a faster decline in attention to non-target objects over time in the predictable condition. These findings were further confirmed by the robustness check with the subsampled dataset (see Appendix 14), although the significance on cubic-term interactions decreased, possibly due to the reduced sensitivity to higher-order interactions with smaller sample size.

Figure 7-11. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the high-quality interpreter group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

The CPA for the AOI effect revealed that in the predictable condition, there were two significant clusters for the TP contrast: from −500 ms to −200 ms (cluster mass statistic = 27.533, p < .001) and from 250 ms to 1150 ms (cluster mass statistic = 80.783, p < .001). Three significant clusters were identified for the TI contrast: from −1000 ms to −700 ms (cluster mass

statistic = 18.748, $p < .001$), from −600 ms to −200 ms (cluster mass statistic = 34.494, $p < .001$), and from the TW onset to 1150 ms (cluster mass statistic = 104.621, $p < .001$). The TD contrast only exhibited one significant cluster, extending from −550 ms to 1150 ms (cluster mass statistic = 153.285, $p < .001$). In the unpredictable condition, significant clusters only emerged after the TW onset extending from 650 ms to 1150 ms for both the TP contrast (cluster mass statistic = 54.617, $p < .001$) and the TI contrast (cluster mass statistic = 61.345, $p < .001$). The TD contrast, however, showed a significant cluster early before the TW onset, from −1100 ms to −800 ms (cluster mass statistic = 21.292, $p < .001$), and a significant cluster after the TW onset, from 650 ms to 1150 ms (cluster mass statistic = 64.885, $p < .001$).

The CPA for the condition effect revealed a significant positive cluster for the targets, extending from −550 ms to 750 ms (cluster mass statistic = 385.306, $p < .001$), suggesting a sustained condition effect even after the TW was explicitly mentioned in the audio input. In contrast, a significant negative cluster was found for the plausible competitors, from −500 ms to 450 ms (cluster mass statistic = 203.653, $p < .001$), indicating a strong preference for the plausible competitors both before and after the TW onset. No significant clusters were identified for the implausible competitors or the distractors.

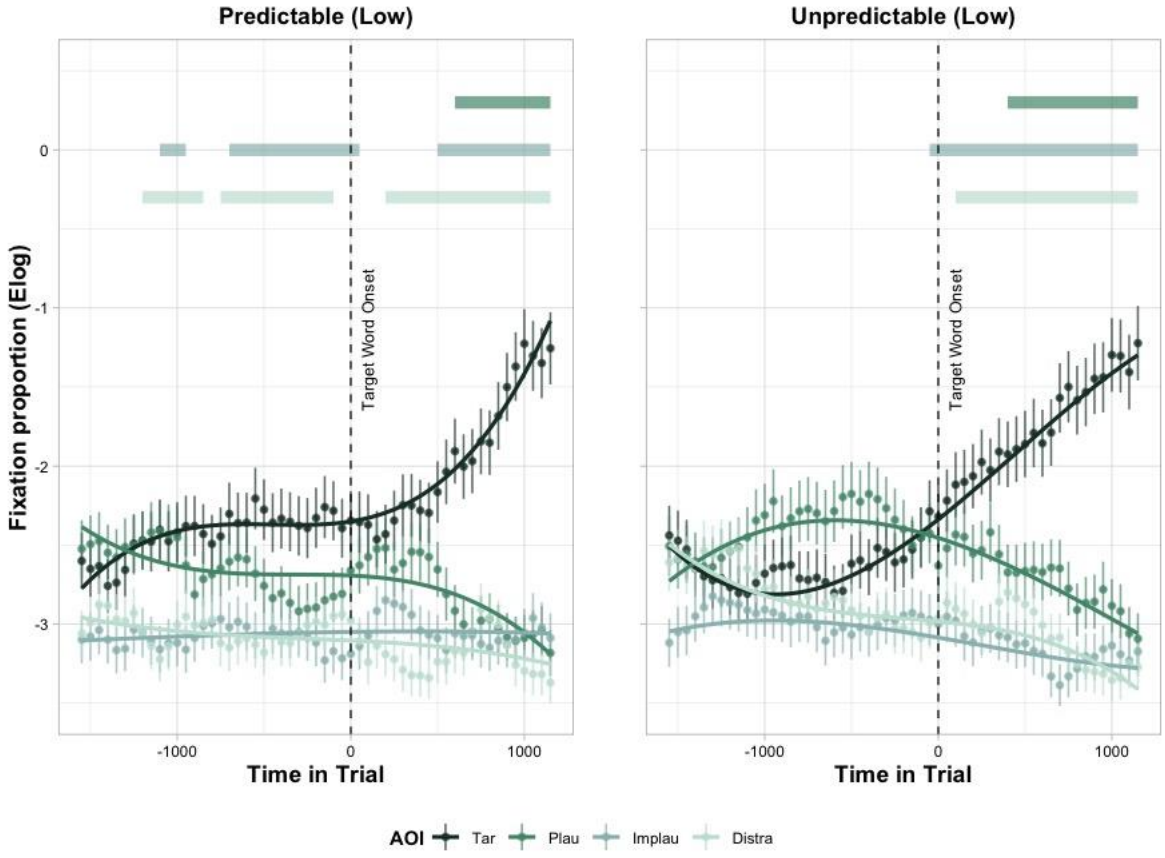
7.6.2.2 Results for the low-quality interpreters

In the low-quality interpreter group, the GCA revealed weaker interactions between the condition and the *AOI* compared to the high-quality group. The interaction between the condition and the plausible competitor was significant ($\beta = –0.231$, $SE = 0.046$, $t = –5.033$, $p < .001$), but notably smaller than that observed in the high-quality group ($\beta = –0.726$). There was no significant interaction between the condition and the implausible competitor in the low-quality group ($p = .271$), in contrast to the strong effect observed in the high-quality group ($\beta = –0.46$), suggesting that similar fixation patterns for the implausible competitors across conditions in the low-quality group. The interaction between the condition and the distractors was also significant ($\beta = –0.214$, $SE = 0.046$, $t = –4.646$, $p < .001$), but again smaller than that in the high-quality group ($\beta = –0.46$).

Temporal dynamics also differed. The predictable condition showed significant interactions with the linear ($\beta = −0.951$, $SE = 0.243$, $t = −3.912$, $p < .001$) and the cubic terms ($\beta = 1.422$, $SE = 0.242$, $t = 5.872$, $p < .001$). The absence of significant condition effect on the quadratic term suggests a lack of significant between-condition differences on the targets around the middle of the time course, in contrast to the pattern observed in the high-quality interpreter

group. The three-way interactions among the condition, the AOI, and the time terms also showed a narrower pattern of significance, particularly on the quadratic term. Only the plausible competitors exhibited a significant positive interaction with the predictable condition (β = 1.271, SE = 0.343, t = 3.71, p < .001). The robustness check also confirmed these key findings (see Appendix 14). These findings suggest that the low-quality interpreters exhibited weaker sensitivity to contextual predictability, with reduced suppression of non-target fixations and weaker facilitation of target fixations compared to the high-quality interpreter group.

Figure 7-12. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the low-quality interpreter group.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

Similar to the results of the late interpreters, the low-quality interpreters exhibited no significant clusters prior to the TW onset in either condition for the TP contrast. One significant cluster was identified after the TW onset in the predictable condition, extending from 600 ms to 1150 ms (cluster mass statistic = 48.086, p < .001). In the predictable condition, three significant clusters were identified for the TI contrast, from −1100 ms to −950 ms (cluster mass

statistic = 9.589, $p < .001$), from −700 ms to 50 ms (cluster mass statistic = 45.850, $p < .001$), and from 500 ms to 1150 ms (cluster mass statistic = 57.157, $p < .001$). Similarly, the TD contrast also identified three significant clusters, from −1200 ms to −850 ms (cluster mass statistic = 20.732, $p < .001$), from −750 ms to −100 ms (cluster mass statistic = 39.632, $p < .001$), and from 200 ms to 1150 ms (cluster mass statistic = 84.552, $p < .001$). In the unpredictable, each contrast had only one significant cluster: for the TP contrast, unpredictable: from 400 ms to 1150 ms (cluster mass statistic = 55.363, $p < .001$); for the TI contrast, from −50 ms to 1150 ms (cluster mass statistic = 11.598, $p < .001$); and for the TD contrast, from 100 ms to 1150 ms (cluster mass statistic = 93.190, $p < .001$).

The CPA for the condition effect also revealed a pattern largely consistent with that of the late interpreters. No significant clusters were identified for either the targets or the implausible competitors. By contrast, one significant negative cluster was identified for the plausible competitors, extending −550 ms to −50 ms (cluster mass statistic = 95.256, $p < .001$), suggesting significantly higher fixation proportions in the unpredictable condition. Two significant negative clusters were found for the distractors: from −1350 ms to −1050 ms (cluster mass statistic = 65.529, $p < .001$), and from 350 ms to 500 ms (cluster mass statistic = 36.747, $p < .001$).

7.6.2.3 Interim discussion

The comparison between the high- and the low-quality interpreter groups revealed substantial differences in both the magnitude and the timing of their gaze responses to contextual predictability. High-quality interpreters demonstrated stronger and more consistent evidence of predictive processing. These interpreters significantly reduced attention to non-target objects in the predictable condition, indicating a stronger ability to exploit contextual constraints to predict upcoming content and plan utterances. Their fixation patterns also reflected more dynamic temporal adjustments, suggesting a more flexible allocation of attention throughout the planning window.

In contrast, the low-quality interpreters exhibited weaker interactions overall and showed less extensive temporal effects, as evidenced by absence of quadratic interactions in the GCA, as well as by fewer significant clusters identified in the CPA. Although the predictable condition interacted significantly with the linear and cubic terms, the absence of a significant quadratic effect suggests that attention to the targets did not peak as clearly as in the high-quality group. Besides, in the unpredictable condition, the low-quality interpreters showed a stronger

preference to fixate on the plausible competitors, indicating a role of lexical association in modulating their fixation patterns. These results indicate that the low-quality interpreters were less adept at suppressing non-target objects through top-down analysis and less responsive to the contextual predictability. Instead, they may have relied more on surface-level semantic associations to predict upcoming content. Taken together, these findings suggest that high-quality interpreting performance is associated with more efficient suppression of non-target information and more precise alignment of visual attention with predictive cues during speech planning.

These findings align with and extend the comparison between the early and the late interpreter groups: the early interpreters showed earlier and more pronounced engagement in predictive processing, whereas the late interpreters demonstrated weaker and delayed sensitivity to predictive cues. The convergence of findings across production timing and interpreting output quality may reflect a shared underlying mechanism. Specifically, the higher-quality and the early interpreters appeared more efficient in processing and integrating predictive cues into utterance production, and they also allocated visual attention in ways that support rapid formulation and fluent delivery. Conversely, delayed initiation or lower-quality interpreting outputs may result from, or contribute to, weaker integration of predictive cues, leading to increased processing load and greater reliance on reactive strategies. However, despite the observed associations among production timing, predictive processing, and interpreting quality, one cannot conclude that earlier initiation or greater synchrony between comprehension and production, necessarily leads to improved predictive processing—or vice versa—and, in turn, finer interpreting output. Notably, more than half of the professionals who produced above-average interpreting outputs were classified as late interpreters, whereas the majority of the students who were classified as early interpreters contributed to below-average outputs. This finding suggests individual differences in the trade-off between production timing and output quality, which may be modulated by factors such as WM or personal strategic choices.

7.7 Relationships between eye movements and retrospection

The retrospective self-reports revealed that participants differed in their attitudes toward predictive processing during SI. While some participants reported active engagement in prediction and anticipation, some stated that they avoided such processing and instead focused on receiving auditory inputs. To explore whether these self-reported attitudes aligned with

participants' actual predictive behaviours, participants were categorised into three groups. Those who explicitly described active engagement were labelled as "anticipators"; those who reported never or rarely anticipating were labelled as "non-anticipators"; and those who expressed neutral or ambiguous attitudes were labelled as "neutral". Table 7-22 presents the distribution of participants across attitudes, EVS pattern, and interpreting quality. In the anticipator group, slightly more participants were classified as early interpreters (around 60%) than as late interpreters (40%), while interpreting quality was almost equally distributed. The non-anticipator group also showed an approximately equal split between early and late interpreters, as well as between the high- and the low-quality interpreters. In contrast, the neutral group included more low-quality interpreters (around 60 %) than high-quality ones (40%), and more late interpreters (about 60%) than early ones (40%). Because neutral participants did not explicitly express their predictive orientation, the by-group GCAs and the CPAs were conducted only for the anticipators and the non-anticipators.

Table 7-20. Distribution of the professional and the student groups in the re-labelled groups

	Anticipator			**Non-anticipator**			**Neutral**		
	Early	Late	Sum	Early	Late	Sum	Early	Late	Sum
High-quality	5	3	8	3	4	7	3	7	10
Low-quality	5	4	9	3	3	6	8	9	17
Sum	10	7	17	6	7	13	11	16	27

7.7.1 The by-group analysis

The GCA for the anticipator group revealed patterns similar to those of the early and the high-quality interpreters. The predictable condition significantly negatively interacted with the Plau AOI (β = –0.410, SE = 0.062, t = –6.579, p < .001), the Implau AOI (β = –0.261, SE = 0.062, t = –4.185, p < .001) and the Distra AOI (β = –0.263, SE = 0.062, t = –4.210, p < .001), suggesting that the anticipators more effectively narrowed their visual attention to the targets in the predictable condition. Regarding temporal dynamics, a significant interaction was found between the predictable condition and the quadratic term (β = –1.085, SE = 0.329, t = –3.302, p < .001), reflecting a sharper rise and fall in fixations. For the plausible competitors, interactions with the predictable condition were significant on both the linear (β = –1.967, SE = 0.465, t = –4.233, p < .001) and the quadratic terms (β = 2.039, SE = 0.464, t = 4.390, p < .001), indicating diverging fixation trajectories relative to the targets. Comparable interaction patterns were also observed for the Implau AOI on the linear (β = 2.297, SE= 0.465, t = 4.944,

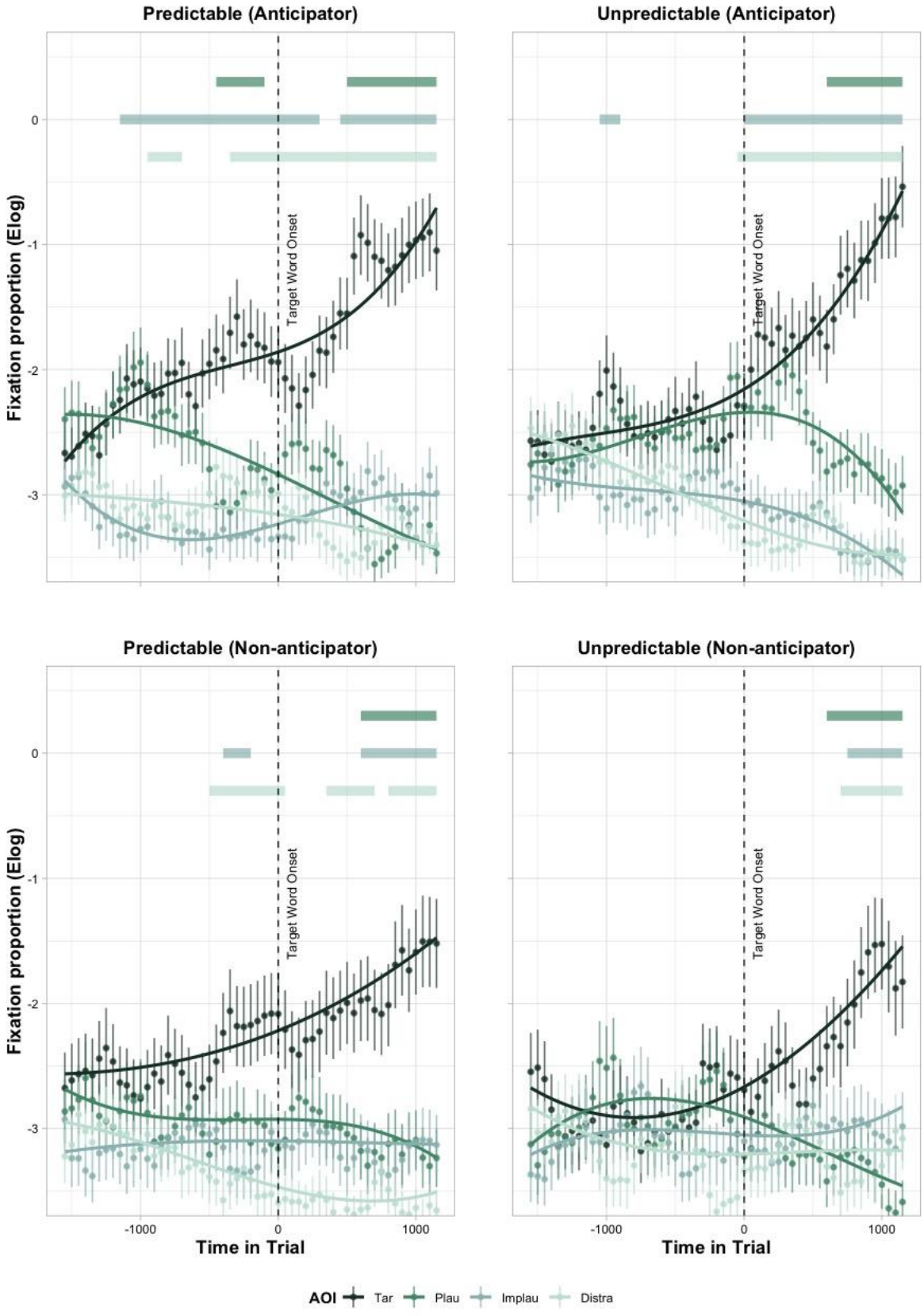
$p < .001$) and the quadratic terms ($\beta = 2.404$, $SE = 0.465$, $t = 5.176$, $p < .001$), and for the Distra AOI on the linear term ($\beta = 1.911$, $SE = 0.465$, $t = 4.114$, $p < .001$).

By contrast, the non-anticipator group showed weaker effects of predictability, like pattern observed in the late interpreters. Although fixation proportions on the plausible competitors were still significantly lower in the predictable condition ($\beta = –0.285$, $SE = 0.069$, $t = –4.119$, $p < .001$), and similar reductions were found for the implausible competitors ($\beta = –0.429$, $SE = 0.069$, $t = –6.207$, $p < .001$) and the distractors ($\beta = –0.549$, $SE = 0.069$, $t = –7.940$, $p < .001$), the fixation proportions to the four object were smaller than those found in the anticipators. The three-way interactions among the condition, the AOI, and the time terms were mostly non-significant. The only exception was a significant positive interaction on the quadratic term for the Plau AOI ($\beta = 1.845$, $SE = 0.514$, $t = 3.592$, $p < .001$), suggesting a less dynamic shift in gaze behaviour over time. The lack of significant interactions on the linear or cubic terms across all AOIs further indicates a more passive or reactive allocation of attention during predictive processing. These findings were further confirmed in the robustness check with the subsampled dataset (see Appendix 15).

The CPAs for the AOI effect were consistent with the GCA findings. In the anticipator groups, significant clusters were found for all three contrast pairs prior to the TW onset in the predictable condition, indicating that contextual predictability facilitated early suppression of non-target objects. In the unpredictable condition, the anticipator groups distributed their visual attention more evenly until the TW onset. A significant cluster for the TP contrast did not appear until 600 ms after the TW onset, suggesting increased hesitation between the targets and the plausible competitors strong predictive cues were absent.

