Week 3: Spoken Language Processing in a Visual Context

Yiling Huo

October 27, 2023

Research on eye movements in reading has a history of more than a century. In contrast, eye movements have only started to become a popular measure in studies of spoken language processing within the last couple of decades. In these studies, participants' eye movements to a visual display are recorded as they follow instructions, listen to sentences, or generate utterances about the "visual world". The visual world paradigm allows researchers to study real-time language comprehension and production in natural tasks. Today we will first talk about components of the visual world paradigm, before discussing (only a fraction of) the most classic or interesting research questions that it has been used to address.

1 The visual world paradigm

In a typical visual world experiment, the participants hear an utterance while looking at an experimental display, while their eye movements are recorded for later analyses.

1.1 The visual display

Typically, the visual display includes the object(s) mentioned in the utterance as well as a few distractors. The visual display can take the form of a semi-realistic scene, an array of objects, or even printed words. The visual display is typically presented 1-2 seconds before the onset of the utterance (preview time) and stays in view until the offset of the auditory stimuli. In some versions of the visual world paradigm, the visual display can be presented first, and a spoken sentence follows while a blank screen is shown. Such a setup is useful in the studies of short-term memory in language comprehension.

1.2 The task

The spoken utterances can be instructions to the participants ("Pick up the/Click on the…") or simply descriptions or comments on the visual display, which distinguish an action-based version of the visual world paradigm and its non-action-based counterpart. In the latter case, participants are usually instructed to look at the screen and to listen carefully to the sentences.

One advantage of the visual world paradigm compared to other psycholinguistic paradigms such as lexical decision or grammaticality judgments is that the listeners do not have to perform meta-linguistic judgments, which may be difficult in populations such as young children.

1.3 The linking hypothesis

Data collected in a visual world experiment is essentially the gaze position at particular time points in each trial. How to link these position data with language processing? (Because you may realise that eye movements are not as simple as "we always look at what's mentioned".) The assumption that provides the link between language processing and eye movements in the visual world is essentially that **the activation of a linguistic representation determines the probability that a participant will shift attention to the corresponding picture and thus make a saccadic eye movement to fixate it**. Therefore, when gaze positions are averaged across multiple trials, researchers can calculate the proportion/probability of looks to the target object, representing activation of the target lexical item.

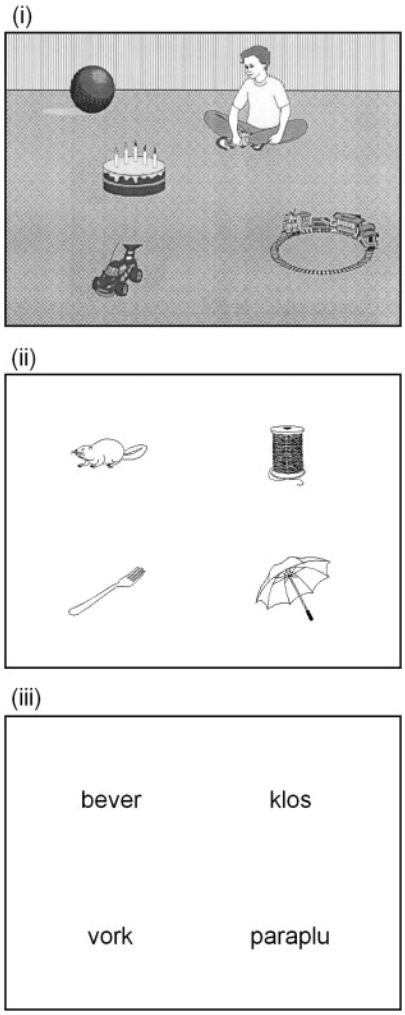

Figure 1: Typical visual world displays. Extract from [1].

