Using the Visual World Paradigm to Reveal the Temporal Dynamics of Spoken Language Processing: Postprint

Authors: Wei Yipu

Date: 2023-10-09T00:00:00+00:00

Abstract

The visual world paradigm is an eye-tracking experimental paradigm that investigates real-time spoken language processing by tracking and measuring fixation patterns on visual objects. The theoretical foundation for applying this paradigm to language comprehension research rests on eye-movement linking hypotheses (e.g., cooperative interaction theory, goal-based linking hypothesis, etc.), which establish meaningful associations between eye-movement trajectories and spoken language processing. Data acquired through the visual world paradigm can provide precise temporal information for spoken language processing, with commonly employed analytical methods including: time-region-of-interest analysis of fixation proportions, divergence point analysis, growth curve analysis, etc. This paradigm furnishes crucial evidence for studying issues such as lexical spoken word recognition, syntactic ambiguity resolution, semantic comprehension, and discourse-pragmatic information processing.

Full Text

Using the Visual World Paradigm to Reveal the Time Course of Spoken Language Processing

WEI Yipu
School of Chinese as a Second Language, Peking University, Beijing 100871, China

Abstract

The visual world paradigm is an eye-tracking experimental method that investigates real-time spoken language processing by tracking and measuring eye movements to visual objects. The theoretical foundation for applying this paradigm to language comprehension research lies in linking hypotheses (e.g., the coordinated interplay account, goal-based linking hypothesis) that establish meaningful connections between eye movement patterns and language processing dynamics. Data obtained using the visual world paradigm provide precise temporal information about spoken language processing, with common analytical methods including analysis of mean fixation proportions within time windows, divergence point analysis, and growth-curve analysis. This paradigm has provided crucial evidence for research on lexical and phonological recognition, syntactic ambiguity resolution, semantic comprehension, and discourse-level pragmatic information processing.

Keywords: visual world paradigm; eye-tracking; spoken language processing

1. Linking Hypotheses Between Eye Movement Trajectories and Language Processing

The theoretical foundation of eye-tracking research using the visual world paradigm rests on linking hypotheses that connect eye movement patterns to the cognitive processes underlying spoken language comprehension (Allopenna et al., 1998; Tanenhaus et al., 2000). Specifically, as listeners process spoken information, they construct dynamic mental representations of the described situations. Their attention to specific entities within these representations shifts with the incoming linguistic input, and correspondingly, their fixation points in visual space move accordingly (Altmann & Kamide, 2007). This concentration and movement of attention, accompanied by changes in pupil position, can be effectively measured through eye-tracking techniques, thereby revealing the time course of spoken language processing.

Over the past two decades, researchers have proposed numerous specific linking hypotheses to explain the relationship between eye movements and spoken language processing, which are used to elucidate how visual attention is allocated to referent objects (see review by Magnuson, 2019). This paper summarizes three influential linking hypotheses to further clarify the theoretical basis for applying the visual world paradigm to spoken language processing research. While these hypotheses do not directly define the time course of specific linguistic elements, they encompass several stages of spoken language processing and provide the foundational premises for investigating processing time courses.

Knoeferle and Crocker (2006, 2007) proposed the **coordinated interplay account**, which divides visual scene-based spoken language comprehension into three main stages: (1) integrating newly input words into the existing sentence structure to form a new sentence interpretation, and using this new information together with prior linguistic information and relevant world knowledge to generate predictions about subsequent utterances; (2) searching working memory (which includes the previously viewed visual scene) for objects referred to by words or objects predictable based on information from the first stage; and

(3) mapping linguistic input (nouns, verbs, etc.) onto objects and actions in the visual scene, revising previously formed sentence interpretations based on visual scene information, and forming new predictions (Knoeferle & Crocker, 2006, 2007; Pyykkönen-Klauck & Crocker, 2016). Notably, although these three processes are presented sequentially in the coordinated interplay account, the theory does not preclude the possibility that they may overlap in time or occur simultaneously. The coordinated interplay account highlights the importance of visual scene information for spoken language comprehension; furthermore, although these scenes gradually fade from working memory after the visual scene disappears, memories of the scenes still exert significant influence on subsequent sentence processing (Knoeferle & Crocker, 2007).

Altmann and Mirković (2009) proposed an alternative linking hypothesis that similarly acknowledges that sentence processing is jointly influenced by linguistic information (e.g., real-time linguistic input, contextual information) and non-linguistic information (e.g., visual scenes, world knowledge). However, unlike Knoeferle and Crocker' s (2006, 2007) coordinated interplay account, Altmann and Mirković (2009) argue that the processes of processing visual scene information and comprehending linguistic input are temporally and psychologically inseparable—because linguistic and non-linguistic information are stored in the same system and together constitute the dynamic representation of the situation. When listeners receive a piece of information, the representation of an object (including experiences and knowledge related to that object) becomes activated. As listeners continuously receive information from different sources (linguistic input, visual scenes, world knowledge, etc.), the representation of the object constantly changes. When information from different sources overlaps, the activation of the object representation strengthens. The different states of this representational system manifest at the mental representation level as attention allocation, which in turn influences eye movement patterns. In other words, accompanying sentence input, participants' fixations on visual objects over time are influenced and driven by a unified representational system that includes linguistic information, contextual information, visual scenes, and world knowledge. Under this theoretical framework, information from different sources has immediate effects on spoken language processing and is rapidly reflected in eye movement trajectories.