The non-anticipator groups exhibited fewer and delayed significant clusters in both conditions. Especially in the predictable condition, only the TI and the TD contrasts showed significant clusters before the TW onset, and even these clusters were both shorter and appeared much later than those found in the anticipator group. The TP contrast exhibited a significant cluster only after 600 ms post-TW onset. These patterns suggest that non-anticipators may have excluded the implausible competitors and distractors before hearing the TW, but the exclusion occurred much later than those found in the anticipator group. In the unpredictable condition, no contrasts were significant until 600 ms after the TW onset, indicating that the non-anticipators delayed narrowing their attention to the targets until the TW was heard.

Figure 7-13. Time course of the fixation proportions of the four objects under each condition with the results of CPAs for the AOI effect for the anticipator (above) and the non-anticipator (below) groups.



#Notes: The solid smooth lines represent GCA model fitting results. The lines in the top (y = c(-0.3, 0, 0.3)) indicate the clusters where fixation proportions on the targets were high than each of the non-target objects, respectively.

7.7.2 Interim discussion

The present findings suggest that participants' self-reported attitudes to anticipation largely aligned with their predictive behaviour. Participants who reported actively anticipating upcoming content also demonstrated earlier and more pronounced suppression of non-target objects when strong predictive cues were available. Such suppression was not observed in those who reported little or no engagement in anticipatory strategies, although they also excluded the implausible competitors and the distractors prior to the TW onset in the predictable condition. The key differences between the two groups mainly reflect a divergence in the strategic use of predictive cues, both linguistic and extra-linguistic. While the anticipators actively utilised predictive cues and conducted top-down analysis to simulate upcoming content, reflecting the operation of prediction-by-production, the non-anticipators seemed to adopt a more reactive approach, either relying on bottom-up associative activation or waiting for information to became explicitly available in the auditory input. Furthermore, the non-anticipator group exhibited generally lower fixation proportions across all objects, suggesting reduced engagement with the visual input and possibly a more uni-modal focus on auditory comprehension. Notably, given that the anticipators and the non-anticipators did not significantly differ in their EVS patterns or interpreting quality, it remains uncertain whether attitudes towards predictive processing directly influence overall SI performance, or whether they instead reflect different but equally effective cognitive styles.

Few SI studies, if any, have integrated self-reported attitudes with eye-tracking and performance metrics to investigate individual differences in metacognitive awareness and strategic use of anticipation. In psycholinguistics, prediction is typically conceptualised as a largely automatic process without an explicit intention, especially when driven by associative mechanisms. Even prediction-by-production, which is more effortful and controlled, is considered optional in the sense that it is activated primarily when certain conditions are met, such as high contextual constraint or sufficient cognitive capacity (Pickering and Gambi, 2018). The present findings, however, offers a nuanced view of this "optionality". Beyond being constrained by stimulus- or participant-level features, predictive processing during SI appears to be modulated by interpreters' deliberate, strategic choices. The ability to suppress or initiate

predictive processing based on task demands or personal strategy implies a level of inhibitory and metacognitive control, which is not often emphasised in prediction models. In this sense, predictive processing in SI is not merely an automatic cognitive process but can also be consciously deployed as a flexible strategy for managing complex language processing under time pressure.

Chapter 8 General discussion

8.1 Presence of predictive processing during SI of coherent discourse

To address RQ1, the present study provides compelling evidence for successful semantic prediction during SI of multi-sentence paragraphs in both professional and student interpreters. Both groups made predictive eye movements to the targets prior to the TW onset in the predictable condition, whereas no such predictive eye movements were observed in the unpredictable condition. This pattern corroborates H1 and suggests the role of contextual probability in guiding predictive processing during SI.

Compared to previous SI studies using the visual-world paradigm (Amos et al., 2022; Liu et al., 2022), the prediction effect observed in the present study appeared slower and smaller. A likely explanation lies in the differences in SI task complexity. Although both the present study and Liu et al. (2022) used verb-mediated sentences to elicit prediction, the key distinction between the two studies is the length and structure of the ST. The present study used coherent multi-sentence discourse, requiring interpreters to maintain coherence and reference across sentences. This inter-sentential continuity meant participants had to retain contextual information for longer stretches sometimes even throughout entire paragraphs. In contrast, in Liu et al. (2022), sentences were isolated, and participants could clear stored information from WM after each one. The richer textual information and longer text duration may have overstretched participants' cognitive capacities (Kuperberg & Jaeger, 2016), who thus exhibited delayed or reduced predictive eye movements. This interpretation is also supported by the observed significant positive effects of preceding sentence length and within-paragraph position on TW-EVS. As the ST progressed, participants tended to extend the latencies between the input and the output. The reduced synchrony between comprehension and production may in turn delay and reduce their engagement of predictive processing.

Besides, participants in Liu et al. (2022) interpreted from their L1 to L2, while in the present study the interpreting direction was the other way around, from participants' L2 to L1. Presumably, participants were more fluent in their L1 than in L2, and interpreting from L2 likely imposed greater cognitive load, which may have slowed predictive processing (Kaan & Grüter, 2021; Ito & Pickering, 2021). This also aligns with findings from Ito et al. (2018), who reported that delayed predictive eye movements in L2 speakers relative to L1 speakers during language comprehension tasks.

Amos et al. (2022) adopted a different approach, using a target-absent design, where targets and competitors were tested in separate trials. In trials where only the target objects were

presented, the absence of competitors allowed more fixations on the targets, especially given that the targets were more strongly associated with the context than unrelated distractors. In contrast, the current study presented all four objects simultaneously, introducing competition and dividing attention. Another potential reason for the lower fixation proportions in the present study is the higher ecological validity of the present study. Participants were encouraged to perform SI as naturally as possible, and some reported in their retrospections that they typically focused more on auditory input in real-life SI practices, sometimes even closing their eyes to reduce visual distraction. It is thus possible that some participants glanced minimally at the visual display, or even completely ignored the visual stimuli by staring at the blank part of the screen. Alternatively, others may have recognised the presented objects quickly and even successfully predicted the targets. To reduce visual interference, they also looked at the blank part of the screen. Since blinks and looks outside the four AOIs were factored into the calculation of fixation proportions, these behaviours contributed to the lower fixation proportions on all objects.

Unlike Hodzik and Williams (2017) and Chmiel (2021), who observed shorter EVSs in highly constraining contexts, the present study found longer EVSs in the predictable condition. Specifically, Hodzik and Williams (2017) observed shorter EVS for sentence-final verbs in contextual constraining contexts for both shadowing and SI tasks, whereas the present study found significantly longer CV-EVS in the predictable condition and a positive correlation between CV-EVS and cloze probability for both the professionals and the students. The difference in verb latency patterns between Hodzik and Williams (2017) and the present study likely arises from differences in syntactic alignment between the source and the target languages. In Hodzik and Williams (2017), participants interpreted from German into English. German follows a subject-object-verb (SOV) structure, in which the main verb is placed at the end of the clause, whereas English follows a SVO structures, in which the verb appears earlier in the sentence. This syntactic mismatch imposes a need for prediction, especially for verbs that are crucial for sentence interpretation and target language reformulation. In highly constraining contexts, preceding contextual cues facilitated prediction for the upcoming German verbs, thereby compensating for syntactic delay in the source language and leading to shorter verb latencies in the target language. In contrast, the present study involved SI from English to Chinese, both of which follow an SVO word order. This syntactical alignment means that the verb appears early and in the same relative position in both the source and the target sentences. The preceding propositional information provided limited predictive cues for the

verbs. Meanwhile, the CVs in the predictable condition in the present study were less frequent and more semantically specific, requiring greater cognitive effort for lexical retrieval and planning. These demands may have led interpreters, especially the professionals, to strategically delay their output in favour of more accurate and contextually appropriate reformulation, thus yielding longer CV-EVS in the predictable condition.

Chmiel (2021) measured the latencies for the sentence-final nouns and also found shorter EVS in highly constraining contexts. The present study, however, revealed marginally longer TW-EVS in the predictable condition for the professionals, and no significant between-condition differences for the students. These differences likely reflect task demands and strategic variation based on expertise. In Chmiel (2021), participants were required to translate the final word of a sentence as quickly as possible. In this task, no extended discourse was involved, and thus the contribution of preceding context was limited. There was no demand for speech planning, and participants naturally focused on the single demands of production speed. The faster translation in high-constraint contexts reflects contextual predictability facilitating rapid access and retrieval of the targets in a simplified, time-pressured task. However, in the present study's SI task, participants had to manage more complicated cognitive processing, including discourse-level coherence, syntactic restructuring, and articulation under time pressure. The TW, which usually appeared in the later part of the experimental sentences, were more vulnerable to these increased cognitive demands. For professional interpreters, who often adopt more strategic planning during SI, high contextual constraint likely enabled them to engage in fine-grained production planning and reformulation. This strategic approach, aimed at optimising fluency and accuracy, may have contributed to the extension of both CV-EVS and TW-EVS in the predictable condition. In contrast, student interpreters may have relied more on surface-level processing and prioritised synchrony between input and output. As a result, their TW-EVS did not vary significantly across conditions, but their CT-span was shorter in the predictable condition, indicating more immediate response to predictive cues without deeper integration or restructuring.

In sum, the present study observed predictive eye movements to the targets when contextual constraints were strong, suggesting the presence of semantic prediction during SI of coherent discourse. However, the predictability effect observed in the present study appeared slower and weaker compared to those reported in previous SI study using the visual-world paradigm (Amos et al., 2022; Liu et al., 2022), possibly due to increased SI task complexity arising from larger and more contextually rich ST. Unlike earlier studies that used latency measures to

examine predictability effects (Hodzik & Williams, 2017; Chmiel, 2021), which found shorter target word latencies in highly constraining contexts, the present studies revealed a trend toward longer EVSs in the predictable condition, and positive correlations between cloze probability and EVS measures. This divergence may be attributed to two key factors. First, syntactic alignment across language pairs (e.g., the SOV-SVO mismatch in German-to-English SI versus the SVO-SVO match in English-to-Chinese SI) affects where and how prediction can be used most effectively. Second, the nature of the production task itself differs. Unlike simpler word translation tasks or short clause-level processing, the present study's discourse-level SI task likely encouraged deeper reformulation and strategic planning, making prediction more effortful and potentially delaying production.

8.2 The underlying mechanism of predictive processing during SI

To address RQ2, the present study revealed a dual route of prediction-by-production and prediction-by-association during SI, supporting H2. The clearest evidence comes from distinct fixation patterns across the two conditions. In the predictable condition, fixation proportions were significantly higher on the targets than on the other three objects, suggesting top-down processing of the global context that overrode local lexical association between the predictive CVs and the implausible competitors. In the unpredictable condition, plausible competitors attracted more fixations, indicating a synergistic effect of both contextual plausibility and bottom-up activation based on lexical associations between the unpredictive CVs and the plausible competitors. These fixation patterns cannot be accounted for by either mechanism alone. Additional support comes from the EVS data. For both the professionals and the students, cloze probability exerted positive effects on CV-EVS, sentence offset EVS, and sentence-span, whereas verb-noun association exerted negative effects on TW-EVS, sentence onset EVS, sentence offset EVS, and the two duration measures. The positive effects of cloze probability indicate effortful top-down processing triggered by higher contextual predictability, sustaining throughout the sentences. In contrast, the negative effects of verb-noun association suggest fast, automatic activation of lexical items via bottom-up processing.

Participants' retrospective reports further reinforced this dual-route account. World-based and discourse-based anticipations suggested that participants actively imitated the speaker's perspective, which was also explicitly mentioned by several participants, derived the communicative intention, and simulated an output speech through the production system.

Visual-based and lexical-based anticipations appeared more reactive and associative, often triggered by local lexical cues. Although association-based prediction is often automatic and less cognitively demanding, it does not mean that such prediction is unconscious. Instead, bottom-up activations of associated representations may be integrated into the mental model even faster because of its low processing cost.

Consistent with Hintz et al. (2017), the present study did not find verb-noun association strength to be a reliable predictor of predictive eye movements. However, the absence of significant association effect should not be interpreted as contradicting the involvement of prediction-by-association. Hintz et al. (2017) reported that functional associations, which were grounded in structured event knowledge, played a more important role than general verb-noun associations in guiding predictive eye movements. On one hand, functional associations likely involve some level of top-down processing, such as analysing typical verb-noun relationships (e.g., "peel" and "banana"), which aligns with the prediction-by-production account (Pickering & Gambi, 2018). On the other hand, more general associative links may only exert influence when the context does not strongly constrain interpretation. In Hintz et al.'s design, for example, the visual presence of "peelable" objects may have reinforced functionally relevant associations and made them more salient for anticipatory eye movements. In the present study, the processing context was more complex, and the evolving mental context may have exerted stronger influence on the directions of association. Therefore, it is possible that the effect of general verb-noun associations was overridden by more specific associations that were most relevant to prediction. Future research should distinguish among different types of associations (e.g., thematic, taxonomic, functional) and clarify how these interact with global discourse context to support predictive processing during SI.

The present study partially contradicts the findings from Kukona et al. (2011), who observed that strong lexical-level associative priming could temporarily override sentence-level contextual constraints, particularly in early time windows. For instance, in the sentence "*Toby arrests the…*", even though the subjective role was already filled (e.g., *Toby*), participants still directed anticipatory fixations to the competitor agent (e.g., *policeman*), which was strongly associated with the predictive verb (e.g., *arrest*) but obviously implausible in the context. In the present study, such post-verb predictive fixations driven by lexical association were only observed in the unpredictable condition but not in the predictable condition. A potential explanation is that the associative strength between the implausible competitors and the predictive CVs was not strong enough to override the global context, which thereby suppressed

the bottom-up lexical activation. This interpretation is consistent with Kukona et al.'s own conclusions that the relative influence of top-down contextual constraint and bottom-up association is gradient, with strong global constraint capable of attenuating or cancelling associative effects. The divergence between the present study and Kukona et al. (2011) also fits with the different task designs across two studies. The SI task in the present study involved extended discourse, possibly building a more coherent and strongly activated global mental model than the shorter, single-sentence contexts used by Kukona et al. (2021), making the interpreter participants even more sensitive to thematic plausibility.

Additionally, the present study observed higher fixation proportions on the targets and the plausible competitors than on the implausible competitors and the distractors prior to the TW onset in both conditions. The early exclusion of the implausible competitor and the distractor likely reflects a thematic priming effect from the global context, which occurred even earlier before the CV onset. Rather than being a result of deliberate strategic prediction, the mental model constructed from prior discourse may have operated more as an associative backdrop, which activated semantically or visually related representations. Items that were congruent with this contextual mental model, e.g., "*bread*" (the target) and "*juice*" (the plausible competitor) at a train station in an early morning, received a processing advantage. Such processing advantage guided visual attention in an associative yet context-sensitive manner. This interpretation aligns with Ferretti, McRae, and Hathere (2001), who showed that verbs can immediately prime typical thematic role fillers (e.g., agents or instruments), indicating that conceptual schemes, such as common situations or events, are rapidly activated during comprehension. These findings suggest a wider conceptualisation of prediction-by-association, which operates not only at the lexical level but also at broader discourse levels.

Taken together, the findings of the present study demonstrate an interactive, layered model of predictive processing in which both prediction-by-production and prediction-by-association operate in parallel during SI of coherent discourse. As the ST unfolds, interpreters incrementally construct and update a mental model of the discourse, which in turn activates networks of semantically and contextually associated representations. In cases when visual inputs are present, these inputs too may activate associated representations and be integrated into the developing mental model. These activated representations, arising from both linguistic and visual cues, are either suppressed or strengthened depending on their relevance to the evolving context. Simultaneously, interpreters engage their language production system to simulate the speaker's communicative intention, incorporating both linguistic and

extralinguistic cues (e.g., world knowledge, situational cues). This simulation guides predictive planning and facilitates coherent reformulation in the target language.

The results are also consistent with the two-system or dual processing streams during predictive language processing (Huettig, 2015; Kuperberg, 2007). On one hand, participants demonstrated active engagement of top-down analysis in contextually rich conditions. This aligns with the "smart" route, which is goal-directed and sensitive to structure linguistic knowledge and thematic-semantic context. On the other hand, the bottom-up activations also played a role in guiding participants' prediction when contextual constraints were weak, which aligns with the "dumb" route that relies on simpler mechanisms through spreading activations based on the semantic memory. Furthermore, the observed expertise-related differences in the engagement of predictive processing during SI cannot be fully explained by the one-system account (Altmann & Mirković, 2009), which assumes a unified constraint-based mechanism underlying prediction. Given that the professionals and the students exhibited similar L2 proficiency, the between-group differences likely reflect variation in cognitive control and strategic attention allocation, which are not explicitly addressed in the one-system account.

8.3 Relationship between predictive processing and interpreting performance

To address RQ3, the present study revealed systematic individual differences in predictive processing during SI, as evidenced by variations in eye-movement patterns, EVS control, interpreting quality, and self-reported anticipatory strategies. The results partly align with H3. Both the early and the high-quality interpreters exhibited more robust and temporally precise prediction effects: they directed more predictive fixations towards the targets and suppressed the non-target objects more effectively in highly constraining contexts. These patterns emerged earlier in the time course and were accompanied by dynamic adjustments to visual attention over time, suggesting both heightened sensitivity to predictive cues and more efficient integration of these cues into real-time sentence planning. In contrast, the late and the low-quality interpreters showed weaker and delayed prediction effects, relying more on reactive processing and exhibiting reduced discrimination between the targets and the non-target objects.

Crucially, participants' subjective attitudes toward anticipation aligned with these behavioural patterns. The anticipators, who reported actively engaging in anticipatory strategies, showed

predictive fixation patterns similar to those of the early interpreters, including early suppression of the non-target objects and greater focus on the targets in the predictable condition. Conversely, the non-anticipators, who expressed a tendency to avoid anticipation, resembled the late interpreters, showed lower overall visual engagement with less differentiated fixations, and relied more on reactive, bottom-up processing. This convergence of subjective and behavioural data suggests that predictive processing during SI is not solely driven by external task features or automatic linguistic mechanisms. Rather, the alignment between EVS control and self-reported strategy use indicates the role of strategic and metacognitive regulation in shaping, or at least reflecting, one's ability to manage EVS and deploy predictive processing during SI.

These findings contribute to recognising the role of individual strategies in modulating predictive processing during SI. While psycholinguistic theories traditionally frame prediction, particularly prediction-by-association, as an automatic and unintentional process, the present study supports an expanded view. Specifically, predictive processing during SI, especially when driven by production-based mechanisms, can be selectively engaged based on the interpreter's goals, attentional control, and task management strategies. High-performing interpreters, whether defined by synchronised EVS, superior interpreting rendition quality, or self-reported use of anticipation, tended to engage in more effortful, top-down prediction. This likely involved the active simulation of speaker intentions, analysis of global discourse context, and use of structural linguistic knowledge to anticipate upcoming content, all of which support strategic alignment between comprehension and production. In contrast, lower-performing groups seemed to rely more on bottom-up mechanisms such as surface-level lexical association or perceptual salience, which may be less cognitively demanding but also less precise. While both routes can facilitate prediction, the production-based pathway enables more strategic reformulation, contributing to fluent and contextually coherent output.

Then why do the high-performing interpreters appear more likely to engage in deliberate top-down analysis? One possibility is that they may have managed the sub-components of SI more efficiently, freeing up cognitive resources to be reallocated toward predictive simulation. Instead of being overwhelmed by immediate input, these interpreters may use their WM and attentional control to operate at a higher level of abstraction. This interpretation aligns with cognitive resource model of SI (e.g., Gile's Effort Model) and broader theories of working memory capacity and executive attention (Baddeley, 2007; Dong & Cai, 2015), which suggests that prediction engagement is moderated by cognitive resource availability.

That said, the present study does not establish a causal relationship between predictive processing and improved SI performance. While the presence of predictive processing often co-occurred with higher-quality output, it was not a necessary condition. Some early interpreters produced relatively poor renditions, and some late interpreters achieved high accuracy. This variability implies that predictive processing during SI may be weighed against other priorities such as accuracy, risk management, or cognitive efficiency. For example, some interpreters may delay output to reduce processing load or avoid premature commitment to specific semantic or syntactic interpretations. This interpretation aligns with the adaptability of prediction proposed by Liu et al. (2022) as well as the "utility view of prediction" proposed by Kuperberg and Jaeger (2016). They argued that predictive behaviour is adjusted dynamically by language users through a cost-benefit analysis of prediction for achieving communicative goals.