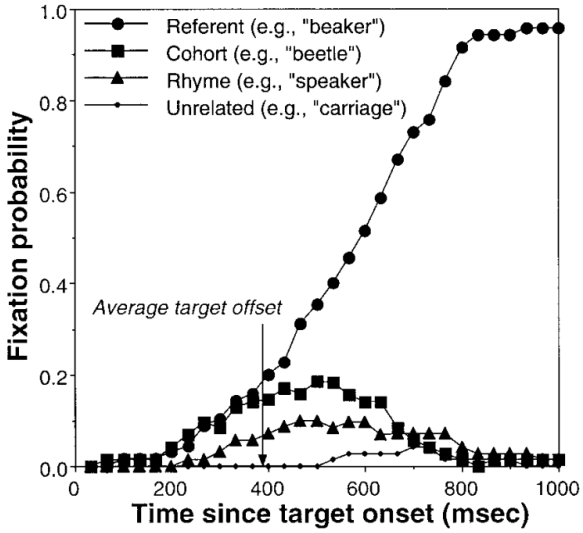

Figure 2: Proportion of looks to each object in the visual display when listening to instructions such as "Pick up the beaker". Extract from [2].

2 Word recognition in the visual world

Parallel activation during word recognition: Allopenna et al. [2]

Allopenna et al. [2] had participants follow spoken instructions to pick out objects shown on the screen (e.g. "Pick up the beaker."). Four objects were shown on the screen: the referent (beaker), a cohort (beetle), a rhyme (speaker), and an unrelated object (carriage). Allopenna et al. observed a (non-linear) rising curve for the probability of fixating on the referent, and a rising-then-falling curve for the probability of fixating on phonologically overlapping objects (the cohort and the rhyme). This provides evidence for a continuous lexical access model during spoken word recognition where all candidates that are temporarily consistent with the speech signal are activated before the speech signal provides enough information to identify the single correct lexical item.

3 Sentence processing in the visual world

3.1 Eye movements induced by sentence processing: Cooper [3]

One of the first classic studies of spoken language in the visual world was by Roger Cooper. Cooper [3] tracked participants' eye movements as they listened to stories while looking at a display of pictures. He found that participants initiated saccades to pictures that were named in the stories, as well as pictures that were associated with words in the story (Africa - lion, zebra, snake). Moreover, fixations were often generated before the end of the word. This provides important evidence that visual attention is highly correlated with spoken sentence processing.

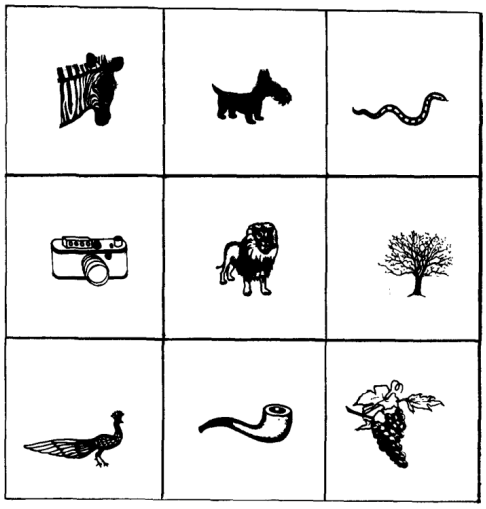

Figure 3: Example visual display in @[3]. Extract from [3].

3.2 Effects of the visual context: Tanenhaus et al. [4]

To fully understand any visual world experiment, we need to be aware that the visual display itself may have an effect on how the listeners interpret the sentence. Tanenhause et al. [4] is one of the most classic studies that demonstrate this. Tanenhaus and colleagues presented participants with sentences such as "Put the apple on the towel in the box", where the first prepositional phrase ("on the towel" in the example) is temporarily ambiguous between denoting the destination of the apple or its current location. In the one-referent condition of the experiment participants saw just one apple on a towel, an empty towel, a box, and a pencil. In the two-referent condition there were two apples: one on a towel and one on a napkin. In this condition, a modifier was needed to inform the listener which of the two apples should be moved. They found that there were significantly more early looks to the empty towel in the one-referent than in the two-referent condition. This is strong evidence that listeners can use visual information immediately to disambiguate sentence structures. Not only does this study tell us to be a bit careful about the visual display when designing a visual world experiment, it also shows that language processing is subject to a broad range of linguistic as well as non-linguistic constraints.