Both of the above linking hypotheses adopt a language comprehension perspective, viewing changes in eye fixations during spoken language processing as the result of joint action between linguistic input and visual information. These hypotheses treat language processing as an independent task, unrelated to the behavioral task goals in the experimental procedure. However, such language comprehension-based linking hypotheses do not address the influence of the actions required to complete the task itself on referential processing (Chambers et al., 2004), nor do they consider that in visual search, eye movements themselves are closely related to the goals of the behavioral task—that is, participants fixate more on objects relevant to their task goals. To better explain the relationship between language processing and eye movements, Salverda et al. (2011) proposed the **goal-based linking hypothesis**, which incorporates the new dimension of "task goals" into eye movement linking hypotheses. Unlike language comprehension-based linking hypotheses, the goal-based linking hypothesis posits that not only context and linguistic input can constrain language processing, but task goals themselves can also serve as constraints—visual objects directly relevant to task execution attract more eye fixations, while objects irrelevant to goal execution do not. According to this hypothesis, spoken language processing in visual contexts first involves a fundamental task of mapping linguistic input onto selectable objects in the visual scene, and eye fixations serve this task goal by locking onto potential referent objects; objects that do not afford the action are rarely fixated. For example, when hearing the instruction "put the cube into the can," only cans whose size can accommodate the cube become target containers that are fixated (Chambers et al., 2004). Salverda et al. (2011) argue that additional tasks such as clicking or moving objects together constitute the task goal structure in spoken language processing tasks and influence eye fixations. For instance, when participants listen to sentences with a truth-value judgment task, they show earlier and more significant predictive fixations compared to listening to the same sentences without such a task (Altmann & Kamide, 1999), locking onto referent targets more quickly in the time course. The goal-based linking hypothesis introduces new requirements for refining and hierarchically structuring task goals during language processing.

Research on spoken language processing using the eye-tracking visual world paradigm is based on these linking hypotheses. According to how visual information is utilized, studies can be divided into two main directions. The first direction uses the visual scene as a backdrop for presenting objects—attention to specific referents in the mental representation is projected onto the visual scene, and listeners form fixations on referent objects accordingly; the eye movement trajectories formed by fixating on objects in the backdrop reveal how different linguistic components are processed in real time (e.g., Cooper, 1974; Cozijn et al., 2011; Kaiser, 2016). The second direction treats visual information as a contextual constraint, primarily exploring how information in the visual environment (e.g., number of candidate objects, object size comparisons, depicted event actions) itself influences language processing (e.g., Chambers et al., 2002; Knoeferle et al., 2005; Tanenhaus et al., 1995). These two types of studies employ similar tasks, but at the theoretical level of linking hypotheses, the first emphasizes the simultaneity and inseparability of the processes of comprehending visual scene information and comprehending linguistic input; the second treats visual scene information processing as a relatively independent process, emphasizing the role of the visual scene itself in spoken language processing. The latest trend in eye movement processing research has begun to focus on the potential role of task goals in language processing. Although linking hypotheses that incorporate the goal dimension have completed preliminary theoretical construction, research comparing processing effects under different task goals remains scarce.

2.1 Paradigm and Tasks

A typical visual world paradigm experiment usually includes spoken language instructions presented auditorily and visual objects presented as stimuli (either in the real world or on a computer screen). While participants comprehend the spoken instructions, the positions of their fixations on visual objects are recorded in real time by the eye-tracker for subsequent analysis (see Figure 1 [Figure 1: see original paper]). Visual stimuli generally appear before the linguistic instructions with a certain preview time; linguistic instructions are presented at a relatively fixed playback speed. Previous research has found that factors such as picture complexity, preview duration, speech rate, and task instruction type (whether participants are explicitly told to predict target objects) all exert certain influences on eye movement results (Huettig & Guerra, 2019; Ferreira et al., 2013).

The visual world paradigm mainly includes two different experimental tasks: First, **active tasks** (action-based experimental tasks) require participants to make behavioral responses to linguistic instructions (e.g., grasping, moving, or clicking objects; see Hanna & Tanenhaus, 2004; Tanenhaus et al., 1995). Second, **passive tasks** (listen-and-look tasks) require participants to simply listen to linguistic instructions and look at pictures or scenes without making behavioral responses (Altmann & Kamide, 1999; Knoeferle et al., 2005). Regarding the differences between these two tasks, Salverda et al. (2011) noted that in active-task visual world paradigm experiments, participants direct a large number of fixations to target objects before grasping, moving, or clicking them; passive-task experiments do not exhibit such fixation patterns—this factor may lead to differences in eye movement patterns between the two experimental tasks. Pyykkönen-Klauck and Crocker (2016) reviewed and compared eye movement experimental results using both task types, concluding that some linguistic effects (e.g., word frequency effects) appear more sensitive in eye movement indices in active tasks, with participants locking onto target objects faster and demonstrating more rapid real-time language comprehension processes. Passive listen-and-look visual world paradigm experiments, which do not require participants to complete additional tasks, have relatively better ecological validity (Huettig et al., 2011a) and can be used to examine which spoken language processing effects are universal in language-vision interaction and which exist only under specific experimental tasks (Huettig et al., 2011b).