However, the evidence does point to an important mediating role of predictive processing, especially when effortfully engaged, in facilitating more refined utterance planning. Specifically, interpreters who engaged in more active, top-down predictive processing appeared to show greater syntactic and lexical flexibility, enabling more accurate and fluent reformulation under pressure. Therefore, even in the absence of a direct performance benefit or universal advantage, predictive processing may serve as one of multiple strategies that interpreters flexibly deploy depending on task demands and personal strengths and that supports higher-order planning and coherence maintenance in SI.

In sum, the present study provides compelling evidence that predictive processing during SI is a dynamic, individualised process that reflect not only the automatic linguistic mechanisms affected by the structure of the input, but also strategic metacognitive control shaped by the interpreter's cognitive resources and task approach. These findings highlight the importance of integrating behavioural, cognitive, and subjective data to understand the variability and adaptability of prediction in demanding language tasks like SI. Future research should explore how automatic and strategic components of predictive processing interact dynamically over time, and whether anticipatory strategies can be enhanced through targeted training or cognitive interventions.

8.4 Expertise-related differences in the cognitive rhythms and strategic control of predictive processing

To address RQ4, the present study reveals distinct expertise-related differences, with the professional interpreters generally outperforming interpreting students, supporting H4.The two groups exhibited similar patterns in predictive fixations to the targets in the predictable condition and more evenly distributed fixations across four objects in the unpredictable condition. However, several key distinctions emerged. The professionals displayed more robust target-focused prediction and more dynamic visual attention shifts even after fixating the target objects in the predictable condition, reflecting heightened vigilance for possible continuations. In contrast, the students showed weaker target bias and had more static gaze distributions, particularly shifting more attention toward the plausible competitors before the TW onset in the unpredictable condition. This suggests that the professionals maintain a broader attentional window and flexibly manage competing interpretations, while the students commit more quickly to a single projection and interpretation.

In terms of EVS, both groups tended to extend EVS in the predictable condition and showed comparable timelines for CV production. Both groups produced higher-frequency CVs more quickly but experienced greater difficulty in producing higher-frequency TWs. They also demonstrated similar positive effects of cloze probability and negative effects of verb-noun association on EVS measures. Discourse level factors modulated EVS in both groups in comparable ways: longer preceding sentences and later sentence positions within a paragraph were associated with longer EVS and duration measures. However, the professionals generally exhibited longer EVSs than the students, and this distinction mainly appeared in the later parts of the sentences in the predictable condition. First, the professional showed significantly longer TW-EVS and marginally longer sentence offset EVS and CT-span than the students, especially in the predictable condition. Second, while the professionals maintained similar CT-span across both conditions, the students showed significantly faster transitions between CV and TW in the predictable condition. These findings suggest that the professionals exhibited greater flexibility in coping with increased processing demands associated with less frequent CV and were better able to utilise predictive cues to support utterance planning and reformulation. By contrast, the students appeared more sensitive to the lexical level variations and allocated fewer cognitive resources to effortful, top-down predictive processing. Instead, they relied more on reactive processing, prioritising input-output synchrony and adopting a streamlined strategy for producing their interpreting rendition.

This interpretation is further supported by their interpreting quality outcomes. The professionals outperformed the students across all three subscale scores and in overall scores. Additionally, the professionals maintained stable lexical- and sentence-level accuracies across conditions, whereas the students exhibited significantly lower CV and sentence accuracies in the predictable condition. Further evidence comes from participants' self-reported accounts. The professionals provided more detailed and reflected retrospections. Despite their higher interpreting quality, they more frequently acknowledged instances of inaccurate interpreting, suggesting improved meta-cognitive control and self-monitoring. In contrast, the students more frequently reported omitting modifiers and focusing on content words, aligning with the use of a streamlined strategy in output production. They also more often mentioned visual-based anticipation, suggesting a greater reliance on association-based predictive processing.

The significant expertise-related distinctions observed in the present study partly contradicts the finding of Amos et al. (2022). While Amos et al. also reported more pronounced predictive eye movements in the interpreter group, they found no significant between-group differences in the prediction patterns. The divergence between the two studies may stem from differences in visual stimuli design: Amos et al. (2022) employed a target-absent design, whereas the present study presented the targets and the competitors within the same visual display. In the target-absent design, participants were more likely focus on the critical objects, as the other three objects were unrelated to the sentence context or the target, semantically, phonologically, or perceptually. Consequently, both groups showed fewer attentional shifts across conditions. By contrast, in the target-present design, the co-occurrence of competitor objects was likely to attract more visual attention, as they were either compatible with the CV or aligned with the broader context. As a result, participants, especially professional interpreters with greater cognitive flexibility, were more likely to shift their visual attention, leading to group-level variations in prediction patterns.

The present findings also partly contrast with Özkan et al. (2023), who reported no evidence of prediction among student interpreters in a language comprehension task. One possible explanation is that, despite the simplicity of the language task, prediction in Özkan et al. (2023) relied on processing case markers, a syntactic cue generally more demanding than semantic prediction. As discussed earlier, prior studies have yielded mixed findings regarding syntactic prediction in L2 comprehension. While advanced or near-native L2 speakers occasionally demonstrated the ability to utilise syntactic information predictively (Foucart et al., 2014; Grüter et al., 2012; Hopp & Lemmerth, 2018), such prediction remains less stable than

observed in L1 speakers and is typically absent among intermediate L2 speakers (Kamide et al., 2003; Kaan et al., 2014; Mitsugi, 2017). Besides, Özkan et al. (2023) used independent sentences as stimuli, limiting the availability of discourse-level context to support prediction. As the present study suggests, student interpreters tend to rely more on superficial, associative processing of predictive cues and thus are less likely to engage in complex analysis of syntactic structures for anticipation.

Consistent with Liu et al. (2022), the present study found significantly lower sentence accuracy in the predictable condition among students. Liu et al. interpreted this reduced accuracy as evidence for the detrimental effect of prediction on SI performance. The present study, however, interprets it as a reflection of expertise differences. Participants in Liu et al. (2022) were bilinguals without interpreting experience, whereas the present study recruited both professional interpreters and interpreting students with at least one year of formal interpreting training. Therefore, the findings of Liu et al.'s more closely resemble those observed among the students in the present study. Although Liu et al.'s participants also exhibited signs of semantic prediction, their prediction was likely based on simpler, associative mechanisms, especially since the targets in their study were more strongly associated with the CVs than the other three distractors. Furthermore, in both studies, predictive CVs were less frequent than unpredictive CVs, suggesting increased lexical retrieval difficulty in the predictable condition. Like the student interpreters in the present study, the bilingual participants in Liu et al. (2022) may have experienced greater processing difficulty associated with less frequent CV in the predictable condition and struggled to utilise predictive cues effectively to facilitate output production. Therefore, the reduced sentence accuracies observed in Liu et al. (2022) likely reflect lower efficiency in lexical access and a limited capacity for top-down contextual analysis to compensate for this difficulty.

The present study observed longer EVS in professionals, consistent with Janikowski and Chmiel (2025). However, the two studies offer different interpretations of this pattern. Janikowski and Chmiel (2025) suggested that the positive correlation between EVS and interpreting experience might have been inconclusive in their study due to a narrow range of experience among participants, in contrast to Timarová et al. (2014), who reported the opposite trend with a broader experience range. The present findings suggest that longer EVS observed in professionals may reflect deeper engagement in higher-order cognitive processes, including more precise lexical retrieval, more refined syntactic structuring, and more coherent semantic

integration. These extended EVS did not impair interpreting quality; rather, they supported more elaborate and contextually appropriate target language production.

Taken together, the present findings reveal substantial expertise-related differences in both the cognitive rhythms and the strategic control of predictive processing. Professional interpreters demonstrated greater efficiency and flexibility in deploying predictive cues and engaged in more extensive predictive processing, particularly through top-down, production-based mechanism. This in turn facilitated more refined utterance planning and reformulation. Although such effortful processes prolonged the EVS, the professionals ultimately produced more syntactically integrated and semantically coherent interpreting renditions. In contrast, student interpreters, while showing evidence of predictive processing, relied more on reactive, bottom-up strategies rooted in simple associative activation. They prioritised input-output synchrony, committing to specific interpretations earlier without adequately preparing for multiple possible continuations. This likely reflects a streamlined strategy that favours immediacy over flexibility and accuracy.

Then what accounts for these expertise-related differences? Although the WM capacity was not directly measured in the present study, previous research has linked higher WM capacity to enhanced prediction in language comprehension (Federmeier & Kutas, 2005; Huettig & Janse, 2016; Ito, Corley, et al., 2018). However, WM effects may be moderated by interpreting experience. Both Lozano-Argüelles et al. (2023) and Özkan et al. (2023) reported interactions between WM capacity and interpreting experience: higher WM capacity benefited predicative processing only in the professionals, not in interpreting trainees or non-interpreter L2 learners. This interaction effect suggests that WM capacity may not directly modulate the predictive processing. Instead, WM may enable broader activation or longer maintenance of potential continuations, but its effectiveness on predictive processing depends on the interpreter's ability to manage and suppress irrelevant representations. In this view, interpreting expertise may facilitate the flexible and goal-oriented allocation of cognitive resources, optimising prediction based on task demands. This view also aligns with the adaptability view of predictive processing during SI discussed in the last section: predictive processing during SI is not shaped only by external task features but also modulated by metacognitive control with a cost-benefit trade-off to achieve communicative goals. Supporting this, neuroimaging studies have shown that interpreting experience is associated with structural and functional adaptations in the brain, particularly in regions related to cognitive control and attention (e.g., García, 2019; Hervais-Adelman, Moser-Mercer, & Golestani, 2015, 2015; Hervais-Adelman, Moser-Mercer, Michel,

& Golestani, 2015). These neural changes may underlie experts' superior ability to regulate prediction strategically during SI, reinforcing the notion that predictive processing in SI is shaped not only by linguistic or memory constraints but also by learned, experience-driven control mechanisms.

Chapter 9 Conclusion

9.1 Summary of the major findings

The present study examined the presence, mechanisms, potential effects, and expertise-related differences in predictive processing during SI of coherent discourse. Using a combination of visual-world eye tracking paradigm, EVS measures, interpreting quality scoring, and retrospective self-reports, the study provided comprehensive evidence that both professional and student interpreters engaged in predictive processing during SI. However, the nature and extent of this engagement vary significantly across individuals and levels of expertise. First, clear evidence was found for semantic prediction during SI: in highly constraining contexts, both the professionals and the students made predictive eye movements to the target objects prior to the TW onset, whereas no such predictive fixations were observed in the unpredictable contexts. These predictive effects were time-locked to the speech and modulated by contextual constraints, confirming the role of discourse-level information in facilitating prediction during naturalistic SI tasks.

Second, the findings support a dual-mechanism account of predictive processing, involving both prediction-by-production and prediction-by-association. Fixation patterns indicated that top-down, production-based mechanisms operated alongside bottom-up associative activations. Specifically, in the predictable condition, interpreters fixated more on the targets and suppressed non-target objects, suggesting effortful, context-driven prediction. In contrast, in the unpredictable condition, fixations were more evenly distributed with slightly more fixations directed to the plausible competitors guided by lexical links, reflecting association-based prediction. These dual mechanisms were further supported by the EVS data: cloze probability was positively associated with EVS and sentence spans, indicating effortful discourse-based planning, while verb-noun association strength was negatively associated with EVS measures, aligning with rapid, automatic activation. Retrospective reports further corroborated this distinction, with interpreters reporting world-based or discourse-based anticipation (indicative of production-based processing) as well as visual or lexical cues triggering more reactive anticipation.

Third, the findings suggest a potential link between predictive processing and SI performance. Early interpreters, high-quality interpreters, and self-reported “anticipators” exhibited more robust, earlier, and more target-focused predictive eye movements, as well as greater flexibility in attentional shifts and more consistent suppression of non-target items. These groups also demonstrated longer EVSs and more dynamic integration of contextual cues into their output,

suggesting strategic engagement in top-down planning. Conversely, late interpreters, lower-quality interpreters, and "non-anticipators" displayed delayed and weaker prediction effects, lower overall visual engagement, and stronger reliance on reactive, bottom-up processing. Importantly, the alignment across production timing, interpreting quality, and self-reported attitudes suggests a critical role of metacognitive control in shaping predictive strategies during SI.

Finally, the study revealed clear expertise-related differences. Professional interpreters exhibited more dynamic gaze patterns and longer and more consistent EVSs, whereas interpreting students showed more static gaze patterns and shorter and less stable EVSs. The professionals outperformed students in overall quality and item-based accuracies, and they also exhibited more reflective self-monitoring. While both groups engaged in predictive processing, the professionals relied more on production-based strategies, integrating linguistic, contextual, and visual cues to anticipate and reformulate the upcoming content. The students, by contrast, relied more on associative cues and adopted a streamlined synchronised input-output strategy. These differences were attributed to greater cognitive flexibility, WM control, and strategic experience among the professionals. Overall, the study provides compelling evidence that predictive processing in SI is not merely an automatic response to linguistic input, but also a flexible, strategic behaviour shaped by individual cognitive resources, professional expertise, and task-specific demands.

9.2 Implications for interpreting practice and pedagogy

The findings of the present study carry important implications for interpreting practice and the pedagogical strategies used in interpreting training. Most notably, while prediction has often been viewed as a passive or incidental by-product of language processing, the current study suggests that skilled interpreters actively and strategically engage in both prediction-by-production and prediction-by-association to manage the complexities of SI. For interpreting practitioners, these findings highlight the value of cultivating flexible, anticipatory skills. These skills enable the interpreter not only to react to the source text in real time but also to actively simulate potential upcoming content. This capacity allows for smoother reformulation, better management of cognitive load, and more contextually appropriate renditions, especially in high-pressure, fast-paced interpreting environments.

For interpreting pedagogy, these findings call for more explicit training in predictive strategies. Traditional interpreting training often prioritises listening comprehension, note-tasking, and memory-based reformulation, with less emphasis on proactive processing. However, the evidence from this study suggests that high-performing interpreters strategically deploy predictive processing to improve output quality and fluency, Therefore, interpreter training programmes could incorporate structured exercises that help students practice anticipatory processing. For example, instructors might present partially completed utterances or discourse contexts and ask students to predict the next phrase, word, or idea. Tasks that train discourse-level inferencing, verb-noun collocation prediction, and scenario-based projection can enhance students' ability to simulate SL content and prepare TL output in advance.

The study also highlights the importance of developing production-based predictive skills. Since production-based prediction involves actively simulating the speaker's communicative intention and formulating output based on both global discourse context and local lexical coherence, it places higher demands on WM and attentional control. Therefore, interpreter training should not only develop linguistic knowledge and vocabulary but also strengthen high-order cognitive capacities, such as WM updating, cognitive flexibility, and goal-directed attention. Pedagogical strategies might include dual-task exercises that train attention splitting, such as shadowing with reformulation under varying contextual constraints, or paraphrasing under time pressure to simulate realistic cognitive demands. Importantly, students should be encouraged to reflect on their processing strategies, as the study found that metacognitive awareness (as evidenced in self-reported attitudes) aligned closely with predictive behaviour and performance. Incorporating guided self-evaluation, peer feedback, and retrospective analysis into training could help students become more conscious of how and when they engage in prediction and how to improve SI performance accordingly.

Another pedagogical implication lies in the integration of visual context in training. The study showed that predictive fixations were influenced not only by linguistic cues but also by visual information. While real-life SI rarely involves visual displays like those in experimental settings, interpreters frequently encounter visual stimuli in the form of slides, speaker gestures, or environmental cues. Training interpreters to use these visual inputs strategically, e.g., anticipating topics from slide titles, deducing speaker intent from body language, can strengthen multi-modal prediction. Classroom tasks that involve interpreting with visual stimuli, or with disrupted visual context to examine compensatory strategies, could enrich students' ability to navigate complex, multimodal interpreting situations.

9.3 Innovations and limitations

The present study offers several methodological and theoretical innovations that advance our understanding of predictive processing during SI. Most notably, it employed coherent multi-sentence discourse rather than isolated sentences as the auditory stimuli. This design mirrors the naturalistic structure of real-world speech more closely than previous studies, thereby enhancing ecological validity. Unlike single-sentence stimuli, coherent discourse requires interpreters to maintain global coherence and integrate information across sentence boundaries, both of which place greater cognitive demands on prediction mechanisms. Therefore, the use of coherent discourse as stimuli allows for a more nuanced examination of how predictive processing operates in realistic SI conditions, where interpreters must balance immediate comprehension with broader discourse-level planning and production.

Another key innovation lies in the triangulation of eye-tracking data, behavioural performance metrics, and retrospective self-reports, allowing for a multi-dimensional investigation of predictive processing during SI. This integrative approach moves beyond the traditional psycholinguistic assumption that prediction is automatic and often unconscious. Instead, the findings suggest that predictive processing in SI can also be consciously regulated and strategically deployed, particularly by professional interpreters. By incorporating both cognitive and metacognitive dimensions of anticipation, the study provides a more holistic account of predictive processing as a flexible, individualised process influenced by task demands interpreter expertise, and strategic intent. In addition, the study used multiple complementary eye-tracking analysis methods, including GCA and CPA. While GCA captured fine-grained temporal dynamics of predictive fixations, CPA provided statistical validation of these gaze patterns without imposing rigid time windows. The convergence of findings across these two methods lends robustness to the study's conclusions.

Despite these strengths, several limitations should be acknowledged. First, while the use of coherent discourse improves ecological validity, the literary style of the ST introduces limitations. Its descriptive narration, infrequent idiomatic structures, and stylistically rich expressions are not representative of the typical speech encountered in real-time SI practice, such as political speeches, technical presentations, or spontaneous dialogue. As such, some findings, particularly those related to the processing of less frequent lexical items or longer sentences, may not generalise to all SI contexts. Interpreters may engage differently with

prediction when dealing with more domain-specific content. Future studies should consider using a range of speech styles and genres to examine how stylistic variation affects predictive processing. Second, while the combination of eye-tracking, EVS measures, and retrospective self-reports offers a richer dataset, the study is still correlational in nature. As such, it cannot establish causal relationships between predictive processing and interpreting performance. Although associations between predictive behaviour and output quality were observed, these links may be mediated by other factors such as working memory capacity, attentional control, or task familiarity, which were not directly measured in this study.

Third, this study did not include transitional probability as an objective measure in assessing predictive processing. Defined as the likelihood of a word given its preceding context, transitional probability has been widely used in psycholinguistic research as a key index of distributional language learning and anticipation (e.g., Hodzik & Williams, 2017). While cloze probability and association strengths capture human judgements of predictability, transitional probability offers a complementary, corpus-driven metric that captures distributional patterns independent of human ratings. Incorporating such a measure in future research could help disentangle the contributions of semantic association, syntactic structure, and distributional frequency to interpreters' predictive behaviour. It would also allow for a more fine-grained analysis of how interpreters implicitly learn and exploit statistical regularities in real-time, particularly when processing domain-specific or recurring discourse structures.

Fourth, the sample size, particularly within relabelled-group comparisons (e.g., anticipators vs. non-anticipators), limits the statistical power to detect more subtle interactions. This limitation is further compounded by imbalanced trial numbers across participants, which may have reduced sensitivity to detect higher-order effects. While robustness checks helped validate core findings, some marginal or subgroup-specific effects should be interpreted cautiously. This limitation is especially relevant for SI research, where trial balancing is difficult to enforce due to the naturalistic nature of stimuli. Future studies could consider more controlled stimulus design or statistical approaches that better account for unbalanced data structures. Finally, while retrospective reports provided valuable insight into strategic awareness, such self-reports are inherently subjective and may be affected by memory biases or participants' self-perception. Incorporating real-time verbal protocols or neurophysiological measures could further clarify the cognitive underpinnings of predictive processing in future research.