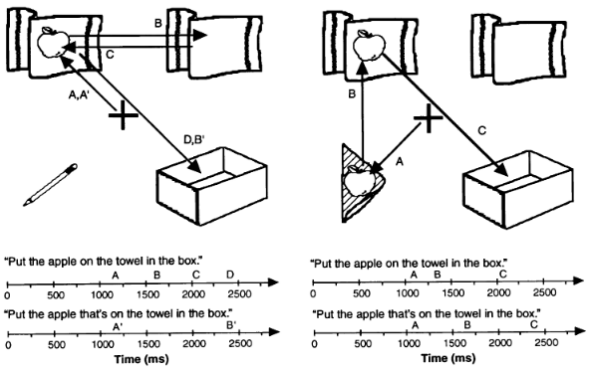

Figure 4: Typical sequence of eye movements in the two conditions of [4]. Extract from [4].

3.3 Syntactic ambiguities: Snedeker and Trueswell [5]

Last week we talked about syntactic ambiguities and the serial vs. parallel processing hypotheses of syntactic ambiguity. One issue in syntactic ambiguity is lexical bias: e.g. the verb *remember* tends to be followed by a direct object (*remembered the story*) while the verb *suspect* tends to be followed by a sentence complement (*He suspects the story is false.*).

Snedeker and Truswell [5] demonstrated this lexical bias in syntactic parsing using the visual world paradigm. Participants listened to sentences whose verb had either a modifier bias, an instrument bias, or neutral (e.g. Choose/Tickle/Feel the frog with the feather) while looking at visual displays of four objects: a target instrument (a feather), a target animal (a frog holding a feather), a distractor instrument (a candle), and a distractor animal (an animal holding a candle). In the one-referent condition, the distractor animal is different from the target animal (a leopard holding a candle) while in the two-referent condition, the distractor animal is the same as the target (a frog holding a candle).

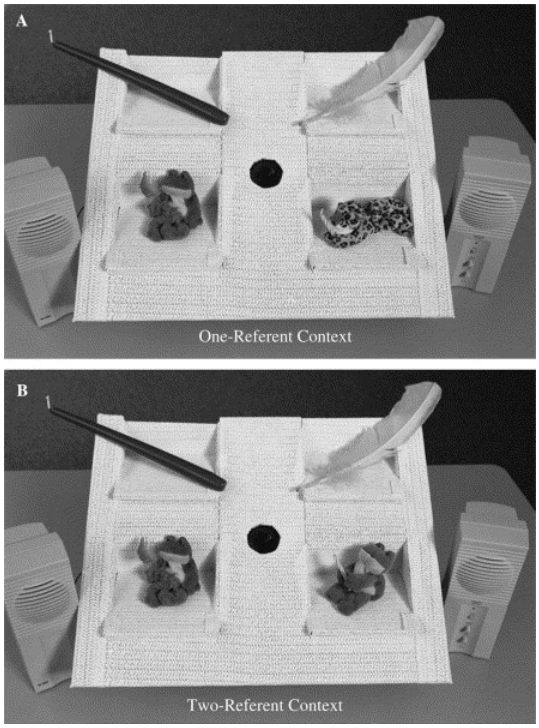

Figure 5: Sample visual display in [5]. Extract from [5].

Results showed that both the visual context and the lexical bias affected listeners' eye movements (in an additive

manner). One referent scenes, as compared with two referent scenes, increased measures of the instrument interpretation and decreased measures of the modifier interpretation. Likewise, as the tendency of the verb to appear with an instrument phrase increased, measures of an instrument interpretation increased and measures of a modifier interpretation decreased. On the issue of lexical bias in syntactic ambiguity, these results clearly show that lexical bias has an influence on the initial syntactic structure comprehenders build for ambiguous sentences. On top of this, these results also show that the visual context has as well an effect on the initial interpretation of these sentences.

3.4 Incrementality of sentence processing

Last week we covered some reading eye-tracking studies that addressed incrementality in sentence processing. A line of studies also addresses this using the visual world paradigm.