The visual world paradigm has two main variants: the **printed-word paradigm** (Huettig & McQueen, 2007) and the **blank screen paradigm** (Altmann, 2004). In the printed-word paradigm, visual stimulus pictures are replaced with words appearing on the screen. Participants hear speech input related to the word while their eye movement fixation trajectories on each letter are recorded for analysis. The printed-word paradigm can be used to examine phonological recognition processes and investigate how orthographic information is processed in real time. The blank screen paradigm is primarily used to reveal the role of short-term memory in real-time language processing.

After visual stimulus pictures are presented for a few seconds, a blank screen appears (generally for 1 second), followed by spoken instructions. Experiments using this paradigm can demonstrate that even after objects in the visual stimulus picture disappear, participants still look toward the locations where relevant objects were when hearing linguistic instructions (Knoeferle & Crocker, 2007). The blank screen paradigm provides evidence for mental representation: once formed, mental representations can be temporarily stored in short-term memory without relying on visual stimuli and participate in subsequent language processing.

2.2 Data and Variables

Common dependent variables in visual world paradigm experimental data analysis include fixations and saccades. The most frequently used fixation index is **fixation proportion**, which is the proportion of fixations falling within a specific region of interest across all trials during a designated time window. Common indices for saccade data include saccade proportion (the proportion of saccades directed to the target region of interest across all trials) and saccade reaction time (the time required to make a saccade to the region of interest after the target word stimulus appears). Independent variables in the data can be within-subject variables from the experimental design (e.g., experimental vs. control conditions, ambiguous vs. unambiguous sentences) or between-subject variables (e.g., different language background groups, age groups).

The advantage of the visual world paradigm lies in the high temporal precision of the data it produces. Current research-grade eye-trackers can achieve a sampling rate of 1000 Hz, capturing eye position every millisecond and providing accurate time course information. Taking fixation proportion as an example, researchers can not only compare the mean fixation proportions of different conditions within a specific time window across groups to identify an effect in spoken language processing; more importantly, they can investigate when an effect emerges (i.e., when fixation proportions begin to differ significantly between conditions) and how the effect changes over time.

2.3 Analyzing Data Using Temporal Information

Temporal precision is the most important characteristic of visual world paradigm data, and how to utilize temporal dimension information is key to data analysis in this paradigm. Based on how temporal information is used, existing analytical methods can be categorized into three types: (1) comparing mean fixation proportions within designated time windows; (2) analyzing the time course of effect emergence and duration; and (3) analyzing curve patterns of effects over time. To better illustrate the application scenarios and analytical logic of these three methods, this paper selects schematic diagrams of experimental objects (Figure 2 [Figure 2: see original paper]) and fixation proportion data graphs (Figure 3 [Figure 3: see original paper]) from Allopenna et al. (1998) as examples (detailed discussion of this study appears in Section 3.1).

The first analytical method, the most commonly used and intuitive approach for visual world paradigm data, involves comparing mean fixation proportions within designated time windows—for example, comparing listeners' fixation proportions on several objects in Figure 2 from the onset to the offset of the target word "beaker" (approximately 375 ms). This analytical method treats fixation proportion, duration, or saccade indices as dependent variables and within- and between-subject variables as independent variables, using statistical methods such as t-tests, ANOVA, or linear mixed-effects models to compare differences in fixation proportions between different objects or between different condition groups. Compared to t-tests and ANOVA, mixed-effects models are currently the most widely used analytical method, as they can incorporate random effects for participants and items, achieving more accurate modeling and testing of effects (application examples: Gardner et al., 2021; Grüter et al., 2020). It should be noted that such statistical methods typically require normally distributed data, while fixation proportions range between 0 and 1 and generally require prior log or logit transformation (Ito & Knoeferle, 2022). Analyzing mean fixation proportions within designated time windows is the simplest method for visual world paradigm data analysis and is suitable for most experimental designs. Its main disadvantage is that the artificially set time window reduces the temporal precision of the data and cannot well capture trends in fixation proportions over time; a compensatory approach is to include different time windows as independent variables in the analysis model to test whether the time window variable itself significantly affects fixation proportions.

The second analytical method involves analyzing the time course of effect emergence and duration. This approach fully utilizes the precise time course information of the visual world paradigm and can be used to investigate the exact time when a spoken language processing effect emerges. **Divergence point analysis** divides the time period when a potential effect might emerge into several small time windows (e.g., 20 ms) and compares fixation proportions between two condition groups within each small window to identify when they begin to differ significantly, thereby finding the earliest time point when two fixation proportion curves begin to diverge. For example, in Figure 3, the fixation proportion curve for the target referent "beaker" and the cohort competitor "beetle" diverge at approximately 400 ms, later than the divergence point between the target referent and the rhyme competitor "speaker." Divergence point analysis can statistically calculate the specific time point when different curves begin to diverge significantly.

Simple divergence point analysis can only identify when an effect begins (the divergence point of curves under two conditions) but cannot test the temporal interval of the divergence point or compare whether two divergence points differ statistically across conditions. Advanced divergence point analysis based on bootstrapping can provide confidence intervals for each divergence time point, enabling cross-condition comparisons (Stone et al., 2021; application example:

Corps et al., 2021). Advanced divergence point analysis provides an effective analytical tool for comparing the time courses of real-time language processing across different populations. For example, first-language (L1) and second-language (L2) learners may not differ in the effect size of a particular language processing effect (e.g., predictive processing), but they may differ in when the effect begins (Kaan & Grüter, 2021). Using this analytical method can effectively test whether L2 learners' predictive processing begins significantly later than that of L1 speakers. In addition to divergence point analysis, cluster-based permutation analysis (Barr et al., 2014) and bootstrapped differences of time-series (Seedorff et al., 2018) can both be used to identify when data from two condition groups begin to differ significantly (see review of eye movement data analysis methods: Ito & Knoeferle, 2022). However, these analytical methods cannot analyze how effects change over time across conditions, requiring the third type of method to analyze changing curves.