9.4 Avenues for future research

Building on the findings and limitations of the present study, several promising avenues emerge for future research on predictive processing in SI. First, future research should extend beyond semantic prediction to investigate more complex forms of syntactic and phonological predictions. While semantic cues often offer the most immediate basis for anticipatory behaviour, syntactic structures, such as gender, case markers, or word order, may also serve predictive functions, particularly in morphologically rich or syntactically flexible languages. Similarly, phonological cues could become especially relevant in low-context or high-speed interpreting settings. These prediction types may operate under different cognitive constraints and timelines and are thus worth examining in greater detail. Future studies should employ more varied stimulus materials, including not only narrative and descriptive texts but also spontaneous speech, argumentative discourse, and technical communication. The accompanying visual stimuli could also be diversified, such as using printed words, schematic diagrams, gestures, or real-time visual aids (e.g., slides), to better reflect the range of inputs interpreters encounter in practice.

Second, further research is needed to disentangle the types and directions of lexical associations that influence prediction during SI. The present study focused on verb-noun association strength, but prior work (e.g., Hintz et al., 2017) has shown that functional associations, which reflect thematic roles or affordances, may have more predictive value than general co-occurrence patterns. Researchers could examine how interpreters prioritise different associative pathways (e.g., from verbs to agents, instruments, or goals) and how these preferences vary with discourse context or expertise. Additionally, the role of cross-linguistic associative asymmetry, particularly in L1–L2 vs. L2–L1 interpreting, warrants further investigation.

Third, although the current study inferred potential links between WM capacity and predictive processing, future research should directly assess how individual cognitive abilities modulate predictive engagement, including (visual-spatial, numeric, and linguistic) WM span, inhibitory control, processing speed, and attentional flexibility. A combined behavioural and neurocognitive approach, using tools such as reading span tasks, pupillometry, or EEG, could yield a deeper understanding of how interpreters allocate and manage cognitive resources. This would be especially informative in situations where prediction is needed most, such as under cognitive overload, during rapid speech, or in syntactically demanding segments.

Fourth, the adaptability of predictive strategies should be a focus of longitudinal and intervention-based studies. While the present findings suggest that experienced interpreters exhibit more strategic, production-based predictive processing, it remains an open question whether such skills can be enhanced through targeted prediction-related training. It is also uncertain whether improvements in predictive processing leads to measurable benefits in SI performance. Longitudinal studies tracking interpreters-in-training over extended periods, ideally with eye-tracking and output analysis at multiple timepoints, could help determine the role of predictive engagement in the development of interpreting expertise.

Fifth, future studies may explore the interaction between predictive processing and modality, such as sign language interpreting, remote interpreting, or multimodal interpreting scenarios involving video feeds and other non-verbal information. These contexts place unique demands on anticipatory skills and may shift the balance between bottom-up and top-down processes in unexpected ways. Similarly, investigating the impact of fatigue, stress, or divided attention on predictive processing could reveal more about the robustness and limits of anticipation in real-world SI settings.

Lastly, more work is needed to explore metacognitive aspects of predictive processing. The present study hinted that interpreters can flexibly modulate their use of prediction based on perceived task demands and personal strategy. Future research should further explore the role of metacognitive awareness and monitoring in shaping predictive behaviour, possibly using strategy elicitation tasks, or adaptive training environments that explicitly instruct and evaluate different predictive approaches.

Taken together, these future directions offer a rich framework for expanding our understanding of predictive processing in SI, from its cognitive underpinnings and linguistic variability to its pedagogical applications and real-world implementation.

Bibliography

Adamowicz, A. (1989). The role of anticipation in discourse: Text processing in simultaneous interpreting. *Polish Psychological Bulletin, 20*(2), 153–160.

Allan, K. (1977). Classifers. *Language, 53*(2), 285-311.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439. https://doi.org/10.1006/jmla.1997.2558

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory and Cognition, 24*(4), 477-492. https://doi.org/10.3758/BF03200936

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*(2), 190-200. https://doi.org/10.1016/j.actpsy.2010.09.009

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583-609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Amini, M. (2018). How to evaluate the TEFL students' translations: Through analytic, holistic or combined method? *Language Testing in Asia, 8*(1), 10. https://doi.org/10.1186/s40468-018-0063-6

Amos, R. M., & Pickering, M. J. (2020). A theory of prediction in simultaneous interpreting. *Bilingualism: Language and Cognition, 23*(4), 706-715. https://doi.org/10.1017/S1366728919000671

Amos, R. M., Seeber, K. G., & Pickering, M. J. (2022). Prediction during simultaneous interpreting: Evidence from the visual-world paradigm. *Cognition, 220*, 104987. https://doi.org/10.1016/j.cognition.2021.104987

Amos, R. M., Seeber, K. G., & Pickering, M. J. (2023). Student interpreters predict meaning while simultaneously interpreting - even before training. *Interpreting, 25*(2), 211-238. https://doi.org/10.1075/intp.00093.amo

Anderson, L. (1994). Simultaneous interpretation: Contextual and translation aspects. In S. Lambert, & B. Moser-Mercer, *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 101–120). Amsterdam: John Benjamins. https://doi.org/10.1075/btl.3.11and

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica, 137*(2), 208-216. https://doi.org/10.1016/j.actpsy.2011.01.007

Anokhin, P. (1978). *Philosophical aspects of the theory of function systems.*

Antón-Méndez, I. (2020). The role of verbs in sentence production. *Frontiers in Psychology, 11*, 189. https://doi.org/10.3389/fpsyg.2020.00189

Ashby, J., Rayner, K., & Clifton, C. J. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Qarterly Journal of Experimental Psychology, 58*(6), 1065-1086. https://doi.org/10.1080/02724980443000476

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2007). *Working memory, thought, and action.* New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528012.001.0001

Barik, H. C. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and Speech, 16*(3), 237–270. https://doi.org/10.1177/002383097301600307

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457-474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J., Jackson, L., & Phillips, I. (2014). Using a voice to put a name to a face: The psycholinguistics of proper name comprehension. *Journal of Experimental Psychology: General, 143*(1), 404 – 413. https://doi.org/10.1037/a0031813

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

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*(4), 343-355. https://doi.org/10.1016/0013-4694(85)90008-2

Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory and Cognition, 8*(6), 631-642. https://doi.org/10.3758/BF03213783

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. (2003). Minding the clock. *Journal of Memory and Language, 48*(4), 653-685. https://doi.org/10.1016/S0749-596X(03)00007-X

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 15*(3), 607-624. https://doi.org/10.3758/s13415-015-0340-0

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology, 64*(3), 545-559. https://doi.org/10.1080/17470218.2010.503374

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*(5), 412-424. https://doi.org/10.1027/1618-3169/a000123

Camblin, C. C., Ledoux, K., Boudewyn, M., Gordon, P. C., & Swaab, T. Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research, 1146*, 172-184. https://doi.org/10.1016/j.brainres.2006.07.033

Canseco-Gonzalez, E., Brehm, L., Brick, C. A., Brown-Schmidt, S., Fischer, K., & Wagner, K. (2010). Carpet or Cárcel: The effect of age of acquisition and language mode on bilingual lexical access. *Language and Cognitive Processes, 25*(5), 669-705. https://doi.org/10.1080/01690960903474912

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 687–696. https://doi.org/10.1037/0278-7393.30.3.687

Charlton, R. A., Schiavone, F., Barrick, T. R., Morris, R. G., & Markus, H. S. (2010). Diffusion Tensor Imaging detects age-related white matter change over a two-year follow-up which is associated with working memory decline. *Journal of Neurology, Neurosurgery and Psychiatry, 81*(1), 13-19. https://doi.org/10.1136/jnnp.2008.167288

Chen, J., Yang, H., & Han, C. (2022). Holistic versus analytic scoring of spoken-language interpreting: a multi-perspectival comparative analysis. *The Interpreter and Translator Trainer, 16*(4), 558–576. https://doi.org/10.1080/1750399X.2022.2084667

Chen, S., Kruger, J.-L., & Doherty, S. (2020). Reading patterns and cognitive processing in an eye-tracking study of note-reading in consecutive interpreting. *Interpreting, 23*(1), 76-102. https://doi.org/10.1075/intp.00050.che

Chen, X., & Dong, Y. (2019). Evaluating objective and subjective frequency measures in L2 lexical processing. *Lingua, 230*, 102738. https://doi.org/10.1016/j.lingua.2019.102738

Chernov, G. V. (1994). Message redundancy and message anticipation in simultaneous interpretation. In S. Lambert, & B. Moser-Mercer, *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 139-155). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.3.13che

Chernov, G. V., Setton, R., & Hild, A. (2004). *Inference and Anticipation in Simultaneous Interpreting : A probability-prediction model.* Philadelphia: John Benjamins Publishing.

Chmiel, A. (2021). Effects of simultaneous interpreting experience and training on anticipation, as measured by word-translation latencies. *Interpreting, 23*(1), 18-44. https://doi.org/10.1075/intp.00048.chm

Chou, C.-J., Huang, H.-W., Lee, C.-L., & Lee, C.-Y. (2014). Effects of semantic constraint and cloze probability on Chinese classifier-noun agreement. *Journal of Neurolinguistics, 31*, 42-54. https://doi.org/10.1016/j.jneuroling.2014.06.003

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*(1), 3–42. https://doi.org/10.1017/S0142716406060024

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-253. https://doi.org/10.1017/s0140525x12000477

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Routledge. https://doi.org/10.4324/9780203774441

Collard, C., & Defrancq, B. (2019). Predictors of ear-voice span, a corpus-based study with special reference to sex. *Perspectives, 27*(3), 431–454. https://doi.org/10.1080/0907676X.2018.1553199

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*(1), 84-107. https://doi.org/10.1016/0010-0285(74)90005-X

Copland, D. A., de Zubicaray, G. I., McMahon, K., Wilson, S. J., Eastburn, M., & Chenery, H. J. (2003). Brain activity during automatic semantic priming revealed by event-related functional magnetic resonance imaging. *NeuroImage, 20*(1), 302–310. https://doi.org/10.1016/s1053-8119(03)00279-9

Cui, Y., & Zheng, B. (2022). Extralinguistic Consultation in English–Chinese Translation: A Study Drawing on Eye-Tracking and Screen-Recording Data. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.891997

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12*(3), 453-459. https://doi.org/10.3758/BF03193787

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia, 50*(8), 1852–1870. https://doi.org/10.1016/j.neuropsychologia.2012.04.011

de Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behaviour Research, 45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

de Groot, A. M. (2011). *Language and cognition in bilinguals and multilinguals.* New York: Psychology Press. https://doi.org/10.4324/9780203841228

de Kloe, Y. J., Hooge, I. T., Kemner, C., Niehorster, D. C., Nyström, M., & Hessels, R. S. (2022). Replacing eye trackers in ongoing studies: A comparison of eye- tracking data quality between the Tobii Pro TX300 and the Tobii Pro Spectrum. *Infancy, 27*(1), 25-45. https://doi.org/10.1111/infa.12441

Defrancq, B. (2015). Corpus-based research into the presumed effects of short EVS. *Interpreting, 17*(1), 26-45. https://doi.org/10.1075/intp.17.1.02def

Dell, G. S., & Chang, F. (2013). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Díaz-Galaz, S., Padilla, P., & Bajo, M. T. (2015). The role of advance preparation in simultaneous interpreting: A comparison of professional interpreters and interpreting students. *Interpreting, 17*(1), 1–25. https://psycnet.apa.org/doi/10.1075/intp.17.1.01dia

Dink, J. W., & Ferguson, B. (2015). *eyetrackingR: An R library for eye-tracking data analysis.* Retrieved from https://github.com/jwdink/eyetrackingR. https://doi.org/10.32614/CRAN.package.eyetrackingR

Doi, K., Sudoh, K., & Nakamura, S. (2021). Large-scale English–Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stuke, & E. Salesky, *Proceedings of the 18th international conference on spoken language translation (IWSLT 2021)* (pp. 226–235). Bangkok: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.iwslt-1.27

Dong, Y., & Cai, R. (2015). Working memory and interpreting: A commentary on theoretical models. In Z. Wen, M. Borges Mota, & A. McNeill, *Working memory in second language acquisition and processing* (pp. 63-82). Bristol: Channel View Publications. https://doi.org/10.21832/9781783093595-008

Dosher, B. A., & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General, 118*(2), 191-211.

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. *Cognition, 110*(2), 284-292. https://doi.org/10.1016/j.cognition.2008.11.012

Dussias, P. E., Kroff, J. V., Guzzardo, T. R., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition, 35*(2), 353-387. https://doi.org/10.1017/S0272263112000915

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641-655. https://doi.org/10.1016/S0022-5371(81)90220-6

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics, 25*(1), 55–98. https://psycnet.apa.org/doi/10.1515/cog-2013-0031

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211. https://doi.org/10.1016/0364-0213(90)90002-E

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology, 68*(8), 1469-1488. https://doi.org/10.1080/17470218.2014.984730

Fan, D. C., Collart, A., & Chan, S.-h. (2022). When two languages are competing: An ERP study of sentence processing in expert and novice interpreters. *Interpreting, 24*(1), 1-37. https://doi.org/10.1075/intp.00069.fan

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469-495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., & Kutas, M. (2005). Aging in context: Age-related changes in context use during language comprehension. *Psychophysiology, 42*(2), 133-141. https://doi.org/10.1111/j.1469-8986.2005.00274.x

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*(2), 164-203. https://doi.org/10.1016/S0010-0285(03)00005-7

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language, 69*(3), 165–182. https://doi.org/10.1016/j.jml.2013.06.001

Ferretti, T. R., McRae, K., & Hathere, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*(4), 516-547. https://doi.org/10.1006/jmla.2000.2728

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2001). Syntactic working memory and the establishment of filler-gap dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research, 30*(3), 321-338. https://doi.org/10.1023/A:1010447102554

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research, 37*(5), 617-631. https://doi.org/10.1016/S0042-6989(96)00218-0

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition, 14*(3), 379-399. https://doi.org/10.1017/S136672891000012X

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning Memory and Cognition, 40*(5), 1461-1469. https://doi.org/10.1037/a0036756

Fowler, Y. (2007). Formative assessment: Using peer and self-assessment in interpreter training. In C. Wadensjö, B. E. Dimitrova, & A.-L. Nilsson, *The Critical Link 4: Professionalisation of interpreting in the community* (pp. 253-262). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.70.28fow

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*(2), 178-210. https://doi.org/10.1016/0010-0285(82)90008-1

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language, 95*, 200-214. https://doi.org/10.1016/j.jml.2017.04.007

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 31(5), 862-877. https://doi.org/10.1037/0278-7393.31.5.862

García, A. M. (2019). *The neurocognition of translation and interpreting.* Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.147

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighborhood density: Effects on spoken word recognition by children and adults. *Journal of Memory and Language, 45*(3), 468-492. https://doi.org/10.1006/jmla.2000.2784

Gee, J. P. (2014). *An introduction to discourse analysis: Theory and method.* London: Routledge. https://doi.org/10.4324/9781315819679

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*(2), 256-281. https://doi.org/10.1037//0096-3445.113.2.256

Gerver, D. (1969). The effects of source language presentation rate on the performance of simultaneous conference interpreters. In E. Foulke, *Proceedings of the Second Louisville Conference on Rate and/or Frequency-Controlled Speech* (pp. 162–184). Louisville: Center for Rate-Controlled Recordings, University of Louisville.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Gile, D. (1992). Predictable sentence endings in Japanese and conference interpretation. *The Interpreters' Newsletter*, 12-23.

Gile, D. (2002). Conference interpreting as a cognitive management problem. In F. Pöchhacker, & M. Shlesinger, *The interpreting studies reader* (pp. 196-214). London: Routledge.

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum, 6*(2), 59-77. https://doi.org/10.1075/forum.6.2.04gil

Gile, D. (2009). *Basic concepts and models for interpreter and translator training.* Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.8

Grüter, T., Lau, E., & Ling, W. (2018). L2 listeners rely on the semantics of classifiers to predict. In A. B. Bertolini, & M. J. Kaplan (Ed.), *Proceedings of the 42nd annual Boston University conference on language development* (pp. 303-316). Somerville, MA: Cascadilla Press.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*(2), 191-215. https://doi.org/10.1177/0267658312437990

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274-279. https://doi.org/10.1111/1467-9280.00255

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267–283. https://doi.org/10.3758/BF03204386

Gunter, T. C., Jackson, J. L., & Mulder, G. (1995). Language, memory, and aging: An electrophysiological exploration of the N400 during reading of memory-demanding sentences. *Psychophysiology, 32*(3), 215-229. https://doi.org/10.1111/j.1469-8986.1995.tb02951.x

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid linguistic ambiguity resolution in young children with autism spectrum disorder: Eye tracking evidence for the limits of weak central coherence. *Autism Research, 8*(6), 717-726. https://doi.org/10.1002/aur.1487

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly, 13*(3), 186-201. https://doi.org/10.1080/15434303.2016.1211132

Han, C. (2022). Interpreting testing and assessment: A state-of-the-art review. *Language Testing, 39*(1), 30-55. https://doi.org/10.1177/02655322211036100

Han, C., & Lu, X. (2021). Interpreting quality assessment re-imagined: The synergy between human and machine scoring. *Interpreting and Society, 1*(1), 70-90. https://doi.org/10.1177/27523810211033670

Henderson, J. M., & Ferreira, F. (2004). *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging, 18*(4), 755-769. https://doi.org/10.1037/0882-7974.18.4.755

Hervais-Adelman, A., Moser-Mercer, B., & Golestani, N. (2015). Brain functional plasticity associated with the emergence of expertise in extreme language control. *NeuroImage, 114*, 264-274. https://doi.org/10.1016/j.neuroimage.2015.03.072

Hervais-Adelman, A., Moser-Mercer, B., Michel, C. M., & Golestani, N. (2015). fMRI of simultaneous interpretation reveals the neural basis of extreme language control. *Cerebral Cortex, 25*(12), 4727-4739. https://doi.org/10.1093/cercor/bhu158

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(9), 1352-1374. https://doi.org/10.1037/xlm0000388

Hodzik, E., & Williams, J. N. (2017). Predictive processes during simultaneous interpreting from German into English. *Interpreting, 19*(1), 1-20. https://doi.org/10.1075/intp.19.1.01hod

Hoffman, J. V. (1980). Studying contextual build-up during reading through cumulative cloze. *Journal of Reading Behaviour, 12*(4), 337-341. https://doi.org/10.1080/10862968009547387

Holmqvist, K., Orbom, S. L., Hooge, I. T., Niehorster, D. C., Alexander, R. G., Andersson, R., . . . Hessels, R. S. (2023). Eye tracking: Empirical foundations for a minimal reporting guideline. *Behavior Research Methods, 55*, 364–416. https://doi.org/10.3758/s13428-021-01762-8

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research, 22*(3), 369-397. https://doi.org/10.1191/0267658306sr272oa

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29*(1), 33-56. https://doi.org/10.1177/0267658312461803

Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching, 53*(3), 277-306. https://doi.org/10.1515/iral-2015-0014

Hopp, H., & Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition, 40*(1), 171-199. https://doi.org/10.1017/S0272263116000437

Huang, H.-W., Meyer, A. M., & Federmeier, K. D. (2012). A “concrete view” of aging: Event related potentials reveal age-related changes in basic integrative processes in language. *Neuropsychologia, 50*(1), 26-35. https://doi.org/10.1016/j.neuropsychologia.2011.10.018

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology, 58*(3), 376-415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition, 200*, 104251. https://doi.org/10.1016/j.cognition.2020.104251

Hudson, P. T., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language, 24*(1), 46–58. https://doi.org/10.1016/0749-596X(85)90015-4

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118-135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual. *Cognition, 96*(1), B23–B32. https://doi.org/10.1016/j.cognition.2004.10.003

Huettig, F., & Altmann, G. T. (2011). Looking at anything that is green when hearing “frog”: How object surface colour and stored object colour knowledge influence language-mediated overt attention. *The Quarterly Journal of Experimental Psychology, 64*(1), 122-145. https://doi.org/10.1080/17470218.2010.481474

Huettig, F., & Brouwer, S. (2015). Delayed anticipatory spoken language processing in adults with dyslexia: evidence from eye-tracking. *Dyslexia, 21*(2), 97-122. https://doi.org/10.1002/dys.1497

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant in structions reveal strong limits on prediction in language processing. *Brain Research, 1706*, 196–208. https://doi.org/10.1016/j.brainres.2018.11.013

Huettig, F., & Hartsuiker, R. J. (2008). When you name the pizza you look at the coin and the bread: Eye movements reveal semantic activation during word production. *Memory and Cognition, 36*(2), 341-360. https://doi.org/10.3758/MC.36.2.341

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes, 25*(3), 347–374. https://psycnet.apa.org/doi/10.1080/01690960903046926

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language,*

Cognition and Neuroscience, 31(1), 80-93. https://doi.org/10.1080/23273798.2015.1047459

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31*(1), 19-31. https://doi.org/10.1080/23273798.2015.1072223

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460-482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., & Pickering, M. J. (2019). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Sciences, 23*(6), 464-475. https://doi.org/10.1016/j.tics.2019.03.008

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Hvelplund, K. T. (2014). Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI*(Special Issue – Minding Translation), 201-223.

