Week 3: Spoken Language Processing in a Visual Context

Yiling Huo

October 24, 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.

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 auditory stimuli

1.3 The task

1.4 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? 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 word.

1.5 Production studies

In production studies, participants see sets of objects or cartoons of events or actions. No spoken input is presented, but instead the participants are asked to describe what they see. Researchers typically determine which objects are inspected, in which order they are inspected, and when they are inspected relative to the participants' speech output. This provides information about the ways speakers coordinate the generation of utterance plans with the overt articulation.

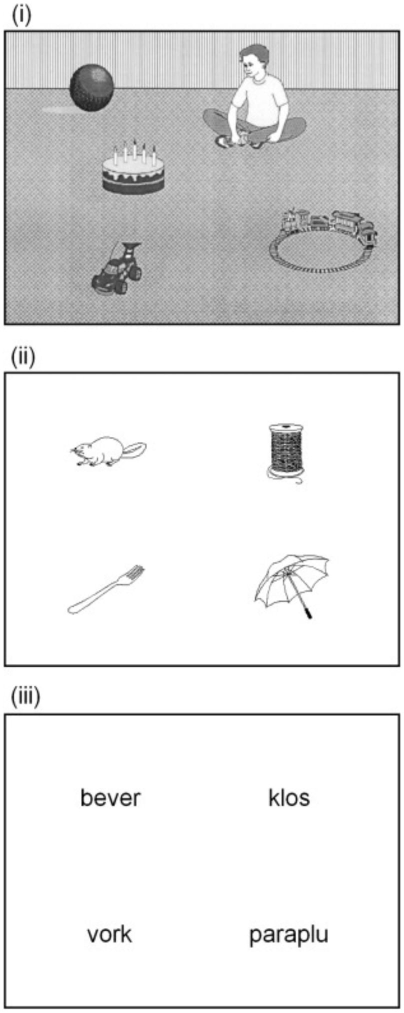

Figure 1: Typical visual world displays. Extract from [1].

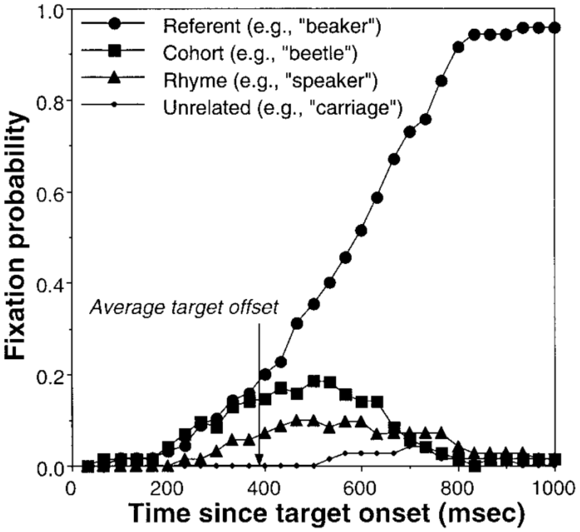

Figure 2: Proportion of looks to each object in the visual display when listening to instructions such as “Pick up the beaker”. Extract from [2].

2 Word recognition in the visual world

Parallel activation during word recognition: Allopenna et al. (1998) [2]

Allopenna et al. (1998) 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

Eye movements induced by sentence processing: Cooper (1974) [3]

One of the first classic studies of spoken language in the visual world was by Roger Cooper (1974). Cooper 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.

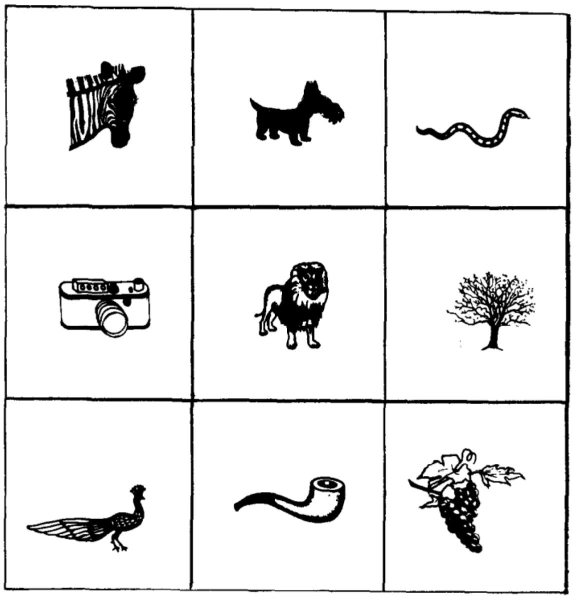

Figure 3: Example visual display in @[3]. Extract from [3].

Effects of the visual context: Tanenhaus et al. (1995) [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. (1995) 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.