The third analytical method primarily focuses on analyzing curve patterns of effects over time in the visual world paradigm. Among these, **growth-curve analysis** models and analyzes the curves of fixation proportions over time for key regions of interest under different conditions, testing whether the patterns of fixation proportion curves differ across conditions and thereby verifying whether effects change over time (Mirman, 2014; Mirman et al., 2008). Unlike the first analytical method, growth-curve analysis includes not only linear models with time as a variable but can also incorporate quadratic and cubic terms of time variables to simulate fixation proportion curves that change parabolically over time. For example, in Figure 3, the fixation proportion for the cohort competitor "beetle" shows a parabolic trend of first rising and then falling, with a different slope than the rhyme competitor "speaker," a pattern that can be analyzed using growth-curve models containing quadratic time variables. During spoken language processing, fixation changes over time are often not linear increases or decreases; modeling and comparing changing curves can more precisely analyze the temporal development of language comprehension (application examples: Henry et al., 2022; Koring et al., 2012; Wei et al., 2019). It should be noted that growth-curve analysis suffers from autocorrelation issues, where adjacent time windows show high correlation in fixation positions, increasing the likelihood of Type I statistical errors (false positives) (Huang & Snedeker, 2020). Therefore, it often needs to be combined with the first and second types of analytical methods to jointly verify effects.

Generalized additive mixed models can also be used to model nonlinear data curves, using thin plate regression splines to more flexibly simulate changing curves and reduce statistical autocorrelation, partially compensating for the disadvantages of growth-curve analysis (Porretta et al., 2018).

3. The Visual World Paradigm and the Time Course of Spoken Language Processing

Early debates in the field about the time course of language processing primarily focused on the issue of processing immediacy. Early experiments mainly used lexical recognition, cued recall, and self-paced reading tasks, with evidence tending to support delayed integration processing (e.g., Garnham et al., 1996; Stewart et al., 2000), suggesting that language users process language by delaying integration until the end of the sentence (delayed-integration interpretation; Millis & Just, 1994). However, with the popularization of measurement methods such as eye-tracking and event-related potentials (ERP), precise measurement of reading times and electrophysiological signals became possible, and increasing evidence supported the immediacy of language processing—that is, language users process encountered information immediately as language input unfolds (incremental interpretation; Traxler et al., 1997; Cozijn et al., 2011; Koornneef & van Berkum, 2006). For visual world eye-tracking measurements, although approximately 200 ms is required from receiving auditory language signal stimuli to making an eye movement response (Matin et al., 1993; Saslow, 1967), numerous spoken language experiments using the visual world paradigm have still found effects of eye fixations on target objects after test word onset but before the next word begins, demonstrating that language users process information in spoken language immediately (see Sections 3.1–3.5).

On the basis of widely recognized immediate processing, recent discussions of language processing time courses have focused on when language users utilize contextual information to understand language. Language users may combine the semantics of test words with prior context for processing immediately as the test word appears, or they may engage in predictive processing of the test word' s phonology, semantics, or even syntactic structure during context processing before the test word appears (expectation-based account; Levy, 2008). In detecting predictive effects, the visual world paradigm has distinct advantages over reading paradigms and ERP measurements (Huettig & Guerra, 2019). Most studies using the latter methods can only capture effects produced by the consistency between test word semantics and contextual information at the test word position, whereas the visual world paradigm can test the influence of context on participants' fixation patterns in the visual scene before the keyword appears, providing crucial evidence for predictive processing in spoken language. The following sections focus on analyzing how the visual world paradigm addresses time course issues at different levels of language processing, including phonology, semantics, syntax, discourse, and pragmatics. It should be noted that information at different levels is not independent in spoken language processing but rather mutually influential (see review: Kuperberg & Jaeger, 2016); this paper separates each level for review purposes only.

3.1 Lexical Recognition and Phonological Prediction

In the visual world paradigm, when listeners hear a word, they search for its referent within the visual domain. Based on this characteristic, the visual world paradigm can be used to test lexical recognition processes and investigate how listeners use available information to predict phonological forms. Allopenna et al. (1998) used this paradigm to test whether the matching process between phonological input and lexical representations in spoken word recognition occurs incrementally over time. If this matching process is incremental, it can be predicted that the cohort competitor "beetle"of the target referent "beaker"would produce stronger interference effects than the rhyme competitor "speaker" (see Figure 2), because "beetle"overlaps phonologically with "beaker"at the beginning of the word, while the overlap between "speaker" and "beaker" occurs later. The visual world paradigm eye-tracking results from Allopenna et al. confirmed this hypothesis: fixation proportions on the target object "beaker" and the cohort competitor "beetle" both increased early in phonological processing (see Figure 3), while fixation proportions on "speaker" increased only at later times in word processing, and the magnitude of increase was relatively smaller. The fixation proportion data provided by the visual world paradigm effectively revealed the matching process between phonological input and lexical representations in word recognition.