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. *Acta Psychologica, 137*(2), 252-259. https://doi.org/10.1016/j.actpsy.2010.07.017

Ito, A. (2019). Prediction of orthographic information during listening comprehension: A printed-word visual world study. *Quarterly Journal of Experimental Psychology, 72*(11), 2584-2596. https://doi.org/10.1177/1747021819851394

Ito, A., & Knoeferle, P. (2023). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods, 55*(7), 3461-3493. https://doi.org/10.3758/s13428-022-01969-3

Ito, A., & Pickering, M. J. (2021). Automaticity and prediction in non-native language comprehension. In E. Kaan, & T. Grüter, *Prediction in second language processing and learning* (pp. 26-46). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/bpa.12.02ito

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition, 21*(2), 251-264. https://doi.org/10.1017/S1366728917000050

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157-171. https://doi.org/10.1016/j.jml.2015.10.007

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience, 32*(8), 954–965. https://doi.org/10.1080/23273798.2016.1242761

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 635-652. https://doi.org/10.1037/xlm0000315

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English. *Journal of Memory and Language, 98*, 1-11. https://doi.org/10.1016/j.jml.2017.09.002

Jörg, U. (1995). Bridging the gap: Verb anticipation in German-english simultaneous interpreting. In M. Snell-Hornby, Z. Jettmarová, & K. Kai, *Translation as intercultural communication: Selected papers from the EST congress, Prague 1995* (pp. 217-228). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.20.22jor

Jaeger, F. T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Janikowski, P., & Chmiel, A. (2025). Ear-voice span in simultaneous interpreting: Text-specific factors, interpreter-specific factors and individual variation. *Interpreting, 27*(1), 28-51. https://doi.org/10.1075/intp.00116.jan

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness.* Cambridge: Cambridge University Press.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism, 4*(2), 257-282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E., & Grüter, T. (2021). Prediction in second language processing and learning: Advances and directions. In E. Kaan, & T. Grüter, *Prediction in second language processing and learning* (pp. 1-24). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/bpa.12.01kaa

Kaan, E., Kirkham, J., & Wijnen, F. (2014). Prediction and integration in native and second-language processing of elliptical structures. *Bilingualism: Language and Cognition, 19*(1), 1-18. https://doi.org/10.1017/S1366728914000844

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133-156. https://doi.org/10.1016/S0749-596X(03)00023-8

Karaca, F., Brouwer, S., Unsworth, S., & Huettig, F. (2021). Prediction in bilingual children: The missing piece of the puzzle. In E. Kaan, & T. Grüter, *Prediction in second language processing and learning* (pp. 115–137). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/bpa.12.06kar

Kempen, G., & Harbusch, K. (2019). Mutual attraction between high-frequency verbs and clause types with finite verbs in early positions: Corpus evidence from spoken English, Dutch, and German. *Language, Cognition and Neuroscience, 34*(9), 1140-1151. https://doi.org/10.1080/23273798.2019.1642498

Kiehl, K. A., Laurens, K. R., & Liddle, P. F. (2002). Reading anomalous sentences: an event-related fMRI study of semantic processing. *NeuroImage, 17*(2), 842-850. https://doi.org/10.1006/nimg.2002.1244

Kim, H.-R. (2005). Linguistic characteristics and interpretation strategy based on EVS analysis of Korean-Chinese, Korean-Japanese interpretation. *Meta, 50*(4), 1-16. https://doi.org/10.7202/019846ar

King, J. W., & Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience, 7*(6), 376-395. https://doi.org/10.1162/jocn.1995.7.3.376

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Kleinman, D., Runnqvist, E., & Ferreira , V. S. (2015). Single-word predictions of upcoming language during comprehension: Evidence from the cumulative semantic interference task. *Cognitive Psychology, 79*, 68-101. https://doi.org/10.1016/j.cogpsych.2015.04.001

Klonowicz, T. (1990). A psychophysiological assessment of simultaneous interpreting: The interaction of individual differences and mental workload. *Polish Psychological Bulletin, 21*(1), 37–48.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30*(3), 481-529. https://doi.org/10.1207/s15516709cog0000_65

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*(4), 519-543. https://doi.org/10.1016/j.jml.2007.01.003

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95*(1), 95-127. https://doi.org/10.1016/j.cognition.2004.03.002

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D.-H., & Gowda, S. M. (n.d.). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering, 57*(11), 2010. https://doi.org/10.1109/tbme.2010.2057429

Kotz, S. A., Cappa, S. F., von Cramon, D. Y., & Friederici, A. D. (2002). Modulation of the lexical-semantic network by auditory semantic priming: an event-related functional MRI study. *NeuroImage, 17*(4), 1761-1772. https://doi.org/10.1006/nimg.2002.1316

Kukona, A., Cho, P., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 326-347. https://psycnet.apa.org/doi/10.1037/a0034903

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen , H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*(1), 23–42. https://doi.org/10.1016/j.cognition.2010.12.002

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: challenges to syntax. *Brain Research, 1146*(4), 23-49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32-59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., & Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience, 15*(2), 272-293. https://doi.org/10.1162/089892903321208204

Kuperman, V., & van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception & Performance, 39*(3), 802-823. https://doi.org/10.1037/a0030859

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behaviour Research, 44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Kurz, I. (1995). Watching the brain at work: An exploratory study of EEG changes during simultaneous interpreting (SI). *The Interpreters' Newsletter, 6*, 3-16.

Kurz, I. (2003). Physiological stress during simultaneous interpreting: A comparison of experts and novices. *Interpreters Newsletter, 12*, 51-67.

Kurz, I., & Färber , B. (2003). Anticipation in German-English Simultaneous Interpreting. *Forum, 1*(2), 123-150. https://doi.org/10.1075/forum.1.2.06kur

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205. https://doi.org/10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161-163. https://doi.org/10.1038/307161a0

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar, *Predictions in the brain: Using our past to generate a future* (pp. 190-207). Oxford: Oxford University Press. https://psycnet.apa.org/doi/10.1093/acprof:oso/9780195395518.003.0065

Kwon, N., Sturt, P., & Liu, P. (2017). Predicting semantic features in Chinese: Evidence from ERPs. *Cognition, 166*, 433-446. https://doi.org/10.1016/j.cognition.2017.06.010

Lamberger-Felber, H. (2001). Text-oriented research into interpreting: Examples from a case-study. *Hermes, 14*(26), 39-64. https://doi.org/10.7146/hjlcb.v14i26.25638

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484-502. https://doi.org/10.1162/jocn_a_00328

Lederer, M. (1981). *La traduction simultanée – expérience et théorie.* Paris: Minard Lettres Modernes.

Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer, 2*(2), 165–184. https://doi.org/10.1080/1750399X.2008.10798772

Lee, T.-H. (2002). Ear voice span in English into Korean simultaneous interpretation. *Meta, 47*(4), 596–606. https://doi.org/10.7202/008039ar

Levelt, W. (1989). *Speaking: From intention to articulation.* The MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63*(4), 447-464. https://doi.org/10.1016/j.jml.2010.07.003

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A Functional Reference Grammar.* Berkeley; London: University of California Press.

Liu, M.-h. (2013). Design and analysis of Taiwan's interpretation certification examination. In D. Tsagari, & R. van Deemter, *Assessment issues in language translation and interpreting* (pp. 163-178). Peter Lang Publishing. https://doi.org/10.3726/978-3-653-02510-1

Liu, Y., Hintz, F., Liang, J., & Huettig, F. (2022). Prediction in challenging situations: Most bilinguals can predict upcoming semantically-related words in their L1 source language when interpreting. *Bilingualism: Language and Cognition*, *25*(5), 801-815. https://doi.org/10.1017/S1366728922000232

Lozano-Argüelles, C., & Sagarra, N. (2021). Interpreting experience enhances the use of lexical stress and syllabic structure to predict L2 word endings. *Applied Psycholinguistics, 42*(5), 1135-1157. https://doi.org/10.1017/S0142716421000217

Lozano-Argüelles, C., Sagarra, N., & Casillas, J. V. (2020). Slowly but surely: Interpreting facilitates L2 morphological anticipation based on suprasegmental and segmental information. *Bilingualism: Language and Cognition, 23*(4), 752-762. https://doi.org/10.1017/S1366728919000634

Lozano-Argüelles, C., Sagarra, N., & Casillas, J. V. (2023). Interpreting experience and working memory effects on L1 and L2 morphological prediction nological prediction. *Frontiers in Language Sciences, 1*, 1065014. https://doi.org/10.3389/flang.2022.1065014

Luck, S. J., & Kappenman, E. S. (2012). *The Oxford handbook of event-related potential components.* Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.001.0001

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*(1), 133-156. https://doi.org/10.1080/03640210709336987

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake — But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843-847. https://doi.org/10.1037/a0029284

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition, 27*(1), 53–78. https://doi.org/10.1017/S0272263105050035

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*(517), 522-523. https://doi.org/10.1038/244522a0

Marslen-Wilson, W. D. (1984). Function and process in spoken word-recognition: A tutorial review. In H. Bouma, & D. Bouwhuis , *Attention and Performance* (pp. 125-150). Lawrence Erlbaum Associates.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific Reports, 8*(1), 1079-1079. https://doi.org/10.1038/s41598-018-19499-4

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*(4), 574-588. https://doi.org/10.1016/j.jml.2013.08.001

Masson, M. E. (1991). A distributed memory model of context effects in word identification. In D. Besner, & G. W. Humphreys, *Basic processes in reading: Visual word recognition* (pp. 233-263). Hillsdale, NJ: Lawrence Erlbaum Associates.

Matsumoto, A., Iidaka, T., Haneda, K., Okada, T., & Sadato, N. (2005). Linking semantic priming effect in functional MRI and event- related potentials. *NeuroImage, 24*(3), 624-634. https://doi.org/10.1016/j.neuroimage.2004.09.008

Mazza, V., Turatto, M., & Caramazza, A. (2009). Attention selection, distractor suppression and N2pc. *Cortex*, *45*(7), 879-890. https://doi.org/10.1016/j.cortex.2008.10.009

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language, 55*(3), 381-401. https://doi.org/10.1016/j.jml.2006.06.006

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30

McKoon, G., & Ratcliff, R. (1989). Semantic associations and elaborative inference. *Journal of Experimental Psychology, 15*(2), 326-338. https://doi.org/10.1037//0278-7393.15.2.326

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology, 60*(5), 661-671. https://doi.org/10.1080/17470210601183890

McRae, K., Hare, M., Elman, J., & Ferret, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition, 33*(7), 1174-1184. https://doi.org/10.3758/BF03193221

Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition, 25*(1), 47-56. https://doi.org/10.3758/bf03197284

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language, 66*(4), 545-567. https://doi.org/10.1016/j.jml.2012.01.001

Meyer, A. S., Roelofs, A., & Levelt, W. J. (2003). Word length effects inobject naming: The role of a response criterion. *Journal of Memory and Language, 48*(1), 131-147. https://doi.org/10.1016/S0749-596X(02)00509-0

Meyer, A., van der Meulen, F., & Brooks, A. (2004). Eye movements during speech planning: Talking about present and remembered objects. *Visual Cognition, 11*(5), 553-576. https://doi.org/10.1080/13506280344000248

Meyer, D. D., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227-234. https://psycnet.apa.org/doi/10.1037/h0031564

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*(4), 475-494. https://doi.org/10.1016/j.jml.2007.11.006

Mitsugi, S. (2017). Incremental comprehension of Japanese passives: Evidence from the visual-world paradigm. *Applied Psycholinguistics, 38*(4), 953-983. https://psycnet.apa.org/doi/10.1017/S0142716416000515

Mitsugi, S., & MacWhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition, 19*(1), 19-35. https://doi.org/10.1017/S1366728914000881

Moser-Mercer, B. (1994). Aptitude testing for conference interpreting: Why, when and how. In S. Lambert, & B. Moser-Mercer, *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 57-68). Amsterdam/Philodephia: John Benjamins. https://doi.org/10.1075/btl.3.07mos

Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (Pilot study). *Interpreting*, *3*(1), 47-64. https://doi.org/10.1075/intp.3.1.03mos

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, 36*(3), 402-407. https://doi.org/10.3758/BF03195588

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . Rousselet, G. A. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*, e33468. https://doi.org/10.7554/eLife.33468

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1-13. https://doi.org/10.1177/1609406917733847

O'Brien, E. J., & Albrecht, J. E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 777-784. https://doi.org/10.1037//0278-7393.18.4.777

Ono, T., Tohyama, H., & Matsubara, S. (2008). Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias, *roceedings of the sixth international conference on language resources and evaluation (LREC'08)* (pp. 3383-3387). Marrakech: European Language Resources Association (ELRA).

Otten, M., & van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes, 45*(6), 464-496. https://psycnet.apa.org/doi/10.1080/01638530802356463

Otten, M., & van Berkum, J. J. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research, 1291*, 92-101. https://doi.org/10.1016/j.brainres.2009.07.042

Otten, M., Nieuwland, M. S., & van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience, 8*, 89. https://doi.org/10.1186/1471-2202-8-89

Özkan, D., Hodzik, E., & Diriker, E. (2023). Simultaneous interpreting experience enhances the use of case markers for prediction in Turkish. *Interpreting, 25*(2), 186-210. https://doi.org/10.1075/intp.00085.ozk

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language, 67*(4), 426-448. https://doi.org/10.1016/j.jml.2012.07.003

Perez, I. A., Hartley, A., Mason, I., & Peng, G. (2003). *Peer- and self-Assessment in conference interpreting training.* Edinburgh: Centre for Languages, Linguistics and Area Studies.

Peters, R. E., Grüter, T., & Borovsky, A. (2015). Anticipatory and locally coherent lexical activation varies as a vunction of language proficiency. *Proceedings of the Annual Meeting of the Cognitive Science Society, 37*, 1865-1870.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002-1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105-110. https://doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language produc- tion and comprehension. *Behavioral and Brain Sciences, 36*(4), 329-392. https://doi.org/10.1017/s0140525x12001495

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memoiy, and Cognition, 24*(4), 940-961. https://psycnet.apa.org/doi/10.1037/0278-7393.24.4.940

Podhajská, K. (2008). Time lag in simultaneous interpretation from English into Czech and its dependence on text type. In I. Čeňková, *Prague translation studies: The next generation* (pp. 87-110). Praha: Univerzita Karlova .

Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research, 1*(2), 73-86. https://doi.org/10.1016/0926-6410(93)90013-u

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*(3), 385-408. https://psycnet.apa.org/doi/10.1037/0033-295X.95.3.385

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514-528. https://doi.org/10.1037/a0020990

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition, 122*, 267-279. https://doi.org/10.1016/j.cognition.2011.11.011

Rommers, J., Meyer, A. S., & Huettig, F. (2015). Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Attention, Perception, & Psychophysics, 77*(3), 720-730. https://doi.org/10.3758/s13414-015-0873-x

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia, 51*(3), 437-447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002

Rossell, S. L., Price, C. J., & Nobre, A. C. (2003). The anatomy and time course of semantic priming investigated by fMRI and ERPs. *Neuropsychologia, 41*(5), 550-564. https://doi.org/10.1016/s0028-3932(02)00181-1

RStudio Team. (2024). *RStudio: Integrated development environment for R (Version 2024.04.2+764) [Computer software].* Retrieved from RStudio, PBC: https://posit.co/downloads/

Ruchkin, D. S., Johnson, R. J., Canoune, H., & Ritter, W. (1990). Short-term memory storage and retention: an event-related brain potential study. *Electroencephalography and Clinical Neurophysiology, 76*(5), 419-439. https://doi.org/10.1016/0013-4694(90)90096-3

Ruiz Rosendo, L., & Galván, M. C. (2019). Coping with speed: An experimental study on expert and novice interpreter performance in the simultaneous interpreting of scientific discourse. *Babel, 65*(1), 1–25. https://doi.org/10.1075/babel.00081.rui

Salverda, A. P., & Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, 36(5), 1108-1117. https://doi.org/10.1037/a0019901

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language, 71*(1), 145-163. https://doi.org/10.1016/j.jml.2013.11.002

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6), e13335. https://doi.org/10.1111/psyp.13335

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language, 24*(2), 232-252. https://doi.org/10.1016/0749-596X(85)90026-9

Seeber, K. G. (2001). Intonation and anticipation in simultaneous interpreting. *Cahiers de Linguistique Française, 23*(4), 61–97.

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories — new models. *Interpreting, 13*(2), 176 - 204. https://psycnet.apa.org/doi/10.1075/intp.13.2.02see

Seeber, K. G. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target, 25*(1), 18-32. https://doi.org/10.1075/target.25.1.03see?locatt=mode:legacy

Seleskovitch, D. (1984). Les anticipations de la compréhension. In D. Seleskovitch, & M. Lederer, *Interpréter pour traduire* (pp. 273–283). France: Didier Erudition.

Setton, R. (1999). *Simultaneous interpretation: A cognitive-pragmatic analysis.* Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.28

Setton, R. (2005). So what is so interesting about simultaneous interpreting? *SKASE Journal of Translation and Interpretation, 1*(1), 70-84.

Setton, R., & Dawrant, A. (2016). *Conference interpreting: A trainer's guide.* Amsterdam/Philadephia: John Benjamins.