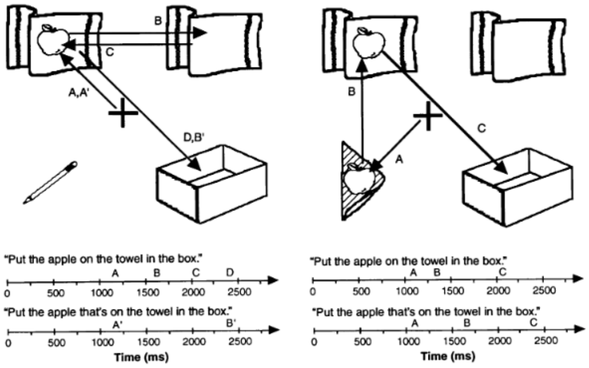
B
C
A,A'
D,B'
B
A
C
"Put the apple on the towel in the box."
A B C D
0 500 1000 1500 2000 2500
"Put the apple that's on the towel in the box."
A' B'
0 500 1000 1500 2000 2500
Time (ms)
"Put the apple on the towel in the box."
A B C
0 500 1000 1500 2000 2500
"Put the apple that's on the towel in the box."
A B C
0 500 1000 1500 2000 2500
Time (ms)

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

3.1 Syntactic ambiguities: Snedeker and Trueswell (2004) [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 (2004) 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).

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.

3.2 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.

3.2.1 Altmann & Kamide (1999) [6]

Altmann and Kamide (1999) 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).

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, [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

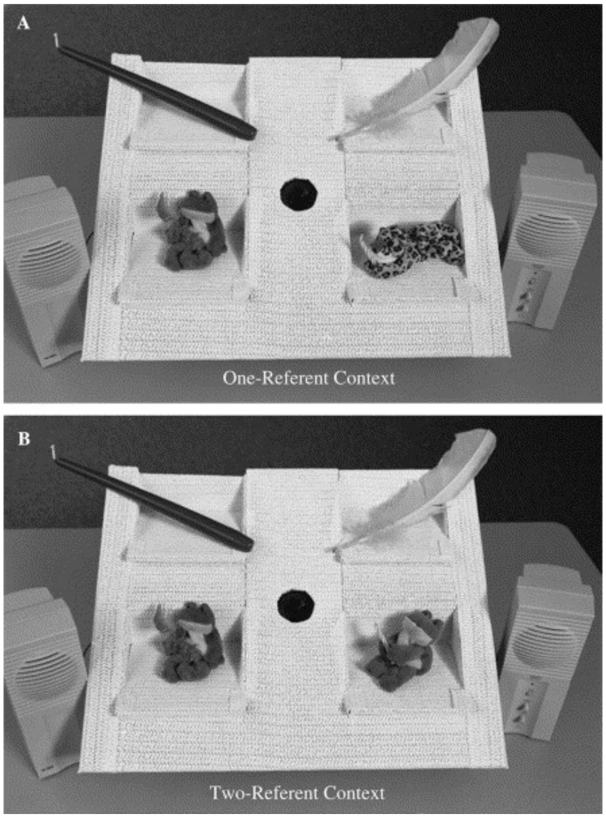

Figure 5: Sample visual display in [5]. Extract from [5].

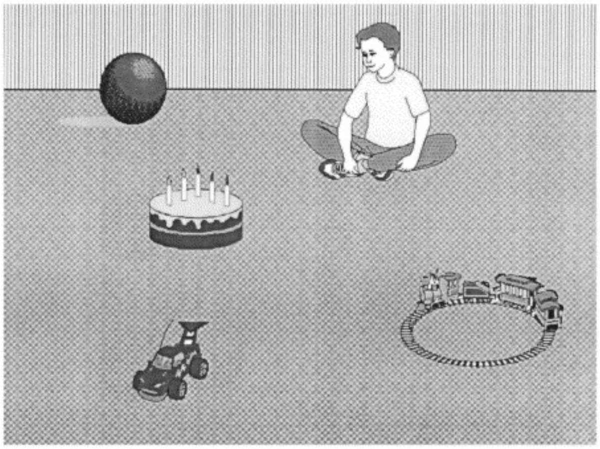

Figure 6: Sample visual display in [6]. Extract from [6].

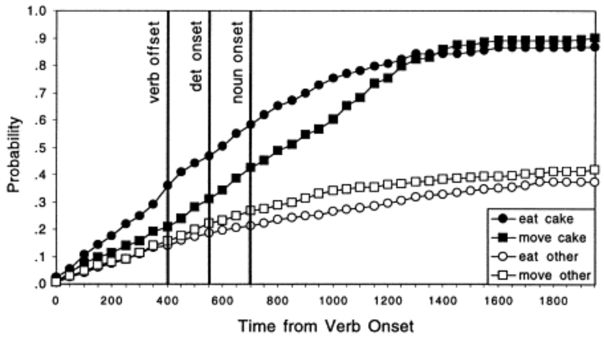

Figure 7: Proportion of looks to the target object (e.g. cake). Extract from [6].

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.

3.2.2 More on prediction

A good number of studies have investigated what types of information are involved in predictive processing during language comprehension.

@@@@@@

3.3 Pragmatic inferencing

4 Speech production in the visual world

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.

[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.

[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. Integration of visual and linguistic information in spoken language comprehension. 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] Tanenhaus MK. Chapter 20 - eye movements and spoken language processing. In: Van Gompel RPG, Fischer MH, Murray WS, Hill RL, editors. Eye movements, Oxford: Elsevier; 2007, p. 443–II. https://doi.org/Mhttps://doi.org/10.1016/B978-008044980-7/50022-7.