Regarding whether language users can predict the phonological information of upcoming words through contextual information, existing ERP studies have yielded divergent results and have not obtained stable, replicable phonological prediction effects (DeLong et al., 2005; Nieuwland et al., 2018), while the visual world paradigm provides strong evidence for investigating phonological prediction. Ito et al. (2018) used visual world paradigm eye-tracking experiments and found that in highly predictable contexts (e.g., "The tourists expected rain when the sun went behind the⋯"), listeners not only predictively fixated on the target object ("cloud") but also fixated more on the phonological competitor of the target object ("clown," which shares the initial syllable with "cloud"). This finding confirmed the existence of phonological form prediction. More importantly, this predictive effect in the visual world paradigm emerged 500 ms before the target word appeared, providing strong evidence that prediction of phonological forms in language processing is proactive. Compared to results from other paradigms that only capture integration effects at the target word position, the visual world paradigm provides more direct evidence for language prediction. Furthermore, the visual world paradigm provides empirical evidence for the mechanism of phonological prediction: like semantic prediction, the mechanism behind phonological prediction is based on association—by processing context, language users activate corresponding semantic and phonological forms in the mental lexicon, thereby forming expectations about upcoming words (Kukona, 2020; comparison of phonological and semantic prediction see: Karimi et al., 2019). Notably, phonological prediction research using Western languages faces an unavoidable problem: target words (e.g., "cloud") and their phonological competitors (e.g., "clown") overlap not only phonologically but also in orthographic information. Li et al. (2022) used Mandarin Chinese, where phonological and orthographic information are relatively separate, and through visual world paradigm experiments also found similar phonological form predictions, verifying the universality of phonological prediction.

3.2 Syntactic Processing

The visual world paradigm makes two main contributions to research on the time course of syntactic processing. First, the paradigm can be used to analyze the process of ambiguity resolution in ambiguous sentences, such as garden-path sentences. Tanenhaus et al. (1995) were the first to use the visual world paradigm to investigate the processing of structurally ambiguous English sentences and the influence of visual scenes on sentence disambiguation. For example, "Put the apple on the towel in the box" is structurally ambiguous before "in the box" appears: "on the towel" can be interpreted either as the direction of the action "put" or as a locative modifier for "the apple." Using visual world paradigm eye-tracking, Tanenhaus et al. found that when there is only one apple in the visual scene, listeners are more likely to interpret "on the towel" as the direction of the action (eye movements move directly from the apple to the towel); when there are two apples in the visual scene, listeners are more likely to interpret it as a locative modifier for "the apple" (after fixating on the apple on the towel, they look directly at the true target location—the box).

Second, the visual world paradigm provides new evidence for when different levels of information are processed in syntactic processing. Early two-stage accounts held that in sentence comprehension, syntactic structure analysis precedes the processing of other non-structural information (including lexical semantics, world knowledge, discourse, etc.) (initial syntactic analysis, Frazier, 1987), whereas constraint-based accounts argued that sentence processing involves joint constraints from multiple levels of information (Trueswell et al., 1994) that influence syntactic structure analysis early in sentence processing. Visual world paradigm experimental research supports the latter hypothesis. For example, Snedeker and Trueswell (2004) studied ambiguous prepositional phrase structures ("Choose the cow with the stick" vs. "Tickle the pig with the fan"). The phrase "with the stick/fan" can be interpreted either as a modifier of the object or as an instrument for completing the action. They found that information in the visual scene (number of objects) and verb biases (verbs biased toward modifier interpretation like "choose" vs. verbs biased toward instrument interpretation like "tickle") both influence syntactic structure analysis early in sentence processing, manifesting as listeners looking at different target objects depending on the number of objects and verb bias. Furthermore, Chambers et al. (2002, 2004) found that world knowledge information related to object shape, size, and properties in the visual scene also influences syntactic structure analysis, and these influences all occur at the very beginning of sentence processing, refuting the theoretical hypothesis that syntactic structure analysis is primary.

3.3 Semantic Predictive Processing

A major contribution of the visual world paradigm to semantic processing research is revealing that semantic processing is not only immediate but in many cases even predictive (Altmann & Kamide, 1999; Kamide et al., 2003; theoretical review see: Pickering & Gambi, 2018). Altmann and Kamide (1999) were the first to use the visual world paradigm to study the time course of verb-argument integration: compared to the unrelated verb "move," listeners hearing the verb "eat" in "the boy will eat···" fixated earlier on the cake in the visual scene. This demonstrates that verb semantic information (i.e., "eat" requires an edible argument) helps listeners predict the referent of the argument. Kamide et al. (2003) summarized the main characteristics of semantic processing in subsequent research: (1) the combination of verb and subject promotes semantic prediction—for example, the combination of subject "the man" and verb "ride" predicts the high-probability object "motorbike" ; (2) besides verbs, case markers attached to arguments also activate predictive processing—for instance, in verb-final Japanese, listeners can predict upcoming argument referents through case markers before the verb appears.