Shelton, J. R., & Martin, R. C. (1992). How semantic Is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1191-1210. https://doi.org/10.1037//0278-7393.18.6.1191

Sherman, M. (1998). Efficiency and robustness in subsampling for dependent data. *Journal of Statistical Planning and Inference, 75*(1), 133-146. https://doi.org/10.1016/S0378-3758(98)00123-2

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal analysis: Modeling change and event occurrence.* New York: Oxford University Press. https://psycnet.apa.org/doi/10.1093/acprof:oso/9780195152968.001.0001

Smith, J. A., & Osborn, M. (2003). Interpretative phenomenological analysis. In J. A. Smith, *Qualitative psychology: A practical guide to research methods* (pp. 51-80). Sage Publications, Inc.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319. https://doi.org/10.1016/j.cognition.2013.02.013

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238−299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281-284. https://doi.org/10.1111/1467-9280.00151

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review, 18*(2), 371–376. https://doi.org/10.3758/s13423-010-0046-9

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Stone, K., Lago, S., & Schad, D. J. (2021). Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism: Language and Cognition, 24*(5), 833–841. https://doi.org/10.1017/S1366728920000607

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*(6), 645-659. https://doi.org/10.1016/S0022-5371(79)90355-4

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634. https://doi.org/10.1126/science.7777863

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International*

Journal of Psychophysiology, 83(3), 382-392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Timarová, Š., Čeňková, I., Meylaerts, R., Hertog, E., Szmalec, A., & Duyck, W. (2014). Simultaneous interpreting and working memory executive control. *Interpreting, 16*(2), 139-168. https://doi.org/10.1075/intp.16.2.01tim

Timarová, Š., Čeňková, I., Meylaerts, R., Hertog, E., Szmalec, A., & Duyck, W. (2015). Simultaneous interpreting and working memory capacity. In A. Ferreira, & J. W. Schwieter, *Psycholinguistic and cognitive inquiries into translation and interpreting* (pp. 101–126). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.115.05tim

Timarová, Š., Dragsted, B., & Hansen, I. G. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiselius, *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 121-146). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.94.10tim

Tiselius, E. (2009). Revisiting Carroll's scales. In C. v. Angelelli, & H. E. Jacobson, *Testing and assessment in translation and interpreting studies* (pp. 95-121). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Treisman, A. M. (1965). The effects of redundancy and familiarity on translating and repeating back a foreign and a native language. *British Journal of Psychology, 56*(4), 369–379. https://psycnet.apa.org/doi/10.1111/j.2044-8295.1965.tb00979.x

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*(3), 285-318. https://doi.org/10.1006/jmla.1994.1014

van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443-467. https://doi.org/10.1037/0278-7393.31.3.443

van de Velde, M., & Meyer, A. S. (2014). Syntactic flexibility and planning scope: The effect of verb bias on advance planning during sentence recall. *Frontiers in Psychology, 5*, 1174. https://doi.org/10.3389/fpsyg.2014.01174

van Overschelde, J. P., Rawson, K., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289-335. https://psycnet.apa.org/doi/10.1016/j.jml.2003.10.003

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews, 35*(3), 407-426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Voeten, C. C. (2022). permutes: Permutation Tests for Time Series Data. Retrieved from: https://rdrr.io/cran/permutes/

Waddington, C. (2001). Should translations be assessed holistically or through error analysis? *Hermes, 14*(26), 15-37. https://doi.org/10.7146/hjlcb.v14i26.25637

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50*(1), 1-25. https://psycnet.apa.org/doi/10.1016/S0749-596X(03)00105-0

Wen, Y., & van Heuven, W. J. (2017). Chinese translation norms for 1,429 English words. *Behavior Research, 49*(3), 1006-1019. https://doi.org/10.3758/s13428-016-0761-x

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience, 16*(7), 1272-1288. https://doi.org/10.1162/0898929041920487

Wilss, W. (1978). Syntactic anticipation in German-English simultaneous interpreting. In D. Gerver, & H. W. Sinaiko, *Language interpretation and communication* (pp. 343-352). New York: Plenum Press. https://doi.org/10.1007/978-1-4615-9077-4_30

Wlotko, E. W., & Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology, 49*(6), 770-785. https://doi.org/10.1111/j.1469-8986.2012.01366.x

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1-14. https://psycnet.apa.org/doi/10.1037/0278-7393.32.1.1

Zhang, X. (2016). Semi-automatic simultaneous interpreting quality evaluation. *International Journal of Natural Language Computing, 5*(5), 1-12. https://doi.org/10.5121/ijnlc.2016.5501

Zirnstein, M., van Hell, J. G., & Kroll, J. F. (2018). Cognitive control ability mediates prediction costs in monolinguals and bilinguals. *Cognition, 176*, 87-106. https://doi.org/10.1016/j.cognition.2018.03.001

Appendices

Appendix 1 Source text

Paragraph 1	**Target**	**Implausible Competitor**	**Plausible Competitor**	**Distractor**
1) As the sun begins to set, I am deeply attracted by the town's beautiful buildings. Trondheim has an old Gothic church. *Its corridor is illuminated/decorated by hundreds of* ***candles.***	Candle	Lightning	Painting	Guitar
2) But perhaps the most distinctive buildings are a series of old wooden houses along the river. These buildings are usually converted to cosy restaurants and pubs. They are painted either brown, red, green, or yellow.				
3) I stay at a comfortable hotel by the river. *To prepare for my journey to the north, I iron/choose a couple of* ***shirts****.*	Shirt	Curtain	Book	Balloon
4) When we approach the train station early on the following morning, the gentle rain still falls. *As we approach the station, which is close to the river, I hear/see a cruise ship's* ***horn.***	Horn	Music	Flag	Sofa
5) Trondheim railway station is known for being an elegant classical building. Its warm and bright booking hall is a relief for passengers. They come through the main doors, completely soaked in the rain, and shake water off their umbrellas.				
6) The place seems a home away from home for some of them, and they call cheerful greetings to one another. *In the station store, commuters are eating/buying freshly made* ***bread.***	Bread	Turkey	Juice	Bone
7) When I arrive at the platform, the train is waiting. Six red coaches are all very well designed and arranged. I take my seat, which is most comfortable with a tall back and plenty of legroom. In front of my seat is a generously sized table.				
8) I am completely happy. There is no need to try and exchange it for another. I plan to fully enjoy myself in the journey. *As we pull away from the station, I see a lady with a crying baby pushing/parking a lovely pink* ***stroller****.*	Stroller	Door	Van	Milk

9) Soon we are out of town, moving through gently waving fields. Herds of cattle are eating fresh and wild grass. In the grey sky above, a white-tailed eagle is turning gracefully in the rain.				
Paragraph 2	**Target**	**Implausible Competitor**	**Plausible Competitor**	**Distractor**
1) The rain gradually stops as the sun comes out from the heavy clouds. Although there are no people in view, I see drifts of smoke from chimneys and warm lights shining in windows.				
2) *Finally, I see a boy flying/chasing a rainbow-coloured **kite**.* Not far away behind him, a couple are walking a spotted dog.	Kite	Airplane	Ball	Computer
3) Over on the other side, the train is passing by what looks like the open sea, on which a giant black ship is moving slowly across the horizon. Sometimes the train seems to move downward almost on to the stone beach. Several boats are waiting for smoother waters.				
4) *At some point along the beach, **I** see a man riding/driving a bright purple **motorcycle**.* When the train passes by, he waves to us. From time to time, I am amazed by the train's approach to the rocky beaches and rolling silvery waves.	Motorcycle	Horse	Truck	Rabbit
5) In the carriage, there are newspapers including English ones and some beverage making facilities. *Feeling a bit chill, I brew/drink a cup of hot **tea.*** As I drink and read, the train makes a few more stops.	Tea	Beer	Milk	Cheese
6) *At one of the stops, a blonde-haired lady gets on board dragging/carrying a large black **luggage**.* She sits on the seat opposite to mine.	Luggage	Chair	Backpack	Washing machine
7) It seems nodding to fellow passengers is a custom on Norwegian trains, so she smiles and nods to me. After settling down, she puts on a pair of headphones and closes her eyes to take a rest.				
8) It is almost lunch time, so I decide to visit the dining car. As I walk along the train, I pass through one carriage that has a sort of playpen for children. There is a climbing frame and scattered toys. *A child is dressing/playing with a well-made **doll**.*	Doll	Window	Teddy bear	Bed
9) Interesting enough, two other children seem to be playing a make-believe game. One plays as a dentist asking the other to open her mouth. Another three are sitting in front of a large TV that is silently showing animation.				

Paragraph 3	**Target**	**Implausible Competitor**	**Plausible Competitor**	**Distractor**
1) The restaurant car is extremely clean and modern. *A waitress is wiping/cleaning the dining* ***tables***. The only other customers are a young couple. They both nod at me.	Table	Eye	Fork	Monkey
2) Seeing that I am looking with interest at the sushi they are eating, the woman explains, 'Seafood is really fresh here. The sushi is made of quality fish. I promise you will like it.' Although I am not really a seafood fan, I order a few and find them quite tasty.				
3) *Still feeling unfulfilled, I toast two slices of bread for myself and spread/add some delicious* ***jam***. As I have my lunch, the rolling green meadows have now become more exciting rocky landscape.	Jam	Virus	Egg	Mask
4) When I walk back to my seat, the blonde lady stops listening to music and removes her wireless headphones. She introduces herself as a painter. She will be going to Bodø to draw inspirations for painting.				
5) Since we are going to spend a few more hours together, she asks if she can paint for me. *As I happily agree, she starts to sharpen/prepare some drawing* ***pencils.***	Pencil	Knife	Brush	Battery
6) While she paints, I watch the TV in the front of our carriage, which is playing a magic show. *The magician bends/grabs a tiny silver* ***spoon.*** As he snaps his fingers, it just disappears completely.	Spoon	Knee	Ring	Banana
7) The train progresses more slowly now as the rocky forests have become steep snow-capped mountains. Gradually as the train approaches the station, trees begin to reappear. Unlike small station buildings that are made of wood, the Bodø station is built of bricks.				
8) Just as I leave the station, I notice a few people queueing in front of a stall selling traditional local food. *The chef heats/fetches a specially designed* ***pan***. As it is dinner time, attracted by its wonderful smell, I join the queue and buy a few pieces to try.	Pan	Water	Bowl	Camera
Paragraph 4	**Target**	**Implausible Competitor**	**Plausible Competitor**	**Distractor**
1) Bodo(ø) is a tranquil town surrounded by lakes and mountains. The hotel I am staying at is a two-floor building with a well-kept garden. As I enter the warm reception hall, a porter enters to throw charcoal into a large burning fireplace.				

2) The porter then accompanies me to my room. Before leaving, he reminds me that in case of an emergency, I should not use the lift and instead, I should use safety stairs. I say thanks and goodnight to him and enter my room.				
3) My room is warm and comfortable with a gentle yellow light. *I **fold/remove** my wool **scarf** and put it on a luggage rack.* I decide I would like some wine before going to bed.	Scarf	Paper	Hat	Pen
4) As I arrive at the cafe, there is only a lady serving at the counter. The wine I order is made from a kind of grapes that only grows in the southern France. *As I sip my wine, the woman **squeezes/picks** a couple of **lemons** and makes a cocktail for herself.*	Lemon	Time	Apple	Telephone
5) Having enjoyed the wine, I have a wonderful sleep and wake up rather early in the morning. *As I walk through the Queen's Garden, a man is **trimming/painting** a beautiful **flower**.*	Flower	Hair	Cat	Watering bottle
6) Unlike the peace and quietness in the night, the town becomes bustling in the day. I go to the market at the town centre. There are diverse shops selling various food, flowers, and groceries.				
7) *When wandering around the market, I find a butcher shop that **bakes/sells** sweet and spicy **chicken**.* Attracted by the smell, I buy myself a piece as lunch.	Chicken	Cake	Steak	Desk
8) *On the central square, a young man, companied by thrilling music, is **beating/playing** a unique kind of **drum**.* I listen for a while and leave some coins as my appreciation to the performance.	Drum	Heart	Keyboard	Mushroom
9) The day is finished by a visit to the post office, where I send myself a letter. 24 hours after my arrival, I am back at Bodo(ø) station, preparing to board the Nordland night train.				

#Note: Experimental sentences are presented in italic type. Predictive/unpredictive critical verbs are highlighted in red and bold, while target words are highlighted in black and bold.

Appendix 2 Robustness check for by window analyses of the eye-tracking data

Professional					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
Prediction window					
(Intercept)	−3.063	< .001 ***	−3.063	< .001 ***	→
Plausible competitor	−0.697	0.282	−0.720	0.264	→
Implausible competitor	−1.467	0.067 †	−1.390	0.084 †	→
Distractor	−1.396	0.064 †	−1.386	0.062 †	→
Unpredictable condition	−0.455	0.146	−0.394	0.222	→
Plau × Unpre	1.029	0.016 *	0.993	0.025 *	→
Implau × Unpre	0.630	0.139	0.542	0.223	→
Distra × Unpre	0.706	0.097 †	0.444	0.144	↓
Post-target window					
(Intercept)	−2.028	< .001 ***	−2.119	< .001 ***	→
Plausible competitor	−2.621	< .001 ***	−2.440	< .001 ***	→
Implausible competitor	−2.754	0.001 **	−2.616	0.002 **	→
Distractor	−2.931	< .001 ***	−2.807	< .001 ***	→
Unpredictable condition	−0.183	0.560	−0.053	0.874	→
Plau × Unpre	0.619	0.148	0.416	0.354	→
Implau × Unpre	−0.229	0.592	−0.387	0.389	→
Distra × Unpre	0.594	0.165	0.343	0.444	→
Student					

Fixed effect	Original Estimate	Original *p*	Subsampled Estimate	Subsampled *p*	Variation
Prediction window					
(Intercept)	−4.042	< .001 ***	−4.017	< .001 ***	→
Plausible competitor	−0.176	0.760	−0.200	0.744	→
Implausible competitor	−1.057	0.121	−1.028	0.144	→
Distractor	−0.897	0.159	−0.993	0.128	→
Unpredictable condition	−0.043	0.860	−0.110	0.666	→
Plau × Unpre	0.068	0.841	0.191	0.586	→
Implau × Unpre	0.136	0.688	0.103	0.768	→
Distra × Unpre	0.253	0.454	0.399	0.254	→
Post-target window					
(Intercept)	−2.862	< .001 ***	−2.878	< .001 ***	→
Plausible competitor	−1.724	0.003 **	−1.689	0.003 **	→
Implausible competitor	−2.749	< .001 ***	−2.744	< .001 ***	→
Distractor	−2.605	< .001 ***	−2.632	< .001 ***	→
Unpredictable condition	0.185	0.426	0.135	0.575	→
Plau × Unpre	−0.323	0.322	−0.259	0.446	→
Implau × Unpre	0.045	0.891	0.121	0.722	→
Distra × Unpre	−0.432	0.185	−0.285	0.401	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or *p*-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 3 Robustness check for by-group GCA of the eye-tracking data

Professional					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
Pre × Plau	−0.677	< .001 ***	−0.565	< .001 ***	→
Pre × Implau	−0.242	< .001 ***	−0.196	.001 **	↓
Pre × Distra	−0.441	< .001 ***	−0.405	< .001 ***	→
Pre × ot1	1.124	< .001 ***	0.907	.005 **	↓
Pre × ot2	−2.338	< .001 ***	−2.493	< .001 ***	→
Pre × ot3	−2.761	.370	−3.656	.257	→
Pre × Plau × ot1	−1.741	< .001 ***	−1.656	< .001 ***	→
Pre × Implau × ot1	−0.394	.367	1.450	.318	→
Pre × Distra × ot1	−1.943	< .001 ***	−1.570	< .001 ***	→
Pre × Plau × ot2	2.863	< .001 ***	3.349	< .001 ***	→
Pre × Implau × ot2	3.727	< .001 ***	4.003	< .001 ***	→
Pre × Distra × ot2	2.942	< .001 ***	3.023	< .001 ***	→
Pre × Plau × ot3	0.505	.247	0.746	.102	→
Pre × Implau × ot3	−1.076	.014 *	−1.141	.012 *	→
Pre × Distra × ot3	1.185	.007 **	1.634	< .001 ***	↑
Student					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Variation**
Pre × Plau	−0.310	< .001 ***	−0.344	< .001 ***	→
Pre × Implau	−0.218	< .001 ***	−0.167	< .001 ***	→

Pre × Distra	−0.183	< .001 ***	−0.227	< .001 ***	→
Pre × ot1	−0.944	< .001 ***	−0.872	< .001 ***	→
Pre × ot2	−1.021	< .001 ***	−0.917	< .001 ***	→
Pre × ot3	0.718	.001 **	0.523	.026 *	↓
Pre × Plau × ot1	0.726	.023 *	0.371	.265	↓
Pre × Implau × ot1	1.392	< .001 ***	1.028	.003 **	↓
Pre × Distra × ot1	1.627	< .001 ***	1.935	< .001 ***	→
Pre × Plau × ot2	2.171	< .001 ***	2.196	< .001 ***	→
Pre × Implau × ot2	1.084	< .001 ***	0.670	.044 *	↓
Pre × Distra × ot2	0.892	.005 **	0.640	.054 †	↓
Pre × Plau × ot3	−1.094	< .001 ***	−0.589	.077 †	↓
Pre × Implau × ot3	−0.332	.297	−0.116	.727	↓
Pre × Distra × ot3	−0.640	.044 *	−0.441	.185	↓

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 4 Robustness check for between-group GCA of the eye-tracking data

Fixed effect	Original Estimate	Original *p*	Subsampled Estimate	Subsampled *p*	Change in significance
Pre × Pro × ot1	1.020	< .001 ***	0.879	< .001 ***	→
Pre × Pro × ot2	−0.664	< .001 ***	−0.796	< .001 ***	→
Pre × Pro × ot3	−0.508	.007 **	−0.454	.021 *	↓
Pre × Pro × Plau × ot1	−1.233	< .001 ***	−1.013	< .001 ***	→
Pre × Pro × Implau × ot1	−0.893	< .001 ***	−0.442	.114	↓
Pre × Pro × Distra × ot1	−1.785	< .001 ***	−1.753	< .001 ***	→
Pre × Pro × Plau × ot2	0.346	.196	0.577	.039 *	↑
Pre × Pro × Implau × ot2	1.322	< .001 ***	1.667	< .001 ***	→
Pre × Pro × Distra × ot2	1.025	< .001 ***	1.191	< .001 ***	→
Pre × Pro × Plau × ot3	0.799	.003	0.667	.016 *	↑
Pre × Pro × Implau × ot3	−0.373	.164	−0.512	.067 †	↑
Pre × Pro × Distra × ot3	0.913	< .001 ***	1.038	< .001 ***	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or *p*-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 5 Robustness check for the by-group *t*-tests of the EVS data

Professional					
	Original *t*	**Original *p***	**Subsampled *t***	**Subsampled *p***	**Change in significance**
CV-EVS	2.183	.030 *	2.059	.040 *	→
TW-EVS	1.640	.102	1.595	.112	→
Sentence onset EVS	−0.991	.322	−1.222	.223	→
Sentence offset EVS	1.893	.059 †	1.697	.091 †	→
CT-span	−1.024	.307	−1.210	.227	→
Sentence-span	2.571	.011 *	2.742	.007 **	↑
Student					
	Original Estimate	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Variation**
CV-EVS	2.944	.003 **	2.715	.007 **	→
TW-EVS	1.001	.317	1.107	.269	→
Sentence onset EVS	−0.470	.638	−0.484	.628	→
Sentence offset EVS	1.621	.106	1.551	.121	→
CT-span	−2.410	.016 *	−2.163	.031 *	→
Sentence-span	2.325	.021 *	2.232	.026 *	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or *p*-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 6 Robustness check for the by-group LME models of the EVS data

Professional					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
CV-EVS					
CV frequency	−0.274	.038 *	−0.284	.052 †	↓
TW frequency	0.219	.238	0.231	.231	→
Cloze probability	1.467	< .001 ***	1.068	.013 *	↓
Verb-noun association	−0.582	.303	0.016	.480	→
Sentence length	0.006	.781	0.006	.781	→
Preceding sentence length	0.002	.849	0.001	.339	→
Within-paragraph position	0.005	< .001 ***	0.005	< .001 ***	→
TW-EVS					
CV frequency	−0.040	.775	−0.051	.743	→
TW frequency	0.221	.231	0.252	.198	→
Cloze probability	0.412	.348	0.062	.897	→
Verb-noun association	−2.014	< .001 ***	−1.563	.013 *	↓
Sentence length	0.029	.162	0.043	.060 †	→
Preceding sentence length	0.027	.012 *	0.032	.005 **	↑
Within-paragraph position	0.004	< .001 ***	0.004	< .001 ***	→
Sentence onset EVS					
CV frequency	−0.202	.034*	−0.152	.149	↓
TW frequency	0.344	.008**	0.380	.007**	→