Altmann & Kamide [6]

Altmann and Kamide [6] presented listeners with visual displays showing, e.g., a boy, a cake, and some toys, while the listeners heard sentences such as "The boy will eat/move the cake.". Eye movements revealed that listeners were more likely to look at the target object (cake) prior to its onset when the verb was constraining (eat) than non-constraining (move).

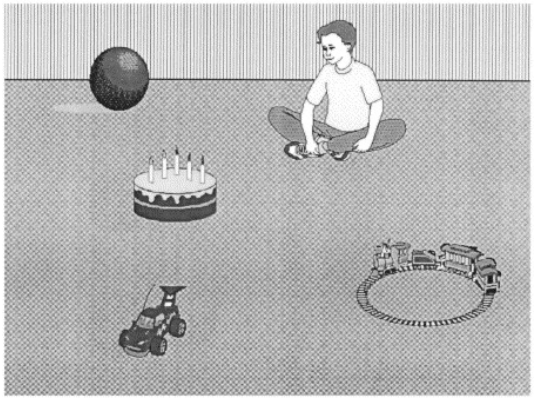

Figure 6: Sample visual display in [6]. Extract from [6].

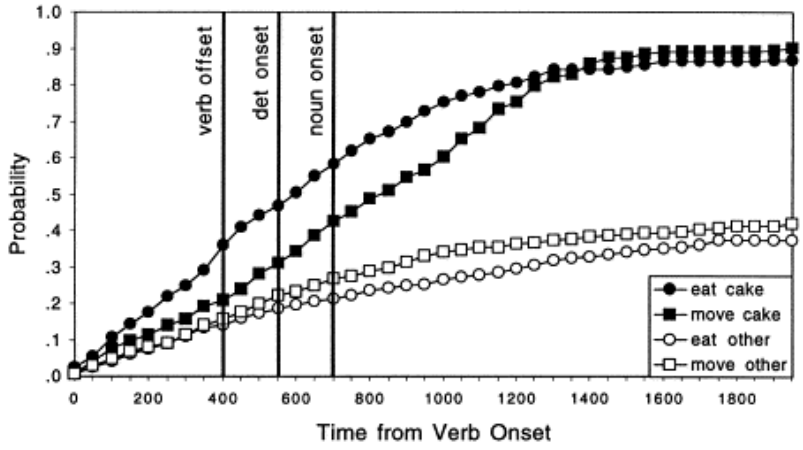

Figure 7: Proportion of looks to the target object (e.g. cake). Extract from [6].

This suggests that not only did listeners interpret the verb and its selectional information immediately after hearing it (incrementality), but they also used the selectional information in the verbs such as eat to actively anticipate what will be referred to next. This phenomenon is later known as prediction during language comprehension.

Similarly, Kamide, Altmann, and Haywood [7] explored whether verb information can be combined with information conveyed by their grammatical subject to drive anticipatory eye movements. They found increased fixations to a

motorbike when listeners heard sentences such as "The man will ride…" and increased fixations to a carousel when they heard "The girl will ride…". This shows that different sources of information can be efficiently combined on the fly during language comprehension to generate predictions of upcoming language constituents.

More on prediction

A good number of studies have investigated what types of information are involved in predictive processing during language comprehension. I will expand a bit on this (because this happens to be what I work on). A wide array of information including real-world knowledge as well as linguistic knowledge has been shown to contribute to predictive processing, including linguistic markers of tense [8], gender [9–11], case [12], number [13], and noun class [14,15]; certain phonological patterns that pose constraints on following syllables [16,17] or carry extralinguistic information [18]; as well as sentential and/or discourse context [6,19,20], and event knowledge [7,21].

For example, Altmann and Kamide [8] found that upon hearing "The man will drink…", listeners showed anticipatory looks to a full glass of beer; while upon hearing "The man has drunk…", listeners looked instead at an empty wine class.

Ito and Speer [18] showed that listeners were quicker to look to a referent when a contrastive pitch accent was congruous with the contrast ("Find the blue ball. Now, find the GREEN ball.") than when it was neutral or incongruous ("…the green ball/the green BALL.").