Research using the visual world paradigm on semantic processing is not limited to verb-argument structures. Chow and Chen (2020) used this paradigm to study the integration of classifier information and world knowledge in context in Mandarin Chinese, finding that Mandarin speakers can form expectations about upcoming nouns based on world knowledge in context early in processing, and these expectations are influenced by classifiers and further revised later in processing. Additionally, Grüter et al. (2020) found in their study of classifier processing in L1 and L2 speakers that both L1 and L2 speakers are sensitive to the grammatical collocational information contained in classifiers and use this information for predictive processing. However, L2 speakers rely more on semantic information in processing (e.g., the classifier "tiáo" collocates with long, thin objects), showing that when a distractor that does not conform to classifier collocational rules but matches the semantic property of being long and thin appears in the visual scene, L2 speakers fixate more on the distractor.

3.4 Discourse-Level Processing

The visual world paradigm can be used to investigate two important issues in discourse comprehension: referential relations and connective relations. First, eye-tracking in the visual world paradigm can effectively test the establishment process of referential relations between pronouns and antecedents. It is generally believed that when listeners hear a pronoun that has a co-referential relationship with previous discourse and fixate on a related object, it indicates that the object is considered a potential target referent (Runner et al., 2003). Based on this mechanism, researchers have used the visual world paradigm to explore numerous time course issues in referential processing. For example, Arnold et al. (2000) were the first to find that both gender cues and the order in which referents are mentioned have immediate effects on referential resolution:

listeners can use gender-marked forms (e.g., English third-person singular "he" or "she") to lock onto target referents early in processing; simultaneously, the first-mentioned person in a sentence (e.g., the subject in an SVO sentence) is more likely to be interpreted as the referent. In research on the influence of implicit causality on pronoun resolution, Pyykkönen and Järvikivi (2010) found that implicit causality effects appear immediately after the verb—listeners fixate more on referents biased by the verb after hearing it. For example, in "John frightened Bill because⋯," the verb "frightened" is more biased toward the first person, so when listeners hear "frightened," they fixate more on John; whereas in "John feared Bill because⋯," the verb "feared" is more biased toward the second person, and when the verb appears, listeners fixate more on Bill. This finding demonstrates that referential processing occurs immediately and even predictively, rather than through delayed integration (see also: Cozijn et al., 2011).

The visual world paradigm also provides rich empirical evidence for the establishment of connective relations in real-time language comprehension. Wei et al. (2019) used the visual world paradigm to investigate the processing of subjective causal relations (argument-argument) and objective causal relations (cause-effect) and the role of Chinese conjunctions. The study found that when listeners heard the subjective causal conjunction "kějiàn"("it can be seen that") as opposed to the objective causal conjunction "yīn' ér" ("therefore"), they fixated more on the speaker in the visual scene. This suggests that the processing of subjective and objective causal relations may differ in terms of confirming and tracking the speaker, and that the process of tracking the speaker occurs immediately with the input of subjective causal conjunctions. Mak et al. (2017) explored the role of two Russian conjunctions in establishing connective relations by providing two alternative referents in the visual scene and tracking listeners' fixation trajectories on the two referents. The study found that the conjunction "i" ("and," marking continuation relations with consistent subjects across clauses) and the conjunction "a" ("and/but," marking shift relations with different subjects across clauses) could help both monolingual and bilingual children predict whether the subject of the second clause would shift, confirming the existence of predictive processing in spoken discourse comprehension.

3.5 Extraction and Processing of Pragmatic Information

When pragmatic implicatures are processed and whether this process precedes semantic analysis are important issues in pragmatics. The literal-first hypothesis (Huang & Snedeker, 2009, 2011) posits that the processing of the literal meaning of scalar terms (e.g., the semantic interpretation of "some" should be: some-but-possibly-all) precedes the pragmatic implicature of the term (some-but-not-all). Levinson (2000) argues that pragmatic implicatures are processed automatically by default; constraint-based processing theories hold that whether pragmatic implicatures are activated preferentially depends on whether sufficient contextual support exists (Degen & Tanenhaus, 2015, 2016).

The visual world paradigm is an important experimental method for comparing the time courses of semantic and pragmatic information processing. Huang and Snedeker (2011) found in their visual world paradigm eye-tracking experiments that when processing "some," listeners first fixated on objects consistent with the semantic interpretation of "some" (some-but-possibly-all), while using the pragmatic implicature of "some" (some-but-not-all) to disambiguate and exclude referents of "all" occurred later than semantic processing of "some" (by about 800 ms). Degen and Tanenhaus (2016) found that the delayed processing of pragmatic implicatures only occurred when number words also appeared in the instructions; when number words were absent, the processing of pragmatic implicatures for "some" did not occur later than the processing of literal semantic meaning. Gardner et al. (2021) improved the visual object set in Huang and Snedeker' s (2011) experiment to better fit the concept of "some" and found that when sufficient contextual support exists, the processing of pragmatic implicatures is rapid and immediate—that is, listeners can quickly lock onto target objects using the pragmatic implicatures of "some." Furthermore, language users' processing of pragmatic information is greatly influenced by speaker credibility —participants can use the pragmatic meaning of scalar adjectives to lock onto target objects earlier when facing credible speakers, but no early pragmatic processing effects appear when facing less credible speakers (Gardner et al., 2021).