Cloze probability	0.351	.204	0.032	.918	→
Verb-noun association	−0.974	.019 *	−0.602	.180	↓
Sentence length	−0.052	< .001 ***	−0.043	.006 **	↓
Preceding sentence length	0.039	< .001 ***	0.043	< .001 ***	→
Within-paragraph position	0.002	.001 **	0.002	.002 **	→
Sentence offset EVS					
CV frequency	0.008	.945	−0.011	.933	→
TW frequency	0.006	.972	0.084	.637	→
Cloze probability	0.842	.016 *	0.568	.153 *	→
Verb-noun association	−1.935	< .001 ***	−1.491	.008 **	↓
Sentence length	0.001	.971	0.007	.723	→
Preceding sentence length	0.004	.718	0.009	.368	→
Within-paragraph position	0.005	< .001 ***	0.005	< .001 ***	→
CT-span					
CV frequency	0.213	.022 *	0.253	.011 *	→
TW frequency	−0.329	.007 **	−0.283	.024 **	→
Cloze probability	−0.467	.107	−0.704	.022 *	↑
Verb-noun association	−1.267	< .001 ***	−0.981	.013 *	↓
Sentence length	0.011	.435	0.019	.199	→
Preceding sentence length	−0.011	.133	−0.011	.133	→
Within-paragraph position	0.002	.004 **	0.002	.004 **	→
Sentence-span					

CV frequency	1.287	< .001 ***	1.163	< .001 ***	→
TW frequency	−1.882	< .001 ***	−1.903	< .001 ***	→
Cloze probability	1.119	.036 *	1.328	.019 *	→
Verb-noun association	−1.782	.008 **	−1.627	.026 *	↓
Sentence length	0.531	< .001 ***	0.533	< .001 ***	→
Preceding sentence length	−0.175	< .001 ***	−0.178	< .001 ***	→
Within-paragraph position	0.011	< .001 ***	0.011	< .001 ***	→
Student					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
CV-EVS					
CV frequency	−0.303	.002 **	−0.271	.011 *	↓
TW frequency	0.222	.111	0.287	.055 †	↑
Cloze probability	1.528	< .001 ***	1.433	< .001 ***	→
Verb-noun association	−2.037	< .001 ***	−2.105	< .001 ***	→
Sentence length	−0.016	.308	−0.016	.328	→
Preceding sentence length	0.003	.668	0.010	.236	→
Within-paragraph position	0.005	< .001 ***	0.005	< .001 ***	→
TW-EVS					
CV frequency	0.131	.170	0.105	.316	→
TW frequency	0.329	.011 *	0.416	.003 **	↑
Cloze probability	0.219	.478	0.210	.515	→
Verb-noun association	−2.016	< .001 ***	−2.028	< .001 ***	→

Sentence length	0.049	.001 **	0.046	.003 **	→
Preceding sentence length	0.025	< .001 ***	0.029	< .001 ***	→
Within-paragraph position	0.005	< .001 ***	0.005	< .001 ***	→
Sentence onset EVS					
CV frequency	−0.375	< .001 ***	−0.325	< .001 ***	→
TW frequency	0.443	< .001 ***	0.425	< .001 ***	→
Cloze probability	0.123	.549	0.149	.489	→
Verb-noun association	−0.746	.009 **	−0.987	.001 **	→
Sentence length	−0.043	< .001 ***	−0.039	< .001 ***	→
Preceding sentence length	0.031	< .001 ***	0.032	< .001 ***	→
Within-paragraph position	0.002	< .001 ***	0.002	< .001 ***	→
Sentence offset EVS					
CV frequency	0.027	.763	0.024	.808	→
TW frequency	0.136	.287	0.234	.089 †	↑
Cloze probability	0.678	.011 *	0.582	.040 *	→
Verb-noun association	−2.456	< .001 ***	−2.517	< .001 ***	→
Sentence length	0.006	.653	0.002	.882	→
Preceding sentence length	0.013	.076 †	0.018	.017 *	↑
Within-paragraph position	0.006	< .001 ***	0.006	< .001 ***	→
CT-span					
CV frequency	0.120	.052 †	0.112	.101	↓
TW frequency	−0.261	.001 **	−0.242	.008 **	→

Cloze probability	−0.379	.058 †	−0.418	.048 *	↑
Verb-noun association	−0.731	.002 **	−0.685	.010 **	→
Sentence length	0.015	.141	0.014	.204	→
Preceding sentence length	−0.007	.161	−0.007	.166	→
Within-paragraph position	0.002	< .001 ***	0.002	< .001 ***	→
Sentence-span					
CV frequency	1.455	< .001 ***	1.479	< .001 ***	→
TW frequency	−1.312	< .001 ***	−1.312	< .001 ***	→
Cloze probability	1.400	< .001 ***	1.313	.002 **	↓
Verb-noun association	−1.900	< .001 ***	−1.990	< .001 ***	→
Sentence length	0.508	< .001 ***	0.500	< .001 ***	→
Preceding sentence length	−0.161	< .001 ***	−0.158	< .001 ***	→
Within-paragraph position	0.012	< .001 ***	0.012	< .001 ***	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 7 Robustness check for the by-group ANOVAs and Tukey's HSD tests for the EVS data

Professional					
ANOVA	**Original *F***	**Original *p***	**Subsampled F**	**Subsampled *p***	**Change in significance**
CV-EVS	24.84	< .001 ***	22.61	< .001 ***	→
TW-EVS	25.46	< .001 ***	24.47	< .001 ***	→
Sentence onset EVS	7.93	< .001 ***	7.02	< .001 ***	→
Sentence offset EVS	12.94	< .001 ***	13.27	< .001 ***	→
CT-span	1.03	.381	0.77	.514	→
Sentence-span	9.24	< .001 ***	9.42	< .001 ***	→
Tukey's HSD	**Original *t***	**Original *p***	**Subsampled *t***	**Subsampled *p***	**Change in significance**
CV-EVS					
Paragraph 2−1	0.401	.248	0.339	.457	→
Paragraph 3−1	0.374	.368	0.349	.481	→
Paragraph 4−1	1.802	< .001 ***	1.800	< .001 ***	→
Paragraph 3−2	−0.026	.999	0.010	.999	→
Paragraph 4−2	1.402	< .001 ***	1.461	< .001 ***	→
Paragraph 4−3	1.428	< .001 ***	1.451	< .001 ***	→
TW-EVS					
Paragraph 2−1	0.079	.985	0.068	.992	→
Paragraph 3−1	0.518	.120	0.562	.109	→
Paragraph 4−1	1.678	< .001 ***	1.745	< .001 ***	→
Paragraph 3−2	0.439	.205	0.494	.160	→

Paragraph 4−2	1.599	< .001 ***	1.677	< .001 ***	→
Paragraph 4−3	1.160	< .001 ***	1.183	< .001 ***	→
Sentence onset EVS					
Paragraph 2−1	−0.136	.852	−0.185	.732	→
Paragraph 3−1	0.309	.298	0.299	.381	→
Paragraph 4−1	0.613	.002 **	0.563	.010 **	→
Paragraph 3−2	0.445	.054 †	0.484	.042 *	↑
Paragraph 4−2	0.748	< .001 ***	0.748	< .001 ***	→
Paragraph 4−3	0.304	.303	0.264	.470	→
Sentence offset EVS					
Paragraph 2−1	0.167	.865	0.151	.915	→
Paragraph 3−1	0.269	.645	0.330	.532	→
Paragraph 4−1	1.204	< .001 ***	1.288	< .001 ***	→
Paragraph 3−2	0.102	.969	0.179	.874	→
Paragraph 4−2	1.037	< .001 ***	1.137	< .001 ***	→
Paragraph 4−3	0.935	< .001 ***	0.958	< .001 ***	→
CT-span					
Paragraph 2−1	0.011	.999	0.059	.984	→
Paragraph 3−1	−0.119	.886	−0.025	.999	→
Paragraph 4−1	−0.224	.517	−0.174	.740	→
Paragraph 3−2	−0.130	.824	−0.084	.950	→
Paragraph 4−2	−0.235	.397	−0.234	.450	→

Paragraph 4−3	−0.105	.910	−0.149	.798	→
Sentence-span					
Paragraph 2−1	0.690	.476	0.973	.228	→
Paragraph 3−1	0.057	.999	0.249	.964	→
Paragraph 4−1	2.105	< .001 ***	2.341	< .001 ***	→
Paragraph 3−2	−0.633	.410	−0.724	.323	→
Paragraph 4−2	1.415	004 **	1.368	011 *	↓
Paragraph 4−3	2.048.	< .001 ***	2.048.	< .001 ***	→
Student					
ANOVA	**Original *F***	**Original *p***	**Subsampled F**	**Subsampled *p***	**Change in significance**
CV-EVS	61.93	< .001***	56.69	< .001***	→
TW-EVS	84.28	< .001***	75.54	< .001***	→
Sentence onset EVS	24.08	< .001***	25.32	< .001***	→
Sentence offset EVS	50.60	< .001***	44.60	< .001***	→
CT-span	3.21	.023 *	3.20	.023 *	→
Sentence-span	16.92	< .001***	16.10	< .001***	→
Tukey's HSD	**Original *t***	**Original *p***	**Subsampled *t***	**Subsampled *p***	**Change in significance**
CV-EVS					
Paragraph 2−1	0.221	.472	0.243	.429	→
Paragraph 3−1	0.704	< .001 ***	0.724	< .001 ***	→
Paragraph 4−1	1.986	< .001 ***	2.013	< .001 ***	→
Paragraph 3−2	0.483	.007 **	0.481	.011 *	↓

Paragraph 4−2	1.766	< .001 ***	1.770	< .001 ***	→
Paragraph 4−3	1.282	< .001 ***	1.289	< .001 ***	→
TW-EVS					
Paragraph 2−1	0.321	.114	0.380	.065 †	→
Paragraph 3−1	0.779	< .001 ***	0.849	< .001 ***	→
Paragraph 4−1	2.084	< .001 ***	2.116	< .001 ***	→
Paragraph 3−2	0.457	.008 **	0.469	.012 *	↓
Paragraph 4−2	1.762	< .001 ***	1.737	< .001 ***	→
Paragraph 4−3	1.305	< .001 ***	1.267	< .001 ***	→
Sentence onset EVS					
Paragraph 2−1	−0.035	.991	−0.098	.864	→
Paragraph 3−1	0.403	.009 **	0.412	.012 *	→
Paragraph 4−1	0.855	< .001 ***	0.878	< .001 ***	→
Paragraph 3−2	0.438	.003 **	0.510	< .001 ***	↑
Paragraph 4−2	0.890	< .001 ***	0.977	< .001 ***	→
Paragraph 4−3	0.452	.002 **	0.467	.003 **	→
Sentence offset EVS					
Paragraph 2−1	0.258	.299	0.286	.265	→
Paragraph 3−1	0.685	< .001 ***	0.739	< .001 ***	→
Paragraph 4−1	1.692	< .001 ***	1.695	< .001 ***	→
Paragraph 3−2	0.427	.026 *	0.453	.025 *	→
Paragraph 4−2	1.434	< .001 ***	1.410	< .001 ***	→

Paragraph 4−3	1.007	< .001 ***	0.956	< .001 ***	→
CT-span					
Paragraph 2−1	0.200	.222	0.256	.102	→
Paragraph 3−1	−0.081	.880	−0.023	.997	→
Paragraph 4−1	−0.016	.999	−0.014	.999	→
Paragraph 3−2	−0.281	.022 †	−0.279	.040 †	→
Paragraph 4−2	−0.216	.129	−0.242	.106	→
Paragraph 4−3	−0.065	.925	−0.037	.987	→
Sentence-span					
Paragraph 2−1	0.866	.080 †	1.135	.012 *	↑
Paragraph 3−1	0.272	.886	0.618	.370	→
Paragraph 4−1	2.252	< .001 ***	2.408	< .001 ***	→
Paragraph 3−2	−0.594	.194	−0.517	.324	→
Paragraph 4−2	1.386	< .001 ***	1.272	< .001 ***	→
Paragraph 4−3	1.980	< .001 ***	1.790	< .001 ***	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or *p*-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 8 Robustness check for the between-group *t*-tests of the EVS data

Overall					
	Original *t*	**Original *p***	**Subsampled *t***	**Subsampled *p***	**Change in significance**
CV-EVS	0.646	0.519	0.580	0.562	→
TW-EVS	2.515	0.012 *	2.203	0.028 *	→
Sentence onset EVS	1.061	0.289	1.136	0.256	→
Sentence offset EVS	1.718	0.086 †	1.382	0.167	↓
CT-span	1.883	0.060 †	1.636	0.103	↓
Sentence-span	–0.689	0.491	–0.958	0.339	→
Predictable					
CV-EVS	0.391	0.696	0.369	0.712	→
TW-EVS	2.349	0.019 *	2.067	0.040 *	→
Sentence onset EVS	0.365	0.715	0.296	0.768	→
Sentence offset EVS	1.666	0.096 †	1.348	0.179	↓
CT-span	1.670	0.096 †	1.321	0.188	↓
Sentence-span	–0.017	0.986	–0.022	0.982	→
Unpredictable					
CV-EVS	0.455	0.650	0.356	0.722	→
TW-EVS	1.236	0.217	1.081	0.281	→
Sentence onset EVS	1.113	0.267	1.300	0.194	→
Sentence offset EVS	0.846	0.398	0.666	0.506	→
CT-span	1.082	0.280	1.082	0.280	→

Sentence-span	–1.001	0.318	–1.378	0.169	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or *p*-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 9 Robustness check for by-group GCA of the early and the late interpreters' eye-tracking data

Professionals						
EVS	**Item-based accuracy**	**Original *r***	**Original *p***	**Subsampled *r***	**Subsampled *p***	**Change in significance**
Overall						
CV-EVS	CV accuracy	.034	.827	.048	.758	→
TW-EVS	TW accuracy	–.149	.335	–.201	.191	→
Sentence onset EVS	Sentence accuracy	–.572	< .001 ***	–.536	< .001 ***	→
Sentence offset EVS	Sentence accuracy	–.256	.093 †	–.320	.034 *	↑
Predictable						
CV-EVS	CV accuracy	.077	.734	.034	.881	→
TW-EVS	TW accuracy	.041	.855	–.144	.887	→
Sentence onset EVS	Sentence accuracy	–.373	.088 †	–.425	.048 *	↑
Sentence offset EVS	Sentence accuracy	–.195	.384	–.240	.281	→
Unpredictable						
CV-EVS	CV accuracy	.045	.841	.038	.867	→
TW-EVS	TW accuracy	–.004	.987	–.090	.692	→
Sentence onset EVS	Sentence accuracy	–.606	.003 **	–.536	.010 *	↓
Sentence offset EVS	Sentence accuracy	–.208	.353	–.276	.24	→
Students						
EVS	**Item-based accuracy**	**Original *r***	**Original *p***	**Subsampled *r***	**Subsampled *p***	**Change in significance**
Overall						

CV-EVS	CV accuracy	–.262	.015 *	–.178	.101	↓
TW-EVS	TW accuracy	–.230	.033 *	–.206	.057 †	↓
Sentence onset EVS	Sentence accuracy	–.397	< .001 ***	–.372	< .001 ***	→
Sentence offset EVS	Sentence accuracy	–.158	.146	–.124	.255	→
Predictable						
CV-EVS	CV accuracy	.061	.695	.179	.252	→
TW-EVS	TW accuracy	.027	.863	–.075	.630	→
Sentence onset EVS	Sentence accuracy	–.435	.004 **	–.370	.015 *	↓
Sentence offset EVS	Sentence accuracy	.066	.674	.062	.691	→
Unpredictable						
CV-EVS	CV accuracy	–.281	.068 †	–.230	.137	→
TW-EVS	TW accuracy	–.057	.715	.034	.827	→
Sentence onset EVS	Sentence accuracy	–.320	.036 *	–.339	.026 *	→
Sentence offset EVS	Sentence accuracy	–.226	.145	–.155	.321	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 10 Codes concerning anticipation

		Professional		Student		Total	
Main themes	**Coding**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Anticipation modes/cues	World-based anticipation	13 (22)	59.1	17 (23)	38.6	30 (45)	45.5
	Discourse based anticipation	8 (18)	36.4	8 (11)	18.2	16 (29)	24.2
	Glossary-based anticipation	5 (7)	22.7	10 (11)	22.7	15 (18)	22.7
	Visual-based anticipation	2 (3)	9.1	8 (10)	18.2	10 (13)	15.2
	Lexical-based anticipation	2 (5)	9.1	4 (5)	9.1	6 (10)	9.1
Contents of anticipation	Topics/events/scenarios	12 (24)	54.5	23 (40)	52.3	35 (64)	53.0
	Specific words	9 (15)	40.9	9 (13)	20.5	18 (28)	27.3
	Semantic categories/hypernym	1 (2)	4.5	3 (4)	6.8	4 (6)	6.1
Anticipation frequency/attitudes	Always/Actively	6 (7)	27.3	13 (18)	29.5	19 (25)	28.8
	Barely	7 (8)	31.8	9 (10)	20.5	16 (18)	24.2
	Occasionally	1 (1)	4.5	7 (8)	15.9	8 (9)	12.1
	Never	4 (4)	18.2	2 (2)	4.5	6 (6)	9.1

#Notes: No. = number of participants (number of codes); % = percentage of participants in respective expertise group.

Appendix 11 Codes concerning visual inputs

		Professional		Student		Total	
Main themes	**Coding**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Viewing pattern	Automatically viewing	10 (11)	45.5	9 (10)	13.6	19 (21)	28.8
	Always viewing	6 (6)	27.3	4 (6)	9.1	10 (12)	(15.2)
	Checking when possible	6 (6)	22.7	4 (7)	9.1	9 (15)	(13.6)
	Intentionally avoid viewing	4 (6)	18.2	5 (5)	11.4	9 (11)	(13.6)
	Ignore	3 (4)	13.6	4 (4)	9.1	7 (8)	(10.6)
Audio-visual input interaction	Locate the target after hearing the TW	11 (12)	50	12 (19)	27.3	23 (31)	(34.8)
	Locate the target before hearing the TW	3 (4)	13.6	6 (12)	13.6	9 (16)	(13.6)
	Intentionally searching for the target	3 (5)	13.6	3 (4)	6.8	6 (9)	(9.1)
	Simultaneously hear and locate the target	4 (6)	18.2	1 (1)	2.3	5 (7)	(7.6)
	Locate and fixate on a wrong object	1 (1)	4.5	2 (2)	4.5	3 (3)	(4.5)
Effects of visual inputs	Distracting	10 (11)	45.5	11 (14)	25	21 (25)	(31.8)
	Misleading	5 (8)	22.7	7 (8)	15.9	12 (16)	(18.2)
	Reminding the ST	4 (4)	18.2	6 (8)	13.6	10 (12)	(15.2)
	Neutral/unclear	4 (4)	18.2	5 (6)	11.4	9 (10)	(13.6)
	Generally helpful	2 (2)	9.1	4 (4)	9.1	6 (6)	(9.1)
	Confirming the ST	1 (1)	4.5s	3 (3)	6.8	4 (4)	(6.1)

#Notes: No. = number of participants (number of codes); % = percentage of participants in respective expertise group.