3.5 Pragmatic inferencing

3.5.1 When do we derive scalar implicatures? [22], [23]

One of the strengths of the visual world paradigm is that eye movements are constantly recorded, thus we do not need to rely on participants' responses to a target or critical word (i.e. we know what happens before and after the critical word). This means that with the visual world paradigm, we get insights into not only that something happens, but also when it happens.

In a series of studies that investigated the time-course of scalar implicatures, Huang and Snedeker [22], [23] asked participants to listen to utterances such as "Point to the girl that has some of the socks" while viewing a display in which one girl had two of four socks and another girl had three of three soccer balls (a phonological onset overlap competitor). The lexical semantics of "some" denote a quantity greater than one (i.e., some-and-possibly-all), but the word is usually interpreted with an 'upper boundary' (i.e., some-and-not-all). Linguistic theories argue that the meaning of "some" includes the meaning of "all" by default while the implication of "some" meaning "not all" is derived later as an inference. Indeed, eye movements showed that "some" is initially interpreted as compatible with "all" (looks to the girl with two socks did not exceed those to the girl with three/all soccer balls upon presentation of "some"), and participants only started to exclude referents compatible with "all" approximately 800ms later.

4 Speech production in the visual world

Visual world eye-tracking has been informative about speech production as well. For example, Griffin and Bock [24] tasked participants to describe cartoons of events and compared their eye movements to other tasks such as scene observation or patient detection. They found that although participants' initial eye movements (0-300ms) were similar among tasks, they quickly showed a distinct agent-patient pattern in the description task. The researchers concluded the existence of two distinct phases in the picture description task: an initial apprehension phase followed by a speech formulation phase when speakers look at each of the objects they name in the order of mention. Following this work, a line of research has investigated the time course of speech generation as well as the incremental interaction between visual information uptake and utterance generation.

5 Summary

Over the last couple of decades, the visual world paradigm has allowed us to gain much insight into spoken language processing as well as speech production. Although the visual world paradigm has its limitations (e.g. the visual contexts can affect language processing in several ways), it continues to be a popular tool for researchers interested in spoken language for its ability to assess the time course of language processing and its ability to address the interplay of language, vision, and attention.

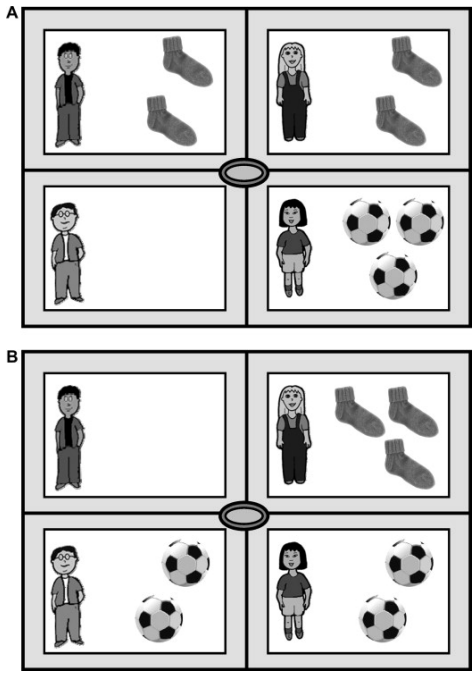

Figure 8: Sample visual displays in [22]. Participants listened to "Point to the girl that has two/some of the socks" in (A) and "…the girl that has three/all of the socks" in (B). Extract from [22].

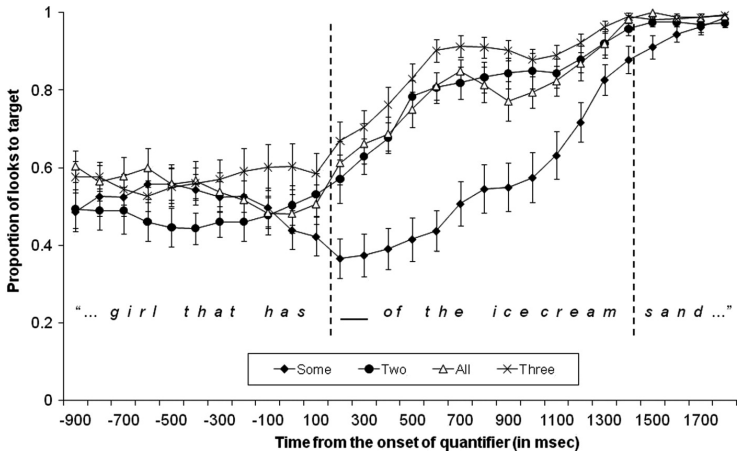

Figure 9: Results from [23]. Extract from [23].