4. Main Contributions, Limitations, and Future Directions

Eye-tracking visual world paradigm research provides two important types of information for language comprehension studies: first, visual dimension fixation indices; second, precise temporal measurement. The former offers rich possibilities for experimental design in psycholinguistics and cognitive psychology; the latter provides accurate time course information for phonological, lexical, syntactic, semantic, discourse, and pragmatic processing at all levels, greatly expanding relevant theories of language comprehension. The combination of the two can effectively reflect how listeners' fixation positions in visual scenes change over time when receiving spoken language input, thereby providing direct evidence for an important issue in language comprehension—the time course of spoken language processing. Experimental research using the visual world paradigm, by analyzing highly time-sensitive eye movement data, has found that processing at all linguistic levels exhibits immediacy and even predictivity, which differs from findings of delayed integration in some early studies, demonstrating that research results on language processing time courses are closely tied to the methods employed. Additionally, because the visual world paradigm mainly relies on listening tasks and does not require participants to have complete literacy skills, it can be used to examine language processing in young children, L2 learners, and elderly populations (research examples see: Canseco-Gonzalez et al., 2010; McMurray et al., 2010; Weber & Cutler, 2004).

One major limitation of the visual world paradigm is its inability to provide processing duration data, thus it cannot answer questions related to language comprehension processing difficulty (Salverda & Tanenhaus, 2018). Moreover, visual world paradigm experiments can only present a limited number of static objects in visual space, which differs from the complex visual environments of everyday language comprehension. Real language comprehension environments may include more objects and dynamic actions and events, leading to certain limitations in the generalizability of results obtained with this paradigm (Huettig et al., 2011a). Additionally, in experimental settings presenting only a limited number of objects, listeners may form certain expectations about linguistic input in advance and strategically fixate on some objects, so eye movement trajectories may not completely reflect language processing (Henderson & Ferreira, 2004). In response to this criticism, Dahan and Tanenhaus (2004) offered a different opinion based on their research on lexical recognition, finding that word frequency effects on lexical recognition were not influenced by whether competitors existed in visual space or by the number of competitors, thereby inferring that presenting a limited number of objects in visual space does not affect the validity of the visual world paradigm.

Eye-tracking research using the visual world paradigm still has considerable room for development. First, although the assumptions about visual and linguistic information comprehension processes proposed in linking hypothesis theories have been confirmed by numerous empirical results, the important role of task goals in language processing remains to be further explored. Comparing how language processing unfolds over time under different task goals will be one future direction for visual world paradigm eye-tracking research. In recent years, eye-tracking research has also begun to use three-dimensional virtual reality (VR) technology, which can highly reproduce natural language communication scenes while maintaining precise control over experimental settings. Some visual world paradigm eye-tracking experiments using VR technology have successfully replicated classic results in language processing, such as predictive language processing (Eichert et al., 2018; Heyselaar et al., 2020). Such technological improvements not only enhance the ecological validity of the visual world paradigm but can also be used to test numerous factors affecting language processing processes in environments approximating real language use. Both theoretical and technological innovations provide new opportunities and greater possibilities for more accurately and effectively collecting and interpreting eye movement data and exploring language processing.

References

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

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The "blank screen paradigm." *Cognition*, 93(2), 79–87. https://doi.org/10.1016/j.cognition.2004.02.005

Altmann, G. T. M., & 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. M., & 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. M., & 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

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76(1), B13–B26. https://doi.org/10.1016/s0010-0277(00)00073-1

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

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., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, 47(1), 30–49. https://doi.org/10.1006/jmla.2001.2832

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

Chow, W. Y., & Chen, D. (2020). Predicting (in)correctly: Listeners rapidly use unexpected information to revise their predictions. *Language, Cognition and Neuroscience*, 35(9), 1149–1161. https://doi.org/10.1080/23273798.2020.1740813

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, 6(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-x

Corps, R. E., Brooke, C., & Pickering, M. J. (2021). Prediction involves stages: Evidence from visual-world eye-tracking. *Journal of Memory and Language*, 122, 104298. https://doi.org/10.1016/j.jml.2021.104298

Cozijn, R., Commandeur, E., Vonk, W., & Noordman, L. G. (2011). The time course of the use of implicit causality information in the processing of pronouns: A visual world paradigm study. *Journal of Memory and Language*, 64(4), 381–403. https://doi.org/10.1016/j.jml.2011.01.001

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498–513. https://doi.org/10.1037/0278-7393.30.2.498

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 39(4), 667–710. https://doi.org/10.1111/cogs.12171

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40(1), 172–201. https://doi.org/10.1111/cogs.12227

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

Eichert, N., Peeters, D., & Hagoort, P. (2018). Language-driven anticipatory eye movements in virtual reality. *Behavior Research Methods*, 50(3), 1102–1115. https://doi.org/10.3758/s13428-017-0929-z

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

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 559–586). Lawrence Erlbaum Associates.

Gardner, B., Dix, S., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2021). Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability. *PLoS ONE*, 16(2), e0245130. https://doi.org/10.1371/journal.pone.0245130

Garnham, A., Traxler, M., Oakhill, J., & Gernsbacher, M. A. (1996). The locus of implicit causality effects in comprehension. *Journal of Memory and Language*, 35(4), 517–543. https://doi.org/10.1006/jmla.1996.0028

Grüter, T., Lau, E., & Ling, W. (2020). How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience*, 35(2), 221–234. https://doi.org/10.1080/23273798.2019.1648840

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, 28(1), 105–115. https://doi.org/10.1016/j.cogsci.2003.10.002

Heyselaar, E., Peeters, D., & Hagoort, P. (2020). Do we predict upcoming speech content in naturalistic environments? *Language, Cognition and Neuroscience*, 36(4), 440–461. https://doi.org/10.1080/23273798.2020.1789005