Appendix 12 Codes concerning SI tasks

		Professional		Student			
Main themes	**Coding**	**No.**	**%**	**No.**	**%**	**No.**	**%**
ST difficulty	Information density	8 (16)	36.4	21 (28)	47.7	29 (44)	43.9
	Unfamiliar contents	8 (8)	36.4	21 (25)	47.7	29 (33)	43.9
	Unfamiliar style	8 (11)	36.4	21 (21)	47.7	29 (32)	43.9
	Slow speech	8 (11)	36.4	5 (5)	11.4	13 (16)	19.7
	Low coherence	3 (3)	13.6	2 (3)	4.5	5 (6)	7.6
	Complicated/idiosyncratic syntactic structure	1 (1)	4.5	4 (4)	9.1	5 (5)	7.6
Challenge encountered	Code switching	13 (21)	59.1	25 (52)	56.8	38 (73)	57.6
	Listening failure	17 (32)	77.3	20 (40)	45.5	37 (72)	56.1
	Omission	12 (33)	54.5	18 (31)	40.9	30 (64)	45.5
	Omission due to time constraints	4 (5)	18.2	6 (9)	13.6	10 (14)	15.2
	Output language polishing	8 (11)	36.4	8 (9)	18.2	16 (20)	24.2
	Inaccurate interpretation	8 (10)	36.4	3 (3)	6.8	11 (13)	16.7
	Misunderstanding	3 (3)	13.6	7 (8)	15.9	10 (11)	15.2
Interpreting strategies	Omitting descriptive contents	1(1)	4.5	5 (5)	11.4	6 (6)	9.1
	Replenishing based on world knowledge	2 (2)	9.1	5 (6)	11.4	7 (8)	10.6
	Note taking	1 (1)	4.5	2 (2)	4.5	3 (3)	4.5
	Syntactic linearity	1 (1)	4.5	1 (1)	2.3	2 (2)	4.0
	Fillers	0 (0)	0	1 (1)	2.3	1 (1)	1.5
Meta-cognitive strategies	Adapting to the SI task	8 (12)	36.4	15 (15)	34.1	23 (27)	34.8

EVS control	7 (8)	31.8	7 (11)	15.9	14 (19)	21.2
Output monitoring	4 (5)	18.2	7 (7)	15.9	11 (12)	16.7
Memory pressure	3 (3)	13.6	2 (3)	4.5	5 (5)	7.6
Cognitive load/fatigue	0 (0)	0	4 (5)	9.1	4 (5)	6.1

#Notes: No. = number of participants (number of codes); % = percentage of participants in respective expertise group.

Appendix 13 Robustness check for by-group GCA of the early and the late interpreters' eye-tracking data

Early interpreters					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
Pre × Plau	−0.578	< .001 ***	−0.549	< .001 ***	→
Pre × Implau	−0.301	< .001 ***	−0.266	< .001 ***	→
Pre × Distra	−0.352	< .001 ***	−0.301	< .001 ***	→
Pre × ot1	−0.824	.002 **	−1.018.	< .001 ***	↑
Pre × ot2	−2.382	< .001 ***	−2.475	< .001 ***	→
Pre × ot3	−0.620	.018 *	3.622	.185	↓
Pre × Plau × ot1	−1.667	.653	−1.497	.698	→
Pre × Implau × ot1	1.497	< .001 ***	1.558	< .001 ***	→
Pre × Distra × ot1	1.634	< .001 ***	2.268	< .001 ***	→
Pre × Plau × ot2	3.816	< .001 ***	4.302	< .001 ***	→
Pre × Implau × ot2	3.581	< .001 ***	3.615	< .001 ***	→
Pre × Distra × ot2	1.931	< .001 ***	1.854	< .001 ***	→
Pre × Plau × ot3	−0.775	.036 *	−0.286	.460	↓
Pre × Implau × ot3	−0.344	.354	−0.195	.614	→
Pre × Distra × ot3	−0.266	.473	0.233	.546	→
Late interpreters					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Variation**
Pre × Plau	−0.308	< .001 ***	−0.300	< .001 ***	→
Pre × Implau	−0.147	. 002 **	−0.834	. 097 †	↓

Pre × Distra	−0.196	< .001 ***	−0.274	< .001 ***	→
Pre × ot1	0.370	.145	0.501	.059 †	↑
Pre × ot2	−0.669	.008 **	−0.568	.032 *	↓
Pre × ot3	0.109	.668	0.035	.895	→
Pre × Plau × ot1	0.125	.727	−0.539	.150	→
Pre × Implau × ot1	0.093	.794	−0.047	.899	→
Pre × Distra × ot1	−0.859	.016 *	−0.833	.026 *	→
Pre × Plau × ot2	1.115	.002 **	1.033	.006 **	→
Pre × Implau × ot2	0.614	.086 †	0.283	.450	↓
Pre × Distra × ot2	1.349	< .001 ***	1.184	.002 **	↓
Pre × Plau × ot3	−0.264	.460	0.090	.810	→
Pre × Implau × ot3	−0.842	.018 *	−0.772	.039 *	→
Pre × Distra × ot3	−0.301	.400	−0.404	.281	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 14 Robustness check for by-group GCA of the high- and the low-quality interpreters' eye-tracking data

High-quality interpreters					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
Pre × Plau	−0.726	< .001 ***	−0.625	< .001 ***	→
Pre × Implau	−0.462	< .001 ***	−0.396	< .001 ***	→
Pre × Distra	−0.362	< .001 ***	−0.321	< .001 ***	→
Pre × ot1	−0.859	.002 **	−0.515	.075 †	↓
Pre × ot2	−2.990	< .001 ***	−2.981	< .001 ***	→
Pre × ot3	−1.038	< .001 ***	−1.131	< .001 ***	→
Pre × Plau × ot1	−1.537	< .001 ***	−1.258	.002 **	↓
Pre × Implau × ot1	−0.606	.120	0.014	.973	→
Pre × Distra × ot1	−1.771	< .001 ***	−1.358	< .001 ***	→
Pre × Plau × ot2	3.921	< .001 ***	4.049	< .001 ***	→
Pre × Implau × ot2	3.998	< .001 ***	4.032	< .001 ***	→
Pre × Distra × ot2	3.370	< .001 ***	3.102	< .001 ***	→
Pre × Plau × ot3	.1551	< .001 ***	2.071	< .001 ***	→
Pre × Implau × ot3	0.670	.085 †	0.658	.107	↓
Pre × Distra × ot3	1.532	< .001 ***	1.778	< .001 ***	→
Low-quality interpreters					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Variation**
Pre × Plau	−0.231	< .001 ***	−0.278	< .001 ***	→
Pre × Implau	−0.051	.271	−0.027	.716	→

Pre × Distra	−0.214	< .001 ***	−0.273	< .001 ***	→
Pre × ot1	−0.951	< .001 ***	−0.741	.003 **	↓
Pre × ot2	−0.360.	.137	−0.359	.155	→
Pre × ot3	1.422	< .001 ***	1.212	< .001 ***	→
Pre × Plau × ot1	0.828	.016 *	0.252	.480	↓
Pre × Implau × ot1	1.736	< .001 ***	1.194	< .001 ***	→
Pre × Distra × ot1	1.879	< .001 ***	2.128	< .001 ***	→
Pre × Plau × ot2	1.271	< .001 ***	1.516	< .001 ***	→
Pre × Implau × ot2	0..550	.108	0..258	.470	→
Pre × Distra × ot2	0.307	.371	0.296	.407	→
Pre × Plau × ot3	−2.093	< .001 ***	−1.754	< .001 ***	→
Pre × Implau × ot3	−1.572	< .001 ***	−1.373	< .001 ***	→
Pre × Distra × ot3	−1.134	. < .001 ***	−0.803	.025 *	↓

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 15 Robustness check for by-group GCA of the anticipators and the non-anticipators' eye-tracking data

Anticipators					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Change in significance**
Pre × Plau	−0.410	< .001 ***	−0.236	< .001 ***	→
Pre × Implau	−0.261	< .001 ***	−0.199	< .001 ***	→
Pre × Distra	−0.263	< .001 ***	−0.151	.020 *	↓
Pre × ot1	−0.457	.164	−0.479	.163	→
Pre × ot2	−1.085	< .001 ***	−1.179	< .001 ***	→
Pre × ot3	0.311	.232	−0.148	.665	→
Pre × Plau × ot1	−1.967	< .001 ***	−2.112	< .001 ***	→
Pre × Implau × ot1	2.297	< .001 ***	2.291	< .001 ***	→
Pre × Distra × ot1	1.911	< .001 ***	2.106	< .001 ***	→
Pre × Plau × ot2	2.039	< .001 ***	2.400	< .001 ***	→
Pre × Implau × ot2	2.404	< .001 ***	2.519	< .001 ***	→
Pre × Distra × ot2	0.617	.184	0.917	.058 †	↑
Pre × Plau × ot3	0.380	.412	1.136	.019 *	↑
Pre × Implau × ot3	−0.570	.219	−0.092	.848	→
Pre × Distra × ot3	−0.514	.267	−0.197	.684	→
Non-anticipators					
Fixed effect	**Original Estimate**	**Original *p***	**Subsampled Estimate**	**Subsampled *p***	**Variation**
Pre × Plau	−0.285	< .001 ***	−0.368	< .001 ***	→
Pre × Implau	−0.429	< .001 ***	−0.393	< .001 ***	→

Pre × Distra	−0.549	< .001 ***	−0.657	< .001 ***	→
Pre × ot1	−0.251	.495	−0.473	.215	→
Pre × ot2	−0.900	.013 *	−1.125	.003 **	→
Pre × ot3	0.091	.804	0.206	.586	→
Pre × Plau × ot1	0.651	.206	0.620	.248	→
Pre × Implau × ot1	−0.029	.956	−0.063	.907	→
Pre × Distra × ot1	−0.781	.130	−0.125	.816	→
Pre × Plau × ot2	1.845	< .001 ***	2.541	< .001 ***	→
Pre × Implau × ot2	0.700	.173	0.667	.212	→
Pre × Distra × ot2	0.816	.112	0.663	.215	→
Pre × Plau × ot3	−0.739	.151	−0.663	.215	→
Pre × Implau × ot3	−0.419	.416	−0.535	.318	→
Pre × Distra × ot3	0.295	.566	0.246	.646	→

#Notes: † $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$. →: No meaningful change in estimate direction or p-value significance; ↑: More significant effect in the subsampled dataset; ↓: Less significant effect in the subsampled dataset. An estimate direction flip would also be marked with ↑ or ↓ depending on whether the effect switched from significant to non-significant, or vice versa.

Appendix 16 Participant information sheet

Participant Information Sheet

Project title: Predictive processing during simultaneous interpreting

Researcher(s): Mingqing Xie

Department: Modern Languages and Cultures

Contact details: mingqing.xie@durham.ac.uk

Supervisor name: Binghan Zheng

Supervisor contact details: binghan.zheng@durham.ac.uk

You are invited to take part in a study that I am conducting as part of my PhD project at Durham University.

This study has received ethical approval from the ethics committee of School of Modern Languages and Cultures of Durham University.

Before you decide whether to agree to take part it is important for you to understand the purpose of the research and what is involved as a participant. Please read the following information carefully. Please get in contact if there is anything that is not clear or if you would like more information.

The rights and responsibilities of anyone taking part in Durham University research are set out in our 'Participants Charter': https://www.dur.ac.uk/research.innovation/governance/ethics/considerations/people/charter/

What is the purpose of the study?

The aim of this study is to investigate the cognitive process of simultaneous interpreting using eye-tracking method. The project is partly supported by the CSC-Durham doctoral scholarship and is planned to be completed by the end of 2024.

Why have I been invited to take part?

You have been invited because you are a master student in translation and interpreting and have received at least a year of simultaneous interpreting training.

Do I have to take part?

Your participation is voluntary, and you do not have to agree to take part. If you do agree to take part, you can withdraw at any time, without giving a reason. If you withdraw from the study during or after data gathering, we will delete your data and there is no penalty or loss of benefits to which you are otherwise entitled. Your rights in relation to withdrawing any data that is identifiable to you are explained in the accompanying Privacy Notice.

What will happen to me if I take part?

If you agree to take part in the study, you will interpret four English paragraphs into Chinese and look on a screen in front of you. Your eye-movements will be monitored while you interpret in order to

record how you process combine visual and auditory inputs. After each paragraph, we may have some questions about your experience. You can omit any questions you do not wish to answer. The experiment take place at the eye-tracker lab at the Elvet Riverside II of MLAC. Your session should last for 45-60 minutes. You will be given full instructions and will be able to ask any questions you may have. You will be paid £12 for your participation in this eye-tracking experiment.

There will be another online session where you need to answer questions mainly regarding your language proficiency. This session should last less than 45 minutes, and you will be paid £10 for your participation in this experiment.

Are there any potential risks involved?

There are no known risks to participation in this study. Other than the payment mentioned, there is no tangible benefits to you. However, you will be contributing to our knowledge about language.

Will my data be kept confidential?

All information obtained during the study will be kept confidential. If the data is published it will be entirely anonymous and will not be identifiable as yours.

Full details are included in the accompanying Privacy Notice.

What will happen to the results of the project?

The results are expected to be included in my PhD dissertation. No personal data will be shared, however anonymised (i.e., not identifiable) data may be used in publications, reports, presentations, and other research outputs. At the end of the project, anonymised data may be archived and shared with others for legitimate research purposes.

All research data and records needed to validate the research findings will be stored for 10 years after publication of the results.

Durham University is committed to sharing the results of its world-class research for public benefit. As part of this commitment the University has established an online repository for all Durham University Higher Degree theses which provides access to the full text of freely available theses. The study in which you are invited to participate will be written up as a thesis. On successful submission of the thesis, it will be deposited both in print and online in the University archives, to facilitate its use in future research. The thesis will be published open access.

Who do I contact if I have any questions or concerns about this study?

If you have any further questions or concerns about this study, please speak to me by email at mingqing.xie@durham.ac.uk. If you remain unhappy or wish to make a formal complaint, please submit a complaint via the University's Complaints Process.

Thank you for reading this information and considering taking part in this study.

Appendix 17 Consent form

Consent Form

Project title: Predictive processing during simultaneous interpreting
Researcher(s): Mingqing Xie
Department: Modern Languages and Cultures
Contact details: mingqing.xie@durham.ac.uk

Supervisor name: Binghan Zheng
Supervisor contact details: binghan.zheng@durham.ac.uk

This form is to confirm that you understand what the purposes of the project, what is involved and that you are happy to take part. Please initial each box to indicate your agreement:

I confirm that I have read and understand the information sheet dated [__/__/__] and the privacy notice for the above project.	
I have had sufficient time to consider the information and ask any questions I might have, and I am satisfied with the answers I have been given.	
I understand who will have access to personal data provided, how the data will be stored, and what will happen to the data at the end of the project.	
I consent to being audio recorded / being video recorded / having my photo taken and understand how recordings / photos will be used in research outputs.	
I understand that my words may be quoted anonymously in publications, reports, and other research outputs.	
I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason.	
I agree to take part in the above project.	

Participant's Signature _____ Date _____

(NAME IN BLOCK LETTERS) _____

Researcher's Signature _____ Date _____

(NAME IN BLOCK LETTERS) _____

Appendix 18 Privacy notice

Privacy Notice

Durham University's responsibilities under data protection legislation include the duty to ensure that we provide individuals with information about how we process personal data. We do this in a number of ways, one of which is the publication of privacy notices. This privacy notice provides a general description of the broad range of processing activity in addition there are tailored privacy notices covering some specific processing activity.

To ensure that we process your personal data fairly and lawfully we are required to inform you:

- Why we collect your data
- How it will be used
- Who it will be shared with

We will also explain what rights you have to control how we use your information and how to inform us about your wishes. Durham University will make the Privacy Notice available via the website and at the point we request personal data.

Our privacy notices comprise two parts – a generic part (common to all of our privacy notices) and a part tailored to the specific processing activity being undertaken.

PART 1 – GENERIC PRIVACY NOTICE

Please access our General Privacy Notice online.

PART 2 – TAILORED PRIVACY NOTICE

This section of the Privacy Notice provides you with the privacy information that you need to know before you provide personal data to the University for the particular purpose(s) stated below.

Project Title: Predictive processing during simultaneous interpreting

Type(s) of personal data collected and held by the researcher and method of collection:
Personal data will be collected through eye tracker, questionnaire, and interviews. This will include name, age, language proficiency, and your experience of English learning and interpreting training. Your eye movements during the experiment will be recorded, and an audio recording of your interpreting output will be produced simultaneously.

Lawful Basis
Under data protection legislation, we need to tell you the lawful basis we are relying on to process your data. The lawful basis we are relying on is public task: the processing is necessary for an activity being carried out as part of the University's public task, which is defined as teaching, learning and research.

How personal data is stored:
All personal data will be held securely and strictly confidential to the research team. You will be allocated an anonymous number for data collection. Information that identifies you will be kept separate from the anonymised data. Signed consent forms will be stored separately to project data. All personal data in electronic form will be stored on a password protected computer, and any hardcopies will be

kept in locked storage. Data will not be available to anyone outside the research team. The conversation will be recorded and stored on an encrypted device until it has been transcribed by the researcher. No-one else will have access to the recording, and it will be erased once the transcript has been completed.

How personal data is processed:
Your eye movement data are collected to analyse how the visual information is used during simultaneous interpreting. The audio recordings of your interpreting outputs are collected to analyse your interpreting performance and the effect of visual information on interpreting performance. Information will be entered into a database for analysis. After six months the data will be completely anonymised and the original records, including any information which can identify you personally, will be destroyed. The recorded conversation will be transcribed by the researcher, and personal information will be coded and anonymised. The original recording will then be erased.

Withdrawal of data
You can request withdrawal of your data until it has been fully anonymised. Once this has happened it will not be possible to identify you from any of the data we hold.

Who the researcher shares personal data with:
No identifiable data will be shared with anyone outside the research team.

Please be aware that if you disclose information which indicates the potential for serious and immediate harm to yourself or others, the research team may be obliged to breach confidentiality and report this to relevant authorities. This includes disclosure of child protection offences such as the physical or sexual abuse of minors, the physical abuse of vulnerable adults, money laundering, or other crimes covered by prevention of terrorism legislation. Where you disclose behaviour (by yourself or others) that is potentially illegal but does not present serious and immediate danger to others, the researcher will, where appropriate, signpost you to relevant services, but the information you provide will be kept confidential (unless you explicitly request otherwise).

How long personal data is held by the researcher:
We will hold personal data for six months, after which it will be anonymised.

How to object to the processing of your personal data for this project:
If you have any concerns regarding the processing of your personal data, or you wish to withdraw your data from the project, please contact me via mingqng.xie@durham.ac.uk.

Further information:
Researcher(s): Mingqing Xie
Department: Modern Languages and Cultures
Contact details: mingqing.xie@durham.ac.uk

Supervisor name: Binghan Zheng
Supervisor contact details: binghan.zheng@durham.ac.uk

Appendix 19 Debriefing sheet

Debriefing Sheet

Project title: Predictive processing during simultaneous interpreting

Thank you for taking part in this study. What I want to find out from this research is whether or not interpreters make prediction during simultaneous interpreting (SI) tasks and if so, what is the underpinning cognitive mechanism. With visual scenes provided on the monitor, your eye movements were recorded during SI tasks. The expectation is that in the predictable condition, upon hearing a critical word (e.g., illuminated) in the experimental sentences (e.g., the corridor is illuminated by hundreds of candles), participants make predictive eye movements to the target item (i.e., candle) before actually hearing the target word (i.e., candles). On the other hand, in the unpredictable condition, the critical word (e.g., decorated) should not trigger predictive eye movements to the target item since there are more than one item that could fit in the context (e.g., the corridor is decorated by hundreds of candles/paintings). The predictive eye movements in the predictable condition are evidence of predictive processing by interpreters.

The data you have provided is automatically anonymised and cannot be traced back to your identity. If you hope to withdraw your data, please notify me as soon as possible and no later than six months from the experiment. This is because the data are expected to be completely anonymised then and it is impossible to identify you from any of the data we hold.

If you would like further information about the study or would like to know about what my findings are when all the data have been collected and analysed then please contact me on mingqing.xie@durham.ac.uk. I cannot however provide you with your individual results.