Read more on this...

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.

Tanenhaus, M. K. (2007). Eye movements and spoken language processing. In *Eye Movements* (pp. 443-II). Elsevier.

References

[1] Huettig F, Rommers J, Meyer AS. Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica 2011;137:151–71. https://doi.org/10.1016/j.actpsy.2010.11.003.

[2] Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. Journal of Memory and Language 1998;38:419–39. https://doi.org/10.1006/jmla.1997.2558.

[3] Cooper RM. 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 1974.

[4] Tanenhaus MK, Spivey-knowlton MJ, Eberhard KM, Sedivy JC, Tanenhaus MK, Spivey-knowlton MJ, et al. Integration of Visual and Linguistic Information in Spoken Language Comprehension Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2888637 JSTOR is a not-for-profit service that helps scholars , r. Science 1995;268:1632–4.

[5] Snedeker J, Trueswell JC. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. Cognitive Psychology 2004;49:238–99.

[6] Altmann GT, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition 1999;73:247–64.

[7] Kamide Y, Altmann GT, Haywood SL. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language 2003;49:133–56.

[8] Altmann GT, Kamide Y. 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 2007;57:502–18.

[9] Lew-Williams C, Fernald A. Young children learning spanish make rapid use of grammatical gender in spoken word recognition. Psychological Science 2007;18:193–8.

[10] Lew-Williams C, Fernald A. Real-time processing of gender-marked articles by native and non-native spanish speakers. Journal of Memory and Language 2010;63:447–64.

[11] Stone K, Veríssimo J, Schad DJ, Oltrogge E, Vasishth S, Lago S. The interaction of grammatically distinct agreement dependencies in predictive processing. Language, Cognition and Neuroscience 2021;36:1159–79.

[12] Kamide Y, Scheepers C, Altmann GT. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from german and english. Journal of Psycholinguistic Research 2003;32:37–55.

[13] Lukyanenko C, Fisher C. Where are the cookies? Two-and three-year-olds use number-marked verbs to anticipate upcoming nouns. Cognition 2016;146:349–70.

[14] Kwon N, Sturt P, Liu P. Predicting semantic features in chinese: Evidence from ERPs. Cognition 2017;166:433–46.

[15] Chow W-Y, Chen D. Predicting (in) correctly: Listeners rapidly use unexpected information to revise their predictions. Language, Cognition and Neuroscience 2020;35:1149–61.

[16] Liu S, Chen X, Wang S. The involvement of phonological information during spoken language prediction: Evidence based on chinese tone sandhi. OSF Preprints; 2023.

[17] Chow W-Y, Huo Y. Listeners can use tone information to predict: Evidence from mandarin chinese. Poster presented at the 33rd annual CUNY Human Sentence Processing Conference; 2020.

[18] Ito K, Speer SR. Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language 2008;58:541–73.

[19] Otten M, Van Berkum JJ. Discourse-based word anticipation during language processing: Prediction or priming? Discourse Processes 2008;45:464–96.

[20] Van Berkum JJ, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition 2005;31:443.

[21] Chow W-Y, Smith C, Lau E, Phillips C. A "bag-of-arguments" mechanism for initial verb predictions. Language, Cognition and Neuroscience 2016;31:577–96.

[22] Huang YT, Snedeker J. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. Cognitive Psychology 2009;58:376–415.

[23] Huang YT, Snedeker J. Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. Language and Cognitive Processes 2011;26:1161–72.

[24] Griffin ZM, Bock K. What the eyes say about speaking. Psychological Science 2000;11:274–9.

[25] Tanenhaus MK. Eye movements and spoken language processing. Eye movements, Elsevier; 2007, p. 443–II.