Henry, N., Jackson, C. N., & Hopp, H. (2022). Cue coalitions and additivity in predictive processing: The interaction between case and prosody in L2 German. *Second Language Research*, 38(3), 397–422. https://doi.org/10.1177/0267658320963151

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). Psychology Press. https://doi.org/10.4324/9780203488430

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

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, 45(6), 1723–1739. https://doi.org/10.1037/a0016704

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, 26(8), 1161–1172. https://doi.org/10.1080/01690965.2010.508641

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions 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., & 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., Olivers, C. N. L., & Hartsuiker, R. J. (2011a). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, 137(2), 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013

Huettig, F., Rommers, J., & Meyer, A. S. (2011b). 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

Ito, A., & Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01969-3

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, 98, 1–11. https://doi.org/10.1016/j.jml.2017.09.002

Kaan, E., & Grüter, T. (2021). Prediction in second language processing and learning: Advances and directions. In E. Kaan & T. Grüter (Eds.), *Prediction in second language processing and learning* (pp. 1–24). John Benjamins.

Kaiser, E. (2016). Discourse-level Processing. In P. Knoeferle, P. Pyykkönen-Klauck, & M. W. Crocker (Eds.), *Visually situated language comprehension* (pp. 151–184). John Benjamins Publishing.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1), 37–55. https://doi.org/10.1023/a:1021933015362

Karimi, H., Brothers, T., & Ferreira, F. (2019). Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions. *Cognitive Psychology*, 112, 25–47. https://doi.org/10.1016/j.cogpsych.2019.04.001

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

Koornneef, A. W., & van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54, 445–465. https://doi.org/10.1016/j.jml.2005.12.003

Koring, L., Mak, P., & Reuland, E. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition*, 123(3), 361–379. https://doi.org/10.1016/j.cognition.2012.02.011

Kukona, A. (2020). Lexical constraints on the prediction of form: Insights from the visual world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(11), 2153–2162. https://doi.org/10.1037/xlm0000935

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

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, X., Li, X., & Qu, Q. (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking in Mandarin Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, 48(5), 531–547. https://doi.org/10.1037/xhp0000999

Mak, W. M., Tribushinina, E., Lomako, J., Gagarina, N., Abrosova, E., & Sanders, T. (2017). Connective processing by bilingual children and monolinguals with specific language impairment: Distinct profiles. *Journal of Child Language*, 44(2), 329–345. https://doi.org/10.1017/s0305000915000860

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, 53(4), 372–380. https://doi.org/10.3758/bf03206780

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, 60(1), 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3), 283–312. https://doi.org/10.1006/jmla.1997.2543

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*, 33(1), 128–147. https://doi.org/10.1006/jmla.1994.1007

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. CRC Press.

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

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ···Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, 7, 1–24. https://doi.org/10.7554/eLife.33468

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

Porretta, V., Kyröläinen, A.-J., van Rij, J., & Järvikivi, J. (2018). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In I.

Czarnowski, R. Howlett, & L. Jain (Eds.), *Intelligent decision technologies 2017. Smart innovation, systems and technologies* (Vol. 73, pp. 268–277). Springer.

Pyykkönen, P., & Järvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology*, 57(1), 5–16. https://doi.org/10.1027/1618-3169/a000002

Pyykkönen-Klauck, P., & Crocker, M. W. (2016). Attention and eye movement metrics in visual world eye tracking. In P. Knoeferle, P. Pyykkönen-Klauck, & M. W. Crocker (Eds.), *Visually situated language comprehension* (pp. 67–82). John Benjamins Publishing.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, 85(3), 618–660. https://doi.org/10.1037/0033-2909.85.3.618

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition*, 89(1), B1–B13. https://doi.org/10.1016/S0010-0277(03)00065-9

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, 137(2), 172–180. https://doi.org/10.1016/j.actpsy.2010.09.010

Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 89–110). Wiley-Blackwell.

Salverda, A. P., & Tanenhaus, M. K. (2018). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 89–110). Wiley-Blackwell.

Saryazdi, R., & Chambers, C. G. (2021). Real-time communicative perspective taking in younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(3), 439–454. https://doi.org/10.1037/xlm0000935

Saslow, M. G. (1967). Latency of saccadic eye movement. *Journal of the Optical Society of America*, 57(8), 1030–1033. https://doi.org/10.2466/pms.2003.96.1.173

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: Using the bootstrapped differences of timeseries (BDOTS) to analyze visual world paradigm data (and more). *Journal of Memory and Language*, 102, 55–67. https://doi.org/10.1016/j.jml.2018.05.004

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

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, 42(3), 423–443. https://doi.org/10.1006/jmla.1999.2691

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

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29(6), 557–580. https://doi.org/10.1023/a:1026464108329

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.1632

Traxler, M. J., Bybee, M. D., & Pickering, M. J. (1997). Influence of connectives on language comprehension: Eye tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology*, 50A(3), 481–497. https://doi.org/10.1080/027249897391982

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

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1), 1–25. https://doi.org/10.1016/S0749-596x(03)00105-0

Wei, Y., Mak, W. M., Evers-Vermeul, J., & Sanders, T. J. M. (2019). Causal connectives as indicators of source information: Evidence from the visual world paradigm. *Acta Psychologica*, 198, 102866. https://doi.org/10.1016/j.actpsy.2019.102866

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